Library eArchiving with ZONTAL Space and the Allotrope Data Format

Dennis Della Corte\textsuperscript{1,2,*}, Wolfgang Colsman\textsuperscript{3}, Ben Welker\textsuperscript{4}, Brian Rennick\textsuperscript{5}

\textsuperscript{1}Brigham Young University, Department of Physics and Astronomy
\textsuperscript{2}ZONTAL, inc Vice President, Marketing & Sales
\textsuperscript{3}ZONTAL, inc Chief Executive Officer
\textsuperscript{4}Brigham Young University, Library, Associate University Librarian for Library IT
\textsuperscript{5}Brigham Young University, Library, Software Engineer

Currently under peer review
Presentation Outcomes

• Gain better understanding of Allotrope business benefits outside of pharma

• See differences an Allotrope based information management solution can make

• Get you excited about exploring how ZONTAL Space can benefit your business
What is ZONTAL Space all about?

INFORMATION LIFECYCLE MANAGEMENT
Build Trust to Maximize Data Reuse

- Data ownership secures data value over time
- Data governance manages documentation standards, ensures interoperability and simplifies collaboration
- Data quality management maximizes data value and compliance
- Data provenance builds trust and encourages reuse
- Configurable workflows allow implementing business specific requirements
- Automated retention schedules minimize efforts of archive management

DATA INTEGRITY, QUALITY AND PROVENANCE
Assessing data for its trustworthiness becomes important in data reuse with the growth in data creation because of the lack of standards for ensuring data quality and potential harm from using poor-quality data.
DATA INTEROPERABILITY
The seamless, secure and controlled exchange of data is a serious challenge in maximizing the potential of the many digital tools and resources. The lack of interoperability means that a complete record often will not be available, compromising compliance and inhibiting innovation.

BUILD ON OPEN STANDARDS
“Drive” Innovation and Sustainable Development

- Allotrope Data Format, GxP compliant Data Container (ADF)
- ISO 14721: Reference Model for Open Archival Information Systems (OAIS)
- ISO 16363: Trustworthy Repositories Audit and Certification (TRAC)
- EMA “Best Archiving Practice” (BAP)
- Public Ontologies (allotrope, prov, foaf, dc terms, qudt, …)
- S3 compatible storage (Amazon S3, Dell ECS, Hitachi HCP, …)
## ADF is superior to BagIt format

<table>
<thead>
<tr>
<th>Functionality</th>
<th>Required by</th>
<th>BagIt</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Data Object</td>
<td>OAIS</td>
<td>Stored in native format only</td>
<td>Stored in native format or converted to Data Cube</td>
</tr>
<tr>
<td>Representation Information</td>
<td>OAIS</td>
<td>Optional: Metadata as key/value pair</td>
<td>Fully open standards: RDF for metadata, HDF5 for Data Cube</td>
</tr>
<tr>
<td>Reference information</td>
<td>OAIS</td>
<td>Optional: Metadata as key/value pair</td>
<td>Optional: Semantic metadata according to RDF.</td>
</tr>
<tr>
<td>Provenance Information</td>
<td>OAIS</td>
<td>Not included</td>
<td>Audit trails in Data Provenance as RDF metadata</td>
</tr>
<tr>
<td>Fixity information</td>
<td>OAIS</td>
<td>Checksums stored in manifests</td>
<td>Checksums are dynamically created and verified</td>
</tr>
<tr>
<td>Hierarchical file packaging</td>
<td>OAIS</td>
<td>“Payload”</td>
<td>Data Package Layer</td>
</tr>
<tr>
<td>Semantic MetaData</td>
<td>FAIR</td>
<td>Not included</td>
<td>Data Description Layer</td>
</tr>
<tr>
<td>Interoperability between user groups</td>
<td>FAIR</td>
<td>Limited by key/value metadata pairs</td>
<td>Semantic connection of user groups enables flexible metadata enrichment</td>
</tr>
<tr>
<td>Vendor agnostic scientific format</td>
<td>FAIR</td>
<td>Not included</td>
<td>Data Cubes</td>
</tr>
<tr>
<td>Migration: Conversion between BagIt and ADF</td>
<td>FAIR</td>
<td>BagIt -&gt; ADF possible</td>
<td>ADF -&gt; BagIt possible, but semantic information will get lost</td>
</tr>
</tbody>
</table>
Ontology Mapping

- Dublin Core ontology is established standard for library application
- Current IT systems do not enforce or check correct usage of ontology

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<xml>
  <create/>
  <contri/>
  <title>[John David Smith?] letter to John Smith</title>
  <altern/>
  <descri>
    Letter from "Jod", possibly John David Smith, to his father, John Smith. Regarding health of the family and thanks for presents when October 7, 1884.
  </descri>
  <date>1884 October 7</date>
  <edito>Electronic version</edito>
  <publis/>
  <physic>1 folder</physic>
  <scale/>
  <extent>1 folder</extent>
  <medium/>
  <dimens/>
  <arrang/>
  <degree/>
  <contra>...\</contra>
  <names/>
  <topics/>
  <geogra/>
  <genre>Leters</genre>
  <occupa/>
  <functi/>
  <subjec/>
  <langua>English; eng; en</langua>
</xml>
```

ZONTAL Space automatically maps XML files containing metadata to the correct ADF and stores them as RDF triples.
BYU Pre-PoC Workflow

Many manual steps and multiple resources are required

Various Producers

Library Personnel
- Collects data from Producers

Excel Spreadsheet
- Template automatically or manually created for objects on FileShare
- Updated with Input from Producers and DPS

CONTENTdm
- Low resolution images and manually entered metadata

FileShare
- Large original images and other objects

Harvest Utility
- Collects objects from FileShare, MetaData from CONTENTdm

Rosetta
- Final Storage location

Oversees archival process

Manual step

Automatic step

Copyright 2019, ZONTAL, INC.
Nearly automatic data preservation with minimal resource overhead
80 percent of the work data scientists perform today, according to this *Forbes Magazine* report, can be automated. This puts great pressure on companies to FAIRify their data today, so that automated systems may be of the highest possible quality and performance.

**Why does this work?**

**Standardized APIs**

Secure and Automate Access to Metadata and Content

- Web service APIs integrate into the existing application landscape
- Agents integrate laboratory instruments to improve data integrity
- Allotrope Data Format (ADF) files transfer raw data, contextual metadata, audit trails and checksums in a vendor neutral format
PoC: One manual ingest command bulk imported hundreds of records and automatically enriched them with metadata.
Little features with big impact
Dashboards monitor disk space usage per collection – admin can now monitor without infrastructure support
Connecting Data in Archives to Data Analytics

Capture
- Reference & Masterdata Management
- Web Clients
- File Shares
- CDS
- LIMS
- ELN
- DMS
- SDMS

Ingest
- Ingest
- Validate
- Convert
- Receive

Manage Information Lifecycle
- Lifecycle Status
- Analyte name
- Sort By
- Search

Access
- REST API
- Tableau
- Qlik Sense
- Microsoft Power BI
- Python

Reuse
- OpenAPI
- SPARQL
- REST API
- Hadoop
- Spark
What can you do with data in ZONTAL Space?

Next steps for library use case:

Enhance existing recommender systems to leverage descriptive information.