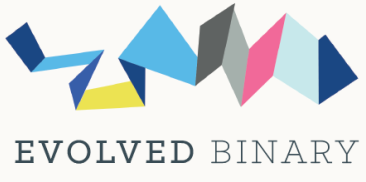
Marymount University 5/12/2016

Digital Preservation and The Digital Repository Infrastructure

Adam Retter

adam@evolvedbinary.com

@adamretter



Adam Retter

Consultant

- Scala / Java
- Concurrency and Databases
- XQuery, XSLT

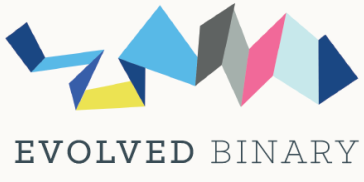
Implementation Lead: The DRI project at The National Archives (UK) 2011-2014

Open Source Hacker

- Predominantly NoSQL Database Internals
- e.g. eXist, RocksDB, Shadoop (Hadoop M/R framework)

W3C Invited Expert for XQuery WG

Author of the "eXist" book for O'Reilly



Talk Disclaimer

1. All opinions are my own!
2. Digital Preservation experience dates from 2011 - 2014
3. Things may have moved on since
4. Some details omitted for security
5. Quickly put together
6. Looking for interaction...

The National Archives

Archive Records of UK from OGDs, NGOs and Special Interest

Excellent at traditional Paper records

- One of the largest collections in the world
- Over 11 million historical Government and Public Records

However, most records today are not created on paper!

- Predicted 2013 - 2020:
  - >6PB of Digital Records to Archive
  - >50% of which will be Born Digital
  - Forecast in 2012, likely increased since!
- 2009: Existing DRS (Digital Records System) will not cope...
  - 2011: Start developing replacement: DRI project

The National Archives

EVOLVED BINARY

Part 1.
Digital Preservation

What is Digital Preservation?

" In library and archival science, digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change. "

" ***The goal of digital preservation is the accurate rendering of authenticated content over time.*** "

- Taken from Wikipedia: https://en.wikipedia.org/wiki/Digital_preservation

EVOLVED BINARY

What is Digital Preservation?

Preservation of a born digital (or digitised) record in the face of:

- File Format Obsolescence
- Software (and Hardware) Obsolescence
- Software and Hardware Failure
- Physical and Technical Degregation / Corruption
- Meeting Sensitivity Requirements (Political/Geo/Human)
- Proving Authenticity
- Providing (meaningful) Access

EVOLVED BINARY

What is Digital Preservation?

No one definition, many Philosophies and open Questions:

- What should you preserve?
- What should you present?
  - Original?
  - Manifestations?
- Emulation vs Migration
  - Software Archive
  - Hardware Archive
  - File Format Selection
  - File Format Risk Identification
  - File Format Transcoding
- Archive of Preservation Software and Config?

EVOLVED BINARY

What is "The Record"?

Influenced by both:

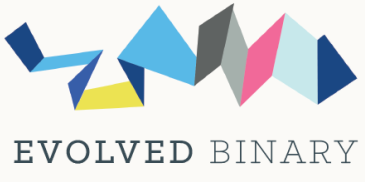
- Organisation Strategy
- The Collection under consideration
  - Physical Considerations
  - Cost/Resource Considerations

Ethereal... however for The National Archives:

- More than just the Digital File
- Metadata
  - Provenance: Source, Transfer, Processing and Accession
  - Technical: Computed Analysis
  - Transcription: Human or Text Extraction
  - Cataloguing
- Manifestations from Migration, Curation, etc.

EVOLVED BINARY

Part 2.
The Digital Repository Infrastructure



What is DRI?

Digital Repository Infrastructure

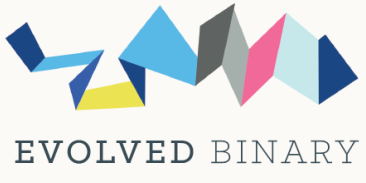
- A new Digital Repository for The National Archives
- 3 Year Project (2011 - 2014)
  - Designed and Developed in house
  - Hardware and Software
- Any File Formats
- Any Metadata (Complex/Structured/Extensible)

Must replace previous DRS (Digital Records System)

- DRS was limited to Collections in the tens of GB (Gigabytes)
- DRI must cope with at least 2PB (Petabytes) per year
- DRI must be able to accession several collections in parallel
- DRI must export Presentation Manifestations to Discovery

EVOLVED BINARY

Security as a Major Factor



Security as a Major Factor

Physical Security

- Dedicated Custom Data Centre separate from Corporate IT
- Data Centre's Physical and Network Organisation was based on Trust Zones
- Policies: Sensitivity Review
- Dedicated Secure Laboratory for Digital Preservation Analysis
- Policies: Access and Handling
- Dedicated Secure Room for Collection Loading

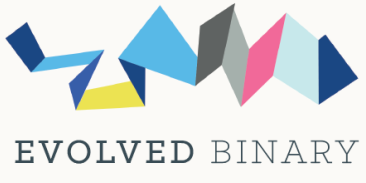
Security as a Major Factor

Technical Security

- Firewalls
- Virus Scanners
- Malware Scanners
- Intrusion Detection
- Access and Authentication
- Encryption
- Network Segmentation
- Physical Separation of Systems and Air-Gaps

Acquisition and File Formats

Acquisition and File Formats



Acquisition and Metadata

Metadata is absolutely essential!

- Allows us to understand the Digital Record
- Collect as rich Metadata as affordable (Cost and Time)
- Minimal Core set required for every accession
- Additional Metadata decided on an accession-by-accession basis (semi-schema free)

Metadata requested by The National Archives

- Is Always in CSV (Comma Separated Value) format with UTF-8
- May be split over several files
- Complex Relationships and Validation are performed using CSV Schema http://digital-preservation.github.io/csv-schema/csv-schema-1.1.html

Acquisition and Metadata

Digitised Records Metadata

- Transcriptions (may be from a external provider)
- File Format Identifation (DROID)
- Extracted Analysis of Image Properties (JHove)
- Provenance recorded from Transfer, and then Digitisation through to Accession

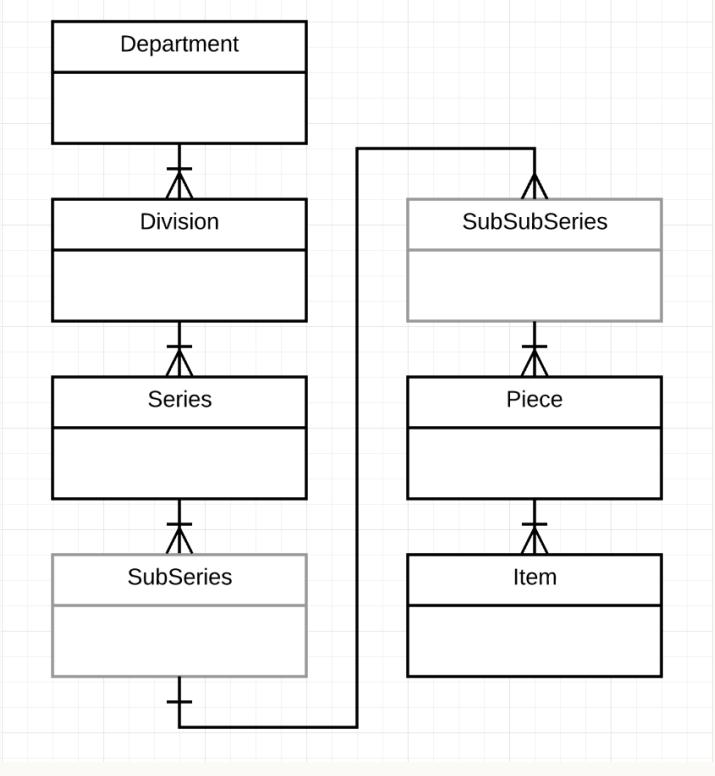
Born Digital Records Metadata

- Transcription is rarer, instead Text Extraction is used
- Fact extraction - Dates/Names/Locations (Gate, Stanbol etc)
- File format identification (DROID)
- File format metadata extraction (e.g. XMP from PDF)
- Metadata Enrichment (e.g. .msg email file -> MBox -> RDF)
- Provenance recorded from Transfer through to Accession

TNA Classic Catalogue Model

Enabled the end-to-end business of Accession

- Basically the Metadata Model for Records
- Designed for Paper Records
- Attempts to adapt to Digital, but does not Scale



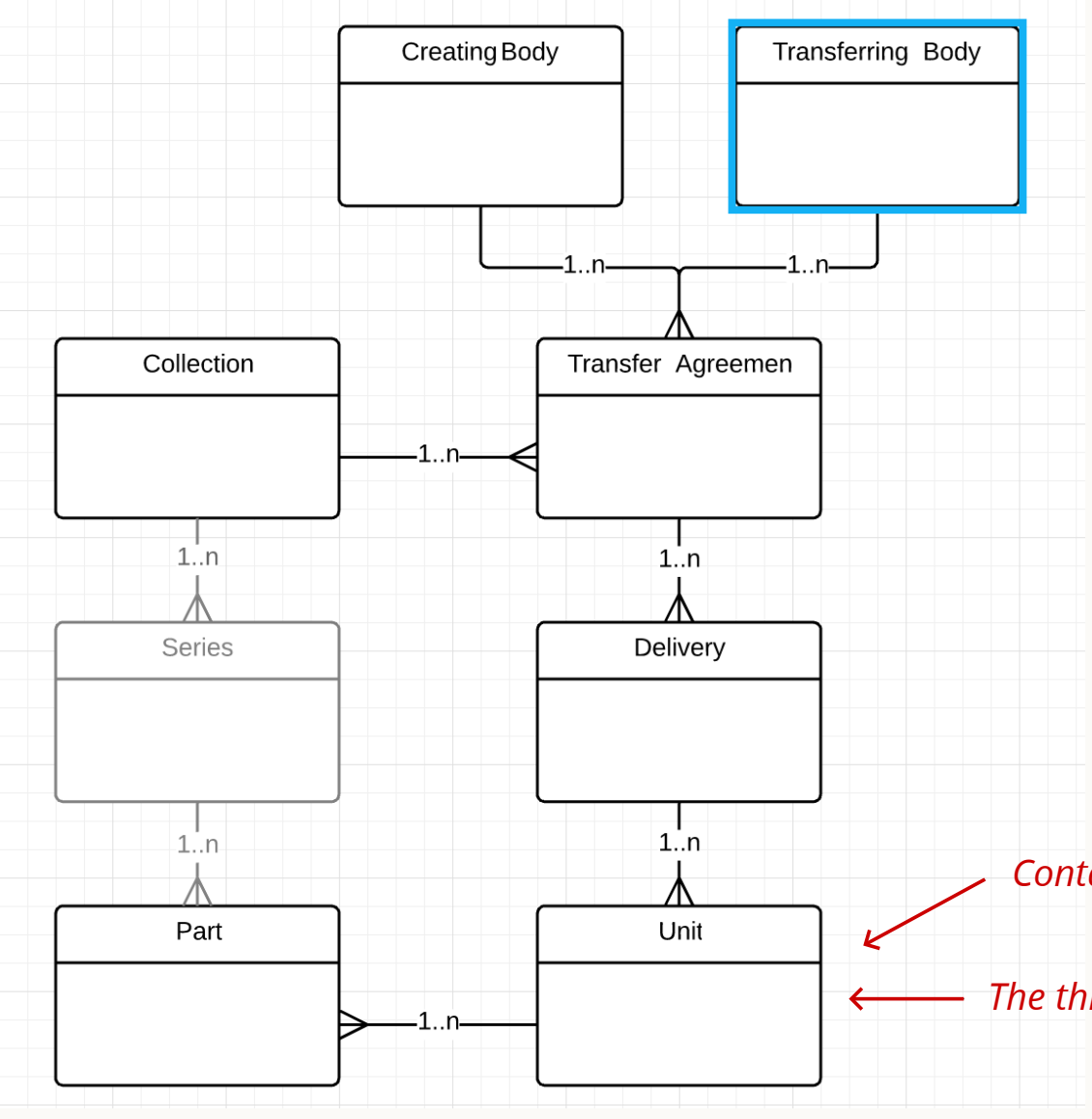
EVOLVED BINARY

DRI Accession Metadata Model

Enables the end-to-end business of Accession

- **Collection**: Distinct Set of Related Records. Composed of one-or-more Series
- **Series**: (TNA Catalogue) Records of the same provenance that were created or used together
- **Delivery**: A group of physical or electronic Units that are delivered to the National Archives as a single consignment at a single point in time
- **Unit**: Either a single item of physical media or a single electronic assembly of files.
- **Part**: Intesection of Series and Unit
  - A Series may be delivered in one or more Parts, across one or more Units.
  - Think of it as a container! It's the thing we process... concurrently!
  - Contains all of the files and metadata

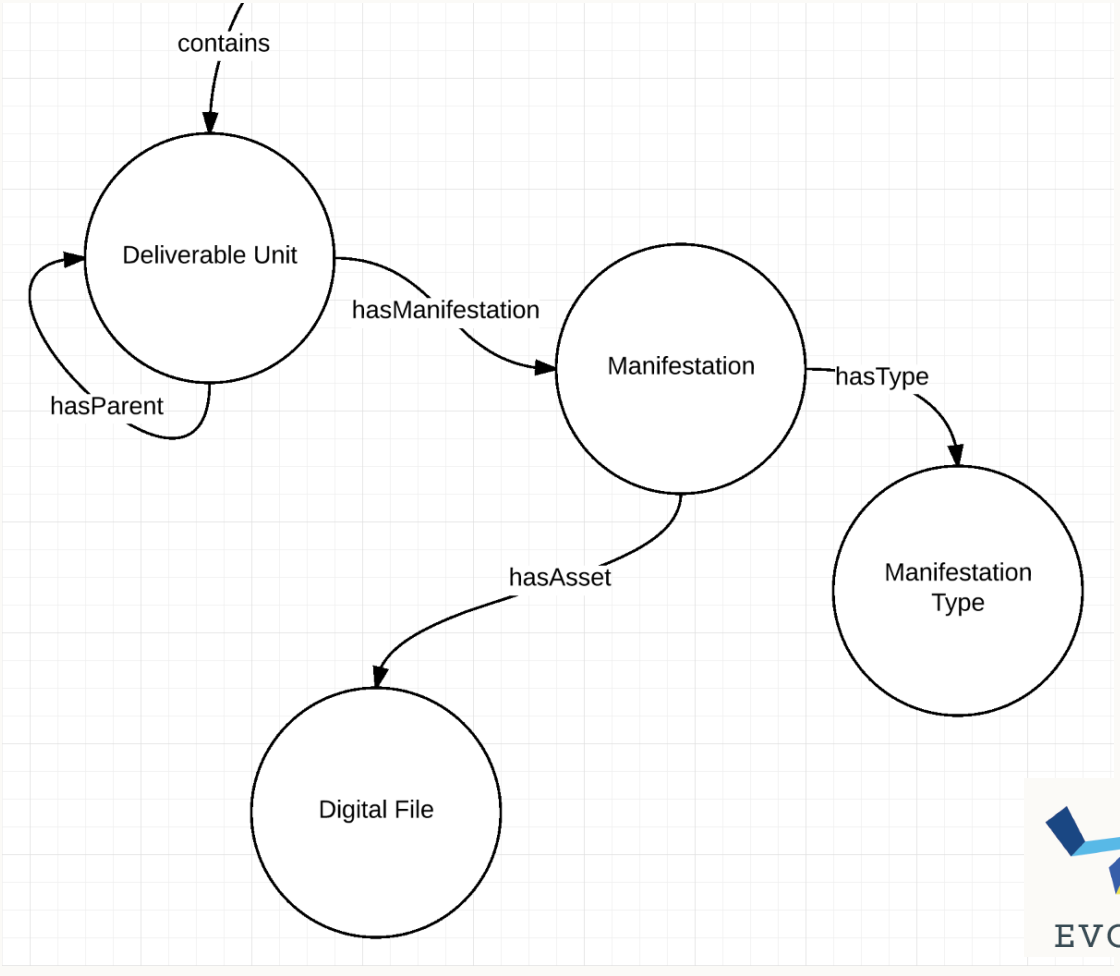
DRI Accession Metadata Model



DRI Metadata Model

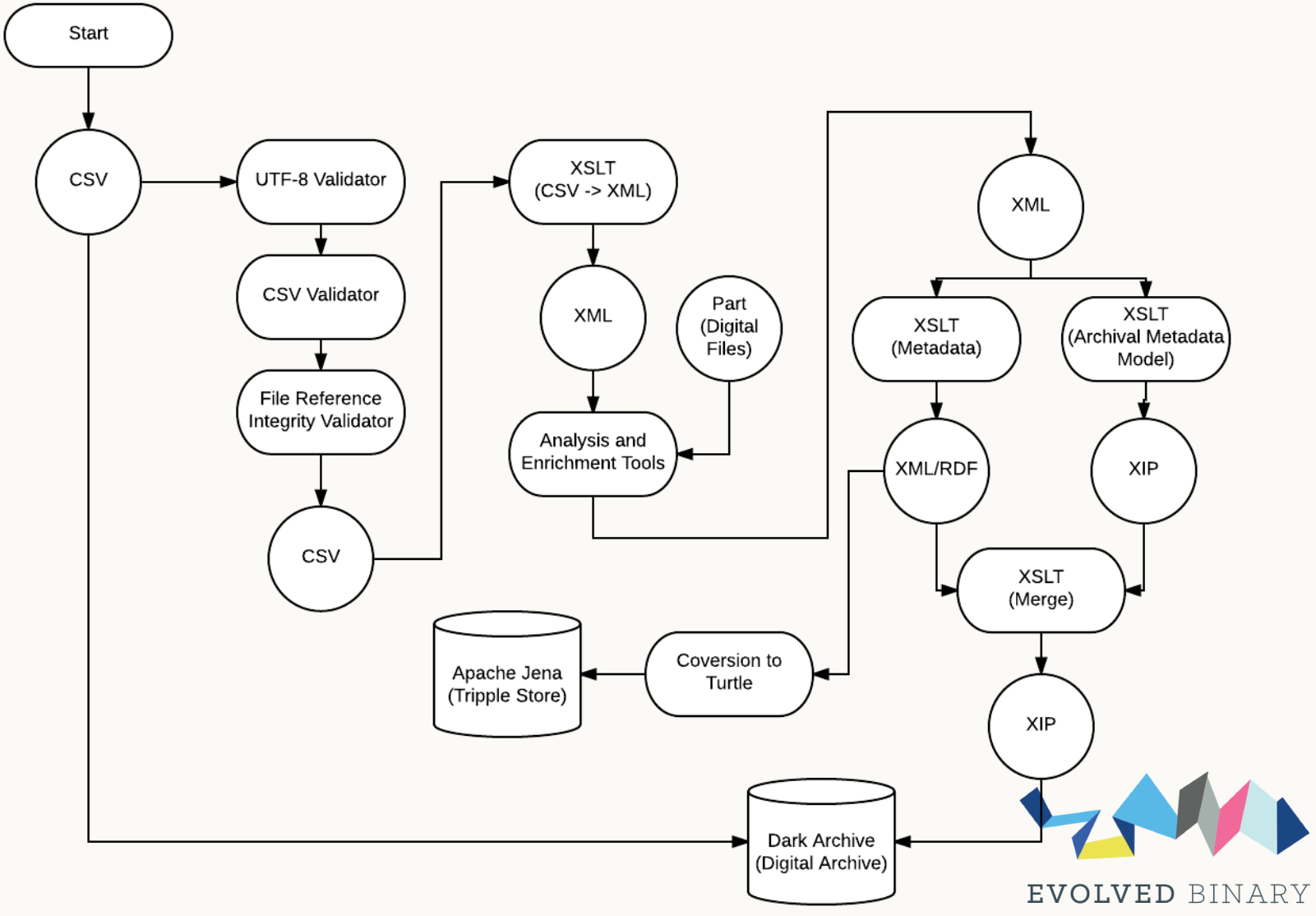
Inside a Part

- **Any Metadata Can be added at Every Level!**



EVOLVED BINARY

DRI Metadata Architecture



EVOLVED BINARY

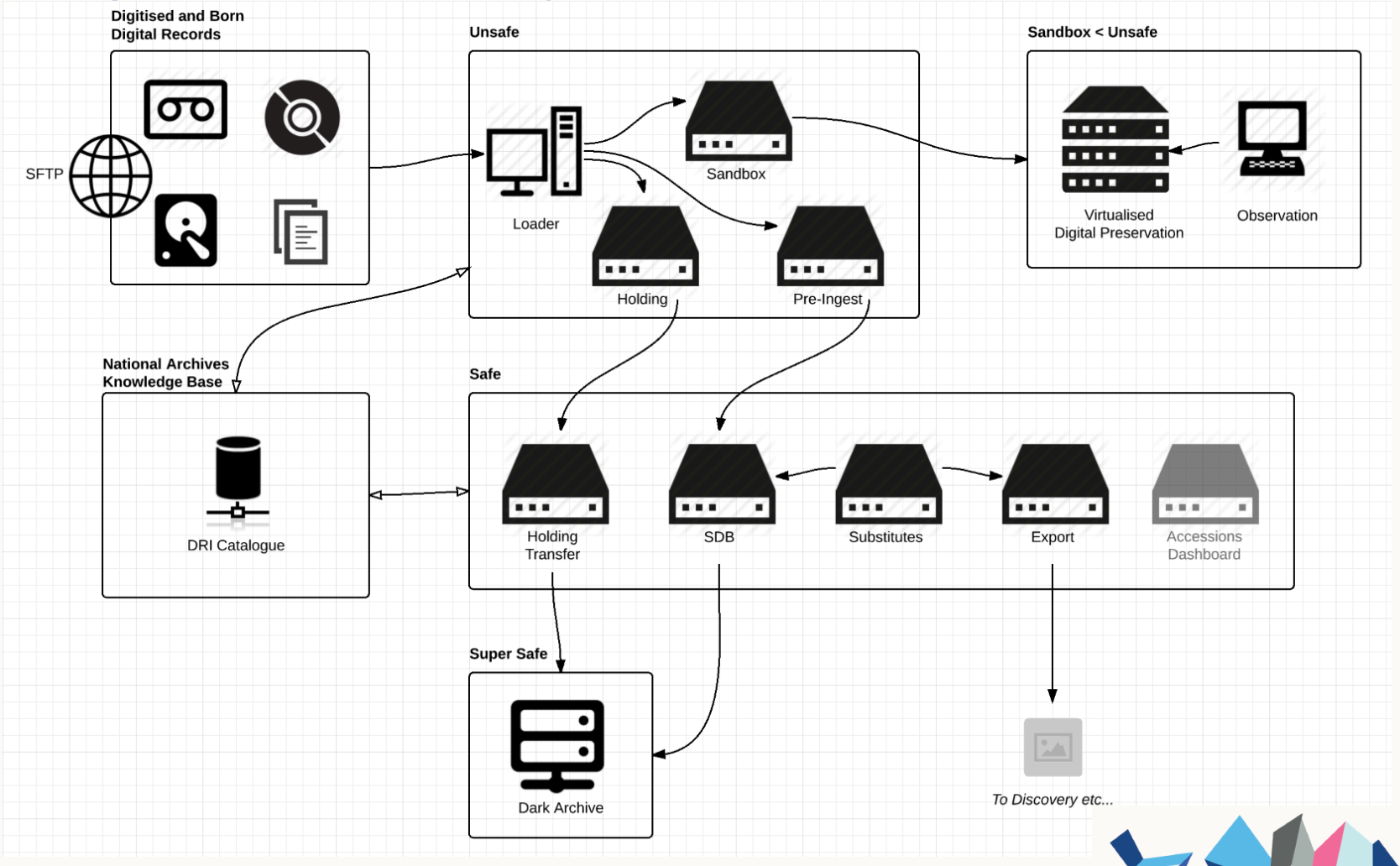
DRI Accession Metadata Model

In Summary

- **Start with CSV files adhering to CSV Schema**
- **Convert these to a simple XML representation**
- **Analysis tools over Digital Files create further XML**
- **Convert the XML into Data Model (XIP) and Metadata (XML/RDF)**
- **Store a copy of the Model and Metadata as Turtle into Apache Jena**
  - Online (non-archival) System for Querying, Presentation and System Activities
- **Inject the XML-RDF into the XIP and Store it in the Digital Archive**
  - We also store the original CSV files!

EVOLVED BINARY

High Level System Overview



EVOLVED BINARY

High Level System Overview

Pre-Ingest (Unsafe)

- For all unknown materials - i.e. Transfers that we receive
- Hadoop: Executes our Security tools and Custom Tools
- Also our Staging area and Digital Analysis/Forensics

Ingest (Safe)

- For Processing Parts
- Tessella SDB Workflows
  - Many many custom Software Components (Scala, Java, XSLT, Python, C++)
  - Many Open Source tools: Akka, ImageMagick, DROID, JHove, etc
- Apache Jena

Dark Archive (Super-Safe)

- A huge Robotic Tape Library!

EVOLVED BINARY

The Dark Archive

Huge Robotic Tape Library

- Presented as Unlimited NFS Storage
- Several Terrabyes of Near-line Disk Cache
- Very Expandable and Configurable

SAM-QFS

- Catalogues Tapes
- Policy Driven
- Knows how to export tapes and retrieve offsite tapes

Presrevation Properties

- Multiple Tape Drives: LTO-4, LTO-6, T10K
- Tape Drives and Tapes from Multiple Manufacturers
- ...Files - Multiple Copies, on Multiple Media, at Multiple Sites :-)

EVOLVED BINARY

Thank You