# Fifteenth International Conference on Grey Literature

## The Grey Audit: A Field Assessment of Grey Literature

Bratislava, Slovak Republic 2-3 December 2013

Slovak Centre of Scientific and Technical Information

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ • SR

CVTI SR

**GL15 Program and Conference Bureau**

TextRelease

**Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands**
**www.textrelease.com ▪ conference@textrelease.com**
**Tel/Fax +31-20-331.2420**

# Digital Preservation at International Nuclear Information System (INIS)

Dobrica Savic,

Head of Nuclear Information Section, IAEA

Germain St-Pierre,

Digital Preservation Technician, Nuclear Information Section, IAEA

## Abstract

Since its creation in 1970 until 1996, the International Nuclear Information System (INIS) collected and converted to microfiche over 312 000 non-conventional literature (NCL) reports received from IAEA member states and international organizations. The microfiche collection contains over 1 million items, with an estimated total of 25 million pages of full-texts.

In 1997, the INIS Secretariat replaced the microfiche-based production system with an imaging system to process and to disseminate all NCL documents in electronic format. That marked the beginning of digital preservation efforts that still continue today.

This paper provides an overview of the digital preservation practices and the technical infrastructure of the International Nuclear Information System (INIS). It describes the technical processes, the standards in place, the hardware and software used, as well as all practices related to scanning, quality control, OCR, preservation and storage.

# 1. Introduction

Since its creation in 1970 until 1996 (INIS, 2010), the International Nuclear Information System (INIS)[1] collected and converted to microfiche over 312 000 non-conventional literature (NCL) reports received from member states and international organizations. The microfiche collection contains over 1 million items, with an estimated total of 25 million pages of full-texts.

In 1997, the INIS Secretariat replaced the microfiche-based production system with an imaging system to process and to disseminate all NCL documents in electronic format. That marked the beginning of digital preservation efforts that still continue today.

This paper provides an overview of the digital preservation practices and the technical infrastructure of INIS. It describes the technical processes, the standards in place, the hardware and software used, as well as all practices related to scanning, quality control, OCR, preservation and storage.

# 2. Technical Infrastructure

The INIS digital preservation technical infrastructure has evolved on a regular basis since the beginning of digital imaging activities. Initial period known as INIS Imaging System lasted from 1997 to 2003, to be followed by a new INIS Imaging System (INISIS2K) period which lasted from 2003 to 2009. Current technical infrastructure was introduced in 2010.

# 3. INIS Imaging System (INISIS) - 1997 to 2003

In 1997, Jouve Systems was selected as a full-scale imaging system to process and disseminate INIS NCL in electronic format (INIS, 1999). This “cradle-to-grave” image-based solution replaced the microfiche-based production system which had been in place at the INIS Secretariat since 1970. The following modules were already part of the original design: workflow monitoring, black and white scanning, image import, image enhancement, quality control, link creation using barcode recognition, link validation against INIS bibliographic metadata and INIS rules, cumulative index creation as well as CD-ROM production according to the INIS NCL Viewer specifications (INISIR). Originally, only the TIFF Group 4 format was supported. In 2002, support for incoming documents in PDF was added, although the Jouve system was phased out only in 2003.

[1] The International Nuclear Information System (INIS) hosts one of the world's largest collections of published information on the peaceful uses of nuclear science and technology. INIS is operated by the International Atomic Energy Agency (IAEA) in collaboration with over 150 member states and international organizations. There are over 3.4 million bibliographic references to publications, documents, technical reports, non-copyrighted documentation and other ‘grey literature’ made available, as well as 350 000 full texts.  INIS offers free and open online access to this unique collection of non-conventional literature through its search application (http://inis.iaea.org/search/).

## 4. INIS Imaging System (INISIS2K) - 2003 to 2009

A study carried out in 2000 by Doculabs recommended building a new INIS Imaging System (INISIS2K) on one of the following “off-the-shelf” 32-bit information capture systems: Kofax Ascent Capture or ActionPoint InputAccel (now part of the EMC-Captiva family) )[2]. InputAccel (IA) was selected, mainly because of its powerful open architecture technology that allowed customization and system integration with Open-Text Livelink[3], the IAEA standard Document Management System. InputAccel also met new requirements such as colour scanning, optical character recognition (OCR) and output to PDF.

Replacement of the INISIS imaging system led to a significant improvement in the production cycle, which was synchronized with the bibliographic database production. All documents were output in PDF and those in Western European, Cyrillic and Slavic scripts were OCRed (INIS, 2004).

From the beginning, INISIS2K was conceived and implemented as one of the components of a larger system, a completely overhauled INIS Processing System (IDPS) based on Livelink technology. All tasks, from the initial imaging request sent to the InputAccel server until the ingestion of its PDF output into the document repository, were monitored through Livelink. This was also the case for the quality control of bibliographic data, the ingestion of NCL input submitted by the National Centres in PDF format, the migration of all new records to the INIS Online Database, and finally for the preparation of an ISO image for distribution of the full-texts on CD-ROM.

In 2006, in order to streamline workflow, improve efficiency and free resources for other activities, the INIS Secretariat issued revised ’Guidelines on How to Submit Full-Text of Non-Conventional Literature (NCL) to INIS‘ (INIS, 2006). The INIS National Centres were strongly encouraged to submit their NCL input directly in PDF and the response from Member States was favourable.

Three new priorities were identified: the digitization of the INIS microfiche collection, the conversion to PDF of all the documents scanned and distributed in TIFF between 1997 and 2003, and the online access to full-texts via the INIS Online Database.

Although highly efficient when introduced in 2003, InputAccel lacked flexibility when it came to the development of workflows tailored for other digitization projects. The maintenance of this modular client/server application was also very expensive and required significant effort from the Systems Development and Support Group (SDSG). Finally, incompatibility of the communication module with Livelink was found during testing of InputAccel v.5.3. This made the migration to this new platform impossible without additional expensive developments. INIS decided to stop the maintenance contract for InputAccel at the end of 2009 and abandoned the system with the migration of all desktops to Windows 7 in 2010.

[2] http://www.emc.com

[3] Livelink was the first Web-based collaboration and document management system made by the OpenText. http://www.opentext.com/2/global/products/products-all/livelink-landing.htm

During this period, the INIS imaging infrastructure consisted of 6 scanning workstations, 3 servers, 4 high performance scanners, 2 flatbed scanners, 1 high performance microfiche scanner and 1 digital camera. The technical characteristics are indicated in the table below.

| Scanner | Type | Paper size | Resolution (dpi) | Bit-in-depth | Speed (A4, 200 dpi) | ADF Page capacity |
|---|---|---|---|---|---|---|
| Fujitsu fi-5750c with VRS Pro | Colour; ADF/flatbed | A8 – A3<br>Up to 34 inches | 50 to 600 | 24 | 110 p/min (simplex)<br>55 p/min (duplex) | 200 p |
| Fujitsu M4099D | B&W; ADF | A7 – A3 | 200, 240, 300, 400 | 10 | 90 p/min (simplex)<br>180 p/min (duplex) | 1000 p |
| Fujitsu M3099GX | B&W; ADF | A7 – A3 | 200, 240, 300, 400 | 8 | 60 p/min (simplex)<br>120 p/min (duplex) | 1000 p |
| Fujitsu M3099G | B&W; ADF | A5 – A3 | 200, 240, 300, 400 | 8 | 55 p/min (simplex)<br>110 p/min (duplex) | 500 p |
| Kodak i260 | Colour; ADF/flatbed | A5 – A3 | Up to 600<br>Optical Resolution 300 | 16-48 | 50 p/min (simplex)<br>100 p/min (duplex) | 150 p |
| SunRise 2000 | Microfiche scanner | A0-A4 reductions 7x-50x | CCD 3600-8800 True Resolution | | Up to 2500 frames/hr | |

Table 1: Imaging Infrastructure 2003 – 2009

## 5. Current Technical Infrastructure

A complete re-evaluation of the technical infrastructure was carried out in 2010, in line with the implementation plan of the 'Desktop 2010' project developed by the IAEA Division of Information Technology (MTIT) (INIS, 2011). An important goal of this project was to ensure security and supportability of all computer systems of the Agency network.

Windows 7 compliance of all equipment and software applications had to be verified through testing prior to the deployment of this new platform. Also, an important reduction in space requirement was an expected outcome of this exercise.

The 3 Fujitsu black and white SCSI scanners, the Kodak i260, the InputAccel system and some small utilities failed this compliance test. Also, several old workstations did not meet the minimum requirements and had to be replaced.

New computers with fast quad-core processors supporting multithreading and multitasking were procured. The number of scanners was reduced to two, both of them supporting color, greyscale and black and white scanning.

### 5.1 Software

The following software and applications are currently used for digitization at INIS:

**Techsoft PixEdit v.7.11.18:** PixEdit was introduced in the imaging workflow in 2000. It is primarily used for its advanced image editing capabilities. This flexible application gradually proved to be an excellent scanning utility. Since the discontinuation of the InputAccel system in 2010, PixEdit is the main scanning application. Five seat licenses are currently available.

**ABBYY FineReader 11 Corporate Edition:** FineReader is used for Optical Character Recognition (OCR). It can process mono or multilingual documents, supports different alphabets including Cyrillic languages and offers an accuracy level of close to 98%. ABBYY policy for this product is to release a new version each year. Version 11 was bought in 2011 together with an upgrade assurance to Version 12 in 2012.

**Adobe Acrobat X Professional** is used for OCR of Chinese (Simplified), Japanese and Korean, as well as for document optimization and conversion to PDF/A[4], when applicable.

**Kofax Virtual ReScan (VRS) + Kodak Perfect Page**: Both technologies have hardware and software components that reduce the need for post-scanning image enhancement.

## 5.2 Hardware

**Scanners -** One of the most important elements in a digitization project is the selection of the appropriate image capture devices, as scanners have great impact on image quality. The choice of equipment depends on a number of factors, including the format, size and condition of the material that will be digitized.

Several types of digitization equipment exists, i.e. flatbed scanners, sheet-fed scanners with automatic document feeder (ADF), drum scanners, open book scanners, digital cameras, and film scanners.

INIS quality scanners are calibrated and maintained regularly. Special methods, including Scanner Test Charts, are used to check image resolution, dynamic range mapping, as well as photographic tone and color reproduction.

There are currently 2 colour scanners with automatic document feeder (ADF) and flatbed, as well as 2 high performance microfiche scanners. The technical characteristics are indicated in the table below.

| **Scanner** | **Type** | **Paper size** | **Resolution (dpi)** | **Bit-in-depth** | **Speed (A4, 200 dpi)** | **ADF Page capacity** |
|---|---|---|---|---|---|---|
| Fujitsu fi-5750c with VRS Pro | Colour; ADF/flatbed | A8 – A3<br>Up to 34 inches | 50 to 600 | 24 | 110 p/min (simplex)<br>55 p/min (duplex) | 200 p |
| Kodak i1440 | Colour; ADF/flatbed | A5 – A3 | Up to 600<br>Optical Resolution 300 | 16 - 48 | 50 p/min (simplex)<br>100 p/min (duplex) | 150 p |
| SunRise | Microfiche | A0-A4 reductions | CCD 3600-8800 | | Up to 2500 | |

[4] *PDF/A* is an ISO-standardized version of the Portable Document Format (*PDF*) specialized for the digital preservation of electronic documents. http://en.wikipedia.org/wiki/PDF/A

|  |  |  |  |  |
|---|---|---|---|---|
| 2000 | scanner | 7x-50x | True Resolution | frames/hr |
| SunRise Apollo | Microfiche scanner | A0-A4 reductions 7x-50x |  | Up to 3600 frames/hr |

Table 2:- INIS Scanner Specifications

**Computers -** Careful consideration was given to the following points when selecting PCs dedicated to digitization work: fast central processor unit (CPU), sufficient random access memory (RAM), fast data transfer rate between components, large disk storage capacity, suitable interface, as well as high quality audio and video cards.

**Monitors -** Large display monitors provide better viewing and image evaluation. As each type, size and quality of monitor interprets and displays values differently, special care is devoted to their adjustment and calibration. Four quality control workstations and the 2 scanning workstations are equipped with widescreen monitors (30-inch LCD monitors, model LP 3065 from Hewlett-Packard).

## 6. General digitization principles at INIS

INIS aims to ensure a consistent, high level of image quality, interoperability and accessibility of digitized materials, as well as long-term preservation for future generations in Member States. To achieve these goals, INIS developed some general principles based on Cornell University's digital imaging tutorial (http://www.library.cornell.edu/preservation/tutorial/index.html), and adjusted them to INIS requirements. The current workflow is based on these principles and described in detail later. The principles can be summarized as follows:

1. benchmarking for image quality and resolution
2. scanning at the level appropriate to the content of the original source
3. digitization of 1st generation material, if available, in order to achieve best possible image quality
4. creation and storage of a master image file
5. use of format and compression techniques that conform to standards (avoiding proprietary formats)
6. creation of backup copies
7. storage of digital files in an appropriate environment
8. off-site storage of the collection
9. metadata for digital resources
10. integration of image files with bibliographic metadata in the INIS Collection

## 7. Image Creation Process

The process of initial capture or conversion of a paper or microfiche based document or object into digital form is known as image creation. Based on the experience gained over the years and through benchmarking, INIS dedicates special attention to the physical nature of the documents to be digitized and applies different measures. Collections available at Member States differ in the ways they are created, used and accessed. The quality and condition of the original material will have a direct impact on the digitization approaches. Therefore, INIS applies the principle to scan at a level that matches the information content of the original.

Before starting a digitization project, it is crucial to obtain copyright permissions from copyright holders. The Intellectual Property Rights laws are comprehensive and complex, and the progress in today's online environment presents serious challenges for copyright compliance.

In this respect, INIS relies on INIS Member States to ensure that appropriate permission is obtained before the full-text of a publication is sent to INIS for inclusion in the Collection.

It is essential that each stage of the digitization process is planned ahead and an appropriate workflow is established. It is not a simple task to create an effective and efficient ***digitization workflow***. However, if properly planned it will support staff performance and enable high quality work. The stages of workflow at INIS are the following:

- Benchmarking
- Source material types
- Preparation
- Scanning
- Quality control
- Image enhancement
- File formats
- Compression
- File naming convention
- Optical character recognition (OCR)
- Storage
- Preservation planning
- Metadata creation

### 7.1 Benchmarking

INIS considers benchmarking for digital capture the first and most important step of the digitizing effort. The results of benchmarking considerably affect all further steps (scanning, enhancement, format, etc.). The purpose of benchmarking is to define and clarify the following:

- Can the informational content of the original material be adequately captured in digital form?

- Does the physical format and condition of material correspond to digitizing requirements?
- What is the type of material to be digitized?
- Which resolution should be applied?
- At which bit-depth?
- Which compression parameters should be set?
- What is the estimated accuracy level for OCR?
- Other considerations?

## 7.2 Source material types

The variety of source material may be categorized, but not limited to:

- Printed text/simple line art
- Rare or damaged printed text
- Manuscripts
- Maps, architectural drawings
- Halftones
- Continuous tone
- Microformats
- Mixed

The majority of the material digitized at INIS is text-based containing illustrations, graphics, photos (black & white, colour), as well as oversized materials with fine details, line drawings, etc., falling mainly into the above cited *Printed Text* and *Mixed* categories. The category of *Printed Text* can be described as distinct edge-based representation that is cleanly produced, with no tonal variation, such as a book containing text and simple line graphics. Documents containing two or more of the categories listed above, such as illustrated books, can be defined as *Mixed.*

## 7.3 Preparation

Good document preparation facilitates scanning and ensures quality results. Materials to be scanned need to be prepared in the following manner:

- Physically (unbinding, removing of staples and clips, separation when glued, etc.);
- Structurally (adding/removing barcodes, separating chapters, sections, parts, covers, etc.);
- According to specific characteristics, e.g. size, thickness, quality (glossy/mat), condition of paper, etc.

Inadequate document preparation can result in paper jams inside the scanner and lead to irreparable damage to original documents. In order to unbind documents in an efficient and safe way, INIS utilizes a professional cutting machine from IDEAL (Model 4850-95).

### 7.4 Scanning

**Capture modes:** It is important to keep in mind that different capture methods are needed depending on the physical form of the original. Capturing is mostly performed in these 3 modes:

- Bitonal (1 bit per pixel) – represents two tones: black and white; best suited to high contrast documents such as printed text.
- Greyscale (8 bits per pixel) – represents 256 shades of grey; best suited to continuous tone documents such as black and white photographs. However, older photos (e.g. sepia tones) may provide better results when captured in colour.
- Colour (24 bits per pixel) – represents 16 million colours & shades of grey; suited to documents with continuous tone colour information.

'*Pixel*' stands for picture elements which make up an image. Each pixel can represent a number of different shades or colours depending on the storage space allocated to it.

***Optical Resolution:*** The optical resolution determines the quality of an image. It is normally expressed in scanner specifications as 'dots per inch' (DPI) or 'pixels per inch' (PPI) and refers to the number of pixels (dots) captured in a given inch. Increasing the resolution enables capturing of finer detail. However, it results in a larger file size. To determine the resolution necessary to capture all significant details present in the source document, Cornell University developed a formula called 'Digital Quality Index' (QI). This formula can be used as guidance for calculating the optimal scanning resolution**.** More information is available on Cornell's Web site at: http://www.library.cornell.edu/preservation/tutorial/

***Bit depth:*** The amount of information that a sensor in an array can capture is represented by the 'bit depth'. Greater bit depths result in a more accurate digital representation of the original. The final decision about resolution and bit depth depends on the goal of digitizing**.**

INIS applies a resolution range of 300 – 600 dpi for bitonal scanning to documents of A4-A5 size, and 200 – 300 dpi with 8 bit depth (256 colours/tones) for greyscale/colour scanning.

### 7.5 Quality Control (QC)

Quality Control is as an integral part of the digitization process in order to retain value, utility and integrity of the resources. QC consists of a set of procedures and techniques to verify the quality, accuracy and consistency of digitized material. QC is conducted by visual inspection of images on-screen with concentration on the resolution, colour, tone, and appearance. It is important to mention that this assessment may be highly changeable depending on the viewing environment and the characteristics of the monitors.

INIS applies a wide range of QC measures to ensure that quality expectations are met. During the QC process, INIS verifies accuracy and completeness of components, data integrity, metadata correctness, form and validity, as well as correct matching of metadata and image files.

In this context, the ‘checksum’ algorithm serves to ensure the authenticity and integrity of digitized files. It is essential to verify that the number and order of bytes in a file remain the same after moving, copying, transferring, burning or other actions.

In addition to the checksum, INIS also compares the number of pages of the original with the digitized product to ensure the completeness of digitized documents.

### 7.6 Image enhancement

Image enhancement is any process that is applied to the raw scan to improve quality or legibility of the resource. INIS applies several procedures and techniques to verify the quality, accuracy, consistency and integrity of digital products, including despeckling, deskewing, noise reduction, black border removal; colour and tone adjustment, etc.

### 7.7 File formats

There are several standard file formats which vary in terms of resolution, bit-depth, colour capabilities, etc. Although there is no clearly recommended archival format in use today, preference must be given to ‘non-proprietary’ formats.

INIS stores master digital images in **TIFF** Group IV which offers longevity and production of a range of delivery versions (e.g. for screen, for print, for web access). For purposes of delivery to Member States, electronic exchange with customers, users, and access via the INIS Online Database, files are converted to PDF (Portable Document Format) and compressed.

**PDF** is one of the most frequently used file formats to preserve electronic documents and ensure their survival for the future. Recently, the International Organization for Standardization (ISO) released the full PDF specification as ‘ISO 32000-1:2008’.

**PDF/A:** The PDF/Archival (PDF/A) standard aims to enable the creation of PDF documents whose visual appearance will remain the same over the course of time. This standard was adopted by the International Organization for Standardization (ISO) in autumn 2005 and published as ‘ISO 19005-1:2005’. INIS is considering adopting this standard to achieve preservation and long-term archiving of the Agency’s and Member States’ nuclear information resources. (that last sentence should perhaps be written a bi differently – not ideal the way it is now written…)

### 7.8 Compression

Compression algorithms are used to reduce image file size for storage, processing and transmission. There are two compression techniques, i.e. ‘lossless’ and ‘lossy’[5]. When lossless compression is applied, the space needed for the storage of an image file is reduced without loss of data. During lossy compression, the least significant information is averaged or discarded. Uncompressed files or compressed files using the lossless compression technique are clearly preferred. There are several standards, as well as proprietary compression software available to create images for web delivery. INIS has chosen the JBIG2 standard for web optimization of black/white resources, and JPEG for colour digital resources.

JPEG2000 uses wavelet compression to achieve small but high quality images and is increasingly being used as repository and archival image format. INIS is considering JPEG2000 as a possible alternative for image delivery.

### 7.9 File naming conventions

For system compatibility and interoperability, it is important to follow an established file naming convention. Unique file names assure consistency and easy retrieval of resources.

### 7.10 Optical Character Recognition (OCR)

In order for an image of a printed text to become searchable as electronic text, raster images are processed with an OCR program to be translated to machine editable text.

For INIS digitization projects, the creation of ‘searchable full text’ has been defined as the primary objective. The quality and condition of the original material will have a direct impact on the OCR result.

INIS uses ABBYY FineReader, an Optical Character Recognition (OCR) software that allows users to convert paper documents, PDF files, and various images including photographs taken by a digital camera to editable formats for changing and repurposing. Close to 98% accuracy is reached at character level when applying OCR to raster images of text printed in Latin and Cyrillic characters.

Recent tests have provided satisfactory results using Adobe Acrobat Professional 8.0 for OCR of documents in Chinese (Simplified), Japanese and Korean. Tests with ABBYY FineReader Pro9 for Hebrew and Thai also provided good results. Further tests are being performed to identify suitable tools for the Arabic language.

Latest developments in OCR technology include recognition of document structure known as ’Logical Form Recognition‘ (Omnipage 16) or ’Adaptive Document Recognition‘ (FineReader Pro 9). While accuracy has greatly improved in font type and font size recognition, OCR technology also makes intelligent use of hardware technologies, such as ‘multi core parallel processors’ for speeding up the OCR

[5] See File compression: An introduction for raster image files. http://www.library.carleton.ca/sites/default/files/help/gis/File_Compression.pdf

process. The trends in OCR technology show that significant developments can be expected in the future.

### 7.11 Storage

In order to provide longevity of digital files, they need to be stored in a reliable, controlled environment (White). Master files should be stored on high quality, industry standard devices, such as CD-R, DVD, or other contemporary reliable media. Backups of master files must be created regularly and stored off-site in a secure location.

A ***RAID*** (Redundant Array of Inexpensive or Independent Disks) consists of a number of drives which collectively act as a single storage system. The production of digital material requires sufficient hard disk capacity to store files at various stages of the preservation process. It may be appropriate to consider a RAID solution if the production environment large.

In 2008, INIS purchased a THECUS N5200B PRO, 5x3,5" SATA Raid. The equipment has 5 disks of 1 TB each and has been configured as local network data storage.

Backups of master files must be created regularly and stored away from the original source in a secure location on a routine basis.

Since the beginning of the system until 1997, INIS converted all full texts of NCL from paper into microfiche for safer long-term storage. In 1997, a complete collection of NCL on microfiche, representing the intellectual knowledge and information of INIS Member States, was donated to the Central Library of Physics of the University of Vienna, which acts as a secure 'off-site‘ storage. As the microfiche collection is being converted to PDF, all digitized resources are also backed-up in PDF at the Central Library of Physics, which is situated less than 5 kilometers from the IAEA.

### 7.12 Preservation planning

In order to ensure that the contents of a digital archive remain a readable and usable information resource for the future, digital files should regularly be refreshed to new media (Hedstrom & Montgomery, 1998). This can be achieved by using different processes. The process of copying files from one storage medium to another medium of the same kind is called ***refreshing.*** This targets media obsolescence. After media refreshing, a verification procedure should be applied (e.g. checksum) to ensure the authenticity and integrity of the files.

Another process is ***migration***, i.e. transferring digital information from one hardware and software setting to another or from one computer generation to subsequent generations. Migration can also be format-based, to move image files from an obsolete file format to a new format.

A third process is called ***emulation*** which involves the re-creation of the technical environment required to view and use a digital resource. This is achieved by maintaining information about the hardware and software requirements so that the system can be reengineered. Due to its cost and the time required for proper emulation, this process is not often used.

At present, INIS applies the technique of **refreshing** the digital files by copying the collection to a new storage media, e.g. CD to DVD, Blu-ray Disc, etc. At the time of implementation of PDF/A as long-term archival format, the **migration** technique will be applied to all INIS digital files.

### 7.13 Metadata creation

Metadata plays a key role in describing, processing, managing, tracking, accessing and preserving digital resources. According to NISO[6] (2004), metadata is key to ensuring that resources will survive and continue to be accessible into the future.

There are different types of metadata that can be associated with digital resources. INIS applies comprehensive 'bibliographic metadata' which describes the intellectual content of the digitized full text and includes an extended set of bibliographic elements for identification and retrieval of the resources. When integrated to the INIS Database, digital resources are accompanied and linked to their corresponding bibliographic records (INIS, 2009). The whole process is carefully reviewed by INIS specialists and validated by computer programs and specially designed algorithms.

At present, technical metadata for digital resources is generated automatically during creation of the PDF files. However, a more sophisticated approach will be considered along with the implementation of PDF/A.

## 8. Microfiche Digitization Project

The in-house digitization of the microfiche collection started in 2002 after the acquisition of a Sunrise 2000 microfiche scanner. Initially only aimed at fulfilling document delivery requests, the digitization of the full collection became an actual topic in 2003, after a release of the new INIS on-line database that supported direct access to full-texts.

It was decided to outsource a substantial part of the microfiche scanning in order to support the existing in-house digitization capabilities. The contracts were issued after formal invitation to bid and the amount of microfiche scanning requests depended on the funds available. Some funding for this project was provided by the Nuclear Knowledge Management Unit (NKM) of the IAEA. Over the years, the following three contractors were engaged: EMD Austria, Prosoft Germany, and PM Dimensions Austria.

It should be mentioned that good coordination and a good strategy are necessary to ensure the success of such a project. It was especially important to avoid duplication of work and to take into consideration the different digitization initiatives by the IAEA Member States. For this reason, INIS chose the country of publication as the

[6] National Information Standards Organization (NISO), a non-profit association accredited by the American National Standards Institute (ANSI), identifies, develops, maintains, and publishes technical standards to manage information. http://www.niso.org

main selection criteria, and an extensive coordination effort with the respective Member State followed each decision to digitize their part of the INIS based microfiche collection. In order to support national document and knowledge preservation efforts, the INIS Secretariat provided Member States with DVD country sets of their digitized non-conventional literature (NCL) from the microfiche.

The following table gives a detailed overview of INIS microfiche digitization activities since its inception in 2003.

| Year | PDF | Pages | Size (GB) |
|---|---|---|---|
| 2003 | 566 | 49 574 | 3.7 |
| 2004 | 19 962 | 1 325 217 | 36.5 |
| 2005 | 36 935 | 1 577 365 | 32.1 |
| 2006 | 23 163 | 1 367 637 | 33.3 |
| 2007 | 9 313 | 668 769 | 16.3 |
| 2008 | 25 675 | 1 228 057 | 29.7 |
| 2009 | 81 221 | 3 939 811 | 77.3 |
| 2010 | 33 881 | 1 969 110 | 45.9 |
| 2011 | 24 027 | 511 990 | 16.2 |
| 2012 | 20 434 | 843 579 | 40.7 |
| Total: | 275 177 | 13 481 109 | 331.8 |

Table 3: Digitization of the INIS NCL Collection on Microfiche

Close to 80% of the INIS microfiche collection has been digitized since the beginning of the project. An estimated 3 million pages still need to be processed before project completion. Depending on available resources, this major project is expected to be completed within the next two years. The ultimate goal is a complete integration of the microfiche-based NCL into the INIS Collection and online access to full-texts provided via the Google-based INIS Collection Search

The INIS Collection Search (ICS) is a free and open web access of the INIS Collection to all Internet users. Currently it holds over 3.4 million bibliographic (metadata) records and over 350 000 full-text NCL documents. This collection of documents on the peaceful uses of nuclear science and technology is now fully indexed and searchable online using Google-based technology. Around 50 000 searches and 3000 downloads are performed monthly. A link to the INIS Collection search is available from the INIS home page[7] or directly from http://inis.iaea.org/search/.

Besides digitization of its microfiche collection of NCL, INIS is also involved with the digitization of old IAEA publications. Examples of INIS in-house efforts include

[7] http://www.iaea.org/inis

the digitization of IAEA Bulletins[8] in all available languages, accompanied by INIS bibliographic metadata; digitization of Member State's Technical Reports and Proceedings Series, done in cooperation with the IAEA Library; digitization of reports from the International Nuclear Data Committee collection (INDC); and the digitization of out-of-print IAEA publications.

## 9. Conclusion

Large digitization projects, such as the digitization of the INIS microfiche collection of historic non-conventional literature, require serious planning, substantial funds, qualified staff, awareness of standards, and well defined purpose. Lack of qualified personnel can be mitigated by outsourcing to companies specialized in large volume digitization projects. Lack of in-house knowledge about the various aspects of digitization can also be alleviated by hiring experts and consultants, but it is important to maintain consistent quality throughout all of the digitization workflow steps, at the maximum possible level. Document preparation, selection of applicable scanning techniques, type of equipment, and adherence to current standards, are all factors which will decide success or failure of any digitization effort.

Digitization should not be a goal in and of itself. Its ultimate use and usefulness must always be taken into account. It is therefore imperative that meaningful and searchable metadata accompany any digitized collection with a goal of making such a repository available through appropriate online search and delivery tools. Once this is achieved, ways and means for long term preservation need to be considered and put in place in order to ensure future sustainability and availability of the digitized collection.

[8] http://www.iaea.org/Publications/Magazines/Bulletin/Bull521/index.html

# References

Headstrom, M. & Montgomery, S. (1998). *Digital Preservation Needs and Requirements in RLG Member Institutions*. A study commissioned by the Research Libraries Group. Mountain View, CA, USA: RLG. Retrieved from http://www.oclc.org/research/activities/past/rlg/digpresneeds/default.htm

INIS (1999). *INIS Status Report 1998. Twenty Seventh Consultative Meeting of INIS Liaison Officers.* 631-L2-TC-441.27/2. Retrieved from http://goo.gl/HQWic

INIS (2004). *INISProgress and Activity Report 2003.* L2.04.01/INIS-PAR/2003. Retrieved from http://goo.gl/nXe4p

INIS (2006). *New Guidelines for Submission of Non Conventional Literature (NCL) full text to INIS.* INIS Technical Note No. 185

INIS (2009). *Guide to Bibliographic Description* (Rev.8). June 2009, Amend. 3. IAEA-INIS-01. Retrieved from http://goo.gl/bXNPm

INIS (2010). *The International Nuclear Information System (INIS): The First Forty Years.* Prepared by C. Todeschini. Retrieved from http://goo.gl/w7hUV

INIS (2011). *INISProgress and Activity Report 2010.* Retrieved from http://goo.gl/s3yQe

NISO (2004). Understanding Metadata. Bethesda, MD, USA: NISO Press. Retrieved from www.niso.org/standards/**resources**/Understanding**Metadata**.pdf

White, J. *Guidelines for Using Electronic Records: Illinois State Archives*. Retrieved from http://www.cyberdriveillinois.com/departments/archives/records_management/electrecs.html