Research Data @ Elsevier

From generation through sharing and publishing to discovery

IJsbrand Jan Aalbersberg
SVP Journal and Data Solutions
NDS, Boulder - June 12, 2014

Contributors:
Anita de Waard
Hylke Koers

Outline

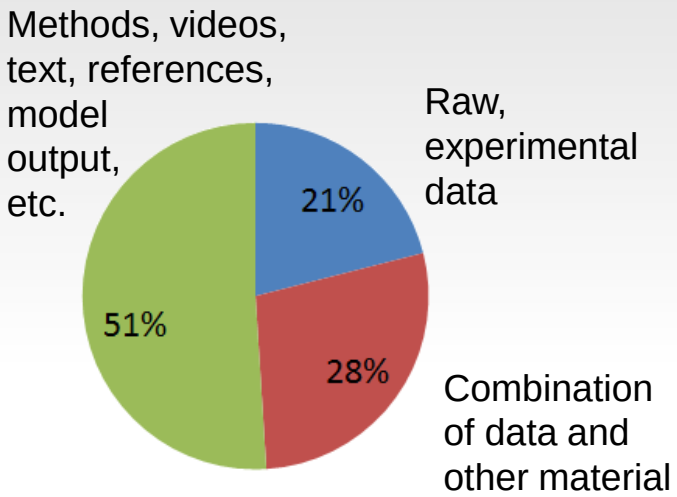
- Research Data – Current Status @ Elsevier
- Researcher Data Workflow
- Research Data Needs
- Experiments

Elsevier data activities started early on, with supplementary files and outward entity linking

- Accepting supplemental files for over a decade:
  - Growth of articles with such files of 25-35% per year
  - Recent snapshot showed 50% of such files contain data
- Suggesting in GfA to post data at repositories:
  - When appropriate data repositories were available
  - When supported by community and editorial boards
- Signing Brussels Declaration on data in 2007:
  - Raw research data should be made freely available. Publishers encourage public posting of such data. Data sets submitted with paper should (wherever possible) be made freely accessible.
- Entity-linking from data identifiers in article text:
  - Author-indicated (initially) or text-mined (sometimes)
  - Examples: GENBANK, PDB, TAIR, mostly Life Sciences

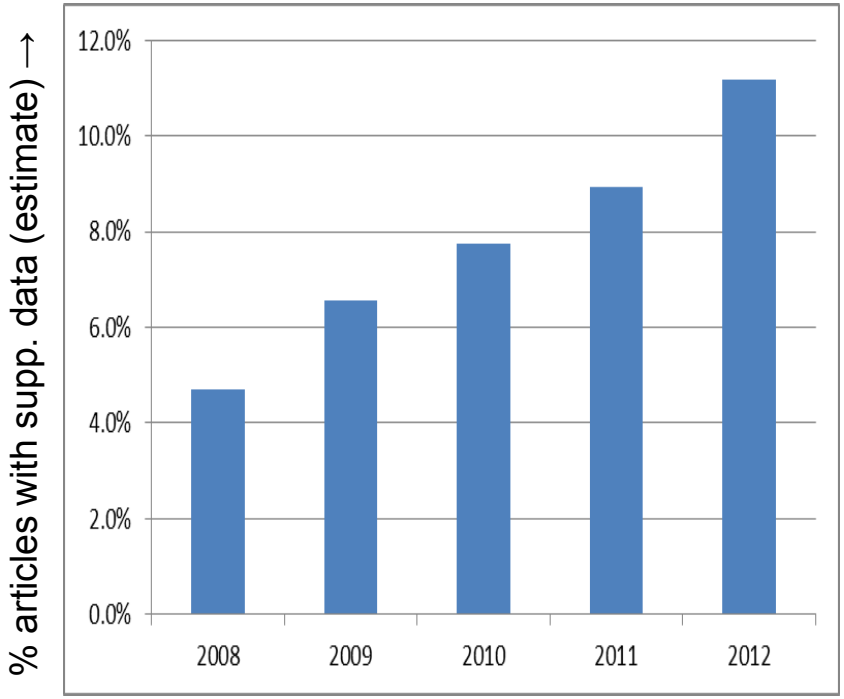
Supplemental files in research articles

What kind of content is submitted as supplementary material?



Most prevalent file types:

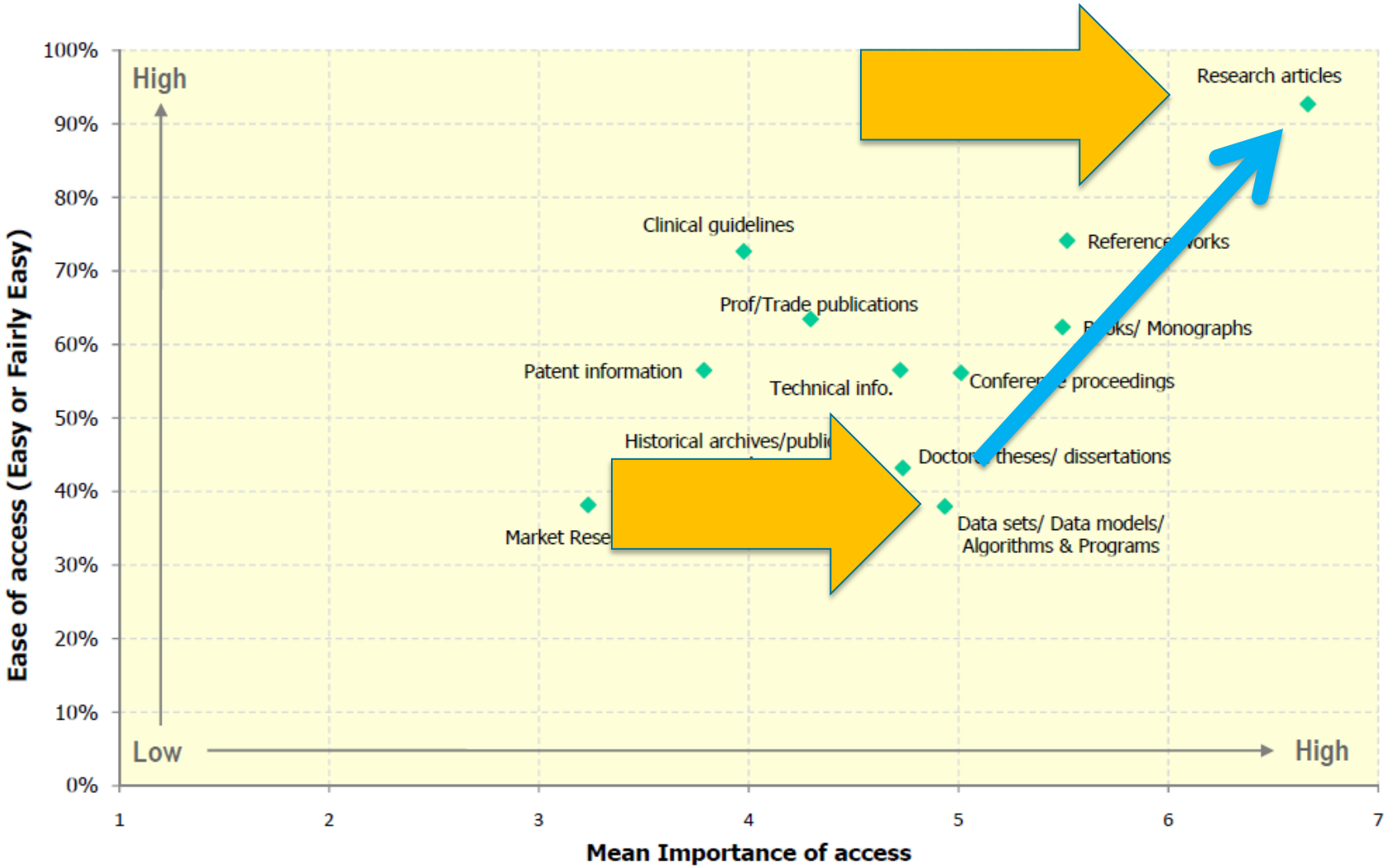
- DOC
- PDF
- ZIP
- XLS
- …



Elsevier data activities started early on, with supplementary files and outward entity linking

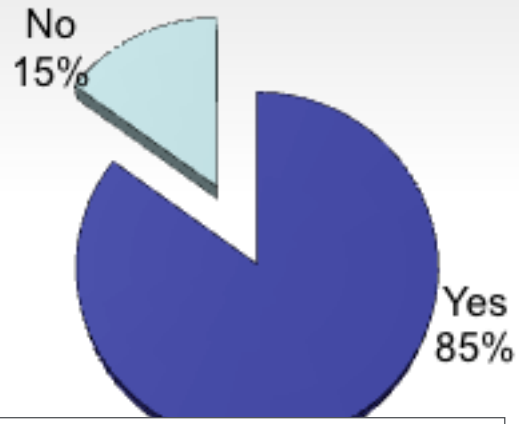
- Accepting supplemental files for over a decade:
  - Linear growth of articles with such files of 30-40% per year
  - Recent snapshot showed 50% of such files contain data
- Suggesting in GfA to post data at repositories:
  - When appropriate data repositories were available
  - When supported by community and editorial boards
- Signing Brussels Declaration on data in 2007:
  - Raw research data should be made freely available. Publishers encourage public posting of such data. Data sets submitted with paper should (wherever possible) be made freely accessible.
- Entity-linking from data identifiers in article text:
  - Author-indicated (initially) or text-mined (sometimes)
  - Examples: GENBANK, PDB, TAIR, mostly Life Sciences

Over time, data sets grew in importance and availability, however they were difficult to find

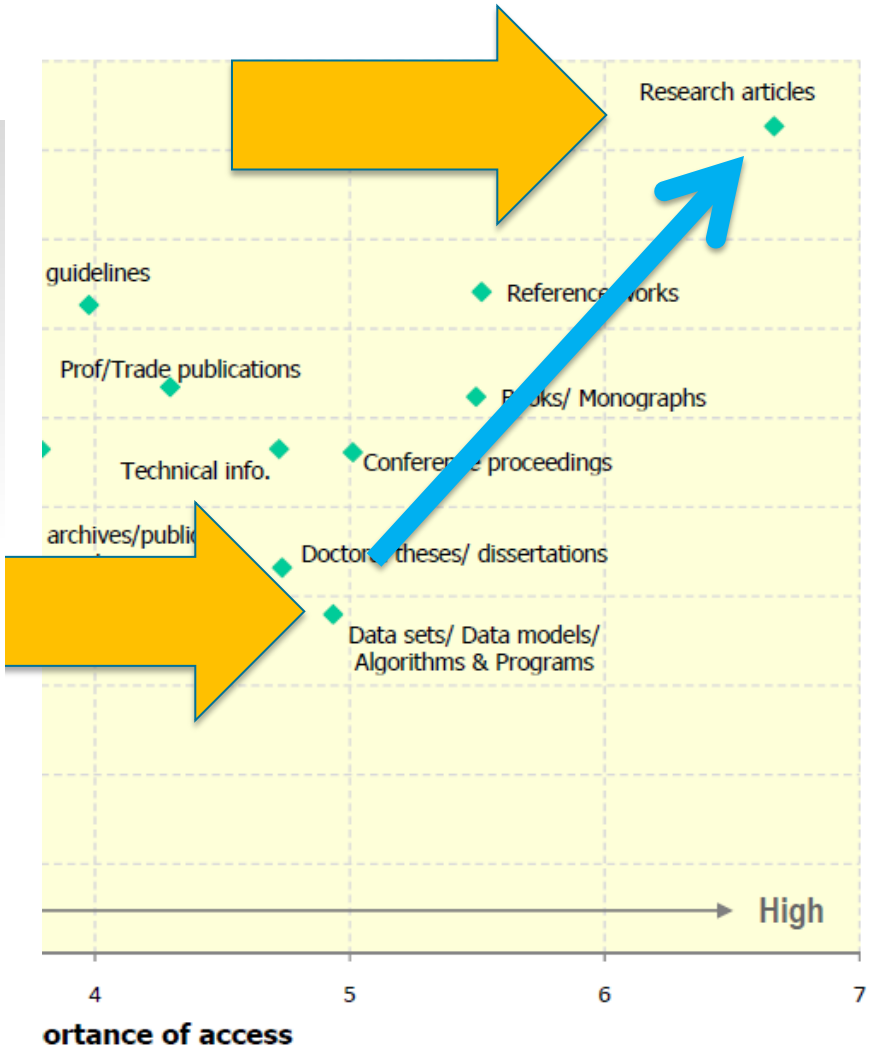


Research data has a data discovery problem

Do you think it is useful to link underlying research data with formal literature?

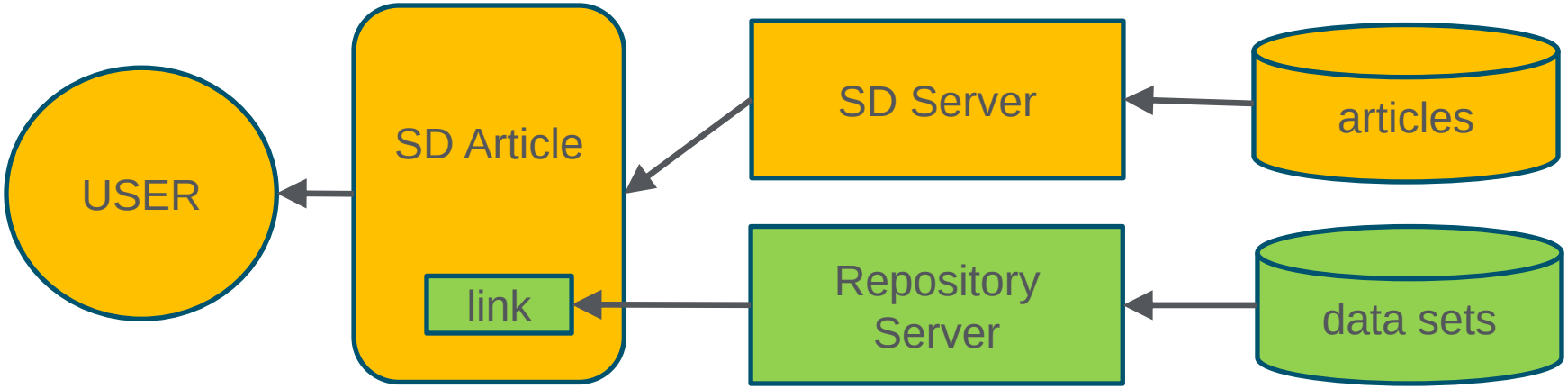


Researcher survey, 1202 respondents (PARSE.insight 2010)



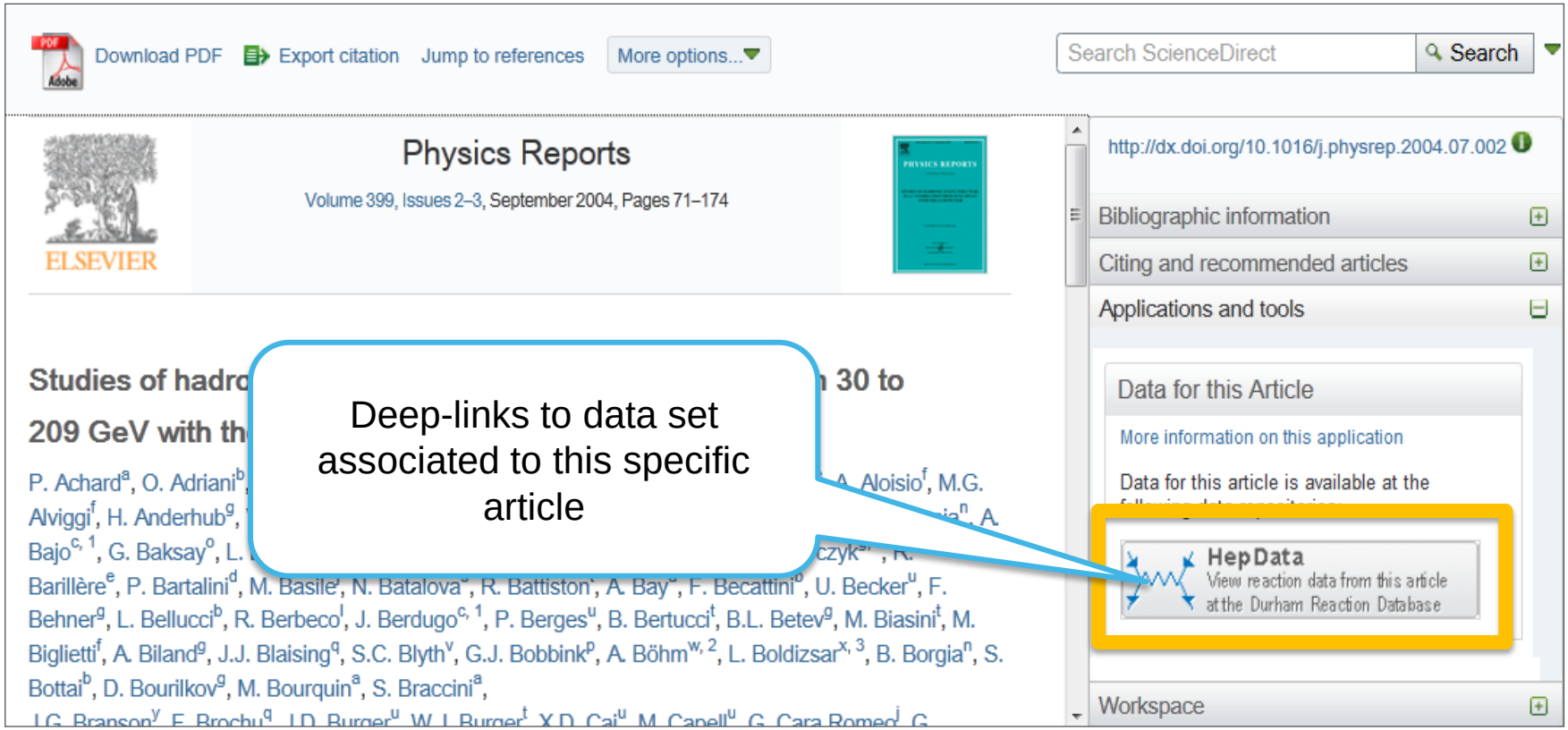
We started a project to use the published articles to better discover associated data sets

- Started collaboration with data repositories – now ~40



- Based on image-based linking – deep link to data set
- SD article asks for a “data set image” from repository
- If data available, repository shows image and link
- If no data available, repository shows 1-pixel transparent image (which is de-facto invisible for the user)

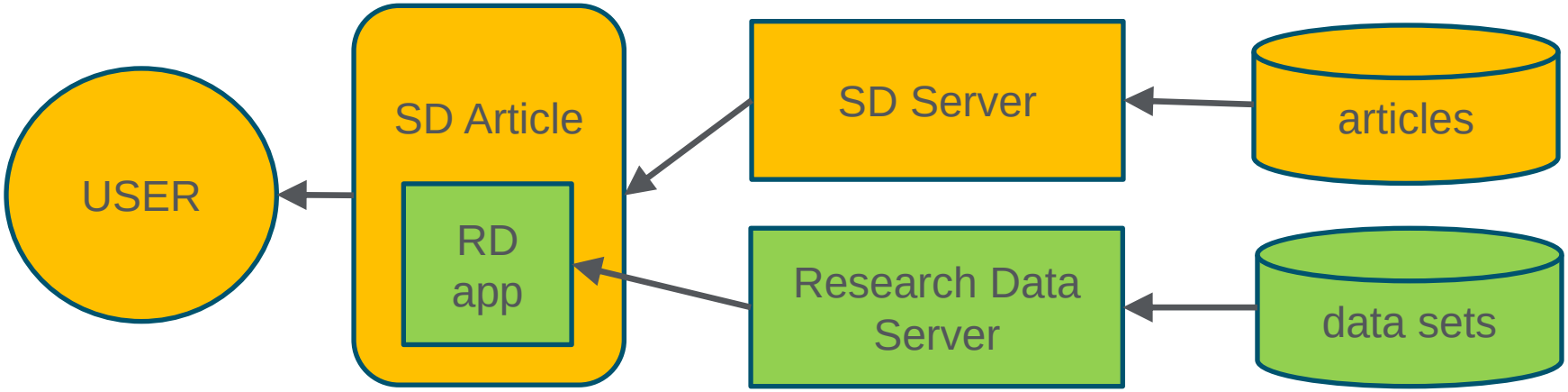
Image link displayed inside published article



http://www.sciencedirect.com/science/article/pii/S0370157304002753

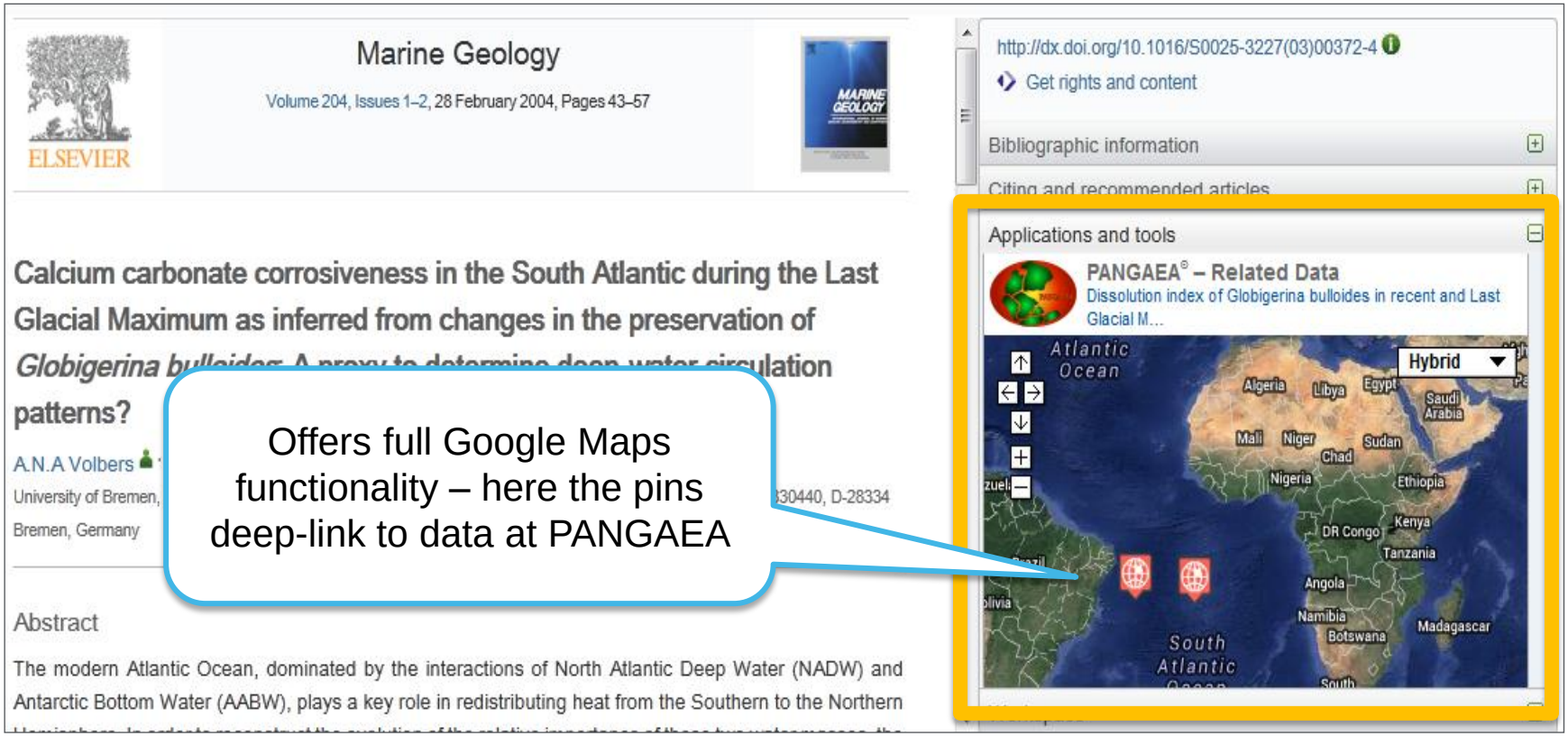
Making data actionable or visualize it inside the article, increases the chance of (re-)use

- Bringing data functionality inside the SD article



- Functionality is specific to data repository / data set
- Application does link to repository for full functionality
- Data set can be interacted with in context of article
- “Research Data Server” can also be supplemental file
- Examples: GenBank, PDB, PANGAEA, MINT

Repository application inside published article



Offers full Google Maps functionality – here the pins deep-link to data at PANGAEA

http://www.sciencedirect.com/science/article/pii/S0025322703003724

Deep link directly goes to associated dataset

ELSEVIER

PANGAEA®
Data Publisher for Earth & Environmental Science

Not logged in (log in or sign up)

Always quote citation when using data!

Data Description

Show Map | Google Earth | RIS | BibTeX

Citation: Lidbury, I et al. (2013): Seawater conditions during the experiment in May 2011 at the sampling sites off Vulcano Island. doi:10.1594/PANGAEA.808529,
Supplement to: **Lidbury, Ian; Johnson, Vivienne R; Hall-Spencer, Jason M; Munn, Colin B; Cunliffe, Michael (2012):** Community-level response of coastal microbial biofilms to ocean acidification in a natural carbon dioxide vent ecosystem. *Marine Pollution Bulletin*, **64(5)**, 1063-1066, doi:10.1016/j.marpolbul.2012.02.011

Abstract: The impacts of ocean acidification on coastal biofilms are poorly understood. Carbon dioxide vent areas provide an opportunity to make predictions about the impacts of ocean acidification. We compared biofilms that colonised glass slides in areas exposed to ambient and elevated levels of pCO2 along a coastal pH gradient, with biofilms grown at ambient and reduced light levels. Biofilm production was highest under ambient light levels, but under both light regimes biofilm production was enhanced in seawater with high pCO2. Uronic acids are a component of biofilms and increased significantly with high pCO2. Bacteria and Eukarya denaturing gradient gel electrophoresis profile analysis showed clear differences in the structures of ambient and reduced light biofilm communities, and biofilms grown at high pCO2 compared with ambient conditions. This study characterises biofilm response to natural seabed CO2 seeps and provides a baseline understanding of how coastal ecosystems may respond to increased pCO2 levels.

Project(s): **Mediterranean Sea Acidification in a Changing Climate** (MedSeA)

Coverage: *Latitude:* 38.416700 * *Longitude:* 14.950000

Minimum DEPTH, water: 1.0 m * *Maximum DEPTH, water:* 1.0 m

Event(s): **Vulcano** * *Latitude:* 38.416700 * *Longitude:* 14.950000 * *Location:* Vulcano, Aeolian Islands, North East Sicily, Italy * *Device:* Experiment

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment

http://www.sciencedirect.com/science/article/pii/S0025322703003724

Application can reveal data underlying plots

higher than that obtained in traditional membrane distillation processes. The maximum value of J_D and GOR could reach 6.98 kg/m^{2}h and 6.44 respectively. The effects of various operation parameters including feed temperatures and feed flow rate on the performance of the AGMD process had been investigated. The effects of various membrane module parameters such as membrane porosity (ε), membrane pore size (d_r), the rate of hollow fibers and membranes (N_d/N_m), the thermal conductivity coefficient of heat-exchange hollow fibers (k_d), the thickness of hollow fibers (d_d) and the air gap width (d_a) were experimentally studied. The high saline water of 70 g/L was concentrated to about 308 g/L in the deep-concentration exper ... um value of the electrical conductivity of t

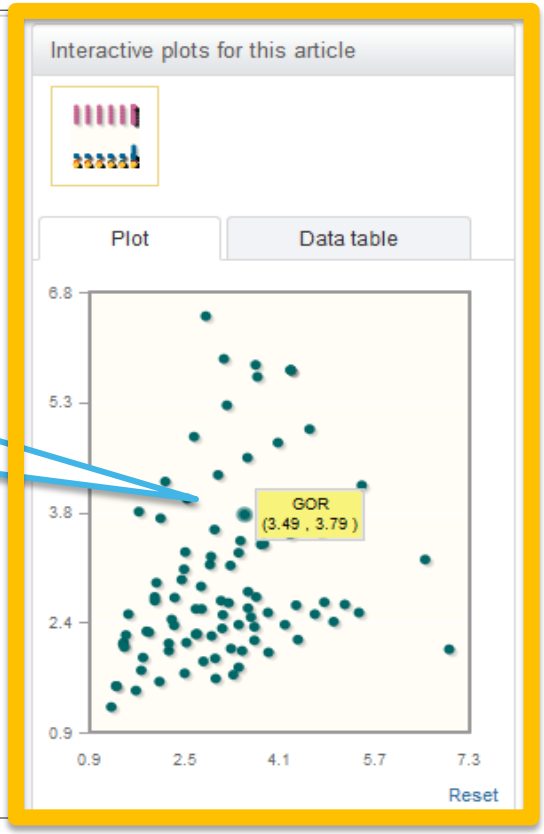
Keywords

Hollow fiber AGMD modu

1. Introduction

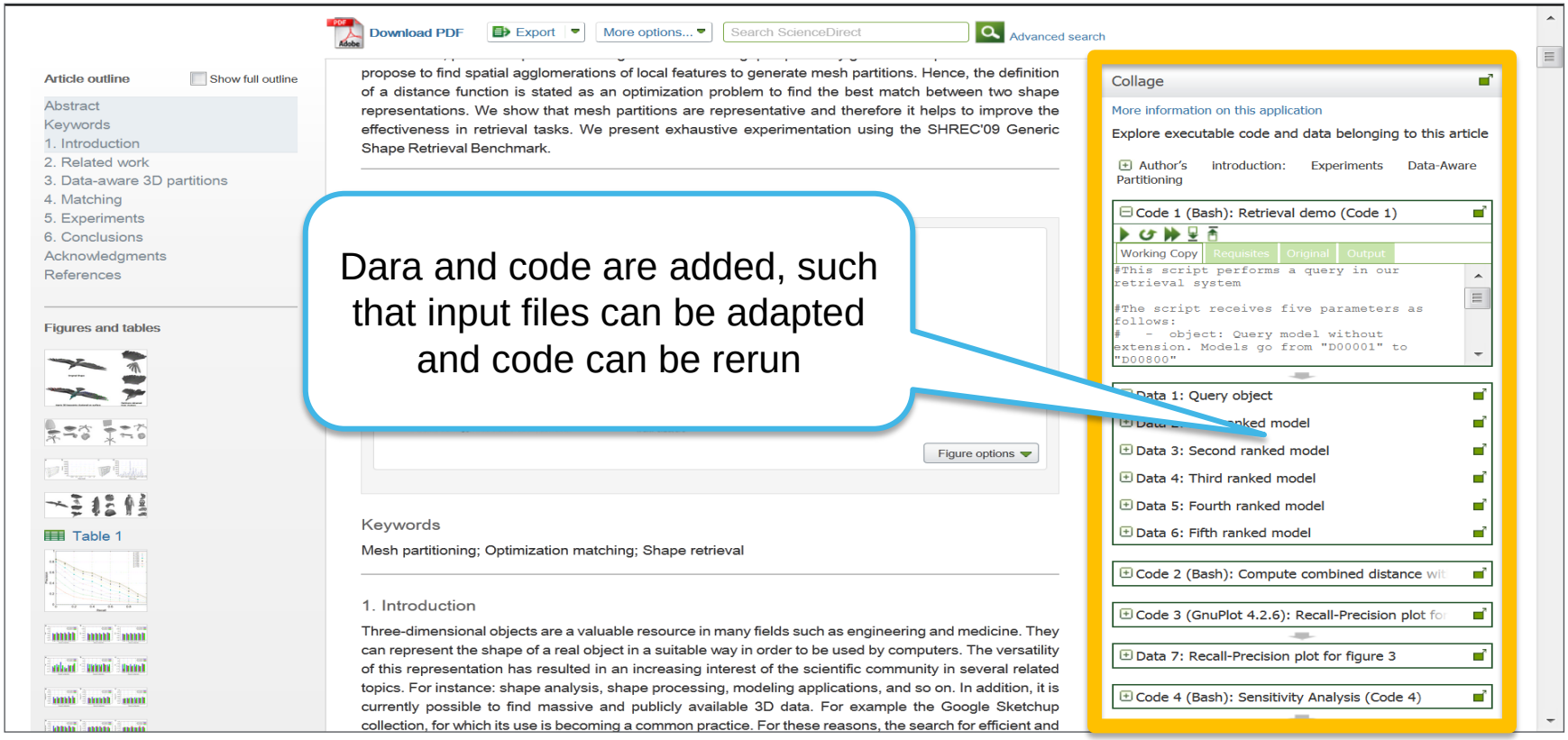
Water shortage is one of the most serious global challenges. Presently, over one-third of the world's population lives in water-stressed areas [1]. Seawater desalination is one of the most economic ways to get freshwater, which is expected to be an effective way to relieve the global water crisis because seawater desalination offers a seemingly unlimited, steady supply of high-quality water [1]. Now the total global desalination capacity is around 66.4 million m^3/d and it is expected to reach 100 million m^3/d by 2015 [2] and [3]. The recovery of the RO process with one stage is only 35%–45% and can reach only 60% if the second stage is applied [4]. The high operation pressure and shortage of high strength membranes limit RO to deal with high saline water. The RO concentrated solution containing chemical substances produced in the seawater pretreatment process for scaling control, fouling, and corrosion preventions, and

Interactive plots show the data behind figures (data also available for download)



http://www.sciencedirect.com/science/article/pii/S0011916414001520

Executable papers allow immediate validation

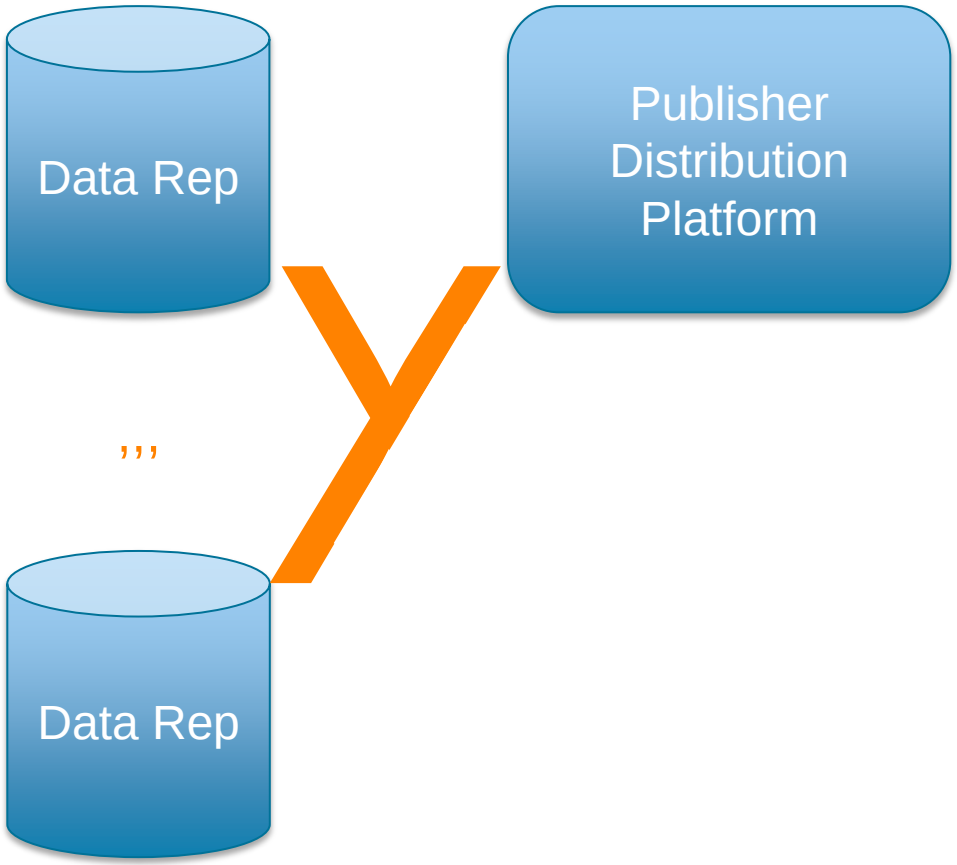


http://www.sciencedirect.com/science/article/pii/S0097849313000484

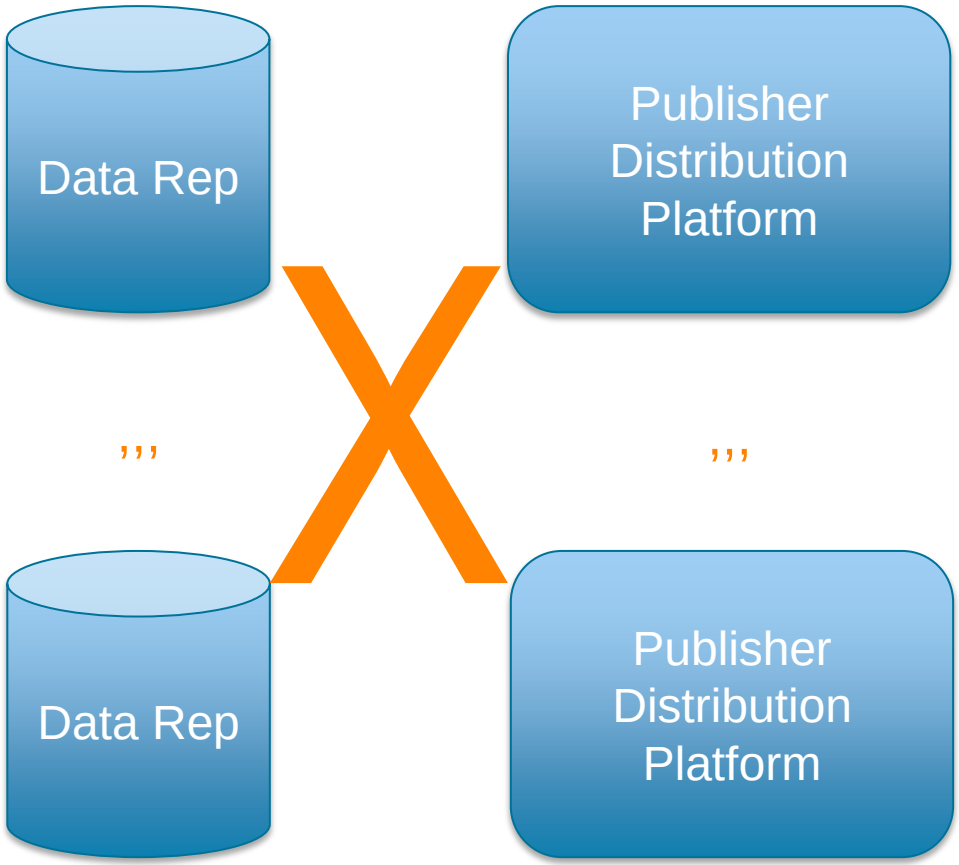
Up to 40 data repository linking partners

http://www.elsevier.com/databaselinking

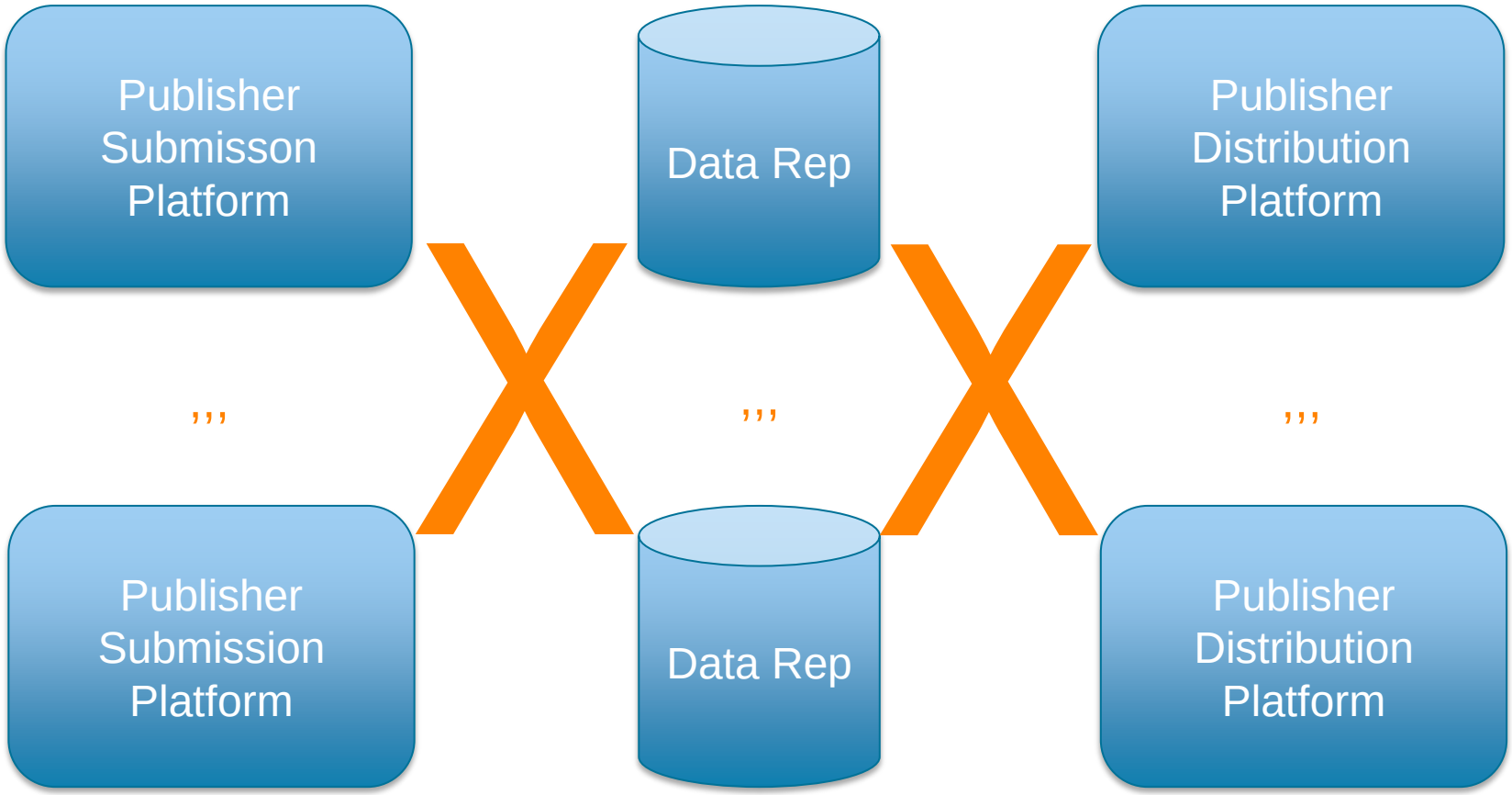
Current linking approach isn't scalable; an increase in repositories requires standards



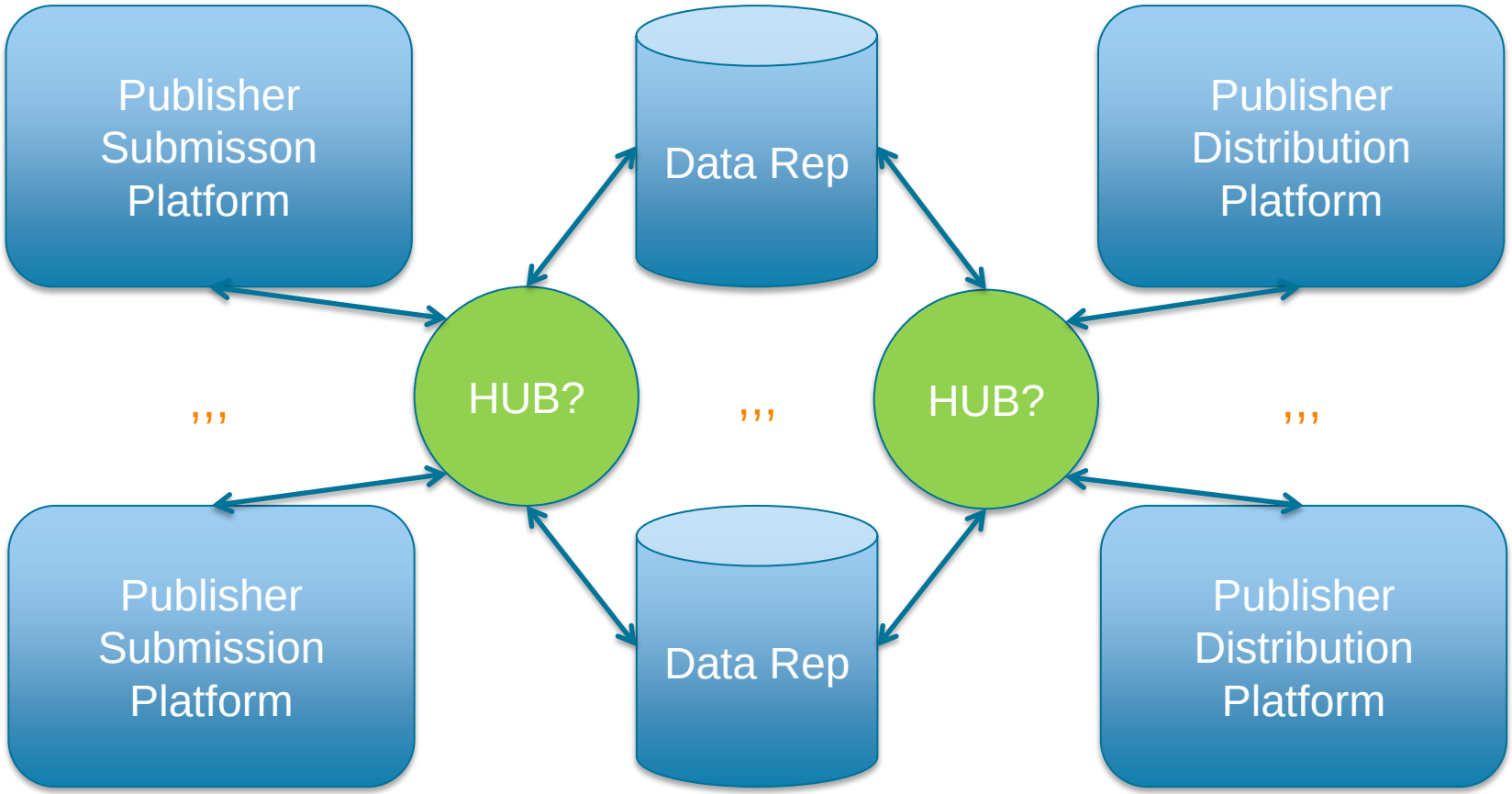
Current linking approach isn't scalable; an increase in repositories requires standards



This especially holds when publishers start to support data posting at variety of repositories



Requires solution for article-data connections, and for data submission interoperability

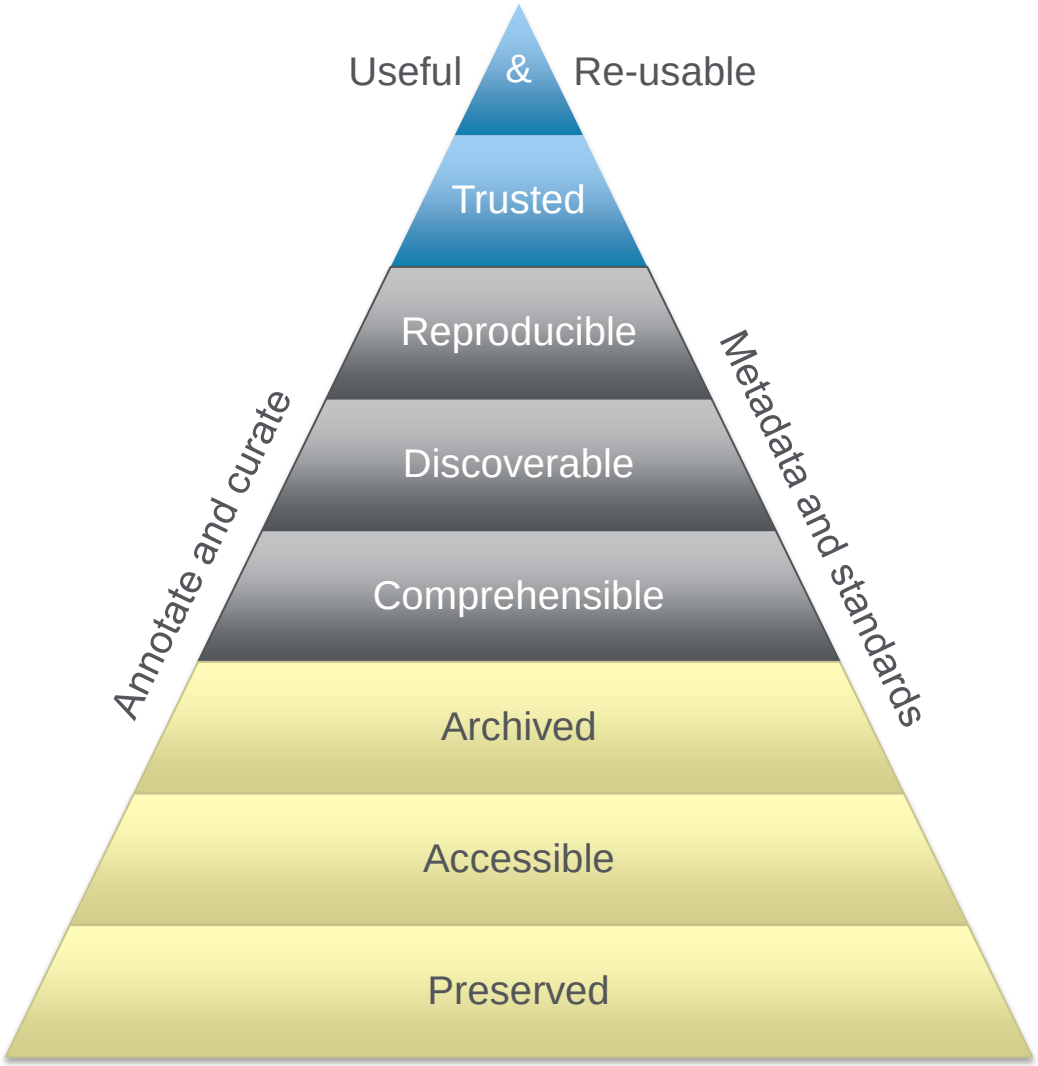


Looking at Researcher Data Workflow shows that also other standards are highly needed

Main Task	Activities	Needs
Experiment	Plan, measure, record, analyze, annotate, store, archive, preserve, …	Workflow and analysis tools, ELN, standards, metadata
Publish	Prepare, post, submit, get reviewed, publish, get cited, get credit, …	Public hosting, data space, standards, metadata
Re-use	Curate, search, access, analyze, …	Standards, metadata, analysis tools

- Publishers (and others) operate in all task areas
- Effective interoperable infrastructure needs standards
- Generic, discipline-specific, and data and metadata
- RDA, WDS, CODATA, Force11 – now data citation

Such standards are also required to move the data from bits and bytes to fully (re-) usable



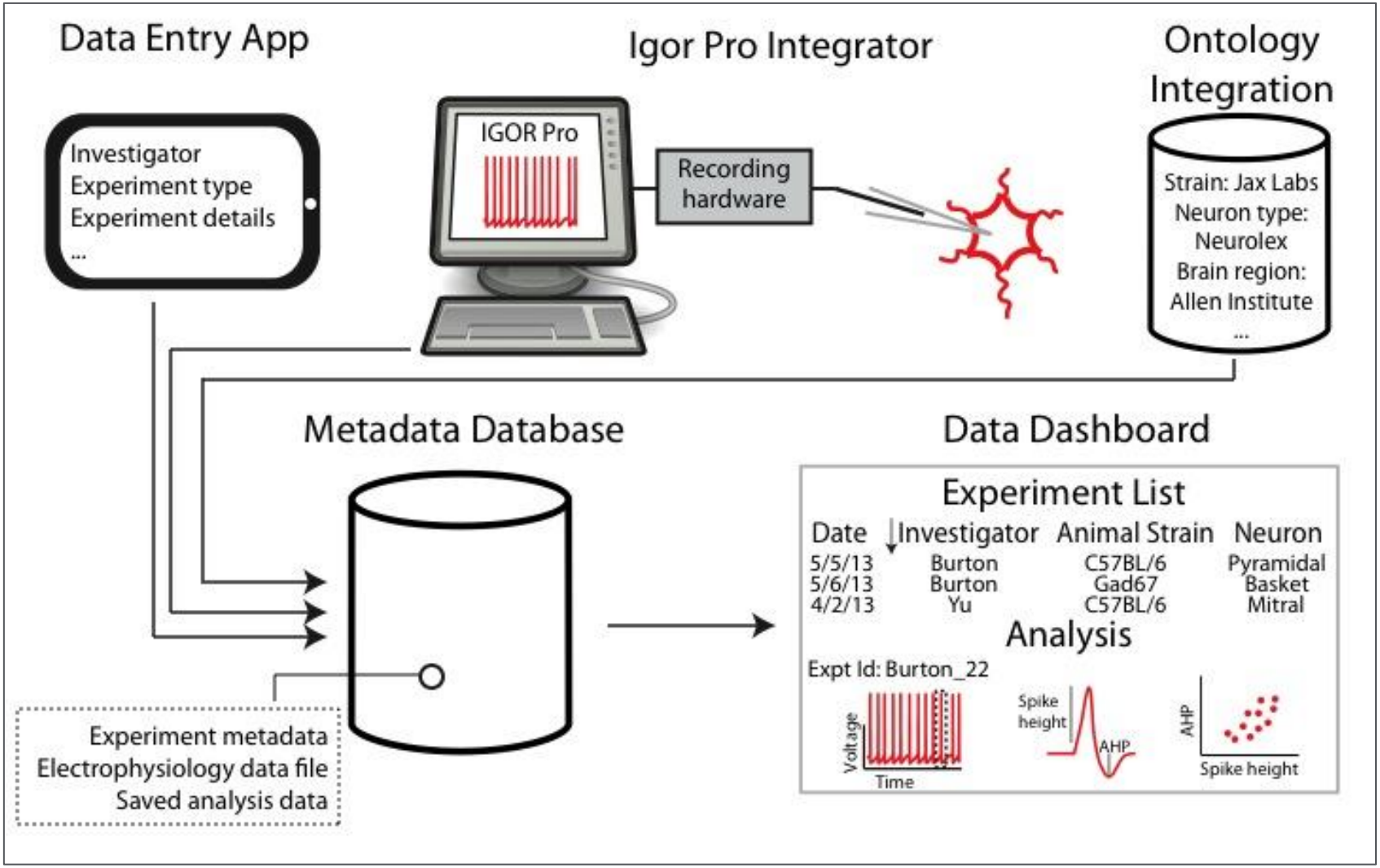
Elsevier initiatives:

- Executable papers
- Microarticles
- Data articles
- Data linking
- Data integration
- Supplem. files
- Standard Cmt
- Pilot projects:
  - Urban Legend
  - Moonrocks

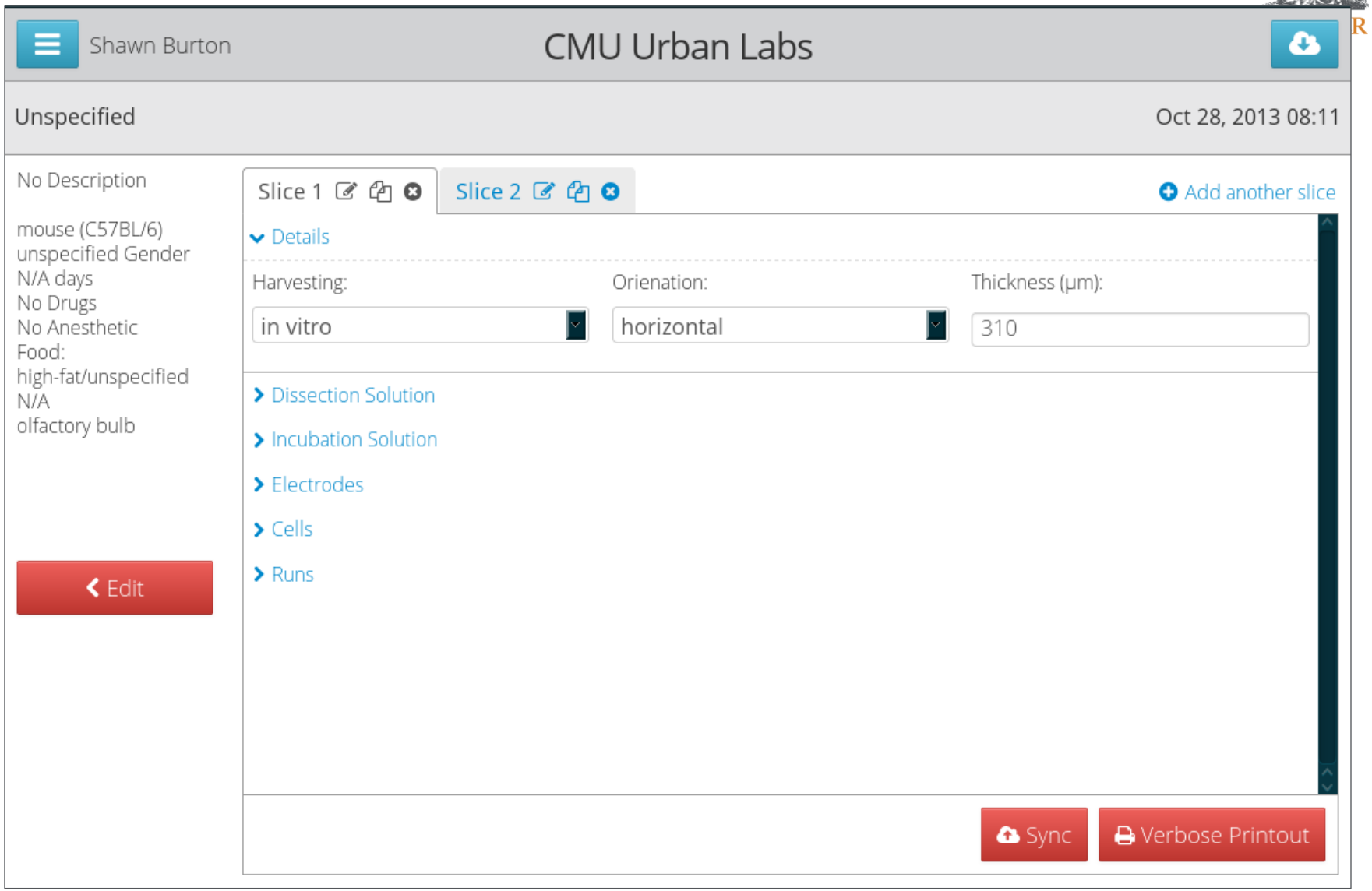
Urban Legend – with CMU

- *How can we make a standard neuroscience wet lab store and share their data?*
- Incorporate structured workflows into the daily practice of a typical electrophysiology lab (the Urban Lab at CMU)
  - What does it take?
  - Where are points of conflict?
- 1-year pilot, funded by Elsevier
  - CMU: Shreejoy Tripathy, manage/user test
  - Elsevier: development, UI, project management

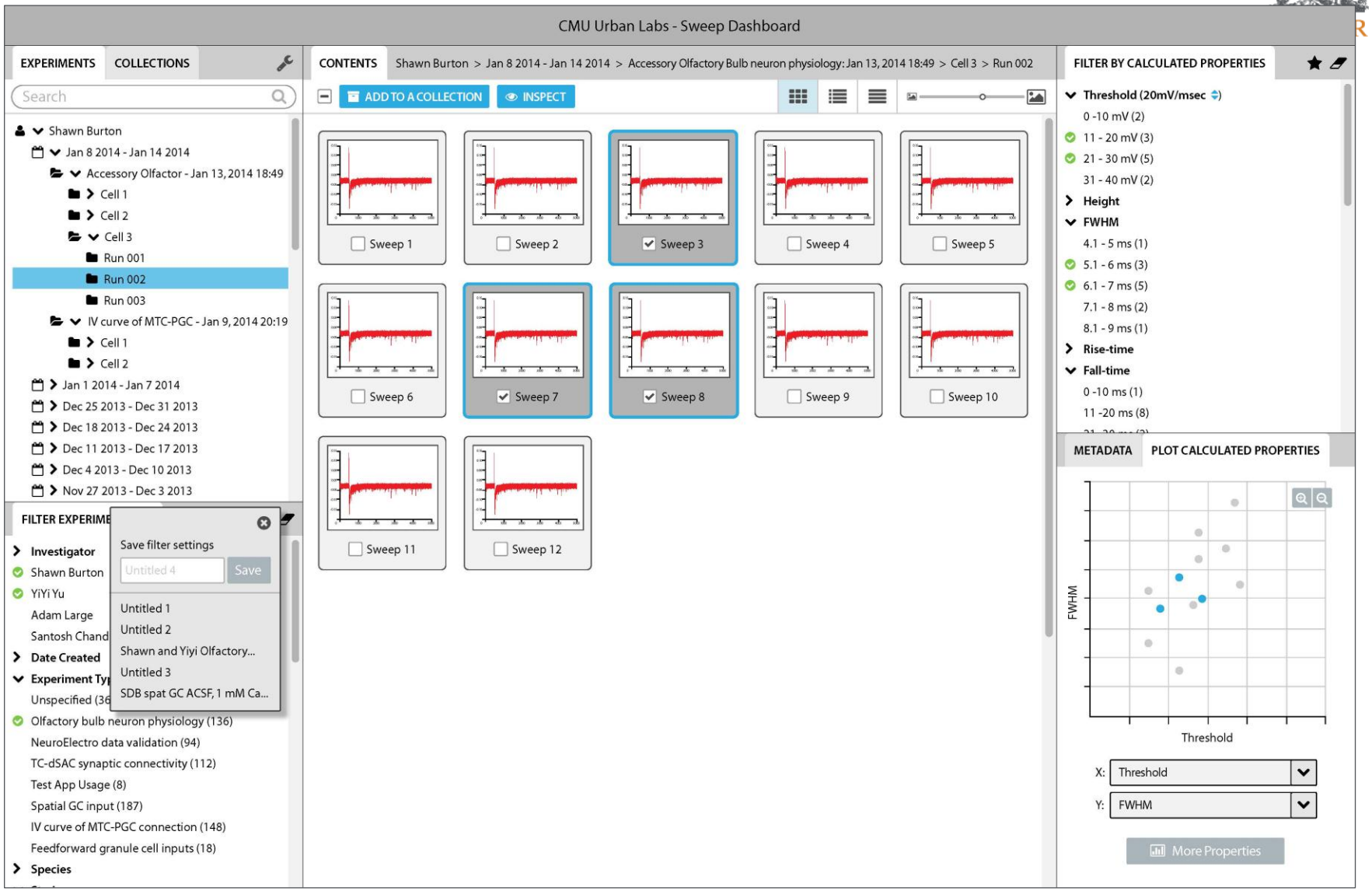
Urban Legend – Components



Urban Legend – Data Entry App



Urban Legend – Data Dashboard

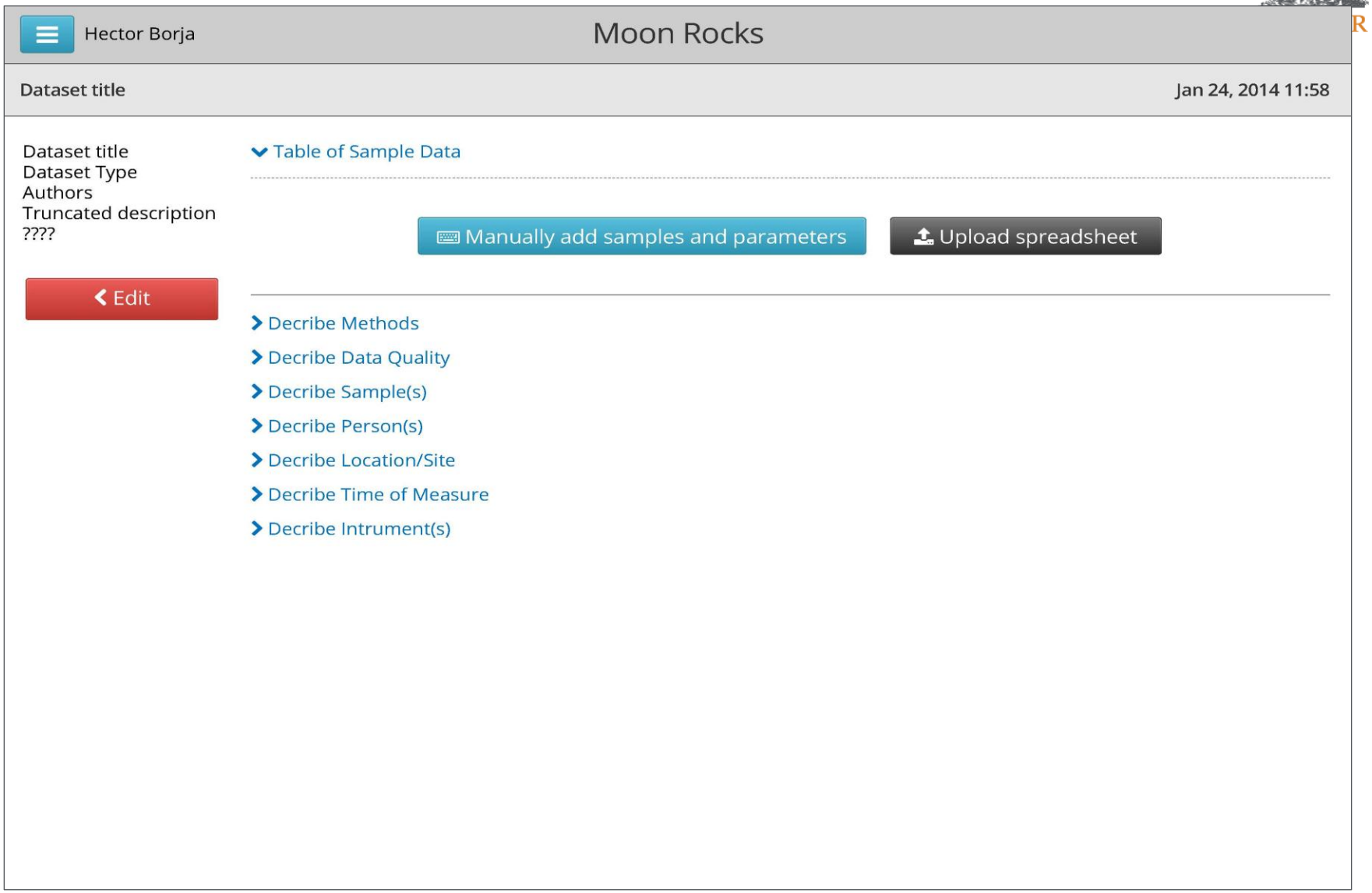


Moonrocks – with IEDA

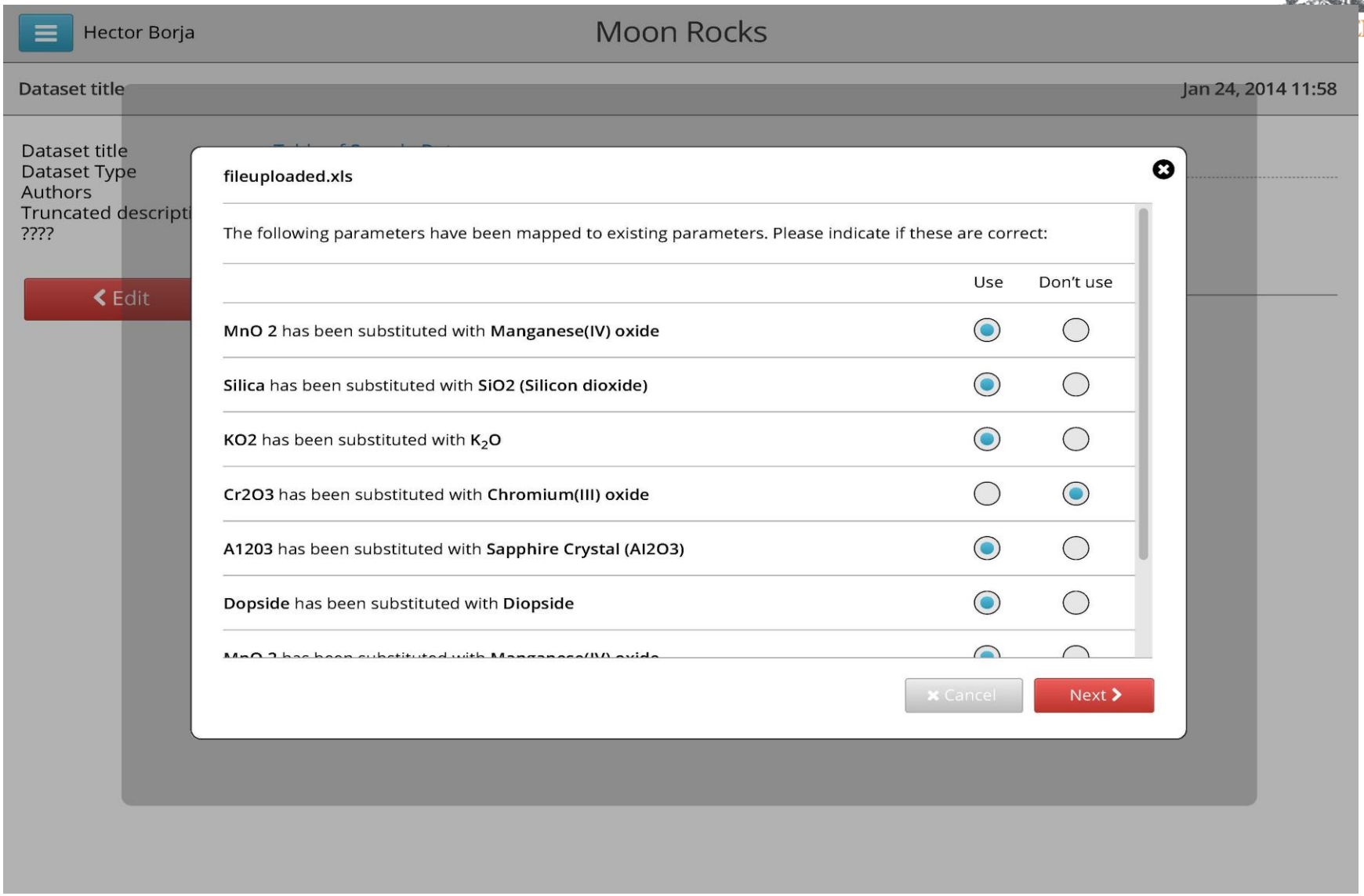
- *How can we scale up data curation?*

- Build a database for lunar geochemistry
- Leapfrog & improve curation time
- Determine best practices and challenges
- Estimate costs
- 1-year pilot, funded by Elsevier

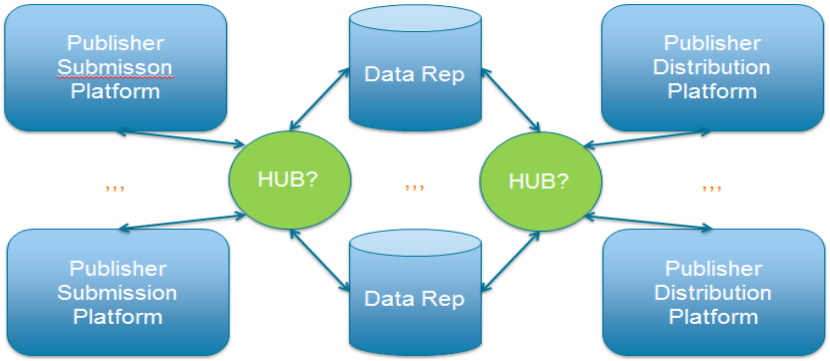
Moonrocks – Data Entry

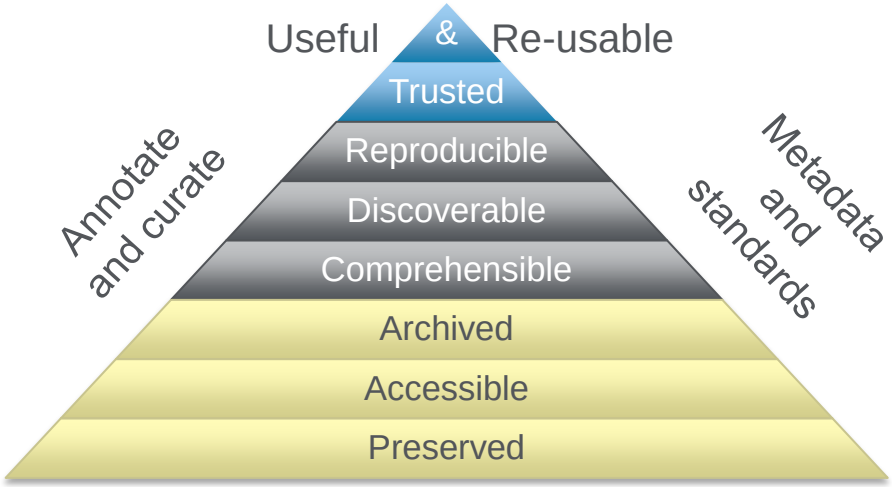


Urban Legend – Data Dashboard



1) Interoperable architecture for subm and disc
2) Metadata and standards for value re-use





Elsevier initiatives:

- Executable papers
- Microarticles
- Data articles
- Data linking
- Data integration
- Supplem. files
- Standard Cmt
- Pilot projects:
  - Urban Legend
  - Moonrocks