gla.smp/smp/CDocS-app2/2001-03-19/1

UNIVERSITY OF GLASGOW

The CDocS Committee Document System

James Currall, and the ERM Project Team

March 2001

The CDocs System

The CDocS system contains three components:-

- A creation process employing: Wizards, Templates and Macros in Microsoft Word to aid the document creator to produce a well structured consistent record with adequate Metadata added at creation time
- A Conversion process which converts the file format produced by Microsoft Word into a more durable Extensible Markup Language (XML) form which can be transformed into a variety of output formats (currently HTML for web delivery, but the ability to transform to PDF is also under development)
- An upload process which, through a web form based interface, allows committee clerks to transfer the documents created in Microsoft Word to the central server where they are transformed and made available on committee web pages.

Document Creation

CDocS seeks to shape document creation practice in two areas:

- the capture of some basic descriptive information (**metadata**) about the record
- an approach to the use of word processor features which adds some very simple description of the **logical structure** of the document

Analysis of the documents in use highlighted a small set of 'document types' (such as 'agenda', 'minutes', 'reports', 'papers' etc.) distinguished by their logical structure, and a corresponding set of Microsoft Word templates was created. The use of the templates was supported by a set of macros.

The first function of the macros was to provide a simple dialogue interface through which the creator of a new document supplied (or selected from drop-down lists) the values of a small set of metadata properties. The values were supplemented by other data generated by the macros and stored in the document as custom document properties (and, where appropriate, as document content). The function of this metadata is not simply for resource discovery, but to enable intellectual control of the record and to provide contextual description. The property set was drawn from the Dublin Core Element Set and from elements based on the National Archives of Australia's Record-keeping Metadata Standard [1].

The second role of the macros was to support the capture of a basic description of logical structure. This required the creators to move away from the practice of applying ad hoc formatting and to adopt a rigorous use of Word paragraph and character styles. A macro performed basic "validation" of the creator's selection of paragraph styles against a structural model when the document was saved. This approach does require a shift in practice, and the project directed considerable efforts to education. In some cases, authors had to simplify the structure of their text to fit within the constraints of the tools and models - it might be argued, however, that such structural simplification resulted in documents of greater clarity!

In addition, macros provided some functionality specific to the use of individual document types (such as the transfer of content between the agenda and minutes of a single meeting, establishing relationships between agenda and minutes of previous meetings etc.).

Document Upload

Once the document has been completed by the clerk, the upload process can begin. The clerk visits a web page where there is the option to upload three types of document:-

- documents created within the committee document system (or converted to that form),
- other documents created in Microsoft Word (which have not been converted to CDocS form),
- other files (such as spreadsheets, etc.).

At Glasgow access to this web page is controlled by an LDAP-based authentication[2] and access control system which is being developed as part of the Scottish Middleware Project. An appropriate upload form is then presented which collects information relating to the names of the files to be uploaded, the committee to which the documents apply, the type of document (agenda, minutes, paper, report, etc.) and the date of the meeting concerned. The first of these is required to select the correct file and the remainder to ensure that the converted documents are put in the correct location and the committee index files are rebuilt in the correct way.

For committees that have reserved business and which have set up systems to ensure that authentication and authorisation are properly taken care of, there is also a checkbox to indicate if the item concerned is 'reserved'.

The presence of the information requested is checked by some simple Javascript processing when the 'submit' button is clicked.

The remainder of this section concentrates on the first category of document (those created within the committee document system) as this involves the full conversion and transformation process. The other forms employ much simpler processing, but do not achieve the archival quality, nor is so much 'down-steam' processing possible, as they do not have a recognisable structure on which that can work.

The committee documents are first checked to ensure that the name of the committee, date of meeting and document type given on the upload form conform to the metadata given in the documents themselves. If this check fails, the clerk is returned a web page detailing the problem, otherwise the series of conversion processes described in the next section take place.

If there are no conversion problems, the clerk is returned a web page which contains an HTML representation of the document that was uploaded, so that the clerk can check it. The clerk is also sent an e-mail to his/her registered e-mail address indicating the next step. The use of e-mail for this purpose allows us to separate the process of upload from the process of approval of the document for publication as the e-mail could be sent to a 'supervisor' or the convenor of the committee to action the next step. This e-mail contains:-

- links to the HTML form of the uploaded document and its metadata (in a temporary holding area),
- if the document is an agenda and the facility is enabled, a link to a copy of the agenda with check boxes for 'starring' of items which should be brought to the attention of the committee members,
- links to a web page which will automatically trigger the process to copy the document to the relevant committee web site and make it available (this link includes security checks to ensure that the correct person is visiting it).

If there is no starring the final process simply copies the HTML files to the committee pages and the XML and RDF files to the archive. If starring is in use the process cycles back to the last step of the conversion to add starring to the agenda.

A final web page is presented to the clerk at the end of the process confirming that everything has worked and providing links to the live copies of the document and its metadata.

Much of the detail concerning how the upload process works is handled in configuration files which can be set up to indicate which features should be employed (starring, reserved business, etc.) and the locations of servers, committee directories, etc.

Document Transformation

The document is saved from Microsoft Word in Rich Text Format (RTF) which is then processed by a tool which exploits this 'structural description' to generate an Extensible Markup Language (XML) document. The tool used is Logictran's RTFtoHTML (now RTF Converter)[3], which offers a table-based configuration model to generate output on the basis of the use of paragraph and character styles in the input document. Although designed to generate HTML, it permits the generation of an XML document conforming to a Document Type Definition (DTD) designed for this application -with careful design, quite complex 'nesting' can be implemented, together with a Resource Document Framework (RDF)/XML representation of the metadata.

This XML document serves as the basis for the generation of other representations as required. In the short term, this has been limited to the use of XSL-T to perform simple transformations, generating HTML renditions of the document content and metadata. Initially James Clark's XT XSL-T processor [4] was used; this was replaced by Michael Kay's SAXON[5] as it offers higher conformance with the XSL-T 1.0 Recommendation. These transformations were performed as processes on a Unix server, initiated by the committee clerk through a Web interface (with appropriate authentication checks), and this became the mechanism of distributing the documents to members. That process utilised the document metadata to update relevant HTML index/navigation pages on the intranet server.

Other forms of indexing, abstraction and 're-presentation' of the XML document set become possible. Indeed, more sophisticated application-specific processing of the document content is feasible, such as analyses of meeting attendance, the triggering of reminder messages to committee members identified as having tasks prescribed at meetings, or the tracking of items through the decision-making process through time. Such processing represents a visible 'return' for the committee clerks and members on the 'investment' made by the document creator in adopting a standardised approach.

Furthermore, as it uses a standards-based syntax rather than a proprietary encoding format, the XML document is inherently more suitable for longer term storage. It should be noted, however, that this separation of logical structure (described in the XML document) and presentation (in the stylesheet or transformation process) does add complexity to questions of testing authenticity, since the user's 'experience' of the record becomes the product of a process operating on a number of independent inputs.

References/Links

[1] National Archives of Australia, Recordkeeping Metadata Standard for Commonwealth Agencies, Version 1.0 National Archives of Australia, May 1999
http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm

[2] Currall, James and Henderson John. Use of LDAP at Glasgow and St Andrews.
http://www.gla.ac.uk/projects/scotmid/gendocs/ldapuse-smp.html

[3] Logictran http://www.logictran.com/

[4] XT http://www.jclark.com/

[5] SAXON http://users.iclway.co.uk/mhkay/saxon/