IAEA International Atomic Energy Agency — Nuclear Information and Knowledge

News from the Nuclear Information Section
International Nuclear Information System (INIS) & IAEA Library
No. 14, June 2013

The History of Digital Preservation at INIS

Since its creation in 1970, the International Nuclear Information System (INIS) collects and disseminates non-conventional literature (NCL) received from Member States and international organizations[1]. Initially, INIS received paper based NCL documents which were microfilmed in-house[2] and stored in the INIS archives. In 1997, the INIS Secretariat replaced the microfiche-based production system with an imaging system to process, preserve and disseminate all NCL documents in electronic format. This marked the beginning of digital preservation efforts that continue today.

The INIS digital preservation technical infrastructure has evolved on a regular basis since 1997. Its history can be divided into three periods:

- 1997–2003: INIS Imaging System (INISIS)
- 2003–2009: INIS Imaging System 2000 (INISIS2K)
- 2010 to present: Current technical infrastructure

This article provides an overview of the digital preservation practices and the technical infrastructure of INIS. It describes the hardware and software used, as well as some practices related to scanning and quality control. The digitization of the INIS microfiche collection, a unique archive containing over 1 million physical assets, including more than 312 000 non-conventional literature reports, was described in detail in INIS Newsletter No. 13, September 2012[3]. The digitization of this archive will lead to a collection of approximately 17 million pages of full texts.

INIS Imaging System (INISIS) — (1997–2003)

In 1997, Jouve Systems was selected as a full-scale imaging system to process and disseminate INIS NCL in electronic format[4]. 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 conceived to support TIFF Group 4 file format only, the system was modified in 2002 to accept a growing number of incoming full texts in PDF. The Jouve system was discontinued in 2003 after a phased migration to the new INIS Imaging System (INISIS2K).

INIS Imaging System 2000 (INISIS2K) — (2003–2009)

In 2000, a study carried out by Doculabs[5] recommended building a new INIS Imaging System (INISIS2K) on a leading 'off-the-shelf' 32-bit information capture system. Among the shortlisted products, INIS selected ActionPoint InputAccel[6] , mainly because of its powerful open architecture technology that allowed customization and system integration with Open-Text Livelink[7], the IAEA standard Document Management System. InputAccel also met new requirements such as colour scanning, optical character recognition (OCR) and output to PDF.

The 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[8].

From the beginning, INISIS2K was conceived and implemented as one of the components of a larger system, a completely overhauled INIS Data 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'[9]. 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 on-line access to full texts via the INIS Online Database[10].

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 in-house IT group. The InputAccel system was phased out during the migration of all desktops to Windows 7 in 2010.

During this period, the INIS imaging infrastructure consisted of 4 scanning workstations, 3 Quality Control 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/flat bed	A8 – A3 Up to 34 inches	50 to 600	24	110 (simplex) 55 (duplex)	p/min p/min	200 p./
Kodak i260	Colour; ADF/flat bed	A5 – A3	Up to 600 Optical Resolution 300	16-48	50 (simplex) 100 (duplex)	p/min p/min	150 p
Fujitsu M4099D	B&W; ADF	A7 – A3	200, 240, 300, 400	10	90 (simplex) 180 (duplex)	p/min p/min	1000 p
Fujitsu M3099GX	B&W; ADF	A7 – A3	200, 240, 300, 400	8	60 (simplex) 120 (duplex)	p/min p/min	1000 p
Fujitsu M3099G (back up)	B&W; ADF	A5 – A3	200, 240, 300, 400	8	55 (simplex) 110 (duplex)	p/min p/min	500 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

Current Technical Infrastructure (2010 to present)

In 2010, a complete re-evaluation of the technical infrastructure was carried out. The objectives and expected outcome of the 'Desktop 2010' project[11] was to ensure security and supportability of all computer systems of the IAEA network and the compliance of all equipment and software applications with Windows 7, the IAEA standard operating system.

The 3 Fujitsu black and white SCSI scanners, the Kodak i260 colour scanner, the InputAccel system and some small utilities failed the compliance tests. Also, several old workstations did not meet the minimum requirements for Windows 7 and had to be replaced.

Special attention was paid to ergonomics while planning the new work environment. The number of workstations was reduced by almost 50% by procuring new computers with fast quad-core processors supporting multithreading and multitasking. The number of scanners was also reduced to two, both of which support colour, greyscale and black and white scanning. This significant reduction of equipment dedicated to digitization, coupled with an efficient optimization of the

digitization tools, helped the INIS Secretariat in their effort to reduce the space required for operations.

Furthermore, the new and flexible technical infrastructure has enabled the INIS Secretariat to support several important digital preservation initiatives within the IAEA. This includes, for instance, the digitization of out of print IAEA publications from Technical Reports, Safety and Proceedings Series, the digitization of the IAEA Bulletin in all official languages, as well as the digitization of historical photographs from IAEA archives. The following software and hardware are currently used for digitization at INIS:

Techsoft PixEdit v.7.11.18: PixEdit was introduced to the imaging workflow in 2000. It is primarily used for its advanced image editing capabilities. This flexible application has proven 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, 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 purchased in 2011, together with an upgrade assurance to Version 12.

Adobe Acrobat X Professional is used to OCR Chinese (simplified), Japanese and Korean documents, as well as for document optimization and conversion to PDF/A[12], 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.

Scanners: To digitize paper documents, INIS uses 2 colour scanners with automatic document feeder (ADF) and flatbed. One big advantage of this new generation of scanners is that they no longer require time-intensive calibration.

To digitize its microfiche archive, INIS uses 2 high performance microfiche scanners.

The technical characteristics of the scanners are indicated in the table below.

Scanner	Type	Paper size	Resolution (dpi)	Bit-in-depth	Speed (A4, dpi)	200 ADF Page capacity
Fujitsu fi-5750c with VRS Pro	Colour; ADF/flatbed	A8 – A3 Up to 34 inches	50 to 600	24	110 p/min (simplex) 55 p/min (duplex)	200 p
Kodak i1440	Colour; ADF/flatbed	A5 – A3	Up to 600 Optical Resolution 300	16 - 48	50 p/min (simplex) 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	
SunRise Apollo	Microfiche scanner	A0-A4 reductions 7x-50x			Up to 3600 frames/hr	

Table 2:- INIS Scanner Specifications

Conclusion

The technical infrastructure in place at the INIS Secretariat has allowed for the conversion to microfiche of over 312 000 non-conventional literature reports received from Member States and international organizations between 1970 and 1996. The migration to a scanning environment in 1997 was an important milestone in INIS history. This marked the beginning of digital preservation and was the first step towards building a durable digital repository which now contains over 475 000 full texts in PDF format, with 21,3 million pages of full texts and 597 gigabytes of data. Large digitization preservation projects, such as the digitization of the INIS microfiche collection of historic non-conventional literature, require substantial funds, qualified staff, and adequate software and hardware tools.

References

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

[2] Except for the NCL from the U.S.A. and from Japan, which were received in microfiche form

[3] Newsletter available at: http://www.iaea.org/inis/products-services/newsletter/INIS-Newsletter-2012-13/2012-13-07/index.html

[4] 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

[5] http://www.doculabs.com

[6] Now part of the EMC-Captiva family (http://www.emc.com)

[7] 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

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

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

2013-07-16 00:00

[10] The INIS Online Database was based on the BASIS Search technology. It has been replaced in April 2011 with the INIS Collection Search (ICS), which is based on Google Search Appliance (GSA) (http://www.iaea.org/INIS/). To find out more about the history of the INIS Collection Search, see http://www.iaea.org/inis/products-services/newsletter/INIS-Newsletter-2012-13/2012-13-04/index.html

[11] The 'Desktop 2010' project was developed by the IAEA Division of Information Technology (MTIT) and implemented in the whole INIS working Unit by the Systems Development and Support Group (SDSG). INISProgress and Activity Report 2010. Retrieved from http://goo.gl/s3yQe

[12] PDF/A is an ISO-standardized version of the Portable Document Format (PDF) specialized for the digital preservation of electronic documents. http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml

Dobrica Savic
Head, Nuclear Information Section

Germain St-Pierre
Digital Preservation Technician, Nuclear Information Section