NISTIR 4560

Government Document Processing Requirements Report

Roger F. Sies

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards and Technology
Computer Systems Laboratory
Office Systems Engineering Group
Gaithersburg, MD 20899

U.S. DEPARTMENT OF COMMERCE
Robert A. Mosbacher, Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

NIST

NISTIR 4560

Government Document Processing Requirements Report

Roger F. Sies

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards and Technology
Computer Systems Laboratory
Office Systems Engineering Group
Gaithersburg, MD 20899

April 1991

U.S. DEPARTMENT OF COMMERCE
Robert A. Mosbacher, Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

GOVERNMENT DOCUMENT PROCESSING REQUIREMENTS REPORT

by

Roger F. Sies

Systems and Software Technology Division
National Computer Systems Laboratory
National Institute of Standards and Technology

March 20, 1991

CY90 CALS TASK	3.1.3.	"Host working meetings for harmonized development of a common government application"
	3.2.1.	"Represent 28001 requirements at OSI Implementors' workshop SGML meetings"
Deliverables		28001/OSI Requirements Report

- ABSTRACT -

This report describes several activities of the Office Systems Engineering Group in the area of electronic publishing standards. It gives an account of the July 30, 1990 workshop on Electronic Information Exchange Standards Used in Document Processing Applications and the list of user requirements that came out of that workshop. The report also discusses other efforts the Office Systems Engineering Group has made to help bring about the harmonization of electronic publishing standards.

- Table of Contents -

1. INTRODUCTION

On July 30, 1990, the Office Systems Engineering Group at NIST presented a workshop on Electronic Information Exchange Standards Used in Document Processing Applications. This workshop was part of our work with the Computer-aided Acquisition and Logistic Support (CALS) project and represents the deliverable of meeting minutes and summary report under the Common Government application objective. The workshop was one of our efforts to bring the various factions such as Standard Generalized Markup Language (SGML) and Office Document Architecture (ODA) together. The second goal of this workshop was to develop a set of user requirements for electronic information exchange standards and document processing applications.

The workshop was held to provide a forum for individuals in private industry and government to exchange information on topics that relate to the selection and application of electronic information exchange standards, such as ODA, SGML, Document Style Semantics and Specifications Language (DSSSL), and Standard Document Description Language (SPDL), among others, that are used in document processing environments. Refer to Appendix A for more definition of these and other terms used in this document.

During the past year, the Office Systems Engineering Group has taken several other steps toward the harmonization of electronic publishing standards through support of the Open Systems Interconnection (OSI) Implementors Workshop. In June 1990, both Allen Hankinson and Lawrence A. Welsch spoke to the X3V1 committee, the standards group for text and office systems, about the importance of resolving differences between task groups within that body involved with SGML and ODA. The clear message was that the government needed one standard or standards that worked well together in this area. Also in June 1990, Lawrence A. Welsch gave a tutorial on SGML to the ODA Special Interest Group (ODA SIG) of the OSI Implementors Workshop. On this same occasion, Mr. Welsch briefed this group on the message given to X3V1 earlier in the month. In addition, we finished and provided an SGML encoding of the Level 3 Document Application Profile (DAP) to the ODA SIG of the OSI Implementors Workshop. The ODA SIG is now working on an Office Document Language (ODL) DAP which they plan to have completed in 1991.

2. EXECUTIVE SUMMARY

The primary requirement articulated by the majority of speakers at the workshop on Electronic Information Exchange Standards was the need to move toward information management in a database environment rather than in separate documents. In conjunction with this requirement is a second strong requirement to have a migration strategy that considers legacy data. For instance, if changes are made to standards or technology, there must be the ability to read and use the presently stored data.

Major issues presented with regard to the database requirement were:

1. ability to encode the information itself to provide an information service independent of the application/device,
2. on-line display,
3. query and retrieval ability,
4. ability to have electronic deliverables and pageless technical manuals.

The database and integration requirements were reiterated by the speakers. They are the requirements that end users are most interested in. We believe that all of the requirements are important and they are contained in a complete list of requirements presented later in the report.

The Office Systems Engineering Group at NIST has delivered the message that the government needs one standard or a group of standards that work well together in the area of electronic publishing. This message has gone to Technical Committee X3V1, Text and Office Systems, operating under The American National Standards Institute (ANSI); and the ODA SIG of the NIST Implementors Workshop. Work has started on the ODL DAP which is a positive step toward harmonization in this area.

3. PROGRAM REPORT

Opening remarks were by Mr. Allen Hankinson, Chief, NIST Systems and Software Technology Division, and Lawrence A. Welsch, Ph.D. Manager, NIST Office Systems Engineering Group presented an overview of the workshop goals. Mr. Welsch said that the two goals of the workshop were to develop a list of user requirements and to see how ODA and SGML fit together. The next set of speakers were Mr. Bruce Lepisto, Deputy Director, CALS and Mr. Frank Gilbane, President, Publishing Technology Management. They provided a status overview of the CALS Planned Applications for Electronic Information Exchange Standards. This presentation started with a definition of CALS terms including Type A data - Information managed as pages -- hard copy or raster; Type B data - information managed as (parts of) documents -- hard copy or electronic; and Type C data - information managed as a database. They then talked about existing standards and technology and the transition to Type C. The strategy was described as 1) encoding only the information itself and 2) encoding the information independently of devices, applications and inappropriate metaphors such as documents or pages. They also talked about the economics of transition and the care needed for planning. Their summary included the following four points:

1. There is a Goal - Technical Information Publishing
2. There is a Transition Strategy - Identify and Encode Information Objects and Structures
3. MIL-M-28001A is on the Transition Path

4. There is a need to encourage Research and/or development of technology and standards that support the goal of technical information publishing.

The next speaker, Mr. John Chapman, Chief, Electronic Dissemination Task Force, U.S. Government Printing Office, talked about the Government Printing Office perspective on plans for using electronic interchange standards. Mr. Chapman started by giving an overview of the service the GPO provides for federal agencies. He discussed the methods used including publishing from a database. Mr. Chapman also talked about how the GPO might handle SGML documents from agencies. He went on to list requirements for the SGML, DSSSL and SPDL standards. These included table handling, revisions and the speed necessary to accommodate large volumes of work. Mr. Chapman finished his talk by listing standards which need to be developed. These included standards for tagging, indexing and formatting text, tabular, database, and image data on Compact Disk - Read-Only Memory (CD-ROMs). He also talked about standards to reduce the number of different routines that end users must learn and the efforts GPO is making in anticipation of the impact from new standards.

Next we had a group consisting of Mr. Norman Scharpf, President, Graphic Communications Association (GCA), James Mason, Ph.D., Head of Standards Development for the Publications Division, Oak Ridge National Laboratories, U.S. Department of Energy, Jon Cunningham, Ph.D, Foundation for Electronic Publishing, and Mr. Robert Barlow, AGFA Compugraphic. They talked about government publishing's needs for national (ANSI), and international (International Standards Organization) Standards for electronic publishing and for compound document applications. In the government needs area, it was pointed out that the government is doing a great deal of its own publishing. They also listed requirements such as the need for timely access and dissemination of information, simplified control over document creation and management, and congruence between electronic interchange standards - beginning within Government Open Systems Interconnection Profile (GOSIP).

Mr. Robert Terrell, Director, Technology and Telecommunications Office of Scientific and Technical Information (STI), U.S. Department of Energy and Spokesperson for the Commerce, Energy, NASA, Defense, Health and Human Services (CENDI) organization, spoke about the interest in the ODA and SGML electronic information exchange standards within the STI. Mr. Terrell said that we must deal with scientific and technical information, equations and formulas within documents. He also said a bridge is needed between ODA and SGML, that there need to be rules for selection and use of these standards and that there needs to be consideration given to the end user of this technology.

Dr. Mason came back to talk about his experiences in using SGML and about evaluating ODA/SGML Standards. One message Dr. Mason shared was that in building SGML applications in Oak Ridge, building the application is simple, designing the applications and understanding what data needs to be managed and understanding the document structures and their relationships are the real problems. Dr. Mason then discussed planning methods and also some differences between ODA and SGML. He said that the two address different markets. SGML manages information and ODA transmits messages. SGML requires more planning because the life of the SGML document is expected to be longer then that of the ODA document by nature. Dr. Mason talked about ODL and the ability to describe ODA structures in an SGML application. He continued by saying that neither ODA or SGML

defines a complete environment and that additional standards are needed. He then listed standards such as DSSSL, SPDL, and font standards as some of those needed. Dr. Mason concluded by saying that none of these standards belonged in GOSIP; the standards are not communications protocols but they all depend on GOSIP services.

The next was a joint presentation with Mr. Frank Dawson, IBM ASD Office Systems, Mr. Robert Corst, and Mr. Mark Smith, Digital Equipment International. They gave an update on the ODA Implementation Agreement and experiences with implementing ODA. Mr. Dawson talked about the ODA SIG Charter, current implementation agreements, the three DAPs: 1) Raster DAP, 2) Level 2 DAP and 3) Level 3 DAP. Mr. Dawson also talked about the work in development of ODL encodings of DAPs which is now taking place within the OSI Implementors Workshop. Mr. Corst provided an ODA tutorial/history and Mr. Smith discussed developer's requirements for compound document support. Some of the requirements he talked about are as follows:

1. Implementation of support of document interchange formats needs to be backed up with tools that provide developers with document access support.
2. Application developers require interchange format support that interfaces with their applications at the application level. They require interchange format that is application and not interchange format oriented.
3. An application-oriented compound document toolkit that has a standard interface for application developers is required. The toolkit should model the objects developers handle. The toolkit must also be able to use multiple interchange formats and be adaptable to accommodate new compound information technologies.

Dr. Charles Goldfarb, IBM Almaden Research Center, gave a talk entitled "ODA, SGML, DSSSL, and SPDL: Having Your Cake and Eating it Too!" Dr. Goldfarb started by asking the audience how many were present to hear about ODA - about 4 people raised their hands. He then asked how many were present to hear about SGML and the remainder of the 400+ raised their hands. He went on to state that SGML is here today and that ODA looks interesting for tomorrow. Dr. Goldfarb then discussed ODA and SGML and concluded by saying "Start with SGML today. Add architectures as you need them within the same interchange format environment. Build linking and referencing capabilities across architectures because of common document structuring language."

Mr. Avi Bender, Director, Image Engineering, Contel Federal Systems, discussed Information and Image Retrieval Systems. He talked about the potential role of using SGML for the creation of searchable text and image retrieval systems. He said that publication is just one aspect of information processing and then listed compound document storage and retrieval, sharing text and images across networks and the use of CD-ROM in electronic publishing as other aspects. He says there is an industry trend to merge text, data, image, and voice and that the emphasis is on open system architecture. Mr. Bender described full text search and retrieval systems, talked about database creation issues, and discussed how he thought SGML could address database creation issues. Mr. Bender listed four items as future trends: 1) text/image data captured and tagged at the source, 2) proliferation of data communication networks to transmit information, 3) networks used to create, process, store, transmit, display and print information, and 4) information is an important resource:

standards and tools are needed to process electronic information from a more global perspective.

Mr. Mel Lammers, Director, CALS Test Network, gave a presentation on the Air Force experience in the use of digital technical manual data. His first four very strong points were 1) Legacy data must be considered, 2) printed data will continue, 3) the move to digital is mandatory, and 4) there is the desire for pageless vs paged data. Mr. Lammers then talked about implementation worries of standards' stability and new standards. His worries centered on the possibility of data and software obsolescence in this regard. He also talked of management worries regarding the three types of data, database corruption and training needs. In his summary he said that there was the need for stable standards and no obsolete data and that although the implementation of CALS was not easy, there was no choice but to remain committed to that course.

Last but certainly not least, Mr. John Karpovich, Special Assistant for Technology, Navy Publishing and Printing Service, talked about the Navy publishing and printing experience in implementing information exchange standards. Mr. Karpovich began his speech by stating that specifications were not specific and that standards were not standard. He also stated problems with validation of CALS compliant delivery data and hardware processing components. He talked about the use of SGML for database development and wanted SGML and ODA to be machine functions - an expert system - not functions that had to be learned and performed by people.

Mr. Welsch then presented a summary of the message received from the speakers. His summary made the following points:

1. There is a large SGML audience in government today - ODA must address that audience in a much more direct and clear manner then it does today,
2. a migration strategy is needed to move to document databases,
3. ODA needs to be improved - in areas such as equations and tables,
4. ODA and SGML must work together.

4. USER REQUIREMENTS LIST

This list has been developed containing user requirements as articulated by the speakers of the July 30, 1990 workshop on electronic publishing. This list represents the deliverable of meeting minutes and summary report under the Common Government application objective.

The primary requirement articulated by the majority of speakers at the July 30th workshop was the need to move toward information management in a database environment rather than through separate documents. In conjunction with this requirement is a second strong requirement to have a migration strategy that considers legacy data. For instance, if changes are made to standards or technology, there must be the ability to read and use the presently stored data.

Major issues presented with regard to the database requirement were:

1. ability to encode the information itself to provide an information service independent of the application/device,
2. on-line display,
3. query and retrieval ability,
4. ability to have electronic deliverables and pageless technical manuals.

The database and integration requirements were articulated many times by many speakers. They seem to be the requirements that end users are most interested in. We believe that all of these requirements are important even if they were not emphasized as much as the two above.

Some of the other requirements are as follows:

1. ability to handle tabular matter "on the fly,"
2. ability to insert graphics "on the fly,"
3. standards for tagging, indexing, and formatting text, tabular, database, and image data on CD-ROMs,
4. standards to facilitate end-user retrieval, display and printout,
5. standards to reduce the number of different routines that end users must learn,
6. simplified control over document creation and management,
7. ability to manage information in a distributed environment,
8. congruence between electronic interchange standards,
9. hypermedia publication,
10. ability to deal with scientific and technical information such as equations and formulas,
11. a bridge between SGML and ODA,
12. ground rules for selection and use of document processing standards and how they are finally implemented,

13. support is needed for application developers to provide:

    a. low cost entry,

    b. an underlying model which is easy to understand such as on object-oriented design base,

    c. capability of catering for the many interchange formats in use today,

    A core set of application-oriented objects should be supported by the toolkit and serve as a stable base on which an application developer can build the application.

14. stability in standards,

15. ability for insuring correctness while using digital media,

16. common set of standards and goals,

17. ability to do conformance testing, eliminate self certification,

18. ODA must address the large SGML audience in government in a much more direct and clear way then it does now,

19. ODA needs to make improvements in formatting issues such as equations and tables,

20. ODA and SGML must work together.

5. SUMMARY

We have achieved significant results from the actions discussed in this report. First, the workshop of July 30, 1990 helped convince many people that changes in the electronic publishing area were needed. We took the resulting list of user requirements from this workshop and presented them to Task Group 1 of X3V1. This task group has the responsibility for developing user requirements for the standards being worked on by X3V1. We have succeeded in encouraging the participation of SGML experts in the ODA SIG of the OSI Implementors Workshop. This has contributed to the work being done on the ODL DAP in that group. All of this has a significant impact on the harmonization of standards within the electronic publishing area.

6. DEFINITIONS

ODA - Office Document Architecture

The definition below is taken from ISO 8613-1.

ISO 8613 applies to the interchange of documents by means of data communications or the exchange of storage media.

It provides for the interchange of documents for either or both of the following purposes:

- to allow presentation as intended by the originator;
- to allow processing, such as editing and reformatting.

The composition of a document in interchange can take several forms:

- formatted form, allowing presentation of the document;
- processable form, allowing processing of the document;
- formatted processable form, allowing both presentation and processing.

ISO 8613 also provides for the interchange of ODA information structures used for the processing of interchange documents.

Furthermore, ISO 8613 allows for the interchange of documents containing one or more different types of content, such as character text, images, graphics and sound.

SGML - Standard Generalized Markup Language

The following definition is taken from ISO 8879.

This International Standard specifies a language for document representation referred to as the "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.

DSSSL - Document Style Semantics and Specification Language - The following definition is taken from WD 10179.

This International Standard defines the Document Style Semantics and Specification Language (DSSSL) for the specification of document processing, such as formatting and data management functions, with the initial focus on formatting to both print and on display media, and data conversion. This International Standard has been structured to permit future sections to be added to this International Standard to cover the areas of data management.

An objective of the DSSSL Standard is to provide a formal and rigorous means of expressing the full range of document production specifications, including high-quality typography, required by the graphic arts industry. These specifications will be expressed using standardized basic semantics or combinations of the basic standard DSSSL semantics. These semantics will allow users to specify fully the characteristics to be applied during document processing, such as composition, pagination, and imposition. The DSSSL typographic semantics may be used to form the basis of a standard style sheet language.

In addition, DSSSL includes General Language Transformation constructs which provide the capability to translate into an existing processing language, such as a database update language (e.g. SQL) or a traditional text formatting language.

SPDL - Standard Page Description Language

The following definition is taken from DP 10180.

This International Standard defines a language for the specification of electronic documents, comprised of black and white, gray scale, or full colour text, images, and geometric graphics, in a form suitable for presentation (printing or displaying on other suitable media).

This International Standard is intended to be extensible in order to accommodate future developments in imaging technology.

This International Standard is intended to be used in a variety of configurations meeting a variety of connectivity needs. It is specifically compatible with use over OSI networks.

In addition to specifying how document images are represented, this International Standard specifies how additional information called Document Production Instructions affects the document image.

ODA DAPs - Office Document Architecture Document Application Profile

The following definition is taken from the "Working Implementation Agreements for Open Systems Interconnection Protocols" dated September 1990.

Development of this document application profile has been done in liaison with several

organizations. These include ODA expert groups within the:

- Asia-Oceania Workshop(AOW);
- CCITT Study Group VIII;
- European Workshop on Open Systems (EWOS).

The liaison between these organizations has occurred within the meetings of the Profile Alignment Group for ODA (PAGODA). These meetings have focused on the development of a single set of Internationally aligned ODA document application profiles.

ODL - Office Document Language

This definition is taken from ISO 8613-5.

ODL is a language in which the constituents and attributes of the document are identified by descriptive tags, and are grouped into one or more storage entities (e.g. files) as the user may require.

For interchange, each ODL entity is represented as a single data structure or data item, specified using ASN.1 (Abstract Syntax Notation) in a data stream constructed according to the SGML Document Interchange Format defined in ISO 9069.

ODL is specified in annex E (normative) of ISO 8613.

Note - ODL is an SGML application conforming to ISO 8879.

NIST-114A
(REV. 3-90)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NISTIR 4560

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
APRIL 1991

4. TITLE AND SUBTITLE

Government Document Processing Requirements Report

5. AUTHOR(S)

Roger F. Sies

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
GAITHERSBURG, MD 20899

7. CONTRACT/GRANT NUMBER

8. TYPE OF REPORT AND PERIOD COVERED

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

10. SUPPLEMENTARY NOTES

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

This report describes several activities of the Office Systems Engineering Group in the area of electronic publishing standards. It gives an account of the July 30, 1990 workshop on Electronic Information Exchange Standards Used in Document Processing Applications and the lsit of User Requirements that came out of that workshop. The report also talks about other efforts the Office Systems Engineering Group has made to help bring about the harmonization of electronic publishing standards.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

CALS, ODA, ODA-DAP, ODL, DSSSL, SGML, SPDL

13. AVAILABILITY

- [X] UNLIMITED
- [ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).
- [ ] ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.
- [X] ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
15

15. PRICE
A02

ELECTRONIC FORM