[Presented at: *Translating and the Computer 14*.,10-11 November 1992]

Interchange of Documents in Electronic Form Utilising International Standards

Edward John Moore
EJM Associates

The paper describes a model for document interchange taking into account all elements of an individual document including text, graphics and tables, while concentrating on the means of interchange for textual matter between dissimilar hardware platforms and application packages.

It particularly addresses the means to supply documents in a neutral form for translation at remote sites and the subsequent retrieval of such documents for publication production.

The Model

The United States Department of Defense CALS initiative addresses many issues in the procurement and support of large volume produced items, the issues related to documentation is particularly pertinent in the more general publishing arena, where documentation is subject to continual updating and re-issue either in hard copy form or in digital form within large corporate organisations.

CALS provides the foundation for those documents to be created on a wide range of equipment in the relevant areas within the organisation but then transferred to a central controlling operation for revision and issue procedures to be carried out.

CALS the initiative

Computer-aided

Acquisition &

Logistic

Support

Background

In 1985 the US DoD approved recommendations to achieve improvements in weapon system design, also the accuracy and availability of technical information.

The strategy was to effect a transition from paper intensive weapon system design, manufacture and support to a largely automated and integrated mode of operation.

Its aim was to establish a means to acquire, process and use technical information in digital form. The development process commenced in early 1989 and is continuing to address the following issues in two distinct phases.

i) The existing lack of integration which makes it difficult to design systems "right first time".

ii) The implementation of costly design changes through the product life cycle to ensure producibi1ity and supportabi1ity.

iii) The many incompatible automated systems being used by contractors to produce data presented on paper, only for it to be re-entered into Government databases.

The volume of paper generated and the potential for errors created the need to improve the acquisition and support processes.

Typically a single weapon system may generate 3,500 new manuals adding 1,000,000 pages to the existing volume, 20% of these pages may need changes each year.

CALS Objectives

As in most organisations cost benefits are the criteria by which changes are judged. In the case of the US DoD a minimal unit saving on a project provides a substantial saving taking into account the large volumes of items being procured. The savings may not only be in the capital cost per unit but in the subsequent price of support in terms of time and provision of spares.

To achieve these savings the following objectives were identified.

i) Reduce lead time by

a) Utilise integrated data to improve industry responsiveness.

b) Utilise shared data environments to shorten

design development, production and re-supply time.'

c) Utilise integrated planning to reduce out of service time.

ii) Reduce product life cycle costs by

a) Eliminate separate processes in design, manufacturing and support.

b) Effectively use digital information for acquisition, logistics and field operations to reduce the use of paper.

c) Inter-operabi1ity achieved by shared data resources.

iii) Improve quality by

a) The integration of key databases, in a near real time environment, considering producibi1ity, reliability and maintainability linked to the use of CAD and CAE tools. This will minimise errors in design and manufacturing at the earliest possible time in the product life cycle.

How does CALS achieve objectives

CALS utilises concurrent engineering and integrated logistic support philosophies to meet its objectives.

Concurrent Engineering CE

CE addresses the CALS requirements from the point of design influenced by manufacturing requirements, this undertakes to interface automated manufacturing processes with the product data and provide feedback to affect new design.

Integrated Logistic Support

Provides a means of supplementing the CE activity by providing information regarding reliability and maintainability of a product and to incorporate such information into the initial design. Also identify information regarding maintenance procedures, spares provisioning and levels of staffing to support the product.

In effect ILS provides information at the outset of a product life of what it is going to take to operate and maintain it through its life span. Between 60 and 80 per cent of a product cost is operating and maintaining. The objective is to do what is possible at the design stage to minimise the life cycle costs by designing a product that is easier to

operate and maintain or repair. In particular identifying what is needed in terms of people, spares and tools and when they are needed. Then, to obtain these things at the right time and have them in the right place in the appropriate amounts to support the product.

CALS Phase 1

Aims to replace the hard copy document flow with digital file exchanges.

1 Each prime supplier has a number of issues to address:

2 The traditional islands of information within his own organisation.

3 The subcontractors own islands of information.

The compatibility of data sources both in terms of operating system and application packages used with in these islands of information.

The DoD recognised that if the information from these sources were to be compiled to provide digital transfer to their organisations a common denominator needed to be identified to enable this to happen.

Indeed a number of common denominators were identified to accommodate the elements within documents i.e. text and graphics.

Text exchange between dissimilar computer systems could well be achieved by the use of ASCII coding although this would not impart information regarding style and structure within the document. Equally each word processing application package utilises differing internal codes to accommodate such information hence restricting the use of any one such product.

Graphics are produced on many applications packages producing varied output dependant on the requirement of the product. For instance in the case of electrical wiring diagrams two dimensional illustrations would be appropriate, where solid modelling is concerned then the illustrations would come from a three dimensional application package and in the case of structures then a freehand illustrative package is used to generate illustrations. Each of the above had to be taken into consideration when determining the common denominators for data interchange.

To address these issues the DoD identified and adopted a particular set of existing standards and incorporated them into their own specifications for data interchange.

These US DoD specifications are:-

MIL-STD-1840A
MIL-D-28000
MIL-M-28001A
MIL-R-28002
MIL-D-28003

The Standards

MIL-STD-1840A

This is the parent document for the family of military specifications through which CALS standards are published.

It provides enveloping rules for organising files of digital data into a complete document. This specification identifies the media and formatting of data for transportation between sites such as telecommunication media like X.25 and X.400 or 9 track 1/2 inch magnetic tape.

MIL-D-28000

Specifies the format for graphical data interchange of CAD generated vector graphics. Presently this utilises specific subsets of the Initial Graphics Exchange Specification (IGES) version 3 but a discussion paper regarding the use of version 5 is circulated within the CALS Test Network, users and relevant vendors for consideration.

MIL-M-28001A

Specifies a document markup language to compile text and reference illustrations to form a complete document. The Standard Generalised Markup Language (SGML) identifies ASCII tags as labels for the elements of the document, it also imposes constraints on the structure of the document lending itself to an object oriented database environment to be considered in CALS Phase II.

MIL-M-28002

Specifies the format for interchange of raster images typically created by scanning hand drawn illustrations, for this application CCITT Fax Group 4 has been selected.

MIL-D-28003

Specifies the use of Computer Graphics Metafile (CGM) for delivery of presentation graphics such as PI charts and the like from business application software.

General Synopsis

All the above specifications utilise international or US standards conforming to or using subsets of ISO, CCITT and ANSI specifications.

The US DoD are progressively considering developments in the use of each standard. The main activity tends to be in the use of SGML where a user defined set of instructions known as a Document Type Definition (DTD) has been created by the DoD. Feedback from the users of the original DTD found that it was in some ways cumbersome to use so now each document has three component parts and corresponding subsets to the main DTD.

CALS Phase 2

It is agreed that the next phase of the CALS initiative is some years off completion, but having digital interchange of information in place means that the foundation is laid for moving to Phase 2.

CALS Phase 2 aims to replace internal interfaces with distributed databases holding information in an object oriented environment which combines sub-contractors, prime contractors and end users information into an open system. This system will control access and revision procedures of data at the same time allowing the philosophies of ILS and CE to be applied in meeting the overall objective of the CALS program.

In meeting these objectives new standards are emerging for use in integrated database applications. These are the Standard for the Exchange of Product Model Data (STEP) being developed in the European standards arena and the US Product Data Exchange Specification (PDES) a parallel standard to STEP. Indeed the US have move toward the European standard and renamed PDES as PSTEP.

So What?

CALS input in the market place is yet to have effect as far as the general user is concerned. But its philosophies provide the basis for what most users are looking for, the means to interchange information between dissimilar systems, e.g. in terms of hardware architecture, operating system and application package.

The major computer manufacturers are seriously addressing the issues in order to keep their position in the US Military market place, a considerable segment. But the benefits are immeasurable to the general user.

The product life cycle costs for a major military acquisition correspond in value to that of a major construction project or volume production of automotive parts. The objective is to reduce costs in terms of design, manufacture or construction and finally in repair and maintenance, CALS provides the means to reduce life cycle costs, hence increase margins.

Most organisations do not have the opportunity to set aside time and capital to consider their operation in such a manner. Here the US DoD have made major strides in providing a framework which can be applied, independently of product type, to achieve savings on life cycle costs. A task which should be more easily achieved in a single environment than having to consider external sources of information.

The initial activity addresses the effective compilation of technical documentation by providing a means of extracting data from various application packages meeting specialist needs within the organisation.

Who else utilises these standards?

Air Transport Association of America
European Standards Bodies
BSI, DIN, AFNOR
HMSO
US Securities and Exchange Commission
German Publishers Association
Association of American Publishers
Automotive Industry

Document Interchange and the Translator

The movement of documents between publisher and translator face the same problems as the US DoD in their initial analysis, that of dissimilar hardware platforms, (PC, Macintosh, UNIX) and application software, (WordPerfect, Word, Wordstar, Interleaf, Aldus PageMaker, Quark Xpress, 3B2) therefore the provision of a means to move data between hardware platforms that ensures the integrity of the document across platforms would provide a cost and time benefit to the publisher. Utilisation of the international standard ISO 8879 SGML addresses this particular issue.

What is SGML (Standard Generalized Markup Language)

The ISO introductory paragraph for SGML states:

> This International Standard specifies a language for document representation referred to as "Standard Generalized Markup Language" (SGML). SGML can be used for publishing in its broadest definition, ranging from single medium conventional publishing to multi-media database publishing. SGML can also be used in office document processing when the benefits of human readability and interchange with publishing systems are required.

Such a general introductory paragraph does not enhance the understanding of SGML but the implications of such a broad statement are far reaching. In short SGML is a neutral language, that is it is format and system independent. It provides a tagging system based on structure and enforced by a set of rules which are both human and machine readable. It identifies where elements can and must occur, identifies the order and frequency of elements. Finally it separates form from content, in other words the document is not restricted to a single use, the single source of information can be utilised in many forms, formatted for hard copy, document image processing or microfiche, database application for online retrieval and hypertext applications.

How Does it Work?

There are three areas for to consider for SGML to operate, the rules for SGML defined by the SGML Declaration, the rules for the document type defined in the Document Type Definition, and the documents themselves referred to as the Document Instance.

The SGML Declaration

The rules for SGML are rarely changed but provision for

customisation is made for in the SGML Declaration. This specifies the SGML applications, the SGML syntax and features supported.

The character set that is used for the type of document is specified, this provides a reference to the validation process of how to interpret ASCII character coding for a particular application.

e.g. BASESET "ISO 646-1983//CHARSET

	Character Number		Action
DECSET	0	9	UNUSED
	9	2	9
	11	2	UNUSED
	13	1	13

To supplement the base character set special characters are accommodated by the use of entity sets organised to reflect the structure of ISO 6937 character sets and include:

2 Latin sets
Greek alpha
Greek symbols
Cyrillic alpha used in Russian
Cyrillic alpha not used in Russian
General technical
Mathematic symbols

The concrete syntax declares the symbols used in the document type definition.

e.g. Role Symbol Description

AND	&	AND connector
SEQ	,	Sequence connector
OPT	?	Optional occurrence

The quantity set declares value parameters such as length of a tag name, number of tokens in a group and the number of nesting level of open elements.

Finally, the features section of the declaration specifies the use of optional features that may be used. These include minimization features such as omit tag and short tag; link features e.g. simple or implicit; other features such as the use of sub-document.

The Document Type Definition (DTD)

Specifies the naming of elements by use of labels or tags to which attributes can be associated e.g.

```
<!ELEMENT report (front, body, rear)
<!ATTLIST report    version   NUMBER    "01"
                    security  <ts|sec|unc) #REQUIRED>
```

The DTD also specifies the rules for determining the structure of a document by the use of the symbols specified in the document declaration in association with the definition of element labels and take the form:

```
<!ELEMENT report    - -  (front, body, rear)>
<!ATTLIST report    version   NUMBER    "01"
                    security  (ts|sec|unc) #REQUIRED>
<!ELEMENT front     - -  (abstract, contents)>
<!ELEMENT body      - -  (title, topic+)>
<!ELEMENT topic     - -  (toptitle, story)>
<!ELEMENT toptitle  - -  (#PCDATA)>
<!ELEMENT story     - -  (item, para+, storytxt?)>
<!ELEMENT storytxt  - -  (subhd, para+)>
<!ELEMENT subhd     - -  (#PCDATA)>
<!ELEMENT para      - -  (#PCDATA)>
```

The above utilise some of the symbols specified in the concrete syntax to determine the structure of the document described as a "report" in SGML terms the tag <report> is the first required element and encompasses every element of an individual report.

In the above example the next structural level allows for a front matter and a body and end matter to be included. At the next level within the body matter there is a main title followed by one or more topics, each topic then contains a topic title and the story. The lowest level of a document is that of PCDATA parsed character data (text to you and me) that is text conforming to the specified character set.

The structure of a document is user defined and can be as severely or loosely constrained depending upon the nature of the document, this is determined by the use of different connector in the DTD. The most extreme example is the use of "sequence" connector given by "," and the "or" connector given by "|".

e.g. 1 (abstract, contents)
indicates there must be an "abstract" followed by a "contents".

e.g. 2 (abstract | contents)
indicates that either "abstract" or "contents" can appear and in any order.

The document instance contains the tagged data with appropriate attributes specified and the textual content for the document.

e.g.

```
<report   version="01" security="ts">
<front></front>
<body>
<title>Standard Generalized Markup Language</title>
<topic>
<toptitle>Document Instance</toptitle>
<story>
<item>
 <para>The  document  instance  comprises  ASCII character codes  to represent the  tags and the  text elements of a document.</para>
< /item></story></topic></body></report>
```

The Parser

All of the above are required to generate an SGML document, the validation of the Document Declaration, the Document Type Definition and the Document Instance is a software program called a parser. The parser verifies the Document Type Definition with regard to its syntax and then compiles it, once compiled the SGML text file (the Document Instance) can be verified against the Document Type Definition.

There are a number of parsers available commercially which function on various platforms including UNIX, MS DOS and in the Macintosh environment. Each has various functionality because the specification (ISO 8879) allows for optional features that supplement the base guidelines. So in choosing a parser it is necessary to be aware of those features that are likely to be used and vet the parser vendors prior to purchase. Alternatively the UK SGML User Group have considerable information for guidance.

It must be said that where complex Document Type Definitions are being used the operation of the parser is memory intensive and the minimum recommended configuration should be met to eliminate frustration in use.

Many SGML editors incorporate a parser to allow validation of the Document Instance to be carried out interactively. Equally tools for the generation of Document Type Definitions support parsers for DTD validation.

To enable the movement of documents the DTD and the Document Instance are ASCII files. DTD's fall into two categories that of Private and Public.

A Private DTD is one generated by an organisation for use amongst its own community and for data to be validated by a

new user would supply the DTD with the initial set of document instances.

A Public DTD is one registered with a recognised body for public access. The recognised body in the UK is the National Computing Centre, Manchester and NIST in the USA where both hard copy and digital supply of Public DTD's are available.

Conclusions

1 SGML is a recognised international standard ISO 8879

2 SGML is a neutral language independent of : -
Form
System
Allowing documents to be moved freely between systems and output in many forms.

3 SGML provides a framework for a document maintaining its integrity for both the translator and the publisher.

4 SGML provides the means to support international character sets.