2014 International ISKO Conference, May 19-22, 2014, Krakow, Poland.

Marcia Lei Zeng[1], Karen F. Gracy[1], Maja Žumer [2],
[1] Kent State University, Kent, USA
[2] University of Ljubljana, Slovenia

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

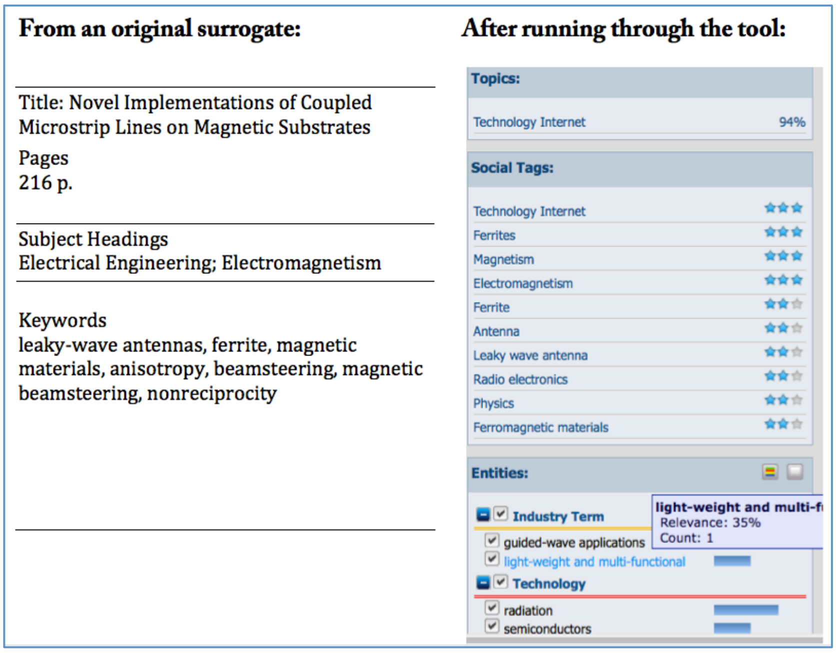
Abstract:
This paper attempts to explore an approach of using an automatic semantic analysis tool to enhance the "subject" access to materials that are not included in the usual library subject cataloging process. Using two research samples the authors analyzed the access points supplied by OpenCalais, a semantic analysis tool. As an aid in understanding how computerized subject analysis might be approached, this paper suggests using the three-layer framework that has been accepted and applied in image analysis, developed by Erwin Panofsky.

1. 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 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 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 six keywords suggested by the dissertation author in the process of submitting to the electronic thesis and dissertation (ETD) repository and two standardized subject headings assigned by a library cataloger in the re-processing procedure. 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").

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).



From an original surrogate:

Title: Novel Implementations of Coupled Microstrip Lines on Magnetic Substrates

Pages
216 p.

Subject Headings
Electrical Engineering; Electromagnetism

Keywords
leaky-wave antennas, ferrite, magnetic materials, anisotropy, beamsteering, magnetic beamsteering, nonreciprocity

After running through the tool:

Topics:

Technology Internet	94%

Social Tags:

Technology Internet	★★★
Ferrites	★★★
Magnetism	★★★
Electromagnetism	★★★
Ferrite	★★☆
Antenna	★★☆
Leaky wave antenna	★★☆
Radio electronics	★★☆
Physics	★★☆
Ferromagnetic materials	★★☆

Entities:

☑ **Industry Term**
- ☑ guided-wave applications
- ☑ light-weight and multi-functional

light-weight and multi-f
Relevance: 35%
Count: 1

☑ **Technology**
- ☑ radiation
- ☑ semiconductors

The process 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. Review of Related Literature

The Cranfield project is considered the first systematic evaluation in information retrieval systems. Led by Cyril Cleverdon, it lasted ten years (from 1957) and focused on the effectiveness of different indexing languages. The project set the stage for further research in Information retrieval – and established subject access as the central topic. A review of the literature shows a long sequence of papers on various aspects of subject access, emphasising its importance and the need to support it in bibliographic information systems in addition to known-item searching. Marcia Bates (2003) points 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: "The second question concerns the use of available software for generating access terms. Anything that can be well done automatically should be" (Bates, 2003, p. 39).

In the last ten years we are 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, 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 et al., 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. 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 et al., 2014, p. 10).

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., Calais), 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 et al., 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, 2006; Harpring, 2009). Panofsky (1939) summarized the coordination of the three layers of object interpretation as (I) primary or natural subject matter; (II) secondary or conventional subject matter; and (III) intrinsic meaning or content. The layers are aligned with the three types of interpretation: act of, equipment for, and controlling principle of interpretation. Simplified by CCO, the three layers become: description, identification, and interpretation. These are to be further discussed in the following section.

3. Research Method and Preliminary Findings

As an aid in understanding how computerized subject analysis might be approached, this paper suggests using the three-layer framework that has been accepted and applied in image analysis, as developed by Erwin Panofsky (Figure 2). 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); III. Interpretation (referring to the meaning or themes represented by the subjects and including a conceptual analysis of what the work is about). We aligned the CCO layers with the summarized Panofsky layers in the following figure:

Figure 2. Panofsky's three-layer framework and the simplified layers used by CCO.
Source: Compiled based on Panofsky, 1939, p.14-15 and CCO Chapter 4.

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).

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 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 OpenCalais semantic analysis tool. The first sample includes 43 archival record groups from sixteen institutions, including university archives, government records archives, and manuscript/special collections repositories in various LAMs. Descriptive information such as creator histories and scope and content notes found in the archival finding aids, as well as abstracts from these descriptions were put into the OpenCalais open service to generate extracted access point candidates. The whole process was automatic. Using an in-house-developed program, the software automatically obtained the archival records and sent them to the semantic analysis service supported by Calais. The output, which was in the JSON format, was then converted directly into a CSV file, which can be viewed as 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).

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). Identifying terms are in general more common than descriptive 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.

Entities correctly identified via Calais analysis (at level one, 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 (level two, or, identification); these topics were often found as social tags or as entities under the "IndustryTerm" or "Product" category. These categorizations 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), Library of Congress Subject Headings (LCSH), and Art and Architecture Thesaurus. 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: topical terms (including occupations and functions represented in the records), genre and form terms; and 2) Identification: 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). 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 may influence the exhaustivity of subject analysis. Under these circumstances, it is difficult to propose that automated semantic analysis will always result in a more exhaustive or accurate list of terms. 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. 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.

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 a selected sub-sample (22) from KentLINK and a random sample (22) from OhioLINK. Abstracts, titles, keywords, and introduction paragraphs 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 a term (e.g., named entity, tag, or general heading), and 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 (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. 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. This level is different from "identification" and "description", seems to be "inferencing". Among the average of 9 tags per abstract in the KentLINK sub-sample, 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 sub-sample are similar to these figures. Using the three-layers as the framework, the research found that 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. Conclusions and Future Research

The paper reports on the analysis of the resulted access point candidates based on Panofsky's three layers, which indicate these subject access points fall at the "description" and "identification" levels, rather than the "interpretation level. At a certain point, we can say that results are also derived by inferencing (e.g., those generalized terms). Since we are particularly focusing on large heterogeneous digital libraries, it would be interesting to analyze typical user queries of such tools. In a future study 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.

Acknowledgement

The authors would like to thank research assistants Sammy Davidson and Laurence Skirvin of Kent State University for assisting with OpenCalais-related processes.

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 (2013).

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(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. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf

Catone, Josh. (2008). "Australian Museum Uses Open Calais to Tag Collection" (blog). http://alturl.com/xv7hb

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(3): 212-230.

Harpring, Patricia. (2009). "Subject Access to Art Works: Using CCO/CDWA & Vocabularies" (educational material). http://www.getty.edu/research/tools/vocabularies/subject_access_for_art.pdf

Klavans, Judith L., LaPlante, Rebecca and Golbeck, Jennifer. (2014). "Subject Matter Categorization of Tags Applied to Digital Images from Art Museums." *JASIST* 65, (1): 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(3): 243-281.

Panofsky, Erwin. (1939). *Studies In Iconology: Humanistic Themes In The Art Of The*

Renaissance. S.l.: Oxford University Press, 1939. 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. http://alturl.com/nehkr

Raimond, Yves and Lowis, Chris. (2012). “Automated Interlinking of Speech Radio Archives.” Linked Data on the Web (LDOW2012), April 16, 2012. Workshop at 21st International World Wide Web Conference, Lyon, France. 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(3):174-184.

Shatford Layne, Sara. (1994). “Some Issues in the Indexing of Images.” *JASIS* 45(8): 583-588.

Tarling, Jeremy and Matt, Shearer. (2013). **“**Unlocking the data in BBC News.” Knowledge Organization: Pushing the Boundaries. ISKO UK Biennial Conference, July 8-9, 2013, London.