Adobe Enterprise Solutions

PDF/A, XMP, and a Metadata Registry Use Case

This paper covers three topics:

1. An overview and progress report of the PDF/A standard for archival.
2. A description of Adobe's eXtensible Metadata Platform (XMP) and how it relates to a registry.
3. An Federal Enterprise Architecture (FEA) based use case for Adobe's registry.

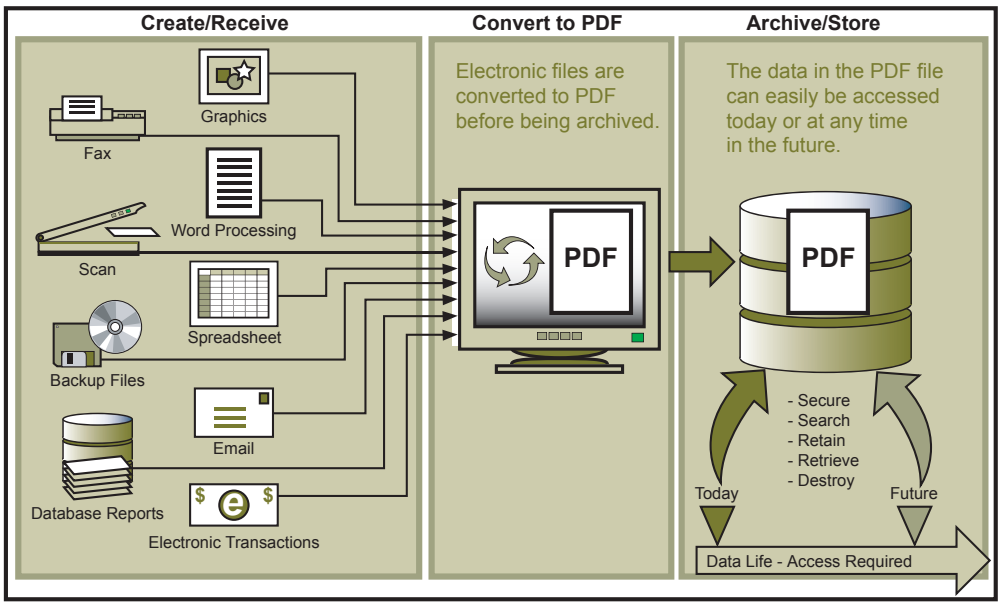
Part One - PDF/A for Archiving

PDF/A is a standard being established to set guidelines for archiving and preserving digital documents in Portable Document Format (PDF). PDF/A-a joint initiative of NPES the Association for Suppliers of Printing, Publishing, and Converting Technologies and the Association for Information and Image Management International (AIIM)-will address the growing need to electronically archive documents in a way that will ensure the preservation of their contents over an extended period of time and that will ensure that those documents can be retrieved and rendered with a consistent and predictable result in the future. For more information about the NPES/AIIM project, visit www.aiim.org/pdf_a.

Why PDF?

There are many electronic formats and technologies to choose from for archiving. These include ASCII (for text), TIFF, PDF, and XML-not to mention word processing, spreadsheets, and other formats. The proprietary nature of some of these formats leads to the criticism that they cannot be guaranteed to continue for the long term. Only one of these formats is uniquely suited to ensuring display preservation over a long period of time. PDF represents not only the data contained in the document but also the exact form the document took. The file can be viewed without the originating application. In fact, ten years from now, and into the future, users will still be able to view the file exactly as it was created.

With the addition of XML metadata to the PDF file, we can have both fidelity and accessibility. Because PDF is a publicly available specification, the information about the file format will always be in the public domain, making it a very attractive format to select for electronic archives.



PDF/A Workflow

Moving from PDF to PDF/A

PDF/A is a subset of the PDF format. PDF has been restricted in several ways to enhance it's suitability for archiving.

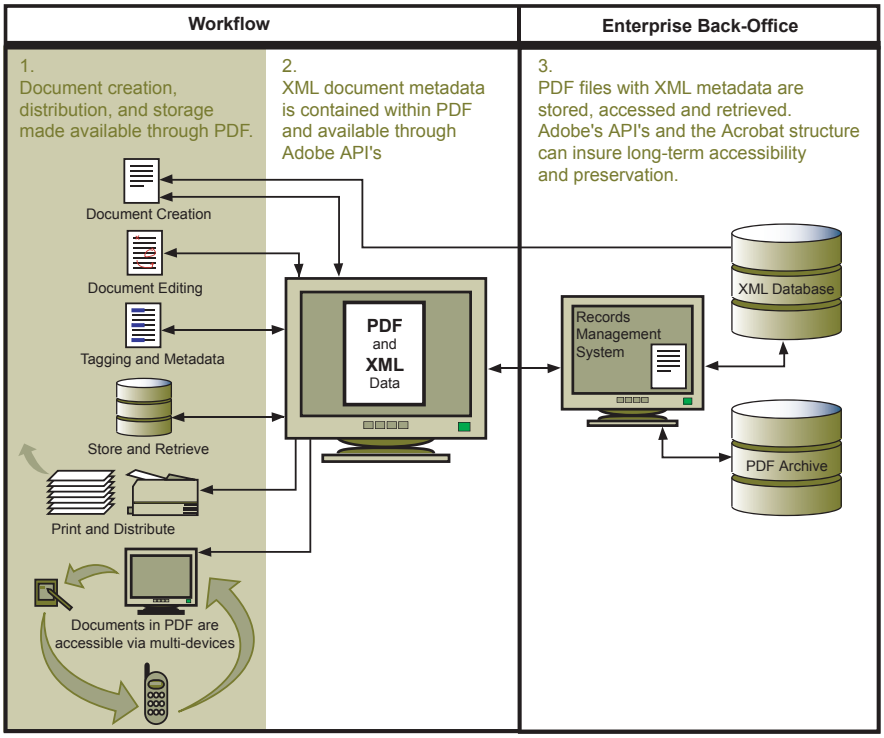
1. **Audio and Video** - Embedded audio and video rely on proprietary decoding software and are not a native part of the PDF format. They should not be included.
2. **Encryption** - Encryption algorithms may not be used since they may prohibit future access.
3. **Compression** - Non-public domain compression methods may not be used since they may leave the file unreadable without proprietary software.
4. **Fonts** - All fonts must be embedded in the document to ensure that they will be available. Proprietary fonts that require licensing may not be used.
5. **Links** - Links to other documents outside of the scope of the archive are not permitted.
6. **Scripts and Executables** - Scripts and executables may rely on a particular platform or virtual machine and should not be used.
7. **Metadata** - Standards for metadata elements should be set, and metadata should be required.

PDF/A and Metadata

In order to create electronic archives, emphasis must be given to the creation of metadata. Metadata means "data about data". It is used to describe information about an object. PDF metadata uses Adobe's XMP (eXtensible Metadata Platform). XMP is based on Resource Description Framework (RDF). RDF allows metadata schemes to be read by humans as well as parsed by machines. The underlying XML simply requires that all namespaces be defined. Once they are defined, they can then be used to the extent needed by the provider of the metadata. Because metadata is in XML format, it can be extended and modified using third-party products.

Part Two - XMP (eXtensible Metadata Platform)

The eXtensible Metadata Platform (XMP) is a framework for adding semantic content to application files, databases and content repositories. Adobe's XMP is one of the first major, comprehensive implementations of Resource Description Framework (RDF). RDF is used to structure the labels for metadata. For machine reading, RDF is implemented in XML expressions.



XMP Workflow

The elements of the Adobe XMP platform are:

1. RDF Framework for expressing metadata from multiple schema - XMP Framework
2. Schema used to describe properties, contained in namespaces - XMP schema
3. Method for embedding XML fragments in binary streams - XMP Packet Technology
4. Support for third party interface and extensions to XMP - XMP SDK

Document Structure

In practice, documents are composed of sub-documents as a page in a magazine is composed of multiple images and sections of text. Adobe's Metadata Framework respects this operational reality: Where a document is assembled from sub-documents, each containing a metadata label, the sub-document label is preserved in the containing document composed of sub-documents.

The RDF rules specify the composition of labels into a sequence of XML statements called a "triple" of data. A triple contains a resource, a property, and a value (alternatively called subject, predicate, and object). The schema expressed with RDF define the vocabularies used in the labels. RDF schema are collections of attribute and corresponding value types that can be specified, allowing for the creation of labeling systems that are appropriate for a particular domain of knowledge.

For the Adobe Metadata Framework, Adobe has created the "Standard XMP Schema." These schema are a starting point, but critical to the value of the XMP framework is the ability to include any RDF schema, provided it is defined according to the specification. Once the new schema is in place, custom property value pairs can be added to the data in the XMP packet, and they will be respected by all XMP processing routines just as if they were property value pairs corresponding to standard schema. In some cases the same concepts are used but the names for the properties are different. XMP provides for this by supporting aliasing from a non-standard name to a name in one of the standard schema.

XMP, Content Management Systems, and Registries

XMP can currently be used in conjunction with content management database systems. In the future, Adobe's XML registry will be able to read XMP directly from incoming documents, automatically populating standard registry metadata fields.

Part Three - FEA Registry Use Case

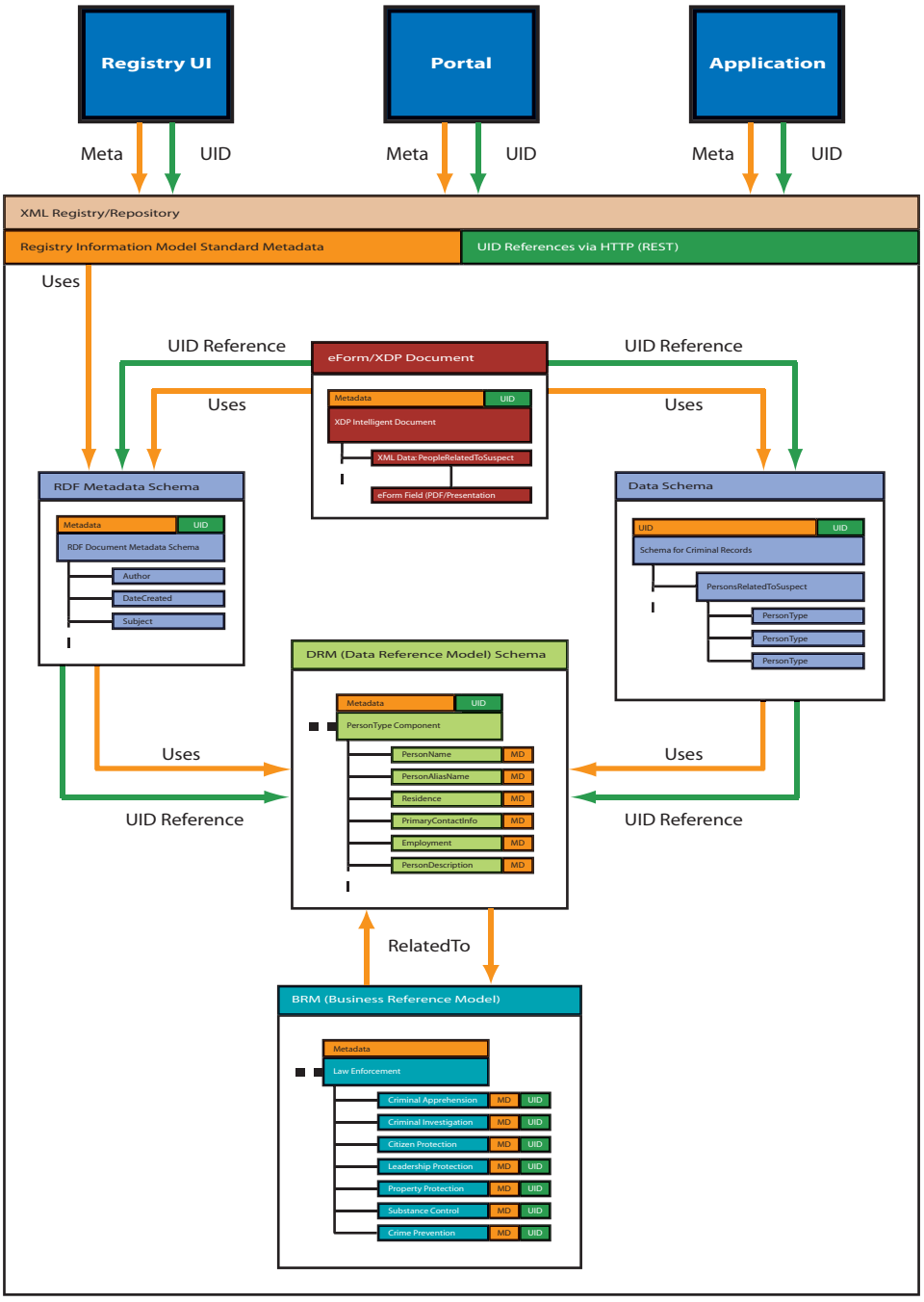
The concepts of the Federal Enterprise Architecture present a challenge to registry and metadata tools in their current state. Mapping reference models, business processes, and interfaces and deploying them in the real world with today's technologies is a daunting task. Adobe's Intelligent Document Platform (IDP) and Metadata Registry technologies can significantly close the gap between concept and implementation.

Mapping the BRM (Business Reference Model) and DRM (Data Reference Model)

The Business Reference Model (BRM) defines and compartmentalizes the business model for the government. It defines the roles undertaken by the government, as a whole, and maps these roles to individual agencies. **The Data Reference Model (DRM)** classifies schema, schema subcomponents, and base datatypes that are used to create a standard vocabulary for the business processes used within and between government agencies, as well as citizens and businesses using government services. The BRM and DRM can be mapped to each other inside a registry by breaking them down into components. Registry metadata can be used by each model to reference the other's components.

The hierarchal structure of the BRM translates readily to a registry, allowing users to easily find registry items by their business type. BRM functions and sub-functions become nodes and sub-nodes in the registry. A single item (schema, component, document) inside the registry can be referenced by both it's relationships within the BRM (such as which business processes use it), as well as inside the DRM (by what components and datatypes it uses, or use it). This enables both a logical, context-based, human readable hierarchy as well as machine readable semantic linking.

In addition, by using references to an item's Unique Identifier within the registry, business processes which are not able to communicate with the registry semantically are still able to link to registry items directly. This two-tiered access model allows for a great degree of flexibility when it comes to implementing a business process and offers compatibility with both schema-based processes, and emerging component assembly solutions.



Use Case Example

A user or application needs to obtain information concerning the metadata of a business document:

For the user, this can be accomplished by visiting the registry and either searching for the document by name or metadata, or by browsing the BRM in the registry and diving down through the functional categories of the model to find the specific business process. The user can also view, by association, all other documents and components that have any relation to their particular item. This creates a logical, visible chain from document to schema to component to datatype.

An application also has two ways to access metadata from the registry. If the application is semantically enabled, it can query the registry using metadata elements specified by the registry's information model. Otherwise, the application can take advantage of each registry item's Unique Identifier (UID) and address content directly. In this case, the XMP's schema reference can include the full registry address and UID of the schema, allowing it to be accessed from outside the registry via HTTP.