



PERGAMON

Information Processing and Management 38 (2002) 363–379

www.elsevier.com/locate/infoproman

**INFORMATION
PROCESSING
&
MANAGEMENT**

Exemplary documents: a foundation for information retrieval design

David C. Blair ^{a,*}, Steven O. Kimbrough ^b

^a *Computer and Information Systems, Graduate School of Business, University of Michigan, Ann Arbor, MI 48109-1234, USA*

^b *The Wharton School, University of Pennsylvania, 1322 Steinberg Hall-Dietrich Hall, Philadelphia, PA 19104-6366, USA*

Received 8 April 2000; accepted 19 January 2001

Abstract

Documents are generally represented for retrieval by either extracting index terms from them or by creating and selecting from an external set of candidate terms. There are many procedures for doing this, but while work continues along these dimensions, there have been relatively few attempts to change this basic process. Of particular importance is the creation of indexing schemes for retrieval systems in non-library contexts. Here, the cost of developing an indexing scheme independent of the documents to be retrieved is often considered too high to implement. As a result, simple full-text retrieval or, to a lesser extent, automatic extractive or associative indexing methods are the predominant methods used in non-library contexts. This paper suggests an alternative document representation method based on what we call *exemplary documents*. Exemplary documents are those documents that describe or exhibit the intellectual structure of a particular field of interest. In so doing, they provide both an indexing vocabulary for that area and, more importantly, a narrative context in which the indexing terms have a clearer meaning. Further, it is much easier to develop an indexing scheme by using exemplary documents than it is to do so from scratch. © 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

Information, or document, retrieval is no longer just the library problem, and both the designers of management information systems and commercial retrieval system providers are coming to see this. Unlike libraries, organizational document retrieval systems are often rapidly

* Corresponding author.

E-mail address: dcblair@umich.edu (D.C. Blair).

built, do not follow traditional library methods of organization and representation, and frequently involve “mission-critical” information – information for which there may be a substantial penalty if it cannot be retrieved in a timely fashion. Certainly there are many non-trivial challenges to be met if we are to build effective information retrieval systems, but one design factor is critical: the level of effectiveness that an information retrieval system can attain is crucially dependent on how the documents in a collection are represented (Blair, 1990). No computerized retrieval system or retrieval algorithm, however complex, can adequately compensate for documents that are represented poorly or inappropriately. But the widespread need for non-library document retrieval systems which require high levels of retrieval effectiveness has meant that system designers are finding themselves dealing with document collections that may have no obvious representation scheme that could provide access to the intellectual content of those documents. It had always been hoped that some underlying universal representation scheme could be discovered that might be used to classify all information or knowledge. But attempts to find such universal classification schemes are as old as Platonic Realism, and the single, timeless classification of knowledge remains elusive. The thoughtful system designer, faced with the demands of building mission-critical information retrieval systems, is soon confronted with the fact that there is not one, ideal classification scheme for textual information, but myriad ones – uncountable, dynamic and ad hoc. This phenomenon was demonstrated most convincingly in Swanson’s important, but not widely circulated study (Swanson, 1966; discussed in Blair, 1990). Swanson found that when indexers were encouraged to assign as many index terms as possible to a document, they could come up with a virtually unlimited number of reasonable descriptions of the intellectual content of those documents – and in some cases, the total number of descriptions for a document was even greater than the length of the document being indexed. The reasons for this phenomenon are not entirely clear, but certainly the inherent creativity or productivity of natural language is, in part, responsible (Blair, 1990, pp. 170–171, Langendoen & Postal, 1984). It is also the case, that even when a reasonably good representation scheme is put together the process is so labor-intensive that only the most persistent systems designer can carry it out; and even a well-constructed representation scheme suffers from an important trade-off: the more complex and finely tuned the system of representation is, the less comprehensible that system may be for the average searcher. Clearly, the challenges of designing a representation scheme for a set of documents where no obvious representation scheme already exists and high levels of retrieval effectiveness are needed, are formidable.

2. The simple full-text retrieval solution

Faced with the significant challenges of formulating an intellectual structure for a collection of documents where no such structure has existed before, the designer of document retrieval systems often utilizes a simple full-text retrieval system. In such a system, the entire text of the collected documents is stored on line and the searchers retrieve documents simply by guessing the words and phrases that might occur in the documents that they want (but do not occur in the documents they do not want). While there are complex versions of this search model, what we are discussing here is the most simple version of this model where the search process is left entirely up to the searcher and is not supplemented by any automatic techniques. This is by far the most common

version of commercial full-text retrieval systems, and is a search option for a majority of computerized document retrieval systems (Delphi Consulting Group Inc., 1993). Of course, this kind of system relieves the system designer from the task of constructing an intellectual framework for the documents by accepting the actual text of the documents as a default representation scheme (non-content words, such as articles and prepositions, as well as duplicates of the words in text are generally not used to represent the documents). Such simple full-text systems, while easy to set up from an intellectual point of view, do not furnish very high levels of recall when providing access to a document collection of realistic size. In short, they trade off lower indexing and design effort for higher search effort. Blair and Maron (1985, 1990) demonstrated a *maximum* recall value of 20% for such a system in an operational setting. In spite of this, there are members of the information retrieval research community who still advocate the use of document retrieval systems with such low performance levels. In response to Blair and Maron (1985), Salton remarked, “. . .not only is this level of performance (i.e., 20% recall) typical of what is achievable in existing, operational retrieval environments, but that it actually represents a *high order* of retrieval effectiveness.” (Salton, 1986, p. 649. Author’s emphasis.) Yet, increasingly, recall levels of 20%, or less, are not seen as acceptable for operational systems, especially in critical applications outside the walls of university libraries and experimental test collections. In more and more current information retrieval applications, the consequences of low levels of access can be disastrous:

The effects of a failure of ‘document control’ can be dramatic. A major utility company was required to shut down four nuclear reactors because of lost repair instructions (at a loss of \$2m per day). The US department of defense estimates that half of all military accidents result from missing or inaccurate technical information. A major airline was fined \$10 k per take-off because of out-of-date maintenance information. A major drug company lost its entire R&D investment owing to inability to provide timely documentation (Fleischer, 1991)

If information retrieval researchers truly believe that a 20% recall level represents a “. . .*high order* of retrieval effectiveness. . .” then they are certainly out of touch with the demands of the market-place.

Efforts to design more effective operational information retrieval systems are certainly proceeding. Commercial systems such as Queria or Thunderstone’s “Metamorph”, among others, are developing and implementing automatic ways of improving searches based on the intellectual content of documents. But another way to improve access to documents is to investigate better ways of representing the intellectual structure and content of those documents. It is in this area that this paper aspires to contribute. Efforts in this direction are no less promising than automatic techniques, and improvements in both areas would be complementary. But progress in this latter area is uncertain – as we mentioned previously, efforts to identify an underlying knowledge structure to map all possible document representations onto have proved difficult and problematic. What is needed are some general principles that would help us to focus on which ways of representing documents are better than others. One suggestion is to relate the representation scheme for a document collection to the activities that use or generate the documents in the collection (Blair, 1990). This paper is an attempt to look at another (though related) basis for the representation scheme of a document collection – *exemplary documents*. (N.B. This section should not be construed as saying that the use of exemplary documents in IR systems will produce recall

values greater than 20%. Without empirical studies, it would be impossible to make such estimates, given our present meager ability to model IR system performance. It may be, too, that the major benefit of exemplary documents will not be an improvement in recall, but a speed-up or facilitation of the indexing and design process.)

3. Exemplary documents

In many document collections some documents are more representative of the intellectual structure of that collection than others are. That is, the intellectual structure of the document is similar (in whole or in part) to the intellectual structure of the discipline or field of study of which the document is a part. These are what we call “exemplary documents.” Exemplary documents provide not only a view of the issues discussed in many of the other documents in the collection, but they also provide a framework which often links those issues together into a more meaningful whole. Further, they also provide the vocabulary in which those issues are discussed. In essence, exemplary documents provide an “intellectual road map” to the semantic content of a given set of documents. Many document collections, then, can be divided into two kinds of documents: the documents to which access is desired, and a much smaller set of exemplary documents which may be used to build an intellectual structure on which some or all of that access can be based. If these exemplary documents do exist in sufficient numbers, and it is our contention that they do, it may be possible to exploit them as an organizing principle for a collection of documents.

4. Exemplary documents: examples

So far, we have said that exemplary documents provide a conceptual “view” of the issues and topics that a set of documents is concerned with. While helpful, such a rough definition must be made more precise if we are to use it to identify actual exemplary documents. Perhaps the best way to do that would be to begin by discussing several examples of exemplary documents.

4.1. *Research literature*

The most obvious kind of exemplary document is the “survey article” that appears in the literature of a particular research area. This kind of article explicitly discusses the major issues or topics in a field of study, and often provides specific references from these issues to significant publications, authors and institutions that are important in this area. But survey articles do not just provide a “topography” of the intellectual landscape of the field, they also discuss the issues of the field *using the professional vernacular of that field*; that is, they are demonstrations of *how* the field refers to, and talks about, these issues and topics. This is not to say that the language of that field is fixed. There are few active fields in which that would be the case. But insofar as there might be preferred or popular ways of discussing the topics and issues of a field, those linguistic preferences would likely be exemplified in the survey article. In this way, the survey article can act as a kind of conservative linguistic process that demonstrates how, out of the many ways a topic *could be* discussed, it usually *is* discussed in that field of inquiry. Survey articles, therefore, provide

several essential elements for representing the intellectual content of documents in a particular field:

- A description of the major issues or topics in the field, or part of the field (often including those issues that are in direct conflict with each other).
- A demonstration of the language used to refer to the issues and topics with which the field is concerned.
- References to some of the significant authors/researchers in the field.
- A description of major techniques, methodologies, ways of working, etc. (where applicable).

N.B. Obviously the citations given in the survey article could also be followed up to find the names of institutions and journals that also figure prominently in that field.

Survey articles are not the only exemplary documents in research literature, but they are usually the easiest kind of exemplary document to identify. Those exemplary documents which are not survey articles share with the survey article the ability to describe the major issues and topics of a field. Such documents might be editorials or opinion papers written by prominent members of a field's research community in which they might draw on their experience in the field to reflect on its conceptual or intellectual structure. This kind of exemplary document tends to be more abstract than a survey article. Ostensibly, it contains the views of a single individual, while the survey article is frequently a synthesis of the work of many individuals. But since the author of an editorial or opinion paper is usually a prominent member of the research community, his or her views, while personal, still often represent a consensus of sorts.

There is also a hybrid exemplary document that can have characteristics of both the survey article and the reflective opinion paper. This is the lead article (usually written by a guest editor) in a special issue of a research journal devoted to a specific topic. These lead articles vary widely in content, and in many cases they may not qualify as exemplary documents at all. But in some instances, such articles may both survey a particular field, or sub-field, as well as propose an abstract intellectual framework in which to fit the topics and issues of that sub-field.

The idea of an exemplary document recalls Kuhn's notion of an "exemplar" (Kuhn, 1970, p. 186ff), so it may be useful to take a moment to describe the relation, if any, between exemplars and exemplary documents. For Kuhn, the principal activity of normal science is puzzle-solving. Classic puzzles that a field has already solved serve as what Kuhn calls exemplars. These exemplars are the most important component of the disciplinary matrix that constitutes the paradigm of a scientific field. They serve as good examples of the kind of puzzles members of that field should work on, and thereby also become models for the kind of new puzzles members of that field should devote themselves to. Further, they also have a pedagogical purpose: the student becomes a full-fledged member of that field by working through and solving these exemplary puzzles. If he does enough of them, he will attain not only the accepted puzzle-solving skills and methods of that field, but will also share a similar point of view towards the subject matter and activities of the field (what Kuhn called a "group-licensed way of seeing" (op.cit. p. 189)). This common point of view will allow the neophyte to select the right new puzzles to work on in his/her field. Anyone who has taken an introductory physics laboratory has had first hand experience with this process. In fact, Kuhn insists, it is easy to identify the exemplars of any stable field – just examine a set of textbooks of the field and see which puzzles they have in common. Those that they have in common are the exemplars of the field.

Although exemplary documents are not puzzles, they *do* share some characteristics of Kuhn's exemplars. Specifically, exemplary documents can also have a pedagogical role in most disciplines. Any document that presents the intellectual structure of a field in a clear and unambiguous way, as many survey articles do, would be an important source from which students could learn the intellectual structure of that field. It would help them to attain the group-licensed way of seeing that Kuhn's exemplars do.

Just as the textbooks of a field are important sources for Kuhn's exemplars, they might also have a role to play similar to exemplary documents. Specifically, the organization of a textbook is likely to be similar to the intellectual organization of the field it represents. The organization of the textbook, then, exemplifies the intellectual organization of the field. Further, the vocabulary used in the textbook should be a good model for the indexing vocabulary needed to represent documents in that field. In short, textbooks may provide the same kind of intellectual organization of a field that exemplary documents do. There are a few differences, though. Textbooks are invariably longer than exemplary documents, so they may serve as less concise access mechanisms than exemplary documents do (although, individual chapters of the textbook might be taken separately as exemplary documents that describe the intellectual structure of a part of the discipline). Also, they may represent a more conservative or established view of the field – they describe the field as it is taught, and would probably not represent newer or more controversial areas of the field. Survey articles, on the other hand, would be more likely to describe newer, less established or less widely accepted work in the field. Further, textbooks generally cover more introductory levels of the field than survey articles do since, of course, they are intended to be used by students. Survey articles, on the other hand, often deal with more advanced concepts and are often intended primarily for established members of the field (or, “aspiring” members of the field, such as graduate students). Finally, because of the textbook's need to represent an accepted view of the field it deals with, it may represent an older, less current view of that field. Survey articles can be published more quickly than textbooks, so they can establish an immediacy unattainable in the longer, more consensus-oriented format of the textbook. These caveats aside, textbooks are certainly candidates for exemplary status.

There is another kind of article in the research literature that may bear some relationship to exemplary documents: the seminal paper. While it is possible that an exemplary document may also be a seminal paper in a given field, it is not the case that every seminal article is an exemplary document. An example should make this clearer. As a recent survey makes clear (Krishnan, 1993), the article by Will (1975) is considered seminal in the field of Model Management since it was instrumental in motivating researchers to begin work in this area. But the article itself, while *seminal*, is not *exemplary* since it offers no enduring guide to the major issues, topics or methods of the field as it developed. In the field of information retrieval, Vannevar Bush's paper “As we may think” (Bush, 1990), is frequently mentioned as a seminal paper in the field and has even found a recent relevancy to hypertext-based information retrieval systems. While anticipating hypertext-type document arrangements, Bush's article, of course, does not anticipate how the intellectual structure of hypertext-based information retrieval has recently evolved. Like the Will article (*supra*), it offers no enduring guide to the issues, topics or methods of the information retrieval field, or even the narrower field of hypertext-based retrieval systems, and is thus *not* an exemplary document.

4.2. Large-scale litigation

Corporate and government litigation are just two areas where there is a dramatically increasing number of large-scale lawsuits – lawsuits in which attorneys may need to have access to thousands, or sometimes millions of documents. Further, the successful conduct of the lawsuit is critically dependent on effective document access. The intellectual structure of the information germane to a particular lawsuit is best exemplified in a document called the “complaint”. The complaint identifies, very precisely, the issues or events being litigated, the individuals involved in the case, and, if relevant, the institutions involved also. In addition, the complaint will often identify the *chronological sequence of events* that tie all the individuals, issues and institutions together, and will often assert *causal links* between certain key events.

Another kind of exemplary document in litigation consists of the “depositions” given during the pre-trial phase of the lawsuit. These depositions are often explicit narratives given by the major participants in the lawsuits in which they answer questions and discuss the actionable events in question. While the complaint gives the legal (i.e., lawyers’) view of the suit, the depositions give the individual participants’ views of the events in question. Depositions and complaints often have a narrative structure – they tell a story. The narrative is not only a good way to organize the issues of a lawsuit, but may, in fact, be a fundamental way in which we structure information. To some researchers in cognitive science or related fields (e.g., Bruner, 1990; Turner, 1996), the narrative is one of the central organizing principles of our experience and knowledge. Hence, the deposition may directly model the way we understand the issues of a lawsuit making its use as an organizing principle in an information system even more compelling.

4.3. Public policy

The formulation and analysis of public policy comprises an interesting body of textual information. The literature itself often has a structure that organizes around explicit arguments or debates. Frequently, in this kind of literature, documents will occur which describe the issues and structure of particular debates. For example, we have been building a prototype information retrieval system based on exemplary documents in the area of “technology policy”. The debate here revolves around the United States’ declining world competitive strength in science and technology. As we gathered documents on technology policy it soon became clear that one particular document exemplified the major issues in this debate, and asserted several causal connections that were the basis for much of the debate in this area. This document was “The competitive strength of US industrial science and technology: strategic issues” (National Science Board., 1992). It satisfied our loose criteria for an exemplary document because it gave a clear presentation of the major issues in the debate over technology policy. It also described the major causal assertions of the debate and cited supporting literature for its view of the debate.

As we looked further into the literature of this debate, we found a second exemplary document – “Technology: The engine of economic growth. A national technology policy for America”, authored by the 1992 Clinton/Gore Presidential campaign committee (Clinton/Gore, 1992). Like the National science Board Report, the Clinton–Gore document provided a broad view of the debate and made several causal assertions of its own. But there were also differences between the NSB report and the Clinton–Gore document. Both represented somewhat different views of

the same debate, so there was not a one-to-one correspondence between the issues and assertions of the two documents. Is this a problem? In other words, is there a difficulty if two exemplary documents make different – even opposing – assertions? Further, is there a problem if the assertions of an exemplary document prove incorrect? The answer to each of these questions is “not necessarily”. What is essential about exemplary documents is not whether their assertions are correct or incorrect, but whether they can provide an intellectual road map to a body of literature. So, for example, an article that passionately advocates a “creationist” view of evolution might be useful for organizing not only the creationist literature, but also the more “scientific” literature of evolutionary theory. This is because while the creationist exemplary document may argue for the creationist point of view, it may also mention and attack the scientific paradigm for evolution theory, thereby providing a schema for both the scientific and creationist views of the field.

4.4. Project management

Long term projects are examples of information intensive activities that can generate a large number of documents, drawings, engineering specifications, etc., all of which are relevant to the conduct of the project and many of which are essential. Important organizing principles of such document collections would be the chronology of the project and the engineering breakdown of the components or systems which comprise the project. Exemplary documents for project document control would most likely consist of project initiation or enabling documents (such as the original contract specifications), major planning and review documents, documents prepared for major decision points in the course of the project, or evaluations of project checkpoints.

We applied these principles in the development of a document-oriented project management system for the Coast Guard as part of the KSS project (see Kimbrough, Clark, Michael, & Hemant, 1990, for early work in this area). Here we identified an important document, OMB Circular A-109, which defines the major systems acquisitions process for the US Government. This document describes the acquisition process in detail and specifies key subsequent documents that must be created and approved as part of the process. In essence, it prescribes how the activity must progress and which specific subsequent documents must be created to enable the process to continue. Some of these subsequent documents may be exemplary documents themselves. In this case, the primary exemplary document is not hard to identify. In fact, it is a *sine qua non* of the project management process itself.

The above examples illustrate the kinds of exemplary documents one might find in several different fields. It would not be hard to give similar examples for other information-intensive activities such as corporate strategy, archival management or the management of museum collections.

5. Exemplary documents: characteristics

Looking at the above examples of exemplary documents, we can start to see some of the characteristics that many of them have:

- (i) They always provide a synoptic view or survey of at least some of the major topics and issues of a field, or part of a field.

- (ii) They often discuss these topics and issues with a vocabulary that is a restricted subset of all the possible ways these issues could be discussed. An exemplary document is a demonstration of the “dialect” of natural language that is used to talk about issues in that field.
- (iii) They often indicate the structure or framework of the field by showing, explicitly or implicitly, the relationship between some or all of the issues or topics they identify (e.g., that certain issues are related *causally*; or, how specific events, such as a series of experiments, are related chronologically).
- (iv) They often provide explicit bibliographic links to individuals, authors, institutions, documents, etc., that figure prominently in a given literature.
- (v) The truthfulness of the assertions made by an exemplary document do not necessarily affect its usefulness as a tool for organizing a body of literature.
- (vi) In some, perhaps rare, instances an exemplary document can even *drive* a field’s literature; that is, an exemplary document may be so persuasive about identifying the issues and structure of a field that subsequent papers that are written in the area may respond specifically to the issues and their relation spelled out in that exemplary document. A clear historical example of this is the mathematician Hilbert’s early 20th century paper that outlined what he believed were 23 major puzzles of mathematics for that time (Hilbert, 1901). His paper became a blueprint for a great deal of the subsequent work in theoretical mathematics for decades afterwards. In our own discussion, exemplary documents for the literature of public policy debate often drive the subsequent literature. In cases like this there may be some correlation between citation rates and exemplary documents, though high citation rates alone would not, we think, identify exemplary documents.

6. Exemplary documents: internal structure

The role of the internal structure of an exemplary document is important enough to discuss in some detail. Such a structure may be explicitly presented in the document, or it may be implicit. There is also a number of ways that the issues identified by an exemplary document can be related: in the complaint of a lawsuit there may often be an explicit assertion of causal relations between specific actionable issues; there may also be a chronological sequence of events that is significant, too. In a paper which surveys the research of a field, the chronological sequence of that research may be a very important part of the structure of that field since it may be used to establish precedent and dependence. The research methods of a given field may also provide an implicit structure for the issues in that field, and these structures may vary from field to field insofar as their respective research methods vary.

There is also another kind of internal structure inherent in exemplary documents that should not be overlooked. This structure is the narrative structure of the document text itself. It almost always contains sections and headings (some survey articles, such as those in *computing surveys*, even have tables of contents for *each* article in an issue). These section headings and divisions offer not only an explicit structuring of the article, they also may offer, implicitly, a structuring of the issues discussed in the article, and this implicit structuring of the article may be used to organize other documents in the field.

On a more subtle level, there is also a structure to the discourse of the document. This discourse may be classifiable, for example, according to categories of Speech Acts (Austin, 1962; Searle,

1969) or “implicatures” (Grice, 1989), and these categories, in turn, imply a structure that may be useful for arranging the issues of a field (see Blair, 1992). The representation of the intellectual content of documents is, fundamentally an issue of language and meaning.

7. Exemplary documents and activity-based document representation

In the first part of this paper, we mentioned that the notion of exemplary documents offered an alternative to representing documents based on the activities that “use or generate the documents in the collection” (Blair, 1990). On closer examination, however, it becomes clear that exemplary documents are more closely related to these activities than may have first been apparent. In fact, it may be the case that exemplary documents actually arise out of the activities that use or generate sets of documents, that is, they may actually play a role in the conduct of the activity. As Xerox Vice-President Steven Kiser put it, “We used to think of a document as the end result of a process; the reality is that the document *is* the process.” (van Kirk, 1992) This is clearly the situation in litigation support, where the complaint – which we identified as an exemplary document – is not just an intellectual road map of the lawsuit, but specifies the goals of the lawsuit, or how the case will be litigated. Likewise, in the area of public policy debate, an exemplary document not only can show how the literature can be organized, but it may also describe and perhaps even influence the course of the debate itself. Even a survey of an academic field can be a natural consequence of the activity of that field. That is, academic fields are not just collections of topical information or research. They are ongoing, active research disciplines that are engaged in furthering research in a particular area. An important component of the activity of research in any area is for researchers to be informed about what has been done and what is being done in their area of research. Survey articles specifically address this need for researchers to view the work in their field, so they are, in effect, actually participants in the ongoing activity of research. It is no surprise, then, that they can be good instruments for organizing the textual information of a field.

It is clear that if exemplary documents arise out of activities, then they are likely to assume many of the characteristics of the activities that they come from. Activities, fundamentally, have goals or purposeful concerns. They reflect an intent to accomplish something more or less specific. These goals or concerns provide a point of view or context for the activity. That is, they provide a framework in which a participant can judge what is pertinent to an activity and what is not. It is a fact of natural language that it, too, requires a point of view or context, otherwise, it is difficult to interpret unambiguously (for the simple reason that the same words can be used in so many different activities – that is, they can be used in so many different ways). If natural language is needed to represent documents, as we have stressed in this paper, then it stands to reason that for a language of document representation to be unambiguous it must have a clear point of view or context. This point of view or context is usually what is missing from traditional indexing schemes, but it is what exemplary documents can provide. This also helps to explain why simple full-text retrieval systems, which we discussed in the beginning of the article, have such consistently low levels of recall. By using all the content-bearing words of the documents in a collection to represent those documents for retrieval, the inquirer is presented with an enormously wide variety of access terms and combinations (see Blair & Maron (1985) for examples of this phenomenon). But this is only half of the problem: not only is there

an enormously large number of terms that could provide access to documents in the collection, but there is no broad context in which to interpret the meanings of those words – their purpose or use may be unclear. Exemplary documents address both of these problems: they provide a selected vocabulary to represent a set of documents, rather than simply using *all* the words in the documents to represent them; exemplary documents also provide a narrative context for the selected vocabulary that helps the inquirer to understand what certain specific words mean – that is, how they might be used to represent non-exemplary documents that he might be interested in.

8. Exemplary documents: practical advantages

The use of exemplary documents as an intellectual organizing principle has several specific practical advantages for system design:

8.1. Exemplary documents embody both “top–down” and “bottom–up” organizing principles

Information retrieval systems are often based on an implicit inconsistency, namely, that while a given system’s intellectual structure (indexing or representation scheme) is usually designed from the bottom–up, the individual searchers often begin their searches from the top–down. By bottom–up design we mean that the system designers construct the intellectual access structure of the system by an examination of the individual documents included in the system. While this structure can be designed either by individuals (indexers) or it can be done automatically, it still begins in the same place – by evaluating individual documents in the same, often random, order in which they are received. The inquirer, on the other hand, often takes what we could call a top–down approach. He/she does not usually begin with a particular document in mind, but with some more-or-less abstract description of the intellectual content he/she is searching for. How well that abstract description of what the inquirer wants matches with the specific indexed descriptions of documents that he would find useful, is problematic. Since the top–down and bottom–up approaches are based on different principles or different points of view, then it is not likely that they will match up very well. The key then, to bridging the inconsistency between top–down and bottom–up approaches is to develop abstract principles of representation that are based on individual documents. This brings together the inquirer’s need for an indexing scheme that can represent high level, abstract concepts with the system designer’s need to represent individual documents.

Clearly, you cannot develop a useful schema for representing the abstract intellectual content of a set of documents by basing it on randomly selected documents. Most documents are too narrowly focused to be representative of the implicit intellectual structure of the document collection to which they belong. This is why it is important to identify exemplary documents. Exemplary documents provide an abstract view of the intellectual structure of a set of documents that is grounded in actual documents in that set. Specifically, such documents provide the system designer with not only the abstract intellectual content of a set of documents, but also the relationship or structure of those intellectual abstractions as well as the specific vocabulary in which they are discussed.

8.2. Exemplary documents provide a common foundation for both indexing and searching

As mentioned in Section 4.1 (*supra*), inquirers and system designers/indexers may often begin their respective tasks from different starting points. A clear advantage of exemplary documents is that they provide a common foundation for both the document representation *and* the search formulation process. The indexer, or system designer, will look to the exemplary documents for the indexing vocabulary and structure of the documents he/she wants to represent, while the inquirer will also go to them for the concepts being represented in the collection and the specific vocabulary used to discuss these concepts.

Although there still may be some divergence between the indexer's and inquirer's conception of how the documents might be represented, there appears to us that there will be less divergence if they both begin their tasks in the same place (the importance of this divergence was discussed in Blair (1986)).

8.3. Exemplary documents enable inquirers to better understand the intellectual structure of a document collection

One of the trade-offs we mentioned earlier in this paper was that the more complex the intellectual structure of an indexing or representation scheme is, the harder it may be for the inquirer to understand. Yet, if we want to provide high levels of retrieval effectiveness for some information retrieval systems, then we will need fairly complex and detailed intellectual structures for representing their documents. But such complex representation schemes will require that inquirers must be able to learn them. Unfortunately, most indexed collections of documents are not constructed to facilitate the "learning" of the intellectual structure (Blair, 1990). Often the only way an inquirer can learn how documents are represented on a system is by trial and error, by submitting queries and retrieving documents (Swanson, 1977). Such a piecemeal process is a very unsystematic and probably flawed way to learn, in any general way, how the documents in the collection have been represented. In the first place, the inquirer may never be able to see enough documents to form any general opinion about how the documents are represented (Blair, 1980); and, in the second place, if the inquirer can only see individual documents it may be hard to infer what the broad intellectual relationships are that may exist among groups of documents. Exemplary documents can help to defeat, or at least mitigate, this problem by providing access to not only the intellectual concepts of a field or sub-field, but also to the structure of those concepts and the specific vocabulary used to discuss them.

If an inquirer wants to understand what the intellectual structure of a document collection is, he should begin his search by reading the most relevant exemplary document that exists on the system. The search process, then, is two-tiered: the inquirer searches first through only the exemplary documents available on the system, he then constructs further search queries based on the parts of the exemplary documents that he has found relevant to his search. (In a hyper text-like document management system such as the World Wide Web, links to other, non-exemplary, documents could be embedded in the exemplary documents. In this way, the inquirer would not have to formulate his/her own search queries at every stage in the search.) An exemplary document gives the inquirer a view of the topics or issues germane to her search, a framework that relates these topics or issues, and a clear vocabulary which discusses these topics and can be used

to formulate search requests. (The search procedure based on exemplary documents is similar to the strategy of “seed searching” proposed by Blair (1986).)

8.4. Exemplary documents help to limit the “productivity” of natural language

One of the major characteristics of natural language is its productivity. Language is productive in the sense that a relatively small vocabulary and a few rules of syntax and semantics can combine to produce an uncountably large number of acceptable sentences or descriptions (this is the notion of a “generative grammar”). The productivity of language has proved to be both a virtue and an obstacle for information retrieval systems (as it has for any other computerized system that has a large semantic component): productivity is a virtue because it is relatively simple to find a reasonable way to represent any document; but this productivity is an obstacle, too, because language is so creative that the number of “reasonable” representations for a document and its content is virtually unlimited (Blair, 1990; Swanson, 1966). The same problem holds for the inquirer, too. The inquirer can easily formulate more semantically similar search requests than he/she can possibly have time to actually use, and, unless the inquirer uses them all he/she will have no idea which ones are the most successful.

Since natural language is also productive, as we said, why don’t we have the same problem in our everyday discourse as we have in representing documents – that is, while there are lots of semantically equivalent ways to say something, we usually do not have a great deal of trouble expressing ourselves clearly and understanding others. The obvious answer to this is that in everyday speech we have a lot of help from the pragmatic context of our utterances. This is how we can say something quite precise with a sentence that, taken “out of context”, would be impossibly ambiguous (see Blair, 1990 for a discussion of this ambiguity problem). For example, suppose that when I arrive at my office I ask a colleague whether so-and-so is at work today. She answers, “I saw a yellow Volkswagen in the parking lot this morning”. Strictly speaking, this is not an answer to my question, yet with some simple assumptions about the context in which the statement is uttered and the intentions of my colleague, I can “make sense” of what might be, taken literally, an impossibly ambiguous statement. Such implicatures are an important part of the pragmatic context of speech (Grice, 1989; discussed in Blair, 1992).

Exemplary documents limit the productivity of language by providing a context of linguistic usage for potential indexing terms. That is, of all the many ways a given word might be used to represent the intellectual content of documents, the exemplary document provides an example of what precise concepts an index term might be used to represent in that collection of documents.

9. Exemplary documents: theoretical justification

While we feel that the practical rationale for the use of exemplary documents in information retrieval is compelling, we also claim that there are sound theoretical reasons that underscore the importance of this kind of document. The theoretical justification for the utility of exemplary

documents is based on the fundamentally linguistic nature of information retrieval. The justification goes like this:

The effectiveness of an information retrieval system is crucially dependent on how the documents in the collection are represented (for example: assigned keywords; full text; automatically generated index terms; etc.). These document representations are fundamentally linguistic in nature, and are drawn from a subset of natural language discourse. But the meanings of keywords or terms extracted from the text of documents are often not exactly equivalent to the meanings of those same words when used in everyday discourse. For example, the document content that is indexed by the term “computers” is similar to, but may (or may not) be equivalent to “what we mean” when we use the word computers in everyday speech or writing. The word computers, when used to represent the intellectual content of a document, is open to a wide variety of interpretations (that is, computers can be used to represent a wide variety of different, though often related, document topics) (Blair, 1986). Of course, it is also true that the word computers, when used in everyday language and considered by itself, can be quite ambiguous, too. But everyday language has one major advantage over document representation: the words that we use in everyday language have a pragmatic context and accompanying implicatures that can be used to disambiguate individual word uses (Blair, 1992; Grice, 1989). Such a pragmatic context has no equivalent in computerized information retrieval. But for an inquirer to understand what an indexing term means he/she must come to understand the “language of document representation” – and he/she must do so in the absence of any easily accessible contextual information about how the words of that “language” are used.

Putting aside, for the moment, the problem of the lack of contextual reference for the language of document representation, let us look briefly at how a mature individual learns natural language (here we will assume that we are dealing with an intelligent adult who is either learning a new language entirely, or learning new words in an already familiar language). It is no great intellectual leap to see that it would be best for the inquirer to learn the language of document representation in the same way that he would learn new words in any natural language. How is this done? While there are many theories of language acquisition, the theory proposed here is based on the later philosophy of language of Wittgenstein. One of the present authors has argued that Wittgenstein’s later philosophy of language is of significant use in understanding the unique problems of document representation (Blair, 1990).

10. Language acquisition and *übersichtliche Darstellungen*

Briefly, Wittgenstein’s theory of language acquisition runs like this: We do not acquire language by definitions and explanations alone, but by having the *use* of the expressions in question demonstrated to us – ideally, in the same circumstances or activities in which they are ordinarily used. Definitions and explanations are not central to language acquisition and can be entirely replaced by examples and demonstrations of word usage. Definitions and explanations *do* have a use in language acquisition, but it is not a central use. They serve to clarify or modify a particular

linguistic expression whose general type of usage is already relatively clear. The examples of language usage which are used in training must be “perspicuous” – they must be clear and compelling. In short, Wittgenstein’s theory of language acquisition is crucially dependent on the identification of “perspicuous representations” of word usage.¹

It stands to reason, that if Wittgenstein’s philosophy of language acquisition is correct, or at least plausible, then to enable inquirers to learn how document representations are used we must provide them with examples of not just *any* index term usage, but with perspicuous term usage. These representations of usage must be exemplary, clear, and embedded in a rich enough linguistic context that their usage is relatively unambiguous. Within the linguistic framework of information retrieval, the usage and context of index terms is not at all clear, in fact, as we mentioned before, the pragmatic context of index term usage is virtually non-existent. (Recalling our previous example, it would be like trying to discover the meaning of my colleague’s statement “I saw a yellow Volkswagen in the parking lot this morning” by just tabulating the meanings of the individual words in the sentence without regard to the circumstances in which it was uttered.) Further, if one were to look at actual index term assignments to documents it would be difficult to distinguish the term assignments that are exemplary from those that are not. This is why exemplary documents are so important. They not only provide a view of a topic area (in whole or in part), but they also are an active demonstration of the language that is used to talk about these areas – in short, exemplary documents provide the perspicuous representations of index term usage that would permit the inquirer to understand what these terms might mean if they were used to represent the intellectual content of other documents. The text of exemplary documents also provides the pragmatic context of index term usage that is missing in contemporary information retrieval systems.

11. Exemplary documents: implementation

As mentioned earlier in this paper, we are currently using the notion of exemplary documents as a basis for the design of information retrieval systems. It might be useful to sketch some of the

¹¹ What has been translated as perspicuous representations is what Wittgenstein originally called “übersichtliche Darstellungen”. This is not an unequivocal translation for there is no simple English equivalent (see Baker and Hacker (1985) for a discussion of the translation problem here), but the essential concepts which constitute übersichtliche Darstellungen are that they are not just perspicuous examples of word usage, but representations that give us a “broad view” of how the word in question is used. That is, any representation of a word usage is not necessarily a perspicuous one. The representation must be, to use another word, exemplary. In addition, the German word “Darstellung” means not just “representation”, but also “performance”. The notion of a performance is crucial to Wittgenstein’s philosophy of language acquisition – since he emphasized the importance of how words are used in the conduct of our daily activities and practices, rather than how they are defined. In short, what the individual learning new words or a new language needs is a good supply of exemplary uses or performances of those words within the context of natural language discourse.

A main source of our failure to understand is that we do not *command a clear view* of the use of our words. Our grammar is lacking in this sort of perspicuity [Übersichtlichkeit]. A perspicuous representation [Darstellung] produces just that understanding which consists in ‘seeing connexions’. (N.B. “connexions” is the translator’s rendering of “Zusammenhänge” which can also be translated as “context”) (Wittgenstein, 1953, para 122).

directions we are taking. Since exemplary documents actually embody the intellectual structure of the literature they are found in, it may be useful to implement at least part of this organization through some kind of hypertext system. In that way, links to non-exemplary documents could be embedded right in the text of the exemplary documents, or could be generated automatically based on the similarity between the text of the exemplary document and the representations of the non-exemplary documents. The advantage of this is clear: when inquirers read a section of the exemplary document that they are interested in, they could activate such a link and move directly to those documents most closely related to the section they were reading. This would save inquirers from having to formulate a search query for each section of an exemplary document that they are interested in. If a number of documents were identified by the same link, then they would, of course, need to be ranked by the closeness of fit between them and the section of the exemplary document to which they are related. There are a number of existing procedures that could be employed to do this.

It is clear that simple hypertext links are just one of the ways in which exemplary documents can be used to organize a collection of documents. Many of the nodes on the WWW are already organized this way. An improvement might be to write or find exemplary documents that could be used as the first level of access at the home page of a web site.

12. A final note on coverage

One important aspect of exemplary documents that may trouble some readers is that there may be cases where exemplary documents exist, but do not provide complete coverage of all the intellectual issues of a field or sub-field. It may also be the case that they might not provide the granularity of description desired for that field. Until a concerted effort is made to identify exemplary documents in a field, it is uncertain how complete the intellectual road map they provide will be. It is hoped that the coverage provided by exemplary documents will be adequate, but short of that, the coverage that they do provide, and the context of representation that they bring with them, will be useful adjuncts to the traditional manual or automatic methods of document representations that we currently employ.

13. Conclusion

This article has addressed the specific problem of constructing document representation schemes in situations where traditional representation schemes do not apply well, and where highly effective access to the documents is required. We have proposed that certain documents, which we call exemplary, may exhibit useful ways for representing a class of documents. These exemplary documents provide several advantageous things: an intellectual road map to the topics and issues of a field; a specific vocabulary that can be used to refer to those topics and issues; and an interpretive context in which the language used to represent a set of documents can be disambiguated. We have also argued that apart from several distinct practical advantages for the use of exemplary documents, there are also several theoretical reasons that also support the use of exemplary documents in information retrieval situations.

Acknowledgements

The authors wish to thank Pat Wilson and M.E. Maron (University of California, Berkeley) for their comments on an earlier version of this paper.

References

- Austin, J. (1962). *How to do things with words*. Oxford: Oxford University Press.
- Baker, G. P. & Hacker, P. M. S. (1985). Übersicht. In *Wittgenstein: Meaning and understanding. Essays on the philosophical investigations. Vol. 1*. The University of Chicago Press; p. 295–310.
- Blair, D. C. (1992). Information retrieval and the philosophy of language. *Computer Journal*, 35(3), 200–207.
- Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.
- Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information and Processing Management*, 22(2), 229–241.
- Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal for the American Society for Information Science*, 31(4), 271–277.
- Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, 28(3), 289–299.
- Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: further analysis and clarification. *Information Processing and Management*, 26(3), 437–447.
- Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.
- Bush, V. (1990). As we may think. *The Atlantic Monthly*, 176, 101–108.
- Delphi Consulting Group Inc. (1993). *Text retrieval systems: A market and technology assessment*. Boston, MA.
- Clinton/Gore '92 Committee. (1992). *Technology: The engine of economic growth. A national technology policy for America*. National Campaign Headquarters, Little Rock, Arkansas. Final Version, September 21.
- Fleischer, R. (1991). Total document control: a text-retrieval perspective. In G. Peter (Ed.), *Text retrieval: information first, Proceedings of the Institute of Information Scientists 1990 Text Retrieval Conference*, London, October 1990. London: Taylor Graham.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hilbert, D. (1901). Mathematische probleme. *Archiv für Mathematik und Physik*. 3 Reihe Bd. 1, S. 44–63; S. 213–237.
- Kimbrough, S. O., Clark, P., Michael, B., & Hemant, B. (1990). The coast guard's KSS project. *Interfaces*, 20(6), 5–16.
- van Kirk, D. (1992). Document management: Destined to become a smash hit. *Infoworld*, November 2, p. 52.
- Krishnan, R. (1993). Model management: survey, future directions and a bibliography. *ORSA CSTS Newsletter*, vol. 14(1), Spring.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
- Langendoen, D., & Postal, T. P. (1984). *The vastness of natural languages*. Oxford, UK: Basil Blackwell.
- National Science Board. (1992). *The competitive strength of US industrial science and technology: Strategic issues*. Committee on Industrial Support for R&D, August.
- Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7), 648–656.
- Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, 47(2), 128–148.
- Swanson, D. R. (1966). *Studies of indexing depth and retrieval effectiveness*. Unpublished report, National Science Foundation Grant GN 380, February.
- Turner, M. (1996). *The literary mind*. Oxford: Oxford University Press.
- Will, H. J. (1975). Model management systems. In E. Grochia & N. Szyperski (Eds.), *Information Systems and Organization Structure*, (pp. 468–482). Berlin: Walter de Gruyter.
- Wittgenstein, L. (1953). *Philosophical investigations* (G.E.M. Anscombe, Trans.). New York: The MacMillan Company.