

Annotating document content: a knowledge-management perspective

Fabio Ciravegna and Daniela Petrelli

Ontology-based annotation provides a method of making document content amenable to automatic searching and the construction of an index even when professional support is not a possibility. Research in this area is currently investigating applications both to the World Wide Web and to specific domains, e.g. in aerospace. Some of the tools developed to enable ontology-based annotation could help professional indexers increase their own efficiency.

Introduction

The sharp fall in computer memory and storage costs and the increase in their speed and connectivity are dramatically changing the way information is stored. In the past, medium-size, mainly textual, centralized archives were the only resource for knowledge management. Nowadays companies handle multimedia information in distributed archives in an order of magnitude that was unimaginable even a few years ago. Large organizations' intranets are now the size of mini Webs, connecting thousands of computers and embracing tens of millions of documents. This trend will continue: it is expected that intranets will soon include *hundreds* of millions of documents – a size comparable to the Internet at the end of the 1990s. Moreover the increased use of the WWW as a source of information has made the boundary between intranet (i.e. the internal knowledge) and Internet (the publicly available material) very thin. This dramatically increases the whole concept of 'information space'.

As an example consider a jet engine: every flight produces about 1 gigabyte of vibration data:¹ if irregularities are found, part of the data is stored; when the engine is serviced and problems are found, pictures are taken and reports are written; financial information is also produced (e.g. recording of service costs on spreadsheets); information on jet engines can be found on intra- and Internets (e.g. engine description and design, service manual). Each individual engine has a potential 'folder' of information describing its whole lifecycle, containing highly interrelated information stored in different media and easily running to several gigabytes, potentially terabytes.²

It is thus obvious that, if an organization is to keep its competitive edge and minimize its costs, it is of fundamental importance to collect, aggregate, share and reuse multimedia knowledge. There are several studies showing how ineffective sharing of information impacts on business: IDC³ has found that Fortune 500 companies lose at least \$31.5 billion a year by failing to share knowledge (Babcock, 2004) due to difficulties in accessing content which is scattered in multiple repositories and databases (Feldman, 2004), often in personal file systems or email. Unstructured information

is particularly critical as it is estimated to make up an average 80–85 per cent of a company's knowledge base expressed in some form of natural language, raw data, images or a mix of these (Raghavan, 2004). In practice, no one knows precisely what is available and where it is (Feldman, 2004).

Unstructured resources cannot be queried in simple ways, and therefore the information they contain cannot be used by automatic systems, or easily retrieved by humans.

To deal with unstructured knowledge the tendency is to search documents by using keywords (as is the case with search engines) or to provide a structure via a taxonomy (as in the case of many knowledge-management platforms) for the limited purpose of document retrieval.

This paper describes some methodologies for acquiring and sharing knowledge that provide a comprehensive support to knowledge management, beyond keyword matching. This technology has the potential to impact on the indexing community as it offers possibilities for the facilitation and enrichment of the work of indexers.

Levels of annotation

In traditional information systems, documents are described using metadata. For example, in a library information system (i.e. a catalogue) a book is identified by its title, author, publisher and ISBN number. The publisher also provides a set of keywords and the library will classify the book according to a specific classification system such as the Dewey Decimal Classification System.

Metadata contain information about the document, but not about the content of the document: this is the job of the index. The index can be used to retrieve content, not only by humans but also by automatic systems. Unfortunately while it is possible to employ a professional indexer for books, it is at present impossible to do the same for every document produced in a large organization: every report, every email, every image and so on. It is therefore necessary to devise methodologies that can capture the knowledge contained in the unstructured sources in the absence of a professional indexer.

Traditionally this is done by automatically extracting keywords or by explicitly associating keywords at the time of editing. When keywords are extracted automatically from documents and made available for searching (as in WWW search engines), the accuracy of the process cannot be guaranteed as the terms selected are not necessarily representative. This can be critical with medium and small archives, because this particular methodology for identifying key words automatically tends to work better with large databases. The result is likely to be better if the author of the document identifies a few associated keywords. However such keywords can be:

- Non-standardized: the same content is described by different keywords, which makes automatic retrieval difficult.
- Unrepresentative: too generic or too specific. This is quite common when a standard classification, such as the ACM taxonomy, is used.
- Not objective: selected keywords represent the author's view of the document, not the reader's. For example a report about faults on a car may be indexed using the fault as seen by the customer (e.g. a wet carpet) as keyword. Service engineers, however, will focus their search on the causes of the fault (e.g. an ill-positioned gasket).

In all three cases there is a mismatch between the keywords used by the author and those assumed by the person seeking to retrieve information. The result is a difficult or unsuccessful search.

To solve this problem, the Semantic Web Community⁴ has devised strategies for the indexing of document content that go beyond the simple use of keywords. The methodology is called semantic (or ontology-driven) annotation (Uren et al, 2006). An ontology is an explicit formal specification of the terms in the domain and relationships between them (Gruber, 1993). 'For example, the ontology of cooking and cookbooks includes ingredients, how to stir and combine the ingredients, the difference between simmering and deep-frying, the expectation that the products will be eaten or drunk, that oil is for cooking or consuming and not for lubrication, and so forth.' (Hendler, 2001).

Ontology can be as different as domains are: an ontology describing faults on jet engines will include the possible faults, their causes, the parts of the engine involved, the situations in which faults can occur (e.g. plane take-off) and so on. The goal of ontology-driven annotation is to associate portions of a document (a word or a group of words) to concepts and interrelationships in the ontology. Annotating a document about faults on a jet engine means finding all the references to faults, parts and the like, and their interrelationship (which fault occurred on what part, etc.) in the document and to make them available for retrieval in the form of indexes to the document. Similarly, an image can be annotated by associating portions of the image with concepts in the ontology. For example an image showing a broken engine part could be annotated by identifying the area of the picture where the damaged part is located and its damaged subparts.

Ontology-based annotation enables better retrieval and reasoning because of:

- The ability to access unstructured material by content, rather than superficial features such as metadata or keywords.
- The use of a standard vocabulary specifically defined for the task in hand. Terms in the ontology are defined *a priori* and shared by the community.
- The possibility of retrieving concepts that are more generic as well as more specific than the one used. For example, the search 'faults on a blade' will retrieve images of blades that are broken, worn, rubbed and so on.

Content, however, is not the only important part of the document. A further way to enrich documents is to incorporate free text annotations in the form of comments (Kahan et al, 2001). This has become quite a standard feature, especially in the knowledge-management community.⁵ Comments are used to integrate the text, adding information and knowledge not explicitly mentioned within the document. For example, a lawyer could add explanations about why he decided to refer to a specific regulation in a document (e.g. an EU directive), rather than to other regulations that might seem, in principle, more relevant. Such knowledge *about* the document and the decision-making process that generated it can be as valuable as the content of the document itself. Ontology-based annotation can be applied also to textual annotations, providing access not only to the document content but also to the additional information in the comments.

A further level of annotation can be used to connect a single document to the rest of the information or document collection, in much the same way as hyperlinks connect pages on the Web. Hyperlinks identify a specific string (the anchor text) and associate it with a document where, for example, more information can be found. The hyperlink concept can be extended beyond being simply a means of navigating among documents. One possibility proposed by Dzbor (2003) is to add 'services' to concepts in an ontology. When an ontological annotation is inserted in a document, the services are associated with the annotation. Figure 1 (from Dzbor, 2003) shows an example: an online Open University module displays the course material annotated using a simple ontology and services are associated to each annotation. By right-clicking on the annotation associated to 'precipitation', the students can access a Web service providing the definition of the term, other parts of the course where the term is mentioned, etc. This kind of annotation enables the connection between the document and the rest of the available information. In the OU example, it simplifies the job of the student and makes learning an easier and more enjoyable experience. From the knowledge-management point of view, it shortens the time (and hence the cost) required for accessing information. It becomes much easier to share and reuse knowledge.

Strategies for annotations

If this methodology is to work, the user must be given the means to annotate documents easily at the three levels (ontology-based, comments and connection with the rest of

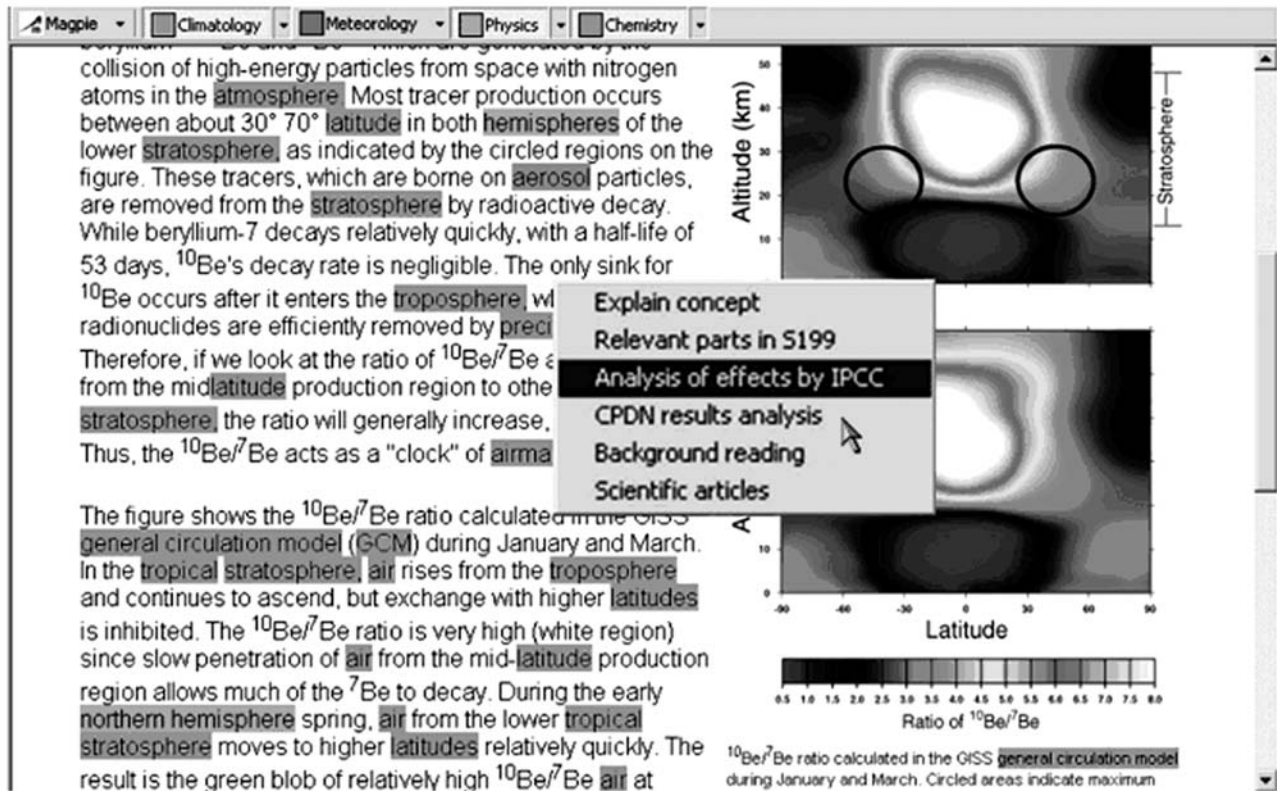


Figure 1 The Magpie Interface (from Dzbor, 2003)

the knowledge). The simplest and most direct way to do this is to provide an easy-to-use interface for accessing the electronic version of a document and for annotating it without the need to learn special languages (e.g. XML or RDF). Different authors have proposed graphical interfaces based

on the concept of text decoration by using the mouse and a tablet of colours (Uren et. al, 2006). In Melita (Ciravegna et al, 2002), each concept in the ontology has a colour; the user selects the appropriate colour and highlights the snippet of text, thus making explicit the text–concept relation (Figure

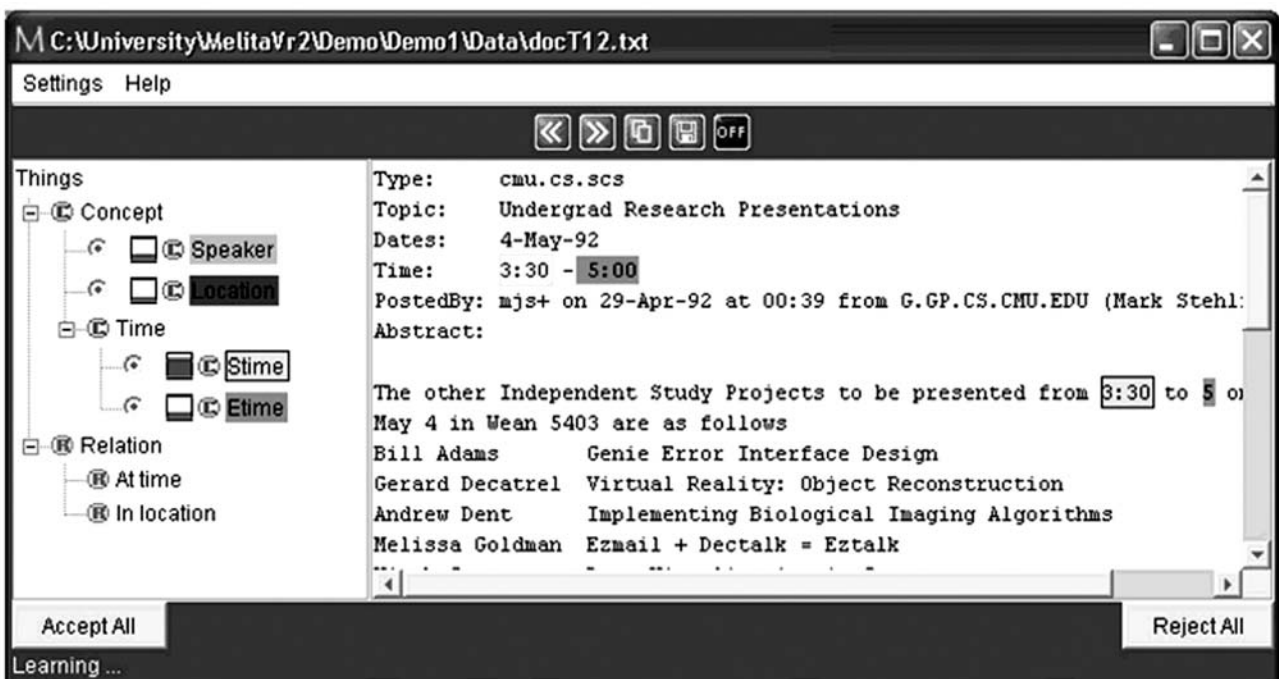


Figure 2 The Melita Interface

2). This association is translated automatically into XML code by the system. In *AktiveMedia*,⁶ images are annotated using the same concept/colour schema, but this time portions of the image are identified by drawing rectangles within the image (Figure 3). As the annotation is done by a human, its quality is in generally very high. However, manual annotation can be a long and expensive process, and becomes error prone if tiredness or hurrying intervenes. For this reason, methodologies have been devised either to help the user identify potential annotation (semi-automatic techniques) or to replace the human annotator altogether (fully automatic or unsupervised techniques). In *Melita*, a Machine Learning module monitors the annotations introduced by users into a set of documents. After a while, the system itself starts suggesting potential annotations for each new document and the role of the user changes from annotator to corrector of annotations, removing wrong annotations and adding missing ones. (Petrelli et al, 2005). Every time there is a correction, the system learns to avoid repeating the mistake. The continuous annotate/learn/correct/retrain process ensures that after a while the system achieves a high level of accuracy. Experiments carried out at the University of Karlsruhe have shown that the use of such

a paradigm can reduce annotation time by 80 per cent and double the quality of annotation compared with manual methods.

In cases where there is a large quantity of un-annotated material and it is not possible to have a human in the annotation loop (e.g. large legacy document collections or the WWW), methodologies for unsupervised ontology-based annotation have been devised (Dill et al, 2003; Ciravegna et al, 2004). Such systems are based on a set of modules dedicated to the recognition of ontological concepts and relationships in documents. Unsupervised annotation is a novelty and has been studied mainly for the WWW using some very shallow ontologies. The quality of the generated annotation is generally lower than the annotation obtained with manual or semi-automatic methodologies, as the systems tend to be imprecise. Such limited precision is acceptable in case of vast archives (e.g. the WWW) as that material would be otherwise inaccessible beyond simple keyword matching. However automatic techniques raise the issue of how trustworthy the extracted information/knowledge is, and are therefore applied as an advanced method for retrieving documents rather than a way of extracting knowledge from them.

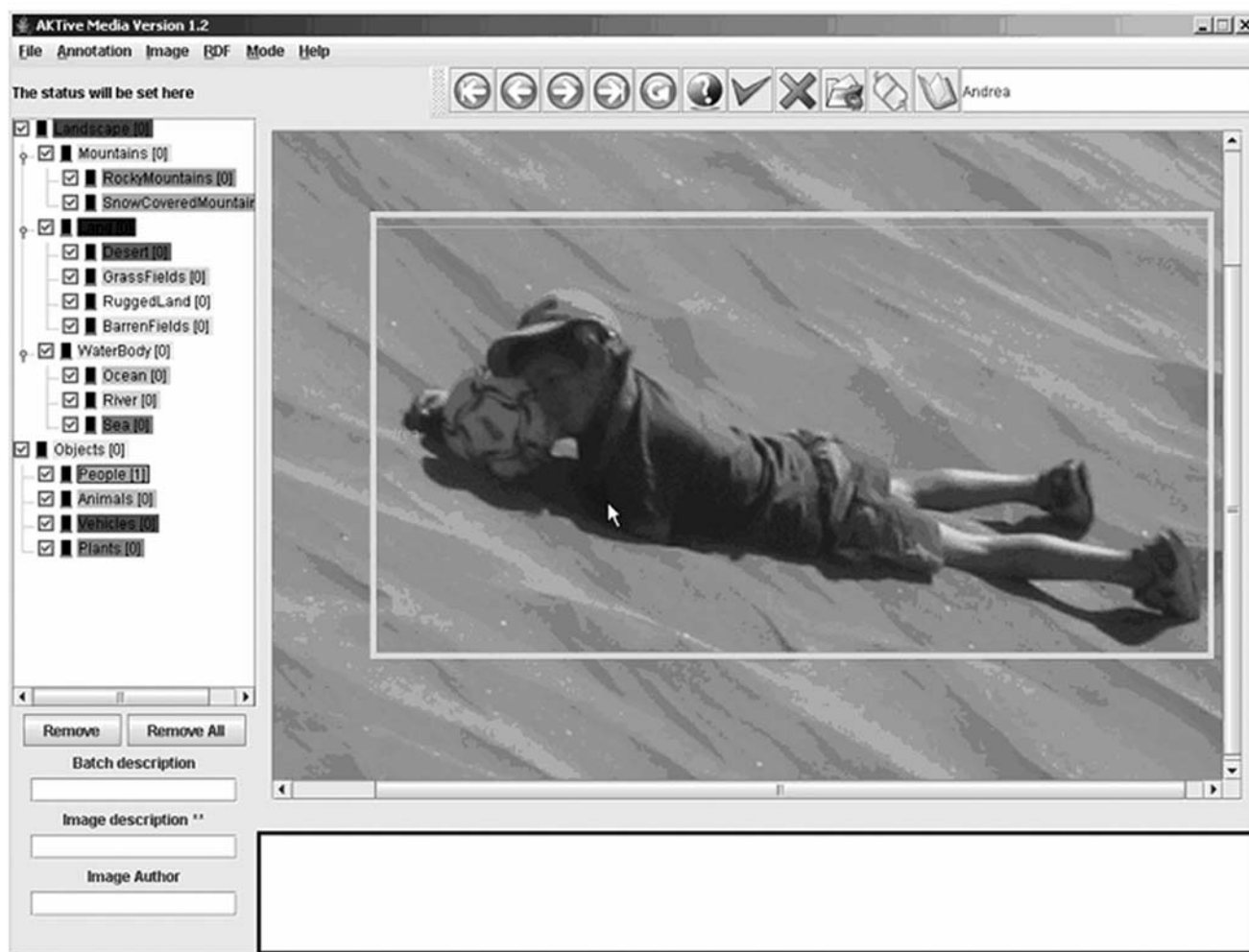


Figure 3 The AktiveMedia Interface

Conclusion and future work

This paper describes ontology-based document annotation as a way of making document content available when professional indexing is not a possibility. If used in a knowledge-management environment, document annotation enables effective capturing, sharing and reuse of knowledge. It is our claim that automatic or semi-automatic methodologies could help traditional indexers do their job more efficiently. Ontologies could be defined for specific domains (plants, cuisine, etc.) to enable ontology-based annotations of classes of books. Semi-automatic systems (like Melita) could help identify terms relevant to the index and – when its suggestions are accepted – generate the index automatically.

Notes

1. A gigabyte corresponds to a billion (10^9) bytes (a megabyte is a million bytes); see for an extended discussion <http://en.wikipedia.org/wiki/Gigabyte>.
2. A terabyte corresponds to 1,000 gigabytes; see <http://en.wikipedia.org/wiki/Terabyte> for details.
3. www.idc.com
4. <http://www.semanticweb.org>
5. Editorial tools like Microsoft Word and Adobe Acrobat also support comments.
6. <http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html>

Bibliography

- Babcock, P. (2004) Shedding light on knowledge management. *HR Magazine*, 49(5). http://findarticles.com/p/articles/mi_m3495/is_5_49/ai_n6041032
- Ciravegna, F., Dingli, A., Petrelli, D. and Wilks, Y. (2002) User-system cooperation in document annotation based on information extraction, in *Proceedings of the 13th International Conference, EKAW 2002*, Siguenza, Spain, October 1–4: Lecture Notes in Computer Science 2473. Berlin and Heidelberg, Springer.
- Ciravegna, F., Chapman, F., Dingli, A. and Wilks, Y. (2004) Learning to harvest information for the semantic web, in *Proceedings of the First European Semantic Web Symposium, ESWS 2004*, Heraklion, Crete, Greece, May 10–12: Lecture Notes in Computer Science 3053. Berlin and Heidelberg, Springer.
- Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A. and Zien, J. Y. (2003) SemTag and seeker: bootstrapping the semantic web via automated semantic annotation, in *Proceedings of the Twelfth World Wide Web Conference*, Budapest, Hungary, 20–4 May.
- Dzbor, M., Domingue, J. and Motta, E. (2003) Magpie: towards a semantic web browser, in *Proceedings of the 2nd International Semantic Web Conference (ISWC)*, Sanibel Island, Florida, USA.
- Feldman, S. (2004) The high cost of not finding information. *KM World*, available online at <http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=9534> (accessed 2 February 2006).

- Gruber, T. R. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition* 5, 199–220.
- Hendler, J. (2001) Agents and the semantic web. *IEEE Intelligent Systems*, 16, 30–7.
- Kahan, J., Koivunen, M., Prud'Hommeaux, E. and Swick, R. (2001) Annotea: an open RDF infra-structure for shared web annotations, in *proceedings of the Tenth International World Wide Web Conference*, Hong Kong, May 1–5.
- Petrelli, D., Lanfranchi, V. and Ciravegna, F. (2005) Working out a common task: design and evaluation of user-intelligent system collaboration, in *Proceedings of INTERACT 2005 International Conference on Human-Computer Interaction*, Rome, 12–16 September.
- Raghavan, P., (2004) Keynote presentation at WebDB 2004 International Conference, accessible online at <http://webdb2004.cs.columbia.edu/keynote.pdf> (accessed 2 February 2006).
- Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. and Ciravegna, F. (2006) Semantic annotation for knowledge management: requirements and a survey of the state of the art, *Journal of Web Semantics*, 4(1).

The authors both work at the University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom. Fabio Ciravegna is a member of the Department of Computer Science, and Daniela Petrelli of the Department of Information Studies. Email: [F.Ciravegna/D.Petrelli]@sheffield.ac.uk.

indexing specialists

Our team of experts have many years' experience in undertaking and designing effective indexes

- Books, journals, manuals ●
- Multi-volume works ●
- From 100,000 to 100 lines ●
- Unfamiliar subjects ●
- Design of style and layout ●
- Short deadlines met ●

Phone +44 (0) 1273 738299 Fax +44 (0) 1273 323309
 email richardr@indexing.co.uk
 website www.indexing.co.uk

INDEXING SPECIALISTS (UK) LTD
 Regent House, Hove Street, Hove
 East Sussex BN3 2DW, UK