Criteria and evaluation of research data repository platforms @ the University of Pretoria, South Africa

Presented by Mr. Johann van Wyk & Mr. Isak van der Walt
Library Services
University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA

Project team
UP IT: Karin, Yzelle, Herman
UP Library: Isak, Johann, Heila

Agenda

- Project Scope & project team
- Research data lifecycle
- E-Research Framework
- Product Investigation
- Criteria & evaluation
- Recommendations
- Next Steps
- Documents produced

Project Scope

The scope of the project was to evaluate products (commercial and open source) which could be utilised as a **Research Data Repository Platform** as part of a total Research Data Management (RDM) solution at UP.

A total RDM solution include all phases of the Research data life cycle, but for the repository solution, the focus was thus on identifying a potential solution for the “Dissemination” phase of the research data life cycle.

RDM Repository Project Team

Business Sponsor – Prof Stephanie Burton (VP: Research)

ITS Sponsor – Andre Kleynhans (Deputy Director: ITS)

Project Team members:

ITS Project Manager and Business Analyst – Karin Meyer

ITS Infrastructure Architect - Dr Yzelle Roets

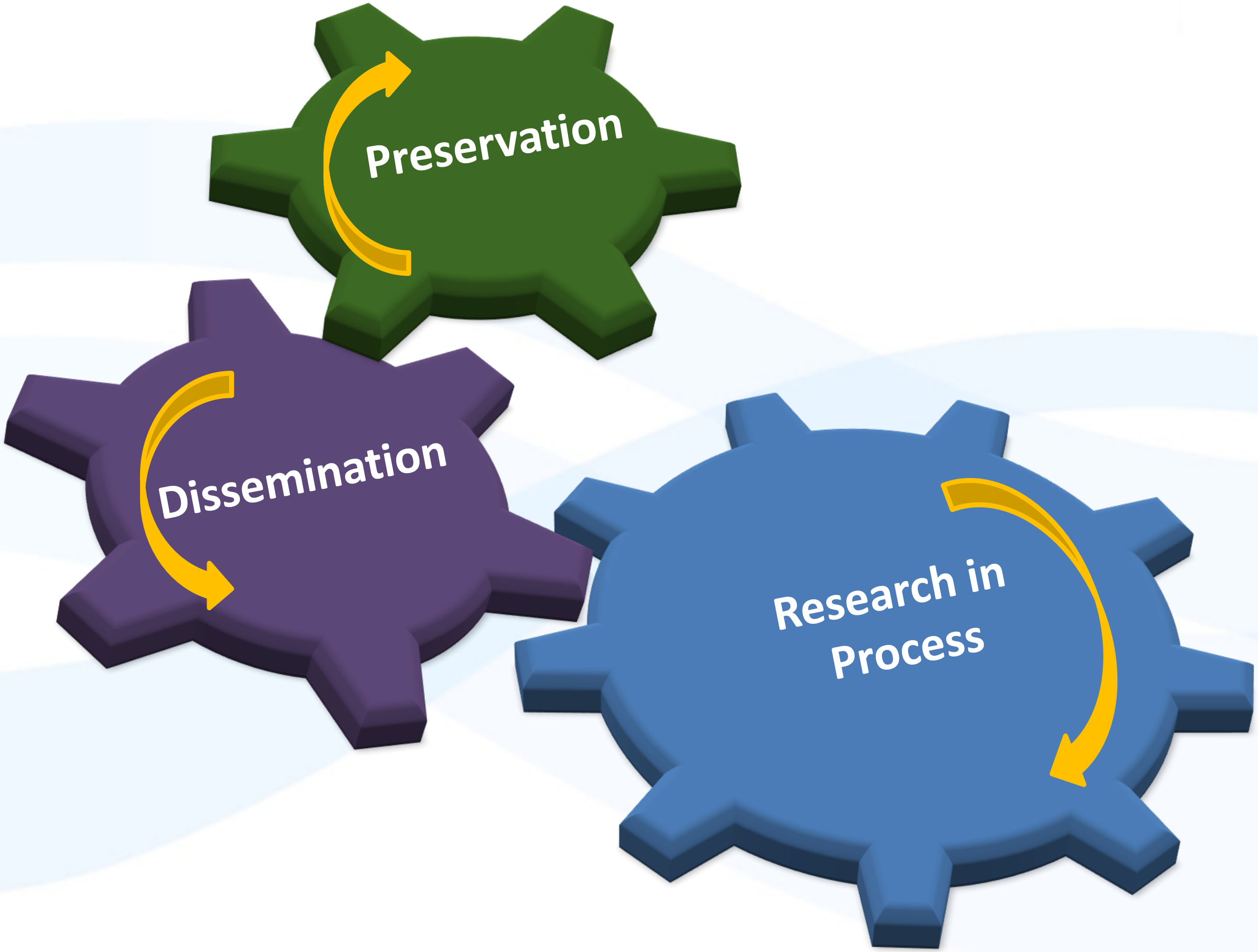
ITS eResearch Support Manager – Herman Jacobs

Library Services: Senior IT Consultant – Isak van der Walt

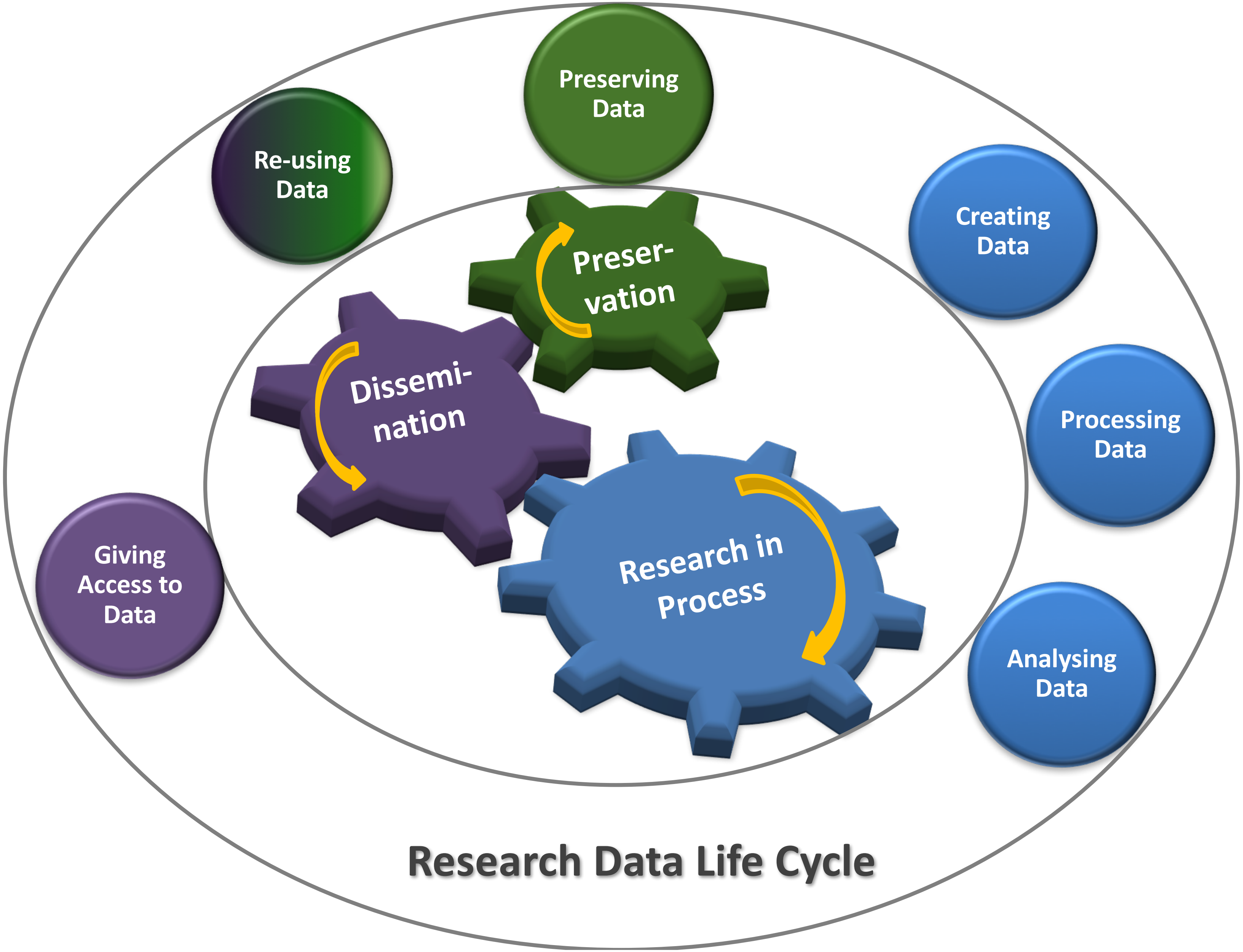
Library Services: Assistant Director: RDM – Johann van Wyk

Library Services: Deputy Director: Strategic Innovation – Dr Heila Pienaar

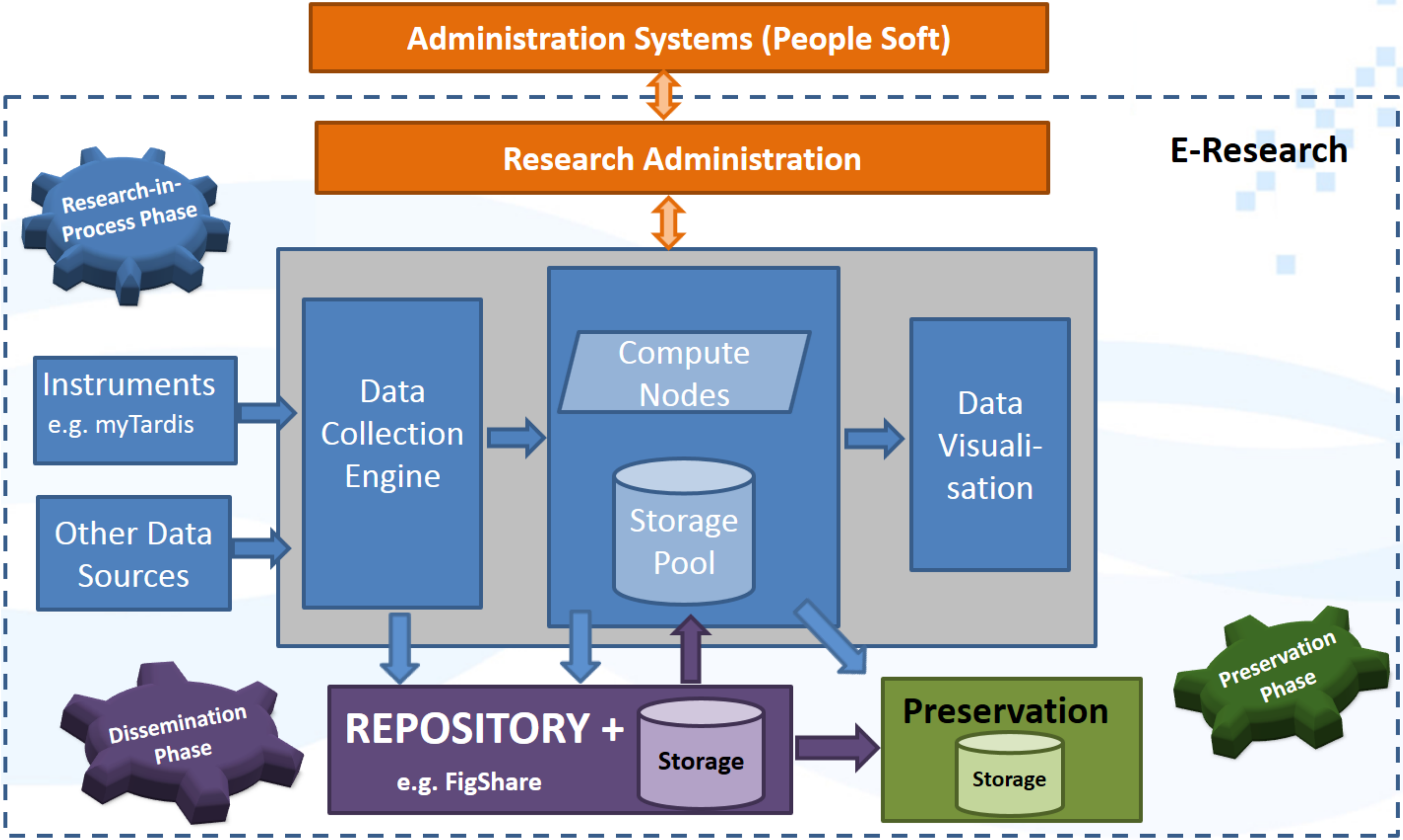
DATA FLOW within the RESEARCH DATA LIFE CYCLE



PROCESSES within the RESEARCH DATA LIFE CYCLE



eResearch Framework



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Product Investigation Methodology

Finalisation of product evaluation criteria

- Consulted with various stakeholders
    - Library and ITS staff
    - External stakeholders at the NEDICC workshop held at the CSIR
    - Peer Universities
- Utilised **various selection criteria** from other institutions e.g. Leeds University, Texas Digital Library and the RDA RPRD IG Matrix (http://tinyurl.com/RPRD-matrix) selection criteria as a basis and adapted it according to UP specific requirements.

Product Short Listing

Products were short listed based on the following:

- Product scan of products being used internationally, and
- Most commonly used products at universities similar to UP (size and research activity).

Product Evaluation

- UP's formal Request For Information (RFI) process was followed
- Product evaluation criteria list was compiled and send to short listed vendors together with standard RFI documentation
- The requested information was received from the vendors and prepared for scoring, and
- Products were scored and evaluated.

Evaluation Criteria

- Functional / Business criteria: Deposit and Upload; Re-Usability; Identity and Access Management; Reporting; Discovery; Preservation
- Non Functional: Repository Architecture; Data Management; Data Governance
- Technical aspects: Back-end Management; Integration; Infrastructure
- Vendor specific: Support, Training, Usage of Product
- Performance requirements
- Integration requirements

Unique ID	Requirement Description	Priority
DU-1	Offer customisable metadata schema as per research area or discipline (including mandatory fields).	H
DU-2	Offer the indexing of metadata.	H
DU-3	Offer sufficient support for geospatial and journal article metadata. Support association of single or multiple files with one metadata record.	H
DU-4	Upload and store metadata at a data object level, where a data object is a folder that contains one or more files.	M
DU-5	Support multiple file types and formats of data, e.g. MS Excel 2007, MySQL database, raw data file from a Campbell CR10 data logger, any multimedia, etc.	H
DU-6	The system should have a simple process for uploading large (multi-TB) data sets, potentially consisting of thousands of files. Must have the ability to upload large data sets (e.g. 2MB, 2 GB, 1 TB).	H
DU-7	Support controlled lists against some metadata fields, either held locally or drawn from an external source e.g. Subject vocabularies.	H
DU-8	Support customisation of out-of-the-box help text and provide context sensitive feedback for the depositor e.g. Highlight missing metadata fields, file upload failure alert.	M
DU-9	Accommodate workflow where data needs to be destructed with an approval process and audit trail.	L
DU-10	Researchers must be able to submit data to repository themselves.	H
DU-11	Process of submitting data to a repository from other systems/instruments.	H
DU-12	Ability to batch upload data into a repository.	H
DU-13	Third party must be able to upload dataset on behalf of researcher.	H
DU-14	Support generation / labelling of persistent unique identifiers for datasets including DOIs.	H
DU-15	Ability to support the submission of data at any research stage (i.e. Initial Data, Working Data, Final Data Stages) to the repository.	M
DU-16	Explain how user interface customisation is achieved.	H
DU-17	Out-of-the-box user interface intuitive (easy to use) to users.	M
DU-18	Out-of-the-box user interface meets accessibility requirements, e.g. W3C WCAG 1.	H
DU-19	Assignment of Intellectual Property (IP) rights and multiple content licensing options with terms and conditions exposed clearly human and machine re-users is possible, such as copyright and creative commons (CC).	H

Table 1: Deposit and Upload functional criteria

Shortlisted Products & RFI Feedback

Product	Vendor / Implementation Partner	RFI Feedback
DSpace	Atmire	Received information on criteria list, proposed implementation options and its associated cost.
Figshare	Digital Science	Received information on criteria list, proposed implementation options and its associated cost.
Islandora	Discoverygarden	Received information on criteria list, proposed implementation options and its associated cost.
Dataverse	Harvard University	Received insufficient information on criteria list, implementation options and cost.
PURR	Purdue University	Failed to respond to RFI.
Redbox	Queensland Cyber Infrastructure Foundation (QCIF)	Received information on criteria list, but Redbox is **only a meta data repository** and **not a data repository**.

Implementation options with most important advantages / disadvantages – Option 1

Option	Advantages	Disadvantages
Option 1 - Locally hosted (both application and storage are locally hosted at UP)	• UP not dependent on internet for access to application • UP able to manage own data • Compliance to legal issues regarding data, i.e. POPI Act • Risk of security is lower (control own storage)	• Resources to be provided (includes Infrastructure and Human resources for application and storage) which increase cost • Required skills set (e.g. web skills) is limited or not currently available in ITS • UP bandwidth will cause restrictions, i.e. indexing of site • Open source product - no legal entity/responsible company for assistance, support, enhancements, new releases, etc.

Implementation options with most important advantages / disadvantages – Option 2

Option	Advantages	Disadvantages
Option 2 - Hybrid (application is cloud hosted, while the storage is locally hosted)	• Collaboration with other institutions in future is easier • No additional resources (HR or infrastructure) are required for the application • Legal entity exist i.e.. the application • Geographic redundancy • High availability on the UP front end – no bandwidth constraints • Meta data as well as data will be always available, searchable and able to be indexed • UP will be in control of their IP (control own storage) • Risk of security will be lower (control own storage)	• Resources to be provided which includes infrastructure and human resources for storage as well as RD, backups, access control, cooling, etc. • Required skills set (e.g. web skills) is limited or not currently available in ITS • Indexing of site dependent on UP's bandwidth

Implementation options with most important advantages/ disadvantages – Option 3

Option	Advantages	Disadvantages
Option 3 - Fully cloud-based (both the application and storage are cloud hosted through the vendor)	• Collaboration with other institutions in future is easier • No additional resources (HR or infrastructure) are required for the application • Legal entity exist i.e. the application • Geographic redundancy • High availability on the UP front end – no bandwidth constraints • Meta data as well as data will be always available, searchable and able to be indexed • UP will be in control of their IP (control own storage) • Risk of security will be lower (control own storage)	• UP does not have control of IP (governance and accessibility to UP's data is in the hands of the vendor) • Possible future sanctions against some countries may result in some users from other parts of the world not being able to reach UP's repository • Growing running cost as UP will have to pay for up-and downloading as well as storage of data

Product Evaluation Results

Criteria	Figshare	Islandora	DSpace
BEEEE	All products and associated vendors/implementation partners are internationally based , therefore no weight was assigned in the scoring exercise.		
Requirements Criteria (incl functional, non-functional, vendor)	85% fit	96% fit	65% fit
Pricing	CONFIDENTIAL		
Preferential criteria: Hybrid Option (option 2)	100% Fit	10% fit – only available through huge custom development which poses huge risks to UP.	0% Fit
Preferential criteria: Consortial pricing	100% Fit	0% fit	0% fit

Recommendations

The following is recommended for implementing of a Research Data Repository platform) solution at UP:

- **Figshare** should be considered as the product of choice
- Implement the **Hybrid** implementation option with the application being cloud hosted and a local storage of 20Tb to start with
- Local storage can be supplemented in future with Cloud storage
- Storage should be investigated in line with the total eResearch initiative and framework of UP
- A business owner needs to be identified to be responsible for a total RDM implementation
- Implementation of a Research Data Repository platform will require a significant increase in Human and Infrastructure Resource components, and
- Consortial pricing can be kept in mind for the future and was not used as a determining selection criterion.

Next Steps

- Appoint a Business owner(s) for a total RDM solution
- Investigate tools that can support the Research-in-Process phase, e.g. myTardis
- Finalise storage solution (eg. African Research Cloud)
- **Business Case to secure resources (financial and human)**
- **Implementation of repository solution**
- **Training of researchers & library staff**

Gap analysis: Figshare (obtained 0 on these criteria)

Functional criteria:

- Must be able to change data formats, although most formats are agnostic.
- Auto-generate preservation metadata, e.g. PREMIS.
- Ability to migrate files in datasets to new/other formats over time.
- Be compliant with the OAIS (Open Archival Information System) reference model.

Non-functional criteria:

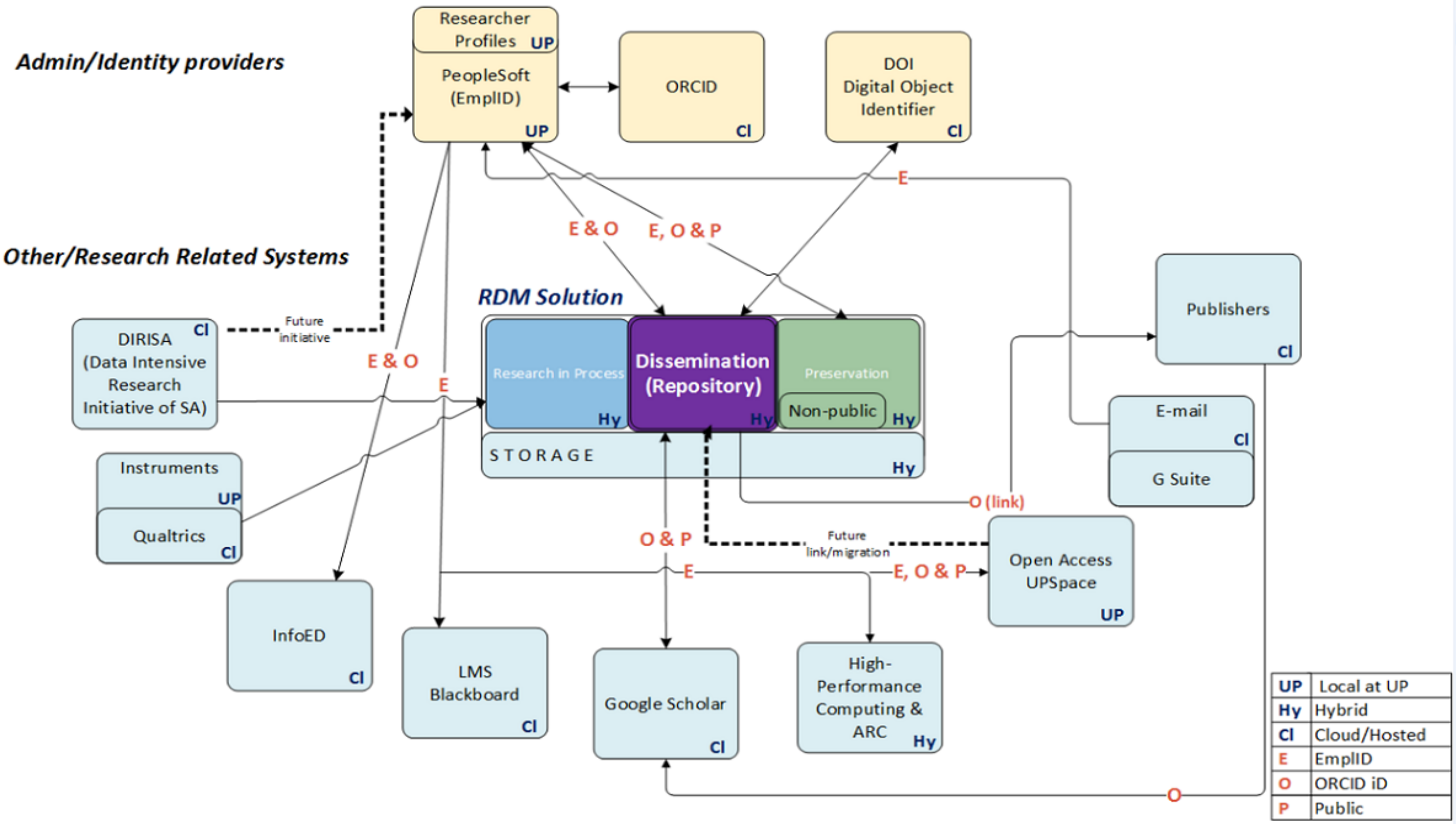
Offer de-duplication of data, metadata

Disadvantages:

- The annual subscription fee for Figshare is relatively high
- Customisation is not possible as it is a proprietary product
- The proprietary product aspect also limits the look and feel customisation of the product to reflect more of UP's footprint, and
- No local support exists within South Africa.

Context Diagram: Research Data Management

Context Diagram: Research Data Management (RDM)



UP	Local at UP
Hy	Hybrid
Cl	Cloud/Hosted
E	EmplID
O	ORCID iD
P	Public

Data can be open for public access or embargoed for public use within the repository, preservation system, researcher profile, ORCID and DOI.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Documents

- UP Research Data Repository Evaluation
- UP Research Data Management Business Requirements Specification
- Executive summary
- RDM Project Progress Feedback
- Context Diagram for RDM
- Islandora, Figshare, Redbox, DSpace, Dataverse, PURR requirements criteria feedback documents

Still a lot of ground to cover

Thank You