The Structured Information Manager: A Database System for SGML documents

R. Sacks-Davis
RMIT, Australia
rsd@mds.rmit.edu.au

One of the important standards for document interchange and representation that has emerged is SGML, the Standard Generalized Markup Language. SGML is designed to capture the logical structure of documents, i.e. the logical components such as titles and paragraphs and their interrelationships. SGML is a complex standard, and the design of a database system for managing SGML documents poses many challenges. In this talk, we describe an SGML conformant database system, called the Structured Information Manager (SIM), and illustrate how the support of document structure can help in many important applications by describing how SIM has been deployed to provide public access to databases of legislation.

The Structured Information Manager (SIM) is a document database system designed to manage multigigabyte collections of documents containing unstructured text (ASCII), structured text (including SGML and MARC), binary objects (such as images and videos) and other kinds of data.

As an information retrieval system, SIM provides a client-server model of processing and supports a wide range of user interface platforms, including command line, MS-Windows, MacIntosh, and X. SIM uses compressed inverted file technology for accessing large text collections using both query and browsing paradigms [ZobMof92]. Both Boolean and natural language queries are supported and response times are sub-second, even for multigigabyte databases.

SIM is standards based. It provides direct support for SGML, the international standard for document representation and interchange and Z39.50, the international standard for client server communication in an information retrieval applications [SacArn95]. For Web access, an HTTP to Z39.50 translation is supported. By directly supporting SGML, documents of arbitrary complexity can be supported by SIM and a collection of documents can be treated as a database of information.

SIM is supported and marketed in Australia and New Zealand by Ferntree Computer Corporation. Research and development of SIM is undertaken by RMIT's Multimedia Database Systems Group. Users of SIM include CSIRO, Australia's national scientific research organization, Macquarie Dictionary, State and Federal Departments.

SIM is used by the Government of Tasmania for the drafting and consolidation of legislation. This system is used by the Office for Parliamentary Counsel within the Department of Premier and Cabinet. Legislation can contain both substantive provisions and provisions which apply textual amendments to the substantive law. To determine the state of law at a particular point-in-time, the legislation has to be consolidated by applying amendments to the substantive provisions up to that point in time. The SGML Markup makes it possible to automate the consolidation of legislation at arbitrary points in time. Thus SIM is able to provide for point in time searches which return the correct state of the law at any point in time [ArnSac].

As well as supporting automatic consolidations, the use of SGML assists in the drafting of new amendments. Drafters are able to modify the current legislation directly, and the text for the appropriate amending legislation can be automatically generated.

Another feature of the system is the drafting environment for new legislation. Drafting of legislation is performed using Microsoft Word with additional templates and macros. Two way translation between RTF (Rich Text Format) and SGML is done automatically using translators developed as part of the SIM software.

References

[ArnSac] T. Arnold-Moore and R. Sacks-Davis. Databases of Legislation: the Problems of Consolidations. *Law Library Journal*, (to appear).

[SacArn95] R. Sacks-Davis, T. Arnold-Moore and A. Kent. A Standards Based Approach to Combining Information Retrieval and Database Functionality. *International Journal of Information Technology*, 1(1):1–16, 1995.

[ZobMof92] J. Zobel, A. Moffat and R. Sacks-Davis. An Efficient Indexing Technique for Full-Text Database Systems *Proceedings of the 18th. Int. Conf. on Very Large Databases*, 352–362, Vancouver, August, 1992.

**Proceedings of the 22nd VLDB Conference
Mumbai(Bombay), India, 1996**