

**STANDARD FORMATS:
WHY THEY ARE NEEDED
AND
CONSIDERATIONS IN DEFINING THEM**

C. H. Malarkey
Data Systems Research and Development Division
Y-12 Plant
Lockheed Martin Energy Systems

May 6-7, 1998

Preprint for submission to:
Inforum '98
Oak Ridge, Tennessee
May 6-7, 1998

Prepared in conjunction with the
**NUCLEAR WEAPONS INFORMATION GROUP
TOOLS WORKING GROUP**

Comprised of:

Bob Baskerville, AlliedSignal Federal Manufacturing & Technologies (AS/FM&T/KC)
Dave Dickison, Defense Special Weapons Agency (DSWA)
Anne Macek, Los Alamos National Laboratory (LANL)
Dave Anderson, Lawrence Livermore National Laboratory (LLNL)
Mack Woolard, Nevada Test Site (NTS)
Bob Donohue, Office of Scientific and Technical Information (OSTI)
Stefanie Elsea, PANTEX Plant
Keith Johnstone, Sandia National Laboratory (SNL)
Connie Malarkey, Lockheed Martin Energy Systems (LMES)

STANDARD FORMATS: WHY THEY ARE NEEDED AND CONSIDERATIONS IN DEFINING THEM

1. INTRODUCTION

The quality and long-term usefulness of the data in our electronic repositories depends on decisions about storage formats, compression algorithms, and resolutions of the electronic images when the data are created and prepared for archive. The Nuclear Weapons Information Group (NWIG) recognized the need to identify standard formats for the exchange of digitized data between sites within the Nuclear Weapons community and for the capture and long-term storage of data for electronic repositories. This presentation concentrates on the business reasons for identifying standard formats and the approach the NWIG working group used to determine which formats were appropriate. In addition, considerations for keeping the electronic repositories up to date as technology changes are presented.

2. NUCLEAR WEAPONS INFORMATION GROUP

The Nuclear Weapons Information Group (NWIG) is a consortium of representatives from each of the Nuclear Weapons Complex (NWC) sites in the United States, the Atomic Weapons Establishment in the United Kingdom, and DoD agencies such as the Defense Special Weapons Agency. They are concerned with archiving and preserving nuclear weapons data, information and knowledge. NWIG is chartered to define, develop and promote requirements, tools, techniques, and formats to enable continuing and appropriate access to such data, information, and knowledge, so that it may enable the NWC to:

- preserve the foundation for science-based stockpile stewardship;
- preserve information critical to stockpile maintenance and dismantlement;
- preserve data and information for training and qualifying future scientists, engineers, and technicians;
- support proliferation analysis;
- preserve fundamental information for reconstitution of nuclear weapons research, development, production, and testing in the future, if necessary; and
- provide immediate critical information for emergency response.

NWIG activities involve metadata preparation, defining requirements, obtaining and/or developing required software tools for managing the metadata records, and developing change management procedures for use by the participating NWC organizations. Working Groups are defined as needed.

The stated charter of NWIG's Tools Working Group (TWG) was to suggest, coordinate, and promote tool development projects among NWIG members with the goal of realizing synergistic economies of scale through leveraging individual member core competencies for the benefit of the NWIG community as a whole. In pursuit of this goal, the working group focused on recommending and promoting tools and protocols which would lead to site information architectures capable of supporting the capture and exchange of electronic nuclear weapons data.

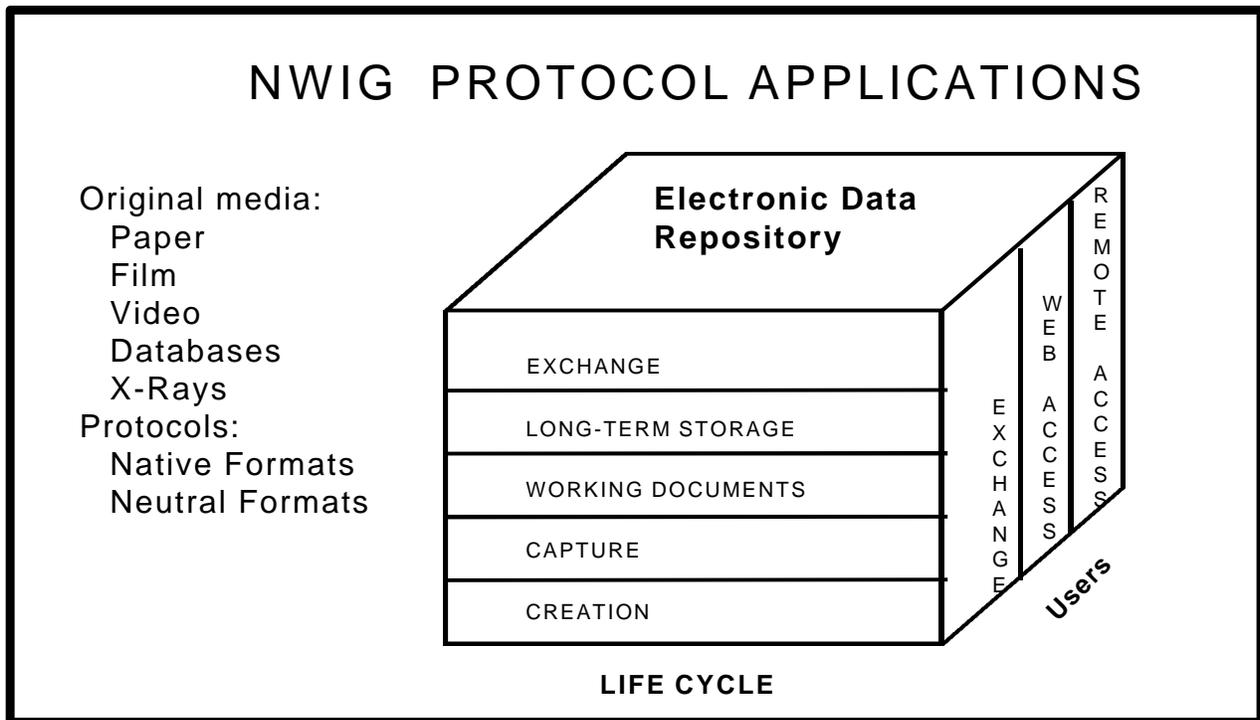
A significant part of the TWG function was the identification of standard formats. Standard formats for the capture and long-term storage of nuclear weapon data and ones to be used during the exchange of archival data were identified and published. These standards are accepted by the NWIG participants and will be communicated throughout DP for use in nuclear weapons archiving programs.

3. STANDARD FORMAT IDENTIFICATION AND RECOMMENDATIONS

At the general NWIG summer meeting in 1996, the TWG was commissioned to define a set of standard formats for information capture and exchange of archived nuclear weapons data. This was viewed as a critical and potentially controversial task. Initially, the group focused on formats for the exchange of archival information and then addressed the identification of recommendations for the capture and long-term storage protocols.

Figure 1 provides a conceptual view of the electronic data repository environment for which the standard formats are identified. Since the primary focus is on archive data, the life-cycle phases of document creation and their utilization as working documents were considered outside the charter for the initial efforts at standards identification. Most of the source media is assumed to be paper, film, videos, archived databases, and radiographs. Although the initial standards focus did not look at software generating documents in native formats, it is understood that working documents created and managed in native formats must be able to be converted to a storage format for long-term archive purposes.

Additionally, decisions about the data repositories require consideration of the projected use of the files within the repositories. With the advent of secure classified communication the file may be exchanged with another site through a file push or pull scenario, it may be accessed using web compatible tools over an Internet or an intranet secure http server, and it may be made available to direct access by log on from a remote user.



Decisions made about formats and resolutions when the archives are designed and the files are created determine the quality of the preserved data. Thus, the capture and long-term storage recommendations are aimed at ensuring long-term quality of the archival data by recommendations which assist in the conversion from storage formats to exchange standards. When determining which format to use, priority should always be given to those standards that are recognized open international standards, since these formats are the most stable over time. In order to enhance a sites ability to make decisions based on its

business requirements, no prioritization of formats is identified for either the exchange formats or for the capture and long-term storage formats.

The intent of having electronic data repository files managed using a minimum number of formats is to facilitate the exchange of archived nuclear weapons information between sites. These standards are to provide a limited number of formats to be supported thereby minimizing future migration and conversion requirements associated with long-term archival.

During group discussions, a working definition of an information standard was agreed upon and used as a measure of the recommended formats. *An Information Standard is an agreed upon specification that enables the interchange of information.*

3.1 NWIG Standard Exchange Formats

The standard formats identified for the exchange of electronic data are being adopted and promoted by Defense Programs for use in all programs concerned with the archiving of nuclear weapons data. The data and information exchange formats are enumerated in Table 1 below. In December 1997, the formats for exchange were identified as a configuration item and placed into the NWIG configuration control procedure.

OBJECT TYPE	FORMAT
Documents	SGML HTML ASCII PDF TIFF (Group 4)
Images: Photos Diagrams Strip Charts	TIFF (Group 4) GIF JPEG
Engineering Drawings	IGES HPGL STEP TIFF
3-D and Solid Models	VRML
Numeric/tabular	ASCII formatted pairs, which may be: comma delimited, space delimited, or columnar fields
Numeric Simulation and Observed Data	HDF

OBJECT TYPE	FORMAT
Video	MPEG 1 MPEG 2
e-mail attachments	MIME

Table 1. NWIG Exchange Standard Formats as of Sept. 29, 1997.

Exchange standards identified for the Data Archival and Retrieval Enhancement (DARE) program sponsored by the Defense Special Weapons Agency (DSWA) and by the Office of Scientific and Technical Information (OSTI) were used as a starting point for discussion. The NWIG formats are consistent with the exchange standards used by these organizations as well as being compatible with standards in use at the Atomic Weapons Establishment (AWE). At the time when the group initiated the discussion of standards, the National Archives and Records Administration (NARA) had not defined a standard for long-term archival except ASCII.

Encryption standards are being addressed by other groups within the nuclear weapon complex and are not included.

Primarily, these formats were defined for the exchange of electronic data that is not going to be processed at the receiving site. The exchange of documents in a native format is appropriate when two sites agree on the exchange format. Table 1 represents standard formats for exchange only. Storage formats used for long-term archival should be easily converted without data loss for the purposes of exchange.

The intent of identifying exchange formats is to facilitate the exchange of archived nuclear weapons information between sites. These standards are to provide a limited number of formats that must be supported thereby minimizing future migration and conversion requirements associated with long-term archival.

3.2 Capture and Long-term Storage Recommendations

While focusing on the recommendations for the capture and storage of archive data, the group was concerned with the long-term archival and future ability to use the electronic data. File format and resolution must be determined by projected use, not by object type. Decisions associated with capture formats should balance the time and cost of converting hard copy into electronic formats against the accuracy and detail requirements for the digital image. The capture resolutions are considered minimum recommendations. Expected business uses for the archived objects may require the use of enhanced resolutions. Conversion to recommended formats or resolutions of legacy data repositories is not expected, but as systems are replaced or upgraded all new repositories should be implemented according to the recommendations.

OBJECT TYPE	FORMAT
Documents, Text	TIFF (group 4), 300 dpi ASCII (includes SGML and HTML) PDF

OBJECT TYPE	FORMAT
Photographs	Gray Scale: JPEG or TIFF (default of 8 bit, 150 dpi or better) Color: JPEG or TIFF (default of 24 bit, 150 dpi or better)
X-Ray and Radiographs	Deferred to watch list
Engineering and CAD Drawings	TIFF (group 4), 200 dpi or better STEP
3-D and Solid Models	STEP
Diagrams	GIF TIFF
Data	Numeric/tabular data: Any ASCII formatted data, which may be: Comma delimited, Space delimited, or Well defined columnar fields Numeric Simulation and Observed Data: HDF
Video	Capture: Beta Cam SP Electronic Storage: MPEG I or II

Table 2. Capture and Long-term Storage Recommendations as of Sept. 29, 1997.

The Capture and Long-term Storage recommendations in Table 2 were submitted to the NWIG configuration control process after review by the NWIG POC and the TWG members.

For documents comprised of mixed media, any combination of recommended formats that best retains the original flavor of the document and meets end user requirements is appropriate.

4. EVOLUTION OF STANDARDS

Standard formats and protocols evolve over time as technology changes. The advantage of focusing on standard formats is that technology changes provide evolutionary paths to the newest protocols. In order to manage the electronic environment and to keep the data repository viable for the five to twenty-five to seventy-five year time frame required for nuclear weapons data preservation, there must be defined transition plans which address changes in technology and the evolution of standard formats. Sites are responsible for transition plans for their electronic archives. Site plans should include monitoring

software and technology changes that would drive requirements for conversion to an upgraded format or to a different storage media.

In order to assist NWIG in the evolution of the standard format recommendations, the group identified watch lists. Formats on watch lists are projected to become globally used and accepted. The NWIG Exchange Working Group (EWG) or its successor is responsible for an annual review of standard recommendations usually at its first meeting in each calendar year. Watch lists are not under configuration control and can be updated and changed at any point when a consensus of members agree that an emerging format should be considered for applicability to Defense Programs archiving effort and to NWIG's mission.

5. COMMUNICATION OF STANDARDS

Both the exchange formats and the capture and long-term storage recommendations are managed through NWIG's configuration control process. The configuration items are the identified formats and recommendations only. The watch lists are NOT part of the configured item, but are available for information and communication purposes.

Once the standards are placed under configuration control, the NWIG Points of Contact (POC) are expected to ensure publication and promotion within their local site. Recommended ways of communication are through an internal intranet and other types of distribution mechanisms. Additional communication can be accomplished through e-mail. For example, the exchange standards have been distributed by e-mail within the DOE DP community. Although the protocols are available on the TWG Web-Site, this site is username and password protected and is not available to the entire community at each NWIG site. Thus, a broader method of communication is required.

6. FORMAT DESCRIPTIONS

The following is a list of the recommended formats for both the exchange and capture and long-term storage of information among government agencies, laboratories, and contractors. Examples of information associated with each format type and a definition of each recommended format is provided.

FORMAT	APPLICATION OBJECT
ASCII	text files
DXF	CAD files
GIF	Images - black and white and color
HDF	Numerical and Spacial Data
HPGL	Engineering Drawings
HTML	formatted information
IGES	Engineering Drawings

FORMAT	APPLICATION OBJECT
JPEG	Photos or images - Black and white AND color
MIME	e-mail
MPEG	video storage
PDF	formatted information and images
SGML	formatted information
STEP	Engineering Drawings and models
TIFF	images - black and white
VRML	3-D and solid models

Table 3. Recommended formats as of Sept. 28, 1997.

ASCII: (American Standard Code for Information Interchange)

The predominant character set encoding of present-day computers. Files containing these codes are commonly referred to as "text files." (Obsolete standard - ANSI X3.4; superseded by ISO 646 and currently by ISO 10646 - known as UNICODE)

DXF: (Data eXchange Format)

AutoCAD's external format. It has become an industry de facto standard for exchanging CAD data between smaller desktop systems.

GIF: (Graphics Interchange Format)

CompuServe's standard for defining generalized color raster images. This Format allows high-quality, high-resolution graphics to be displayed on a variety of graphics hardware. It is the predominant graphics image format on the World Wide Webb.

HDF: (Hierarchical Data Format)

A library and multi-object file format for the transfer of graphical and numerical data between machines. HDF is versatile in that it supports several different data models; it is self-describing, allowing an application to interpret the structure and contents of a file without any outside information; it is flexible, allowing mixing and matching of related objects together in one file for access as a group or as individual objects. The HDF interchange format was developed by National Center for Supercomputing Applications (NCSA) at the University of Illinois.

HPGL: (Hewlett-Packard Graphics Language)

A vector graphics file format from HP that was developed as a standard plotter language. It has since emerged as a popular vector graphics output format. It is the official format specified for the exchange of NWC standard plot files according to the U.S. Department of Energy Albuquerque Operations Office Development and Production Manual.

HTML: (Hypertext Markup Language)

HTML is a collection of platform-independent styles (indicated by markup tags) that define the various components of a World Wide Web document. HTML documents are plain-text documents that use English language editing markup elements to define specific formatting for individual sections or the complete document. HTML can also link text and/or an image to another document or document section. HTML is a subset of SGML.

IGES: (Initial Graphic Exchange Specification)

IGES specifies file structure and syntactical definition, and defines the representation of geometric, topological, and non-geometric product definition data. The standard permits the compatible exchange of product definition data used by various computer-aided design and computer-aided manufacturing (CAD/CAM) systems.

JPEG: (Joint Photographic Experts Group)

JPEG compression economizes on the way data is stored and also identifies and discards extra data, that is, information beyond what the human eye can see. Because it discards data, the JPEG algorithm is referred to as "lossy." This means that once an image has been compressed and then decompressed, it will not be identical to the original image. In most cases, the difference between the original and compressed version of the image is indistinguishable. You do not need to decompress images saved in the JPEG format. They are automatically decompressed when they are opened. (ISO/IEC 10918-1)

MIME: (Multipurpose Internet Mail Extensions)

MIME is an Internet standard (RFC 1521 and 1522) that helps transfer different files across e-mail. It provides a common method to transfer data and then tells the computer how to deal with it. MIME is designed to include multiple objects in a single message; to represent body text in character sets other than ASCII; to represent formatted multi-font text messages; to represent non-textual material such as images and audio fragments; and generally, to facilitate later extensions defining new types of Internet mail for use by cooperating mail agents.

MPEG: (Moving Pictures Experts Group)

An ISO/CCITT standard for compressing full-motion video. MPEG is supported by hardware and software decompression. Software decompression is limited, requiring systems to decompress the moving images through the player software. Hardware decompression is capable of displaying a full-color, 30 fps video image with CD-quality stereo sound. (Low bandwidth - ISO/IEC 11173, and high bandwidth - ISO/IEC 13818).

PDF: (Portable Document Format)

A unique cross-platform PostScript-based file format developed by Adobe. A PDF file can describe documents containing any combination of text, graphics, and images in a device and resolution independent format. These documents can be one page or thousands of pages, very simple or extremely complex, and make rich use of fonts, graphics, color, and images. PDF is an Adobe proprietary format.

SGML: (Standard Generalized Markup Language)

A system for organizing and tagging elements of a document. SGML was developed and standardized by the International Organization for Standards (ISO) in 1986. SGML itself does not specify any particular formatting; rather, it specifies the rules for tagging elements. These tags can then be interpreted to format elements in different ways. SGML is used widely to manage large documents that are subject to frequent revisions and need to be printed in different formats. (ISO 8879:1986)

STEP: (Standard for the Exchange of Product Data)

An international standard for product data modeling and data exchange. It is used to describe engineered products in a standard, vendor independent manner. The standard defines all aspects of describing technical diagrams and documents in a neutral format for transmission over communication networks and for processing by numerically controlled machining and assembly tools. (ISO 10103)

TIFF and TIFFg4: (Tagged-Image file Format)

A standard file format for storing images as bit maps. TIFF is used mainly for exchanging documents between different software applications and different computer platforms. TIFF stores width, length, and resolution information in the file-header for use by a computer for image display. TIFF can differentiate between black and white, gray scaled, and colored images. TIFF (group 4) is preferred for black and white images.

VRML: (Virtual Reality Modeling Language)

VRML is a rich text language for the description of 3-D interactive virtual worlds. It supports the building of complex, realistic 3-D environments, complete with shiny materials, textured surfaces, multiple light sources, animations, and sound tracks. VRML worlds can sense a viewer's touch, position, and gaze direction, trigger sounds and animations on viewer proximity, fly the viewer on a guided tour of the world, and even communicate with other applications and users on the Internet. All the major CAD vendors (Pro/Engineer, Unigraphics, Intergraph, etc.) have VRML translators available.

BIOGRAPHY

Connie Malarkey is the head of the Electronic Document Management System Section in the Applied Computing Technology Department at the Y-12 Plant. Connie has 20 years of experience in all phases of software design and development. During the last two years she was a member of the Y-12 team which prepared the Weapons Record Archive and Preservation (WRAP) Plan. Another of her major activities was as one of the Y-12 representatives on the Nuclear Weapons Information Group (NWIG). In addition, she acted as the chair of NWIG's subgroup to identify tools and technologies for use in the exchange of nuclear data within the nuclear weapons community. Connie has a bachelor's degree in Computer Sciences and a master's degree in Industrial Engineering from the University of Tennessee.