Preservation of Nuclear Information and Records

- Main components of knowledge preservation
- Digital preservation (management issues)
- Review of Main IAEA Knowledge Preservation Projects
Goals of Preservation

• Select the most valuable information to convey to the future
• Ensure that it remains readable, accessible and understandable
• Manage technological change so that those objectives are met

Type of Information

• Text (book, journal article, brochure, listing …)
• Image (photo, film, picture …)
• Sound
• Data (numerical, formulas, graph …)
• Interactive (rule-based, training, database …)
• Multimedia
• Computer code
• Sample (physical object)
• Tacit knowledge
• …
Main Components of Knowledge Preservation

- Select
  - Capture
    - Describe/classify
    - Store
    - Provide access
- Maintain (longevity)

Selection of Information for Preservation

- Why Select?
  - Storage is not equal to Preservation
  - High costs and limited budget
  - Maintenance mortgage
  - Legal issues
- Evaluation
- Prioritization by Value, Use and Risk
Copyright Issues

- Copyright protects the actual expression of an idea, not the idea itself
- The absence of copyright notice does not mean absence of copyright protection
- Possession or ownership of physical item does not mean the possessor or owner owns the copyright
- Copyright does not apply to all works, and it does not last forever

Information Capture

- Purchasing
- Copy (the same media or different), digitize
- Interview (tacit knowledge)
Describe and Classify Information

Create metadata

• Metadata is structured data about data

• Metadata is a summary of information about the form and content of resource to facilitate identification and retrieval

Type of Metadata

• Administrative
• Descriptive
• Structural
• Semantic
Administrative Metadata

- Management information needed to maintain, retrieve and display an object
- Rights and permissions
- File format, size compression, etc.
- Hardware, software
- Physical location
- Etc.

Descriptive Metadata

- Information that provides access to the subject of an object
- Author or Creator
- Title
- Subject terms
- Classification
Structural Metadata

- Information used to display and navigate an object
- Structural divisions of an object
- Sub-object relationships (internal links)

Semantic Metadata

- Subject
- Descriptors (controlled, multilingual)
- Semantic links
- Information audience
- Related sources of information
Store

- Environment
- Media
- Format
  - Text
  - Image
  - Text + Image

PDF (text+image, hypertext, sound, video, metadata)
XML

Provide access

- On-line
  - Web
  - Z39.50
- Off-line
  - CD, DVD
- Full-text and/or Metadata
- Portability
- Multilingual Interface
Maintain. Ensure longevity.

- Control
- Refreshing (media)
- Migration (format)
- Emulation (application software)

Type of Media

- Paper
- Film, photo materials
- Gramophone record/plate
- Magnetic tape
- Diskette, CD/DVD …
- Hard disk, flash memory
- Magneto-Optical
- Glass, metal … (holography)
- Etc.
INIS records management 1970 to present

- 1970: first generation of the Bibliographic Database (paper based INIS Atomindex)
- 1978: available on-line
- 1991: available on CD-ROM
- 1996: available on Internet
- 1997: migration from magnetic tape to CD-ROM
  - migration from EBCIDC to ASCII
  - transition from microfiche to digital images
- 2002: migration of archive from microfiche to digital images, OCR
- 2003: migration from tag-text format to XML
  - transition from TIFF image format to image+text PDF

Preservation. Analog versus Digital

### Analog

- ‘Simple’ climate - controlled environment
- Long life
- No special equipment needed
- Simple maintenance technology
- Readability even after partial damage

### Digital

- Space
- Metadata Search only
- Manual maintenance
- Not easy access
## Preservation. Analog versus Digital

### Digital

- Easy access and search
- Content and semantic search
- Automated maintenance
- Easy duplication and distribution
- Multilinguality

### Analog

- High risk of damage
- Short life
- Special equipment and software needed
- Too many different formats
- Dependency on digital technology
- Non-stop maintenance
- Legal constrains

### Preservation of information

- Volume of information published in digital form is growing up dramatically (x2 every 3 years)
- Young generation preference is digital information
- New possibilities: Electronic document analysis, translation and data mining
High Density Analog Storage Devices (extreme longevity)

- Developed by Los Alamos Laboratories and Norsam Technologies
- Analog images on a 3” nickel disk or on a 3” square plate at densities of up to 350,000 pages per disk
Analog versus Digital

• ~65% digital preservation projects failed

Part 2

Digital Preservation
Digital Preservation

• **Organizational Infrastructure**: consistent, systematic management; comprehensive policy framework; co-operation

• **Technological Infrastructure**: technology anticipates needs; open architecture; well defined standards

• **Resources**: sustainable funding

Two main standards

**OAIS** – Reference Model for an Open Archival Information System

**TDR** - Trusted Digital Repositories: Attributes and Responsibilities
Open Archival Information System (OAIS)

- **Was initiated by NASA in June 1995**
- **To define an archive reference model and service categories for the intermediate and indefinite long term storage of digital data obtained from, or used in conjunction with, space missions.**
- **To provide a framework and common terminology that may be used by Government and Commercial sectors in the request and provision of archive services. This will also encourage commercial support for the provision of archive services which would truly preserve our valuable data, not only for space related data but also for all long term data archives**
- **Became an ISO standard in June 1999**

### OAIS Functional Entities

- **SIP** = Submission Information Package
- **AIP** = Archival Information Package
- **DIP** = Dissemination Information Package
- **DI** = Descriptive Information

1. **Ingest**
2. **Archival Storage**
3. **Data Management**
4. **Access**
5. **Requests**
6. **Administration**
Trusted Digital Repositories

- March 2000 – start: to establish attributes of a digital repository for research organizations, building on international standard of the Reference Model for an Open Archival Information System (OAIS)

- A trusted digital repository is more than just organization responsible for storing and managing digital files.

  A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.

TDR: Attributes

- Compliance with the Reference Model for an Open Archival Information System (OAIS)
- Administrative responsibility (standards for physical environment, backup and recovery procedures, and security system …)
- Organizational viability (commitment to the long-term retention, management of, and access to digital assets on behalf of depositors and users)
- Financial sustainability
- Technological and procedural suitability (preservation strategies; h/w, s/w, storage, access; comply with all relevant standards and best practices)
- System security (should be designed to assure the security of the digital assets; authentication systems, firewalls, backup system; policies and plans for disaster preparedness; data integrity)
Issues to Consider

- Clear mandate?
- Defined scope?
- Policy framework, procedures, standards?
- Multi-year plan?
- Relationship between various stakeholders within your organisation?
- Terms and conditions for access and use?
- Preservation planning?
- Appropriate technology?
- Designated, sustained resources?

Principles of Responsibility

- Everyone doesn’t have to do everything
- Everything doesn’t have to be done at once
- Someone must be willing to take a lead on almost all steps
- Small steps are usually better than no steps
- Preservation should not be postponed until a perfect solution appears.

Collin Webb
“Digital Preservation – A Many Layered Thing”
Part 3

Review of Main IAEA Knowledge Preservation Projects

Main Preservation Activities

• INIS NCL Production
• Digitization of INIS NCL Microfiche
• Digitization of older IAEA and Member States Information
• Preserving Web-based information resources (evaluation project initiated)
INIS NCL (Non-Conventional Literature) full-text collection

- Contains knowledge about peaceful nuclear sciences & technologies (collected by Member States for over 30 years)

- Contains over 600,000 documents (many of them can’t be found anywhere else!)

History

1970 – 1996 Microfiche Technology

- NCL Hard Copy
- Photo Imaging
- Microfiche

1997 Electronic Technology

- NCL Hard Copy
- Scanning
- OCR - 2002
- NCL on CD-ROM

INIS Bibliographic Data
INIS NCL Collection

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total NCL documents</td>
<td>791,642</td>
</tr>
<tr>
<td>NCL available from INIS</td>
<td>614,971</td>
</tr>
<tr>
<td>NCL in electronic form</td>
<td>183,298</td>
</tr>
<tr>
<td>Pages (electronic)</td>
<td>&gt; 4,000,000</td>
</tr>
<tr>
<td>Total NCL pages</td>
<td>~25,000,000</td>
</tr>
</tbody>
</table>
INIS NCL Collection

- 63 languages
  - Western languages - 83%
    - English - 70%
  - Cyrillic and Slavic – 12%
    - Russian – 10%
  - Asian languages – 4.5%
    - Japanese – 3.5 %
  - Arabic – 0.4 %

INIS NCL Microfiche Archive Digitizing Project

- 2002 – 12,000 documents
- 2003 – 13,000 documents
- 2003 – 45,000 documents

Total NCL in electronic form: 183,298

Format
- PDF (image + hidden text)
### Digitization of Older IAEA and Member State Documents

- **Documents and Records of the IAEA Board of Governors (~5,000; 1957 – 1996)**
- **IAEA Technical Documents (~1,500)**
- **IAEA Fast Reactors Initiative (~400 docs + …). → Knowledge Package**
- **IAEA Technical Reports Series (~3,000)**
- **Nuclear Data Reports (~3,500)**

### Digitization of Older IAEA and Member State Documents

- **IAEA Nuclear Safety Series**
- **IAEA Bulletin**
- **IAEA Conference Proceedings**
- **Legal Documents (~1,000)**
- **French CEA-R Collection (~4,500; 1946-1970)**
Next Step

Nuclear Knowledge Packages

Small steps are usually better than no steps!

Thanks for your attention!
Supplementary

Digital Preservation. Technical Primer

*Digitization is not preservation*

Main steps in digitization process

*(Digitizing Workflow)*

- Document benchmarking
- Scanning
- Quality Control
- Image Enhancement
- OCR
- Output Formats and Compression
- Archiving & Longevity
Document benchmarking

- the first and very important step in digitizing. The results of document benchmarking effect further steps very much (scanning, enhancement, format, etc.)
- The purpose of document benchmarking is to define/clarify the following:
  - Can the informational content of a document be adequately captured in digital form?
  - Do the physical formats and condition of material correspond to digitizing requirements?
  - Document type
  - Resolution
  - Bit-Depth for colour and grayscale, and threshold for bitonal
  - Output file format and compression

Document Types

- **Printed Text/Simple Line Art**—distinct edge-based representation, with no tonal variation, such as a book containing text and simple line graphics
- **Manuscripts**—soft, edge-based representations that are produced by hand or machine, but do not exhibit the distinct edges typical of machine processes, such as a letter or line drawing
- **Halftones**—reproduction of graphic or photographic materials represented by a grid of variably sized, regularly spaced pattern of dots or lines, often placed at an angle. Includes some graphic art as well, e.g., engravings
- **Continuous Tone**—items such as photographs, watercolors, and some finely inscribed line art that exhibit smoothly or subtly varying tones
- **Mixed**—documents containing two or more of the categories listed above, such as illustrated books
Document Types

Resolution

Zoom
Resolution

- is determined by the number of pixels used to represent the image, expressed in dots per inch or as pixel dimensions.
- Increasing resolution enables the capture of finer detail. *At some point, however, added resolution will not result in an appreciable gain in image quality, only larger file size.* The key is to determine the resolution necessary to capture all significant detail present in the source document.
- **Main approach to imaging: “No More, No Less”**
Resolution

• Electronic Access and Display
  • Screen resolution (800x600; 1024x768)
  • 50 – 150 dpi

• Reproduction/Printing
  • 300-400dpi (8-bits for greyscale and 16/24-bits for colour)

• Preservation
  • 400 dpi for text
  • 600 dpi for photographs
Colour Systems

- Save colour images as RGB files
- Avoid CMYK for master image files!

Colour/Greyscale/Bitonal

- Bit Depth – number of bits of data representing each pixel (dot) of image
- Number of tones for colour and greyscale images = \(2^{\text{Bit Depth}}\)
  
  - 1 bit – black & white (bitonal) = \(2^1\)
  - 2 bits – 4 tones = \(2^2\)
  - 4 bits – 16 tones = \(2^4\)
  - 8 bits – 256 tones = \(2^8\)
  - 16 bits – 65,536 tones = \(2^{16}\)
Colour/Greyscale/Bitonal

**Bit Depth:** When a 24-bit image (left) is reduced to an 8-bit one (right), the color reduction may result in quantization artifacts.

Bitonal/Greyscale/Colour

**Bit Depth:** Left to right - 1-bit bitonal, 8-bits grayscale, and 16-bits color images.
Size of file with scanned image

- file size (in bytes) = 
  - \( H \times W \times (\text{Bit depth}) \times (\text{dpi})^2 / 8 \)

- \( H \) – height of image (in inch)
- \( W \) – width of image (in inch)
- \( \text{dpi} \) – resolution (dots per inch)

Examples:
- 1. A4, 300 dpi, Bitonal
  - file size (in bytes) = 
  - \( 8.5 \times 11 \times (1) \times (300)^2 / 8 = 1,05 \text{ MB} \) (uncompressed)

- 2. A4, 300 dpi, 256 tones
  - file size (in bytes) = 
  - \( 8.5 \times 11 \times (8) \times (300)^2 / 8 = 8.4 \text{ MB} \) (uncompressed)
Image Enhancement

- Deskewing (100% INIS documents)
- Despeckling
- Black border removing
- ...

OCR

- **OCR (Optical Character Recognition)**
  - No longer based on optical processing
  - OCR s/w algorithms process Raster bit maps
- **ICR (Intelligent Character Recognition)**
  - Became synonymous with OCR
- **3D OCR**
  - Uses greyscale/colour information to improve character recognition of low resolution images (50-150 dpi)
Required OCR Accuracy

• For full text searching
  • Above 75%

• For republishing documents
  • Above 99.9% (5 errors per 5000 characters)

Output Formats and Data Compression

• To ensure necessary level of quality
• To save space and time

• Lossless technology
  • File reconstruction is identical to original image

• Lossy technology
  • A certain amount of original information discarded during imaging (compression) process
Output Formats and Data Compression

- **TIFF** - *Tag Image File Format*
  - Most common standard for archiving
  - TIFF – G4 (group 4 fax compression – lossless) – for black & white images
- **PDF** - *Portable Document Format*
  - Most common standard for electronic publishing
- **JPEG** – *Joint Photographic Experts Group*
  - For colour images (allows lossless option)
- **JPEG2000**
  - New wavelet technology
- Many others

PDF - *Portable Document Format*

- **PDF/A**
- **PDF/X**
- **XML**

- Text/Hypertext
- Image
- Image + Text
Output Formats and Data Compression

100 dpi low JPEG compression
File size: 248K

100 dpi medium JPEG compression
File size: 49K

100 dpi high JPEG compression
File size: 22K