Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples†

Marcia Lei Zeng*, Karen F. Gracy**, and Maja Žumer***

*/**School of Library and Information Science, Kent State University, P.O. Box 5190, Kent, OH 44242, *<mzeng@kent.edu>, **<kgracy@kent.edu>
***Department of Library and Information Science and Book Studies, University of Ljubljana, Slovenia, Aškerceva 2 1000 Ljubljana, Slovenija, <Maja.Zumer@ff.uni-lj.si>

Marcia Lei Zeng is a Professor of Library and Information Science at Kent State University. She holds a Ph.D. from the School of Information Sciences at University of Pittsburgh (USA) and M.A. from Wuhan University (China). She has been involved in the development and research of knowledge organization systems (KOS) for over 25 years and contributed to the related standards including NISO Z39.19 and ISO 25964 for structured vocabularies. Her major research interests include knowledge organization systems (taxonomy, thesaurus, ontology, etc.), linked data, metadata and markup languages, database quality control, multilingual and multiculture information processing, and digital collections for cultural objects.

Karen F. Gracy is an associate professor of library and information science at Kent State University. She possesses an MLIS and a Ph.D. in Library and Information Science from the University of California, Los Angeles and an M.A. in critical studies of film and television from UCLA. Her scholarly interests are found within the domain of cultural heritage stewardship, which encompasses a broad range of activities such as preservation and conservation processes and practices, digital curation activities that consider the roles of heritage professionals and users in the lifecycle of objects and records, as well as knowledge representation activities such as definitions of knowledge domains, development of standards for description, and application of new technologies to improve access to cultural heritage objects.

Maja Žumer is professor of library and information science at University of Ljubljana (Slovenia). Her M.L.S. degree from Kent State University (USA) in 1993 was a turning point in her career: she became first a systems librarian, then head of Research and Development Department at the National and University Library in Ljubljana. After completing her Ph.D. in information science in 1999 (University of Zagreb, Croatia), she joined the faculty of Department for Library and Information Science and Book Studies at University of Ljubljana. Her research interests include design and evaluation of information retrieval systems, end-user interfaces, usability, and, recently, conceptual modelling. She has been involved in IFLA working groups and several EU projects.

Zeng, Marcia Lei, Gracy, Karen F., and Žumer, Maja. **Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples.** *Knowledge Organization. 41(6)*, 440-451. 22 references.

Abstract: The problem addressed by this study is the assessment of alternative approaches of generating subject access points to materials that are usually not made available through regular library catalog routines. As an aid in understanding how computerized subject analysis might be approached, we suggest using the three-layer framework that has been accepted and applied in image analysis. The hypothesis is that the computer-assisted semantic analysis has great potential in generating subject access at the "description" and "identification" levels. Two research samples were used to analyze the access points supplied by the OpenCalais semantic analysis tool. The first sample includes 43 archival record groups from 16 institutions, including university archives, government records archives, and manuscript/special collections repositories in various LAMs. The analysis resulted in dozens and, at times, hundreds of potential entities and social tags that could be used to provide additional points of entry to

these archival records. These entities and tags correspond almost exclusively to the first two layers of subject analysis (description and identification). The second sample contained 44 philosophy theses. In this part of the research, it was found that the semantic analysis based on the abstracts generated more successful tags than those based on the titles. The research based on the two samples indicate these subject access points fall at the "description" (referring to the generic elements depicted in or by the work) and "identification" (referring to the specific subject) levels, rather than the "interpretation" (referring to the meaning or themes represented by the subjects and including a conceptual analysis of what the work is about) level.

† The authors would like to thank research assistants Sammy Davidson and Laurence Skirvin of Kent State University for assisting with OpenCalais-related processes. This research is a sub-project of Metadata-Vocabulary Junction Project (http://lod-lam.slis.kent.edu), which was funded by the Institute of Museum and Library Services (IMLS) National Leadership Grants for Libraries program (2011-2013).

Received: 28 July 2014; Revised: 7 November 2014; Accepted: 11 November 2014

Keywords: semantic analysis, subject access points, terms, tags, Panofsky's theory

1.0 Introduction: the research question

The problem addressed by this study is the assessment of alternative approaches of generating subject access points to the materials that are usually not made available through regular library catalog routines. Subject access is critical for both individual digital collections and cross-institutional digital libraries, such as Europeana, which hold and provide access to a variety of information resources provided by libraries, archives, and museums (LAMs). LAMs have invested huge amounts of human resources in subject analysis, as evidenced in catalogs, finding aids, indexes, etc. As the size and variety of accessible open resources grow exponentially, LAMs are recognizing the impracticality and impossibility of conducting exhaustive traditional subject analysis. Yet, without providing good quality subject access, LAMs will find that users' search requests cannot often be satisfied. Limited subject access points are particularly critical with very large-scale resources of cross-institutional collections.

Using computerized subject analysis may prove to be promising in improving subject access to large heterogeneous collections. For example, advanced technologies in natural language processing and semantic annotation have resulted in enhanced, software-suggested access points (both named entities and topics) and even relations of the contents of a given resource. The following figure (Figure 1) is a screenshot showing manual and automatic subject analysis results.

On the left is an original doctoral dissertation's metadata, including seven keywords and two standardized subject headings entered by the dissertation author in the process of submitting to the electronic thesis and dissertation (ETD) repository. On the right is about 1/3 of the returned result after running the abstract of the dissertation through the semantic analysis tool OpenCalais (free version). The online software also displays the relevance ranking and count for each suggested tag (which the Calais called "social tag"). The processes of obtaining the original text, running it through the analysis, converting the resulting output into a database, cleaning up the data, and reconciliation can all be automated via a set of programs. Some portions needing judgment (e.g., merging synonyms, selecting preferred labels, or judging the appropriateness of a tag or entity name) would need either human assessment or further automatic processing. This sounds very promising. But what kinds of "subject" matters can such tools identify? Are they applicable to assist in subject analysis and indexing, or even be used as a primary solution to enhance subject access for existing resources?

2.0 Review of related literature

The Cranfield project is considered the first systematic evaluation in information retrieval systems. Led by Cyril Cleverdon, it lasted for about ten years (from 1957) and focused on the effectiveness of different indexing languages and methods. The project set the stage for further research in information retrieval—and established subject access as the central topic (Cleverdon 1960). A review of the literature shows a long sequence of papers on various aspects of subject access, emphasizing its importance and the need to support it in bibliographic information systems in addition to known-item searching. Husain and O'Brien (1992) confirmed that research conducted in the early 80's showed that subject access was one of the most dominant approaches in OPACS. Entering the so-called third-generation OPAC, the subject searching capability was greatly enhanced by the inclusion of additional controlled and uncontrolled access points, for example, letting the words taken from the table-of-contents and in-

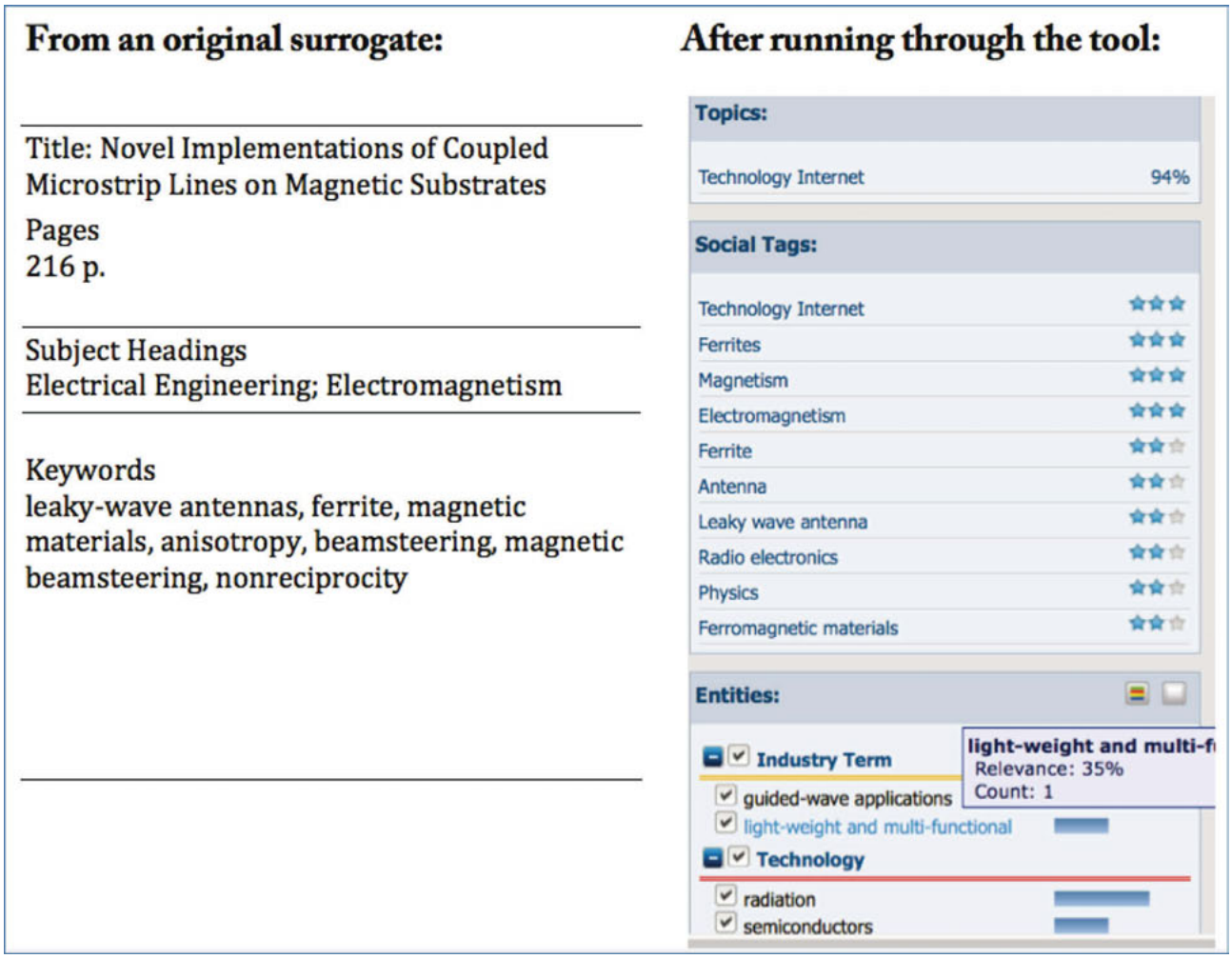


Figure 1. Subject headings and keywords provided by original catalog record, (left), and topics, tags, and entities provided by an automatic semantic analysis tool (right).

dexes of the books supplement The *Library of Congress Subject Headings* (*LCSH*). Bates pointed out problems end-users have when searching on a topic and proposes an entry vocabulary as a complement to controlled vocabularies, but also encourages the use of automated methods (2003, 39): "The second question concerns the use of available software for generating access terms. Anything that can be well done automatically should be."

In the last ten years we have been witnessing heated discussions on whether controlled vocabularies—subject headings in particular—are still worth the investment. Many researchers and practitioners argue that keyword searching or user-generated tags make controlled vocabularies obsolete, inefficient, and unnecessary. Yet, Gross and Taylor (2005) discovered out that over one third of records retrieved through keyword searches are those where keywords were found in subject headings. The lack of controlled vocabularies would therefore seriously affect keyword searching, which is the predominant way users now search for information. William Badke (2012) sees the solution in user education, particularly in the academic environment, and concludes rather pessimistically: "If we fail to advocate and if we do not restore the prominence of such vocabularies, they will disappear because of disuse and a negative cost-benefit analysis."

The growing use of user-generated tags in information systems has spurred numerous studies of tags' efficacy in improving access to materials (Rolla 2009; Klavans, LaPlante, and Golbeck 2014). The conclusion of the first study, which compared LibraryThing tags and *LCSH*, are that both have strengths and weaknesses and the author suggests that libraries should combine both in supporting their users (Rolla 2009). The second study is an analysis of the nature of tags according to two facets based on Panofsky (1939) and Shatford (1986): subject matter (who, what, where, and when) and specificity (general, specific, abstract). While the researchers found that their test collection of digital art images was most likely to generate generic tags that describe people or things found in the images, they also suggest that this was not a universal finding

for how people tag, and that "tag sets largely depend on the type of collection and the needs of the user" (Klavans, LaPlante, and Golbeck 2014).

A comprehensive literature review on automatic classification of documents in library environments revealed three basic types of research or efforts; one of them is text categorization and document categorization using different types of classifiers with or without using training documents (Desale and Kumbhar 2013). Recently reported applications in applying automatic or machine-assisted semantic analysis in LAM collections, especially those not in the routine cataloguing coverage or in the analytical level subject indexing, have focused on semantic annotation, entity extraction, and relationship description. The theories and methods can be traced from the field of automatic summarization and semantic analysis involving many linguistics researchers (Mani 2001). One of the theories of Text Coherence is the Rhetorical Structure Theory (RST) that brought up four rhetorical relations: Circumstance, Motivation, Purpose, and Solutionhood. Among those, the circumstance means that the satellite sets a temporal, spatial, or situational framework in the subject matter within which the reader is intended to interpret the situation presented in the nuclear text span (Mann and Thompson 1988). On the other hand, Robert Allen (2013a, 2013b) explains that RST does not seem well suited to large volumes of complex texts. Allen's team proposes that the event-entity fabric be overlaid with additional structures to present causation, generalization, explanation, argumentation, and evidence. Using rich content such as historical texts as the case, the two articles by Allen suggest that schematic models, which describe the content of documents rather than descriptions about the documents, are the key for a new generation of descriptive systems.

For entity extraction, pioneer works include BBC's automated interlinking of speech radio archives (Raimond and Lowis 2012) and experiments of entity extraction for BBC News (Tarling and Shearer 2013). Whether used to embed annotations inside the text (e.g., Brat and Pundit annotation tools) or to extract entities out of the text (e.g., OpenCalais), these tools "type" the entities according to classes or categories pre-defined or defined in the analytic processes. They present a great potential in subject analysis workflow in LAMs, combined with the ontologies, conceptual and data models, and metadata schemas developed in related domains and applicable to processing LAM materials. Examples include using Calais to enhance access to oral history materials (Perkins and Yoose 2011) and museum online collections (Catone 2008).

Erwin Panofsky's three-layers theory has been widely used by the researchers and practitioners examining subject access to images, particularly iconological themes found in the art of the Renaissance as well as art images in general (Panofsky 1939; Shatford Layne 1994; Klavans, LaPlante, and Golbeck 2014). The theory has also been extended to be the basis for subject analysis of all cultural objects, as suggested by the content standard *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images* (*CCO*) (Baca et al. 2006; Harpring 2009). Panofsky (1939) summarized the coordination of the three layers of object interpretation:

- Layer I is the Primary or Natural Subject Matter, subdivided into (A) factual and (B) expressional. The equipment for interpretation is the *practical experience*, i.e., familiarity with *objects* (e.g., human beings, animals, plants, etc.) and *events* (i.e., their mutual relations). The example of *expressional* qualities given by Panofsky included the mournful character of a pose or gesture, or the homelike and peaceful atmosphere of an interior (Panofsky 1939, 5). In a broader sense CCO considers Level I referring to the generic elements depicted in or by the work.
- Layer II is the Secondary or Conventional Subject Matter, constituting the world of images, stories and allegories. The equipment for interpretation is *knowledge of literary sources*, i.e., familiarity with specific *themes* and *concepts*. For example, the subject is apprehended by realizing that a group of figures seated at a dinner table in a certain arrangement and in certain poses represents the Last Supper (Panofsky 1939, 6).
- Layer III is Intrinsic Meaning or Content, constituting the world of '*symbolical*' *values*. It is apprehended by ascertaining those underlying principles which reveal the basic attitude of a nation, a period, a class, a religious or philosophical persuasion – unconsciously qualified by one personality and condensed into one work (Panofsky 1939, 7). The equipment for interpretation is *synthetic intuition* (familiar with the *essential tendencies* of *the human mind*), conditioned by personal psychology and "*Weltanschauung*" (Panofsky 1939, 15).

The layers, or the three strata of subject matter or meaning, are aligned with the three types of interpretation: act of, equipment for, and controlling principle of interpretation (Panofsky 1939, 14-15). Simplified by *CCO*, the three layers become: description, identification, and interpretation. These are to be further discussed in the following section.

3.0 Research method and preliminary findings

As an aid in understanding how computerized subject analysis might be approached, we suggest using the three-

layer framework that has been accepted and applied in image analysis, as developed by Erwin Panofsky (Table 1). In the previous section we indicated the wide use of Panofsky's three-layers framework. When the three layers of object interpretation are simplified by *CCO*, they become: I. Description (referring to the generic elements depicted in or by the work); II. Identification (referring to the specific subject); and, III. Interpretation (referring to the meaning or themes represented by the subjects and including a conceptual analysis of what the work is about) (Beca et al. 2006). We aligned the *CCO* layers with the summarized Panofsky layers in Table 1.

This paper reports on part of the analysis of two research samples from the point of view of Panofsky's theory. The hypothesis is that computer-assisted semantic analysis has great potential in generating subject access at the "description" and "identification" levels.

3.1 Research sample 1: archival descriptions

As noted above, two research samples were used to analyze the access points supplied by the OpenCalais semantic analysis tool. The first sample includes 43 archival record groups from 16 institutions, including university archives, government records archives, and manuscript/special collections repositories in various LAMs. The sample finding aids were identified by searching in several websites and catalogs containing descriptions of archival collections, including the Library of Congress' American Memory Project, the OhioLINK Finding Aid Repository, the Online Archive of California, and the University of Michigan's Bentley Historical Library finding aid repository. The aim was to gather finding aids that represented the full spectrum of archival description, including many different types of collections and archives. Personal and family papers, corporate records, government records, and artificial collections (collections of materials with mixed provenance) were represented evenly in this sample. The breakdown of types of collections among the 43 chosen for the sample is as follows: personal and family papers (10), corporate records (12), government records (11), and artificial collections (10).

Descriptive information such as creator histories and scope and content notes found in the archival finding aids, as well as abstracts and container listings from these descriptions, often contain unstructured text blocks containing many potential personal, corporate body, or geographic names. The following extract, drawn from the finding aid of the Edward and Clara Steuermann Collection, provides a good example of how these names are embedded in narrative text (potential entities marked with an asterisk):

> Edward Steuermann* was born June 18, 1892, in Sambor*, a small Polish city in eastern Galicia* (now part of the Ukraine*). His study of the piano began in 1904 with the Czech pianist and teacher Vilem Kurz* and continued, first in Basel* in 1910 and then in Berlin*, with Ferruccio Busoni*. His first composition teacher of note was Engelbert Humperdinck*, but Steuermann's inclinations towards the modern idiom made him seek instruction elsewhere. At Busoni's suggestion, Steuermann began studying with Arnold Schoenberg* in 1912,

Object of Interpretation	Act of Interpretation	Equipment for Interpretation	Controlling principle of Interpretation	Simplified layers [2]
I-*Primary or natural* subject matter – (A) factual, (B) expressional-, constituting the world of artistic motifs	*Pre-iconographical description* (and pseudo-formal analysis).	*Practical experience* (familiar with *objects* and *events*).	History of *style* (insight into the manner in which, under varying historical conditions, *objects* and *events* were expressed by *forms*).	I–Description (refer to the generic elements depicted in or by the work).
II-*Secondary* or *conventional* subject matter, constituting the world of *images*, *stories* and *allegories*.	*Iconographical analysis* in the narrower sense of the word.	*Knowledge* of *literary sources* (familiar with specific *themes* and *concepts*).	History of *types* (insight into the manner in which, under varying historical conditions, specific *themes* or *concepts* were expressed by *objects* and *events*).	II–Identification (refer to the specific subject).
III-*Intrinsic meaning* or *content*, constituting the world of '*symbolical' values*.	*Iconographical interpretation* in a deeper sense (*iconographical synthesis*)	*Synthetic intuition* (familiar with the *essential tendencies* of *the human mind*), conditioned by personal psychology and '*Weltanschauung*.'	History of *cultural symptoms* or '*symbols*' in general (insight into the manner in which, under varying historical conditions, *essential tendencies of the human mind* were expressed by specific *themes* and *concepts*).	III – Interpretation (refer to the meaning or themes represented by the subjects and includes a conceptual analysis of what the work is about).

Table 1. Panofsky's three-layer framework and the simplified layers used by *CCO*. Based on Panofsky (1939, 14-15) and Chapter 4 of Baca et al. (2006, 208).

thus initiating a professional association that was to figure prominently in Steuermann's career as both composer and pianist. Beginning with *Pierrot lunaire*,* Steuermann performed in the premiere of almost every Schoenberg work for which a pianist was required. While in Vienna*, he served as the pianist for the Verein für Musikalische Privataufführungen (Society for Private Musical Performances)* founded by Schoenberg in 1918 to introduce newer works there. Concurrent with these activities, Steuermann began a distinguished teaching career that would continue through the remainder of his life.

Text blocks such as the one noted above were drawn from each finding aid and processed via the OpenCalais semantic analysis service to generate extracted access point candidates. The whole process of extracting descriptive data from finding aids and inputting it into the OpenCalais service was automatic. Using an in-house-developed program, called the Semantic Analysis Method (SAM) Tool, the software automatically obtained the archival records by batch upload of text files and sent them to the semantic analysis service supported by Calais. More information about the SAM Tool may be found on the following website: http://lod-lam.slis.kent.edu/SemanticAnalysis.html. The source code for the SAM Tool is available at https://github.com/sammysemantics/SAM.

The JSON output generated from the SAM Tool was then converted directly into a CSV file, which could be viewed as a Microsoft Excel spreadsheet. The resulting database contained the following fields: Entity-type, Entity-name, Relevance-ratio, and File-source. Using the OpenRefine tool, the data were clustered automatically to allow the researchers to clean up the data manually (e.g., merge the synonyms and delete incorrect extractions). Figure 2 illustrates this multi-step process.

The analysis resulted in dozens and, at times, hundreds of potential entities and social tags that could be used to provide additional points of entry to these archival records. These entities and tags correspond almost exclusively to the first two layers of subject analysis (description and identification). Entity-based terms are in general more common than topical terms; it is very rare to find any terms at the third level of analysis (interpretation) in descriptions of archival materials, due to their evidentiary nature (see Figures 3 and 4).

Entities correctly extracted via Calais analysis (at level I, or, description) included personal names (Person), corporate names (Company, Facility, Organization), and geographic names (City, Continent, Country, Natural Feature, ProvinceOrState, Region), and events (Holiday, PoliticalEvent). Calais provides relevance scores for each identified entity, which may be used as a valuable clue about the importance of that entity to the overall scope of the archival collection. While it is difficult to predict exactly what the cut-off relevance score might be for a system to include an entity as an indexed term, given the differences in description exhaustivity among different institutions, the relevance scores could certainly be used to suggest possible indexing terms. LAMs may also choose to perform analysis and generate relevance scores only on particular parts of the finding aids (such as the creator history and the scope and content note) to improve reliability of the scores.

In addition to entities, Calais also generated many topical terms describing the subject matter of the records (at level II, or, identification); these topics were often found as social tags or as entities under the "IndustryTerm" or "Product" category (see Figure 5). These cate-

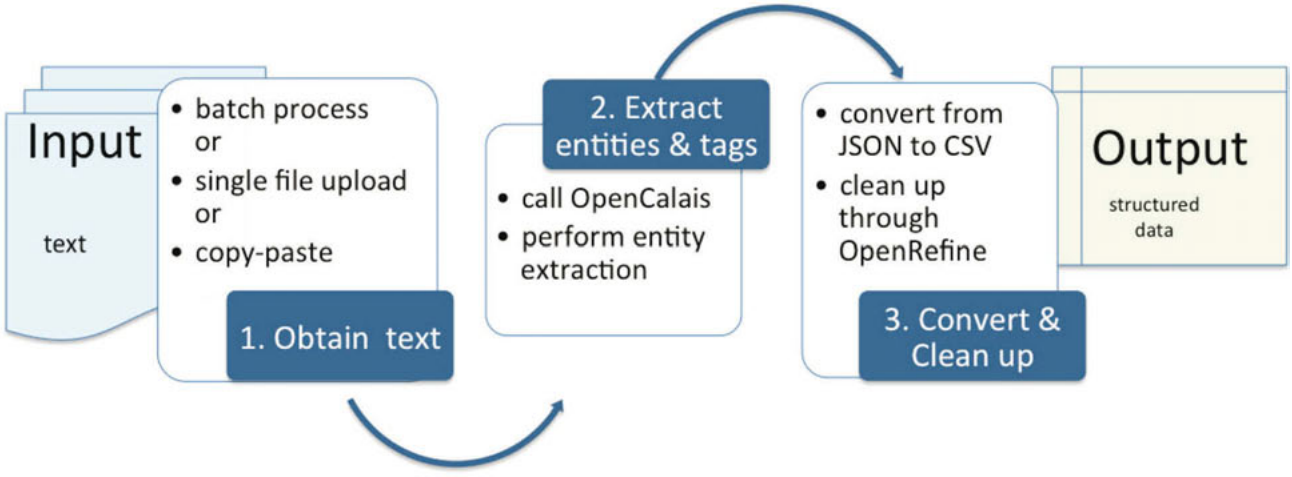


Figure 2. Process for creating structured data from unstructured text found in descriptions of archival materials.

Figure 3. Personal, corporate, and geographic entities generated by semantic analysis of an archival finding aid. Pearl Harbor Attack (Dec 6 – Dec 8, 1941), a collection found at the Franklin D. Roosevelt Presidential Library. Finding aid located at: http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf.

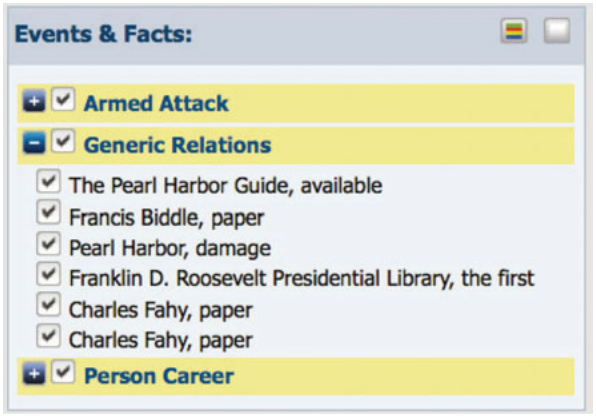


Figure 4. Event entities generated by semantic analysis of the same archival finding aid (refer to Figure 3). Pearl Harbor Attack (Dec 6 – Dec 8, 1941), a collection found at the Franklin D. Roosevelt Presidential Library. Finding aid located at: http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf.

Figure 5. Topical terms, called "social tags," generated from the semantic analysis of an archival finding aid. Pearl Harbor Attack (Dec 6 – Dec 8, 1941), a collection found at the Franklin D. Roosevelt Presidential Library. Finding aid located at: http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf.

gorizations were the least reliable in terms of accuracy; the Calais analytic engine often incorrectly identified text strings from the finding aids as products or industry terms. Many of these errors can be attributed to the raw data that was fed to the engine: the entire finding aid was used and this unedited text often included physical location information for the records and document formatting that generated significant noise for the analysis engine to sort through. Targeted analysis of particular areas of the finding aids may result in better accuracy for topical analysis.

As a point of comparison to the automated analysis of the finding aids, the researchers also examined the controlled vocabulary topical terms and names assigned to the archival records. These terms and names are typically drawn from controlled vocabularies such as Library of Congress Name Authority File (LCNAF), *LCSH*, and the *Art and Architecture Thesaurus* (*AAT*). As with the entities and social tags generated by Calais, the headings can be primarily categorized according to the first and second

layers of analysis: 1) Description: personal, family, corporate, and geographic names (note that the first three types of names can also be encoded as records creators in addition to being subjects depicted in the records); and, 2) Identification: topical terms (including occupations and functions represented in the records), genre and form terms. The depth of subject analysis is wildly variable—while some archival records groups were assigned dozens of headings, others received a minimal number. Government records are often not assigned subject headings at all, while personal papers and special collections are more likely to have a sizeable number of headings (at least five or six, and often many more).

As noted above, certain factors such as the size of archival collections, varying institutional practices, and different approaches to the indexing of different types of archival materials, even within the same institution, may influence the exhaustivity of subject analysis. As an example, compare the number of names and genre terms found in two finding aids to the number of controlled name and genre access points in their corresponding catalog records. These collections, drawn from the Library of Congress' American Memory Project finding aid repository, contain material of a similar type (personal papers of artists—one a musician, and the other a dancer). They are of similar size (in terms of linear feet) (see Table 2).

In this example, the Steuermann Collection finding aid contains almost four times as many potential name entities as the finding aid of the Bacon Collection, and almost 1.5 times as many genre terms. Similarly, significantly more name access points (controlled vocabulary) are assigned to the Steuermann Collection than the Bacon Collection, despite their similar size and domain. Thus, even within the same repository, differing approaches to description by archivists may lead to wide variation in the number of names mentioned in the finding aid and assigned as controlled vocabulary. For a more detailed analysis of the name entities found in this set of finding aids, the authors refer readers to Gracy's (2014) research on challenges of implementing linked data for archival description.

Under these circumstances, where there are varying levels of specificity in indexing archival materials, it is difficult to propose that automated semantic analysis will always result in a more exhaustive or accurate list of terms, as some archivists may describe material more extensively than others. This study suggests, however, that it would be well worth the effort for institutions to experiment with semantic analysis methods as either an initial step to suggest key entities and topics, or as a final check to ensure that important concepts or entities have not been overlooked. As many archives do not provide topical indexing of archival collections at all, it was difficult to analyze assignment of topical terms for the sample used in this study. For certain types of records, particularly those for which subject indexing is not common, semantic analysis may provide entry points to archival records that were not previously available. Such techniques will enhance subject analysis at the first two levels (description and identification), but are unlikely to be useful for interpretation of the material.

3.2 Research sample 2: philosophy theses

In contrast with the methods used in the archival data sample, the second sample used manual processes in most of the procedures. The sample contains 44 philosophy theses consisting of two sub-samples from KentLINK and OhioLINK. OhioLINK is the state-funded consortium of university and college member libraries in Ohio and the State Library of Ohio. The Electronic Theses and Dissertations (ETD) Center is one of the free online databases provided by OhioLINK. It contains Ohio's undergraduate,

Collection Title	*Size of Collection*	*# of names in finding aid*	*# of name access points in MARC record*	*# of genre terms in finding aid*	*# of genre access points in MARC record*
Edward and Clara Steuermann Collection*	Circa 1,800 items (43 boxes, or approximately 16 linear feet)	471	17	70	2
Ernst Bacon Collection†	Around 6,000 items (54 boxes, or approximately 16 linear feet)	81	6	46	3

Table 2. Comparison of entities found in finding aid and catalog descriptions for two archival collections drawn from study sample.

* = Finding aid found at http://hdl.loc.gov/loc.music/eadmus.mu004007; catalog record found at: http://lccn.loc.gov/2010563514.
† - Finding aid found at http://hdl.loc.gov/loc.music/eadmus.mu003006; catalog record found at: http://lccn.loc.gov/2003561021.

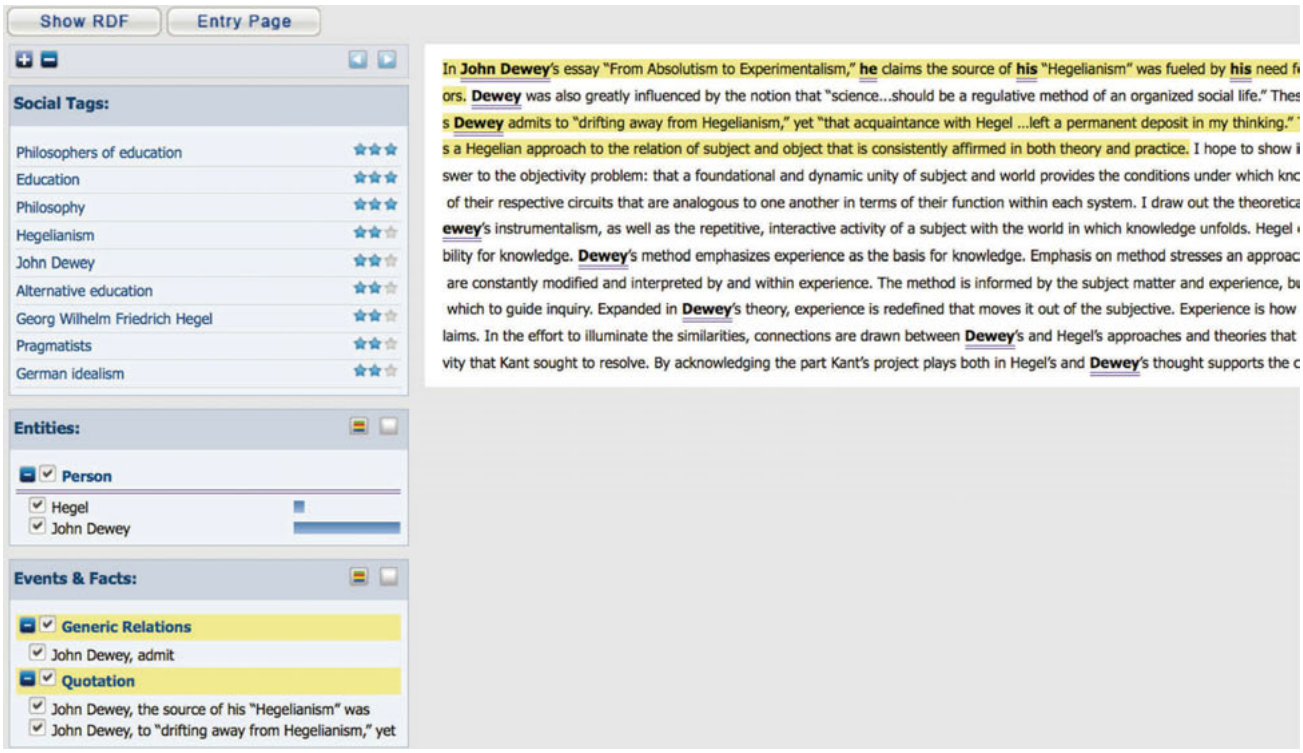


Figure 6. Topical terms, called "social tags," Entities, and Event and Facts (on left) generated from the semantic analysis and the entities highlighted in the text by OpenCalais (on right, partial view).

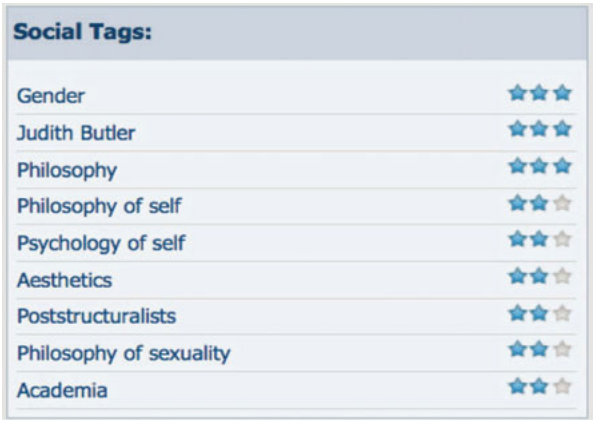
master's, and doctoral theses and dissertations from participating OhioLINK member schools. An author is required to provide a simple metadata description when submitting the final version of the thesis or dissertation, including an abstract and any number of keywords. University libraries in Ohio often re-catalog the ETDs authored by their graduates in order to integrate these ETDs into the university online catalogs. For example, KentLINK is the library catalog of the Kent State University (KSU). The Technical Services department of KSU Libraries catalogs the theses of KSU graduates that have been published on the OhioLink ETD Center using the MARC format in order to include them in the KentLINK.

For this study, the first sub-sample was a selected set (22) of philosophy theses and dissertations. They were published in the OhioLINK ETD Center and also included in KentLINK. The second sub-sample contained philosophy theses and dissertations found in OhioLINK. They were randomly picked from the results of a search in the OhioLINK ETD database and further checked from the WorldCat. A total of 22 philosophy theses and dissertations for which MARC records could be found in WorldCat were selected in the study. It should be noted that these WorldCat records may or may not include controlled subject headings assigned in the process of being converted into MARC records. In other words, the keywords and subject headings submitted by the authors (without using controlled vocabularies) may be the only values in the "subject" fields.

Abstracts, titles, keywords, and introduction paragraphs from the theses and dissertations sample were submitted to OpenCalais separately to obtain the results. All of the candidate terms were counted according to Agent Names, Geographic Names, Corporate Name, and Topic Terms. They were manually validated to determine 1) the relevance to the thesis, 2) the type of term (e.g., named entity, tag, or general heading), and 3) its availability in LCNAF, *LCSH*, Wikipedia (as an entry), and the *Stanford Encyclopedia of Philosophy*.

In this part of the research, it was found that the semantic analysis based on the abstracts generated more successful tags than those based on the titles. Focusing on the tags generated by the software, it is interesting to see that the entity names missed in the Entity section (e.g., singular names such as Plato and Aristotle, or instances where the first name was not included) were often correctly extracted into the tags section. Major concepts were correctly identified in most cases (see Figure 6).

However the software often over-generalized the subjects by assigning very general terms (e.g., "philosophy," for almost every philosophy thesis) and some terms that were unrelated to the subject of the thesis (Figure 7a and 7b). This level is different from "identification" and "description," and seems to be more akin to "inferencing." Among



7a.

(Note that no Topic was identified from the abstract of this thesis.)

7b.

Figure 7a & 7b. The term "philosophy" appears among tags, generated from the semantic analysis of two theses.

the average of 9 tags per abstract in the KentLINK subsample, an average of 1.64 were overly broad topical terms and 3.45 were unrelated topical terms (slightly more than 1/3). The results for the tags in the OhioLINK ETD subsample are similar to the results in the KentLINK subsample.

Using the three-layers as the framework, the tags did very well in level I "description" and adequately in level II "identification." The tags that could be categorized as "inferencing" results seemed to be less valid according to the best practices of cataloging and subject indexing. The overly-broad topic terms are not wrong (e.g., philosophy, knowledge, science) but their relevance in terms of subject access is questionable. The promising news is that among the topical terms (including named entities as topics), *LCSH* together with LCNAF could match about 75% of them closely (we used the degree as closeMatch, in comparison to broadMatch, narrowMatch or noMatch), and DBpedia matches almost 98% with closeMatch degree for both sub-samples. These vocabulary sources hold great potential for these subject access points to become the linking point to the Linked Data datasets that use DBpedia and LC vocabulary URIs as their basis.

4.0 Conclusions and future research

There has been increased attention to the needs of reusing existing LAM data found in catalogs, finding aids, indexes, and author-contributed descriptions for the purpose of improving access, findability, and linkability of LAM materials. Yet, current circumstances also force us to face the gap between such needs and current practices of cataloging and indexing, as well as the significantly increased human resources required to achieve value-added results of enhanced subject access to LAM collections. This study took a journey of experimenting with alternative approaches of generating subject access points to LAM materials. While computerized subject analysis seemed to hold much promise for improving subject access to both individual resources and large heterogeneous collections, the authors also kept in mind concerns about the quality of such alternative mechanisms for generating additional entry points into LAM collections.

We reported on the analysis of access point candidates generated by the OpenCalais semantic analysis engine using the framework of Panofsky's three layers, which were simplified by *CCO*. The research based on the two samples indicates these subject access points fall at the "description" (referring to the generic elements depicted in or by the work) and "identification" (referring to the specific subject) levels, rather than the "interpretation" (referring to the meaning or themes represented by the subjects and including a conceptual analysis of what the work is about) level. At a certain point, we can say that results are also derived by inferencing (e.g., those generalized terms), a layer that was not in Panofsky's or *CCO*'s

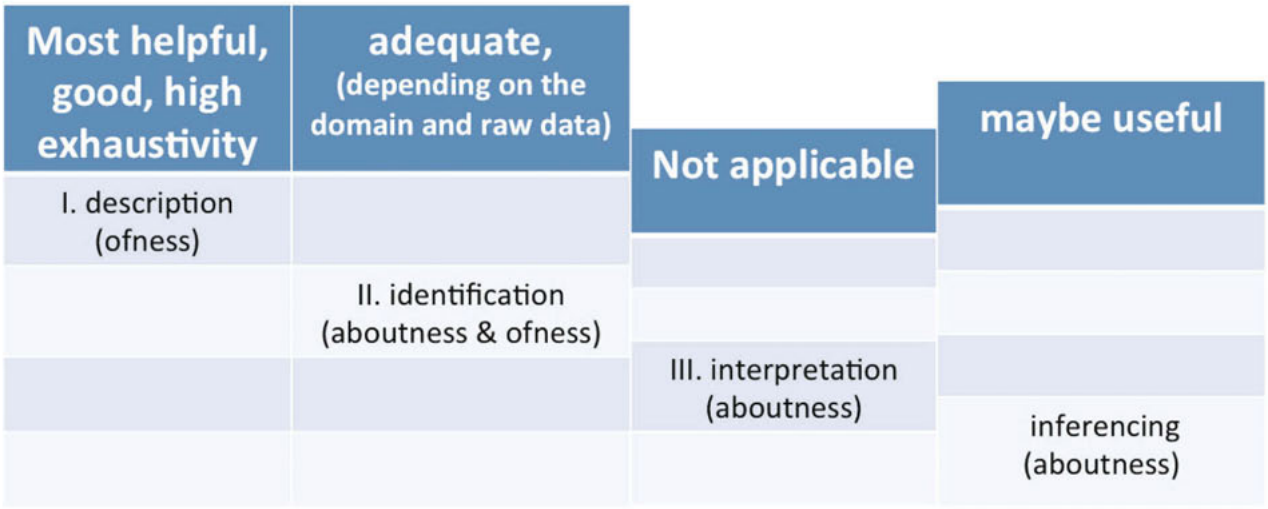


Figure 8. Summary of the usefulness of access point candidates based on the two sample results.

framework. The inferencing results may be useful for large heterogeneous collections that cover a wide range of disciplines and domains. The usefulness of access points at each level of analysis for users are summarized in Figure 8.

Since we are particularly interested in large heterogeneous digital libraries, it would be interesting to analyze typical user queries of such tools; for example, we could analyze user needs according to the three layers (or substitute the "interpretation" with "inferencing") and thus understand their nature. This knowledge would help us predict the usefulness of existing semantic analysis tools.

References

Allen, Robert B. 2013a. Model-oriented information organization: part 1, the entity-event fabric. *D-Lib Magazine* 19 no. 7/8.

Allen, Robert B. 2013b. Model-oriented information organization: part 2, discourse relationships. *D-Lib Magazine* 19 no. 7/8.

Baca, Murtha et al., eds. 2006. *Cataloging cultural objects: a guide to describing cultural works and their images (CCO)*. Chicago: American Library Association.

Badke, William. 2012. Save the subject heading. *Online* 36 no. 6: 48-50.

Bates, Marcia J. 2003. Task force recommendation 2.3, research and design review: improving user access to library catalog and portal information. Final report (version 3). June 1, 2003. Available http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf.

Catone, Josh. 2008. Australian museum uses Open Calais to tag collection [blog]. Available http://alturl.com/xv7hb.

Cleverdon, Cyril. 1960. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, UK: College of Aeronautics.

Desale, Sanjay K. and Kumbhar, Rajendra. 2013. Research on automatic classification of documents in library environment: a literature review. *Knowledge organization,* 40: 295-304.

Gracy, Karen F. 2014. Archival description and linked data: A preliminary study of opportunities and implementation challenges. *Archival science* (forthcoming, 2014). doi: 10.1007/s10502-014-9216-2. Published online 27 February 2014.

Gross, Tina and Taylor, Arlene. 2005. What have we got to lose? The effect of controlled vocabularies on keyword searching results. *College & research libraries* 66: 212-30.

Harpring, Patricia. 2009. Subject access to art works: using CCO/CDWA & vocabularies [educational material]. Available http://www.getty.edu/research/tools/vocabularies/subject_access_for_art.pdf.

Husain, Shabahat and Ann O'Brien, Ann. 1992. Recent trends in subject access to OPACs: an evaluation. *Knowledge organization* 19:140-45.

Klavans, Judith L., LaPlante, Rebecca, and Golbeck, Jennifer. 2014. Subject matter categorization of tags applied to digital images from art museums. *Journal of the American Society of Information Science & Technology* 65: 3-12.

Mani, Inderjeet. 2001. *Automatic summarization*. Amsterdam: John Benjamins Publishing Company.

Mann, William. C. and Thompson, Sandra A. 1988. Rhetorical structure theory: towards a functional theory of text organization. *Text* 8: 243-81.

Panofsky, Erwin. 1939. *Studies in iconology: humanistic themes in the art of the Renaissance*. S.l.: Oxford University Press. Reprint, New York: Harper Torchbooks, 1962.

Perkins, Jody and Yoose, Becky. 2011. Mining oral history for enhanced access. Poster presentation at the Society of American Archivists Annual Conference, Chicago, Illinois, August 22-27, 2011. Available http://alturl.com/nehkr.

Raimond, Yves and Lowis, Chris. 2012. Automated interlinking of speech radio archives. *Linked Data on the Web (LDOW2012); workshop at the 21st International World Wide Web Conference, Lyon, France, April 16 2012*. Available http://events.linkeddata.org/ldow2012/papers/ldow2012-paper-11.pdf

Rolla, Peter. 2009. User tags versus subject headings: can user-supplied data improve subject access to library collections? *Library resources & technical services* 53: 174-84.

Shatford, Sara. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6 no. 3: 39-62.

Shatford Layne, Sara. 1994. Some issues in the indexing of images. *Journal of the American Society of Information Science* 45: 583-88.

Tarling, Jeremy and Shearer, Matt. 2013. Unlocking the data in BBC News. *Knowledge organization: pushing the boundaries*; *ISKO UK Biennial Conference, London, July 8-9 2013*. Available http://www.iskouk.org/conf2013/slides/TarlingSlides.ppt and http://www.iskouk.org/conf2013/mp3/ISKOUK_08-2013_BBC_Tarling Shearer.mp3.