# Tagging = Linking
# Descriptive annotations by expert and lay user communities

Jacco van Ossenbruggen

## 2a. Research topic

On the Web, searching, organizing, visualizing and otherwise interacting with visual media assets is facilitated by associated descriptive annotations. Despite recent progress in fields such as content-based image retrieval and computer vision, many applications still require human generated descriptions, and will continue to do so in the near future [Smeulders et al., 2000].

In many domains, experts routinely provide such annotations. Examples range from applied medical domains to fundamental research in fuel cells. In the first, patients' MRI and other scans are annotated by professionals with very specific medical expertise. In the latter, large databases of annotated electron-microscopic images of cell surfaces provide a rich information source [Hunter et al., 2005]. Methods and tools that improve the quality of such annotations positively impact a wide range of disciplines and applications.

Cultural heritage will be used as the application domain, as in this domain the applicant has built up connections necessary to get access to the real world data needed for realistic experimentation. It also allows the research to build on the experiences and software platform developed in the context of the successful BSIK MultimediaN project [van Ossenbruggen et al., 2007, Schreiber et al., 2006]. In this domain, large numbers of cultural artifacts are currently being made accessible on line, in the Netherlands as well as internationally. On line access requires more than the availability of digital images of the artifacts. Search and other meaningful interaction with these images also requires meaningful descriptions of the artifacts to be digitally available [Smeulders et al., 2002].

Such descriptions are traditionally provided by experts of the museums, archives or libraries maintaining the artifacts. Since their institute's website is typically regarded as an authoritative information source about the artifacts being annotated, the experts strive to produce high quality annotations, and new annotations tend to make it to the public website only after a quality assessment procedure. Quality can here best be defined in terms of the Gricean conversational maxims [Grice, 1975]. For example, experts tend to state only what they believe to be true, and avoid claims lacking adequate evidence. Their annotations tend to be directly relevant to the artifact being annotated, and they formulate them without ambiguity and as brief and as specifically as possible. Their terminology, however, may be inaccessible to the general public, and they tend not to annotate aspects of artifacts that users wish to search on. For example, the year of creation of a painting is one of the aspects typically included in every minimal expert description record, while relatively few users actually search paintings by the year of creation. In contrast, in figurative art, users frequently search on the persons or event depicted on the painting, while this aspect is seldom included in professional annotations [Hollink, 2006].

Quantitatively, there is only a small number of domain experts and they are not able to provide sufficient annotations for the large numbers of artifacts being digitized. Current "tagging" platforms on the Web have shown that lay users are able to annotate far larger numbers of assets. Their annotations, or "tags", tend to be free-form textual descriptions, often with a personal bias [Ames and Naaman, 2007]. Compared to expert annotations, tags tend to be more about those aspects of an asset that users search on. In addition, the words used in tags often provide a better match with the terminology used by users in search tasks than the specific jargon often used by experts [Chun et al., 2006]. The downside is, however, that tags that function well in a personal retrieval context, may work less well for search in other contexts. With respects to the institute's authoritative role, publishing lay user annotations on the public website might raise issues when they violate the Gricean conversational maxims. For example, lay user annotations may be factually incorrect, but still be useful for retrieval on the public website (e.g. annotating Rembrandt's Nightwatch as depicting a scene at night is arguably false from a factual point of view but may increase the painting's findability). Annotations that violate the Gricean maxim of manner (e.g. annotations using obfuscated or offensive language) might also harm the institute's reputation.

We focus on three key aspects of annotation: findability, communicative value and ease of creation. Findability refers to the degree to which a particular asset is easy to discover or locate. It depends on the annotations of the asset itself and of other assets, but also on the underlying retrieval system and the search behavior of the users of the system. Communicative value refers to the degree to which the annotations effectively communicate the contextual information that users need when interacting with the asset once it has been found. It depends on the user's task and the effectiveness of the design of the presentation interface. Findability and communicative value are distinct dimensions that need to be measured differently. To improve readability, however, we sometimes refer to these two aspects together simply as the *descriptive value* of an annotation. Finally, ease of creation refers to the degree to which a user can add a specific annotation without requiring unrealistic levels of expertise, time and effort. It depends on the expected level of expertise of the target user, the user's incentives and the design of the annotation interface.

The scientific challenge is to improve upon the state of the art by casting annotation *value* in terms of *findability* and *communication* into a single model, and to support this model by tools that decrease annotation *cost* by deploying Web-scale lay user annotation in combination with user interface components that guide users to enter high valued annotation types. We seek to combine the benefits of having well-defined expert annotations with larger numbers of free-form, personal lay user tags. Based on an analysis of query behavior, web site navigation logs and expert annotation guidelines, we will develop a domain-specific model that predicts the descriptive added value of a specific annotation. We empirically evaluate the model by combining appropriate information retrieval metrics and expert rating.

Based on this model, we design and develop annotation interfaces that combine the ease of use and incentives of current tagging platforms with innovative interface elements aimed at improving annotation quality. We empirically evaluate the interfaces by formal usability testing on lay users with a non-professional interest in art or other areas related to cultural artifacts.

## 2b. Approach

Our method is to develop and evaluate a descriptive model for visual media assets. This model combines the advantages of both professional and lay user descriptions, and covers the full range from, at the one extreme, an annotating strategy based on a small number of annotations, deploying objective, specific and clearly defined roles and terms [Schreiber et al., 2001], to, at the other extreme, descriptions based on larger numbers of personal, free form tags.

The hypothesis is that, for a given domain and specific asset, we can indicate those aspects best served by a few objective and to-the-point annotations, using as specific as possible and well-defined terminology. At the same time, we indicate aspects that are best served by multiple, loosely defined and possibly personally biased tags. The model should have sufficient predictive capabilities that it can, given an analysis of the visual media assets in a certain domain, determine, for each aspect of the media assets, what type of annotation maximizes the asset's findability and/or the conversational quality of the annotations.

For example, in many domains people tend to search primarily on the "who, what, where and when" (or "wh*") properties of an asset. If, for a given asset, a particular property lacks annotations, then supplying an annotation for that property has a higher descriptive value than adding extra annotations to already annotated properties. The missing property annotation would

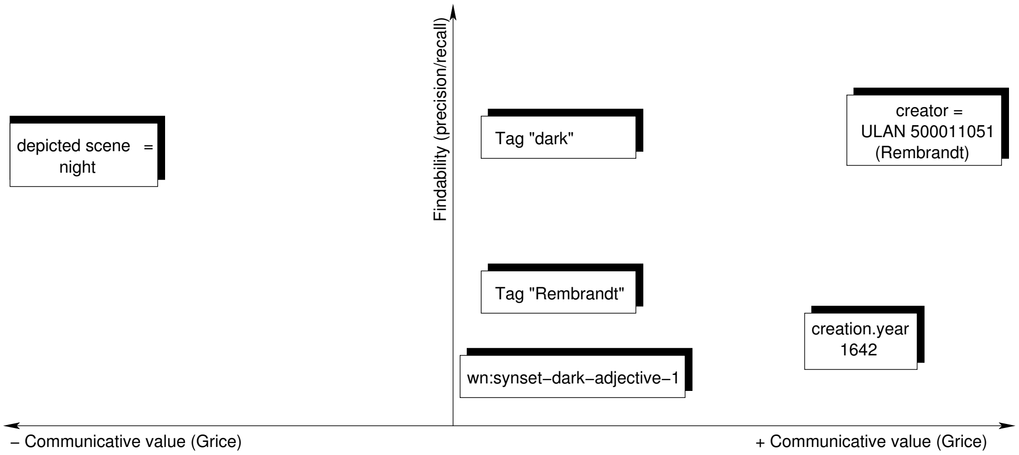
Findability (precision/recall)
depicted scene = night
Tag "dark"
creator = ULAN 500011051 (Rembrandt)
Tag "Rembrandt"
creation.year 1642
wn:synset−dark−adjective−1
− Communicative value (Grice)
+ Communicative value (Grice)

Figure 1: The figure shows example positions for the "Nightwatch" annotations mentioned in the text. Note that annotations might also have a negative impact on findability (not shown in the figure).

thus score higher on the vertical axis shown in Figure 1. In many domains, the wh* properties also tend to communicate the key aspects about an asset effectively, in this case, these annotations would also score high on the horizontal axis.

To give a more domain-specific example, for relatively "objective" properties such as the creator role in the art domain, our model may rate a simple tag "Rembrandt" for representing the painter as low. It may instead give a much higher rating to an annotation that deploys a unique reference to a well defined and authoritative record describing the painter, e.g. as provided by ULAN [Getty Research Institute, 2000b]. An annotation referring to the record would allow dealing with known retrieval issues such as different spelling variants (higher findability). It also allows looking for paintings using other known properties of the painter, e.g. a painting annotated as created by Rembrandt but not annotated by a place and year of creation, can still be found by a query for Dutch 17th century art – just by looking up the times and places where Rembrandt was active. The record might also be used to look up the preferred spelling of Rembrandt's name or be a source of extra navigation paths when the asset is presented on the institute's website (higher communicative value) [van Ossenbruggen et al., 2007]. In contrast, for more subjective roles such as "depicted scene", the model might rate string-based tags higher on findability, for example because it is known that such tags better take advantage of useful ambiguity in natural language, e.g. a tag "dark" can mean both that the scene depicted is lacking in light, and that it expresses a dark and depressing mood. Both senses of dark might apply to the painting, and a disambiguated annotation explicitly referring to a single sense of the adjective "dark", for example in WordNet [Fellbaum, 1998], might be thus lower valued than the tag "dark" (see also Figure 1). The same tag "dark" might provide little or no extra information during the display of the asset, in which case it would score low on communicative value.

Note that the examples above are just that: examples. What makes an effective rating model is part of the research question. The effectiveness of the model is empirically evaluated on real world data. We approximate an asset's findability by standard information retrieval metrics (e.g. precision and recall), based on queries and search algorithms that are representative for the domain and data set. High findability implies that the query terms typically used by users searching for the asset a) return the asset (recall) and b) return few other results (precision). Note that this is an approximation, since we abstract from the impact of the search and navigation interface on an asset's findability. The communicative value of a (set of) annotations will be empirically measured by lay and expert user ratings and by example production experiments. Participants in experiments will be asked to, in the context of a specific task, either directly rate the value of given

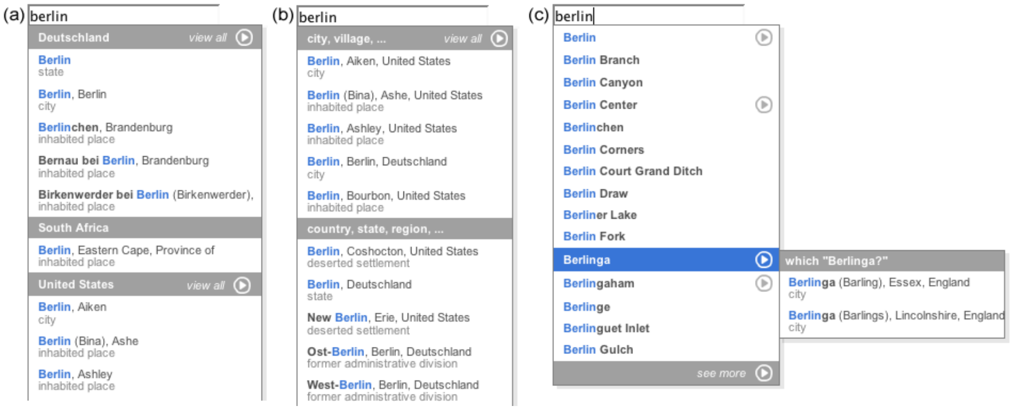

Figure 2: Example screen shots from different interface designs for annotating place names. While typing, users are presented with suggestions grouped on (a) country, (b) place type or (c) place name. Note that none of the designs succeeds in displaying the more than 50 places named "Berlin" in the US. (Suggestions are based on data from TGN [Getty Research Institute, 2000a]. Figures have been adapted from [Hildebrand et al., 2007].)

pieces of descriptive information, or, given a specific asset with no descriptions add the top-N most valuable pieces of information. Again, we strive to evaluate the model itself and abstract from the design of an associated presentation interface.

The model should thus be able to quantify annotation *value* and predict the types of annotation that will yield the highest value. We also claim, however, that users can create such high value annotations against a sufficiently low *cost* in terms of usability. To substantiate this claim we perform user annotation experiments in a realistic setting. These experiments will be based on prototype interfaces that combine the best of both lay user tagging and professional annotation interfaces. Development of the prototypes requires a balanced mix of model-driven design with the iterative design, evaluate, redesign cycle that is common in user-centered interface design (UCD).

Interface components to be researched include autocompletion interfaces [Hildebrand et al., 2007, Hyvönen and Mäkelä, 2006], such as the example components depicted in Figure 2. These interfaces allow guided term disambiguation and targeted suggestion of more or less specific terms in an interaction style many lay users are accustomed to. While autocompletion interfaces have the potential to significantly enhance annotation descriptiveness and reduce annotation effort, a crucial success factor is their ability to maximize the likelihood that the suggestions contain a) the term the user is looking for, and/or b) a term that the user recognizes as an improvement upon the term she was intending to enter. Screen real estate in this type of interface is limited and the number of potential suggestions that syntactically match the partial user input is, in all real world applications, too large to display. The same applies to the display of potential suggestions that aim at improving upon the term intended by the user. State of the art autocompletion interfaces ignore ranking or simply rank on popularity, not quality. The proposed model provides a strong foundation for suggestion filtering and ranking based on the descriptive value of the suggestions.

Forcing or even encouraging users to disambiguate tag roles and values can either be beneficial or harmful. The interface needs to take the descriptive model into account to address this and only guide users to more specific terms if this actually improves the overall descriptiveness of the set of annotations associated with the asset. It may thus need to steer users to common, pre- and well-defined terms in one case, while in other cases it may stimulate users to find new, alternative descriptions, potentially using terms that are not (yet) part of the system's internal vocabularies. It might also steer users towards properties that are known to be often used in search but are currently not part of the annotations, or change annotation strategy depending on the dimension. For example, for an objective dimension such as the creator example used earlier, the interface might show the user that many users have already annotated the painting as created by Rembrandt, and offer an interface where the user can express her agreement with the previous users, or explicitly state why she disagrees and who she thinks the creator should be. In contrast, for more subjective properties such as the scene depicted by the painting, the interface might choose to hide annotations made by other users because a ”fresh” look might yield new and valuable terms.

More in general, the social structures in the data set could provide another source of useful metrics. Traditionally, annotated data sets form (only) a bipartite graph where assets are connected to one another through the annotations they share. This graph can, for example, be used in a navigation interface to provide links to other assets with similar annotations. In tagged data sets, however, for each annotation the user, the tag and the target asset is known, and the data set can thus be seen as a richer, tripartite graph of users, media assets and tags. Because the users play an essential role in this graph, it can be regarded as a type of social network, and existing metrics for social network analysis can often be applied. In the interface, such metrics can be used, for example, to rank users (e.g. in terms of the amount of their content tagged by others, or in terms of the number of tags that they share with others, etc) or to provide social navigation paths (e.g. link to other users that use similar tags). These social interface examples can both be seen as a potential annotation incentive, but also as a from of feedback on the effect of an annotation (e.g. adding an annotation that unexpectedly turns out to have never been used before may indicate to the user that the annotation is using too obscure jargon, while an unexpected high number might indicate that the annotation is too general).

Both the bipartite graphs of traditional expert annotation and the tripartite graphs of current tagging platforms are “closed” data models, to the extent that user profiles, annotations and media assets are contained and controlled from within a single application. Due to recent progress in Web technology for the interoperable representation of heterogeneous linked data sets, annotation strategies may now go beyond this closed model. In the BSIK MultimediaN project, for example, we experiment with annotations that are modeled as URI-based references to concepts defined by multiple, heterogeneous RDF vocabularies [W3C, 1999]. A natural next step is to allow URI-based annotations to arbitrary resources on the Web, so annotations are no longer limited to simple tags or internal vocabulary entries, but can refer to terms defined by vocabularies that are external to the application, and to related HTML pages or other audio-visual assets. For example, an ethnological museum may use its own vocabulary to annotate the cultural provenance of a contemporary African work of art, but may use the vocabulary of a modern art museum to indicate that the work is painted in the “Pop art” style. Lay users may annotate a Picasso painting with a link to an on-line news video, reporting on the painting’s return after having been stolen from the museum. In such an open annotation environment, the quantification of the impact of the annotation on the findability, its communicative value in describing the asset, and its ease of creation remain key aspects.

Again, all examples given are above just that, the actual design of the interface components is part of the research question. The developed interfaces should provably perform comparably to, or better than, state of the art tagging platforms in terms of usability; and produce significantly more descriptive annotations than state of the art tagging platforms. To substantiate these claims, we will perform experiments with lay user participants comparing the usability of representative tagging interfaces against our prototype interfaces. We also compare the descriptive value of the annotations produced by the different interfaces.

## 2c. Innovation

We strive to reconcile the usability aspects of current tagging interfaces with the quality requirements of annotations used in a professional context. Other state of the art approaches either: a) apply current tagging interfaces directly into a professional context (e.g. steve.museum, PowerHouseMuseum.com), thereby ignoring quality issues and losing the advantages of annotation based on pre-defined, existing terms; or b) force formal and strict annotation interfaces on lay users that are used to “free” tagging, thereby losing the advantages of tagging, and, potentially, losing a significant fraction of the user community.

Second, our approach allows tagging and searching across the boundaries of individual institutes (e.g. users may annotate works from one museum with relevant terms from the vocabulary from another museum). We reuse information already available from the Web, where comparable approaches are based on a closed world view, and all descriptions needs to be (re)created inside isolated systems. This not only gives the annotator more freedom in choosing the most appropriate term, but it also makes meaningful relationships explicit across distributed collections) and across the boundaries of domains. In addition, where closed approaches can only exploit the tag/content/user information network structure that is local to the system, we can also exploit the open link network structure of the Web.

## 2d. Plan of work

The research will be carried out by two PhD students and the applicant. The first PhD student should have a strong background in human computer interaction and preferably experience in performing user studies. In year 1, the work will focus on analysis of website query logs and navigation behavior to better understand how users search and interact in the test collection. Together, we will design and perform experiments to measure the communicative value of different types of annotations in the test domain.

The second PhD student should have a strong modeling background and sufficient technical skills to develop the annotation interface prototypes that support the model. Together, we will develop the model that predicts the descriptive value of different types of annotation. The associated software development is expected to build on and significantly contribute to the open source experimentation platform and interface components initially developed within the MultimediaN E-culture project.

The research will be carried out in the research group of the applicant, the Semantic Media Interfaces group (INS-2) at CWI. The work on measuring and evaluating the findability metrics will be done in close cooperation with information retrieval researchers of INS-1. User studies and usability experiments will be done by continuing existing cooperations with the Human-Computer Studies (HCS) Laboratory at the University of Amsterdam. Software development and the data engineering of the annotations will be based on existing relationships with both the HCS lab and Guus Schreiber's research group at the computer science department of VU University Amsterdam.

The applicant and his research group continue to be involved in future standardization activities around multimedia descriptions of the World Wide Web Consortium.

## 2e. Literature

## References

[Ames and Naaman, 2007] Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In Rosson, M. B. and Gilmore, D. J., editors, *CHI*, pages 971–980. ACM.

[Chun et al., 2006] Chun, S., Cherry, R., Hiwiller, D., Trant, J., and Wyman, B. (2006). Steve.museum: An ongoing experiment in social tagging, folksonomy and museums. In *Museums and the Web*, Albuquerque.

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database.* Language, Speech, and Communication Series. MIT Press.

[Getty Research Institute, 2000a] Getty Research Institute (2000a). The Getty Thesaurus of Geographic Names (Online). http://www.getty.edu/research/tools/vocabulary/tgn/.

[Getty Research Institute, 2000b] Getty Research Institute (2000b). Union List of Artist Names (Online). http://www.getty.edu/research/conducting_research/vocabularies/ulan/. Version 2.0.

[Grice, 1975] Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics*, volume 3. Academic Press, New York.

[Hildebrand et al., 2007] Hildebrand, M., van Ossenbruggen, J., Amin, A., Aroyo, L., Wielemaker, J., and Hardman, L. (2007). The design space of a configurable autocompletion component. Technical Report INS-E0708, CWI. Submitted to the 17th International World Wide Web Conference — WWW2008.

[Hollink, 2006] Hollink, L. (2006). *Semantic annotation for retrieval of visual resources.* PhD thesis, Vrije Universiteit, Amsterdam.

[Hunter et al., 2005] Hunter, J., Cheung, K., Little, S., and Drennan, J. (2005). FUSION - a knowledge management system for fuel cell optimization. In *International Conference on Solid State Ionics*, Baden.

[Hyvönen and Mäkelä, 2006] Hyvönen, E. and Mäkelä, E. (2006). Semantic Autocompletion. In *Proceedings of the first Asia Semantic Web Conference (ASWC 2006)*, pages 739–751, Beijing.

[Schreiber et al., 2001] Schreiber, A., Dubbeldam, B., Wielemaker, J., and Wielinga, B. (2001). Ontology-based Photo Annotation. *IEEE Intelligent Systems*, 16(3):66–74.

[Schreiber et al., 2006] Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenjko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., and Wielinga, B. (2006). MultimediaN E-Culture Demonstrator. In *The Semantic Web - ISWC 2006*, pages 951–958.

[Smeulders et al., 2000] Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval: the end of the early years. *IEEE trans.*, 22(12):1349–1380.

[Smeulders et al., 2002] Smeulders, A. W., Hardman, L., Schreiber, G., and Geusebroek, J.-M. (2002). An integrated multimedia approach to cultural heritage e-documents. 4th Intl. Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia 2002.

[van Ossenbruggen et al., 2007] van Ossenbruggen, J., Amin, A., Hardman, L., Hildebrand, M., van Assem, M., Omelayenko, B., Schreiber, G., Tordai, A., de Boer, V., Wielinga, B., Wielemaker, J., de Niet, M., Taekema, J., van Orsouw, M.-F., and Teesing, A. (2007). Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. In *Museums and the Web 2007*.

[W3C, 1999] W3C (1999). Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendations are available at http://www.w3.org/TR.