UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Going Full Circle: Research Data Management @ University of Pretoria

Presentation at eResearch Africa 2014 Conference,
held at University of Cape Town, Cape Town,
South Africa, 23-27 November 2014

By Johann van Wyk and Isak van der Walt

Introduction

Internationally research data is increasingly recognised as a vital resource whose value needs to be preserved for future research. This places a huge responsibility on Higher Education Institutions to ensure that their research data is managed in such a manner that they are protected from substantial reputational, financial and legal risks in the future. This presentation will focus on the Research Data Lifecycle, with an overview of RDM at the University of Pretoria, and a demonstration of pilot projects implemented at the institution.

Research Data Management: A (Brave) Complex New World

Messy

Data Sharing

Open Data

Complex

Data Repository

DMP

Various formats

Data Preservation

Data Archiving

METADATA

Various devices

Various Versions

Copyright License

DATA CITATION

DOI

Data Policy

Data Publishing

RDM

Data Analysis

Linked Data

Small Data

Data Journals

Big Data

Data Anonymisation

Sensitive Data

Visualisation

Data Formats

Data Provenance

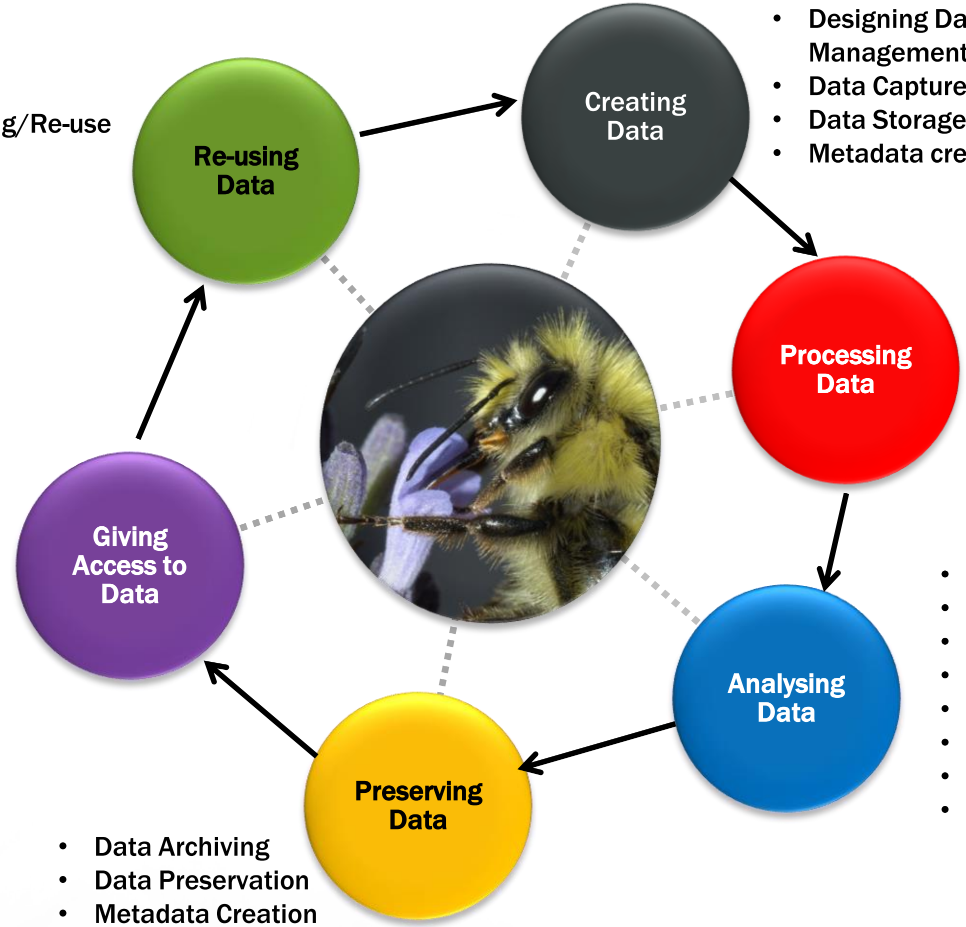
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Why Manage Research Data?

By managing research data you will:

- Meet funding body grant requirements, e.g. NSF, NIH;
- Meet publisher requirements
- Ensure research integrity and replication;
- Ensure research data and records are accurate, complete, authentic and reliable;
- Increase your research efficiency;
- Save time and resources in the long run;
- Enhance data security and minimise the risk of data loss;
- Prevent duplication of effort by enabling others to use your data;
- Comply with practices conducted in industry and commerce; and
- Protect your institution from reputational, financial and legal risk.

Research Data Lifecycle



- Designing Data Management Plans
- Data Capture
- Data Storage
- Metadata creation

- Data Storage
- Metadata Creation
- Data Cleansing
- Data Verification
- Data Validation
- Data Anonymisation

- Data Storage
- Metadata Creation
- Data Cleansing
- Data Verification
- Data Validation
- Data Anonymisation
- Data Interpretation & Analyis
- Data Visualisation

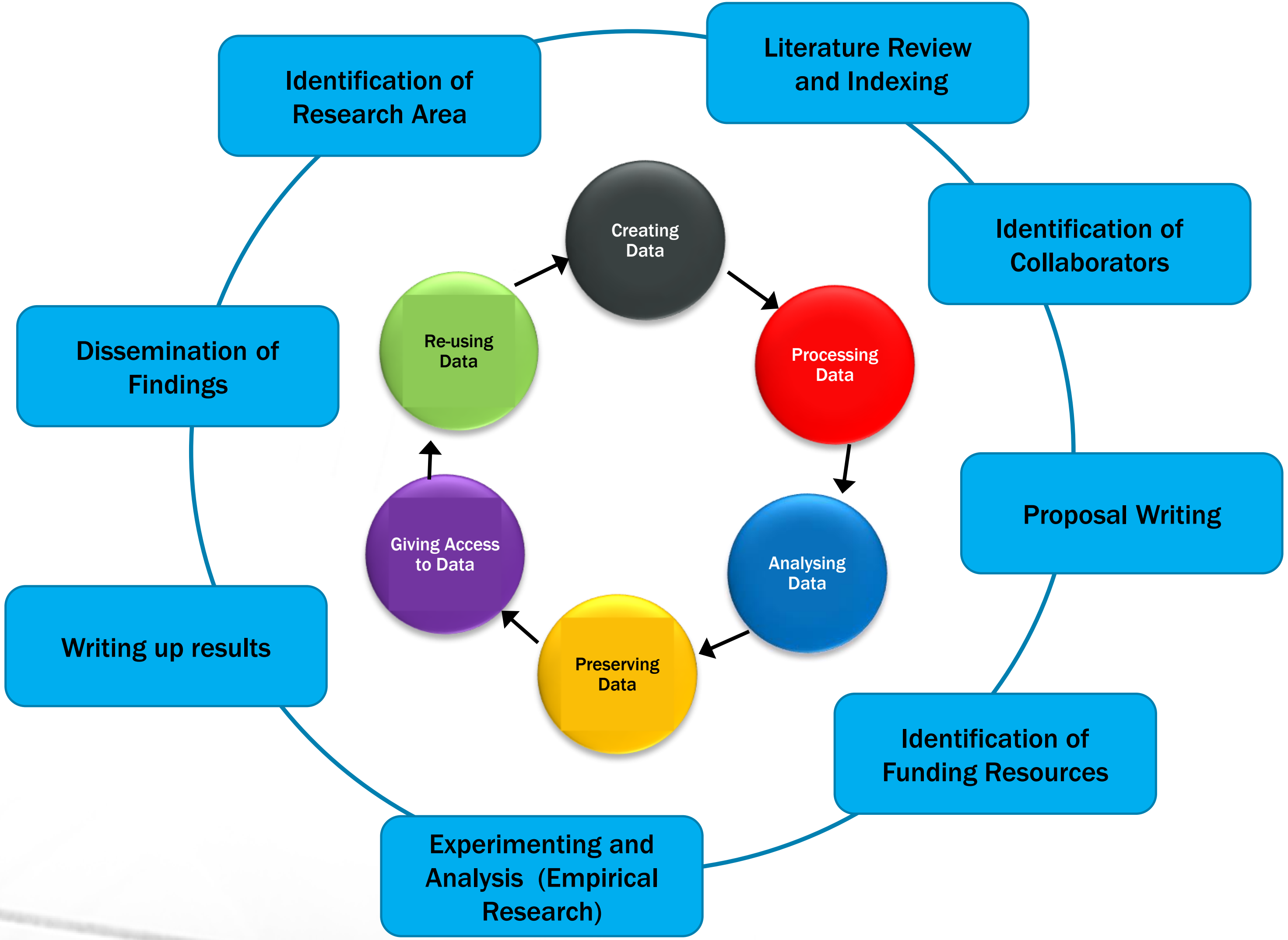
- Data Archiving
- Data Preservation
- Metadata Creation
- Link data to outputs

- Data Sharing
- Data Publishing
- Link Data to Outputs

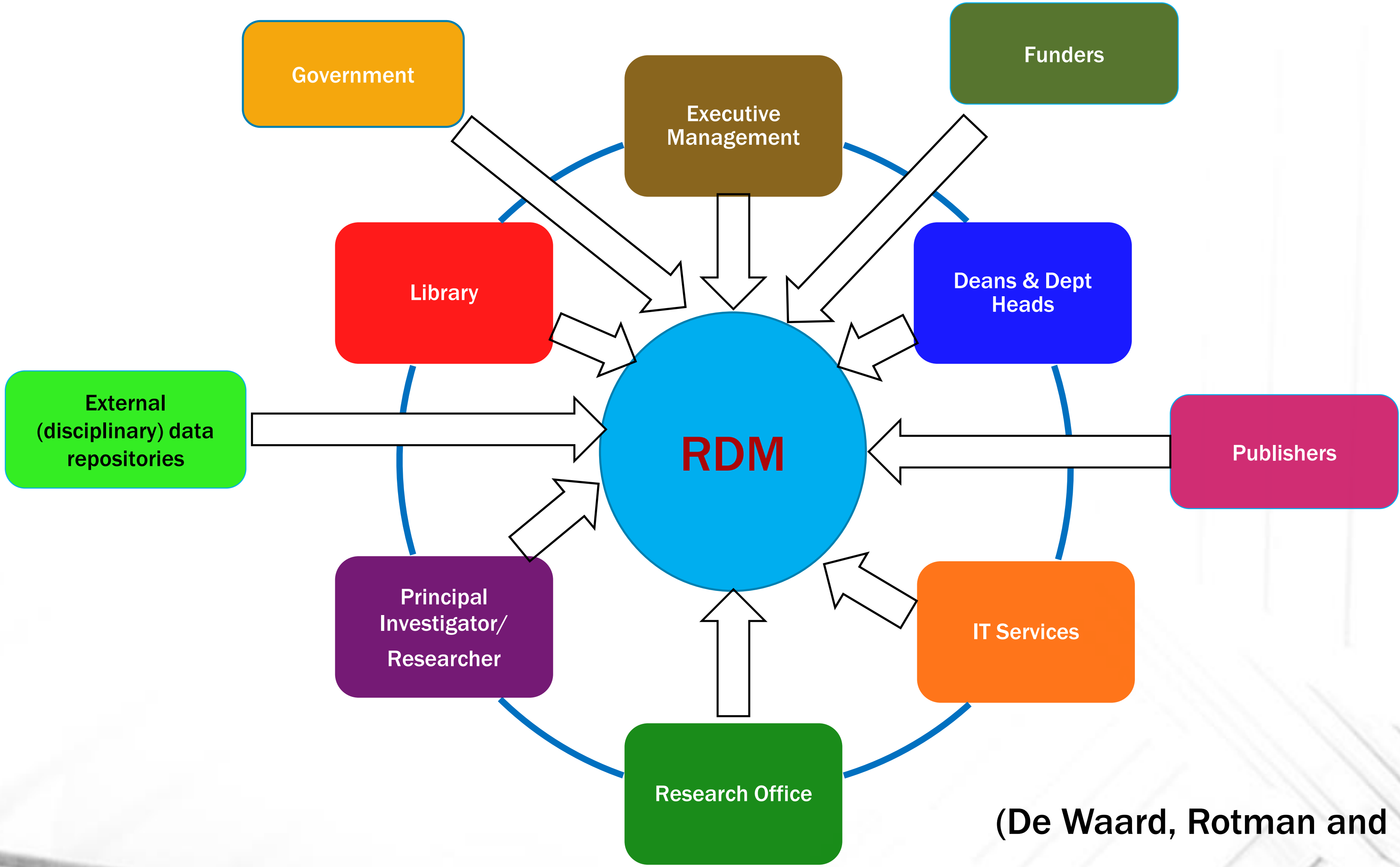
- Data Repurposing/Re-use
- Data Citation

Based on UK Data Archive Lifecycle

Research Data Lifecycle in the context of the Research Life Cycle


Identification of Research Area
Literature Review and Indexing
Identification of Collaborators
Proposal Writing
Identification of Funding Resources
Experimenting and Analysis (Empirical Research)
Writing up results
Dissemination of Findings
Creating Data
Processing Data
Analysing Data
Preserving Data
Giving Access to Data
Re-using Data

Various stakeholders in RDM



(De Waard, Rotman and Lauruhn, 2014)

Chronological Development of RDM at University of Pretoria

Mike Prins Wikipedia

By JMK on Wikimedia Commons

- **2007** - Policy for the preservation and retention of research data
- **2010** - Survey of RDM practices at UP (October 2009 – March 2010)
- **April 2013** – Meeting - Director of Institute for Cellular and Molecular Medicine (ICMM) on possibility of pilot project with students, which was subsequently implemented
- **August-November 2013** - Interviews with Deputy Deans Research Faculties to determine the "Essential Research Data that the University should manage"
- **December 2013** – 2nd Pilot project – Neuro-Physiotherapy

Chronological Development of RDM at University of Pretoria

Mike Prins Wikipedia

By JMK on Wikimedia Commons

- **April 2014** – Visit by Deputy Director Innovation and Technology and Library IT Specialist (UP Library Services) to Purdue University in USA: investigate Purdue's Research Data Repository (PURR) and long-term preservation processes as possibility for replication at UP
- **June 2014** - Assistant Director RDM attended CODATA International Training Workshop in Big Data for Science for Researchers from Emerging and Developing Countries, in Beijing, China.

Chronological Development of RDM at University of Pretoria

Mike Prins Wikipedia

By JMK on Wikimedia Commons

- **July 2014** - High Level Report on RDM sent through to University Executive for review
- **August 2014** – Proposed new University policy on RDM sent through to University Executive for review
- **Jan 2015** – Task Team to investigate infrastructure needed for RDM across the University

Survey of RDM practices at University of Pretoria, October 2009 – March 2010

- 52 interviews conducted by 15 information specialists from relevant Faculty Libraries
- At least 3 lecturers and one Postgraduate student from each Faculty were interviewed
- The information specialists received formal training in interview techniques.
- Interviews were conducted according to a semi-structured interview framework.

Findings of Survey - October 2009 – March 2010

- **Funding:** In most cases no need for data management or data sharing plans, depending on funding agency requirements
- **Data Collection:** Wide variety of data collection methods used. Both primary and secondary data. Data sets are often small.
- **Data Storage:** Ad hoc storage of data, both on paper and electronically
- **Publishing:** In general raw data not published
- **Support:** Lack of support with regard to storage of data (physical and electronic).
- RDM does not exist in any formal manner (with the exception of one or two departments) at the University of Pretoria

Recommendations: Survey - Oct 2009 – Mar 2010

- Investigate Very Large Database initiative from the Department of Science and Technology as possibility to support UP's RDM needs
- Central UP server or repository
- Address the need for physical storage space.
- Create a formal staff position of 'research data manager' to drive RDM at UP

UP Survey of RDM at UP – August – October 2013

- Interviews with Deputy Deans Research of each of the Faculties
- Focus: Determine what is seen as the essential data of the faculty that must be managed
- Conducted eleven interviews, August – October 2013
- Trends were then identified

Trends

- **Essential Data**

Interview data, Questionnaires, Spread sheet data, Lab books, Experiment /laboratory data, Images (e.g. graphs, models, Sketches, X-Rays, scans etc.), Literature reviews, Sequencing data, Computer-generated data.

- **Level of Data**

Some see only raw data as essential, some see processed /analysed data as essential, while others see both raw and processed/analysed data as essential

- **Volume of data**

A small number work only with small data sets.

Majority work with small and big data sets, with exponential increase in big data sets

Trends

- **Data Formats used**

Excel, Pdf, MS Word, Text Format, images in various formats, video, sound, various computer generated formats, SPSS, SAS, AMOS, Qualtrics data, SurveyMonkey data, simulation data formats, and even data from social media. Some in paper format.

- **RDM Plans**

None have Research Data Management Plans in place.

DMP ONLINE

https://dmponline.dcc.ac.uk/

https://www.flickr.com/photos/rosefirerising/6776182890/

- **Uploading capacity**

No capacity to upload these data sets to a repository

- **Willingness to share data**

Majority willing to share their data under certain conditions.

Health Sciences not willing to share their data

http://en.wikipedia.org/wiki/File:Open_Data_stickers.jpg

Recommendations

- A new Research Data Management Policy for UP
- Establishment of a central research data management office
- Establishment of a RDM presence in each Faculty
- Consider impact of RDM on workload and time of researchers and students
- Establishment of data repository for UP
- Investigate necessary IT infrastructure for RDM – (handle small and big data sets, and HPC)
- Determine a time frame for the roll-out of a RDM system for UP

Pilot Projects at University of Pretoria

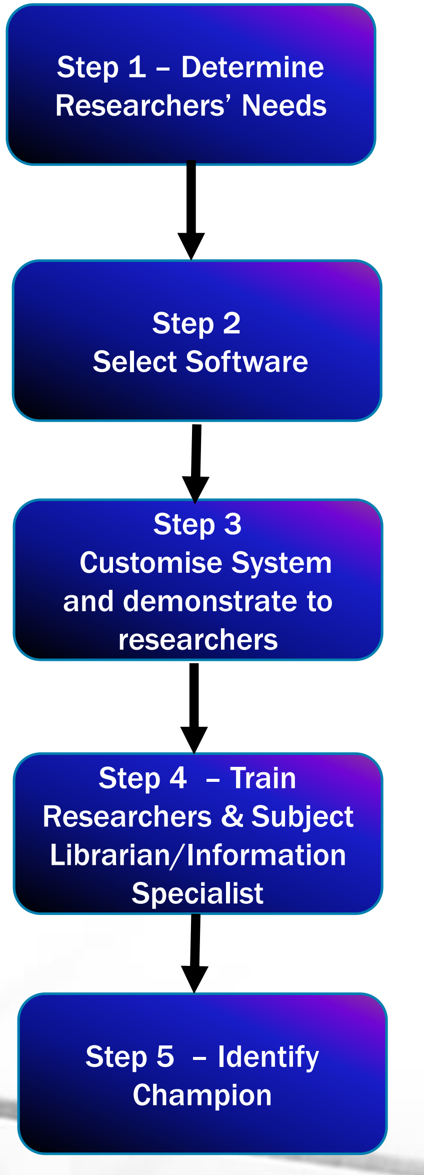
- Two data management pilot projects in 2013-2014:

  Institute for Cellular and Molecular Medicine (ICMM) and the Neuro-Physio-Group.
- Currently also implementing more pilot projects: Potato Pathology Programme, Powdery Scab, and Psychiatry Dissociation
- An Open Source Document Management System Alfresco was customised for this purpose

Why Alfresco?

Open Source
Captured provenance of data
Has a versioning function
Good metadata function
Easy to integrate with other software
Workflow function gave supervisor overview of progress of students
Sync function with Dropbox and Google Drive
Drag and Drop function
File Sharing function
Mobile App

Process followed to implement Pilot Projects



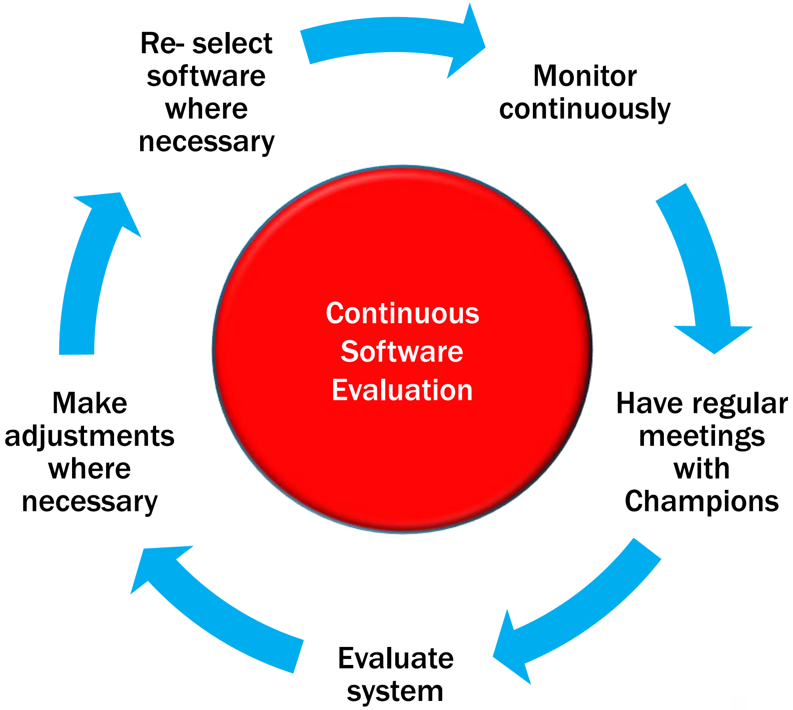
Step 1 – Determine Researchers' Needs

Step 2
Select Software

Step 3
Customise System and demonstrate to researchers

Step 4 – Train Researchers & Subject Librarian/Information Specialist

Step 5 – Identify Champion



Continuous Software Evaluation

Monitor continuously

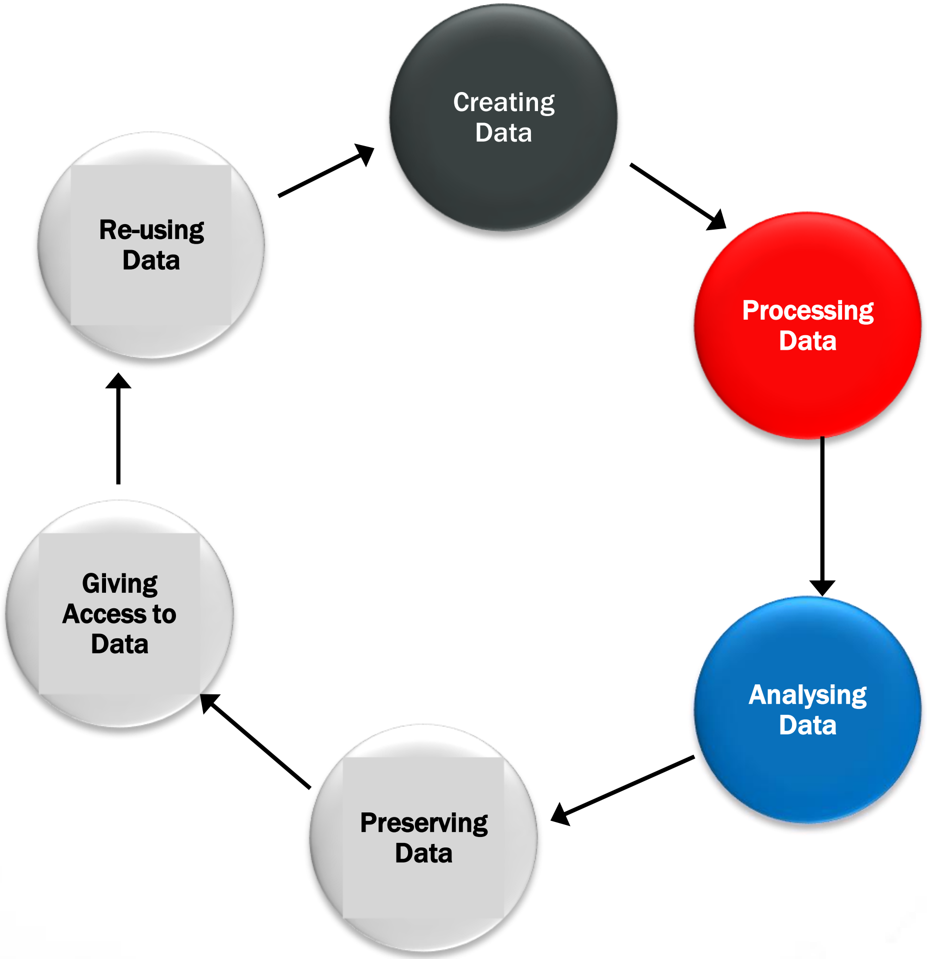
Have regular meetings with Champions

Evaluate system

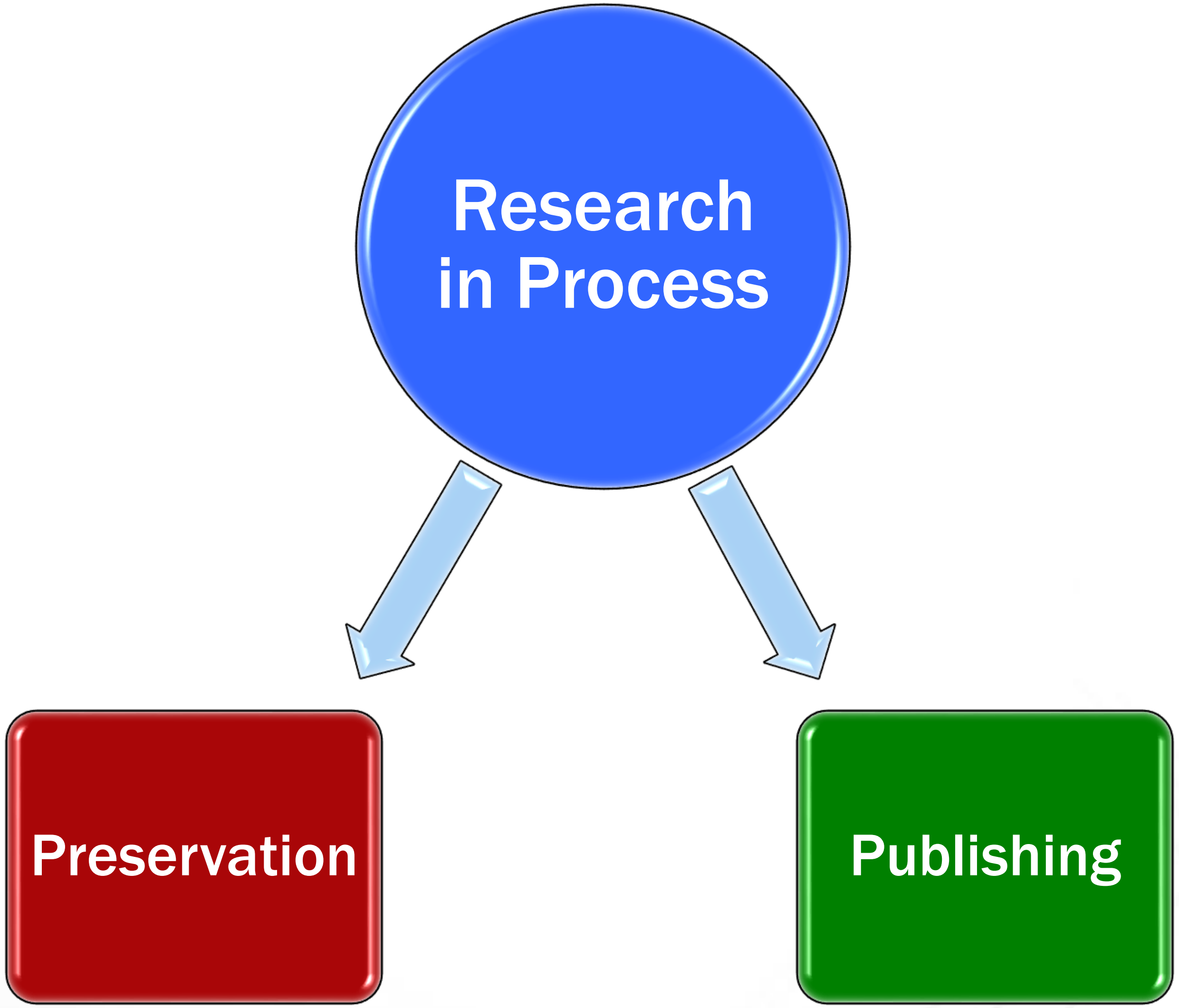
Make adjustments where necessary

Re- select software where necessary

Research in Process part of RDM



The Next steps in our RDM pilot studies



Research in Process

Preservation

Publishing

Overview

1. ECM Approach

2. Supporting the Research Data Lifecycle

3. Research in Process

4. Dissemination

5. Preservation

6. Further Development and Hurdles

ECM Approach

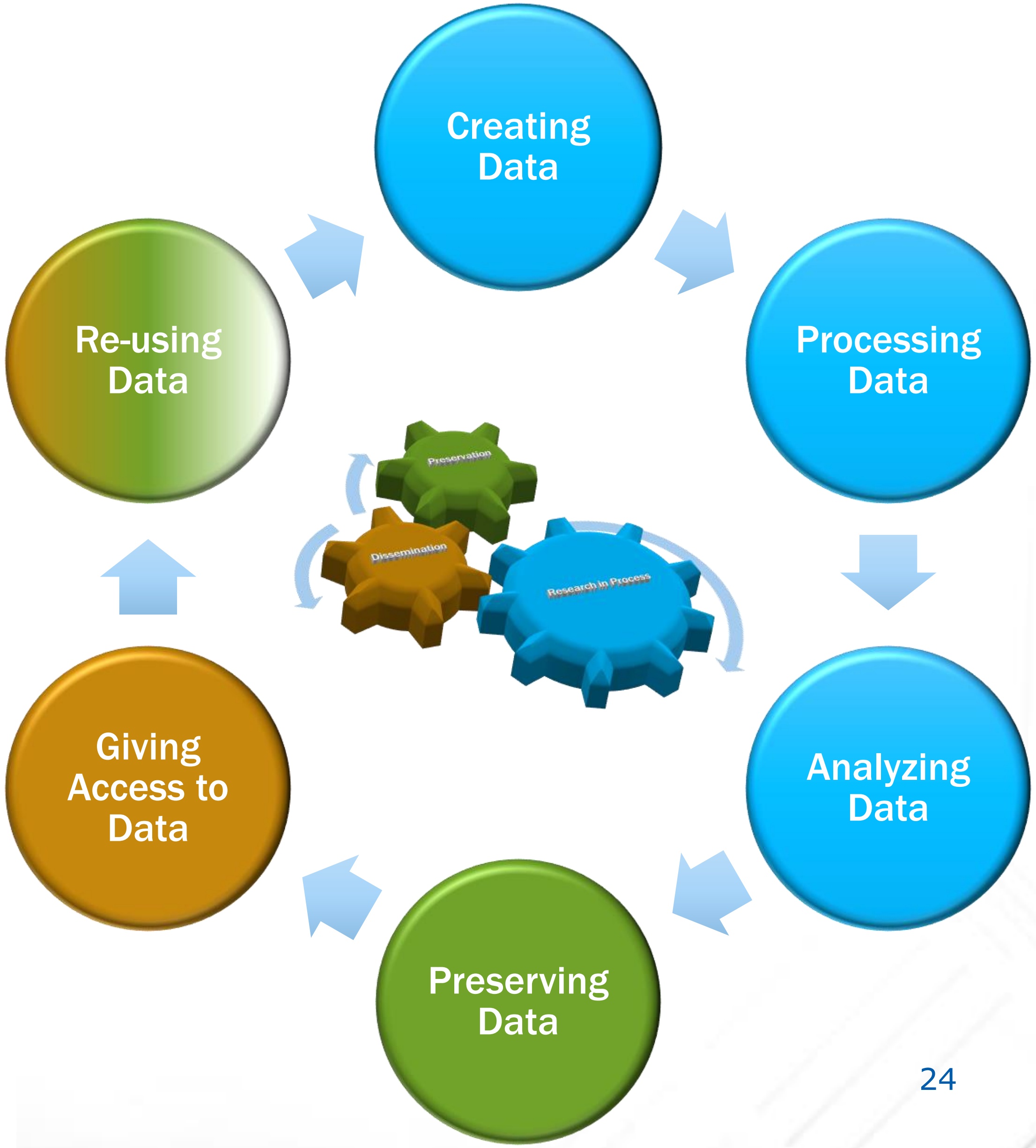
“Enterprise Content Management (ECM) is a formalized means of organizing and storing an organization's documents, and other content, that relate to the organization's processes. The term encompasses strategies, methods, and tools used throughout the lifecycle of the content”

(What is Enterprise Content Management (ECM)?". AIIM. Association for Information and Image Management)

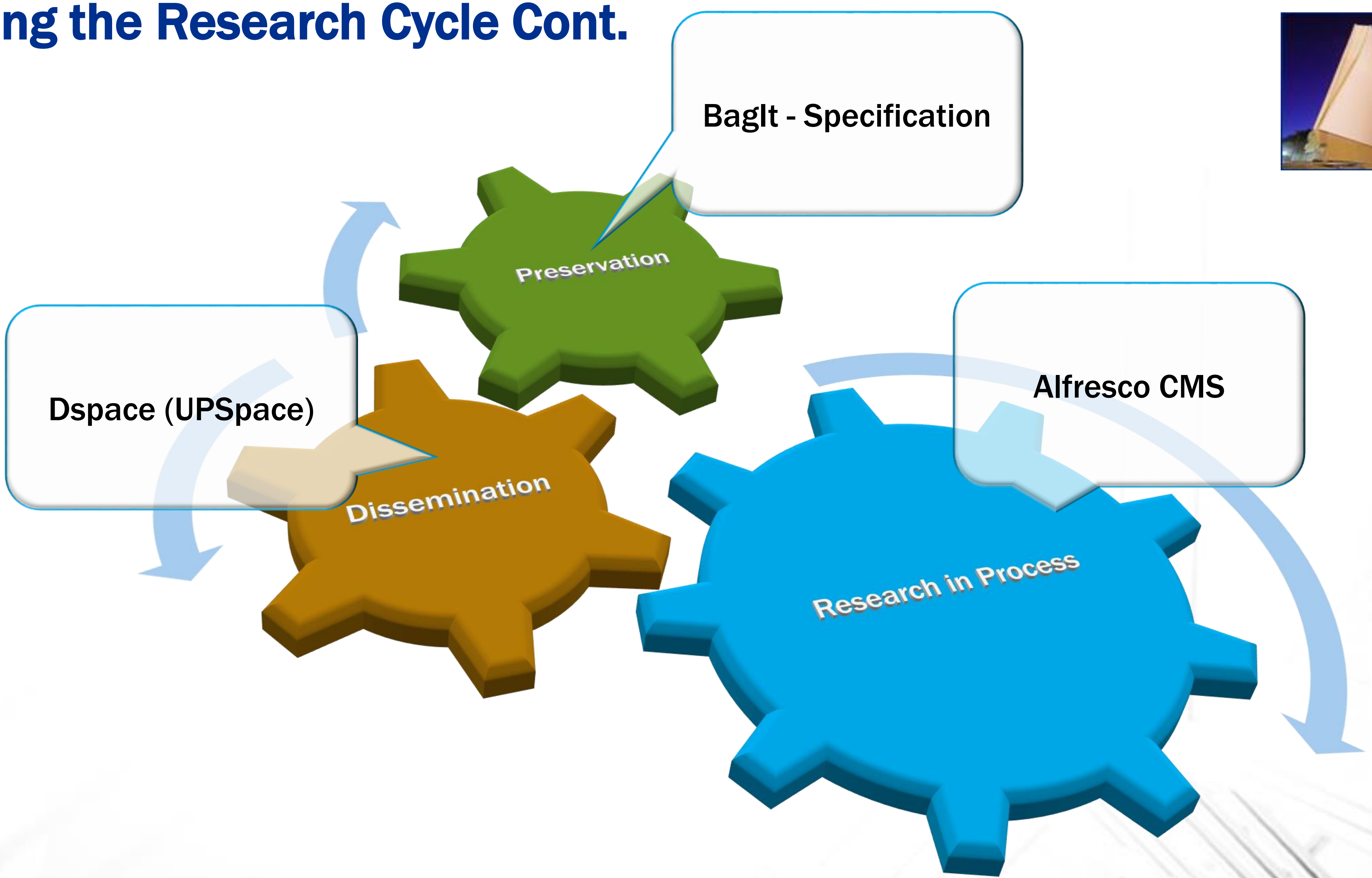
To Note:

- **Enterprise content management is not a closed-system solution or a distinct product category.**
- **Focus should be focused on your environment.**

Supporting the Research Data Lifecycle



Supporting the Research Cycle Cont.



Alfresco – Case Study 1

NASA Langley Research Center

The Challenge

The NASA Langley Research Center conducts hundreds of tests each year designed to make aircraft and spacecraft safer and more efficient. These tests – which are performed by different teams of engineers, researchers, technicians, managers and customers – are highly collaborative, as teams need to be able to share ideas and view each other’s test documentation.

NASA built a homegrown collaborative portal, aeroCOMPASS, which allowed Langley to create individual team sites for sharing and commenting on documents, notes and other research files. But after 10 years of use, the aeroCOMPASS software had become outdated and could no longer meet NASA’s strict security guidelines.

With over **800 sites in use**, NASA needed to move aeroCOMPASS to a new collaboration and document management environment with a similar look and feel, but using a **more secure, modern architecture.**

“We have users coming into the system from all over including researchers, Lockheed Martin engineers, and researchers wanting to build teams. When they come back to the system, all of their documents and research is there waiting for them.” — David Cordner, IT Architect, Research Directorate

Alfresco – Case Study 2

KLM

A key objective of the project was to keep document management simple for users, while meeting the company's technical and budget requirements. Solutions evaluated by KLM needed to provide users with the following framework:

- Personal documents – a personal document management space accessible to all employees on the Web, **anywhere at any time**, with Web Attached Secure Storage Anywhere (WASSA);
- **Project/team documents** – a place for project, departmental or team documents to be shared, stored and collaborated on;
- **Company shared documents** – a company-wide document repository that included search capabilities without a complex taxonomy; and
- Operational documents – a **structured repository with restricted authorization** and only the last version of operational documents that included taxonomy.

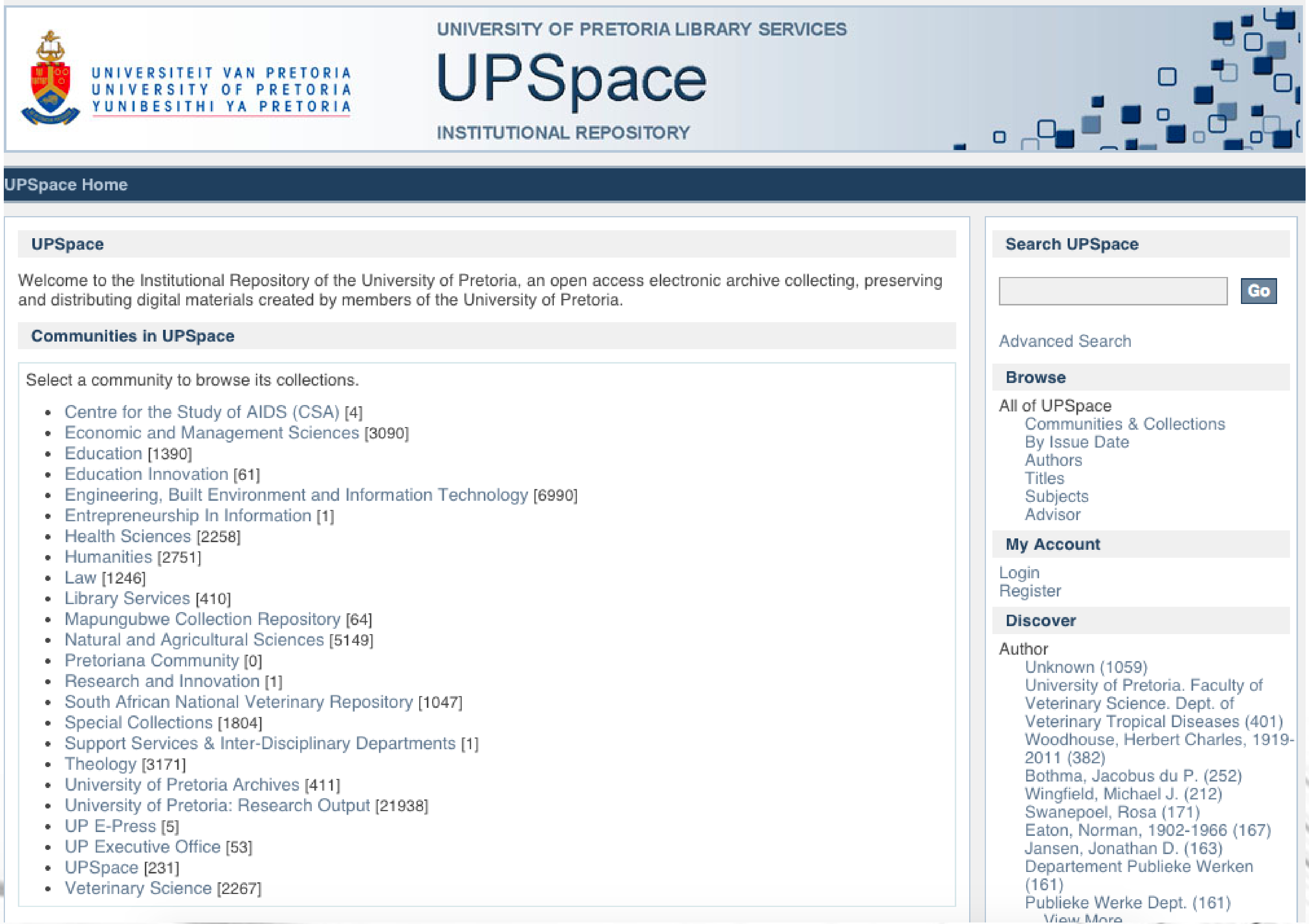
"Alfresco was easy to implement on our standard infrastructure and was rolled out to all 30,000 employees within six months. Any preconceived ideas we had about working with an open source vendor were quickly dismissed. In terms of technical support, working with Alfresco was just like engaging in any commercial service level agreement, but with the added benefits of the open source architecture and cost structure."
Pieter Janssen, Chief Architect at KLM

Alfresco – Research in Process Demo

[Institute For Cellular and Molecular Medicine]

DSpace – Dissemination

DSpace – Dissemination



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

UNIVERSITY OF PRETORIA LIBRARY SERVICES

UPSpace

INSTITUTIONAL REPOSITORY

UPSpace Home

UPSpace

Welcome to the Institutional Repository of the University of Pretoria, an open access electronic archive collecting, preserving and distributing digital materials created by members of the University of Pretoria.

Communities in UPSpace

Select a community to browse its collections.

- Centre for the Study of AIDS (CSA) [4]
- Economic and Management Sciences [3090]
- Education [1390]
- Education Innovation [61]
- Engineering, Built Environment and Information Technology [6990]
- Entrepreneurship In Information [1]
- Health Sciences [2258]
- Humanities [2751]
- Law [1246]
- Library Services [410]
- Mapungubwe Collection Repository [64]
- Natural and Agricultural Sciences [5149]
- Pretoriana Community [0]
- Research and Innovation [1]
- South African National Veterinary Repository [1047]
- Special Collections [1804]
- Support Services & Inter-Disciplinary Departments [1]
- Theology [3171]
- University of Pretoria Archives [411]
- University of Pretoria: Research Output [21938]
- UP E-Press [5]
- UP Executive Office [53]
- UPSpace [231]
- Veterinary Science [2267]

Search UPSpace

Go

Advanced Search

Browse

All of UPSpace
- Communities & Collections
- By Issue Date
- Authors
- Titles
- Subjects
- Advisor

My Account

Login
Register

Discover

Author
- Unknown (1059)
- University of Pretoria. Faculty of Veterinary Science. Dept. of Veterinary Tropical Diseases (401)
- Woodhouse, Herbert Charles, 1919-2011 (382)
- Bothma, Jacobus du P. (252)
- Wingfield, Michael J. (212)
- Swanepoel, Rosa (171)
- Eaton, Norman, 1902-1966 (167)
- Jansen, Jonathan D. (163)
- Departement Publieke Werken (161)
- Publieke Werke Dept. (161)
- View More

Preservation – Bagit Specification

What is BagIt ?

- „BagIt is a hierarchical file packaging format designed to support disk-based storage and network transfer of arbitrary digital content.“

- „A "bag" consists of a "payload" (the arbitrary content) and "tags", which are metadata files intended to document the storage and transfer of the bag.“

- „They are also well-suited to the export, for archival purposes, of content normally kept in database structures that receiving parties are unlikely to support.“

Contents of a Bag

1. **Data Directory**
    - Contains the data payload
    - Can be single or multiple files and directories

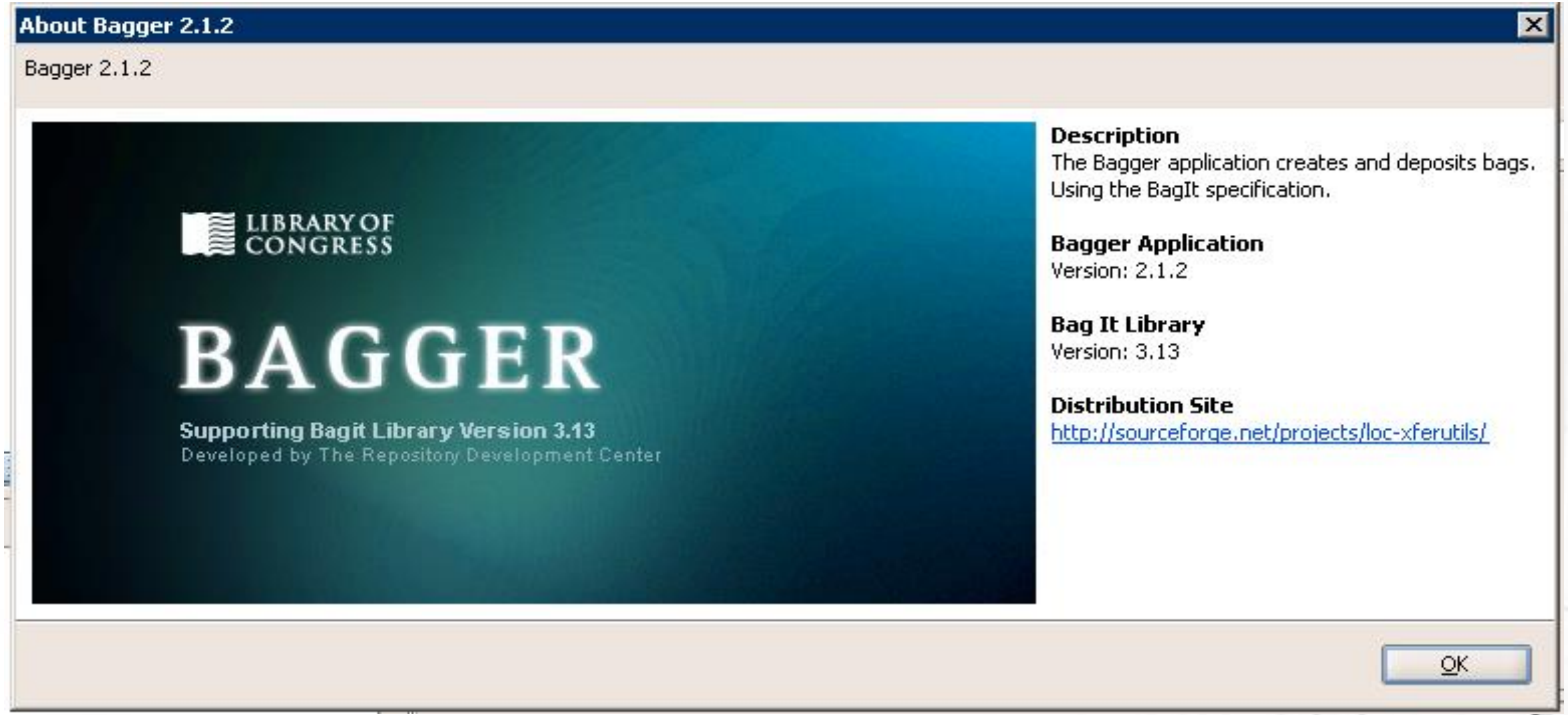
2. **Manifest File**
    - Text file with listed items in the payload with their checksums

3. **BagIt File**
    - Contains info on BagIt version and encoding

4. **BagIt Info File**
    - Contains the Metadata for the bag

5. **Tag Manifest File**
    - Contains checksums to verify the above mentioned txt files

Example



Bagger 2.1.2

Supporting Bagit Library Version 3.13
Developed by The Repository Development Center

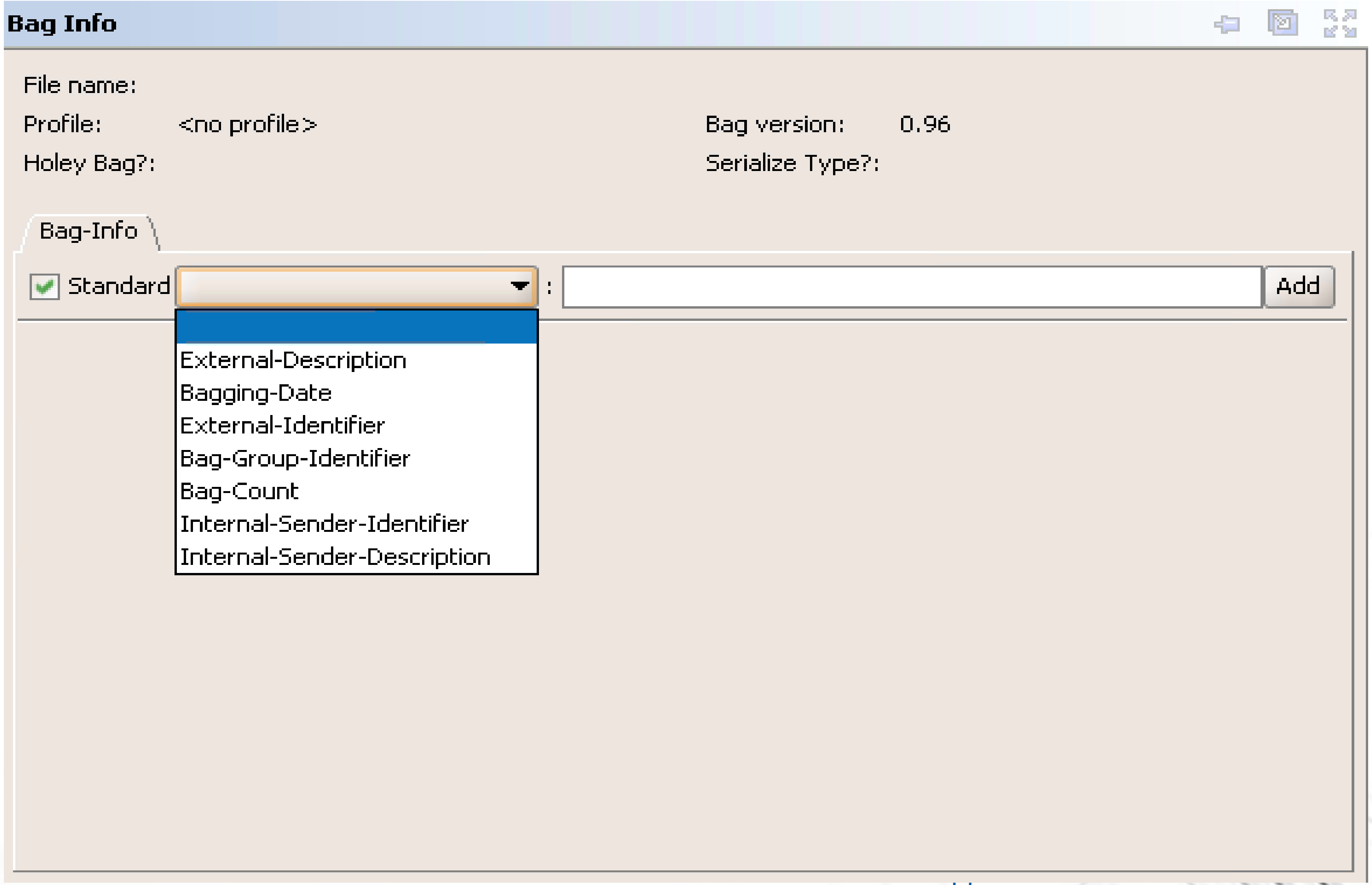
Description
The Bagger application creates and deposits bags.
Using the BagIt specification.

Bagger Application
Version: 2.1.2

Bag It Library
Version: 3.13

Distribution Site
http://sourceforge.net/projects/loc-xferutils/

Example



Bag Info

File name:

Profile: <no profile>

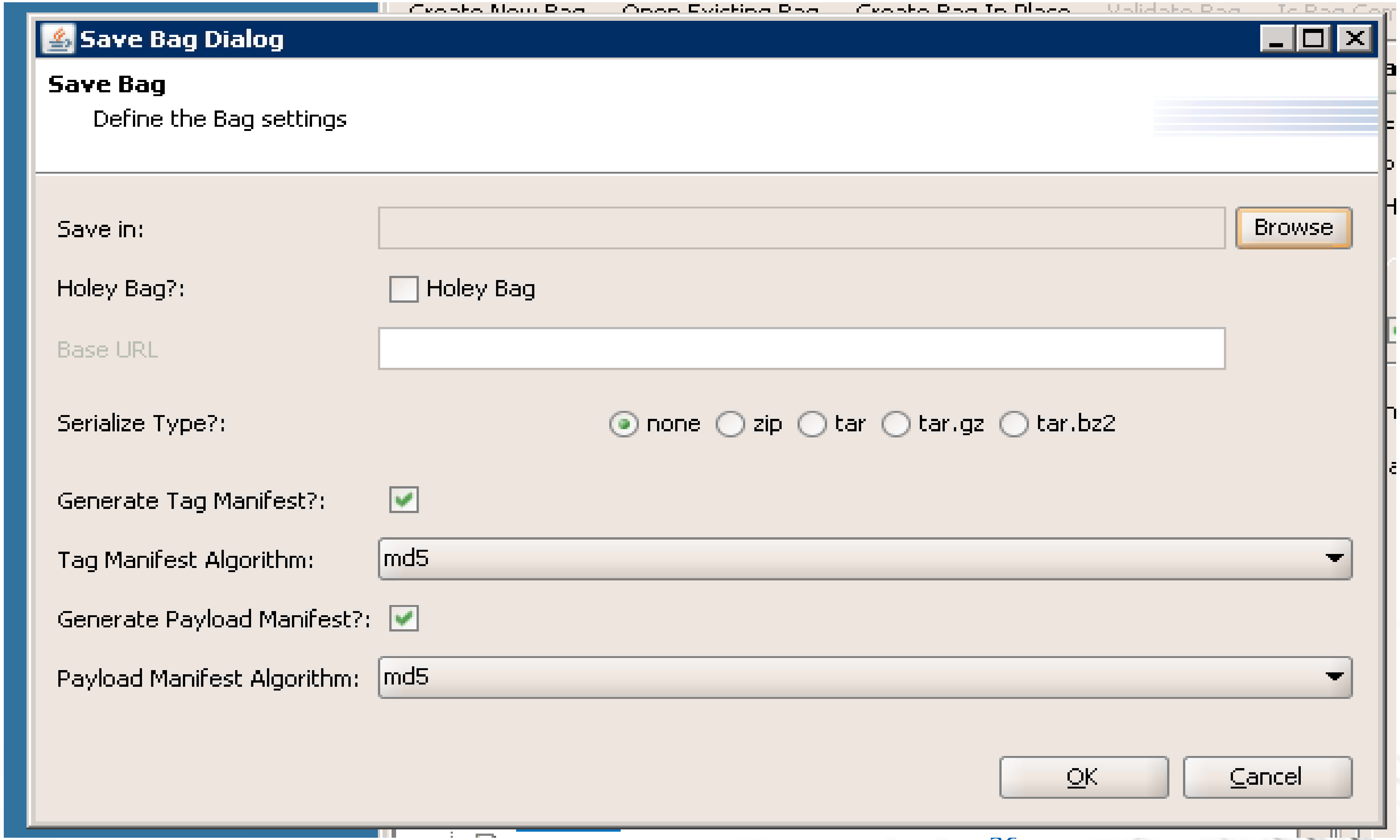
Bag version: 0.96

Holey Bag?:

Serialize Type?:

☑ Standard

Example



Save Bag

Define the Bag settings

Save in: [Browse]

Holey Bag?: □ Holey Bag

Base URL

Serialize Type?: ● none ○ zip ○ tar ○ tar.gz ○ tar.bz2

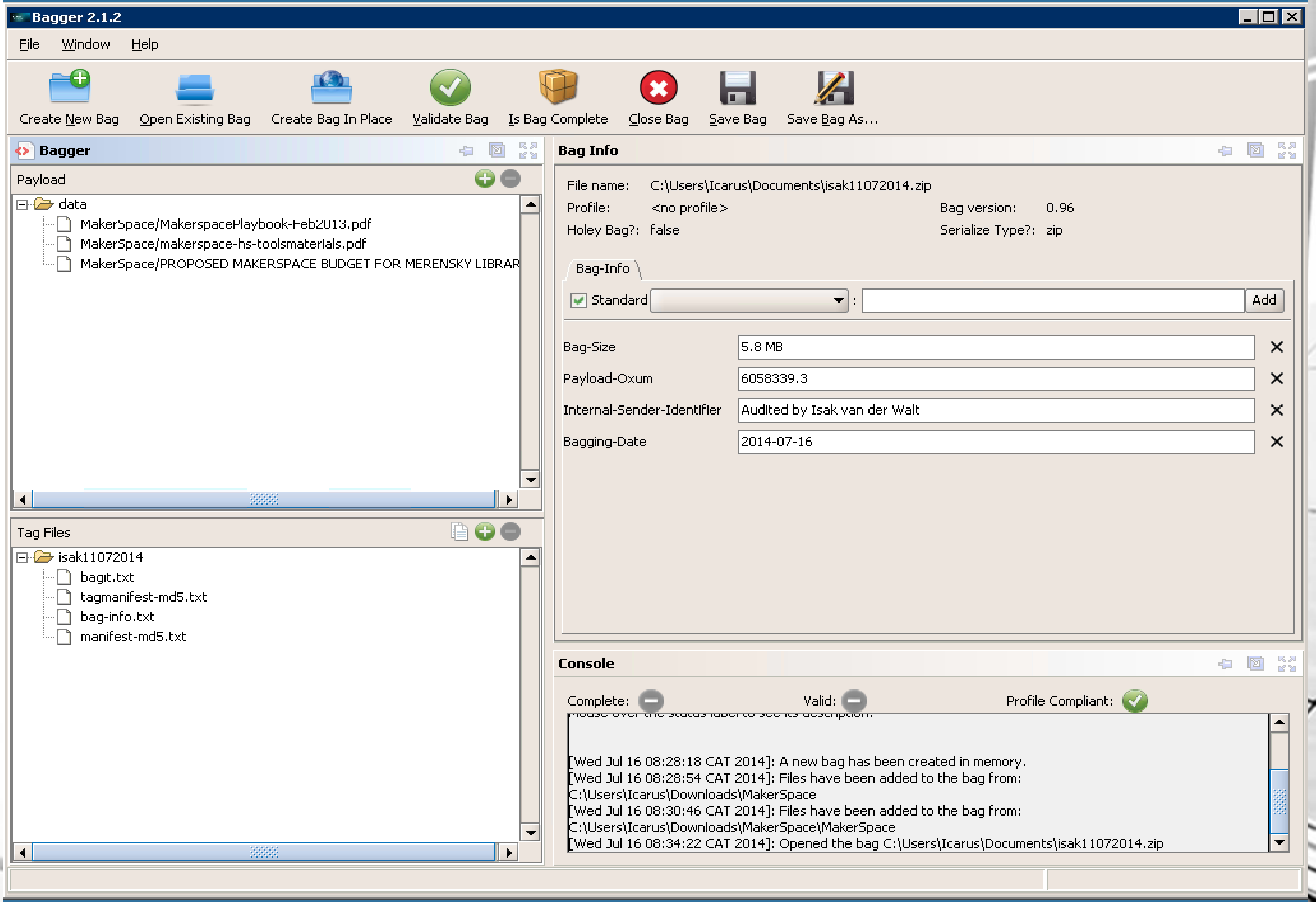
Generate Tag Manifest?: ☑

Tag Manifest Algorithm: md5

Generate Payload Manifest?: ☑

Payload Manifest Algorithm: md5

Example



Payload

- data
  - MakerSpace/MakerspacePlaybook-Feb2013.pdf
  - MakerSpace/makerspace-hs-toolsmaterials.pdf
  - MakerSpace/PROPOSED MAKERSPACE BUDGET FOR MERENSKY LIBRAR

Tag Files

- isak11072014
  - bagit.txt
  - tagmanifest-md5.txt
  - bag-info.txt
  - manifest-md5.txt

File name:	C:\Users\Icarus\Documents\isak11072014.zip		
Profile:	<no profile>	Bag version:	0.96
Holey Bag?:	false	Serialize Type?:	zip

☑ Standard :

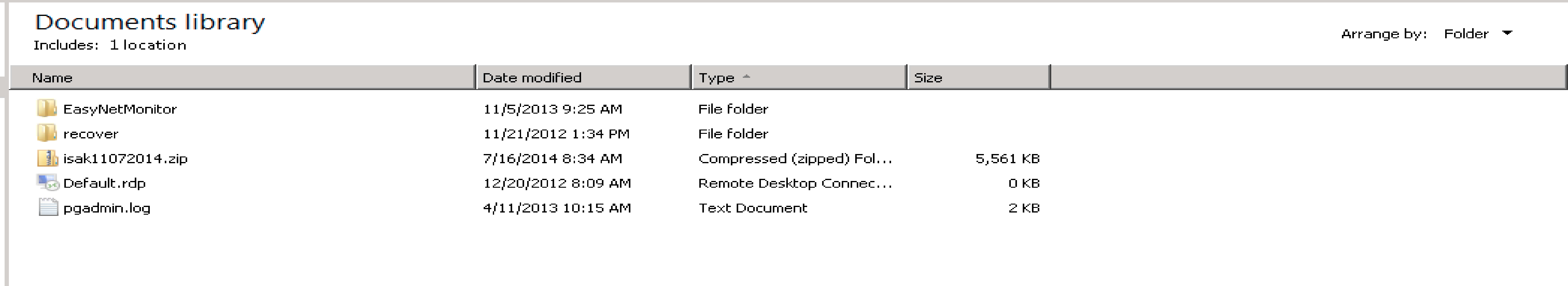
Bag-Size	5.8 MB
Payload-Oxum	6058339.3
Internal-Sender-Identifier	Audited by Isak van der Walt
Bagging-Date	2014-07-16

Complete: ⊖ Valid: ⊖ Profile Compliant: ✓

```
Mouse over the status label to see its description.

[Wed Jul 16 08:28:18 CAT 2014]: A new bag has been created in memory.
[Wed Jul 16 08:28:54 CAT 2014]: Files have been added to the bag from:
C:\Users\Icarus\Downloads\MakerSpace
[Wed Jul 16 08:30:46 CAT 2014]: Files have been added to the bag from:
C:\Users\Icarus\Downloads\MakerSpace\MakerSpace
[Wed Jul 16 08:34:22 CAT 2014]: Opened the bag C:\Users\Icarus\Documents\isak11072014.zip
```

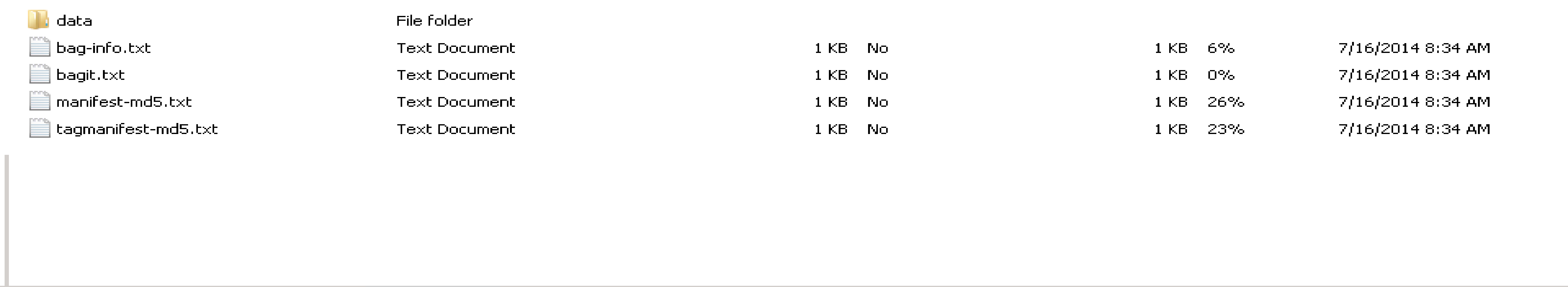
Example



Documents library

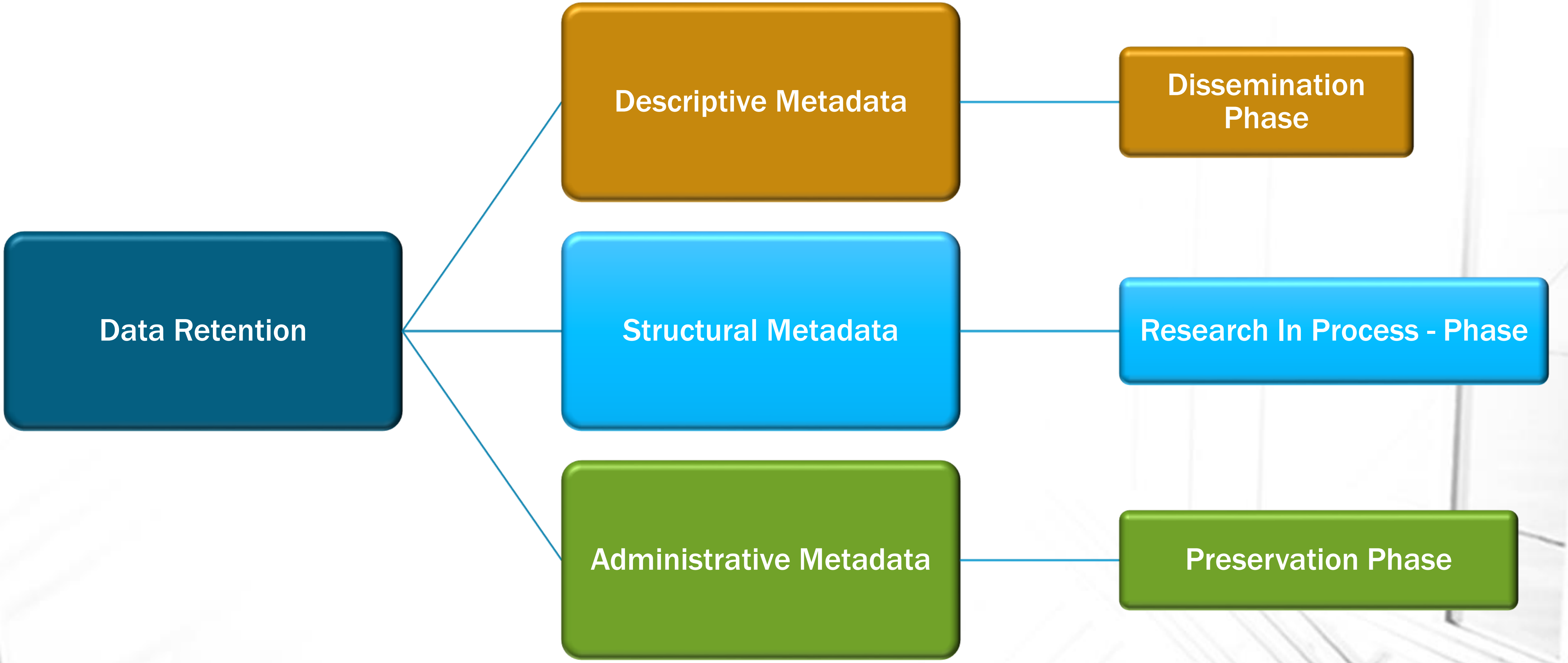
Includes: 1 location

Name	Date modified	Type	Size
EasyNetMonitor	11/5/2013 9:25 AM	File folder	
recover	11/21/2012 1:34 PM	File folder	
isak11072014.zip	7/16/2014 8:34 AM	Compressed (zipped) Fol...	5,561 KB
Default.rdp	12/20/2012 8:09 AM	Remote Desktop Connec...	0 KB
pgadmin.log	4/11/2013 10:15 AM	Text Document	2 KB



data	File folder					
bag-info.txt	Text Document	1 KB	No	1 KB	6%	7/16/2014 8:34 AM
bagit.txt	Text Document	1 KB	No	1 KB	0%	7/16/2014 8:34 AM
manifest-md5.txt	Text Document	1 KB	No	1 KB	26%	7/16/2014 8:34 AM
tagmanifest-md5.txt	Text Document	1 KB	No	1 KB	23%	7/16/2014 8:34 AM

Metadata Captured During The Data Lifecycle



Data Retention

Descriptive Metadata

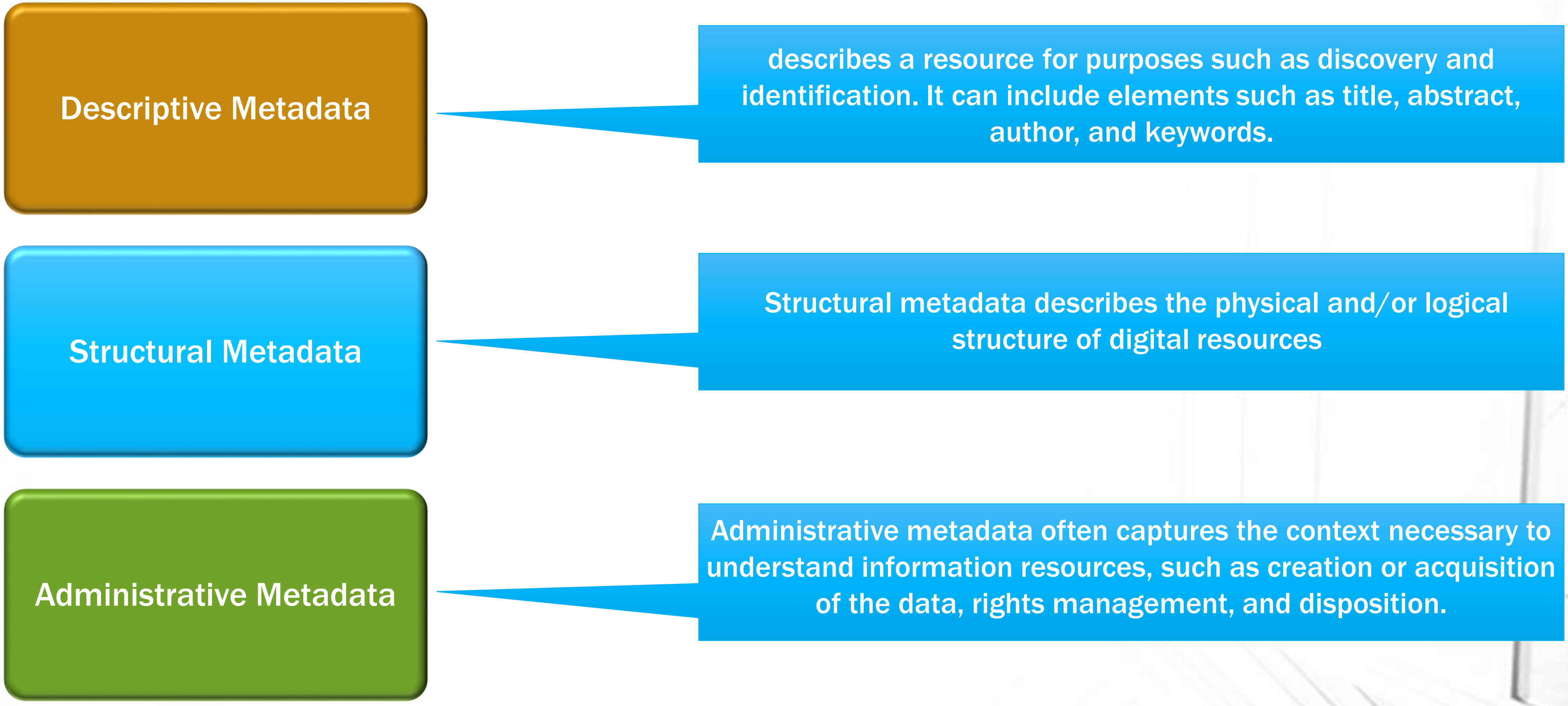
Structural Metadata

Administrative Metadata

Dissemination Phase

Research In Process - Phase

Preservation Phase



Descriptive Metadata

describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

Structural Metadata

Structural metadata describes the physical and/or logical structure of digital resources

Administrative Metadata

Administrative metadata often captures the context necessary to understand information resources, such as creation or acquisition of the data, rights management, and disposition.

Levels of digital preservation

Level 1 – Protect Your Data

Level 2 – Know Your Data

Level 3 – Monitor Your Data

Level 4 – Repair Your Data

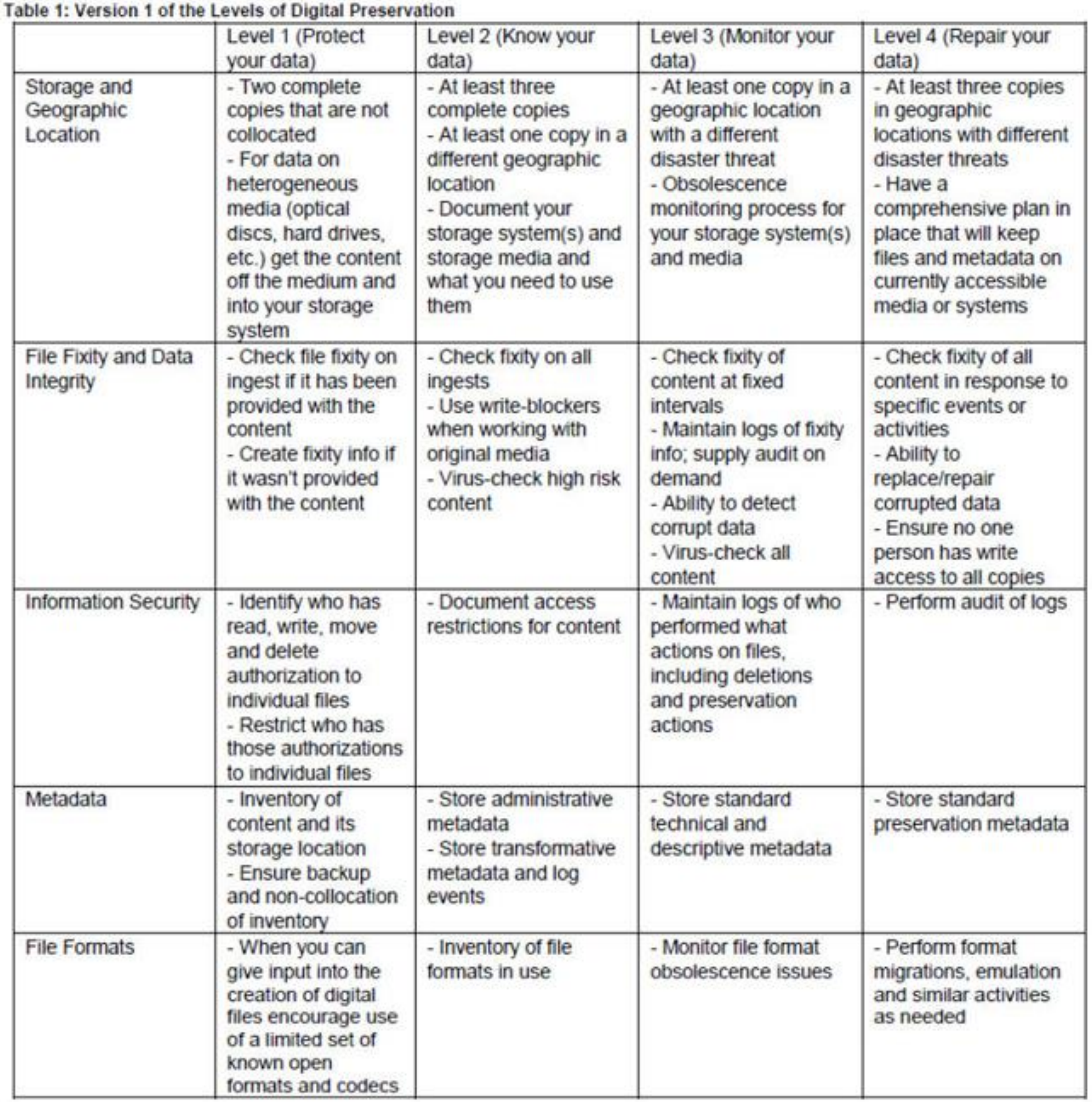


Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

Further Development

Identify a Campus-wide database /repository for the publishing of open access data sets

e.g. Dspace, Fedora, Purr

Further Development

- **Investigation and development of Data Publishing Platform**
  - Fedora (DuraSpace Organization)
- **Creation of DMP (data management planning) tool**
- **Automatization of certain processes**

Hurdles

- **IT infrastructure support**
- **Understanding of the concept and processes**

References

- ***Alfresco – NASA Langley Research Center Case Study.*** [Online] available at http://www.alfresco.com/customers/nasa-langley-research-center (Accessed 19 November 2014).

- ***Alfresco – KLM Case Study.*** [Online] available at http://www.alfresco.com/customers/klm (Accessed 19 November 2014).

- CORTI, L. et al. 2014. ***Managing and sharing research data: a guide to good practice***. Los Angeles: SAGE.

- ***Data Management Planning Tool (DMPTool).*** Oakland, CA: University of California Curation Center of the California Digital Library, 2014. [Online] available at https://dmptool.org/ (Accessed 24 September 2014).

- DE WAARD, A. AND ROTMAN, D. AND LAURUHN, M. 2014. Research data management at institutions: part 1: visions. ***Elsevier Library Connect***, 6 February 2014. [Online] available at http://libraryconnect.elsevier.com/articles/2014-02/research-data-management-institutions-part-1-visions (Accessed 5 October 2014)

References

- ***DMPonline tool.*** Edinburgh, UK: Digital Curation Centre, 2014. [Online] available at https://dmponline.dcc.ac.uk/ (Accessed 22 September 2014).

- PIENAAR, H. AND VAN DEVENTER M. 2009. To VRE Or Not to VRE?: Do South African Malaria Researchers Need a Virtual Research Environment? ***Ariadne***, Issue 59. [Online] available at http://www.ariadne.ac.uk/issue59/pienaar-vandeventer/ (Accessed 13 November 2014).

- UK DATA ARCHIVE. 2014. ***Research Data Lifecycle.*** Colchester Essex: UK Data Archive, University of Essex. [Online] available at http://www.data-archive.ac.uk/create-manage/life-cycle (Accessed 13 November 2014)

- ***What is Enterprise Content Management (ECM)?*** 2010. Silver Spring. MD: AIIM (Association for Information and Image Management). [Online] available at http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx (Accessed 20 September, 2010)

- ***Table 1: Version of the levels of preservation NDSA Levels of Preservation.*** [Online] available at http://www.digitalpreservation.gov/ndsa/activities/levels.html (Accessed 19 November 2014)

Thank You