Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata

[1,2]Judith L. Klavans, [1]Tandeep Sidhu, [1]Carolyn Sheffield, [1]Dagobert Soergel
[1,2]Jimmy Lin, [3]Eileen Abels and [4]Rebecca Passonneau

[1]College of Information Studies (CLIS), USA
[2]University of Maryland Institute for Advanced Computer Science (UMIACS)
University of Maryland, College Park, Maryland, USA
[3]College of Information Science and Technology, Drexel University, Phila PA, USA
[4]Center for Computational Learning Systems, Columbia University, NY, USA

Abstract. In this paper, we present a fully-implemented system using computational linguistic techniques to apply automatic text mining for the extraction of metadata for image access. We describe the implementation of a workbench created for, and evaluated by, image catalogers. We discuss the current functionality and future goals for this image catalogers' toolkit, developed under the Computational Linguistics for Metadata Building (CLiMB) research project.[1] Our primary user group for initial phases of the project is the cataloger expert; in future work we address applications for end users.

1 The Problem: Insufficient Subject Access to Images

The CLiMB project addresses the existing gap in subject description in metadata for image collections, particularly for the domains of art history, architecture and landscape architecture. Within each of these domains, image collections are increasingly available online yet the availability of subject-oriented access points for these images remains minimal, at best. In an initial observational study conducted with six image catalogers, we found that typically between one and eight subject terms are added to catalog records for images and many legacy records lack subject entries altogether.
The literature on end users' image searching practices indicates that this level of subject description may be insufficient for some user groups. In a study of the image-searching behaviors of faculty and graduate students in the domain of American history [3], found that 92% of the thirty-eight participants considered the textual information associated with the images inadequate for the images searched in the Library of Congress' American Memory Collection. The individual records in this collection typically contain quantities of subject descriptors comparable to—or exceeding- those found in the exploratory CLiMB studies. Furthermore, this study found that this

[1] This project was first funded by the Mellon Foundation to the Center for Research on Information Access at Columbia University.

L. Klavans J., Sidhu T., Sheffield C., Soergel D., Lin J., Abels E. and Passonneau R. (2008).
Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata.
In *Metadata Mining for Image Understanding*, pages 3-12
DOI: 10.5220/0002338100030012

group of searchers submitted more subject-oriented queries than known author and title searches. Similar results demonstrating the importance of subject retrieval have been reported in other studies including [6], [4] and [2].

2 Solutions

The CLiMB project was initiated to address the subject metadata gap under the hypothesis that automatic and semi-automatic techniques may enable the identification, extraction and thesaural linking of subject terms. In particular, the CLiMB Toolkit processes text associated with an image through Natural Language Processing (NLP), categorization using Machine Learning (ML), and disambiguation technologies to identify, filter, and normalize high-quality subject descriptors. Like [9] we use natural language processing techniques and domain specific ontologies, although our focus is on associated text rather than captions.

For this project, we use the standard Cataloging Cultural Objects (CCO) definition of subject metadata[2]. According to this definition, the subject element of an image catalog record should include terms which provide “an identification, description, or interpretation of what is depicted in and by a work or image.” The CCO guidelines also incorporate instructions on analyzing images based on the work of Shatford-Layne (formerly Shatford).[11], building on [8], proposed a method for identifying image attributes, which includes analysis of both the generic and specific events, objects, and names that a picture is “of” and the more abstract symbols and moods that a picture is “about”. Panofsky describes the pre-iconographic, iconographic, and iconologic levels of meaning found in Renaissance art images. Shatford's specific and generic level corresponds to Panofsky's pre-iconographic and iconographic level, respectively, which encompass the more objective and straightforward subject matter depicted in an image. The iconologic level (Shatford's about) addresses the more symbolic, interpretive, subjective meanings of an image. To aid user access, catalogers are encouraged to consider both general and specific terms for describing the objective content of an image (the “of-ness”) as well as to include the more subjective iconologic, symbolic, or interpretative meanings (the “about-ness”). Iconologic terms may be the hardest for catalogers to assign but occur often in texts associated with images.

3 Preparatory Studies of Cataloging

In order to get a better sense of the cataloging process and to inform our system design, we performed some fundamental studies on the process of subject term selection as it is currently undertaken. Our goal was to collect data on the process as a whole in order to improve both our system function (either through rules or statistical methods) and our system functionality (i.e. how to incorporate our results into an existing

[2] http://vraweb.org/ccoweb/cco/parttwo_chapter6.html.

workflow and how to perhaps replace that portion of the workflow with automatic techniques). In this section, we briefly discuss two of these formative studies.

The first study was designed to identify the types of subject terms a cataloger may assign to a given image. Identifying these expert term preferences will help guide the development of heuristic rules for automatically identifying high-quality descriptor candidates and filtering out term types which are rarely assigned manually. Participants were given four stimuli: 1) a hypothetical query for an image; 2) an image; 3) another image—this time with associated text; and 4) an image paired with a list of CLiMB-extracted terms. For the first two stimuli, catalogers were asked to generate subject terms. For the third and fourth stimuli, catalogers were asked to select terms from the associated text and list of terms, respectively. We selected four image/text pairs from the National Gallery of Art Collection. To control for varying textual content which may occur with different art historical genres, we chose one image each from the genres of landscape, portrait, still life, and iconography. Employing a Latin Square Design, the study was administered to 20 image catalogers, recruited through the Visual Resource Association. Each cataloger completed 1 stimuli of the study for each image for a total of 5 participants per version (stimuli-image pair).

Through a combined quantitative and qualitative approach, we analyzed the number of terms assigned per task, the types of terms assigned, and the level of agreement between catalogers in formulating terms for a single concept.

In analyzing the types of terms catalogers assigned to this image, we identified seven categories (in order of frequency): content, place, artist names, period/date, type, style and color. Results for landscape art showed 13 terms for content, 9 for place, 8 for artist names, 7 for period/date, 6 for type, and 4 for style and color. This distribution is typical of the other images, and will help guide the priorities placed on term selection in CLiMB.

For the second study, we took a broader look at the overall image-indexing workflow, including standards, local policies, and actual practices, to determine how the CLiMB Toolkit fits into the cataloging process as a whole. This study not only enabled us to define interface parameters and necessary functionality, it also confirmed the lack of subject access currently provided by human indexers. We examined the similarities and differences in image cataloging practices both within a single institution and across three separate institutions, and at the number and types of subject terms added per catalog record. Within and across these academic visual resource centers, we found that general practices and workflow patterns varied little, and that the number of subject terms entered per catalog record varied but typically fell somewhere between one and eight. One of the primary differences across institutions was the use of different software and metadata schemas, some of which were locally developed. These results indicate that, with flexible export functionality built in to a generic workbench, the CLiMB Toolkit should integrate smoothly with existing practices and different work environments, with little or no tailoring required.

4 CLiMB Architecture

This section describes the techniques we have developed to semi-automatically identify terms which qualify as potential subject descriptors. Our techniques exceed simple keyword indexing by:

- applying advanced semantic categorization to text segments,
- identifying coherent phrases,
- associating terms with a thesaurus, and
- applying disambiguation algorithms to these terms.

CLiMB combines new and pre-existing technologies in a flexible, client-side architecture which has been implemented into a downloadable toolkit, and which can be tailored to the user's needs.

Figure 1 shows the overall architecture of the CLiMB Toolkit. The upper left shows the input to the system, an image, minimal metadata (e.g. image, name, creator), and text. The first stage of CLiMB's processing pipeline associates portions of the input text with images. Note that this requires segmentation, and association of segmented text with the image referred to and described. In clear cases, such as online image captions or in exhibition catalogs, association of image with text is a given. However, in cases where there is a more diffuse relationship between text and image (as in art history texts, for example), it is a computational challenge to ensure that text is accurately associated with the correct image, and not with an image in close proximity (which might or might not be described by the text). This logic creates high-quality associations between text and image, unlike a broader approach which simply grabs text surrounding an image in the hopes that it is relevant to that image.

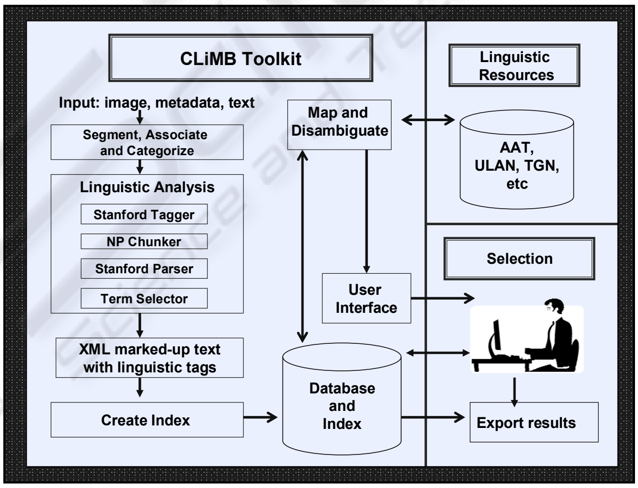


Fig. 1. CLIMB Architecture.

In addition to segmentation, we are developing methods to categorize spans of text (e.g., sentences or paragraphs) as to their semantic function in the text. For example, a sentence might describe an artist's life events (e.g. "during his childhood", "while on her trip to Italy", "at the death of his father"), geographical reference in a work (e.g. "Lake Cuomo"), or style ("impressionism".) A set of categories has been initially proposed through textual analysis of art survey texts, and piloted by asking a range of users to label sample text; using this labeling, initial experiments in machine learning have been conducted to extract features which will permit categorization of sentences. These features can then be used in disambiguation to select between different senses of a term according to its category.

The next phase, Linguistic Analysis, consists of several subprocesses. After sentence segmentation, a part-of-speech (POS) tagger labels (i.e. tags) the function of each word in a text, e.g., noun, verb, preposition, etc. Complete noun phrases can then be identified by the NP chunker based on tag patterns. For example, a determiner, followed by any number of adjectives, followed by any number of nouns, is one such pattern that identifies a noun phrase, as in "the impressive still life drawing". The tagger used for CLiMB, the Stanford tagger[3] provides sentential analysis of syntactic constructions, e.g., verb phrases, relative clauses. The output of Linguistic Analysis consists of XML-tagged words which now contain substantial part of speech tagged and syntactic parsed labels. Lucene is used to create an efficient index for these tagged words.[4]

At this point, the noun phrases stored in the index are input to the disambiguation algorithm, which then enables sense mapping, so that the proper descriptor can be selected from a controlled vocabulary. Words and phrases often have multiple meanings which correspond to different descriptors in a controlled vocabulary but only one may be relevant in context. The ability to select one sense from many is referred to as lexical disambiguation. We map to the appropriate descriptor from the Getty Art and Architecture Thesaurus (AAT), the Getty Union List of Artist Names (ULAN), and the Getty Thesaurus of Geographic Names (TGN).[5] The AAT is a well-established and widely-used multi-faceted thesaurus of terms for the cataloging and indexing of art, architecture, artifactual, and archival materials. In the AAT, each concept is described through a record which has a unique ID, preferred name, record description, variant names, and other information that relate a record to other records. In total, AAT has 31,000 such records. Within the AAT, there are 1,400 homonyms, i.e., records with same preferred name. For example, the term "wings" has five senses in the AAT (see Table 1 below).

Table 2 shows the breakdown of the AAT vocabulary by number of senses with a sample lexical item for each frequency. As with most dictionaries and thesauri, most items have two to three senses, and only a few have more.

[3] Both the tagger and parser are available at: http://nlp.stanford.edu/software.

[4] Lucene is a search engine library: http://lucene.apache.org.

[5] Getty resources can be accessed at:
http://www.getty.edu/research/conducting_research/vocabularies/aat

Table 1. Selection of AAT records for term "wings".

Wings (5 senses):

- Sense#1: wings (costume accessories) Used for accessories that project outward from the shoulder of a garment and are made of cloth or metal.
- Sense#2: wings (visual works components) Lateral parts or appendages of a work of art, such as those found on a triptych.
- Sense#3: wings (theater spaces) The areas offstage and to the side of the acting area.
- Sense#4: wings (furniture components) The two forward extensions to the sides of the back on an easy chair.
- Sense#5: wings (building divisions) Subsidiary parts of buildings extending out from the main portion.

What Table 2 does not illustrate, however, is the semantic distance between senses, i.e. how close the two or three senses might be. This is a measure of the difficulty of disambiguation. Table 2 also does not illustrate the frequency of occurrence in the corpus of these items. Thus, although there may seem to be few terms (types) to disambiguate, they occur with very high frequency (tokens) across the data set. Words with one sense tend to be more rare and highly specialized.

Table 2. Scope of the disambiguation problem in the AAT Thesaurus.

# of Senses	# of Terms	Example	# of Senses	# of Terms	Example
1	29576	scaglioni	8	2	Emerald
2	1097	bells	9	1	Plum
3	215	painting	10	1	emerald green
4	50	alabaster	11	1	Magenta
5	39	wings	12	1	Ocher
6	9	boards	13	1	Carmine
7	5	amber	14	2	Slate

Following standard procedure in word sense disambiguation tasks [7], two labelers manually mapped 601 subject terms to a controlled vocabulary. Inter-annotator agreement for this task was encouragingly high, at 91%, providing a notional upper bound for automatic system performance [5] and a dataset for evaluation. We have used SenseRelate [1], [10] for disambiguating AAT senses. SenseRelate uses word sense definitions from WordNet 2.1, a large lexical database of English nouns, verbs, adjectives, and adverbs.[6]

First, we use all modifiers that are in the noun phrase to find the correct AAT record (Lookup Modifier). We search for the modifiers in the record description, variant names, and the parent hierarchy names of all the matching AAT senses. If this technique narrowed down the record set to one, then we found our correct record. For example, consider the term "ceiling coffers." For this term we found two records:

[6] http://wordnet.princeton.edu/

“coffers” (coffered ceiling components) and “coffers” (chests). The first record has the modifier “ceiling” in its record description, so we were able to determine that this was the correct record. Next, we use SenseRelate to help select the correct WordNet sense of the noun phrase (or its head noun). Using that sense definition from WordNet, we next examined which of the AAT senses best matches with the WordNet sense definition. For this, we used a word overlapping technique which takes senses of WordNet for each polysemous term in AAT and selects the highest value of word overlaps. If none of the AAT records received any positive score (above a threshold), then this technique could not find the best match. Other techniques, Best Record Match and Most Common Sense, are presented in [12].

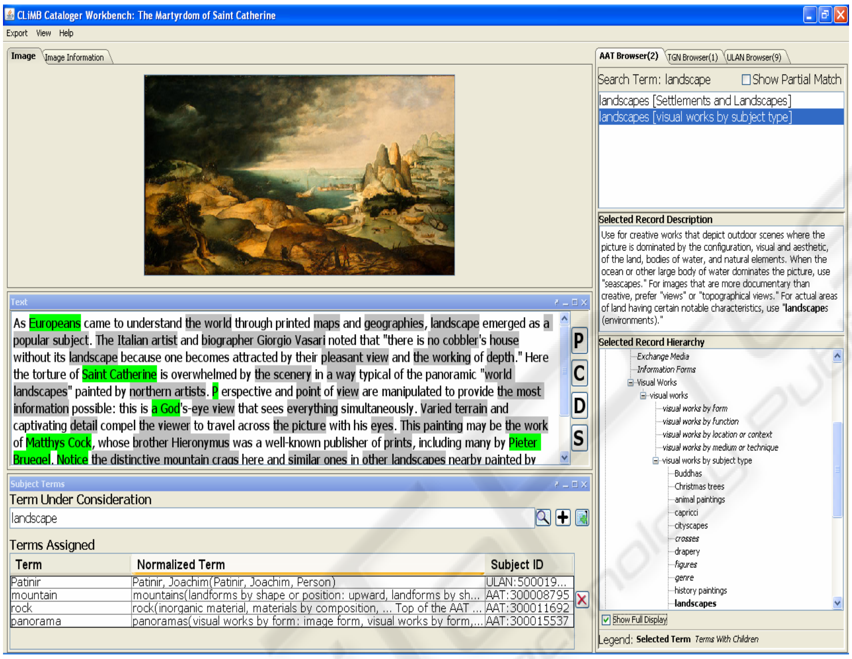
5 Evaluation

Table 3 below shows the breakdown of the data set terms based upon each disambiguation technique. Row 1 in Table 3 shows how few terms were mapped by the lookup modifier technique. In fact, only one was mapped under the Training Set.

Table 3. Breakdown of AAT mappings by Disambiguation Technique.

Row	Technique	Training (n=128)	Test (n=96)
1	Lookup Modifier	1	3
2	SenseRelate	108	63
3	Best Record Match	14	12
4	Most Common Sense	5	18

Rows 2 and 3 show that most of the terms were labeled using the SenseRelate technique followed by the Best Record Match technique. The Most Common Sense technique (Row 4) accounted for the rest of the mappings. An analysis of results and errors shows that our overall accuracy is between 50-55%. General disambiguation tends to run at about 70%, which gives us room for improvement. We are working in a challenging domain with a highly specialized vocabulary. Currently we depend on the external program SenseRelate to perform much of the disambiguation. Furthermore, SenseRelate maps terms to WordNet and we then map the WordNet sense to an AAT sense. This extra step is overhead, and it causes errors in our algorithm. In future work, we will explore the option of re-implementing concepts behind SenseRelate to directly map terms to the AAT. We will explore additional approaches to employ hybrid techniques (including machine learning) for disambiguation. At the present time, we have tested disambiguation and the module will be integrated with into CLiMB as soon as we test results with users. Our plan is to first use results to rank and select a sense for mapping that the user will confirm; once we collect enough feedback from users, we can apply learning to actually map fully and eliminate senses with greater confidence than at present. Figure 2 shows a screen shot of the CLiMB user interface, after having performed a search over images in the National Gallery of Art, and having run the text through the Toolkit.[7] Note that the center top panel contains the image, so the user can look at the item to be approved. The center panel contains the input text, with proper and common nouns highlighted. Under this is the term the user has selected to enter. The right-hand panel gives the thesaural information.



As Europeans came to understand the world through printed maps and geographies, landscape emerged as a popular subject. The Italian artist and biographer Giorgio Vasari noted that "there is no cobbler's house without its landscape because one becomes attracted by their pleasant view and the working of depth." Here the torture of Saint Catherine is overwhelmed by the scenery in a way typical of the panoramic "world landscapes" painted by northern artists. P erspective and point of view are manipulated to provide the most information possible: this is a God's-eye view that sees everything simultaneously. Varied terrain and captivating detail compel the viewer to travel across the picture with his eyes. This painting may be the work of Matthys Cock, whose brother Hieronymus was a well-known publisher of prints, including many by Pieter Bruegel. Notice the distinctive mountain crags here and similar ones in other landscapes nearby painted by

Term Under Consideration

landscape

Terms Assigned

Term	Normalized Term	Subject ID
Patinir	Patinir, Joachim(Patinir, Joachim, Person)	ULAN:500019...
mountain	mountains(landforms by shape or position: upward, landforms by sh...	AAT:300008795
rock	rock(inorganic material, materials by composition, ... Top of the AAT ...	AAT:300011692
panorama	panoramas(visual works by form: image form, visual works by form,...	AAT:300015537

Search Term: landscape □ Show Partial Match

landscapes [Settlements and Landscapes]
landscapes [visual works by subject type]

Selected Record Description

Use for creative works that depict outdoor scenes where the picture is dominated by the configuration, visual and aesthetic, of the land, bodies of water, and natural elements. When the ocean or other large body of water dominates the picture, use "seascapes." For images that are more documentary than creative, prefer "views" or "topographical views." For actual areas of land having certain notable characteristics, use "**landscape**s (environments)."

Selected Record Hierarchy

- *Exchange Media*
- *Information Forms*
- Visual Works
  - visual works
    - *visual works by form*
    - *visual works by function*
    - *visual works by location or context*
    - *visual works by medium or technique*
    - visual works by subject type
      - Buddhas
      - Christmas trees
      - animal paintings
      - capricci
      - cityscapes
      - *crosses*
      - drapery
      - *figures*
      - *genre*
      - history paintings
      - **landscapes**

☑ Show Full Display

Legend: **Selected Term** *Terms With Children*

Fig. 2.CLiMB User Interface for Term "landscape".

At the top of the right are the two senses for the word "landscape" with an indication of where they occur in the AAT hierarchy. Next is the text description of the sense selected. Finally, the entire hierarchy is displayed, bottom right, for the user to view and used to identify any related terms.

As part of the evaluation, we have established a series of test collections with CLiMB partners, in the fields of art history, architecture, and landscape architecture. These three domains were selected in part because of the existing overlap in domain specific vocabulary. Testing with distinct but related domains enables us to test for disambiguation issues which arise in the context of specialized vocabularies. For example, the Art and Architecture Thesuarus (AAT) provides many senses of the term "panel" which apply to either the fine arts, architecture, or both, depending on context. In the context of fine arts, "panel" in the AAT may refer to a small painting on wood whereas in the context of architecture, the same term may refer to a distinct section of a wall, demarked by a border or frame.

[7] In the interest of space, we have included a full screen shot, accompanied by text explanations. If reviewers prefer, this can be enlarged or split into two Figures.

We are currently working with five image-text sets and one image collection for which we are conducting experiments with dispersed texts located online. These six collections will be used for different phases of evaluation, discussed under Future Work. The texts and images for two of the collections, the National Gallery of Art (NGA) Online Collection and the U.S. Senate Catalogue of Fine Arts, can be found online and are in the public domain. For three of the other image collections, The Vernacular Architecture Forum (VAF)[8], The Society of Architectural Historians (SAH)[9], and The Landscape Architecture Image Resource (LAIR)[10], we have secured digital copies of relevant texts along with permissions for use in our testing. The final collection is the Art History Survey Collection, made available to us through ARTstor[11].

6 Future Work

For future work, we have designed a series of studies to test the toolkit in situ. We have partners from several museums and libraries, mentioned in the preceding paragraph, that will first test CLiMB with their cataloging staff, and then who will work with us to design evaluations of Toolkit success in three areas: 1) staff perception on Toolkit ease of use for cataloging within their collections; 2) end user satisfaction with these enhanced records; and 3) several component evaluations, including the named entity recognizer, the noun phrase selector, and the disambiguation component. The proverbial tradeoff between precision and recall may be different for different sectors of the image community; we believe our research in different venues will provide insights on this critical issue. Finally, we intend to explore new directions for integrating CLiMB with current social networking technologies, including social tagging, trust-based ranking of tags, and recommender systems. These technologies offer CLiMB the potential to achieve more personalized results.

Acknowledgements

We acknowledge: the Mellon Foundation; Dr. Murtha Baca, director of the Getty Vocabulary Program and Digital Resource Management, Getty Research Institute for providing us with research access to resources; collections partners, including Jeff Cohen, Bryn Mawr College and University of Pennsylvania; Jack Sullivan, University of Maryland; the Senate Museum and Library and ARTStor.

[8] http://www.vernaculararchitectureforum.org/

[9] www.sah.org/

[10] www.lair.umd.edu/

[11] www.artstor.org

References

1. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, (2003) 805–810 [7].
2. Chen, H.: An Analysis of Image Retrieval Tasks in the Field of Art History. Information Processing & Management, Vol. 37, No. 5 (2001) 701-720.
3. Choi, Y., Rasmussen, E. Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. Journal of the American Society for Information Science and Technology, Vol. 54 (2003) 498-511.
4. Collins, K.: Providing Subject Access to Images: A Study of User Queries. The American Archivist, Vol. 61 (1998) 36-55.
5. Gale, W., Church, K., Yarowsky, D.: A Method for Disambiguation Word Senses in a Large Corpus. Computers and Humanities, Vol. 26 (1993) 415-439.
6. Keister, L.H.: User Types and Queries: Impact on Image Access Systems. In: Fidel, R., Hahn, T.B., Rasmussen, E., Smith, P.J. (eds.): Challenges in Indexing Electronic Text and Images. Learned Information for the American Society of Information Science, Medford (1994) 7-22.
7. Palmer, M., Ng, H.T., Dang, H.T.: Evaluation. In: Edmonds, P., Agirre, E. (eds.): Word Sense Disambiguation: Algorithms, Applications, and Trends. Text, Speech, and Language Technology Series, Kluwer Academic Publishers (2006).
8. Panofsky, E. Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Rowe, New York (1962).
9. Pastra, K., Saggion, H., Wilks, Y.: Intelligent Indexing of Crime-Scene Photographs. In: IEEE Intelligent Systems: Special Issue on Advances in Natural Language and Processing, Vol. 18, Iss. 1. (2003) 55-61.
10. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003).
11. Shatford, S.: Analyzing the Subject of a Picture: A Theoretical Approach. Cataloging & Classification Quarterly, Vol. 6, Iss. 3 (1986) 39-62.
12. Sidhu, T., Klavans, J.L., Lin, J.: Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007), 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic (2007).