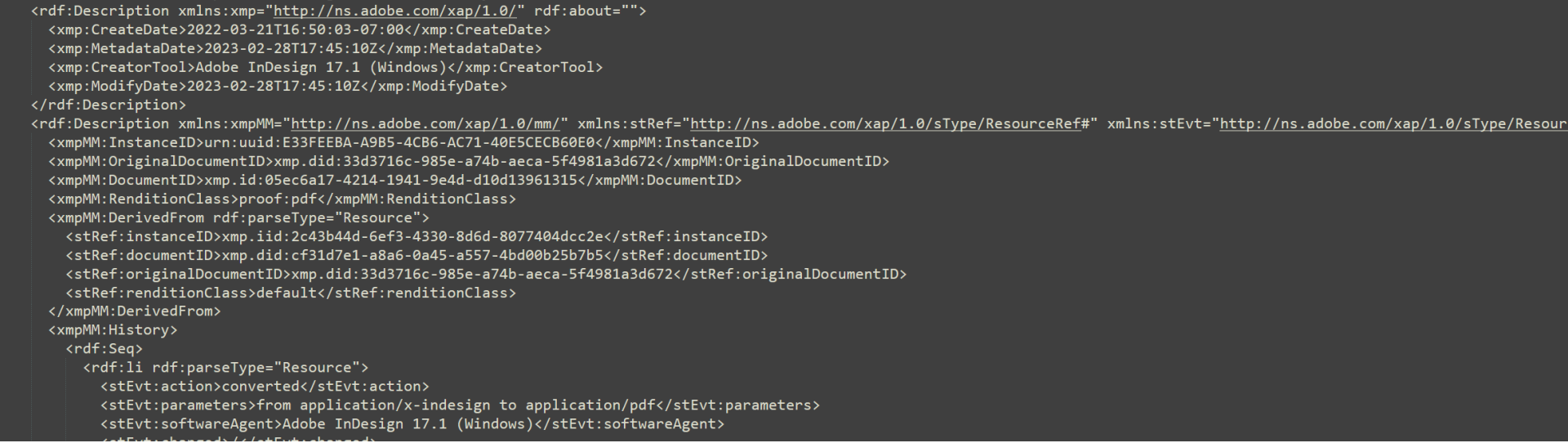


PDF Days Europe 2025

PDF Forensics & the Metadata Conundrum

Metadata? What metadata?

Cherie Ekholm | Product Strategy Lead | Verisk

Overview

- Quick intro
- What is PDF forensics?
- What is PDF metadata? What types are there?
- How is PDF metadata used in forensics?
- How can the PDF community better support forensics?
- What are some tools for looking at PDF metadata?

Quick bio

- Product Strategy Lead at Verisk focusing on Digital Media Forensics (Images and PDF)
- Co-Chair of the PDF Forensics LWG
- Former Project Leader for ISO 14289 and 32000
- 8+ years at Microsoft as a Software Lead in the Office Division focusing on international standards, including PDF, ODF, OOXML, WCAG, ATAG, and UUAG (and other Microsoft roles before that)
- Author, editor, publisher
- Want to hear about my dogs?

PDF Forensics? What's that?

- Involves the examination of PDF files or groups of PDF files to detect possible anomalies used for deceptive purposes
- Files to examine may include invoices, driver's licenses, passports, receipts, reports, or any other kind of document that might be shared from one person or entity to another
- In other arenas it may also include examining PDF files for software security risks
- Common uses for forensics includes detecting fraudulent ID by border agencies or other government agencies, finding altered invoices, sales receipts, or other document for use in insurance claims, banking applications, or court cases, or tracing provenance of files used in almost any situation

What is Metadata

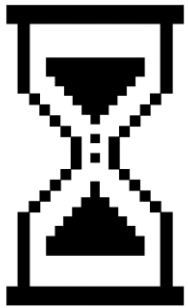
met·a·da·ta [ˈmedəˌdadə, ˈmedəˌdādə]

Noun

1. **a set of data that describes and gives information about other data.**

What metadata is required in 32000?

Section 14.3 in ISO 32000-2 …



No, really. What metadata is required in 32000?

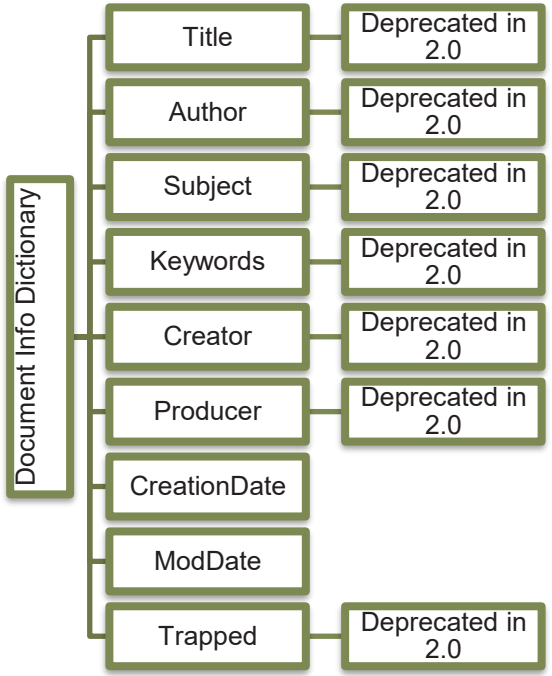
- If you just read the metadata sections of the standard (ISO 32000-2, 14.3, there is no metadata that is required in 32000-1 or 32000-2.
- However, metadata available in a PDF or that can be specified in a PDF extends well beyond anything specified in 14.3
- Dictionaries, the header, and the footer all contain a wealth of additional metadata, some required and some optional
  - PDF version number
  - DocumentID and InstanceID
  - Linearized
  - Marked content
  - ICC Profile info
  - Font info
  - Etc.

Metadata is not the whole answer

- Minimal metadata can’t tell us a file is fraudulent or even slightly suspicious
- But it’s a really good partner for detecting issues
  - Metadata often shows the first hints of problems in a file
  - Metadata can support other findings
  - Metadata can give us breadcrumbs showing where we need to dig deeper
  - The absence of metadata can undermine other investigations

PDF metadata

- Metadata stored in the **document info dictionary**
  - Title,
  - Author
  - Subject
  - Keywords
  - Creator
  - Producer
  - CreationDate
  - ModDate (Required only if ***PieceInfo*** is present)
  - Trapped
- Metadata is optional and easily editable
- Deprecated in ISO 32000-2 in (except CreationDate & ModDate)
- Equivalents for all exist in XMP metadata



Dublin Core metadata

- Developed by the Dublin Core Metadata Initiative (DCMI)
- Dublin Core is a standardized metadata schema (now) often embedded in PDFs via XMP.
- Among the most commonly used elements in PDFs are:
  - **Title** – Name of the document
  - **Creator** – Original author or entity responsible
  - **Subject** – Topic or keywords
  - **Description** – Abstract or summary
  - **Publisher** – Entity making the document available

```
</rdf:Description>
<rdf:Description rdf:about="" xmlns:dc="http://purl.org/dc/elements/1.1/">
   <dc:format>application/pdf</dc:format>
   <dc:title>
      <rdf:Alt>
        <rdf:li xml:lang="x-default">Annual report 2014</rdf:li>
        <rdf:li xml:lang="en">Annual report 2014</rdf:li>
        <rdf:li xml:lang="de">Jahresbericht 2014</rdf:li>
      </rdf:Alt>
   </dc:title>
   <dc:creator>
      <rdf:Seq>
        <rdf:li>John Doe</rdf:li>
        <rdf:li>Mary Miller</rdf:li>
      </rdf:Seq>
   </dc:creator>
</rdf:Description>
```

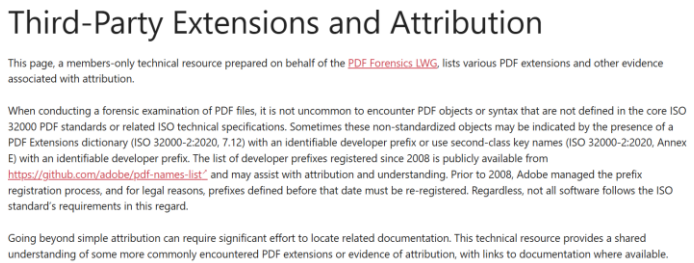
XMP metadata

- Introduced in **PDF 1.4** along with the **metadata stream**
- ISO 32000-2 deprecated use of most of the **document info dictionary** in favor of **XMP** and the metadata stream
- Metadata **can be attached to any PDF object**, not just the document
- Supports **AES-256 encryption** of metadata
- **Aligns** with other ISO standards (PDF/A, PDF/X, PDF/UA), where some metadata may be required
- **Improves interoperability** and indexing for forensic and archival use

```
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="Adobe XMP Core 8.0-c001 79.328176e, 2022/08/01-19:10:29
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
      <rdf:Description rdf:about=""
            xmlns:xmp="http://ns.adobe.com/xap/1.0/"
            xmlns:xmpMM="http://ns.adobe.com/xap/1.0/mm/"
            xmlns:dc="http://purl.org/dc/elements/1.1/"
            xmlns:pdf="http://ns.adobe.com/pdf/1.3/"
            xmlns:pdfx="http://ns.adobe.com/pdfx/1.3/">
         <xmp:ModifyDate>2023-01-10T16:51:47-08:00</xmp:ModifyDate>
         <xmp:CreateDate>2023-01-10T16:51:45-08:00</xmp:CreateDate>
         <xmp:MetadataDate>2023-01-10T16:51:47-08:00</xmp:MetadataDate>
         <xmp:CreatorTool>Acrobat PDFMaker 22 for Excel</xmp:CreatorTool>
         <xmpMM:DocumentID>uuid:3c99cb66-1d68-4d3c-b783-3cedac5ca755</xmpMM:DocumentID>
         <xmpMM:InstanceID>uuid:a6e68fe8-d4ea-403b-a685-e6d2040fa353</xmpMM:InstanceID>
         <dc:format>application/pdf</dc:format>
         <dc:creator>
            <rdf:Seq>
               <rdf:li/>
            </rdf:Seq>
         </dc:creator>
         <pdf:Producer>Adobe PDF Library 22.3.58</pdf:Producer>
         <pdfx:Company/>
         <pdfx:ContentTypeId>0x01010079F111ED35F8CC479449609E8A0923A6</pdfx:ContentTypeId>
      </rdf:Description>
```

Custom metadata

- Metadata fields **defined** by the **document creator or organization**
- **Not** part of **standard schemas** like Dublin Core or PDF Info Dictionary
- Embedded via **XMP extensions** or **custom namespaces** (via the PDF extensibility model)
- May be **inconsistently applied** or undocumented
- **Requires schema awareness** for proper interpretation
- **New guidance** in an application note from the PDF Association: “Including custom metadata structures in PDF”



Third-Party Extensions and Attribution

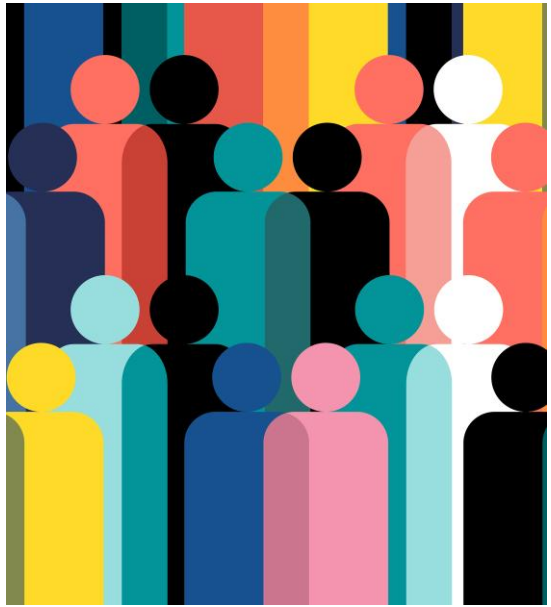
This page, a members-only technical resource prepared on behalf of the PDF Forensics LWG, lists various PDF extensions and other evidence associated with attribution.

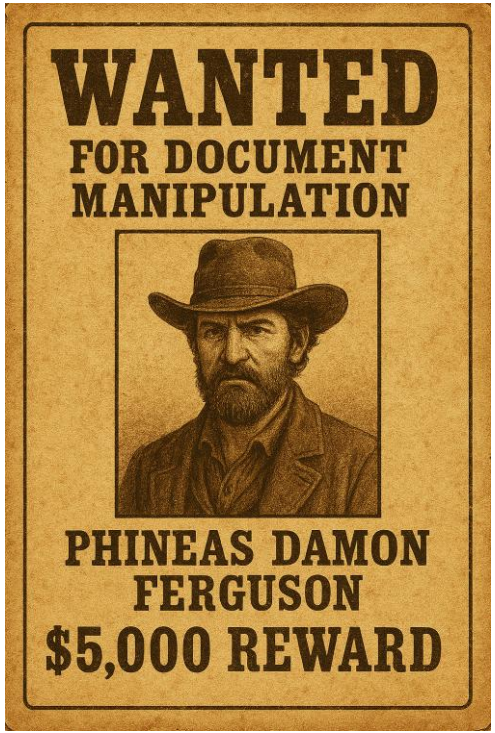
When conducting a forensic examination of PDF files, it is not uncommon to encounter PDF objects or syntax that are not defined in the core ISO 32000 PDF standards or related ISO technical specifications. Sometimes these non-standardized objects may be indicated by the presence of a PDF Extensions dictionary (ISO 32000-2:2020, 7.12) with an identifiable developer prefix or use second-class key names (ISO 32000-2:2020, Annex E) with an identifiable developer prefix. The list of developer prefixes registered since 2008 is publicly available from https://github.com/adobe/pdf-names-list/ and may assist with attribution and understanding. Prior to 2008, Adobe managed the prefix registration process, and for legal reasons, prefixes defined before that date must be re-registered. Regardless, not all software follows the ISO standard's requirements in this regard.

Going beyond simple attribution can require significant effort to locate related documentation. This technical resource provides a shared understanding of some more commonly encountered PDF extensions or evidence of attribution, with links to documentation where available.

https://github.com/adobe/pdf-names-list

Does PDF have a rich, robust community?

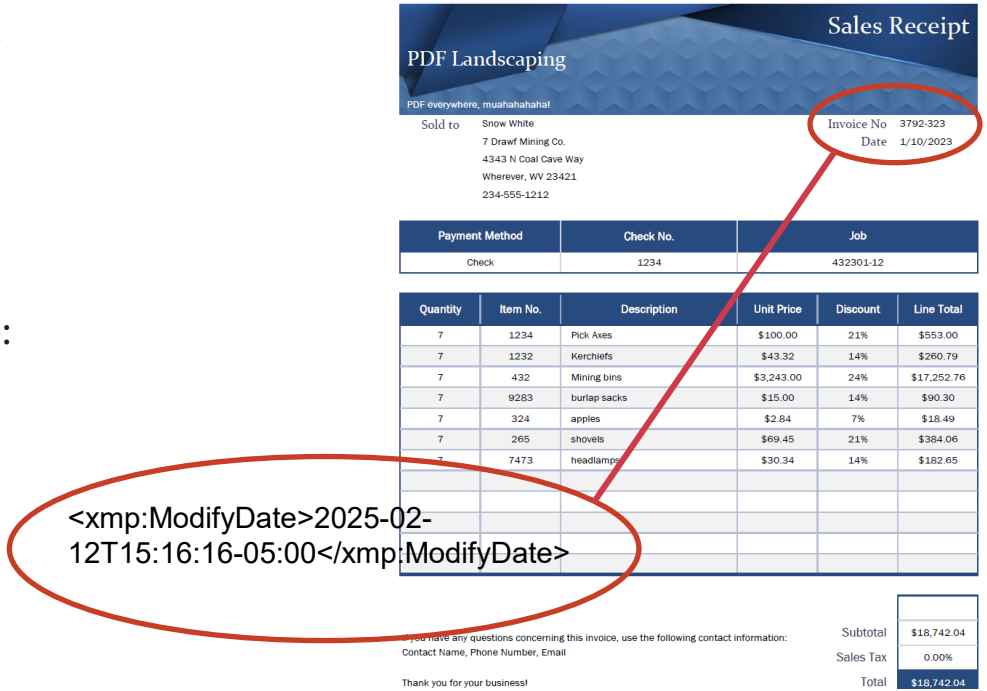




Or is PDF the wild, wild west for forensics?

Why is metadata important for Forensics

- Helps forensic analysts detect inconsistencies
  - Mismatched authorship
  - Implausible editors
  - Mismatched dates
- Metadata can be compared to:
  - Other metadata in the file
  - Document contents
  - Claim information
  - Related documents
  - Other resources



How can the PDF community better support forensics?

- Output some
- Output it consistently
  - Output for every file you create or edit
  - Understand the meaning of the fields and use them in an interoperable manner
- Output it when editing content from another creator/producer
- DO NOT overwrite metadata that is not yours to change

So what metadata should be considered

- **Dates**
  - Of the file itself
  - Of actions taken on the file
  - Of any entities contained in the file
- **Info on the tool(s) used to create or modify the file**
  - What company or person created the file?
  - What tools were used to make edits to the file?
  - What tools were used to remediate or repackage the file?
- **Info on actions taken on the file**
  - Was it scanned? Reformatted from another file type?
  - Was it OCRed? Redacted? Edited in any trackable way?
  - Was it reordered? Had pages added? Recombined?
- **Info on any entities contained within the file**
  - Fonts
  - ICC profiles
  - Signatures
  - Associated files

Minimum output for newly created files

- CreateDate
- ModifyDate
- CreatorTool
- Producer
- DocumentID
- InstanceID
- Params dictionary metadata for File Attachment, Embedded Files, etc.
- Form-related XMP
  - Acroform
  - Has_XFA
  - Form action fields
- Signature info
- Future: C2PA manifest

```
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="Adobe XMP Core 9
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-synta
      <rdf:Description rdf:about=""
            xmlns:xmp="http://ns.adobe.com/xap/1.0/"
            xmlns:dc="http://purl.org/dc/elements/1.1/"
            xmlns:xmpMM="http://ns.adobe.com/xap/1.0/mm/"
            xmlns:pdf="http://ns.adobe.com/pdf/1.3/">
         <xmp:ModifyDate>2025-08-25T17:27:07-04:00</xmp:Modif
         <xmp:CreateDate>2024-08-20T14:16:34-04:00</xmp:Creat
         <xmp:MetadataDate>2025-08-25T17:27:07-04:00</xmp:Met
         <xmp:CreatorTool>Created PDF (32-bit) 2.20895</xmp:C
         <dc:format>application/pdf</dc:format>
         <dc:creator>
            <rdf:Seq>
               <rdf:li>Smythe, Joan</rdf:li>
            </rdf:Seq>
         </dc:creator>
         <xmpMM:DocumentID>uuid:662db7b0-6e83-46f7-bea6-b417e
         <xmpMM:InstanceID>uuid:b79e6a16-ac7a-4bf6-a686-8db44
         <pdf:Producer>Created PDF (32-bit) 2.20895</pdf:Prod
      </rdf:Description>
   </rdf:RDF>
</x:xmpmeta>
```

Minimum effort when editing files

DO NOT MODIFY ON FILE EDIT

- **CreatorTool** – This should remain the tool that originally created the file
- **DocumentID** – The DocumentID should remain static once created for the document

DO MODIFY ON FILE EDIT

- **Producer** – This should be the name of your tool that has done the editing. Preferably you should append your tool name to the existing Producer info – For instance, “TheirTool modified by MyEditor”
- **InstanceID** – incremented from the original InstanceID
- **ModifyDate**

ADD ON FILE EDIT

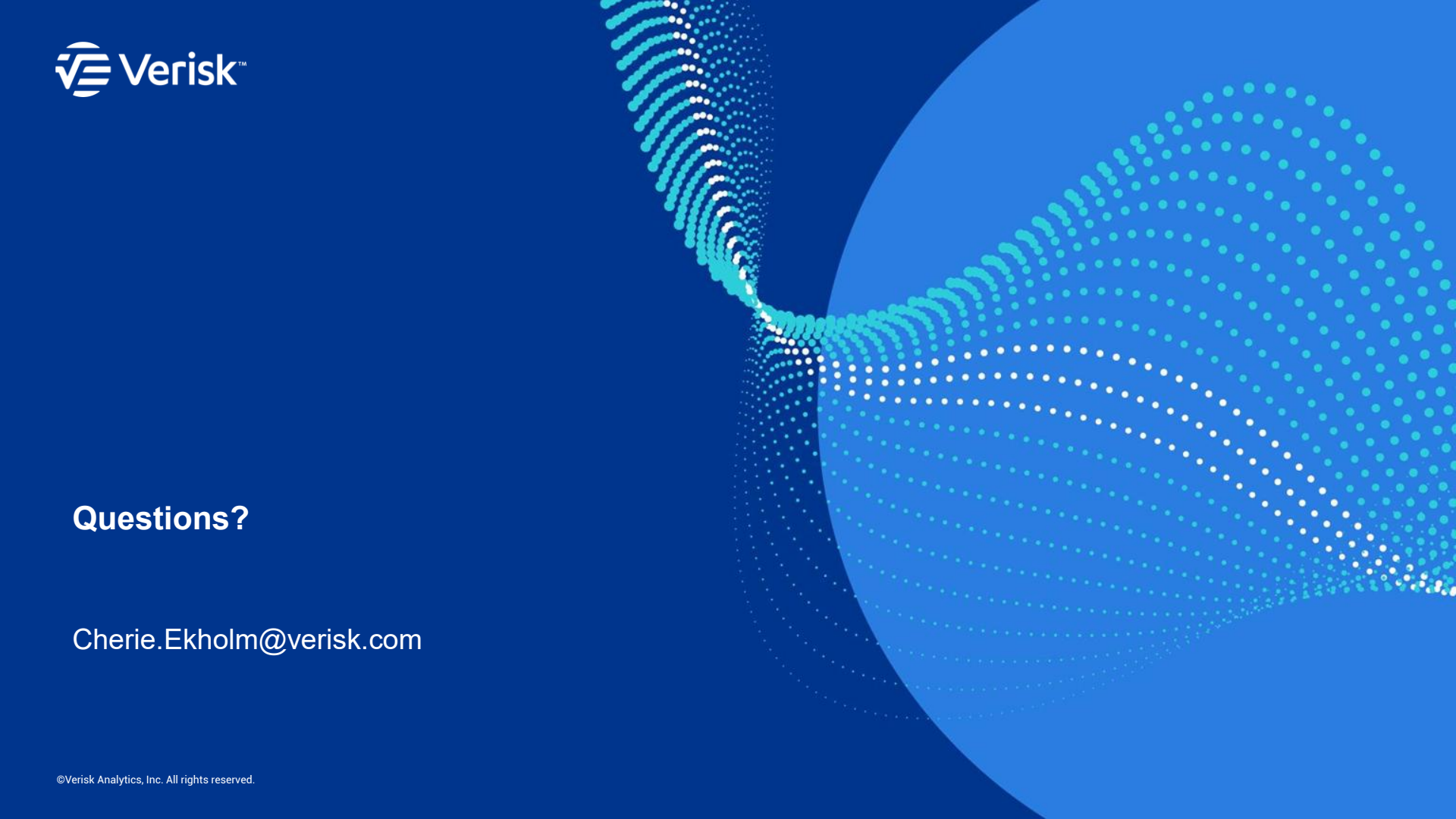
- **History Actions**
  - History
  - HistoryAction
  - HistoryChanged
  - HistoryInstanceID
  - HistoryParameters
  - HistorySoftwareAgent
  - HistoryWhen
- **Derived from Actions**
  - DerivedFromDocumentID
  - DerivedFromFilePath
  - DerivedFromInstanceID
  - DerivedFromLastModifyDate
  - DerivedFromLastURL
  - DerivedFromLinkForm
  - DerivedFromOriginalDocumentID
  - DerivedFromRenditionClass
  - DerivedFromRenditionParams

Tools for extracting metadata

- PDF extraction tools that can expose fonts, structures, metadata
  - ex. pdfinfo from XpdfReader
    - Good for extracting the bare minimum pdf-specific fields
- General purpose file information extraction tools
  - ex. Apache Tika
    - Good for extracting an extended list of metadata for multiple formats including PDF
    - Good to use as a wrapper with other tools
  - ex. Exiftool
    - Good for extracting the largest set of metadata from several formats including PDF

Where do we go from here?

- Define a minimum set of metadata and publish it to the community
- Better define what the various metadata fields mean
- Encourage PDF producers to output this minimum set of metadata
- Update PDF Next to include the minimum set of metadata
- Urge PDF producers to update metadata properly



Questions?

Cherie.Ekholm@verisk.com