Illustrations by mwienerarts.com

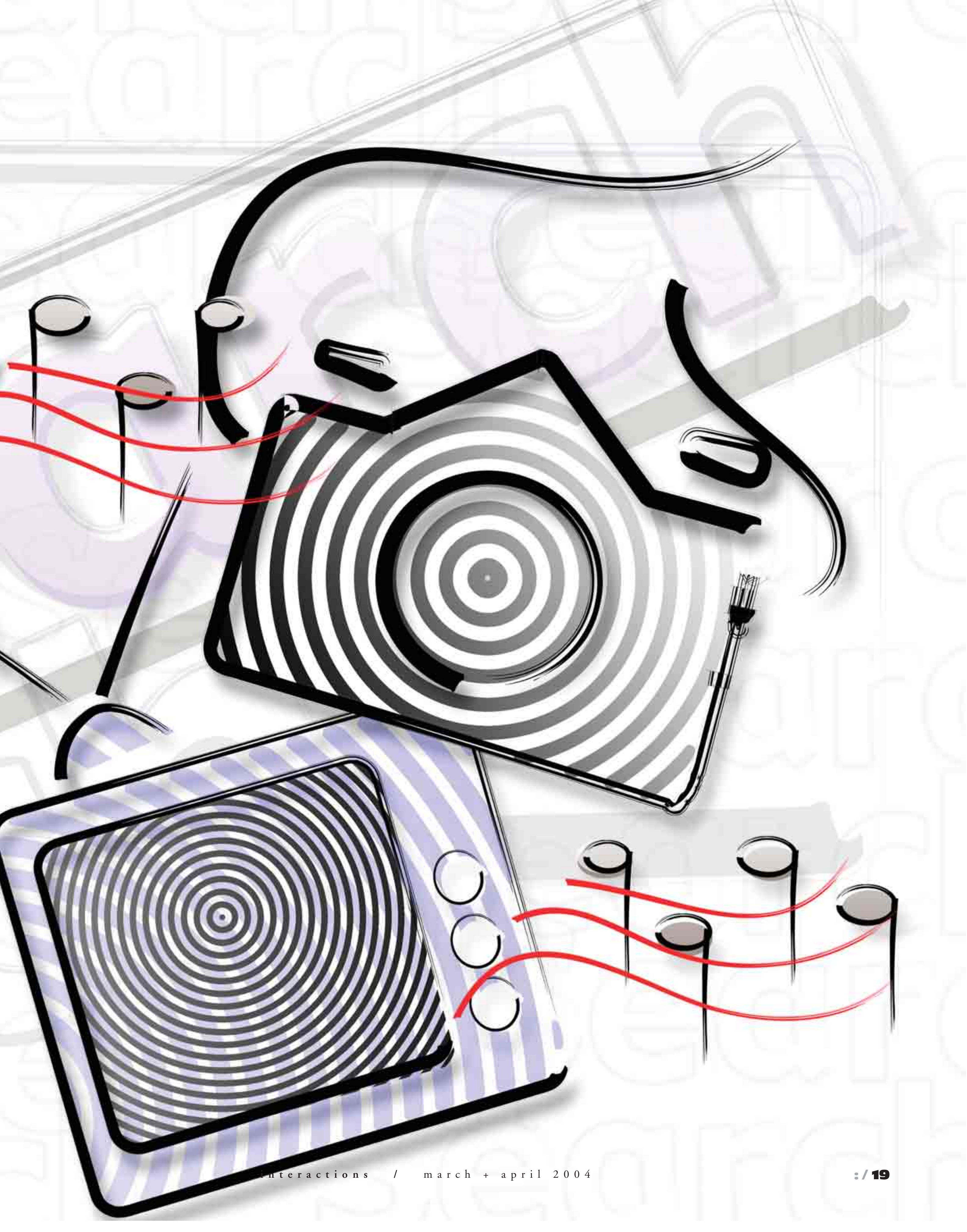
BabelVision

Better Image Searching Through Shared Annotations

By Ken Haase and David Tamés
beingmeta, inc.
kh@beingmeta.com
david@beingmeta.com

The Problem with Search Technology: I Still Haven't Found What I'm Looking For

As the volume of material online increases, it becomes more difficult and time consuming for people to find what they are looking for [5]. Finding the right document depends on how well query terms match keywords or other document descriptors. In the case of text documents, descriptors can be generated from an automatic analysis of the documents' content and structure [1]. But what happens when the "document" is not text but video, sounds, photos, or other images? These media types do not lend them-



selves to automated content analysis. Instead, these nontextual documents are generally annotated by professional librarians or archivists who are skilled at identifying what is most important about the document (salience) and in selecting descriptive keywords that are precise and specific enough for most searches. Although this approach generates high-quality descriptors, it does so at a cost. With the increase in images stored by individuals and businesses, there is a pressing need for an alternative method of finding images and other media documents.

BabelVision offers a counterintuitive twist in the development of search technologies: It helps people find images by making it easier for nonexperts to annotate the image. This article describes BabelVision, a concept-based image annotation and search prototype, and a trial of the technology with an annotation team made up of inner-city high school students in Boston.

Keyword Roulette

The problem with nonexpert annotation is that keywords chosen by nonlibrarians tend to be ambiguous and at different levels of description. This results in similar documents being described in different ways, and searchers are forced to guess which descriptive keywords might have been used to describe the content they are seeking.

The traditional solution to this problem is for both annotators and searchers to use a controlled vocabulary[1] of terms. Unfortunately, traditional controlled vocabularies can be difficult for nonlibrarians to use and often do not allow the kind of precise description that users like to provide while annotating or searching. Controlled vocabularies can be extended but, if not carefully managed, can tend toward the ambiguity and confusion present in natural languages.

An extensible controlled structured vocabulary (ECSV), consisting of terms and their relations, provides a more systematic approach to creating an easy-to-use, controlled vocabulary. New terms are added when they relate to existing terms and are included in the search computation, making descriptions more precise and still usable. BabelVision uses an ECSV called BRICO [6] (after the French word *bricolage*) that contains roughly half a million concepts capturing (in ambiguous relationships) roughly the same number of words. BRICO includes substantial (though incomplete) mappings into a number of other Western languages: Spanish, Portuguese, French, German, Italian, and Dutch.

Annotating Images with Babelvision

In order to annotate or search images, users access BabelVision via a Web browser. The home page (Figure 1) presents the user with a collection of images to annotate.

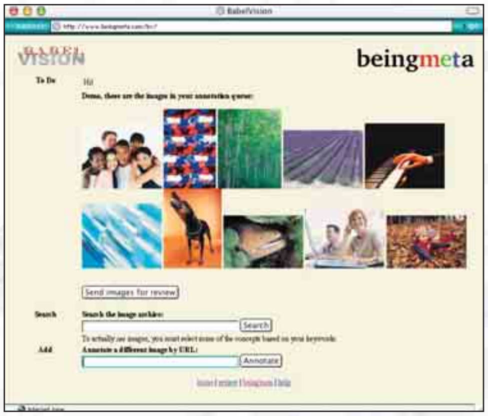
For example, clicking the image on the far right of the top row takes the user to a page with the image flanked by an "Add Concepts" text field. When the user types a word or phrase in this field, BabelVision returns the concepts related to that word or phrase. For example, if the user types "piano," BabelVision returns four concepts (Figure 2).

BabelVision is telling us that "piano" is ambiguous—that is, four distinct concepts are associated with "piano." By clicking the check boxes next to the concepts, the user can disambiguate the term, that is, tell BabelVision specifically which of the four concepts to associate with the image. For the image in our example, the user selects the third concept, "piano...a stringed instrument...," and adds several concepts associated with the word "music." After adding some other concepts, the annotation page eventually looks like Figure 3. BabelVision is quite flexible; it accepts a wide range of words and phrases and finds concepts from a very large vocabulary.

For the user, the disambiguation task is not much harder than typing keywords and requires no knowledge of a controlled vocabulary. BabelVision performs the hard work behind the scenes to associate precise concepts with the image. In this manner, BabelVision transforms the complex task of annotation into free association and concept recognition.

Searching for Images with Babelvision

Search works similarly to annotation. For example, if the user types "piano player" to search for an image of a piano being played, BabelVision returns the concept for "piano player, pianist, a person who plays the piano." In this case the term is not ambiguous and thus maps to a single concept (unlike "music" and "piano," which have many concepts associated with them). Clicking the "piano player" concept will reveal thumbnails of images in the database that have



Hi!

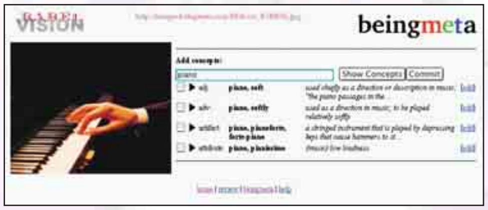
Dema, these are the images in your annotation queue:

Search the image archive:

To actually see images, you must select some of the concepts based on your keywords.

Annotate a different image by URL:

Figure 1. BabelVision starting page.

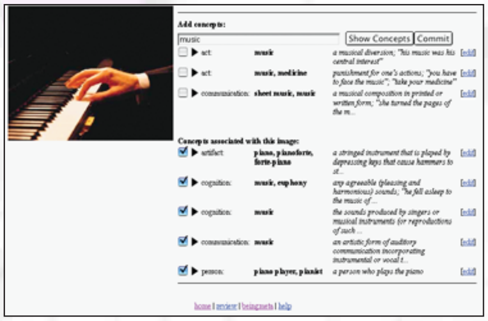


Add concepts:

piano

□ ▶	adj	piano, soft	used chiefly as a direction or description in music; "the piano passages in the ...	[edit]
□ ▶	adv	piano, softly	used as a direction in music; to be played relatively softly	[edit]
□ ▶	artifact	piano, pianoforte, forte-piano	a stringed instrument that is played by depressing keys that cause hammers to st...	[edit]
□ ▶	attribute	piano, pianissimo	(music) low loudness	[edit]

Figure 2. Concepts associated with "piano."



Add concepts:

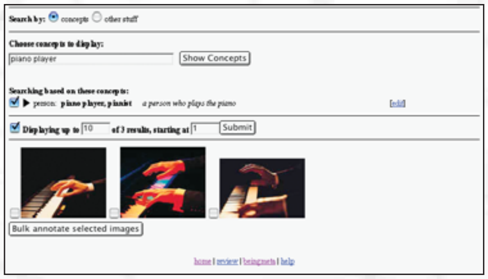
music

□ ▶	act:	music	a musical diversion; "his music was his central interest"	[edit]
□ ▶	act:	music, medicine	punishment for one's actions; "you have to face the music"; "take your medicine"	[edit]
□ ▶	communication:	sheet music, music	a musical composition in printed or written form; "she turned the pages of the m...	[edit]

Concepts associated with this image:

☑ ▶	artifact:	piano, pianoforte, forte-piano	a stringed instrument that is played by depressing keys that cause hammers to st...	[edit]
☑ ▶	cognition:	music, euphony	any agreeable (pleasing and harmonious) sounds; "he fell asleep to the music of ...	[edit]
☑ ▶	cognition:	music	the sounds produced by singers or musical instruments (or reproductions of such ...	[edit]
☑ ▶	communication:	music	an artistic form of auditory communication incorporating instrumental or vocal t...	[edit]
☑ ▶	person:	piano player, pianist	a person who plays the piano	[edit]

Figure 3. Image annotated with various concepts.



Search by: ● concepts ○ other stuff

Choose concepts to display:

piano player [Show Concepts]

Searching based on these concepts:

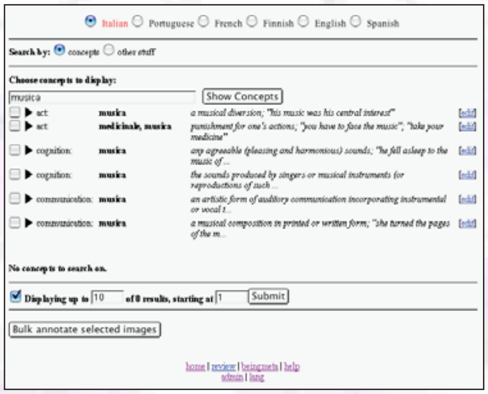
☑ ▶	person:	piano player, pianist	*a person who plays the piano*	[edit]

☑ Displaying up to 10 of 3 results, starting at 1 [Submit]

[Bulk annotate selected images]

home | review | beingmeta | help

Figure 4. Searching by concept "person: piano player."



● Italian ○ Portuguese ○ French ○ Finnish ○ English ○ Spanish

Search by: ● concepts ○ other stuff

Choose concepts to display:

musica [Show Concepts]

□ ▶	act	musica	*a musical diversion; "his music was his central interest"*	[edit]
□ ▶	act	medicinale, musica	*punishment for one's actions; "you have to face the music"; "take your medicine"*	[edit]
□ ▶	cognition:	musica	*any agreeable (pleasing and harmonious) sounds; "he fell asleep to the music of ...*	[edit]
□ ▶	cognition:	musica	*the sounds produced by singers or musical instruments (or reproductions of such ...*	[edit]
□ ▶	communication:	musica	*an artistic form of auditory communication incorporating instrumental or vocal t...*	[edit]
□ ▶	communication:	musica	*a musical composition in printed or written form; "she turned the pages of the m...*	[edit]

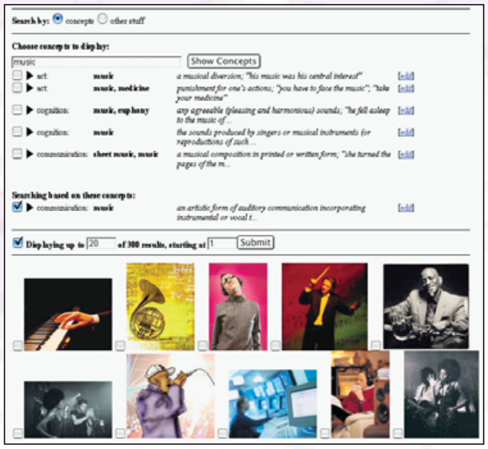
No concepts to search on.

☑ Displaying up to 10 of 0 results, starting at 1 [Submit]

[Bulk annotate selected images]

home | review | beingmeta | help
admin | lang

Figure 5. Search using the Italian word "musica."



Search by: ● concepts ○ other stuff

Choose concepts to display:

music [Show Concepts]

□ ▶	act:	music	*a musical diversion; "his music was his central interest"*	[edit]
□ ▶	act:	music, medicine	*punishment for one's actions; "you have to face the music"; "take your medicine"*	[edit]
□ ▶	cognition:	music, euphony	*any agreeable (pleasing and harmonious) sounds; "he fell asleep to the music of ...*	[edit]
□ ▶	cognition:	music	*the sounds produced by singers or musical instruments (or reproductions of such ...*	[edit]
□ ▶	communication:	sheet music, music	*a musical composition in printed or written form; "she turned the pages of the m...*	[edit]

Searching based on these concepts:

☑ ▶	communication:	music	*an artistic form of auditory communication incorporating instrumental or vocal t...*	[edit]

☑ Displaying up to 20 of 300 results, starting at 1 [Submit]

Figure 6. Search results using concept "communication: music."

been annotated with the concept "piano player" (Figure 4).

BabelVision is interlingual, that is, users can perform annotation in one language and search using a different language. For example, searching for "musica" in Italian yields a similar, yet not identical, set of concepts as the search would for "music" in English (Figure 5). The concepts BabelVision presents to the user are slightly different, reflecting the continuities and cultural differences in the meanings of the term "musica" in Italian and the term "music" in English.[2]

Note that previously, when the user annotated the image of the hand and the piano keys, the only terms they had to think of were "music" and "piano" and the rest of the work involved selecting from concepts related to those terms. Searching works similarly. The user types a familiar term and then disambiguates the search by choosing additional concepts from a list presented by BabelVision. In return for this simple additional step, the user is rewarded with high-precision search results. Figure 6 shows the result of searching on a general concept such as "music"—a lot of images, but high-precision images, are returned. The first image was the one originally annotated by our user. This is not a trick. The system has a preference for those images that were most recently annotated. BabelVision, it must be noted, uses concepts not keywords in annotations. "Music," for example, is a concept not a keyword.

Behind the Scenes

The preceding example used an extensive knowledge base to enable a nonexpert annotator to provide rich descriptions of image content. By replacing the librarian's *recall task* (finding relevant controlled terms) with a simpler *recognition task*, the nonexpert can annotate images unambiguously and with great precision.

An ECSV solves two problems at the same time: ambiguity and specificity. It addresses ambiguity by separating out different core meanings of a term. For example, the word "fire" would have descriptions for meanings that include destructive burning, involuntary termination, the onslaught of projectile weapons, and more metaphorical interpretations. It addresses specificity by connecting concepts to one another so that the concept of destructive burning is tied to more general concepts (such as fiery combustion, which includes explosions or internal combustion) and more specific concepts (such as camp fires, forest fires, and house fires).

Use of an ECSV (extensible controlled structural vocabulary) can enable precise recall while not hampering more general recall.

The annotation example dwelled on the ambiguity problem, but in looking at search we saw how the use of an ECSV can enable precise recall while not hampering more general recall.

How Well Did It Work?

We conducted a five-month study [7] to determine how well the annotations created by non-experts using BabelVision compared with those created by professional library scientists. We used a diverse collection of images[3] characteristic of contemporary stock photo collections. The images were already annotated by experts with keywords from a controlled vocabulary, thus serving as the basis for comparison. We evaluated the results in two ways. First, we looked at whether different annotators using BabelVision were describing the same image in different ways. Second, we compared their annotations with "expert annotations" produced by mapping the original controlled vocabulary keywords into the BRICO conceptual language [6] used by BabelVision.

Students ranging from 14 to 17 years old were drawn from the racially, ethnically, and economically diverse community around the beingmeta headquarters in Dorchester, Massachusetts (a part of Boston). The students were compensated for their involvement in an after-school annotation program—based at the Codman Square Technology Center[4]—that ran throughout the spring and early summer of 2003. Annotators logged approximately 1,900 hours and were supervised by a trainer/manager; experienced annotators frequently trained those new to the project.

The primary evaluation metric was based on comparing different annotators' descriptions of the same images. We looked at a number of direct and derived measures, but the two most revealing metrics were the following:

- *Overlap*: a count of the common annotations between two annotators, weighted to reflect concepts that are not identical but are closely related in the concept language
- *Failure rate*: the percentage of images between two annotators that have no common annotations

BabelVision creates a partnership between the human skills that are well suited to categorizing images and the computational power of adding precision to the description.

Over all of the annotators, we found an average overlap of 1.16 annotations and an average inter-annotator failure rate of 20 percent. These numbers were statistically significant and indicated that different annotators were generally describing the same images in similar ways. However, two patterns were clear as we looked at the data. First, the convergence of annotations between different annotators was much higher when we looked at the results of the six best annotators. Second, low overlap tended to reflect the annotators' choice of different features to describe rather than their selection of different concepts for the same features. Regarding the first observation, reducing the scope of the analysis to the six annotators who seemed to do best yielded a mean overlap of 1.48 concepts and a failure rate of 15 percent.

The foregoing numbers give some idea of how much different annotators' descriptions diverged or converged, but one important question is how their annotations compare with expert annotations. We examine this by converting the controlled vocabulary keywords for the DVO collection into BRICO concepts and comparing the annotators with this "expert annotator."

In this comparison, the average overlap score was 1.52 and the average failure rate was 15 percent. Restricted to the best annotators, the overlap came in at 1.78 and the failure rate averaged 10 percent. Interestingly, when the annotators were taken as an aggregate (as though their collective annotations were done by a single person), the overlap score was

2.7 and the failure rate fell to five percent.

We were disappointed by the relatively high failure rate (15 percent), that is, images for which there was no overlap in annotations between the two groups. When we looked at the data more closely, we found that one source of this problem was the existence in BRICO of fine-grained distinctions such as distinguishing {city | metropolis} (the social entity) from {city | metropolis} (the location). These distinctions had different concepts in BRICO but these concepts were not systematically connected to one another. Similarly, some of our participants repeatedly used the concept {tree, tree diagram} to describe botanical trees whereas the experts used the {tree: woody plant} for these images. These gaps might be addressed with both extensions to the ontology[5] itself and changes to the user interface for BabelVision that would make it easier to find the right concept using, perhaps, iconic representations as well as text.

Do non-experts using BabelVision do as well as expert archivists using a controlled vocabulary? No, but for a variety of different reasons. First, the experts tended to produce much more detailed descriptions (10 to 20 keywords), whereas our annotators tended to use only three or four concepts. Second, the experts only occasionally selected the wrong term from the controlled vocabulary, and BabelVision annotators more frequently chose the wrong term, resulting in the 10 to 20 percent failure rates cited earlier.

Can nonexperts using BabelVision usefully annotate images? Yes. The overlap numbers clearly indicate that the annotators picked many of the same concepts (from a huge space) to describe the same images. This convergence demonstrates that different annotators or (presumably) searchers would find their way to the same concepts and to the images annotated with them.

What still needs to be addressed? Reducing the failure rate—using some of the preceding strategies—is of paramount importance. In addition, although BRICO is a large knowledge base, it can’t represent everything and users will need to extend it to address new domains and more precise categorizations. One crucial question is whether nonexperts can also accomplish the extension of ontologies. Finally, the nonexperts in this study were supervised and it will require some work—at the knowledge, interaction, and interface levels—to enable the system to work for isolated nonexperts.

Disruptive Potential

Information technologies, including digital imaging, word processing and the Internet, are disruptive in that they confer a level of access and performance to ordinary individuals that was previously only afforded to the privileged few. BabelVision, by providing easy access to an intelligent concept database, has the same potential to augment the skills of regular users to approach those of the skilled librarian. The skills of librarians will still be needed for large, specialized image libraries for which expertise in providing precise terms has significant value.

The specific problems addressed by BabelVision are harbingers of the design issues involved in the evolution of the semantic Web [3] and especially in description standards like RDF [2], RDF-Schema [4], and efforts to standardize access to ontologies [8]. These standards provide the foundation for document descriptions that automated search tools can use to find items on the Web. For images, which are opaque to automated annotation methods, the problem remains: How do we create the descriptions in the first place? BabelVision creates a partnership between the human skills that are well suited to categorizing images and the computational power of adding precision to the description.

BabelVision brings the benefits of expert annotation to the growing flood of images, such as personal photo blogs, for which expert annotation is not economically feasible but for which there is a desire to exchange with others.

EDITORS

Kate Ehrlich
Collaboration User Experience Group
IBM Research
One Rogers Street, Cambridge, MA 02142
617-693-1170 katee@us.ibm.com

Austin Henderson, Director,
Systems Laboratory Advanced Concepts & Design Pitney Bowes
35 Waterview Drive MS 26-21, Shelton, CT 06484
203-924-3932 austin.henderson@pb.com

ACKNOWLEDGMENTS

BabelVision and BRICO were initially developed by Ken Haase and the students of the Machine Understanding Group at the MIT Media Laboratory and funded in part by the News in the Future Consortium. Ben Guaraldi contributed to the workflow management tools and the version of BabelVision used in the study. We would like to thank John Evans and the staff of the Codman Square Technology Center for their support, and special thanks are given to the participants in the study who annotated thousands of images and helped evolve the workflow tools and interface: Terrence Galloway, Teneria Walton, DaVaughna Johnson, Rhyheema Burns, Shemere Johnson, Ollie Bodden, Tola Abraham, Shakeyah Scroggins, Ernest Greer, Kimberly Mosby, Timeka Bembridge, Natasha Henry, Tenisha Raspberry, Shauna Pulley, and Sheralee Martin. This project was funded in part by an SBIR Phase I grant from the National Science Foundation, Project number 0232731.

FOOTNOTES

1. Some of the best known controlled vocabularies are those used by the Library of Congress; for example, the Thesaurus of Graphic Materials I (TGM-I) consists of thousands of terms for indexing visual materials. See www.loc.gov/lexico/servlet/lexico/
2. In Figure 5 the terms are in Italian but the concept descriptions are in English as they have not been translated to Italian in the current deployment of BabelVision.
3. One of the image collections, the DVO Collection, was provided to beingmeta for research purposes by Digital Vision Online, www.digitalvisiononline, a provider of stock photography.
4. The Codman Square Technology Center is a neighborhood center providing services to both children and adults and is run under the auspices of the Codman Square Health Center, www.codman.org
5. An ontology defines the terms used to describe and represent an area of knowledge. An ontology represents semantics and enabling the semantics to be used by computer programs, which is particularly important for applications that search across or merge information created by diverse communities. BabelVision uses BRICO as its ontology.

REFERENCES

1. Bazea-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press, New York, 1999.
2. Beckett, D. ed. RDF/XML Syntax Specification." W3C Proposed Recommendation. Available at www.w3.org/TR/2003/PR-rdf-syntax-grammar-20031215
3. Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American* (May 2001).
4. Brickley, D. and Guha, R.V., eds. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Proposed Recommendation. Available at www.w3.org/TR /2003/PR- rdf-schema-20031215
5. Feldman, S. and Sherman, C. The High Cost of Not Finding Information. IDC White Paper, IDC Group, 2000.
6. Haase, K. Interlingual BRICO. *IBM Systems Journal 39*, 3/4 (2000).
7. Haase, K., Guaraldi, B., and Tamés, D. SBIR Phase I Report: Non-Expert Conceptual Annotation. Report submitted to the National Science Foundation, Project Number 0232731, July 28, 2003. Available at www.beingmeta.com/pub/nonexpertannotation.pdf
8. McGuinness, D.L. and van Harmelen, F., eds. "OWL Web Ontology Language Overview." W3C Proposed Recommendation. Available at www.w3.org/TR/2003/PR-owl-features-20031215