Controlled Vocabulary Enhancement through Crowdsourcing: Project Andvari, Micropasts, and Public Quality Assurance

YOUNGOK CHOI
The Catholic University of America

JOSEPH KOIVISTO[1]
University of Maryland College Park

Abstract: This paper presents an experimental approach of using crowdsourcing to test controlled vocabularies for digital collections of cultural objects. For a digital humanities initiative project, *Project Andvari*, which is intended to create a digital portal of early medieval northern European artifacts, it was recognized that there was a need to develop a semantically structured iconographic thesaurus to describe the iconographic content of distributed artefactual collections from a variety of contributing institutions. This paper discusses a workflow of planning and development process of controlled vocabularies for the project and a testing process of the vocabularies to determine both the usability of controlled vocabularies and the feasibility of quality assurance approach. This paper demonstrates an applicability of crowdsourcing in developing controlled vocabularies.

Introduction

Many museum collections of cultural objects have been transformed into digital collections as new channels of dissemination and wider access. For access and retrieval of objects in digital collections, textual description of objects is necessary[2] and such textual description needs to be meaningful and accessible to potential users' searching. However, providing detailed and descriptive textual description of cultural objects poses many challenges due to the non-textual nature of cultural artifacts. Visual elements and meanings of objects should be captured and translated into verbal expressions. This process is very difficult and complex as subject matters and interpretation of visual features are subjective and not straightforward.[3] On the other hand, it is relatively easy and direct to characterize descriptive features of objects like title, creator, or date in description. In order to guide this process, there are guidelines and tools for cataloging and indexing practice.[4] However, many professionals still find it difficult and challenging to apply such guidelines and tools to particular local or domain-specific collections.[5]

---

[1] The authors would like to thank many people who have been involved in *Project Andvari*. Those are Dr. Lilla Kopár, Dr. Nancy L. Wicker, Daniel Pett, Kirsten Mentzer, Katherine Kane, Jeremy Withnell, and Jenna Tenaglio.

[2] Murtha Baca. "Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information." *Cataloging & Classification Quarterly* 36, no. 3-4 (2003): 47-55.; Antonie Issac, Stefan Schlobach, Henk Matthezing, and Claus Zinn. "Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies", *Library Review* 57, no. 3 (2008): 187 – 199.

[3] Layna White. "Interpretation and representation: The who, why, what, and how of subject access in museums." *Art Documentation* 21, no. 1 (2002): 21-22.

[4] Murtha Baca. *Cataloging cultural objects: A guide to describing cultural works and their images.* American Library Association 2006.; Patricia Harpring. *Introduction to controlled vocabularies: terminology for art,*

Thus, textual description of objects often centers in providing factual information of objects like creator, medium, credit lines, etc.[6] This practice of textual description in object records seems problematic for users' search and access. Studies show that users tend to describe pictorial representation and visual meanings of artworks and such user's terms are not often available on object records.[7] These findings of cultural object records and users' description indicate that object records of digital collections are not sufficient or appealing to users and that they should contain a greater level of subject matter description. In other words, there is a need to have a dedicated knowledge organization system, like a thesaurus, to identify and describe content of objects in digital collections. When a digital repository is designed to support integrated and seamless access to cultural objects from multiple collections, such a thesaurus of controlled vocabularies is a must-have requirement.[8] The first step of developing a thesaurus is to gain insight on domain users' practice with resource use and interpretation of objects and invite them into the course of developing controlled vocabularies during a digital repository creation process.[9]

The purpose of this paper is to address the approach of a digital portal in ensuring capture of subject description of cultural objects and building a controlled vocabulary for subject description by utilizing crowdsourcing tools based on domain-specific scholarly practices. The digital portal addressed in this paper is *Project Andvari*, a digital humanities initiative to aggregate early medieval northern European artifacts. A main research question is what aspects of cultural objects signify in the construction of annotation and thus whether crowdsourcing helps identify terms to add to a list of controlled vocabularies.

The remainder of this paper is organized as follows: the *Project Andvari* Overview section discusses the project's background and its goal; the Methods section presents the design of the study and data collection and analysis; the Results section describes the findings from the analysis of data; and the Conclusion section summarizes the study.

Project Andvari Overview and Background

Founded in 2013, *Project Andvari* is a digital humanities initiative to create an aggregate research portal for distributed collections of material culture originating from the northern European periphery of the early medieval period (4th-12th Centuries Common Era) as a means of addressing limitations to research of materials of varied media, object types, and locations.[10] By creating a digital portal, *Project Andvari*

---

architecture, and other cultural works. Getty Publications, 2010. ; Sarah Shatford. "Analyzing the subject of a picture: A theoretical approach." *Cataloging & Classification Quarterly* 6, no. 3 (1986): 39-62.

[5] Alison Gilchrest. "Factors affecting controlled vocabulary usage in art museum information systems." *Art Documentation* 22, no. 1 (2003): 13-20.; Chandana Patra. "Digital repository in ceramics: a metadata study." *The Electronic Library* 26, no. 4 (2008): 561-581.

[6] Martha Kellogg Smith. "Viewer tagging in art museums: Comparisons to concepts and vocabularies of art museum visitors." *Advances in Classification Research Online* 17, no. 1 (2006): 1-19.

[7] Jennifer Trant, "Exploring the potential for social tagging and folksonomy in art museums: Proof of concept." *New Review of Hypermedia and Multimedia* 12, no. 1 (2006): 83-105.; Jennifer Trant, and Bruce Wyman. "Investigating social tagging and folksonomy in art museums with steve. museum." In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*. 2006.

[8] Issac, Schlobach, Matthezing, and Zinn, "Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies"

[9] National Information Standards Organization Standards (NISO). *ANSI/NISO Z39.19—Guidelines for the construction, format, and management of monolingual controlled vocabularies*. 2005. Retrieved on November 23, 2015, from http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a

[10] "About Project Andvari," Retrieved on November 24, 2015 from http://www.andvari.org/index.php

was intended to allow humanities scholars to study the aggregated material in an interdisciplinary fashion, promoting analyses of relationships (both of artifacts and of cultures) hitherto unrecognized; and to promptly disseminate information about new archaeological discoveries and provide a shared virtual workspace for researchers to examine new finds.[11]

Although contemporary research into the Vikings and their predecessors has been supported by the recent availability of digital data sets and museum collections from the Swedish National Heritage Board's (Riksantikvarieämbetet) *Kringla* platform[12] and institutions such as the British Museum,[13] the project directors found that the inability to perform metasearches across collections — many of which are arranged based on idiosyncratic methodologies based on national and scholarly traditions — proved a serious impediment to novel research approaches.[14] Thus, the goal of the project was to create a centralized platform through which researchers can access the metadata records and digital surrogates of numerous collections from around the world.

In order to examine the scholars' information needs and the portal's object record design, the project team convened two planning workshops with funding from the NEH Office of Digital Humanities in the form of a Digital Start-Up Grant. Workshops were held in November 2013 and 2014 at the Catholic University of America in Washington, DC and was attended by representatives from the British Museum, Portable Antiquities Scheme, Swedish National Heritage Board, Norwich Castle Museum & Art Gallery, Royal Commission on the Ancient and Historical Monuments of Scotland, the Institute for Advanced Technologies in the Humanities (IATH) at the University of Virginia[15] as well as representatives from international universities in Norway, Sweden, and Germany. The initial project team consists of numerous researchers and practitioners from a wide array of scholarly domains including medieval studies, art history, archaeology, information technologies, museum administration, and library and information science.

During the workshops, discussion revealed that enhanced metadata in the form of semantically-structured description of iconographic content was a highly desirable platform feature for medieval research specialists. The implementation of such a semantically-structured vocabulary would not only establish a hierarchy of conceptual relationships and terminologies for the complex iconographic language of the identified period — promoting intellectual access to iconographic content for both seasoned and inexperienced researchers,[16] but would also help to further align access to digital surrogates for a wide array of heterogeneous objects that make up the collections of contributing institutions, objects such as runestones, decorative objects, and other artifacts. Furthermore, implementation of a multilingual controlled vocabulary would further facilitate international access to the collection by initially aligning iconographic description in English, German, and Swedish — with the potential inclusion of additional languages in the future.

However, this initial agreement was tempered by concern over the availability of extant resources that would both appropriately describe the iconographic content of the identified material set based on the

---

[11] Ibid.

[12] "Kringla," Retrieved on November 24, 2015 from http://www.kringla.nu/kringla/

[13] "The British Museum: Collection Online," Retrieved on November 24, 2015 from https://www.britishmuseum.org/research/collection_online/search.aspx

[14] Lilla Kopár, Nancy L. Wicker, and Joseph Koivisto. *Project Andvari: A digital portal to the visual world of early medieval northern Europe*. Washington, D.C.: National Endowment for the Humanities, 2015. Retrieved from https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51640-13

[15] http://www.iath.virginia.edu/

[16] Baca, "Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information."

practices of the medieval studies discipline and structure concepts in Linked Open Data (LOD) - compliant encoding protocols. Preliminary discussions revealed that medieval researchers and museum practitioners were disappointed by existing digital authority resources such as ICONCLASS[17], the Library of Congress Subject Headings[18], and the Index of Christian Art[19] as the vast majority were created with the intended purpose of describing and structuring Christian artwork and iconography. This cast doubts on resource applicability to the domain-specific intent of the project platform as the methodological and conceptual foundation of their development differed considerably from the best practices adhered to by researchers and museum specialists when dealing with pre-Christian iconographical works.

Furthermore, existing resources frequently did not feature concepts specific to the mythological or decorative content of pre-Christian iconographic artwork of northern Europe — such as the absence of particular character names from the otherwise useful Name Authority File[20] — thereby highlighting conceptual gaps that would hinder effective description of the identified object set. This, coupled with the stated need to provide description at the non-interpretive, pre-iconographical level[21] in order to ensure the greatest possible level of access to objects, presented an issue to the project team: how can a digital humanities project such as *Andvari* promote intellectual access in a manner that may not be supported by currently available controlled vocabulary resources?

At the initial workshop, scholars and professionals agreed and confirmed the need for a domain specific thesaurus to address the idiosyncratic nature of the pre-Christian iconography of the medieval north. The subsequent project workshop produced an initial word list of general topics that were necessary in order to describe the iconographic content of the identified subject matter. Furthermore, the workshop led to a partnership with representatives from the Micropasts project[22] — a crowdsourcing initiative sponsored by the British Museum — that would ultimately lead to the development of the *Project Andvari* thesaurus crowdsourcing application. As a result, collaboration on developing a simple, web-based interface in which to test the usability of our initial thesaurus concepts began shortly after the conclusion of the second project workshop.

Method

Using the initial word list developed during the 2014 workshop, a draft thesaurus structure was encoded in the Simple Knowledge Organization Systems (SKOS) standard model using the Stanford *Protégé* application, an open source ontology authoring platform.[23] With the initial semantic hierarchy developed for the thesaurus, RDF structured output was forwarded to a partner at the British Museum and the Micropasts team who used the PyBossa-based[24] backend infrastructure of the *Micropasts* web application to develop a user interface for the proposed Andvari crowdsourcing task. In addition to the RDF data, the Micropasts team was provided with a set of over 250 records of representative objects from the online holdings of the British Museum and the Swedish National Heritage Board's *Kringla* interface, consisting of item URIs, URLs from online collections, and short titles to be displayed during tasks. Items were

---

[17] http://www.iconclass.org/
[18] http://id.loc.gov/authorities/subjects.html
[19] http://ica.princeton.edu/
[20] http://id.loc.gov/authorities/names.html
[21] Roelof van Stratten, *An introduction to iconography: Symbols, allusions and meaning in the visual arts*. New York, NY: Taylor & Francis, 1994.
[22] micropasts.org
[23] http://protege.stanford.edu/
[24] http://pybossa.com/

selected according to three criteria that align with the conceptual focus of the Project Andvari initiative: creation dates determined to be during the early medieval era; culture of origin was identified as the Viking north; and the inclusion of sufficient artistic or iconographic content to allow for description using the thesaurus concepts provided.

Development of the interface was completed through the use of a GitHub repository, allowing for collaboration between teams located in both the UK and Washington, DC. Unique tasks were developed for each of the items included in the object list along with a brief tutorial explaining to users how tasks were to be completed. In each task, users were shown an object image displayed within a small frame (Figure 1). Alongside the image, users were provided with a set of fields that prompted them to enter or select descriptive terms based on six semantic hierarchies – *Abstract, Built environment, Figure, Natural world, Object type, and Subject matter* [25]– that corresponded to the top concepts of the draft thesaurus. Users were also asked to provide additional terms that they believed to provide meaningful description of the selected objects as well as any comments they had regarding the object, the task, or the general usability of the supplied terms. All completed task data was made immediately available through the Micropasts's Data Centre[26] in bulk CSV and JSON data formats.

The final version of the *Project Andvari* crowdsourcing web application[27] was made available in March 2015 at http://crowdsourced.micropasts.org/project/andvari/. Information on the application was circulated through various social media outlets. For purposes of this research, data on task runs was collected at the end of June, 2015, allowing for four months of user activity to accrue prior to assessment. Data was collected using the csv output available through the Data Centre.

---

[25] *Abstract:* Iconographic content that can be described as either decorative elements and patterns or as style markers that point to an artistic style as established by scholarship on the art and artistry of early medieval northern European culture; *Built environment:* Iconographic content depicting man-made structures; *Figure:* Iconographic content depicting human figures or specific elements of the human body; *Natural world:* Iconographic content related to flora, fauna, celestial bodies, or natural elements (i.e. fire, air); *Object type:* Iconographic content depicting manufactured objects such as vehicles, weapons, clothing, &c; *Subject matter:* Iconographic content depicting identified historical or mythological figures.

[26] http://micropasts.org/data-centre/

[27] Pett, Daniel and Joseph Koivisto. "projectAndvariLOD: Initial Release." Zenodo. doi: 10.5281/zenodo.33171. Retrieved from https://zenodo.org/record/33171#.VlR783arS02
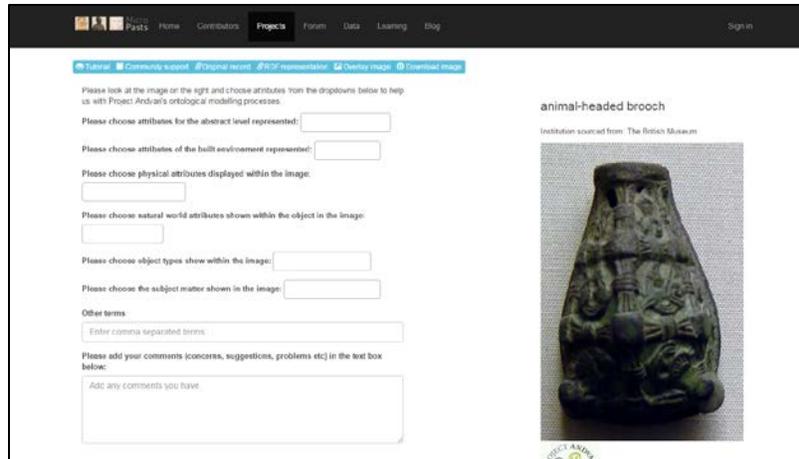
Figure 1. Sample Task on the Project Andvari crowdsourcing application, developed and hosted by the Micropasts initiative (micropasts.org)

Results

The total number of sample object records annotated by users was 426. From each record, we collected selected terms in each facet for analysis. In collecting data from the *Micropasts* server, there was a technical glitch in capturing terms selected from the *Built Environment* field, rendering this data subset irrecoverable. Thus, our analysis was not able to include data for *Built Environment* field while all other fields contained terms selected.

Figure 2 shows that among six categories, *Abstract* is the one which users most frequently selected terms to annotate the object. On the other hand, more than 60% participants did not select terms to annotate a figural attribute of an object from a list of vocabularies. There are two possible reasons of this lack of *Figure* facet term selection: one is that an object may not contain any figural attribute, the other might be a lack of sufficient vocabulary concepts to select.
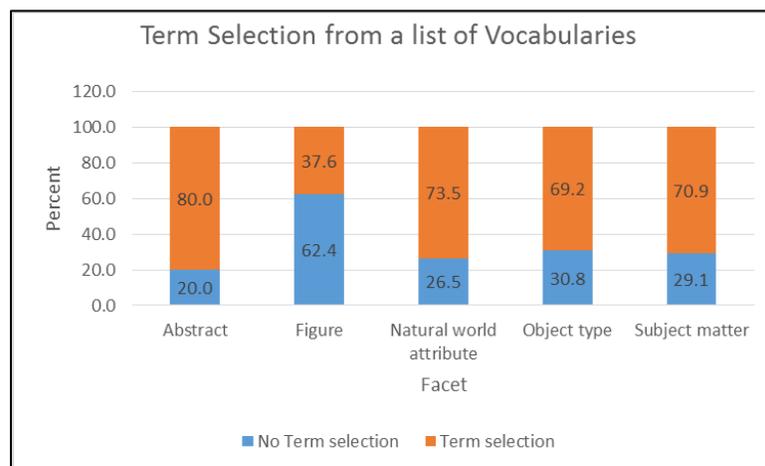

Figure 2. Percentage of Term Selection among Semantic Categories.

The total number of tags generated by users' own words was 312 terms for 170 items, an average of 1.9 terms per item. To gain insight into what object aspects and details users described using terms not found in the initial controlled vocabularies, we performed content analysis of tags based on conceptual facets. For the analysis of this comparison, a graduate student of the Department of Library and Information Science was hired as a research assistant and trained to analyze this data. Once the analysis was completed, one of the authors examined the results from a random sampling (30% of the records where users suggested additional tags) to check reliability. The intercoder reliability was 0.8 according to Holsti's (1969) reliability formula.

Results of the analysis of user-generated tags show that about a half of user-generated terms are categorized into the *Abstract* (about 26%) or *Built Environment* (22.4%) facets, followed by *Object Type (17.6%),* followed by *Subject matter* (13.4%), and *Natural world* (12.5%) (See Figure 3). A few users' comments also were made suggesting additions of object type terms into a list of controlled vocabularies. This finding suggests that users tend to annotate artifacts from a point of pictorial representation (*Pre-iconographical description)*, and a thesaurus will need to expand vocabularies for such descriptions.
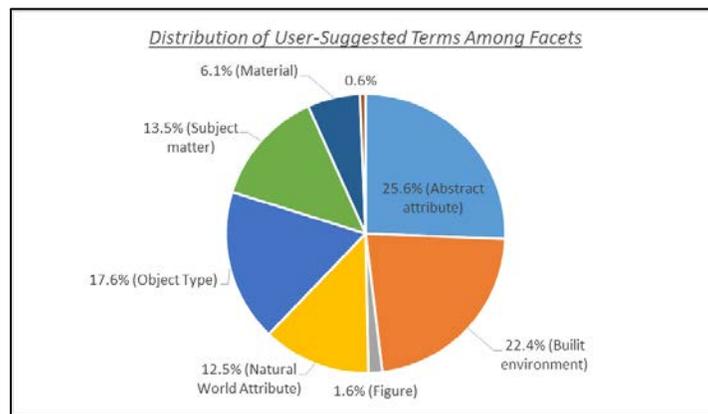

Figure 3. Distribution of User-Suggested Terms among Semantic Facets.

This finding is somewhat aligned with previous findings collected by the Steve Museum project. Results of tagging experiments in museum artworks show that art viewers provide descriptions of visible pictorial element and visual/abstract concepts[28]. This indicates that in representing cultural objects, not only direct physical characteristics of an object but also interpretation and symbolic meanings of an object are very important[29]. A thesaurus of controlled vocabularies should contain various terms for both aspects within a domain of interest.

As shown in Figures 2 and 3, the least selected and tagged conceptual facet is *Figure* and the most frequently selected and tagged conceptual facet is *Abstract.* Such consistency of term selection and additional tags from users among facets is very interesting to note. It is possible to conclude that users seemed less interested in identifying figural elements of iconographic content while simultaneously harboring a deep interest in the description of abstract and symbolic aspects of these cultural artifacts.

---

[28] Trant and Wyman. "Investigating social tagging and folksonomy in art museums with steve.museum."

[29] Muh-Chyun Tang. "Representational practices in digital museums: A case study of the National Digital Museum Project of Taiwan." *The International information & library review* 37, no. 1 (2005): 51-60.

To observe tags' potential as a source for indexing terms to retrieve relevant resources, we compared the tag terms to corresponding keywords in the resource description of the object tagged. User-generated tags totaled 312 and were annotated to 170 items out of 426 (39.9%). As each item has a link to the original metadata description by an individual hosting institution, users' tags were compared to original metadata records to examine in which metadata fields users' tags appeared. In comparing users' tags to metadata descriptions, we excluded metadata description written in languages other than English (48 items' descriptions), leading to 122 records. When terms were mapped to keywords in metadata descriptions, they appeared mostly in fields such as *Description* (36.2%, 113 out of 312 terms), *Title* (18.9%, 59 terms), *Object Type* (12.8%, 40 terms), *Materials* (3.5%, 11 terms), and the other fields including *Technique*, *Subject*, and *Inscription*. This finding may suggest that a description field of museums' metadata description provides rich additional information about objects with natural keywords. Based on these findings, description fields should both be indexed for retrieval and be made available for advanced searching options in digital collections.

Additionally, the alignment of user-generated concept terms with object metadata records indicates a dedicated focus on the description of primary physical characteristics for object classification. Even when tasked with identifying concepts and descriptive terms related to the iconographic content of objects, the tendency to focus on the general description, object types, materials, and techniques of creation reveals that iconographic description and considerations may be secondary to searches focused along lines of object and material type. It is possible to conclude that while the inclusion of semantically-structured iconographic description is of value to narrowly-defined investigation of collections, it is of equal importance to provide structured descriptions of physical object attributes through novel or existing thesauri such as the Portable Antiquities Scheme object type classifications[30] or the Getty Art & Architecture Thesaurus.[31]

Conclusion

In providing access to cultural objects in digital collections or repositories, a controlled vocabulary is an important tool to describe content and support information search and retrieval. In order to support essential functionality, a controlled vocabulary must be sufficient and appropriate to represent the content of the documents and accommodate semantic terms from the intended users. This concept is guided as a principle, *User Warrant*, in designing a thesaurus of controlled vocabularies.[32]

The approach described in this paper demonstrates how a design team of a digital portal was able to follow the core principle of developing controlled vocabularies working with the intended group of people who would use the controlled vocabularies to search for and describe cultural objects. In this process, a new approach of crowdsourcing and tagging was utilized as it was useful in selecting and testing which semantic terms and tags could be a source for new terms to describe existing concepts that should be considered for inclusion as preferred or nonpreferred terms in extant vocabulary resources.[33] Trant also points out that tags contain additional information that could be used to enlarge the structured

---

[30] https://finds.org.uk/datalabs/terminology/objects
[31] http://www.getty.edu/research/tools/vocabularies/aat/
[32] Harping, Introduction to Controlled Vocabularies; National Information Standards Organization Standards (NISO). ANSI/NISO Z39.19—Guidelines for the construction, format, and management of monolingual controlled vocabularies.
[33] Isabella Peters. Folksonomies: indexing and retrieval in Web 2.0. Berlin and Boston: K. G. Saur. 2009

vocabulary used by the professional museum experts for characterizing the subject domain of digital cultural heritage resources.[34] Findings described in this paper validate the argument of tagging's potential to enhance terms in controlled vocabularies. The paper also proves that by embracing a notion of collaborative participatory practice of crowdsourcing, professionals can design controlled vocabularies to serve intended users in digital collections and repositories.

Results of our study show that potential users would be interested in abstract aspects of cultural objects beyond pictorial features of objects. This means that indexing of cultural objects should capture abstract and symbolic elements of iconographic content and a proposed domain-specific thesaurus will need to be expanded in order to provide adequate conceptual coverage. In addition, tagging would be a useful feature to add to the digital collection in order to gather potential natural keywords from users. These findings will be incorporated in the future design of the proposed *Andvari* portal and applicable project-centric controlled vocabularies.

Resources

"About Project Andvari," Retrieved on November 24, 2015 from http://www.andvari.org/index.php

Baca, Murtha. *Cataloging cultural objects: A guide to describing cultural works and their images*. American Library Association, 2006.

Baca, Murtha. "Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information." *Cataloging & Classification Quarterly* 36, no. 3-4 (2003): 47-55.

Gilchrest, Alison. "Factors affecting controlled vocabulary usage in art museum information systems." *Art Documentation* 22, no. 1 (2003): 13-20.

Harpring, Patricia. *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications, 2010.

Holsti, Ole R.. *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley, 1969.

Isaac, Antonie, Stefan Schlobach, Henk Matthezing, and Claus Zinn, (2008),"Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies", *Library Review*, 57 no. 3 (2008): 187 – 199.

Kopár, Lilla, Nancy L. Wicker, and Joseph Koivisto. *Project Andvari: A digital portal to the visual world of early medieval northern Europe*. Washington, D.C.: National Endowment for the Humanities, 2015. Retrieved from https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51640-13

National Information Standards Organization Standards (NISO). *ANSI/NISO Z39.19—Guidelines for the construction, format, and management of monolingual controlled vocabularies*. 2005. Retrieved on November 23, 2015, from http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a

---

[34] Trant, Jennifer, "Exploring the potential for social tagging and folksonomy in art museums: Proof of concept."

Patra, Chandana. "Digital repository in ceramics: a metadata study." *The Electronic Library* 26, no. 4 (2008): 561-581.

Peters, Isabella. *Folksonomies: indexing and retrieval in Web 2.0*. Berlin and Boston: K. G. Saur. 2009.

Pett, Daniel and Joseph Koivisto. "projectAndvariLOD: Initial Release." Zenodo. doi: 10.5281/zenodo.33171. Retrieved from https://zenodo.org/record/33171#.VlR783arS02.

Shatford, Sarah. "Analyzing the subject of a picture: A theoretical approach." *Cataloging & Classification Quarterly* 6, no. 3 (1986): 39-62.

Smith, Martha Kellogg. "Viewer tagging in art museums: Comparisons to concepts and vocabularies of art museum visitors." *Advances in Classification Research Online* 17, no. 1 (2006): 1-19.

Tang, Muh-Chyun. "Representational practices in digital museums: A case study of the National Digital Museum Project of Taiwan." *The International information & library review* 37, no. 1 (2005): 51-60.

Trant, Jennifer, "Exploring the potential for social tagging and folksonomy in art museums: Proof of concept." *New Review of Hypermedia and Multimedia* 12, no. 1 (2006): 83-105.

Trant, Jennifer, and Bruce Wyman. "Investigating social tagging and folksonomy in art museums with steve. museum." In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*. 2006.

van Stratten, Roelof. *An introduction to iconography: Symbols, allusions and meaning in the visual arts*. New York, NY: Taylor & Francis, 1994.

White, Layna. "Interpretation and representation: The who, why, what, and how of subject access in museums." *Art Documentation* 21, no. 1 (2002): 21-22.