



*The Abdus Salam*  
**International Centre for Theoretical Physics**



**SMR.1769-11**


## ***SCHOOL OF NUCLEAR KNOWLEDGE MANAGEMENT***

*18-22 September 2006*

### **Preservation of Nuclear Information and Records**

**A. TOLSTENKOV**  
**International Atomic Energy Agency**  
**Knowledge Preservation Group - INIS & NKMS**  
**Department of Nuclear Energy**  
**A-1400 Vienna**  
**AUSTRIA**

# Preservation of nuclear information and records



International Atomic Energy Agency


## Preservation of nuclear information and records

*Anatoli Tolstenkov*

School of Nuclear Knowledge Management  
Trieste, Italy, 18-22 September 2006


## Preservation of Nuclear Information and Records

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

School of Nuclear Knowledge Management 22 – 28 September 2006, Trieste, Italy    9/8/2006 Preservation of nuclear information    2    International Atomic Energy Agency 

## 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

School of Nuclear Knowledge Management 22 – 28 September 2006, Trieste, Italy    9/8/2006 Preservation of nuclear information    3    International Atomic Energy Agency 

## Example

### IAEA Board of Governors documents

- 1985 – 1995 BGOV Database on Mainframe
- 1995 – DB was lost during migration from IBM Mainframe to Client/Server Architecture
- 2002 - 2003 – Creation of BGOV Database from scratch

School of Nuclear Knowledge Management 22 – 28 September 2006, Trieste, Italy    9/8/2006 Preservation of nuclear information    4    International Atomic Energy Agency 

# Preservation of nuclear information and records

## Example

### IAEA Board of Governors documents

Records were prepared and stored:

- WANG Word Processor (8" Diskette) - Not Readable
- IBM DisplayWrite and WordStar (5" Diskette) – Not Readable, Not Understandable
- IBM Stairs (Magnetic Tape) – Readable but Not Understandable



## 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



# Preservation of nuclear information and records

## 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

School of Nuclear Knowledge Management  
22 - 28 September 2006, Trieste, Italy

9/8/2006 Preservation  
of nuclear information

9

International Atomic Energy Agency



## Information Capture

- Purchasing
- Copy (the same media or different), digitize
- Web information harvesting
- Interview (tacit knowledge)

School of Nuclear Knowledge Management  
22 - 28 September 2006, Trieste, Italy

9/8/2006 Preservation  
of nuclear information

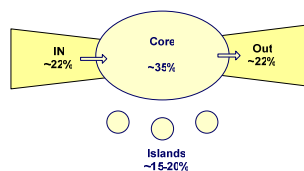
10

International Atomic Energy Agency



## Web information harvesting

### Web topology (Bow Tie)



School of Nuclear Knowledge Management  
22 - 28 September 2006, Trieste, Italy

9/8/2006 Preservation  
of nuclear information

11

International Atomic Energy Agency



## Web Search Services

- Google ([www.google.com](http://www.google.com))
- Yahoo ([www.yahoo.com](http://www.yahoo.com))
- All the Web ([www.alltheweb.com](http://www.alltheweb.com))
- Ask ([www.ask.com](http://www.ask.com))
- Altavista ([www.altavista.com](http://www.altavista.com))
- Licos ([www.likos.com](http://www.likos.com))
- ...
- National Web Search Services

School of Nuclear Knowledge Management  
22 - 28 September 2006, Trieste, Italy

9/8/2006 Preservation  
of nuclear information

12

International Atomic Energy Agency



# Preservation of nuclear information and records

## Deep/Hidden/Invisible Web

[www.brightplanet.com](http://www.brightplanet.com)

- On-line Databases
- Protected Sites
- Grey Web Sites (Dynamic Content Management System)
- “Non-legal” Information Sites
- ...



## Deep Web

is made up of hundreds of thousands of publicly accessible databases and is approximately 500 times then the surface Web, and composed of very high quality information



## Deep/Hidden/Invisible Web

- Dialog [www.dialog.com](http://www.dialog.com) (>900 databases)
- LexisNexis [www.lexisnexis.com](http://www.lexisnexis.com) (>35K Information Sources)
- Nuclear Explosions Database  
[www.ga.gov.au/oracle/nucexp\\_query.html](http://www.ga.gov.au/oracle/nucexp_query.html)
- MedLine
- INIS
- ...



## Deep/Hidden/Invisible Web

How to search

- BigHub [www.bighub.com](http://www.bighub.com)
- Invisible Web [www.invisible-web.net](http://www.invisible-web.net)
- CompletePlanet [www.completeplanet.com](http://www.completeplanet.com)
- Infomine Multiple Database Search  
[infomine.ucr.edu](http://infomine.ucr.edu)
- [www.10kwizard.com](http://www.10kwizard.com)
- ...



# Preservation of nuclear information and records

## Deep/Hidden/Invisible Web

### Solution:

- Semantic Web
- (Semantic) Web Portal



## Describe and Classify Information

### Special Tools/Methods

- Taxonomy
- Thesaurus



## Describe and Classify Information

### Taxonomy

- A system for naming and organizing things into groups that share similar characteristics
- The purpose of taxonomy is to group content into a controlled set of categories

*Jean Graefel*



## Thesaurus

A thesaurus is a **terminological control device** used in translating from the natural language of documents, indexers or users into a more constrained system language. It is a **controlled and dynamic** vocabulary of **semantically and generically related** terms which covers a specific domain of knowledge

*From UNESCO*

### SOLUTIONS

(For chemical solutions only. For mathematics see the word block of MATHEMATICAL SOLUTIONS.)

BT1 homogenous mixtures

BT2 mixtures

BT3 dispersions

NT1 aqueous solutions

NT1 hypertonic solutions

NT1 isotonic solutions

RT solubility

RT solvents



# Preservation of nuclear information and records

## Thesaurus

- Tool to describe knowledge/information in structured form
- Communication language between user and computer



## Thesaurus Structure

- Controlled dictionary
  - Descriptors
  - Forbidden terms
- Semantic relationships (*language independent!*)
  - Hierarchical
  - Associative
- Synonyms
- Definitions, comments, scope notes



## 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



# Preservation of nuclear information and records

## 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





# Preservation of nuclear information and records

## 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.



# Preservation of nuclear information and records

## 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 EBCDIC 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
- 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
- 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. Analog versus Digital

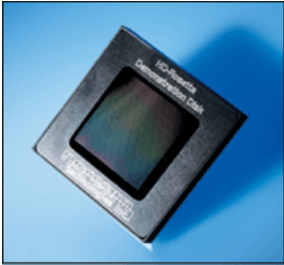
- Volume of information published in digital form is growing up dramatically (x2 every 3 years)
- Young generation preference is digital information
- New possibilities: Electronic document analysis, translation and data mining



# Preservation of nuclear information and records

## 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



School of Nuclear Knowledge Management  
22 – 28 September 2006, Trieste, Italy

9/8/2006 Preservation of nuclear information

37

International Atomic Energy Agency




School of Nuclear Knowledge Management  
22 – 28 September 2006, Trieste, Italy

9/8/2006 Preservation of nuclear information

38

International Atomic Energy Agency

## Part 2

## Digital Preservation

School of Nuclear Knowledge Management  
22 – 28 September 2006, Trieste, Italy

9/8/2006 Preservation of nuclear information

39

International Atomic Energy Agency

## 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*

School of Nuclear Knowledge Management  
22 – 28 September 2006, Trieste, Italy

9/8/2006 Preservation of nuclear information

40

International Atomic Energy Agency

# Preservation of nuclear information and records

## Two main standards

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

**TDR** - Trusted Digital Repositories: Attributes and Responsibilities

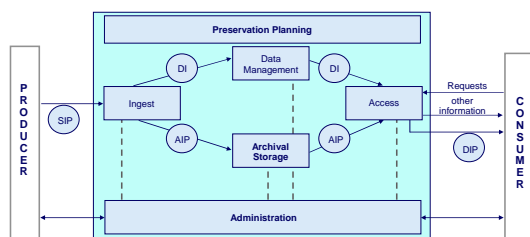


## Open Archival Information System (OAIS)

- 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



## 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.



# Preservation of nuclear information and records

## 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?**



## Issues to Consider

- **~65% digital preservation projects failed**



## 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"



## Preservation of nuclear information and records

**Small steps are usually better than no steps!**

**Thanks for your attention!**

