

# Research Issues for Digital Libraries

David A. Forsyth  
Robert Wilensky  
UC Berkeley

We offer a number of research topics that we feel are worthy of attention. Of course, many other problems, notably, preservation, data integrity and data provenance, are far from being solved, but believe these are at least better recognized. We focus here on some problems that we feel have received insufficient attention thus far.

## Living Without Metadata

Attempts to produce large, useful digital libraries run into key problems associated with metadata. Typically, objects are either missing key metadata or the metadata is difficult to interpret—it has been supplied according to a different schema or the interpretation with respect to a schema is unclear. We note that the most successful tools, e.g., web indices like Google, work as well as they do precisely because they do not rely on any metadata whatsoever. Of course, this approach undeniably limits usefulness.

We see two important possible research directions arising.

The first is building an understanding of methods to translate metadata between schemas. Current strategies for manual translation founder in a morass of detail because they aim for accuracy. However, an inaccurate system may still be useful, particularly when there is no practical alternative, particularly when the volume of data is very large, such as when one is attempting to federate a set of collections. These observations suggest studying various statistical methods to build lexicon-like structures that can automatically align different types of metadata annotation.

The second is to explore various methods for obtaining metadata that are absent from an existing digital object. The current approach—in which skilled personnel attach metadata to objects—is infeasible for large collections. The first natural strategy to explore involves using existing, labeled collections to build classifiers that can attach tags to unlabelled collections. This approach is viable, because there are many large collections that are (partially) labeled, sufficient that it is practical to learn and evaluate classifiers for many kinds of metadata tag.

This class of strategies is likely to be successful only for metadata that can be inferred from the object itself. For example, it may be possible to determine the names of those present from a picture, but it is likely to be impossible to determine the time and data at which the picture was taken. One can attempt to deal with this difficulty by attempting to compel creators or editors of objects to attach metadata at the time of creation or editing, but in practice, this very likely will not happen. Instead, we propose to explore lightweight metadata collection. Objects should become “sticky”; when an object is created, edited, downloaded, read, etc., it should opportunistically collect various forms of information that might be helpful in building classifiers later. This is a generalization of the approach where current digital cameras insert a date and time into picture header files; in the not-too-distant future, we can expect to find GPS information there, too.

## Content Analysis for Collection and Meta-Collection Organization

The issue of content analysis of non-textual objects is well recognized, albeit still immature. However, while the issue of content analysis of images and the like for search has been the subject of substantial investigation, content analysis for organization of collections has received much less attention. We feel that this allocation of research resources is misguided, at least in the case of image collections. This is because, first, there is very little prospect of supplying search tools that can do what *users* of search want, and second, there are a number of other interesting activities, including, but not restricted to, browsing, organizing and data mining, that are more likely to be successful using the technology available in the foreseeable future. These activities have not been sufficiently well studied.

The reason for our caution concerning searching non-textual collection is that, according to what studies we have, users typically query images on semantics. For example, Enser's study (Enser 2000) reveals users requesting from the Hulton-Deutsch collection (now the Hulton-Getty collection) images of hangovers, physicists and the smoking of kippers. Identifying any of these concepts is well beyond the reach of current image analysis. Enser cites as a particularly compelling example a request to a stock photo library for "Pretty girl doing something active, sporty in a summery setting, beach - not wearing lycra, exercise clothes - more relaxed in tee-shirt. Feature is about deodorant so girl should look active – not sweaty but happy, healthy, carefree - nothing too posed or set up – nice and natural looking". Our own experience with users of image databases is, unfortunately, quite similar to this example.

Being able to perform such queries would be valuable, but difficult. However, there are studies that show being able to browse such collections is also of great value, and some reasons to believe that organization for browsing may be tractable. In addition to using metadata or textual data, one might attempt to cluster on a combination of image and text data, as Barnard et al. (2001) do. No option is clearly best at the moment, but some form of structure appears to be required, and it is at least plausible that such structures can be automatically inferred in the near future. In addition, there are many issues of mechanics, of how one must transmit, layout and display images efficiently, as well as choose what to display, that must be addressed.

There are two difficulties in organizing collections of visual material: First, it is difficult to obtain representations of what is depicted. Second, once one has this representation, organizing the material can be tricky. For example, one might be able to browse a museum collection that had been laid out in a manner that "made sense" to the user. In other applications, the organization would be relatively simple if the content could be obtained. For example, in military intelligence applications one might find imagery that showed large recent changes in where forces are concentrated. Similarly, one might examine a large reference collection of digitized images of Buddhist art for trends in the depiction of the human figure across space and time. These examples all depict uses of image collections that are not primarily about image search.

Data mining image collections is another example of a useful organization of collections. For example, it may be possible to build catalogs of objects, such as collections of images of a given individual, by automatic means. The number of applications of such derivative products is considerable.

## **Collection Characterization and Federation**

Content analysis for organization related to our previous point. It may facilitate the federation of collections along lines that haven't been anticipated by the collections' respective creators. Moreover, it is related to another topic that has received scant attention, the automatic characterization of collections. Prior work has focused on the suitability of a collection for the purpose of satisfying a textual query. This avenue of work has not been entirely satisfying, largely to the success with which union searches can be carried out. However, there is a large set of possible tasks for which characterizing collections might prove useful. For example, if one's goal is to find a good picture of George W. Bush, the image collection of the Fine Arts Museum of San Francisco is a poor place to look, whereas a collection of news articles is likely to be of greater utility. Being able to characterize diverse collections for the suitability of particular goals would also

be especially useful when the process to be carried out is computationally expensive, and hence applying a process to a union catalog is infeasible.

Being able to characterize diverse collections for the suitability of a variety of goals appears challenging, but has not been investigated to any great extent. Hence we recommend doing so.

In sum, we suggest focusing on the technology to support content analysis for organization of non-textual collections.

## Information Lifecycle Models

The shift to digitized media has thus far still largely left unchanged our traditional *models* of information use. Individual scholars still submit papers to conferences and journals, which still review them in the traditional manner, perhaps making the product available digitally. Our view is that most of the benefits of technology change will not be realized unless publishing, i.e., peer review, dissemination, and the use of primary and secondary data sources, is fundamentally changed from its current, centralized, linear, binary, expensive, “filter-then-disseminate” model, to a much less costly, powerful, fully distributed “disseminate-filter-collaborate” cycle, without sacrificing good organization, peer review and the like.

Progress in this direction has been slow, although we believe the realization of the importance of such a transition is clear. Progress could be accelerated by exploring further alternative models for content evaluation and filtering (to replace and perhaps majorize peer review), for distributed annotation (to support collaboration), and for distributed organization and manipulation of information substrates (to support individual and groups imposing new organizations on available material).

## References

Barnard, K. and Forsyth, D., 2001. Learning the Semantics of Words and Pictures, International Conference on Computer Vision, pp. 408-415.

Enser, P.G.B., 2000. Visual image retrieval: seeking the alliance of concept based and content based paradigms. *Journal of Information Science*, vol. 26, no. 4, pp 199-210.