Library eArchiving with ZONTAL Space and the Allotrope Data Format

Dennis Della Corte
Department of Physics and Astronomy, Brigham Young University, Provo, Utah, USA
Wolfgang Colsman
Zontal, inc, Provo, Utah, USA, and
Ben Welker and Brian Rennick
Library Department, Brigham Young University, Provo, Utah, USA

Received 27 September 2019
Revised 31 October 2019
Accepted 7 December 2019

Abstract

Purpose – The purpose of this technical paper is to evaluate the emerging standard "Allotrope Data Format (ADF)" in the context of digital preservation at a major US academic library hosted at Brigham Young University. In combination with the new information management system ZONTAL Space (ZS), archiving with the ADF is compared with currently used systems CONTENTdm and ROSETTA.

Design/methodology/approach – The approach is a workflow-based comparison in terms of usability, functionality and reliability of the systems. Current workflows are replaced by optimized target processes, which limit the number of involved parties and process steps. The connectors or manual solutions between the current workflow steps are replaced with automatic functions inside of ZS. Reporting functionalities inside of ZS are used to track system and file lifecycle to ensure stability and data preservation.

Findings – The authors find that the target processes leveraging ZS drastically reduce complexity compared to current workflows. Archiving with the ADF is found to decrease integration complexity and provide a more robust data migration path for the future. The possibility to enrich data automatically with metadata and to store this information alongside the content in the same information package increases reusability of the data.

Research limitations/implications – The practical implications of this work suggest the arrival of a new information management system that can potentially revolutionize the archiving landscape within libraries. Beyond the scope of the initial proof of concept, the potential for the system can be seen to replace existing data management tools and provide access to new data analytics applications, like smart recommender systems.

Originality/value – The value of this study is a systematic introduction of ZS and the ADF, two emerging solutions from the Pharmaceutical Industry, to the broader audience of digital preservation experts within US libraries. The authors consider the exchange of best practices and solutions between industries to be of high value to the communities.

Keywords Digitization, Data preservation, Digital archiving, OAIS, Allotrope, ZONTAL space

Paper type Technical paper

1. Introduction

Today, most industries are feeling the pressure to digitize their data and their processes (Kagermann, 2015). Despite the broad impact of advances in technology, it is still quite common that similar solutions are developed either in parallel or subsequently in orthogonal

Funding research: States in separate document – for double blind peer review.

2059-5816
DOI 10.1108/DLP-09-2019-0036

branches of industry. Many synergistic effects are thus left untouched by foregoing more inter-industrial exchanges of how to implement best practices. Here we will show how advances in the pharmaceutical approach to deal with data can positively benefit university libraries.

At first, it is important to highlight that data is of very little value after its initial creation. The problem with data is that it remains in a fixed context for which it was created, and this context rarely repeats itself with sufficient detail as to render reuse feasible. Actual value can only be derived from data if additional efforts are undergone to enrich data with descriptive information to transform it into information. Figure 1 shows a representation of the information life cycle and illustrates how data can still be of value after initial use. To be shared, archived and reused, it is mandatory that additional metadata is gathered that makes the initial data understandable by other users. In this work, we will focus on evaluating solutions that have enabled unprecedented digital preservation and archiving of information in the pharmaceutical industry.

Some examples of best practices that are recognized by many industries but separately dealt with are the Open Archival Information System (OAIS) (Lavoie, 2014) standard that provides a framework for best practice data archival and the Findable-Accessible-Interoperable-Reusable (FAIR) data principles (Wilkinson *et al.*, 2016). In this case study, we will investigate two solutions developed in the pharmaceutical industry and apply them to a university library use case. We will highlight the technological advancements of the pharma solutions over a current standard format used in many library archives. We will conclude with an outlook of new functionalities and features that a university can provide after upgrading to an OAIS and FAIR compliant information management system.

2. The Allotrope Data Format

In 2012, the Allotrope Foundation (www.allotrope.org/) formed as an international consortium of pharmaceutical, biopharmaceutical, and other scientific research-intensive industries. The aims of the foundation are developing advanced data architectures to transform the acquisition, exchange and management of laboratory data throughout its complete lifecycle. The first tangible product developed by the consortium was the Allotrope Data Format (ADF) (Oberkampf *et al.*, 2018), released in 2017. From the long-term preservation perspective, ADF may be summarized as a PDF for scientific data. ADF is a vendor neutral format that implements the information package as described in the OAIS

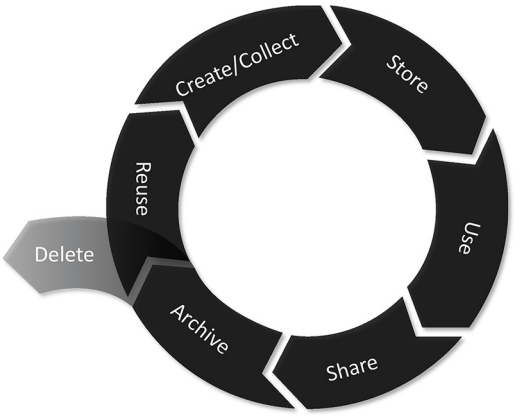


Figure 1.
Information life cycle

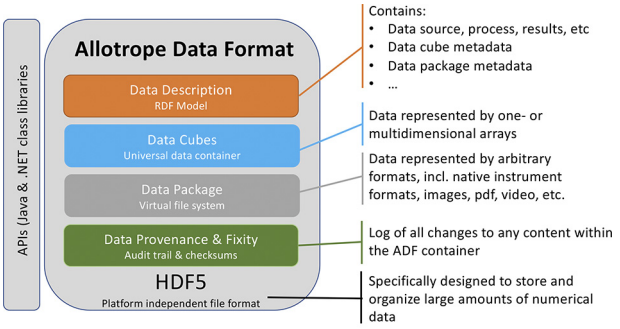
standard (Giaretta, 2012). ADF is built on HDF5 (Folk *et al.*, 2011) core libraries that have been used for effectively storing large satellite data since the 1970's.

The architecture of the ADF is shown in Figure 2. The ADF is implemented in HDF5 with a set of APIs that enable interaction with the content. The content is structured into 4 containers. The first is the Data Description container, storing all relevant metadata according to the Resource Description Framework (RDF) (Klyne and Carroll, 2006). RDF enables storage of metadata in semantic triples, which increases the ability to interoperate between files. The Data Cube container can be optionally filled with a vendor agnostic representation of structured data. It is mainly of benefit for effective long-term storage of scientific measurements. The Data Package container is a virtual file system that can store all types of source data, especially useful for images, PDFs and videos that cannot be converted to Data Cubes. Hierarchical data can also be stored in the Data Package. The Data Provenance and Fixity container stores a log file of all changes that are done on an ADF file as well as checksums for its content.

After the initial release as a data exchange format, the benefits of ADF as a long-term preservation format became apparent. Strongly regulated departments from pharmaceutical companies have since investigated opportunities to preserve their ADF files in accordance with FAIR data principles (Wilkinson *et al.*, 2016). While ADF files are by definition reusable, the major challenge is to make them findable, accessible, and interoperable. For this purpose, the architects of the ADF framework have developed an information management system named ZONTAL Space (ZS) that implements a full archival system as specified by the OAIS. Usage of ADF and ZS are not limited to scientific laboratories. Here, we will explore how the ADF standard in combination with ZS can support digitization and preservation of information created from data stored at the Brigham Young University Library (BYUL).

3. Comparison of Open Archival Information System, BagIt, Allotrope Data Format and user requirements

Many libraries are currently using data management tools that predate the release of the latest OAIS standard. Many of these solutions implement the BagIt standard (Kunze *et al.*, 2018) that resembles core functionalities of the information package as prescribed by the OAIS. The question that forces itself upon each operator of such a legacy system is, whether such a solution can be considered OAIS compliant. Here we will provide an overview of core



Note: The ADF consists of multiple containers, wrapped in HDF5 and accessible through APIs

Figure 2.
ADF architecture

functionalities and features suggested by the OAIS standard and FAIR data principles and evaluate the ability of BagIt and ADF standards to deliver these features. The results are shown in Table I.

Our analysis shows that the BagIt standard describes a minimalistic information package that complies with the OAIS in most, but not all aspects. BagIt allows for storage of source data in native format and optional XML schema-based metadata as key/value pairs. The main strength of the format is an XML manifest storing the checksums of all files in the "payload" of a "bag" to provide fixity information in the future. From an OAIS perspective, the lack of audit trails reduces the ability to track a bag through its lifecycle. Systems that manage bags must keep a second record independent of the file and future dissemination or migration will become more difficult to achieve.

The main shortcomings of the BagIt format become apparent when comparing it to newer requirements as set forth in the FAIR data principles. The most crucial aspect in transforming data into information is the systematic enrichment with meaningful semantic metadata. BagIt does not provide a technical solution for storing metadata triples according to RDF. Without the ability to link between controlled vocabularies and taxonomies, the metadata in a bag remains static and one-dimensional. This becomes an issue when considering interoperability of files between different user groups in an organization. A bag with key/value pairs from one group in an organization might not be interpretable by a second group, as the used labels commonly change their meaning between departments. Even in case of a controlled master data system in place, the key/value storage of metadata does not allow for sophisticated quality checks that semantic reasoners allow on data triples governed by ontologies. The reusability of a bag is completely dependent on the data owner's ability to keep all legacy data processing tools in operation. Without built-in functionality to convert data into a long-term, stable, vendor agnostic format, a bag will only be valuable as long as the correct IT infrastructure is upheld.

Table I.
Comparison of BagIt and ADF format for compliance with OAIS standard and FAIR data principles

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

The ADF, in contrast, offers all of the same features as BagIt with the addition of semantic metadata, audit trails, and a vendor agnostic format for storing scientific data. A key difference is the "openness" of the formats, with BagIt being released under MIT license agreements and many established reader and writer implementations available. Allotrope, on the other hand, is still a developing format closely controlled by the Allotrope Foundation and the Allotrope Partner Network. While it is free of charge for research institutes to join Allotrope, businesses need to purchase developer licenses if they want to sell products using Allotrope technology. If Allotrope wants to claim a similar market position for scientific data as PDF for text files, this business model will have to be adjusted in the future.

In conclusion, BagIt offers a minimal standard, suitable for storing data in small organizations over short periods of time, mainly with the focus on dark archiving. Allotrope, on the other hand, incorporates new standards and principles to lay a strong foundation for living archives that can scale through global organizations and enable long-term, compliant information preservation. An ADF file will be a self-contained dissemination information package after removing it from its management system, ready for migration, analysis, or reuse.

4. ZONTAL Space: a library use case

A full OAIS requires effective communication between data producers, consumers, and management. To submit, archive, and disseminate information packages, an information management system should provide highly customizable, ideally automatic, workflows. ZS is an implementation of an OAIS that attempts to achieve all of the above. ZS was designed to be completely format agnostic and interoperable with existing IT landscapes. It can connect through restful APIs or through a user interface with the outside world to ingest data and to trigger automatic metadata enrichment. Throughout the lifecycle of a file, ZS keeps track of the status and can prompt data stewards for managerial intervention. Internally, ZS converts ingested data to ADF files and stores them initially as submission information packages. After promotion to archival information packages, the rights and permissions of the files are updated in compliance with OAIS. ZS provides basic and detailed customizable report features that enable effective management of the information system. Dissemination is achieved through manual downloads via the user interface, via restful APIs, or through connection to other downstream systems, like business analysis tools.

ZS was initially released in November 2018 and the current release is version 2.2.6. Here, we will perform a proof of concept with BYUL to demonstrate the ability of ZS to ingest data from the content management system CONTENTdm, extract meaningful metadata, and store it in a dark archive. We will also use ZS reporting functionality to produce insights not available with the current archiving system, Rosetta from Ex-Libris (Alter and Peled, 2015), at BYUL.

The main motivation for this proof of concept is the difficulty in establishing connections between data sources and the archival system. Most recently, the integration of BEpress Digital Commons required over 40 hours of work to establish successfully. The BYUL internally developed harvester tools require too much manual overhead to remain as a long-term solution. Further, BYUL requires periodical reporting of system disk usage on a collection level, which is not available in the current system. Analysis of current workflows also suggested that the process landscape could be drastically simplified by replacement of manual steps through automatic workflows in an information management system.

4.1 Detailed proof of concept methodology

The proof of concept consisted of four stages. First, a ZS installation was configured for BYUL. Second, the content management system CONTENTdm (Sager and Ladd, 2016) was directly connected to ZS via restful APIs. Third, a complex manual archiving workflow was replaced with a simple manual ZS workflow. Fourth, custom reports with greater detail than available in Rosetta were generated with ZS.

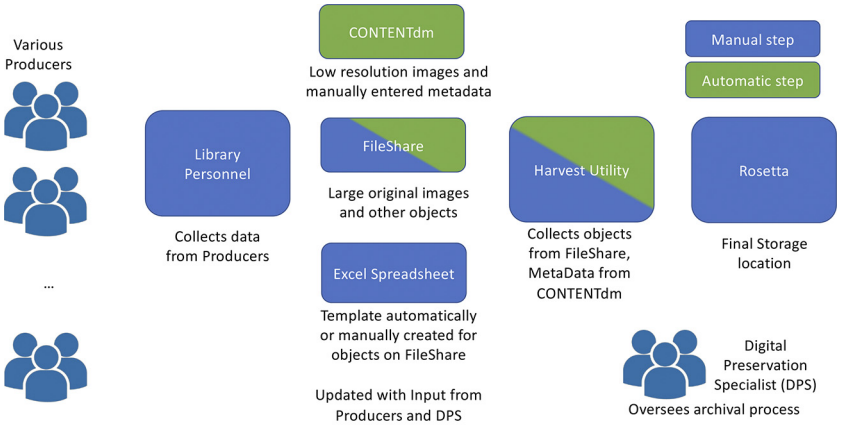
A clean ZS environment was hosted at OSTHUS, and user logins were created for BYUL employees. Initially, an information package profile was created for the users together with ZONTAL. This profile contained mandatory and optional metadata fields as specified by the Dublin Core Ontology (Weibel and Koch, 2000), shown in Table II.

The current workflow of manual archiving of data and of data stored in CONTENTdm was analyzed and is shown in Figure 3. Data producers at the library have to go through a multi-stage process, simplified for the purposes of this publication. From a high-level perspective, multiple iterations with trained Digital Preservation Specialists (DPS) are necessary to store content and metadata in intermediate repositories. The final archival step is performed through a self-developed script, referred to as the harvester utility. Data stored in the content management system CONTENTdm also need to be periodically moved to the long-term storage in Rosetta. Here again the harvester utility is employed to move content and metadata appropriately between systems. Due to changing workflows and standards, the metadata is often asynchronous and changed between collocations and time frames. This makes it difficult to find and retrieve data, or to produce systematic reporting or analysis on the information managed by the current systems.

For purpose of this proof of concept, a new target process was designed as displayed in Figure 4. Here, most of the complexity is taken away from the as-is workflow, by leveraging the user interface of ZS. Data producers are able to ingest files directly with a few clicks and can check standardized metadata as suggested by the system. Upon upload of new archival data, a submission information package is created in the staging area of ZS and an

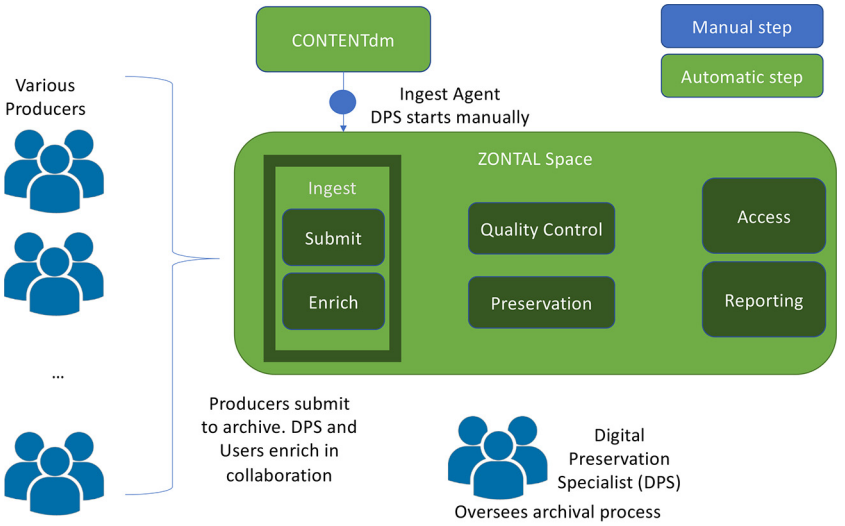
Table II.
Metadata as configured for the proof of concept BYUL archiving within ZS. Automatically filled entries represent metadata extracted or inferred by the system, and manual metadata is entered by suitable roles inside the archiving workflow

Dublin core term	mandatory/optional	Entry type
dc:title	m	Free Text
dcterms:created	o	Date Selection
dc:date	o	Date Selection
dc:coverage	o	Free Text
dcterms:extent	o	Free Text
dcterms:rightsHolder	o	Free Text
dc:type	o	Free Text
dc:language	o	Pick List
dc:relation	o	Free Text
dcterms:bibliographicCitation	o	Free Text
dc:identifier	o	Free Text
dc:rights	o	List of Free Text
dcterms:license	o	Free Text
dcterms:rightsHolder	o	Free Text
dcterms:accessRights	o	Free Text
dc:publisher	o	Free Text
dc:format	o	Free Text
dc:description	m	Free Text
dcterms:available	o	Free Text
dcterms:isPartOf	o	Free Text



Note: For manual as well as automatic archiving multiple manual steps are required, involving a variety of users and archivists to ensure data quality and process compliance

Figure 3.
Analysis of as-is archival workflows at BYU library



Notes: Data producers interact directly with the ZONTAL Space user interface to submit and enrich data. The DPS performs quality control and ensures that preservation storage is sufficient through periodical reporting. CONTENTdm is automatically synchronized with the archive through launching of a ZONTAL ingest agent. For this proof of concept, the DPS launches the agent manually; automation in production use can be configured

Figure 4.
Archival target process with ZS

automatic metadata extraction process is triggered as part of the ingest. During this process, the metadata specified in Table II is automatically extracted from the xml file and mapped against the Dublin Core ontology. The DPS is mainly in the role of quality control and ensures proper annotations are included with each submission information package in the system before promoting it to an archival information package. For automatic archival from CONTENTdm, a ZONTAL agent is launched by the DPS to trigger automatic bulk import and synchronization of CONTENTdm and the archive. Only in a case of missing mandatory metadata is manual intervention and enrichment required, before SIPs are promoted to AIPs.

One pressing issue frequently encountered at BYUL is the correct assessment of disk space used by the various data collections within Rosetta. To write budget requests, detailed statistics are required, but not accessible within the current solution. ZS reporting was customized to provide an in-depth report on the files stored in each collection, detailing file types, file numbers, and file sizes. An example report is shown in left panel of Figure 5.

This proof of concept showed that, by taking advantage of the ZS interface, much of the work currently done by the DPS can be off-loaded to content producers throughout the library. This results in a much higher throughput for the library's digitization and preservation efforts. Part of this is due to the ease with which content can be uploaded to ZS compared to existing systems at BYUL.

In addition to offloading work to content producers throughout the library, the automated processes for ingesting content from CONTENTdm further reduce the workload of the DPS, increasing the library's throughput in this area even further. Furthermore, ZS reporting was used after a second run of the automated CONTENTdm ingest script, which picked up new content from the collection. The result is shown in right panel of Figure 5.

5. Conclusion and outlook

Allotrope is rapidly growing in adaption in the pharmaceutical area and was here shown to bring new functionality to digital archives. ZS is currently the only information management system that converts all ingested data into the ADF and leverages the abilities of this standard. Usage of both enables more efficient workflows, long-term data preservation, and opens opportunities for additional data reuse benefits.

Investigation of the ZS capabilities suggests that it could replace other legacy document management tools. After the successful initial proof of concept, we will evaluate how CONTENTdm functionality might be replaced by ZONTAL workflows to further simplify the IT landscape at the BYUL. After ingesting multiple records, it also became apparent that search results within ZS could enrich current recommender systems developed at BYUL. Further work is necessary to evaluate these opportunities.

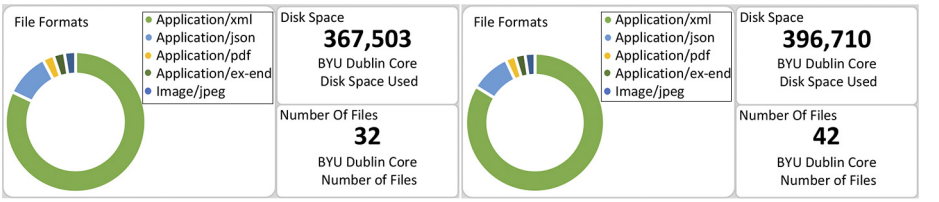


Figure 5.
Example report from ZS showing the disk space used by BYUL during the PoC

Note: On the left hand after an initial run of the automated CONTENTdm ingest script, on the right hand after a second run of the script

References

Alter, A. and Peled, I. (2015), *Managing and Preserving Research Data in Ex Libris Rosetta*, iPRES, p. 215.

Folk, M., Heber, G., Koziol, Q., Pourmal, E. and Robinson, D. (2011), "An overview of the HDF5 technology suite and its applications", *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*, ACM, New York, NY, pp. 36-47.

Giaretta, D.A.D.W.G. (2012), CCSDS AND PANEL, CCSDS, Reference model for an Open Archival Information System (OAIS).

Kagermann, H. (2015), *Change through Digitization – Value Creation in the Age of Industry 4.0. Management of Permanent Change*, Springer.

Klyne, G. and Carroll, J.J. (2006), "Resource description framework (RDF): concepts and abstract syntax", available at: www.w3.org/TR/2004/REC-rdf-concepts-20040210/

Kunze, J., Scancella, J., Adams, C. and Littman, J. (2018), "The bagIt file packaging format (v1. 0)".

Lavoie, B.F. (2014), *The Open Archival Information System (OAIS) Reference Model: introductory Guide*, Digital Preservation Coalition.

Oberkampf, H., Krieg, H., Senger, C., Weber, T. and Colsman, W. (2018), *20 Allotrope Data Format–Semantic Data Management in Life Sciences. Pdf*, Figshare.

Sager, P. and Ladd, M. (2016), "Using CONTENTdm".

Weibel, S.L. and Koch, T. (2000), "The Dublin core metadata initiative", *D-Lib Magazine*, Vol. 6 No. 12, pp. 1082-9873.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L.B. and Bourne, P.E. (2016), "The FAIR guiding principles for scientific data management and stewardship", *Scientific Data*, Vol. 3 No. 1.

Corresponding author
Dennis Della Corte can be contacted at: dennis.dellacorte@byu.edu