Grey Literature and Digital Preservation: Standards in Practice
GL17 Pre-Conference Workshop
30 November 2015, Amsterdam

Digitization and Preservation of Non-Conventional Literature for the International Nuclear Information System (INIS)

Germain St-Pierre
g.st-pierre@iaea.org

Nuclear Information Section
IAEA, Vienna

IAEA

Scope of this Presentation

- Introduction: What is INIS?
- Overview of the INIS System Architecture
- Non-Conventional Literature (NCL) in the INIS Collection
- Preservation and Digitization of NCL (1970 to now)
- NCL Collection Management System
- Microfiche Digitization Project
- Other Digitization Activities
- Practical Session
- Conclusions

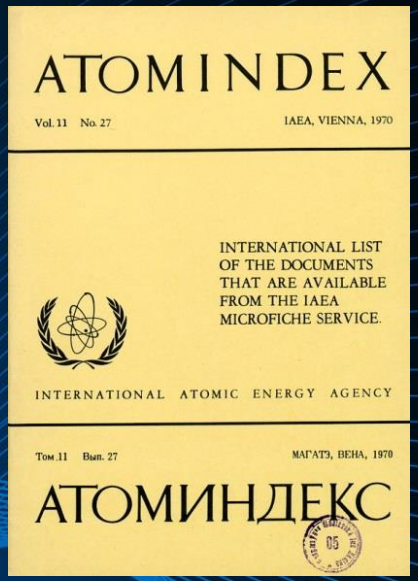
International Nuclear Information System (INIS)

- One of the world's largest collections of published information on the peaceful uses of nuclear science and technology
- Operated since 1970 by the IAEA in collaboration with 130 countries and 24 international organizations
- 45 years of successful international cooperation

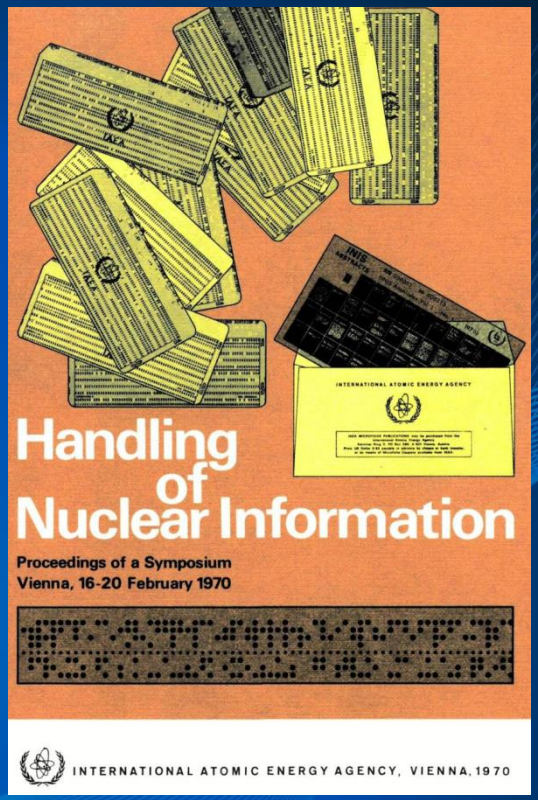
INIS in 1970 …

INIS Atomindex on magnetic tape

INIS Atomindex in hardcopy



NCL on microfiche



Handling of Nuclear Information
STI/PUB/254

… and NOW

- *Free online access* to a unique collection of non-conventional (grey) literature
- Google-based INIS Collection Search (ICS): *inis.iaea.org/search*

INIS Path through History

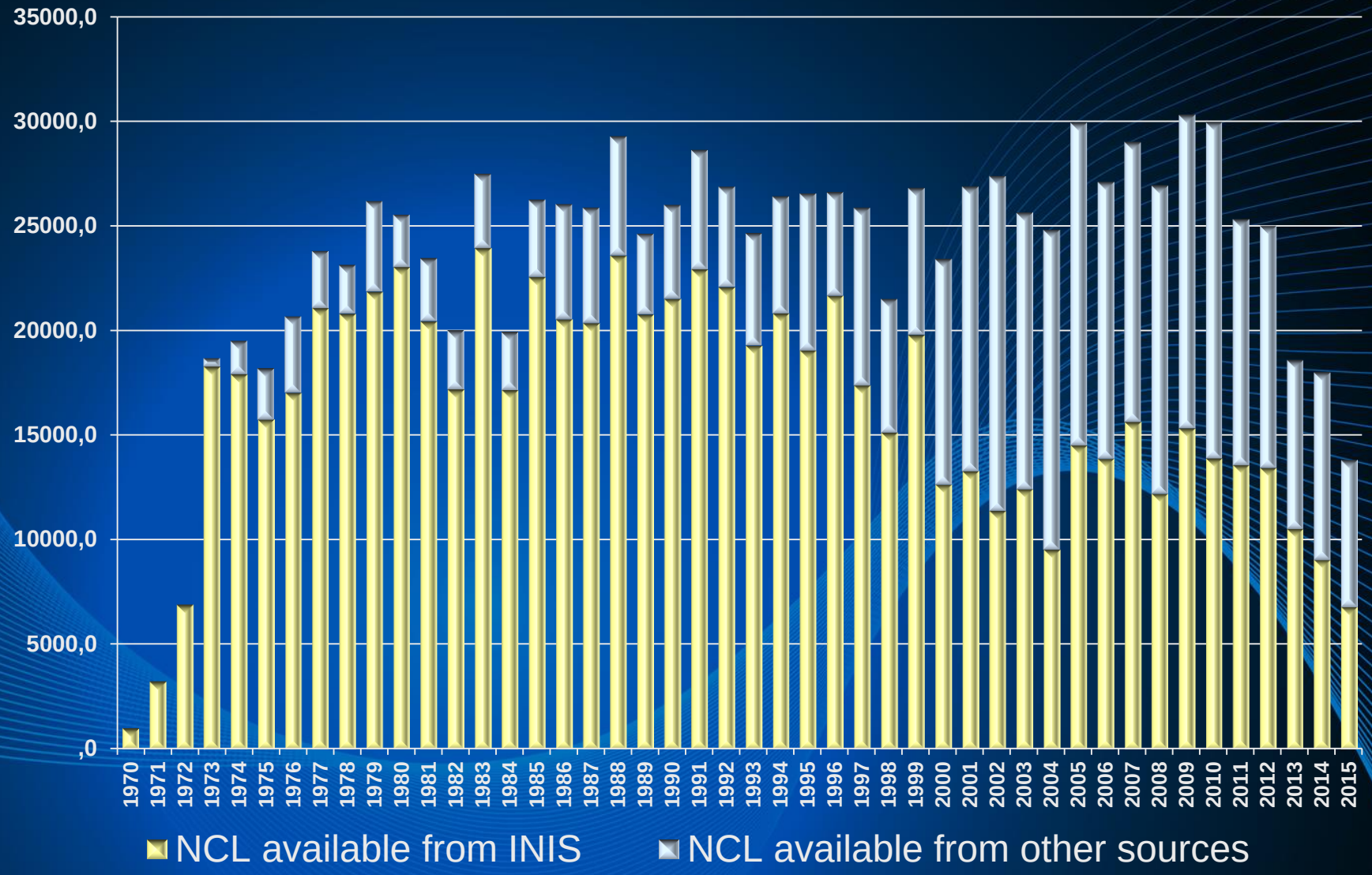
- 1970 First Atomindex in printed and electronic form. NCL distributed on microfiche
- 1971 INIS Thesaurus for document indexing
- 1991 INIS Database available on CD-ROM
- 1997 NCL full-texts on CD-ROM. Start of electronic document delivery service
- 1998 INIS Database on Internet (for subscribers only)
- 2004 Start of Computer-assisted indexing (CAI)
- 2009 Open Access of INIS Database on Internet
- 2011 Google-based INIS Collection Search (ICS)
- 2012 INIS Products made Unicode compliant
- 2013 Publicly accessible PDFs searchable through Google.com and Google Scholar
- 2014 Link to INSPIRE HEP, OCLC's WorldCat service, IAEA Library catalogue and NUCLEUS databases through ICS interface
- 2015 Over 190,000 page views and 230,000 PDF downloads in October 2015

INIS Milestones: https://www.iaea.org/inis/about-us/history.html

Non-Conventional Literature in INIS

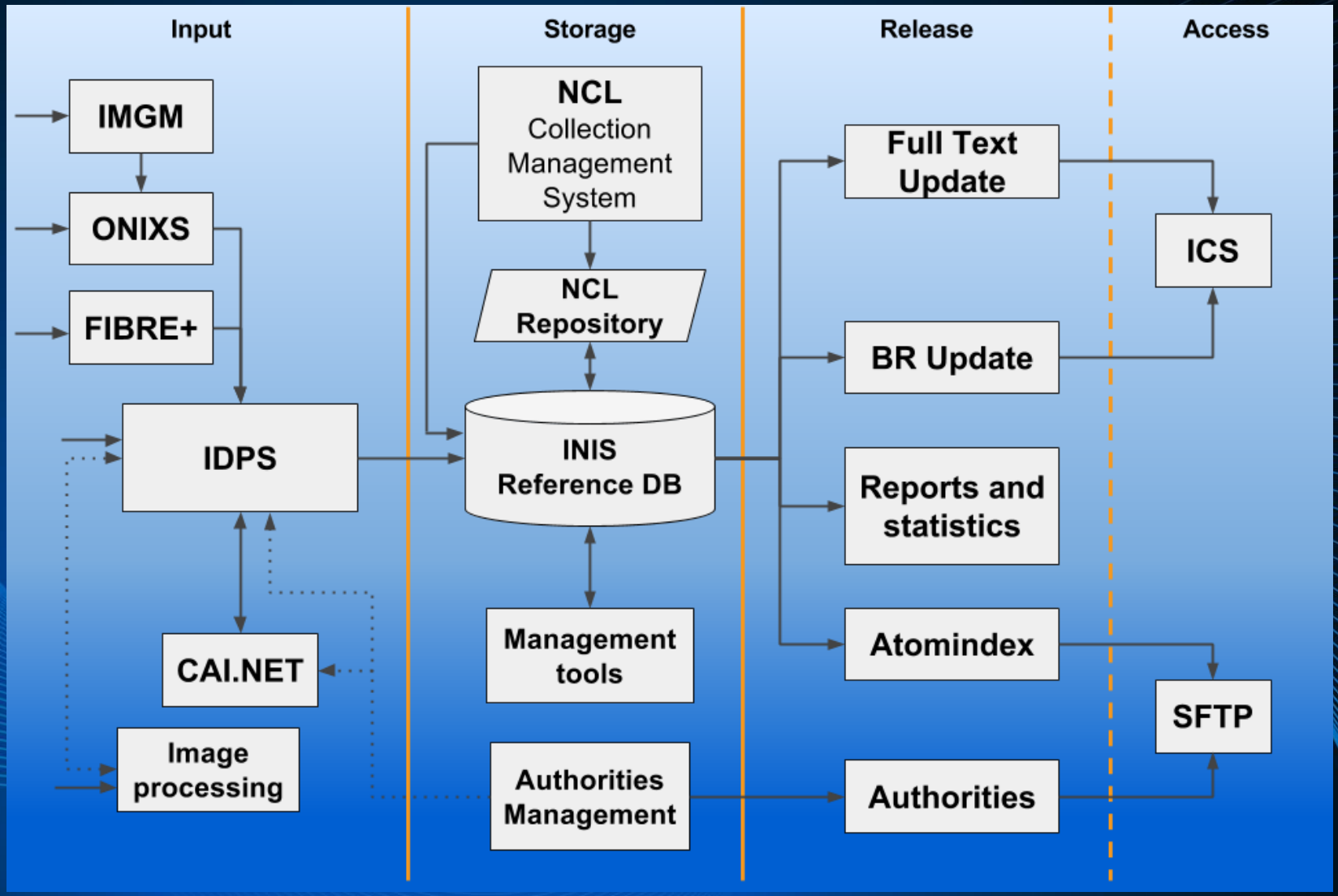
- 3,86 million bibliographic records in the INIS Collection
- 27% (over 1 million records) are NCL
- > 750,000 NCL are *“Available from INIS”*
- > 515,000 full-texts in PDF
- > 371,000 full-texts are PUBLIC

Non-Conventional Literature in INIS 1970 - 2015



IAEA

INIS System Architecture



NCL Processing and Distribution (1970 – now)

<table>
<tr><th></th><th>Production System</th><th>NCL Processing</th><th>NCL Distribution</th></tr>
<tr><td>1970-1996</td><td>Microfiche-based
(B&H, TDC Documate, Anacomp)</td><td>Conversion from paper to microfiche
(Note: NCL from U.S.A. and Japan received on microfiche)</td><td>Diazo duplicates</td></tr>
<tr><td>1997-2003</td><td>INISIS
Image-based
(JouveScan)</td><td>Conversion from paper to TIFF</td><td>NCL CD-ROM with INISIR viewer</td></tr>
<tr><td>2003-2009</td><td>INISIS2K
(InputAccel)
IDPS (Livelink)</td><td rowspan="3">Conversion from paper to PDF + OCR
Ingestion of PDF files</td><td>NCL CD-ROM
Online access
(subscribers only)
E-mail / FTP</td></tr>
<tr><td>2010-2013</td><td rowspan="2">NCL Collection System
(PixEdit + ABBYY FineReader)</td><td>NCL DVD
Free online access
E-mail / FTP</td></tr>
<tr><td>2014 to present</td><td>Free online access
E-mail / FTP</td></tr>
</table>

INIS NCL Collection on Microfiche (1970-1996)

- 312,000 NCL reports
- 500,000 bibliographic records
- Conversion from paper to microfiche and diazo duplication after database production
- "NCL Check" ensured accuracy of the "eye-readable" microfiche header information
- Over 1 million master microfiches (≈ 17 million pages)
- Diazo duplicates sent to customers

INIS Imaging System (INISIS) (1997-2003)

- Image-based production system developed by Jouve
- Imaging started only after database production
- B/W scanning, image enhancement, link to bibliographic metadata using bar codes, validation and CD image creation
- Format supported: single-page TIFF Group 4
- 2 b/w Fujitsu production scanners, 5 workstations, 2 servers

INIS Imaging System 2000 (INISIS2K) (2003-2009)

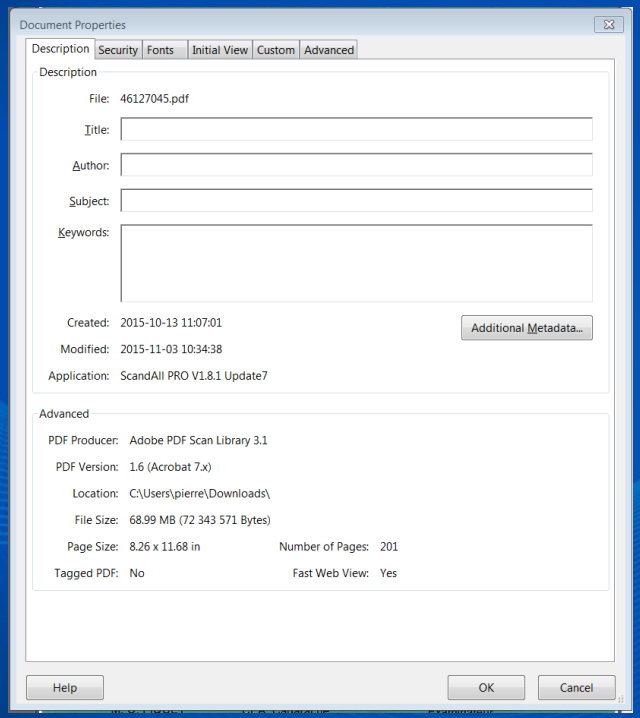
- 32-bit modular Capture System (EMC InputAccel)
- B/W, Greyscale and Color Scanning, OCR, output to PDF
- Integration with Livelink-based INIS Data Processing System (IDPS)
- Bibliographic Control of NCL from PDF output
- 2 B/W Fujitsu production scanners, 1 Kodak i260 color scanner, dedicated workstations for image enhancement and OCR

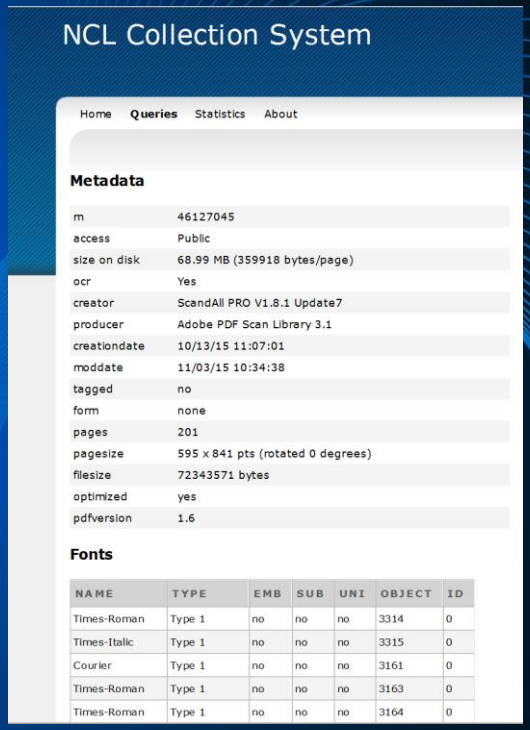
Current Technical Infrastructure (2010 to present)

- Flexible tools to support different projects:
  - PixEdit 8 (scanning, image enhancement)
  - ABBYY FineReader Corporate Edition 12 (multilingual OCR)
- 2 scanners (Fujitsu fi-5750c with VRS, Kodak i1440)
- 2 Sunrise microfiche scanners
- 4 workstations with quad-core processors

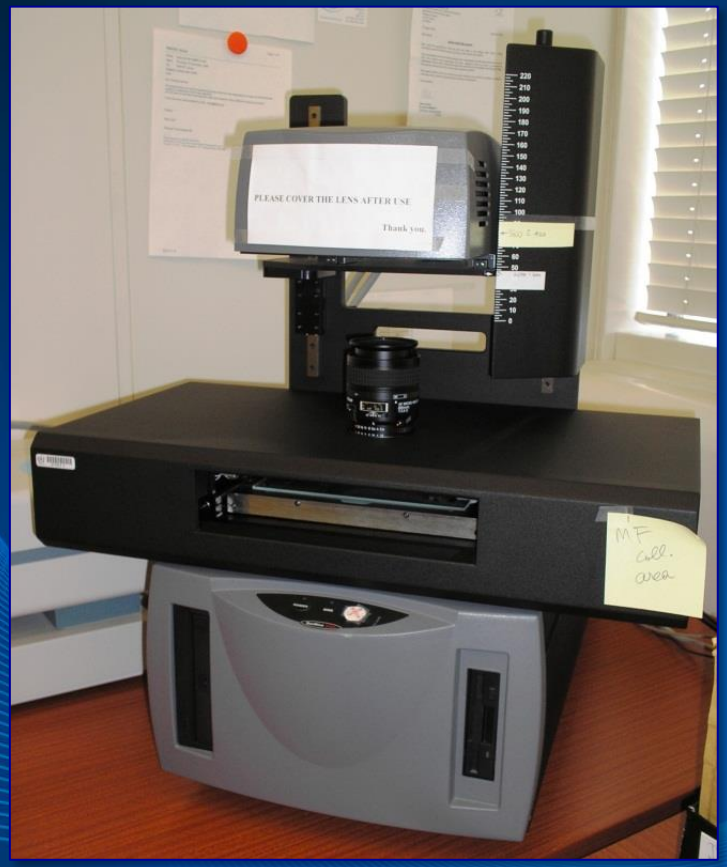
NCL Collection Management System

- Strong web-based data management system
- PDF Metadata is collected during ingestion into the NCL Collection
- Helpful to detect problems, e.g. PDF files needing OCR, poor PDF creation tools





Microfiche Digitization Project



- Project started in 2002 as support to document delivery service
- In close collaboration with the INIS National Centres to avoid duplication of work
- Scanning part was outsourced to 4 different companies
- Image enhancement, OCR and output to PDF done in-house

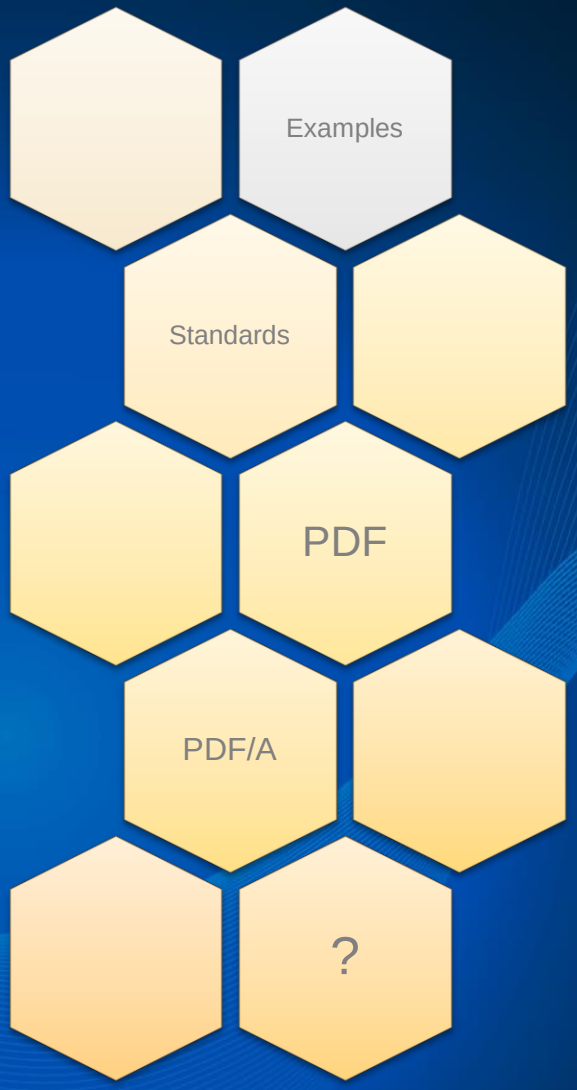
Support / Take Advantage of Digitization Initiatives

- Win-Win strategy. Some examples:
- Digitization of old collections from Mexico, Sweden, the Netherlands and Serbia
- CEA, France: Ongoing digitization of old theses from hard copy
  - Availability of full-texts in PDF adds new value to the INIS bibliographic records
  - PDF scanned from paper may replace the version scanned from microfiche in our NCL Collection
  - If gaps are identified, new INIS records are created
- US DoE: Digitized reports (from paper) available from SciTech are harvested to avoid duplication of work and speed up the digitization of our microfiche collection

Digitization of IAEA documents

- Agreement with IAEA Publications to digitize and make available on-line out-of-print books published before 2000: Proceedings Series, Safety Series, IAEA Bulletin
- Digitized old documents from the IAEA Board of Governors
- Digitized old documents of the International Nuclear Data Committee (INDC) for the IAEA Nuclear Data Section

Practical Session



For more information

- INIS Newsletters: *https://www.iaea.org/inis/products-services/newsletter/*
- INIS Training Seminar Presentations: *https://www.iaea.org/inis/events/training/TS-2015/Presentations/index.html*
- INIS Publications: *https://www.iaea.org/inis/products-services/publications/index.html*

Conclusions

- Digitization projects require:
  - serious planning
  - substantial funds
  - qualified staff (internal or external)
  - awareness of standards
  - well defined purpose
- Maintenance of high quality digitization process is a must
- Document preparation, selection of proper scanning techniques, type of equipment, and adherence to current standards, determine digitization success or failure
- Digitization should not be a goal in itself - its ultimate use and usefulness must always be taken into account
- Meaningful and searchable metadata should accompany any digitized collection, making online search and retrieval efficient
- Long term preservation needs to be considered in order to ensure future sustainability and availability of the digitized collection

Thank you !