Bridging the Gap Between Small and Large Research Repositories

(There is No Dumb Data!)

Anita de Waard
VP Research Data Collaborations
a.dewaard@elsevier.com

http://researchdata.elsevier.com/

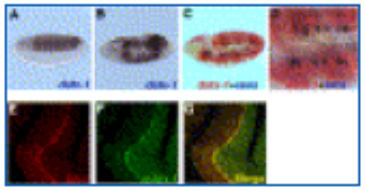
Background:

- Background:
  - Low-temperature physics (Leiden & Moscow)
  - Joined Elsevier in 1988 as publisher in solid state physics
  - 1991: ArXiV => publishers will go out of business *very soon*!
- 1997- 2013: Disruptive Technologies Director, focus on better representation of scientific knowledge:
  - Identifying key knowledge elements in articles (linguistics thesis)
  - Building claim-evidence networks (collaborations on e.g. CKUs!)
  - Help build communities to accelerate rate of change (Force11)
- Per 1/1/2013 Research Data Collaborations:
  - Data is the evidence that the claims are built on!
  - Doug Engelbart: connected minds augment collective intelligence
  - Can a publisher play a useful role?

Claimed Knowledge Update Usually Refers to Data (*or lack thereof!*):

Sens and Gfi-1 Are Coexpressed with Atx-1 Homologs in *Drosophila* and Mice

The findings that fly Atx-1 and Sens as well as mammalian Atx-1 and Gfi-1 physically interact prompted us to examine if Atx-1 and Sens/Gfi-1 are coexpressed in vivo. In situ hybridization and Northern analyses show that *datx-1* is expressed in embryonic stages (Figures 3A–3D and data not shown). The expression of *datx-1* is first observed in the dorsolateral region in the stage 5 embryos (Figure 3A). During gastrulation, *datx-1* is expressed in the dorsolateral ectoderm that encompasses the peripheral neuroectoderm (Figure 3B). *sens* mRNA is first expressed in presumptive sensory organ precursor (SOP) cells at stage 10 (Nolo et al., 2000). We found that *sens* is expressed in a subset of cells within the region of *datx-1* expression (Figures 3C and 3D). In mice, Gfi-1 is expressed in many areas that give rise to neuronal cells during embryonic development (Wallis et al., 2003). However, our data show that, in the adult cerebellum, Gfi-1 expression is mainly confined to PCs, where Atx-1 is most abundant (Figures 3E–3G) (Banfi et al., 1996).



Full-size image (90K)
High-quality image (1040K)

Figure 3. Fly and Mouse Atx-1 Colocalize with Sens and Gfi-1 in Certain Cell Types

There are many data preservation efforts:

- There are many different research databases– both generic (Dryad, Dataverse, DataBank, Zenodo, etc) and specific (NIF, IEDA, PDB)
- There are many systems for creating/sharing workflows (Taverna, MyExperiment, Vistrails, Workflow4Ever,)
- There are many e-lab notebooks (LabGuru, LabArchives, LaBlog etc)
- There are scores of projects, committees, standards, bodies, grants, initiatives, conferences for discussing and connecting all of this (KEfED, Pegasus, PROV, RDA, Science Gateways, Codata, BRDI, Earthcube, etc. etc)
- *You can make a living out of this ;-)! (and many of us do…)*

…but this is what scientists do:

Using antibodies

and squishy bits

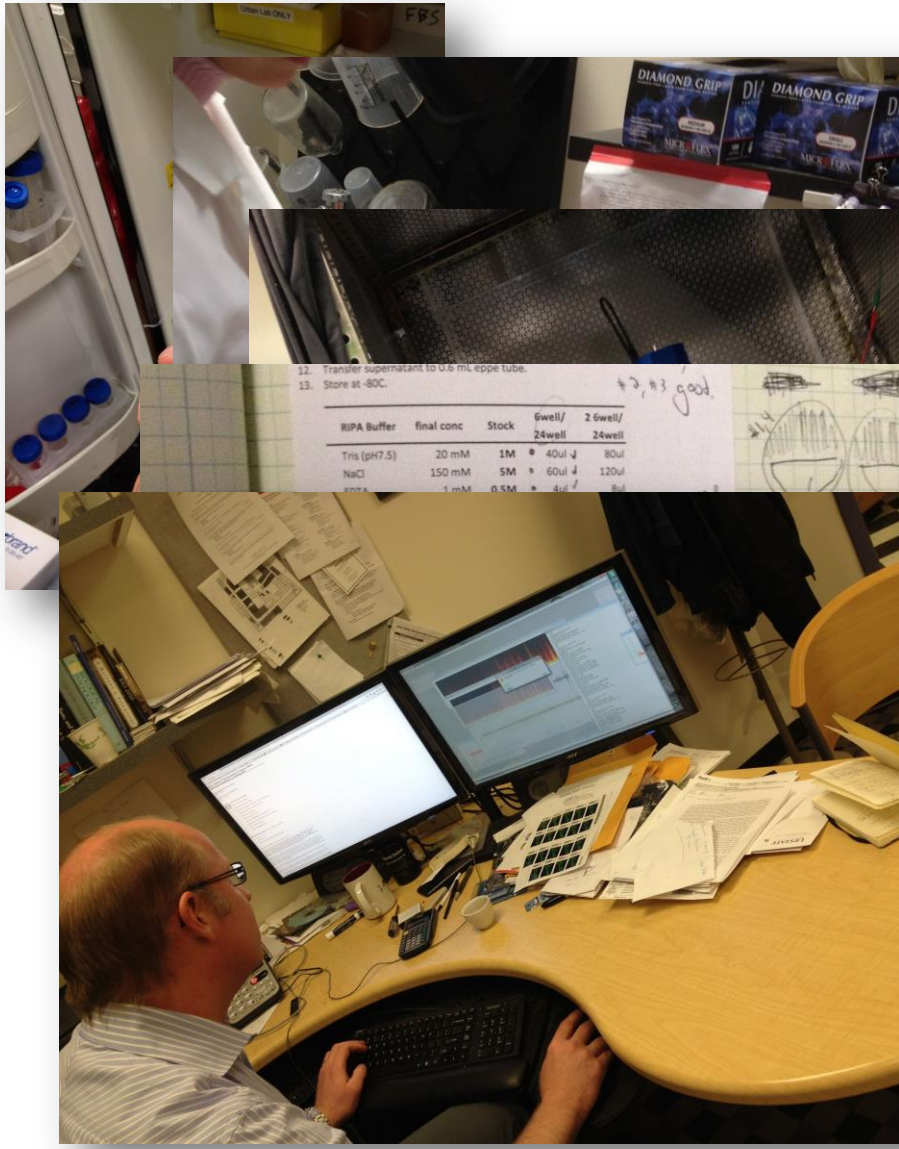
Grad Students experiment

and enter details into their lab notebook.

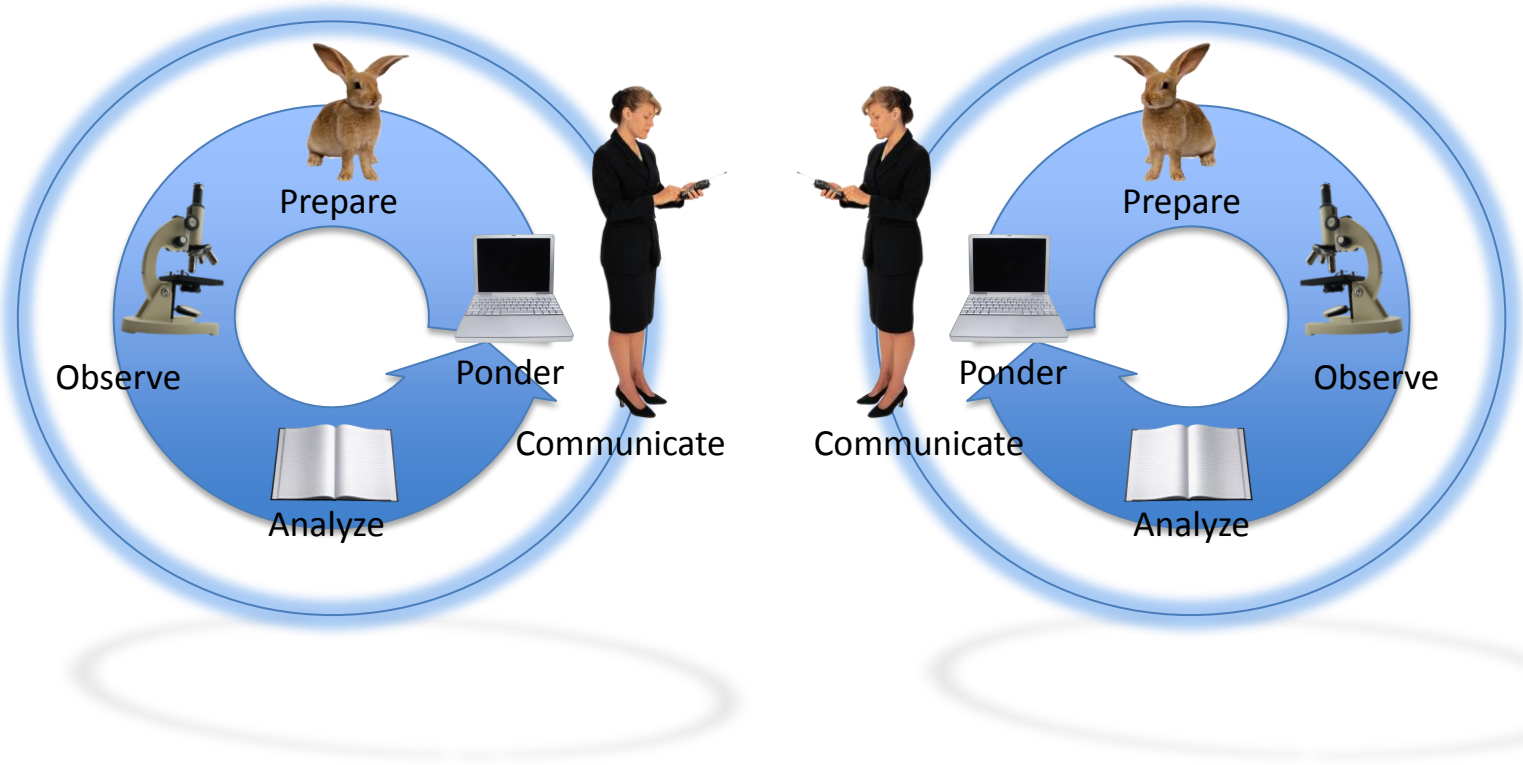
The PI then tries to make sense of this,

and writes a paper.

End of story.



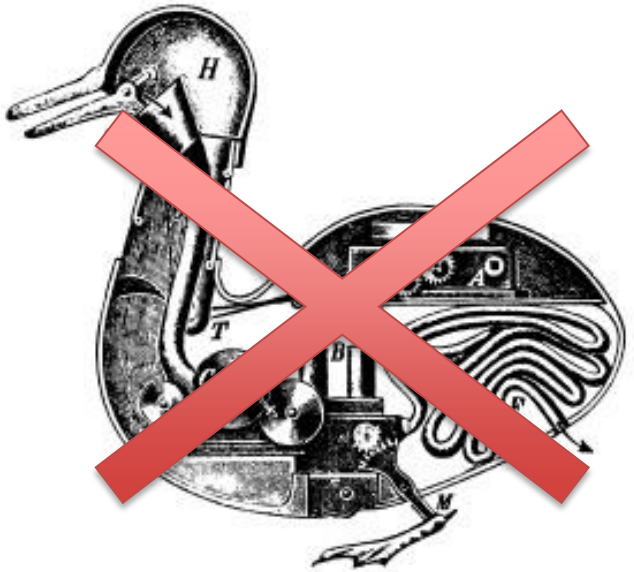
As a result of this practice, e.g. most of biology is quite insular



But also VERY complicated:

- Interspecies variability: *A specimen is not a species*
- Gene expression variability: *Knowing genes is not knowing how they are expressed*
- Microbiome: *An animal is an ecosystem*
- Systems biology: *A whole is more than the sum of its parts*

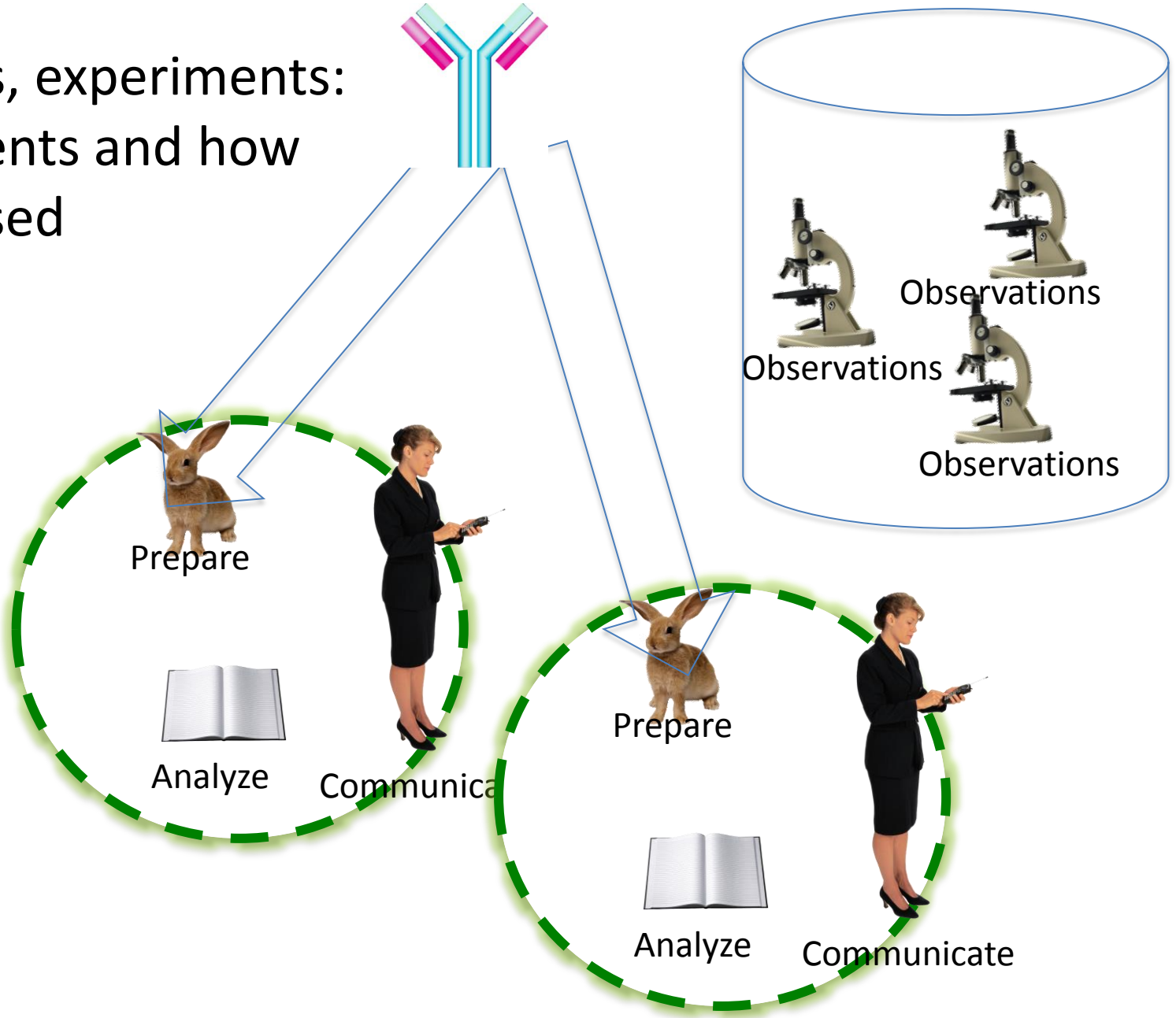
Reductionist science does not work for living systems!



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

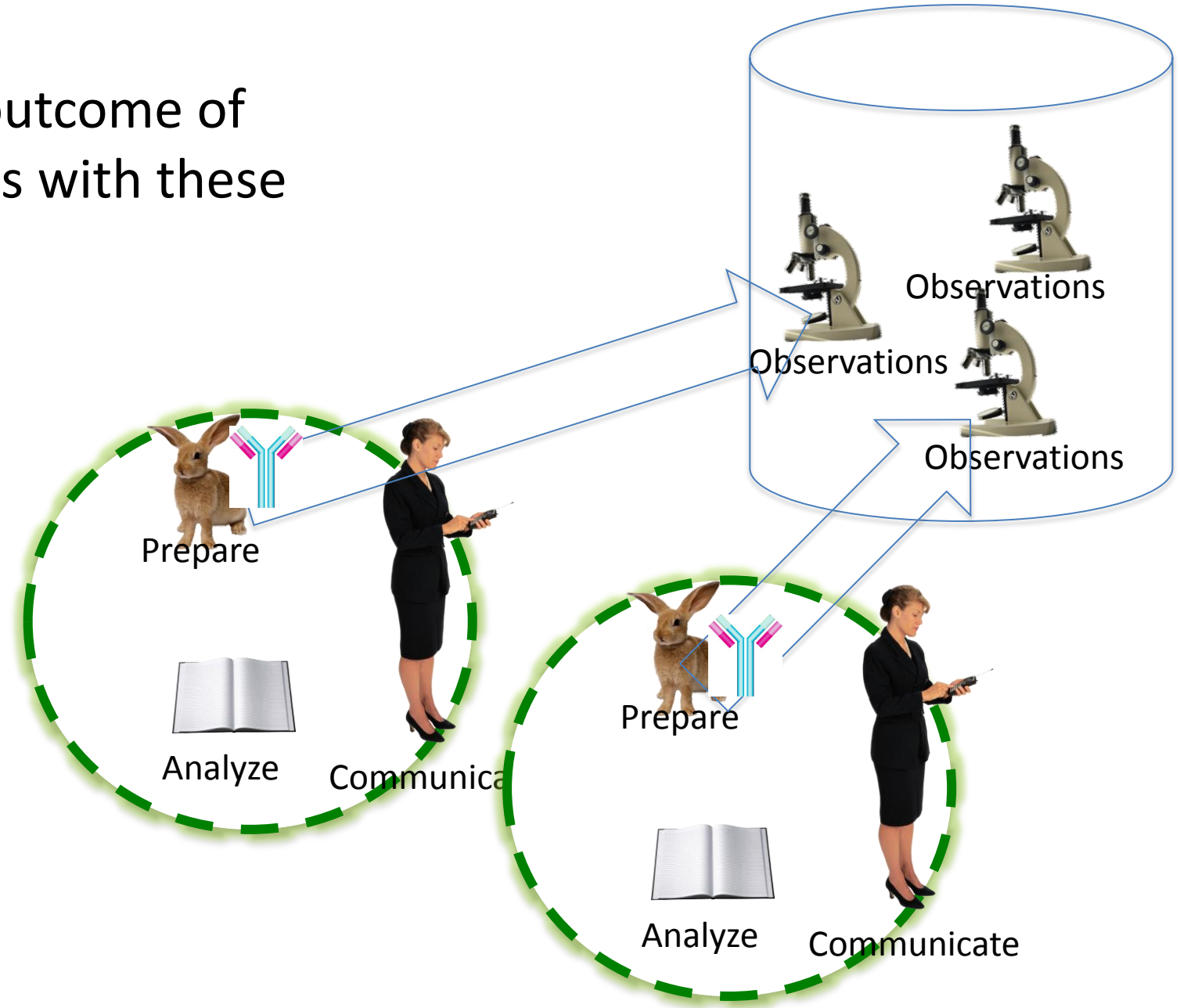
What if the research data was connected?

Across labs, experiments: track reagents and how they are used



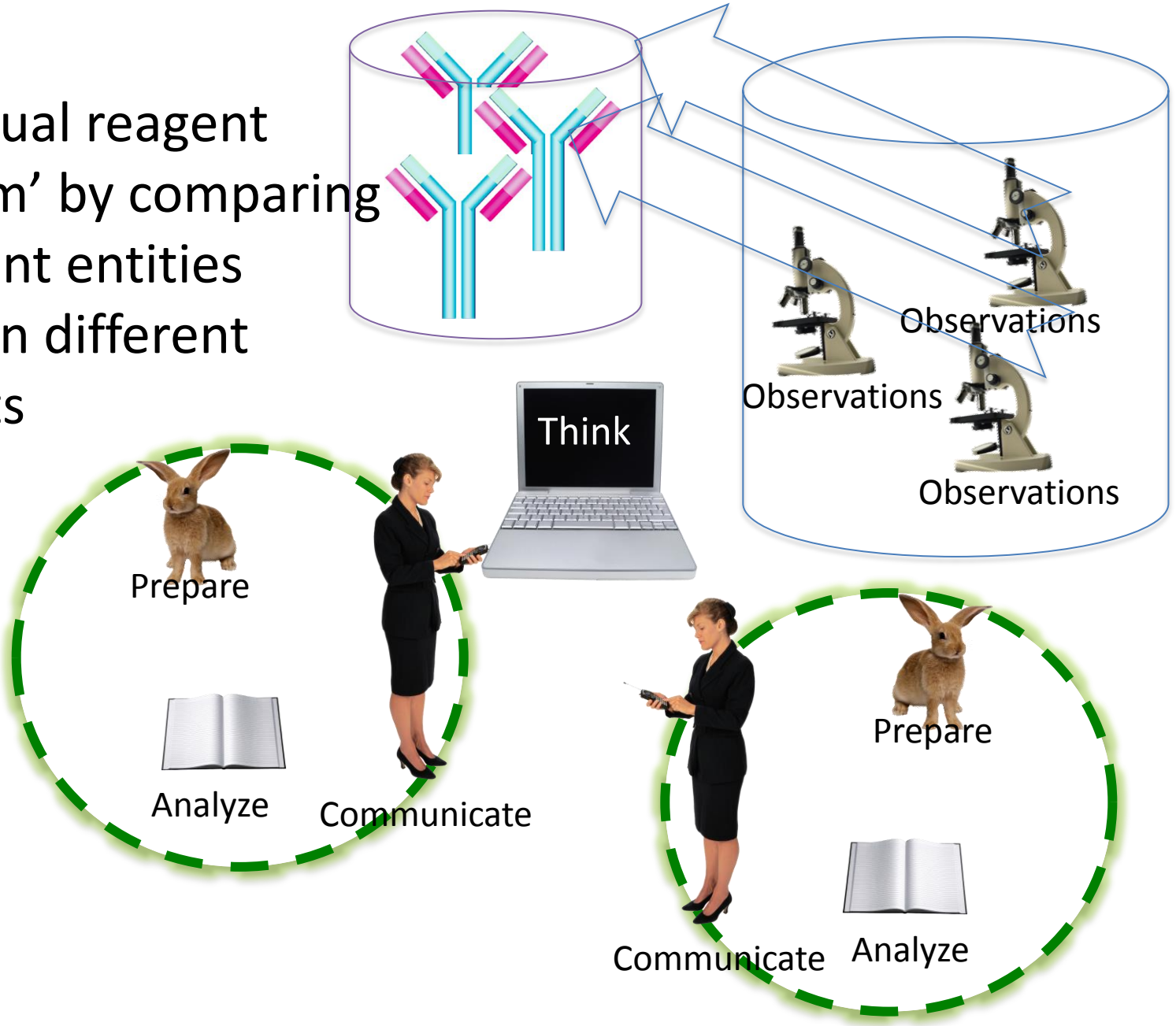
What if the research data was connected?

Compare outcome of interactions with these entities

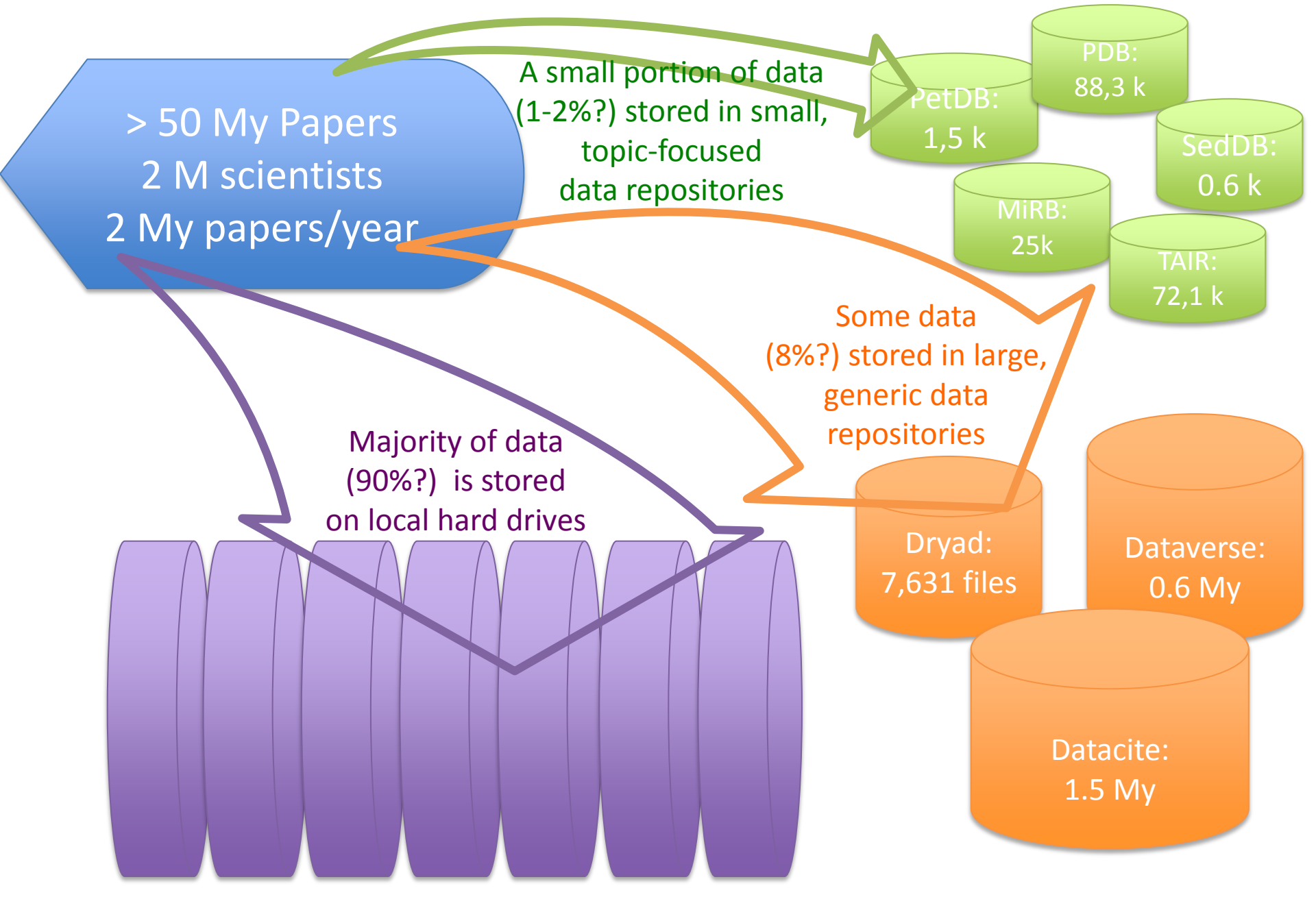


What if the research data was connected?

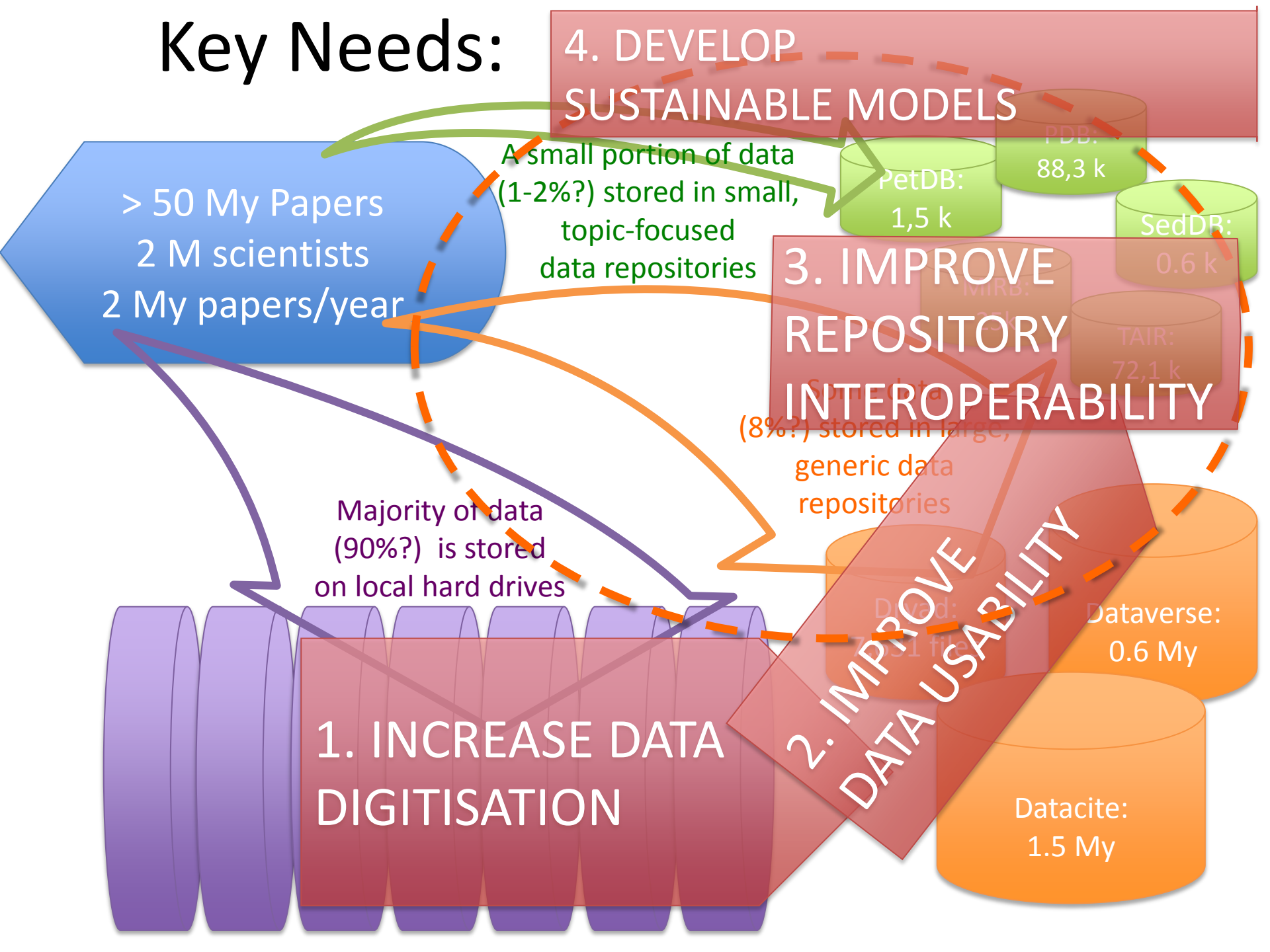
Build a ‘virtual reagent spectrogram’ by comparing how different entities interacted in different experiments



Where The Data Goes Now:



Key Needs:


4. DEVELOP SUSTAINABLE MODELS
> 50 My Papers
2 M scientists
2 My papers/year
A small portion of data (1-2%?) stored in small, topic-focused data repositories
PetDB: 1,5 k
PDB: 88,3 k
SedDB: 0.6 k
3. IMPROVE REPOSITORY INTEROPERABILITY
TAIR: 72,1 k
(8%?) stored in large, generic data repositories
Majority of data (90%?) is stored on local hard drives
Dryad:
Dataverse: 0.6 My
2. IMPROVE DATA USABILITY
1. INCREASE DATA DIGITISATION
Datacite: 1.5 My

Elsevier Research Data Services: Goals

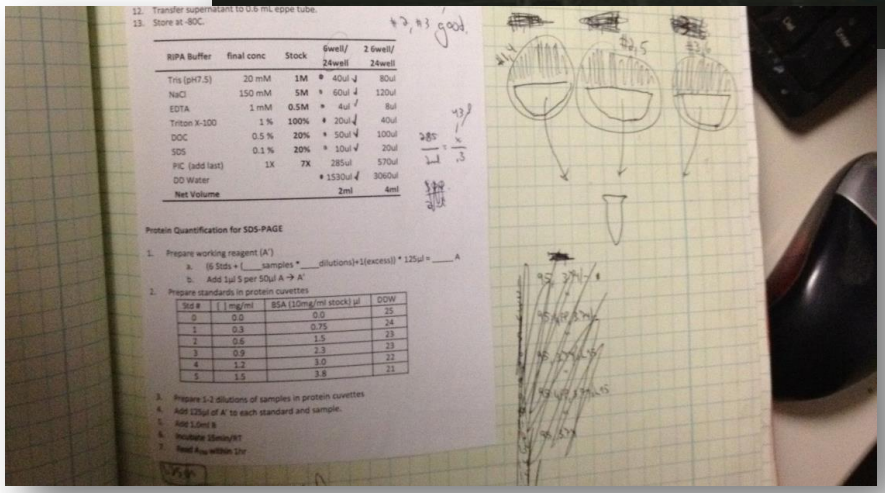
1. Increase Data Preservation:
   Help increase the amount and quality of data preserved and shared
2. Improve Data Use
   Help increase the value and usability of the data shared by increasing annotation, normalization, provenance
3. Enhance Interoperability:
   Help improve interoperability between systems and data
4. Develop Sustainable Models and Systems:
   Help measure and deliver credit for shared data, the researchers, the institute, and the funding body, enabling more sustainable platforms.

Elsevier RDS: Guiding Principles

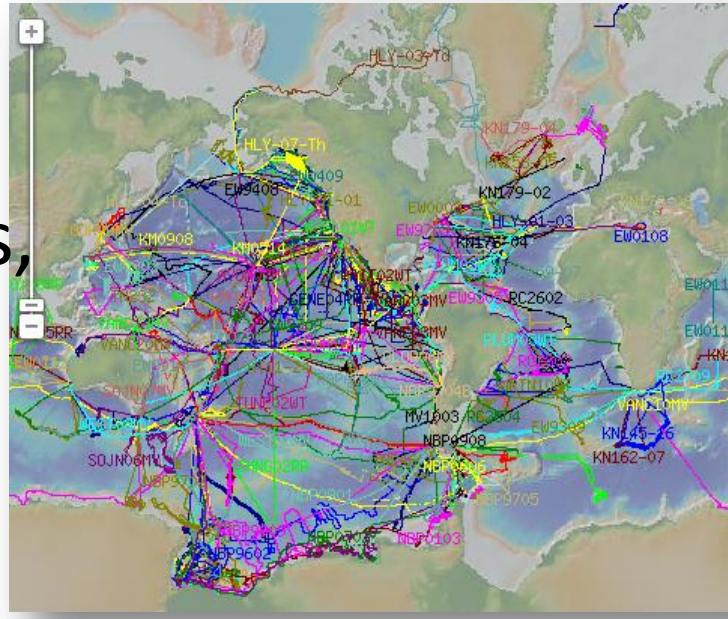
- In principle, all data stays open
- Work with existing repositories – URLs, front end etc stay where they are
- Collaboration is tailored to partner's unique needs:
  - Aspects where collaboration is needed are discussed
  - A collaboration plan is drawn up using a Service-Level Agreement: agree on time, conditions, etc.
  - Working with domain-specific and institutional repositories
- 2013: series of pilots to enable feasibility study:
  - What are key needs?
  - Can Elsevier play a role: skillsets, partnerships?
  - Is there a (transparant) business model for this?

1. Data Digitisation

- Goal: enable access, reproduction
- Issue: much of the research data is simply not digitized!
- Example: Magellan Observatory's paper records
- Example: CMU Electrophysiology Lab: lab notebooks are kept on paper



1. Data Digitisation

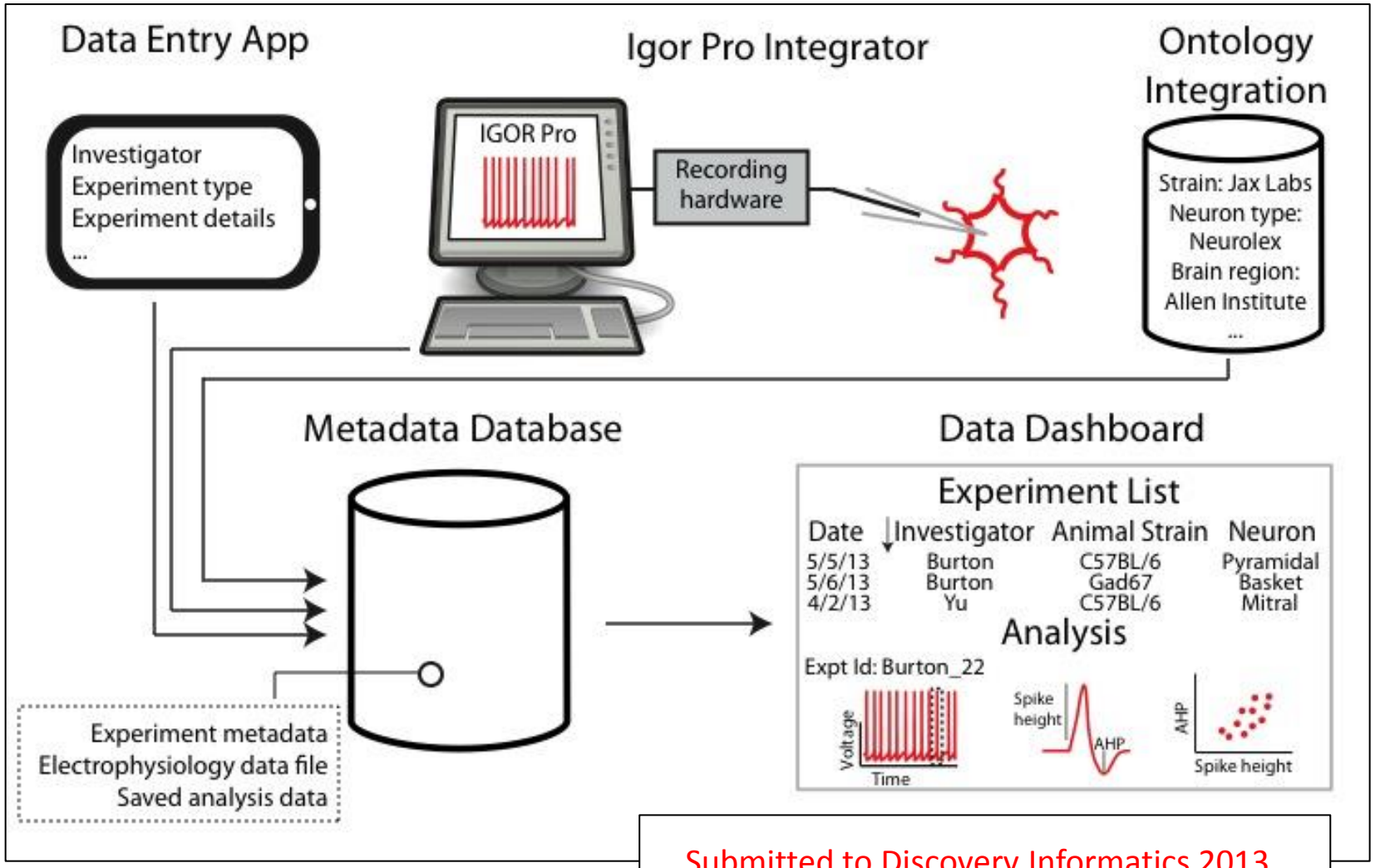


- Example: Marine geophysics suggests: convince instruments, not researchers! http://www.marine-geo.org/
- Prize: IEDA/Elsevier Data Rescue Award in the geosciences:

  $ 5000 award for best data rescue attempt

The 2013 International Data Rescue Award in the Geosciences
Organised by IEDA and Elsevier Research Data Services

1. Pilot: CMU Urban Legend App



Submitted to Discovery Informatics 2013

Creating an Urban Legend:

A System for Electrophysiology Data Management and Exploration

Anita de Waard[1*], Shawn D. Burton[2,3], Richard C. Gerkin[2,3],
Mark Harviston[1], David Marques[1], Shreejoy J. Tripathy[3,4], Nathaniel N. Urban[2,3]

[1]Research Data Services, Elsevier, US, [2]Department of Biological Sciences, [3]Center for the Neural Basis of Cognition, and [4]Program in Neural Computation, Carnegie Mellon University, Pittsburgh, PA; *Correspondence: a.dewaard@elsevier.com

2. Data Curation

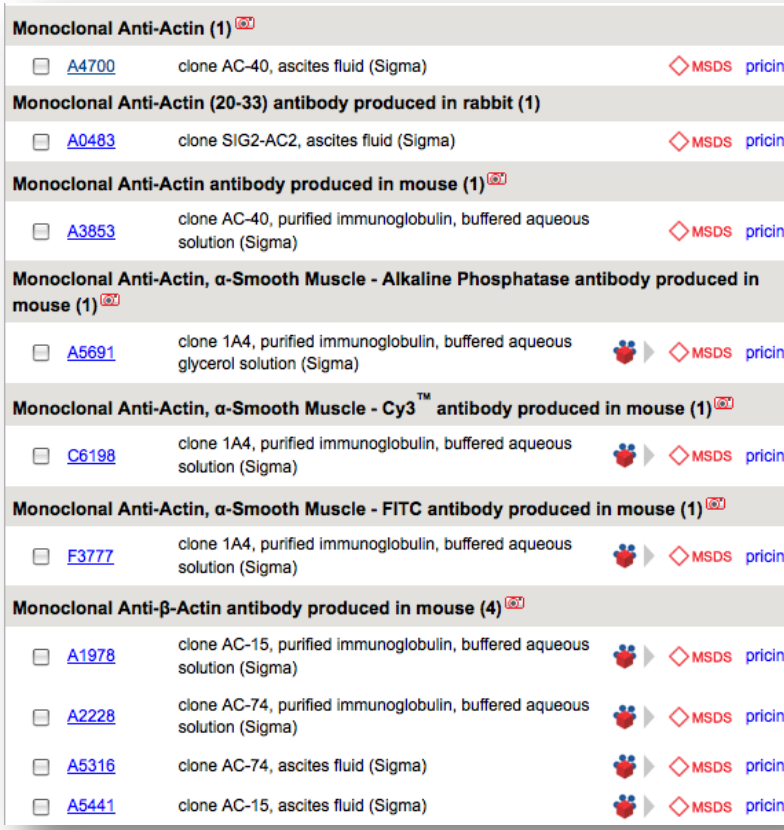
- To allow reuse, data needs to be enriched: why and how was it created?
- Issue: Dropbox and Figshare most popular tools
- Example: moon rock data is stored as PDFs with tables from different papers
- Pilot: lunar samples: curate geochemistry to allow data use

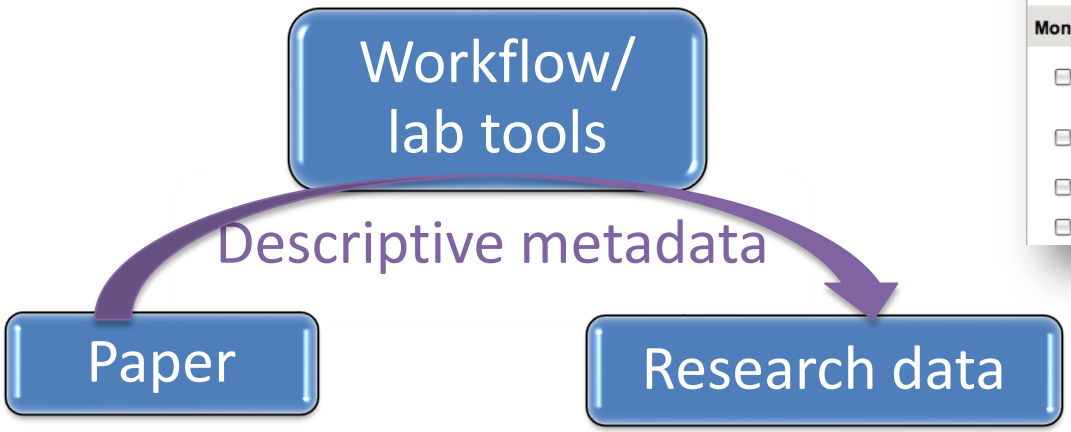
Plagioclase in 15415
electron probe (wt.%)

	Hargraves 72		Hanson 79	McGee 93	Stewart 72	Dixon 75	Papike 97
SiO2	44.19	43.92			43.36	44.8	43.2
Al2O3	35.77	36.24			36.04	34.5	37
FeO	0.16	0.09	0.102	0.085	0.08	0.08	0.086
MgO			0.05	0.071	0.07	0	0.042
CaO	19.66	19.49			19.34	20.1	19.5
Na2O	0.22	0.26			0.32	0.35	0.375
K2O			0.023		0.05	0.02	0.01
Ab			3.5		2.9		3.34
An	97	97		96.9	96.5		96.6
Or					0.3		0.059

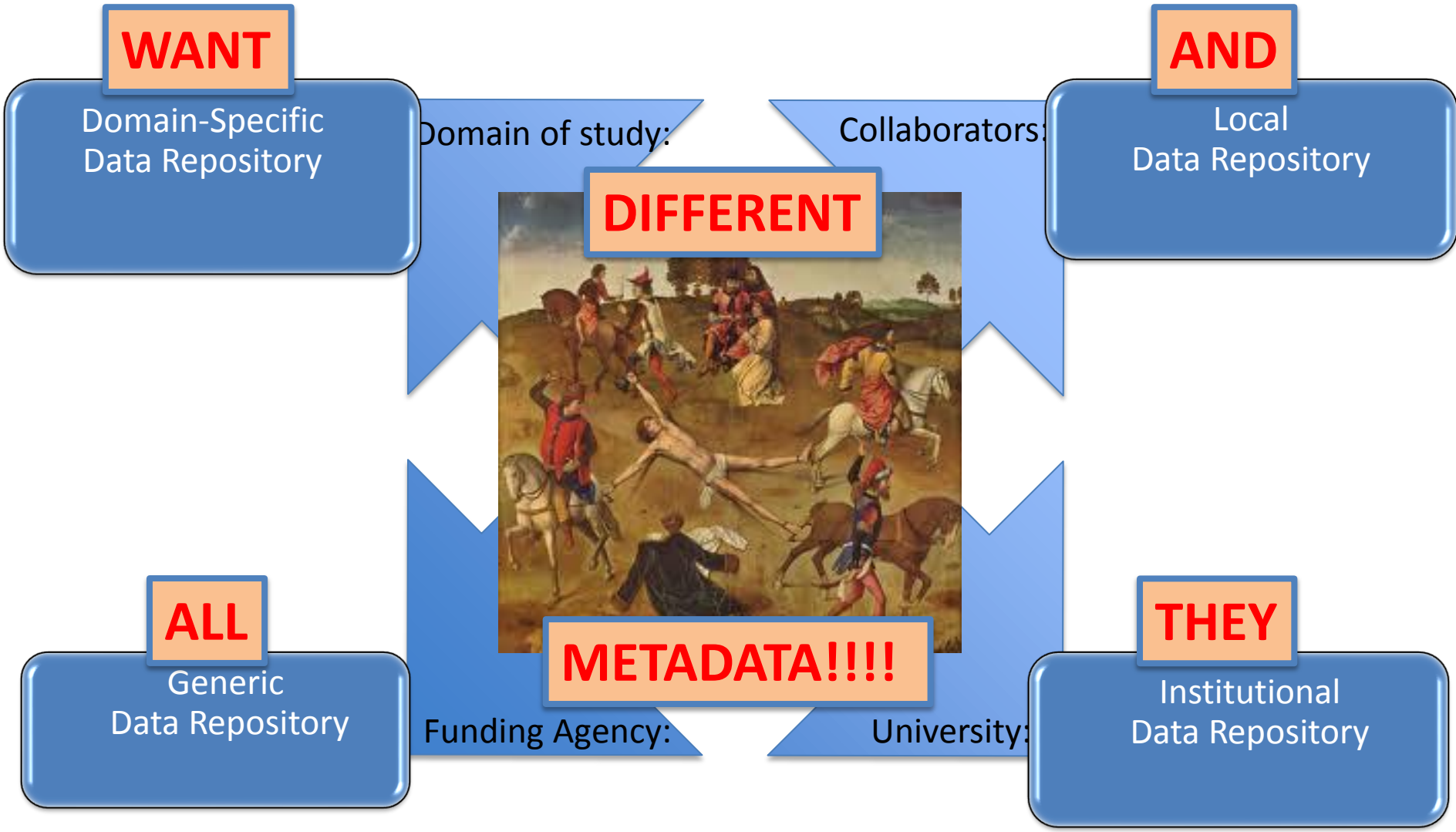
2. Data Curation

- Issue: hard to find right antibodies in papers (NIF)
- Pilot: Properly Annotated Data Sets (PADS) for biology: shared 'cloud' of metadata, describes why, what, how of experimental procedure:



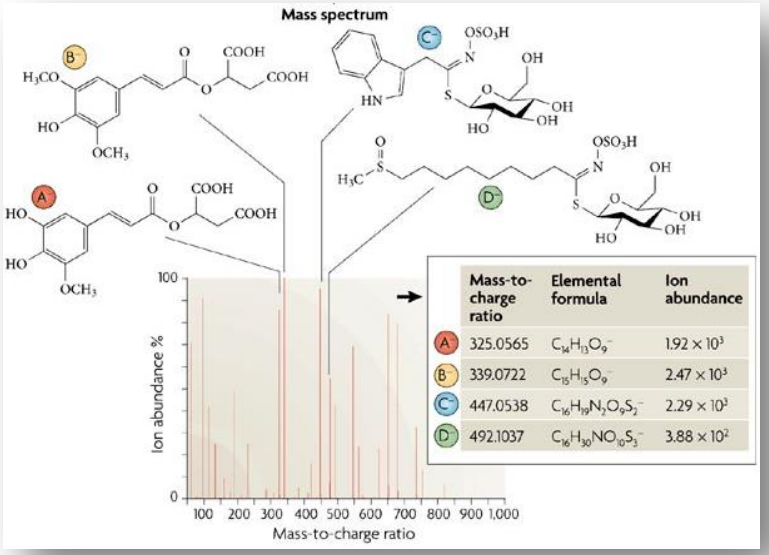


2. Data curation as seen by the researcher:



3. Repository Interoperability

- Pilot: find metabolomic compounds from mass spectrometry data: need biological understanding of chemical results



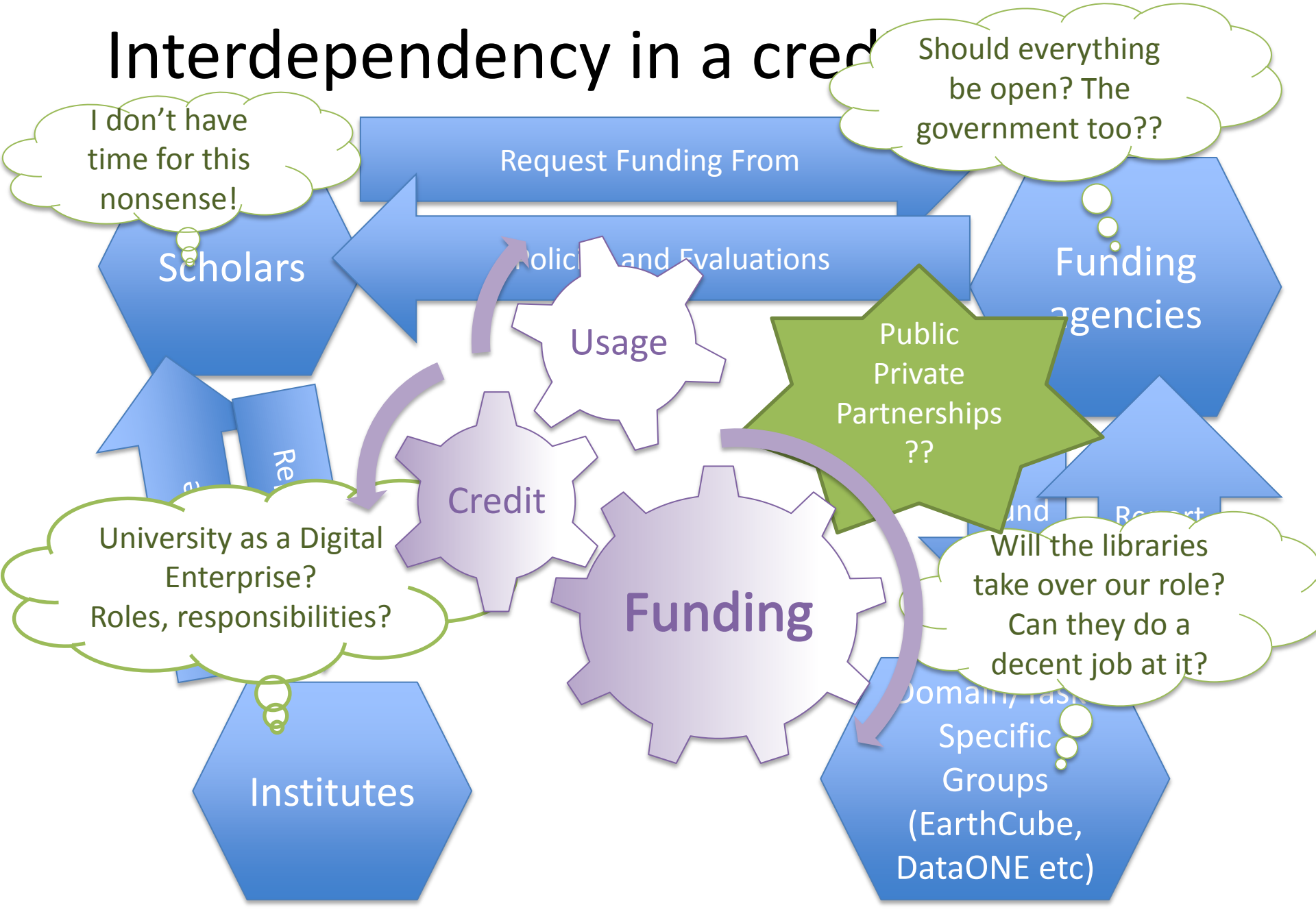
- Issue: battle between domain-specific and 'domain-agnostic' repositories: who is a better data curator? (Example: geochemistry)

3. Repository Interoperability

- Counterexample: MERRITT system:
  CDL builds generic infrastructure – domain-specific content curators

- Planned report with UCL, UCSD, MIT Libraries and CNI: best practices re. role of libraries?
  - How does a researcher decide where to place his/her content?
  - How does a library decide what digital data curation/preservation efforts to invest in?

4. Sustainable Data Models: Interdependency in a cred



4. Sustainable Data Models: (One Fool Can Ask) Many Questions!

- Cost:
  - Who pays for hosting the data?
  - Who pays for data curation?
  - Who pays for long-term preservation?
  - Who pays for data integration?
- Infrastructure:
  - Where does the metadata live?
  - What is the entry point to metadata cloud – the paper, the data?
  - Who is responsible for fulfilling DMP requirements?
  - Who decides on the data storage requirements?
- Usage:
  - Who wants to know where/what data is stored?
  - Who needs to know how data was accessed/used?
  - Who gets credit for data storage, data use?
  - Who needs/pays for credit-metric reporting?

In Summary:

1. Data digitisation:
   - Multiplicity of content, how to reach ‘small data’ creators?
   - Working with equipment could be key to success?
   - Pilots with CMU, Lunar Samples, Data Rescue Award.
2. Data curation:
   - Essential for reuse, but who does the work?
   - Each use case/user has own metadata requirements
   - Pilots with Metabolomics MS, Properly Annotated Data Sets.
3. Repository integration:
   - Domain-specific vs. domain agnostic?
   - Various domains have different requirements
   - Study and report re. best-practices for libraries.
4. Sustainable models in a credit economy:
   - Cost, infrastructure, usage: who needs what?
   - Who pays for what?
   - Interviews/discussions with number of institutions.

Thank you!

Collaborations and discussions gratefully acknowledged:

- CMU: Nathan Urban, Shreejoy Tripathy, Shawn Burton, Ed Hovy
- UCSD: Phil Bourne, Brian Shoettlander, David Minor, Declan Fleming, Ilya Zaslavsky
- NIF: Maryann Martone, Anita Bandrowski
- MSU: Brian Bothner
- OHSU: Melissa Haendel, Nicole Vasilevsky
- California Digital Library: Carly Strasser, John Kunze, Stephen Abrams
- Columbia/IEDA: Kerstin Lehnert, Leslie Hsu
- CNI: Clifford Lynch
- Harvard: Michael Kurtz, Chris Erdmann
- MIT: Micah Altman
- UVM: Mara Saurle

Elsevier http://researchdata.elsevier.com/
Research Data Services