

## Metadata without tears?

By Glenn Sanders BA, Dip Lib, GDDM, MBII, ARMA

This article is based on work submitted in partial completion of the MBII degree, RMIT University.

### Biographical Details

Glenn is one of Australia's leading consultants in document and records management. He has worked on software development for three commercial systems, written several books and articles, and been document manager for Tyndall Australia, Deloitte Touche Tohmatsu and (currently) EnergyAustralia. He is an active supporter of the RMAA and RECMGMT Listservs.

### Abstract

Capturing metadata about documents and people can add significant overheads to business processes. This article suggests that there is considerable explicit and implicit metadata available in documents and the systems used to create and access them, and that this metadata can be captured with little or no human intervention.

---

Documents contain explicit information. This can be minimal on a handwritten note, through to quite elaborate metadata in formal reports or articles submitted to professional journals. In many organisations, any major report, business case, budget proposal or the like has a standardised document control page, setting out ownership and sponsorship, approvals, status, and version control, in addition to the usual author, title and date. To this, the document management system should add security, document life cycle and disposal information, functional descriptors and so on. But even this amount of metadata is insufficient for most metadata standards, some of which specify a daunting number of mandatory fields.

In many cases, metadata is captured automatically, though not usually in a form suitable for document management. E-mail systems automatically record sender and recipient, transmission times, and in some cases transmission routes and errors. PC operating systems record date and time of file creation or last amendment (but not both, a major drawback). Office suites like Microsoft Office record various document properties such as title and author, and may record revisions and changes, but this requires specific user intervention, and if you copy someone else's document as a pseudo-template and 'save as' without changing its properties, you end up with very misleading data.

To date, document management systems have relied on such metadata, plus the document content itself, for management and retrieval. However, as document databases grow, even unusual terms will occur frequently enough that the results of a search will span more than the two or three screens usually deemed to be the tolerance threshold for useability<sup>1</sup>. There is a corresponding increase in both the amount and the level of human intervention required to create and maintain adequate metadata.

However, we cannot require that document creators and users spend additional time creating metadata. Users are rightly under pressure to focus on core, cash-earning business, and often lack the indexing skills needed for proper subject analysis. In virtually every report I have written in my consulting business for the last 18 years has been a statement that good document management follows as a byproduct of good business practice, and that it must be seamless and unobtrusive. Any attempt to require users to fill out more than two or three metadata fields, is courting failure<sup>2</sup>, particularly if any of the fields are seen to relate to abstract statutory obligations or long term archiving<sup>3</sup>.

There is much effort at present in Australia to overcome these problems automatically. The main thrust is on inheritance - the idea that when you save a document in a directory or folder, it automatically inherits the attributes of that folder, such as functional descriptors, security, and disposal status. However, there is a considerable overhead in maintaining the inherited attributes and ensuring they are closely aligned with business needs. Much of the required information cannot be derived from the documents themselves: you must consult users, thesauruses and disposal schedules as well.

The situation is even worse for tacit information. This takes two forms: information about the document, and information about the people using and creating the document. However, we do know more about both than we might think, and capturing much of this information does not require additional human intervention - it is implicit in the systems we use to create and access the documents.

Privacy issues aside, and assuming no more than a modern EDM system and a corporate intranet for access to organisation charts, skills registers and resumés, we know quite a lot about the people involved:

- Functional role
- Who else they work with
- Each project they work on
- Who else works on each project
- Which documents they create, modify, access and delete
- Who else they communicate with (at least by e-mail)
- Skills, qualifications
- All of the above historically

And we know even more about documents:

- Dates created, accessed, modified, deleted
- System used to create, access, modify, delete
- Which people create, access, modify, delete, and when
- Which process the document is part of
- Language or terminology used
- People, organisations, projects referred to
- Other documents explicitly referenced (eg footnotes, hyperlinks)
- Other documents explicitly related (eg compound documents, encapsulation)
- Other documents tacitly related (eg via thesauruses, classifications, file plans, linguistic analysis, pattern matching, sequence, chronology etc)
- All of the above historically

Why is this tacit information important? Apart from the useability and metadata maintenance issues mentioned earlier, organisational and personal information needs change - dramatically - over time<sup>4</sup>, and this in itself can generate new tacit information<sup>5</sup>, hence the last item in each list above.

That we can derive much information from these tacit sources is nothing new. Suitable techniques, particularly relating to chronology, have been used by investigators for many years<sup>6</sup>. The techniques are little different from those used for business intelligence analysis, and are already creeping into knowledge management systems<sup>7</sup>.

Workflow has particular potential here. Documents do not exist in a vacuum, and investigators and historians stress the importance of sequence and chronology. Workflow systems can record much of this, without human intervention, once properly set up. Again, this is not new technology.

My argument is simple: once, long ago, we could afford the resources to manually intervene (usually with ledger and quill pen) on every business document, to record and manage it. Now we cannot, nor would it be effective if we did, because user needs are much more complex and change frequently.

The business intelligence and investigatory models, plus workflow, show that we can derive, automatically, much useful information from the documents - and from what we know of the people. The two in combination, plus the increasingly sophisticated search techniques available<sup>8</sup> mean that we should be able to develop effective document management systems, including both explicit and tacit information, without requiring unjustifiable overheads to create metadata, or get inside people's heads.

All we need now are the systems, and the management skills to exploit the information.

---

<sup>1</sup> For example, don't even try a Web search for 'document management'

<sup>2</sup> Personal discussions with software vendors and clients.

<sup>3</sup> And no-one will dare suggest, in this increasingly electronic and decentralised age, that we try and do it centrally, will they? Please?

<sup>4</sup> Boyd, Stowe "Rethinking knowledge management: this time it's personal" *Message* no.1, 2001 ([www.knowledgecap.com](http://www.knowledgecap.com))

<sup>5</sup> Borghoff, U and Pareschi, R "Information technology for knowledge management" *Journal of Universal Computer Science* 3(8) p835-

<sup>6</sup> Personal discussions with ASIC investigators and DPP lawyers (look, *I* was interviewing *them*)

<sup>7</sup> Tkach, Daniel (ed) *Text Mining Technology: Turning Information Into Knowledge*. IBM Software Solutions, February 17, 1998; Seiner, Robert S "Knowledge management: it's not all about the portal" *The Data Administration Newsletter* downloaded from [www.tdan.com/i014fe04.htm](http://www.tdan.com/i014fe04.htm) 15 Dec 2000.

<sup>8</sup> Moad, Jeff "In search of knowledge: new tools aim to turn unstructured data into a corporate resource" *PC Week Online*, 7 Dec 1998; Ananthaswamy, Anil "You hum and I'll find it" *New Scientist*, 17 March 2001