Sustainability and Finances

Strategic Workshop for Research Data Management (RDM)

University of Alberta; November 17, 2015

Financing and Sustainability – the burning questions

- To what ends?
- Why is RDM a priority?
- How much will it cost?
- How will we fund it?
- How will we sustain it?

RDM infrastructure– a snapshot of today

Component	Who leads, funds & sustains (Capital/development and operations)
"Physical" infrastructure – compute, network, storage (short term and archival)	CFI, CC, CANARIE, institutions (e.g. TSpace), grants
Curation infrastructure – e.g. preservation systems; metadata standards	Institutions (individually and collectively), in particular libraries (e.g. OCUL), DataCite; CASRAI
RDM infra underlay ; services – e.g. for ingest, discovery, visualization, training	CFI cyber pilot; institutions; CARL – Portage project; CASRAI
Managing data as infrastructure – often for domain specific utilization	Diverse – NRC (astronomy and particle physics), Universities and the GoC (Canadian Polar Data Network CPDN), international orgs
System connections	RDC (with CANARIE support);

"System" Assessment

The good news

- Many pieces of the RDM ecosystem exist
- The engagement of multiple players

The bad news

- The patchwork quilt of players is without an overarching vision, policy framework or effective coordination. No acceptance of roles and responsibilities
- There is little attention paid to the deeper level of infrastructures required for identification, storage, metadata and relationships that enable research and scholarship.
- There is little recognition of the real locus of costs – data curation and RDM services (the human dimension); existing digital infrastructure programs are capital not human intensive
- Funding agencies are avoiding the question of who pays for what
- Few institutions are deeply engaged; yet they have RDM responsibilities
- Most researchers do not appreciate the benefits from good RDM, nor do they have the requisite skills
- OVERALL – an as yet fragile foundation for sustainability of RDM

An effective RDM ecosystem

- Good policy framework, governance and incentives
- Distributed stewardship, management & funding
- A focus on getting the deeper layers of infrastructure right (stuff that is invisible when it works and stuff that is not 1:1 aligned with project funding – it is underpinning)
- Recognizes the human capital intensity of the RDM infrastructure – pre-ingest, ingest, archival and access
- Seeks scale economies through cooperation, collaboration and coordination of activities

The challenges

- Motivation and culture (incl disciplinary)
- Technical – having the infrastructures, services, processes and training in place
- Program rigidities, both “capital” and operating
- Costs and cost uncertainties
- Legal and ethical provisions, e.g. IP, confidentiality
- Interoperability

What do we know about costs (1)? The UK Archeological Service

"Looking at the distribution of staff costs over five major cost categories… (pre-archive, acquisition, ingest, archive, and access), the largest proportion is accounted for by the access category (31%). However, the activities leading up to and including ingest of the materials into the archive collectively account for 55% of total staff costs. … the process of actually preserving the materials (archive category) accounts for only 15% of total staff costs."

Beagrie et al 2010

What do we know about costs (2)?

- The UK report "Science as an Open Enterprise" – sample costing of operating data initiatives

Data initiative	Annual cost	Staff levels
Tier 1 – major international data initiatives with well-defined protocols for the selection and incorporation of new data and ensuring access Tier 2 – data centres and resources managed by national bodies or prominent research funders		
Worldwide Protein Data Bank (wwPDB)	$11-12M of which $6-7M is for data deposition and curation	69 staff
UK Data Archive	£3.43M	64.5 staff
arXiv.org	$810,000	6 staff
Dryad	$300,000	4-6 staff
Tier 3 – curation at the level of individual universities and research institutes, or groupings of them		
ePrints Soton at U of Southampton	£116, 318	3.5 staff
D-Space at MIT	$260,000	1.25 + 1.5 FTE
Oxford University Research Archive and DataBank	Under development; costs not available	2.5 FTE +?

What do we know about costs (3)?

- UK – Jisc suggests that up to 5% of the project costs will be for RDM where there is i) high re-use potential and ii) data complexity
- Another UK estimate: that curation is 1.4% -1.5% of the total research expenditure of the research councils (definition of what is included in curation is unclear)
- There are <u>also</u> real costs in setting up the necessary layers of infrastructure (the capital expenditure) for effective use and re-use of existing data
  - Example – LINCS (Linked Infrastructure for Networked Cultural Scholarship) – that has the potential to transform humanities research
    - $5M capital project to create an innovative platform
  - Example – CASRAI semantic standards for administrative research data and RDA for research data – invisible but key parts of the ecosystem

What do we know about ROI?

Macro-economic studies

Date	Study	Scope	Benefit of open data (% GDP)
2011	EU Commission	Europe (public sector data only)	1.5
2013	Shakespeare	UK (public sector data only)	0.4
2013	McKinsey	Global	4.1
2014	Lateral Economics	G8 countries	1.1

Moving forward (1)

The context

- Don't underestimate the importance of an enabling policy framework
- Learn from our experiences with diverse models of genesis, funding, delivery and governance of infrastructures
  - Federally mandated (e.g. CANARIE)
  - Community driven & governed; federal contributions (e.g. Compute Canada, CRDCN)
  - Consortia – regional, national, international (e.g. OCUL, Portage, CASRAI)

(and each model has its strengths and weaknesses)

Moving forward (2)

Take a page out of the UK Concordat

- "…consideration of cost forms an important part of any obligation arising from the move to open research data. *Such costs should be proportionate to real benefits*."
- "*The costs should not fall disproportionally on any part of the research community*. Rather, all parties should work together to identify the appropriate resource provider whilst recognising the obligation to reduce costs through sensible design of both obligations and infrastructure."

Moving forward (3)

Some directions to consider

- Redeployment/efficiency – Reassess how digital research/research infrastructure resources are deployed (institutional, regional and national levels)
- Incremental investment – RDM has a real cost (with commensurate ROI)
- Consider:
  - A "top-slice allotment" in which the enabling infrastructure funding is not tied to project costs
  - Innovation in redesign of existing funding mechanisms

Back to the burning questions

- How much will it cost?
  - Limited evidence; some from the UK
- How will we fund it?
  - Think global, act local and regional
  - Proportional-cost funding models
- How will we sustain it?
  - Importance of the local institution
  - Importance of regional organizations
  - Importance of national funding – for innovation, incentive and sustaining (regional and national levels)

So – to move towards sustainability

- A national policy framework
- A consensus on what infrastructures are required for RDM
  - National
  - Regional
  - Local
- Articulation of roles and responsibilities in stewardship, managing and funding RDM at all levels
- Reform of how we fund RDM – at national and at local levels

Over to you

Janet E. Halliwell
jehalli@telus.net