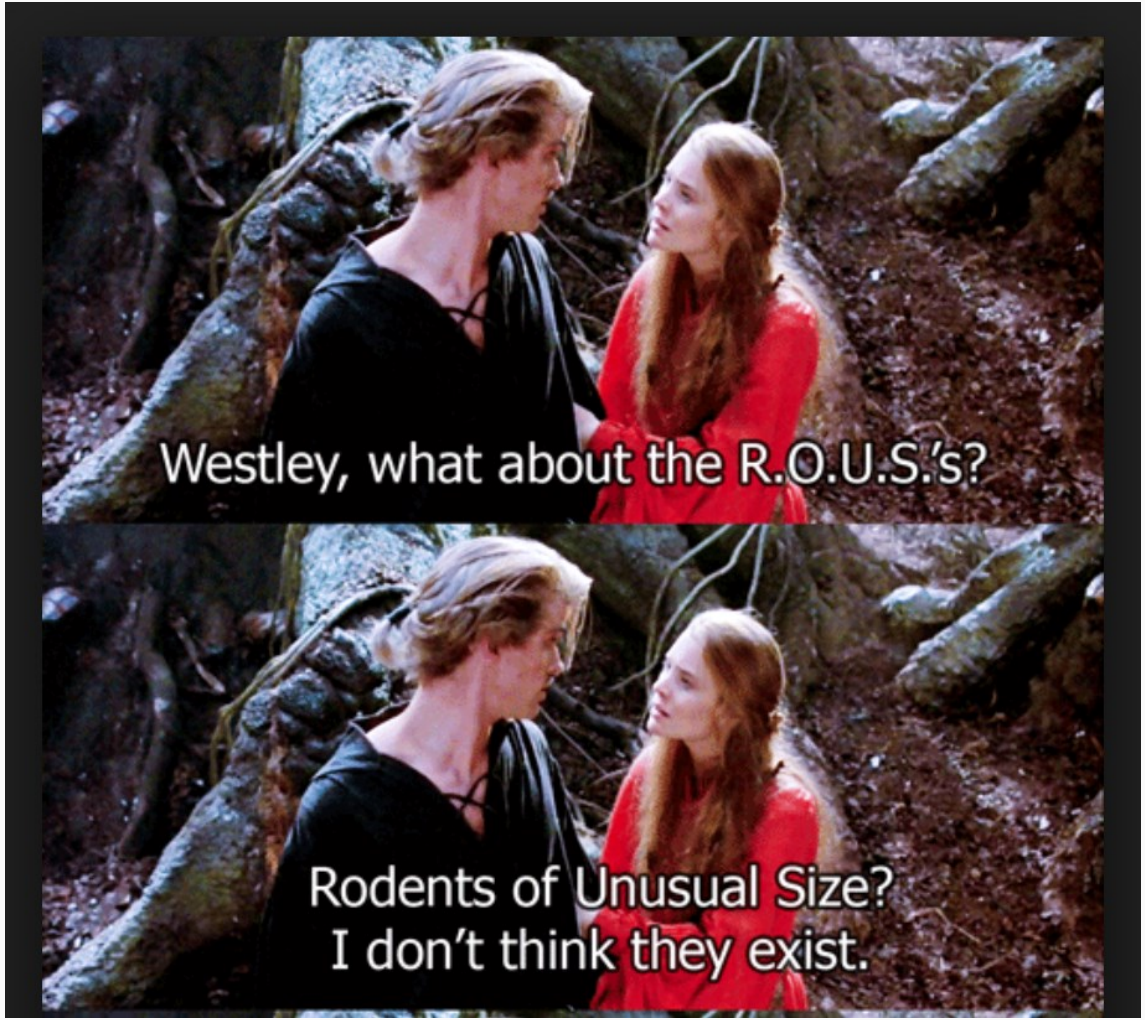
Data of Unusual Size: A Policy Framework Investigation

Laura Gerlitz and Leah Vanderjagt
University of Alberta Libraries
Netspeed Conference, 2017


Westley, what about the R.O.U.S.'s?
Rodents of Unusual Size?
I don't think they exist.

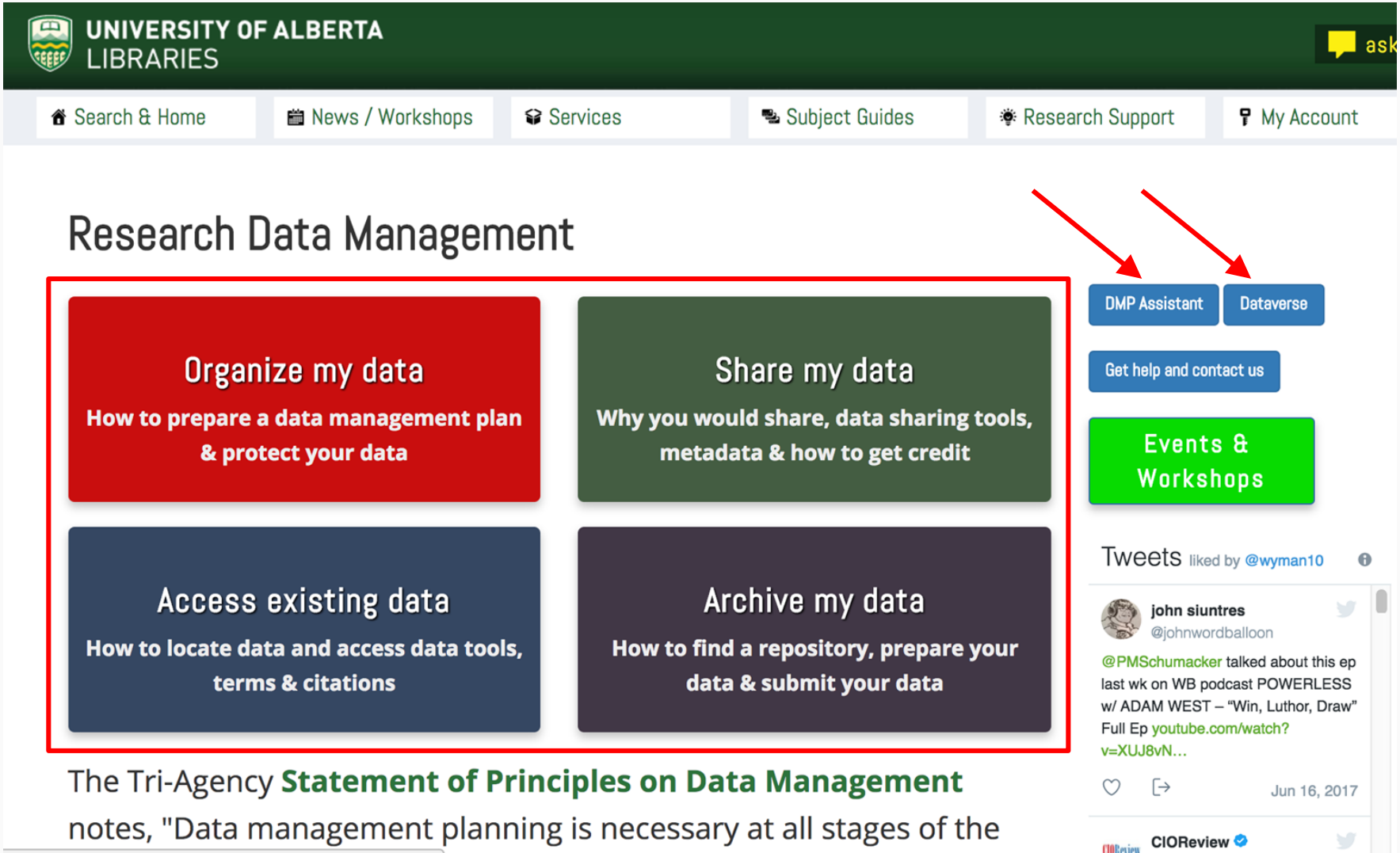
Background

UAL defined Research Data Management support as a key strategic priority for development and outreach

We created training, held many workshops, spoke to faculty, grad students and others responsible for data management

We promoted our work in this area within the university

We established a data repository (Dataverse)


UNIVERSITY OF ALBERTA
LIBRARIES
ask
Search & Home
News / Workshops
Services
Subject Guides
Research Support
My Account
Research Data Management
Organize my data
How to prepare a data management plan & protect your data
Share my data
Why you would share, data sharing tools, metadata & how to get credit
Access existing data
How to locate data and access data tools, terms & citations
Archive my data
How to find a repository, prepare your data & submit your data
DMP Assistant
Dataverse
Get help and contact us
Events & Workshops
Tweets liked by @wyman10
john siuntres
@johnwordballoon
@PMSchumacker talked about this ep last wk on WB podcast POWERLESS w/ ADAM WEST – "Win, Luthor, Draw" Full Ep youtube.com/watch?v=XUJ8vN…
Jun 16, 2017
CIOReview
The Tri-Agency Statement of Principles on Data Management
notes, "Data management planning is necessary at all stages of the

Resulting (good) problem

Researchers wrote data management plans into their grant applications

...And then completed their research

Named our services in the plan

And are now bringing us data for sharing and preservation!

Much of the data is **LARGE**. **REALLY REALLY LARGE.**

How large is large?

FOR US?

Capacity preservation storage: many-terabytes of preservation storage (access interfaces connecting to this through paced development, and we are collecting many use cases)

Application upload/download limits (2 GB)

FOR RESEARCHERS?

Large is a rather subjective concept, but much of it terabyte-scale

Scratching the surface...

We also discovered these challenges:

Who is the **data steward**?

Tackling **sensitive or protected** information

Finding a place for **data preparation**

Eligibility: filesize cap limits, fees, agreements, application processes?

De-accessioning processes?

The problem felt like this

We need for it to feel like this

Environmental review

To develop a service plan, we reviewed website communications and policy statements from a sample of mature/robust data archiving services.

Final report covered 10 representative institutions.

Sites we examined

- University of British Columbia EduCloud Server
- Ontario Library Research Cloud
- Purdue University Research Repository
- DuraCloud (Chronopolis)
- Indiana University Bloomington Science Data Management Storage and Preservation (SDA)
- University College London Research Data Storage Service
- Data Archiving and Network Services EASY and DataverseNL
- Data.uel (University of East London)
- Research Data Storage Facility at the University of Bristol
- Research Data Store at the University of Sydney

What we found

“There is a lot of **valuable information available online** for long-term large storage demand archiving services, and we should be able to **pull together a framework** for questions we need to ask ourselves about **service objectives**, **decisions**, and **management** based on the documentation elements we have identified here.”

Elements

Governance
Levels of access
Price models
Application processes
Storage capacity

Support services
Data organization and metadata requirements
Data retention policies
Partnerships

Governance?

Who is in charge of these services?

Some libraries make decisions about service in collaboration with other campus organizations.

1. Ethics and legal compliance
2. Managing personal information
3. Data decommissioning

E.g. At Purdue, the Dean of Libraries, VP-Research, and VP-IT/CIO are collectively responsible for policy development and preservation.

Levels of access?

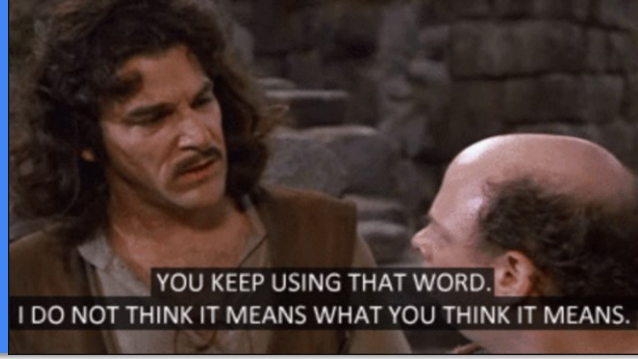
Can people access their large data? Under what circumstances?

No ‘dark archives’ without researcher access - **all services stated that users could continue to access their data**.

We did find **some restrictions on what users could do with the data** in the storage/archiving environment.

Primarily: **archive space is not your active project space** for generating new data.

Levels of access?



How is this space defined? What terminology is used?

We found that services make a distinction between:

a) **private** “project space” vs. **public** space;

b) “**small** data” vs. “**large** data” (with different thresholds for where the cutoff is), and

c) “**fixed**” data (data that will not be updated) vs. “**active**” data (data is still being created, organized and described).

Locations and personal roles for access?

Where does the data go and who's responsible for it?

Location prescribed by **storage thresholds** and **fixed vs. active** data status

Roles for access were defined, especially at Bristol: "**Data steward**" = Principal Investigator, full access to all data, "**Data User**" = authorized consumer, "**Data Controller**" = person who handles processing of personal info in data

Pricing models?

Are people charged for this service?

Most express '**basic account**', and call this **free**, but **don't define what 'basic' is**
- if you need more than X TB/GB, contact us for a quote

Some sites present **tiered account levels** according to different ranges of **storage** consumption and **fixed vs. active** status. Active storage is more expensive.

Fees are for storage consumption. Found no fees for any other aspect of service.

Application process?

Does anyone use an application process to triage requests?

No cases of this. **Institutional affiliation = eligibility.**

We were surprised that **no sites discussed** data organization, rights or metadata status **requirements for service**. One site wanted a research data management plan.

Another service asked for storage consumption estimate for future additions and a 500 word description of the data.

Storage capacity?

Was there a common total capacity allowed for one client, and common total capacity available for entire service?

We observe a **common threshold for users of 5 TB**, beyond which… “we need to talk.”

Only 2 sites expressed total current capacity -

> **Indiana**: 15PB of usable tape capacity, and 1800TB of total disk capacity (cache). **Bristol** has 2.3PB of storage that can be expanded as needed, and mirrored IBM DCS9900 systems populated with 2TB disks, providing a usable 900TB per site.

Terms of Use / Usage Policies?

Any legal requirements in common use to adopt?

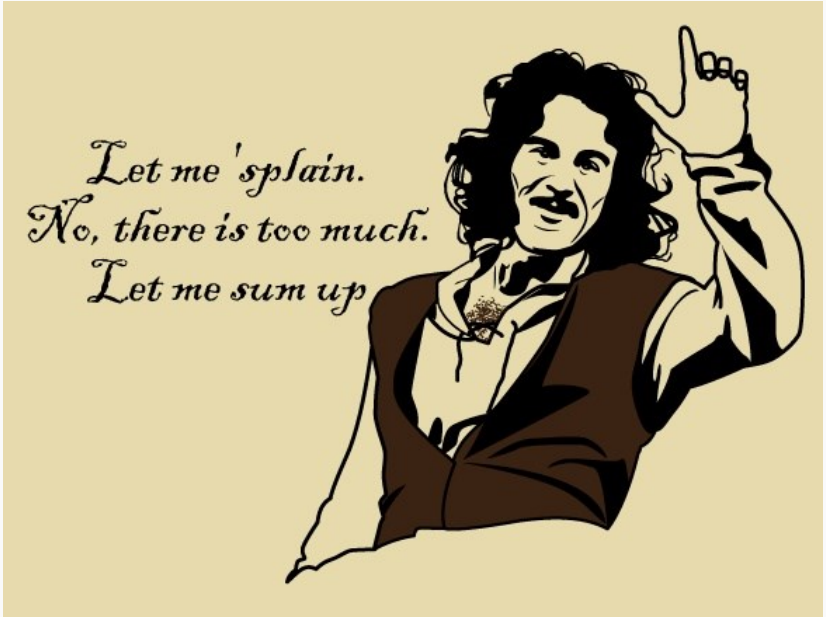
Two big themes:

Sensitive information: Most have been explicit within Terms of Use for the handling of sensitive information

User responsibilities: Standard misuse-of-service language, but with ‘adherence to ethical principles” added; detailed user and institution responsibilities available at Bristol

Technical Infrastructure

Can we borrow any frameworks or technology implementations?



Technical Infrastructure

Implementation structure varied widely, but the combination involved:

Most used commercial supplier for back end storage/preservation (large file storage). One used Swift/OpenStack (Open source).

Several offered **high-performance computing** space.

Different utilities employed to provide **front-end access** to large data in back end storage (e.g. NAS header, Heirarchical Storage Interface).

Support services

How are users supported through the process of submitting these files?

Information on this is a little thin.

Email/ticketing systems, libraries, units, and **individuals to contact**; **FAQs** - standard, but added:

Procedures for handling data: This included content guidelines, file format requirements or suggestions, instructions for preparing and depositing data, and procedures for handling sensitive data.

Data organization and metadata requirements?

How do sites guide the contributors in preparing data to become archive-ready?

Few give specifics for this!

Purdue, DANS, Indiana and Sydney offer suggestions and guidance for preservable **file formats**.

Some sites require **README files** that contain metadata about the data files, such as file descriptions, noting of any distinctions between the files, and whether the files contain sensitive information.

Data retention policies?

Do any sites have policies on period of retention or review?

Very few do this.

Where present, expressed as “**at the discretion of the library**”, or after **5-10 years**, after which time the **data is reviewed**.

Who reviews for retention? How?

Bristol: **Research Data Storage and Management Board** reviews annually.

Purdue: Beyond their **ten year retention period**, data is reviewed. May deselect b/c data

- does not relate to the **teaching and research mission** of uni
- **is archived elsewhere** in a trustworthy repository

University College London offers different retention policies **based upon the subject of the data** and detail their terms in a PDF link.

Partnerships?

What partnerships are in place for service delivery?

All reviewed partner with university IT services to deliver large storage demand archiving services.

OCUL Cloud is a service based on a network of **partnered universities** in Eastern Canada.

DANS (Netherlands) is partnered with a number of **European institutions**.

DuraCloud is partnered with other non-profits to offer extended services, such as automatic offsite backups for Archive-It partner organizations.

Conclusions for us

What can we conclude?

Policy models should drive infrastructure after use case review.

Partnerships are important to archiving large data.

Large data archiving infrastructure and policies are built around data in various stages of ‘readiness’ for archiving.

Eligibility for service probably should not be determined by data-readiness, rather data- “willingness”.

Difficulties of large data

Relating and rationalizing development of local services in light of **new national infrastructure** (Federated Research Data Repository, Portage)

Ensuring our **requirements for digital preservation** are met in our service design; who will migrate files, create metadata, structure files? Are we adequately **resourced to meet these demands**?

Designing **predictable and reliable service paths**, so that each request is not a 6, 8, 12-month project - and doing so on shifting sand, with more and more data.

Thanks! Questions?

Link to report:

http://bit.ly/2xYISD2