PDF Week 2025 Recap

Workshop,

PDF Days Public Conference

and

ISO meetings

Herm Fischer
Exbee Ltd

Workshop (2 days)

- PDF Ecosystem
  - History, standards and future
- PDF as page description language
  - Why PDF
  - ISO
  - Data model, syntax and grammar
  - Validation
- Correctness
  - Lexical
  - PDF versions
  - Data integrity, threats and forgery
- Living standard, future
  - Near term changes (compression & objects)

(slide deck available)

PDF Days (2 day public conference)

- Keynote: “Role of AI in content creation, management and authenticity”
- PDF Tables – difficulties in auto-recognizing tables from span/embedded text
- Collaboration PDF
  - It was not google-docs like collaboration but AI mgmt. of team instead
- HTML representation in PDF
  - Some easy map up, list semantics, table semantics
- Forensics
  - PDF embeds change logs
  - Cache
- Validation and threats
  - PI and conf information not removed
  - Malicious content
  - Spec violations which happen to work in some but not all tools

My Poster - PoC for XBRL

PDF Days Europe 2025

PDF/A Proof of Concept for XBRL

Herm Fischer – Exbee Ltd

XBRL is a global standard for financial and prudential reporting • PDF is a globally adopted format for information sharing

Current Practices

Tagged HTML visuals linked to XBRL

- Viewer — tool required
- AI Ingestion — facts, inconsis-tent semantics

Cost can vary depending on filer needs and complexity
Processes can involve multiple steps
Semantics may not be consistently captured

Tagged HTML links visuals to XBRL but requires a viewing tool. Inline XBRL can present scaling challenges for large, table-heavy filings. AI ingestion often involves multiple conversions and may not always fully reflect filer intent due to XML taxonomy constraints.

Exploring New Possibilities

PDF/A Inline XBRL
Layout
Structure
Embedded XBRL JSON

- ✅ Viewable
- ✅ AI Ingestible

The PDF/A proof-of-concept investigates whether visual layout and filer intent might be better preserved by mapping document structure to XBRL facts and embedding XBRL JSON data and semantic JSON taxonomy within the file.

Proof of Concept

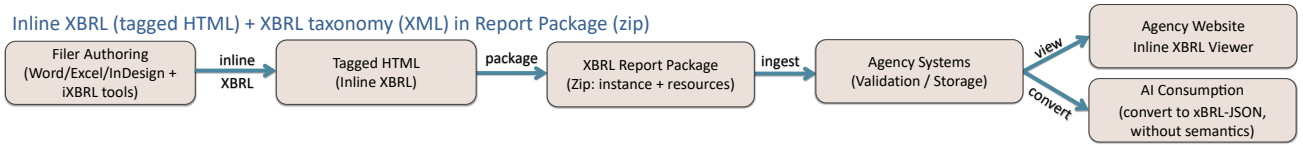
- Arelle and PDF plugins

Use Cases

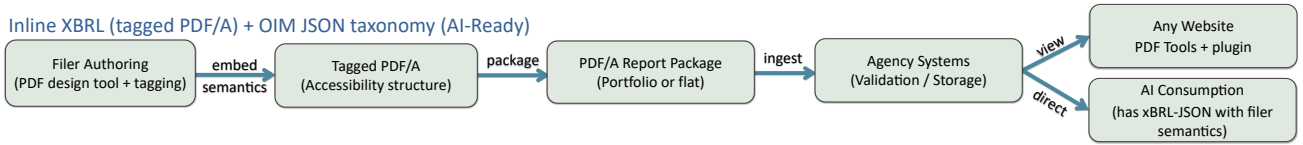
- Explore tagged PDF as an additional submission option
- Pre-tagged PDF forms
- Large tables (xBRL-CSV)

Tagged HTML vs PDF Submissions

Inline XBRL (tagged HTML) + XBRL taxonomy (XML) in Report Package (zip)



Inline XBRL (tagged PDF/A) + OIM JSON taxonomy (AI-Ready)



Key differences: • PDF/A structure + XBRL JSON inside → fewer conversions • Pixel-perfect rendering. • Easier AI ingestion

ISO Working Groups

- WGs on standards extensions
  - Major participation on HTML mappings
  - Personal contacts for low level integration

Key Take-aways

- Conf focus on AI for creating and managing PDF
  - XBRL focus is AI **consuming** reports from multiple sources (in PDF?)
- Multiple presenters focus on PDF embedding semantics:
  - Invoices (and other “GL” like applications)
- Feeling that PDF will subsume semantic documents if left on their own
  - Nobody will need us if we self-obsolesce ourselves
- Clear request to use current standards technology
  - Forget CSV -> use CBOR array streaming
  - Avoid JSON -> use CBOR object (streamable)
  - Consider XBRL JSON semantics using PDF object model instead
- EDGAR realization
  - We MUST validate PDFs and apply EFM 5.5 like criteria to PDFs

My next steps

- PDF Tool Plugins
  - Both Acrobat and Foxit GUI Tools
  - XBRL Viewer pane on tagged PDF (like Arelle Viewer pane)
- Arelle Viewer addition for PDF
  - Intermix PDF and HTML tabs
  - (PoC from Madrid meeting)
- OIM Taxonomy Integration for AI consumability
  - Required that major LLMs to discover XBRL semantics on loading PDF