



ELSEVIER

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

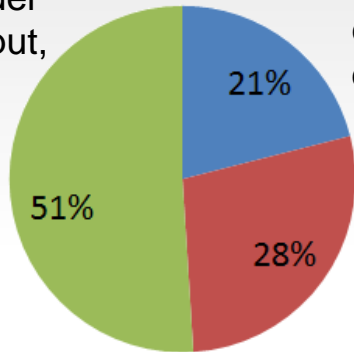


- Accepting supplemental files for over a decade:
 - Growth of articles with such files of 25-35% per year
 - Recent snapshot showed 50% of such files contain data
- Suggesting in GfA to post data at repositories:
 - When appropriate data repositories were available
 - When supported by community and 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?

Methods, videos,
text, references,
model
output,
etc.

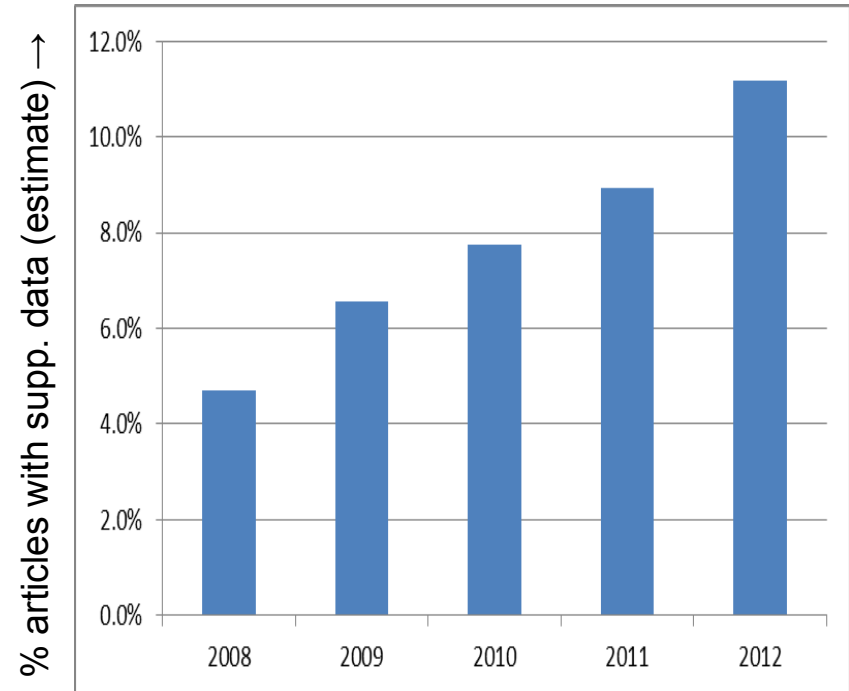


Raw,
experimental
data

Combination
of data and
other material

Most prevalent file types:

- DOC
- PDF
- ZIP
- XLS
- ...

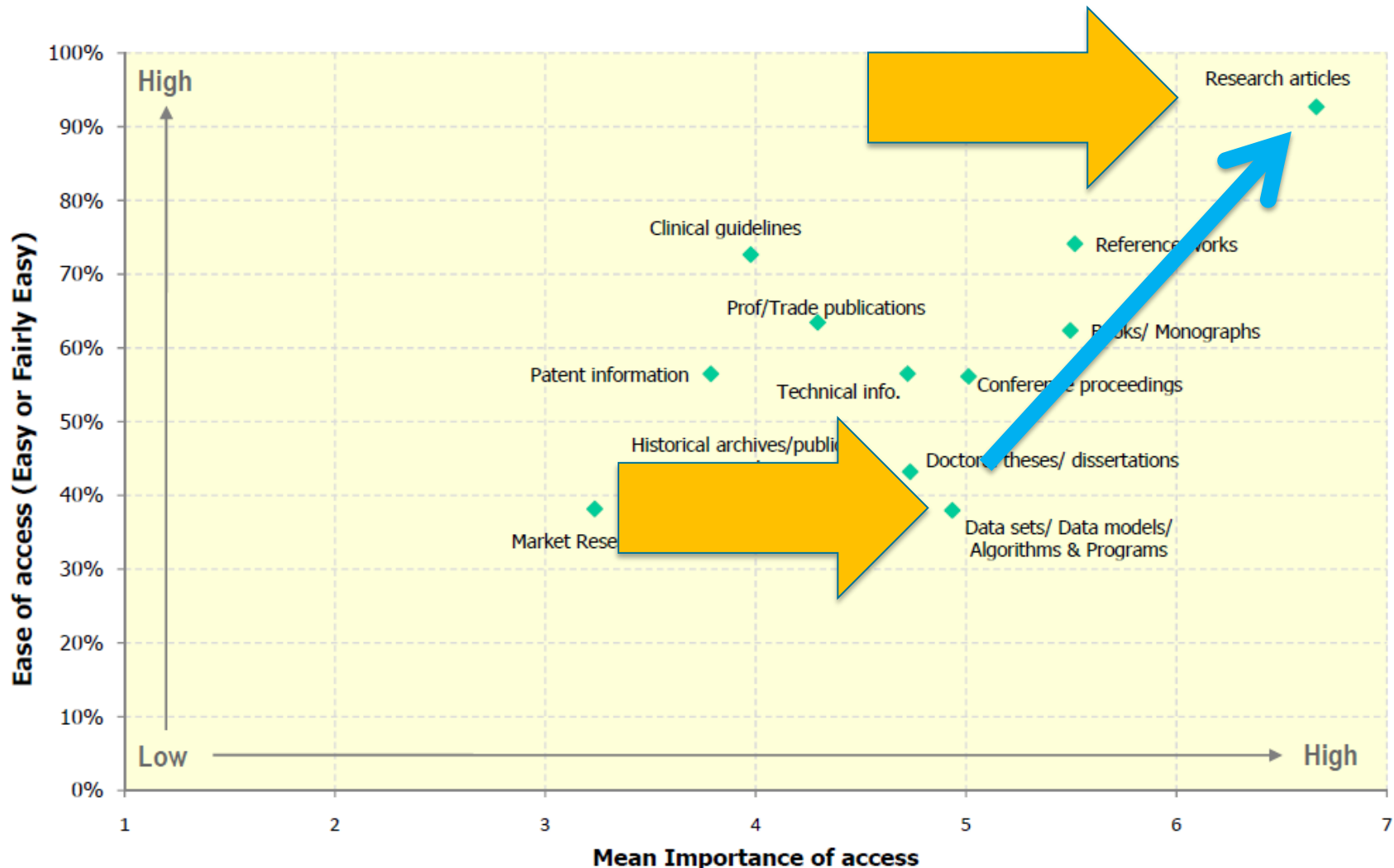


Elsevier data activities started early on, with supplementary files and outward entity linking



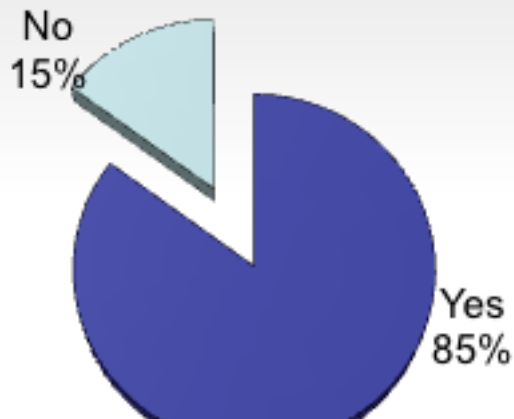
- Accepting supplemental files for over a decade:
 - Linear growth of articles with such files of 30-40% per year
 - Recent snapshot showed 50% of such files contain data
- Suggesting in GfA to post data at repositories:
 - When appropriate data repositories were available
 - When supported by community 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?



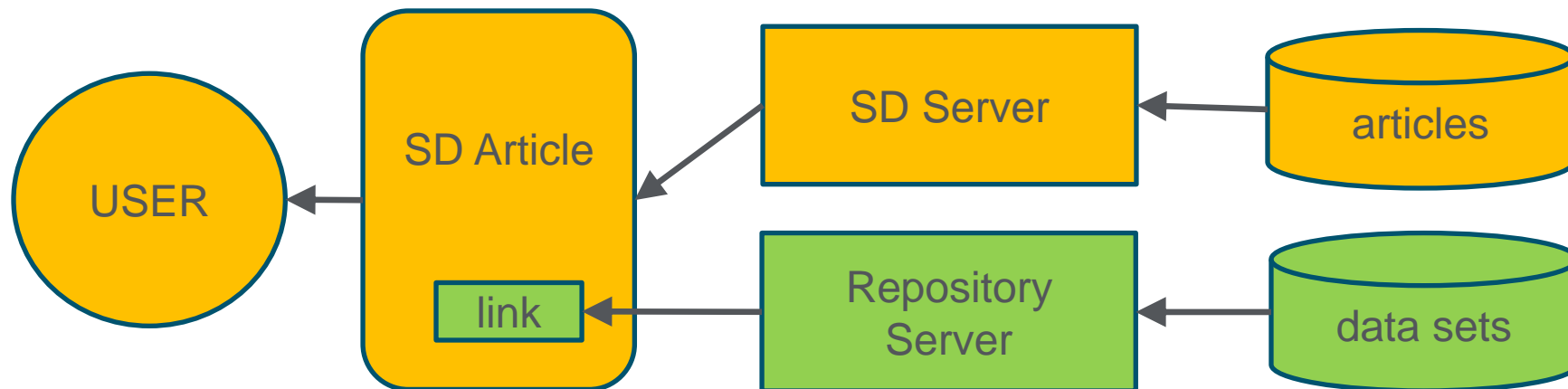
Researcher survey, 1202 respondents (PARSE.insight 2010)



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

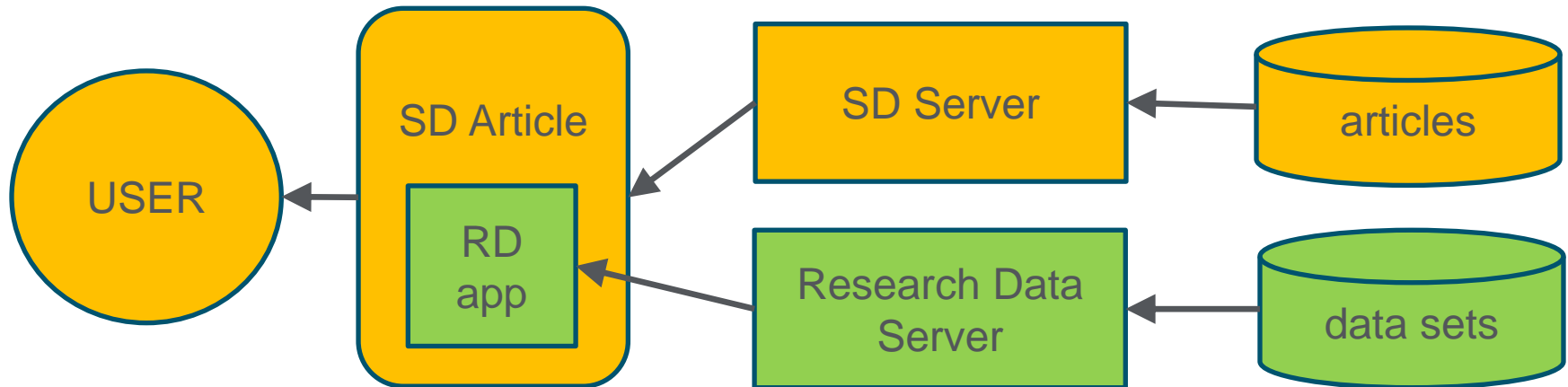


The screenshot shows the ScienceDirect interface for a Physics Reports article. At the top, there are navigation options: 'Download PDF', 'Export citation', 'Jump to references', and 'More options...'. A search bar is located in the top right corner. The article title 'Physics Reports' is prominently displayed, along with the volume and issue information: 'Volume 399, Issues 2-3, September 2004, Pages 71-174'. The main content area shows the beginning of the article text, including the title 'Studies of hadro...' and the first few lines of the abstract. A callout box with a blue border and white background points to a 'HepData' link in the right-hand sidebar. The callout box contains the text: 'Deep-links to data set associated to this specific article'. The 'HepData' link in the sidebar is highlighted with a yellow box and contains the text: 'HepData View reaction data from this article at the Durham Reaction Database'. The sidebar also includes sections for 'Bibliographic information', 'Citing and recommended articles', 'Applications and tools', and 'Data for this Article'.

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

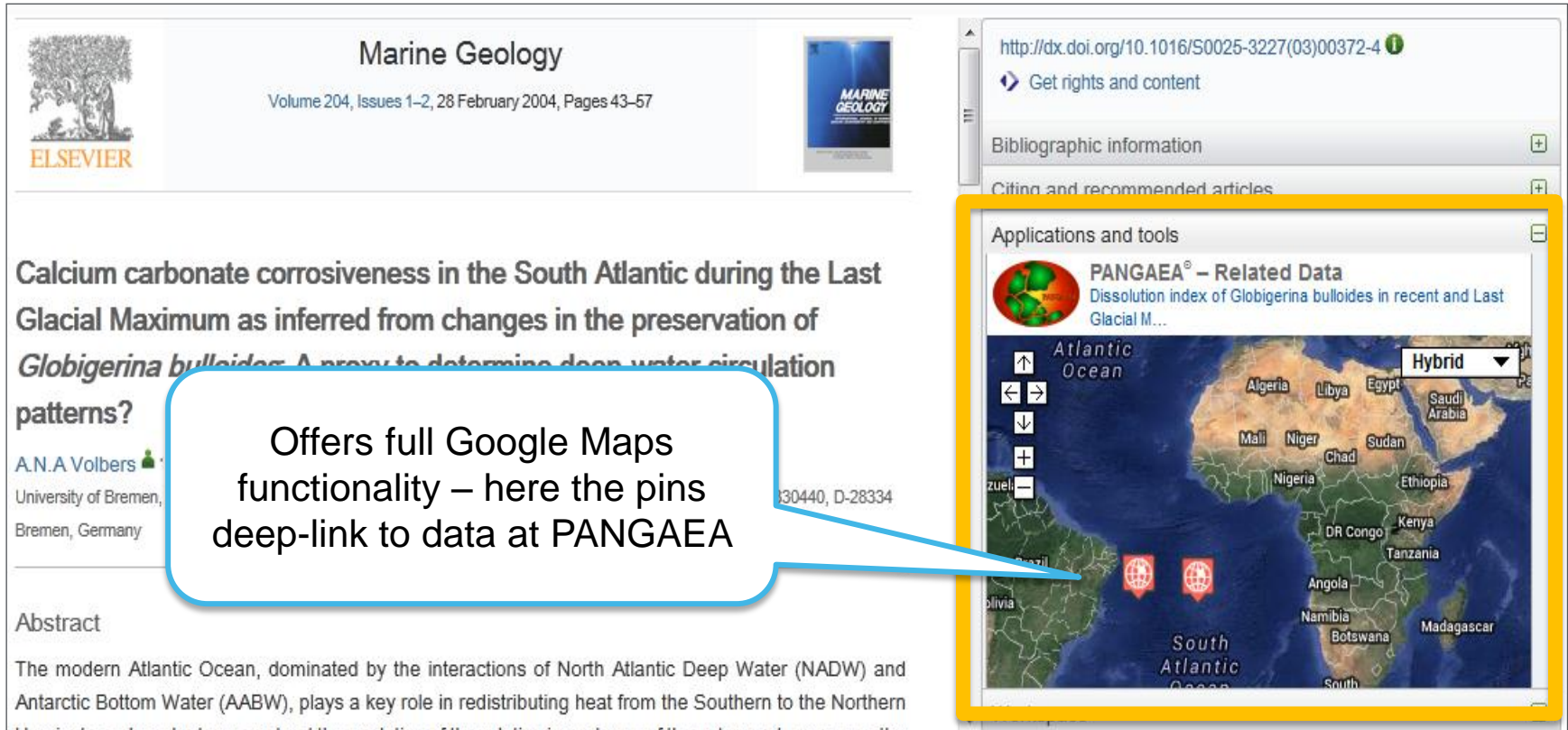
Making data actionable or visualize it inside the article, increases the chance of (re-)use

- 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



The screenshot shows a scientific article page from Elsevier. The journal is 'Marine Geology', Volume 204, Issues 1-2, 28 February 2004, Pages 43-57. The article title is 'Calcium carbonate corrosiveness in the South Atlantic during the Last Glacial Maximum as inferred from changes in the preservation of *Globigerina bulloides*. A proxy to determine deep water circulation patterns?'. The author is A.N.A. Volbers, University of Bremen, Bremen, Germany. The article ID is S0025322703003724. The 'Applications and tools' section is highlighted with a yellow border and contains a link to 'PANGAEA® - Related Data' with the description 'Dissolution index of *Globigerina bulloides* in recent and Last Glacial M...'. Below the text is a map of the Atlantic Ocean showing the South Atlantic region with two red pins indicating data locations. The map includes labels for various countries and regions like Algeria, Libya, Egypt, Saudi Arabia, Mali, Niger, Chad, Sudan, Nigeria, Ethiopia, DR Congo, Kenya, Tanzania, Angola, Namibia, Botswana, Madagascar, and South Africa. The map is titled 'Atlantic Ocean' and 'South Atlantic Ocean'.

Marine Geology
Volume 204, Issues 1-2, 28 February 2004, Pages 43-57

Calcium carbonate corrosiveness in the South Atlantic during the Last Glacial Maximum as inferred from changes in the preservation of *Globigerina bulloides*. A proxy to determine deep water circulation patterns?

A.N.A. Volbers
University of Bremen,
Bremen, Germany

Abstract
The modern Atlantic Ocean, dominated by the interactions of North Atlantic Deep Water (NADW) and Antarctic Bottom Water (AABW), plays a key role in redistributing heat from the Southern to the Northern

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

Applications and tools

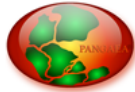
PANGAEA® - Related Data
Dissolution index of *Globigerina bulloides* in recent and Last Glacial M...

Atlantic Ocean
South Atlantic Ocean

Algeria, Libya, Egypt, Saudi Arabia, Mali, Niger, Chad, Sudan, Nigeria, Ethiopia, DR Congo, Kenya, Tanzania, Angola, Namibia, Botswana, Madagascar, South Africa

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

Deep link directly goes to associated dataset



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 $6.98 \text{ kg/m}^2\text{h}$ 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_p), 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 experiment. The maximum value of the electrical conductivity of the

Keywords

Hollow fiber AGMD mod

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 $66.4 \text{ million m}^3/\text{d}$ and it is expected to reach $100 \text{ million m}^3/\text{d}$ 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

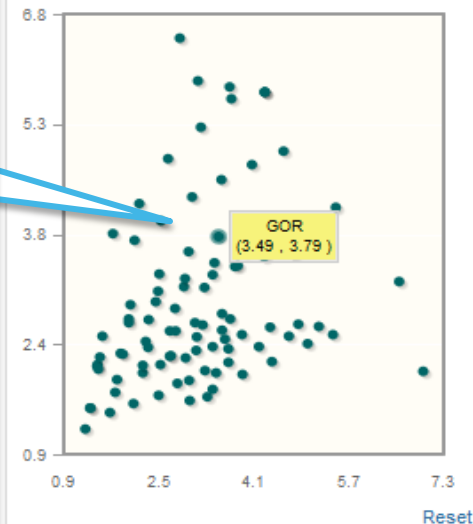
Interactive plots show the data behind figures (data also available for download)

Interactive plots for this article



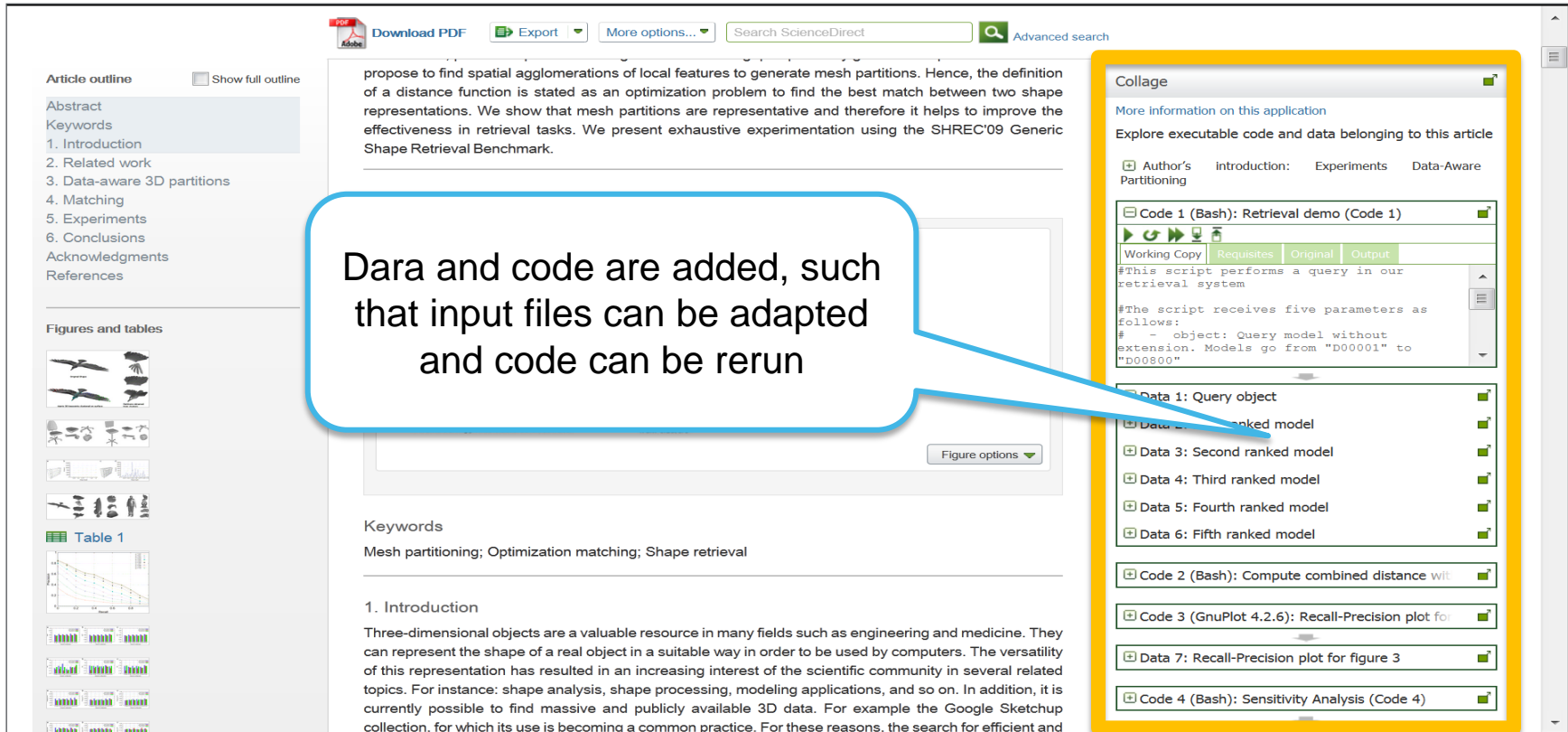
Plot

Data table



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

Executable papers allow immediate validation



Article outline Show full outline

Abstract
Keywords
1. Introduction
2. Related work
3. Data-aware 3D partitions
4. Matching
5. Experiments
6. Conclusions
Acknowledgments
References

Figures and tables

Table 1

Download PDF **Export** **More options...** Search ScienceDirect **Advanced search**

propose to find spatial agglomerations of local features to generate mesh partitions. Hence, the definition of a distance function is stated as an optimization problem to find the best match between two shape representations. We show that mesh partitions are representative and therefore it helps to improve the effectiveness in retrieval tasks. We present exhaustive experimentation using the SHREC'09 Generic Shape Retrieval Benchmark.

Keywords
Mesh partitioning; Optimization matching; Shape retrieval

1. Introduction
Three-dimensional objects are a valuable resource in many fields such as engineering and medicine. They can represent the shape of a real object in a suitable way in order to be used by computers. The versatility of this representation has resulted in an increasing interest of the scientific community in several related topics. For instance: shape analysis, shape processing, modeling applications, and so on. In addition, it is currently possible to find massive and publicly available 3D data. For example the Google Sketchup collection, for which its use is becoming a common practice. For these reasons, the search for efficient and

Collage

More information on this application
Explore executable code and data belonging to this article

Author's Introduction: Experiments Data-Aware Partitioning

Code 1 (Bash): Retrieval demo (Code 1)

Working Copy Requisites Original Output

```
#This script performs a query in our retrieval system  
  
#The script receives five parameters as follows:  
# - object: Query model without extension. Models go from "D00001" to "D00800"
```

Data 1: Query object

Data 2: Ranked model

Data 3: Second ranked model

Data 4: Third ranked model

Data 5: Fourth ranked model

Data 6: Fifth ranked model

Code 2 (Bash): Compute combined distance with

Code 3 (GnuPlot 4.2.6): Recall-Precision plot for

Data 7: Recall-Precision plot for figure 3

Code 4 (Bash): Sensitivity Analysis (Code 4)

Dara and code are added, such that input files can be adapted and code can be rerun

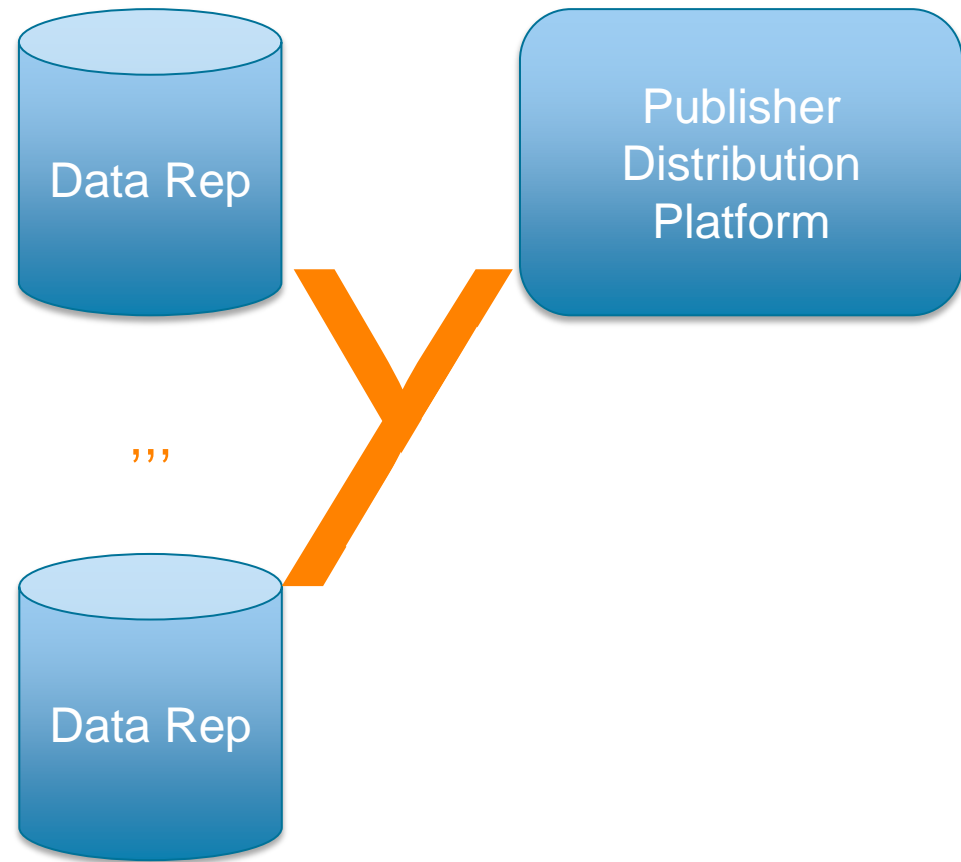
<http://www.sciencedirect.com/science/article/pii/S0097849313000484>

Up to 40 data repository linking partners

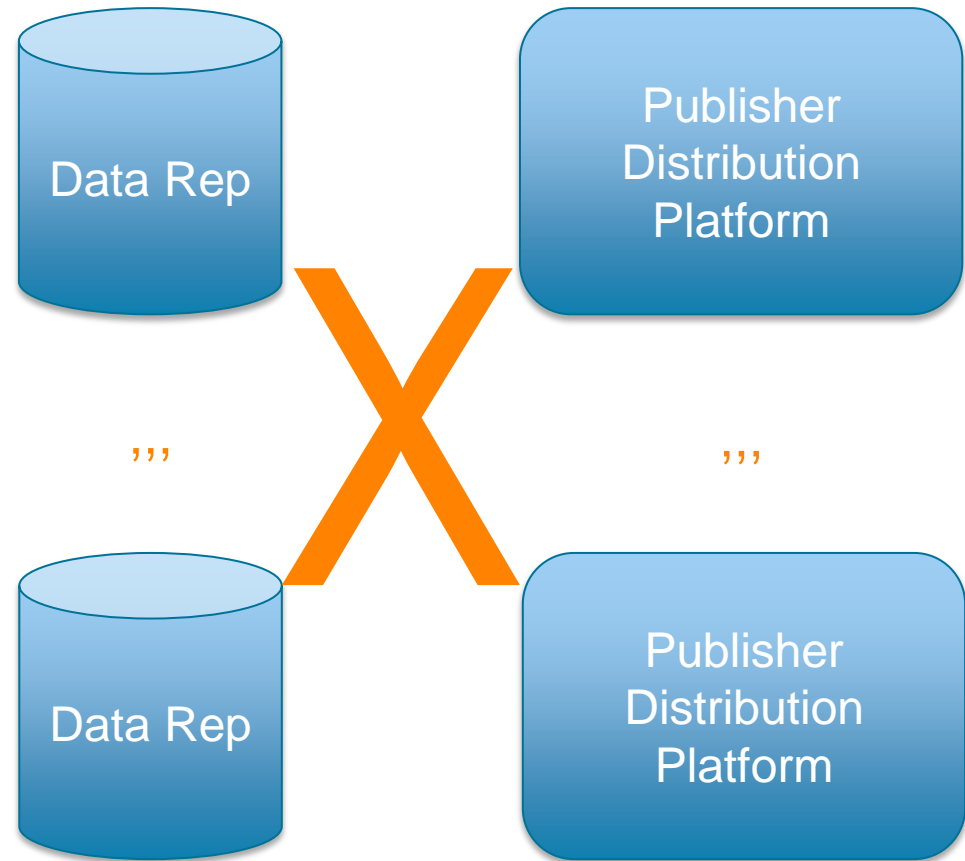


<http://www.elsevier.com/databaselinking>

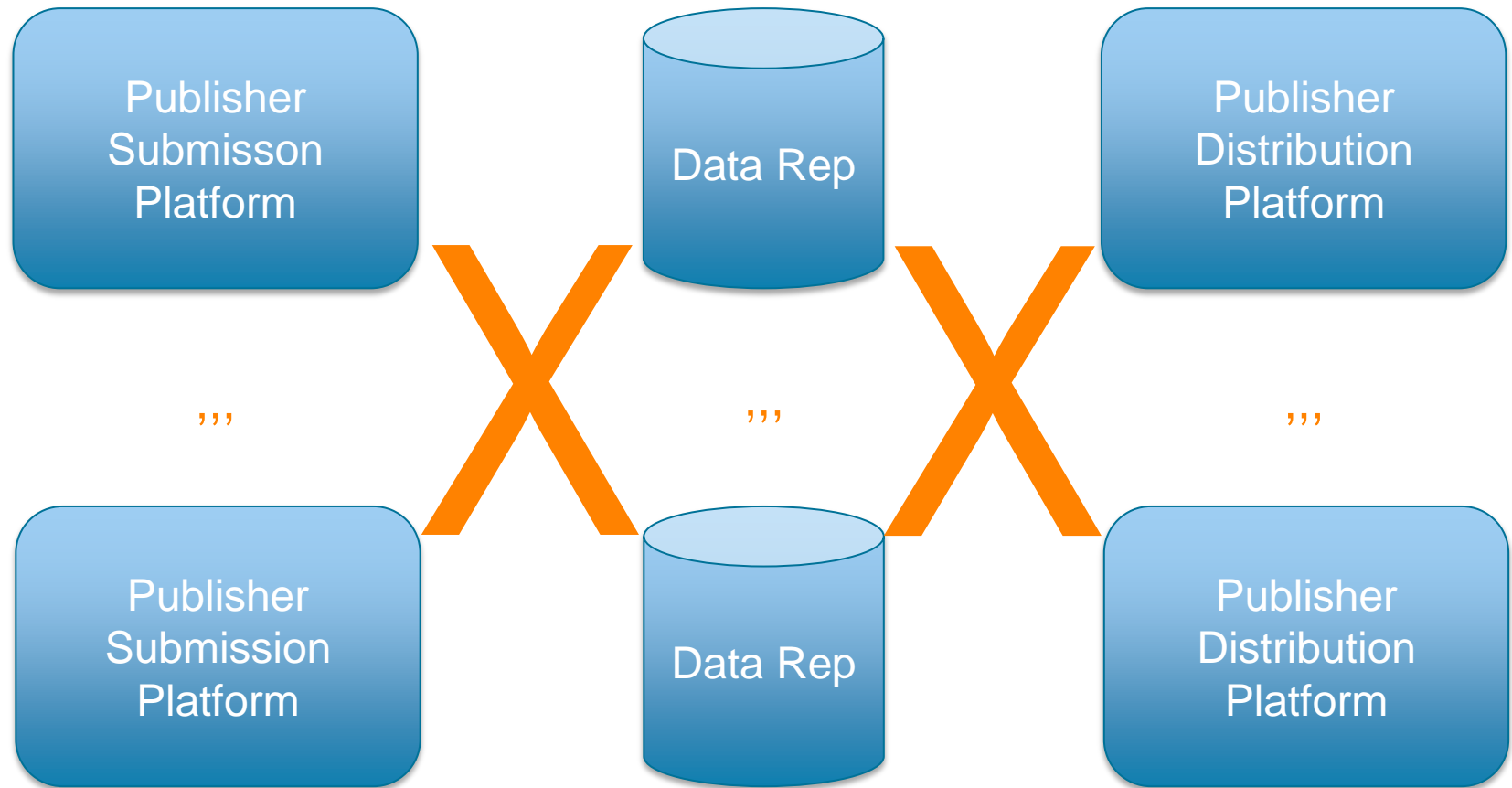
Current linking approach isn't scalable; an increase in repositories requires standards



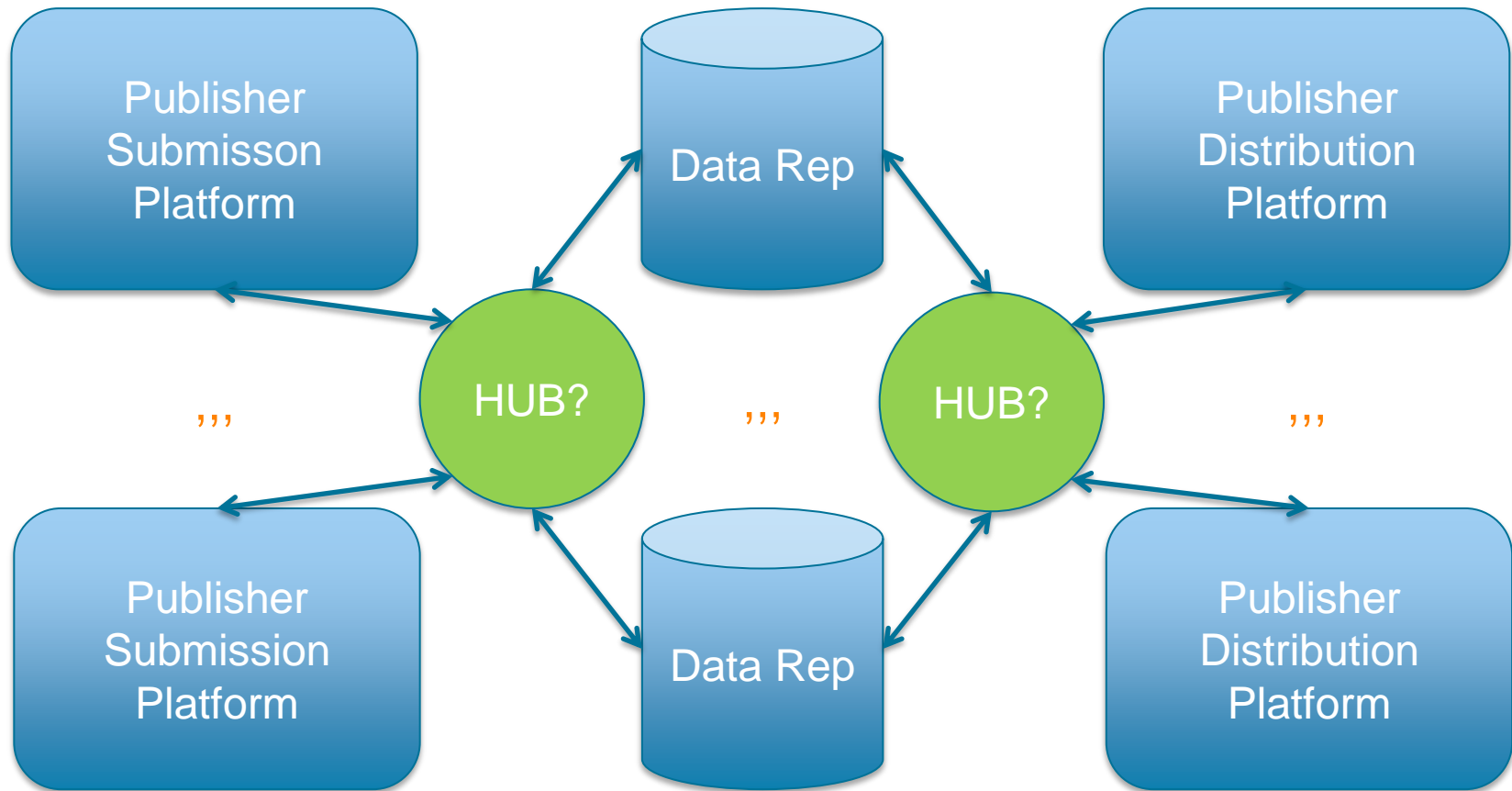
Current linking approach isn't scalable; an increase in repositories requires standards



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



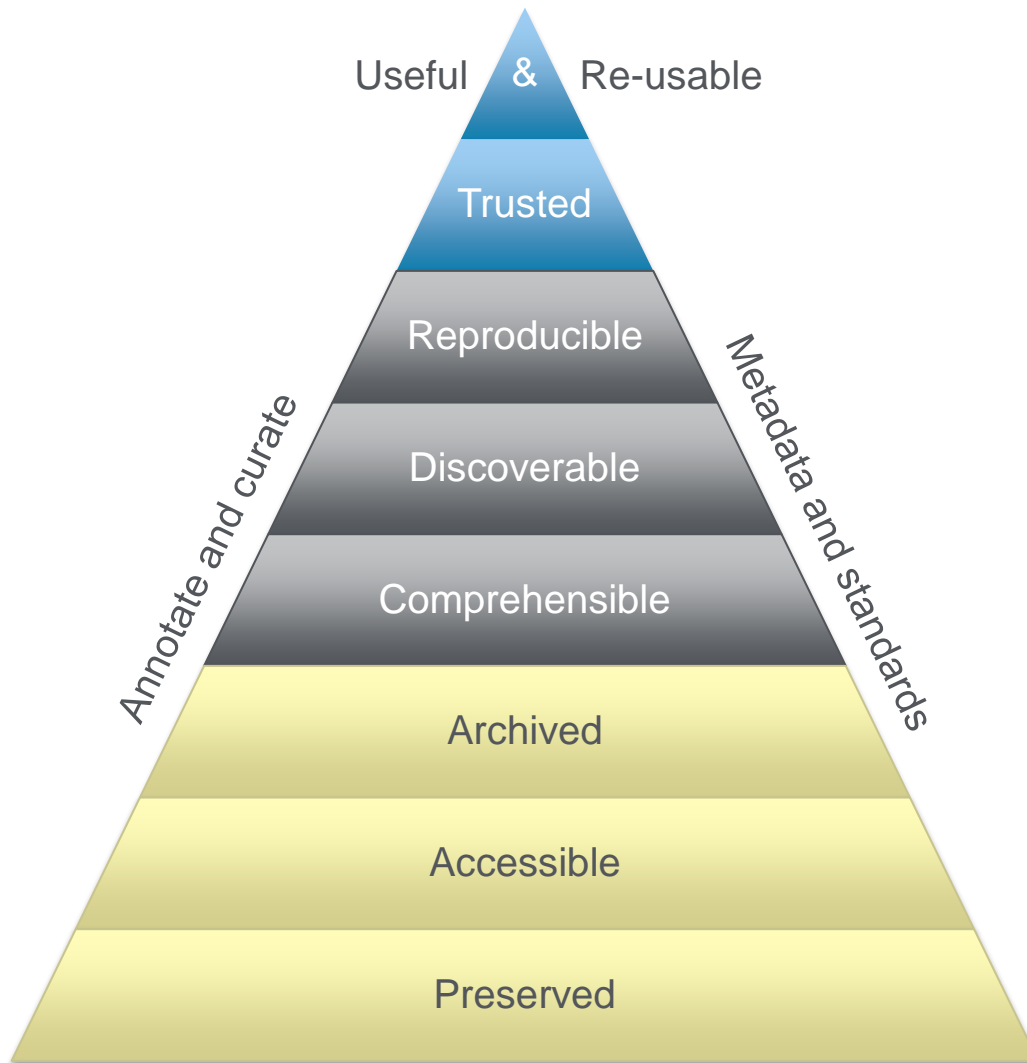
Looking at Researcher Data Workflow shows that also other standards are highly needed



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:

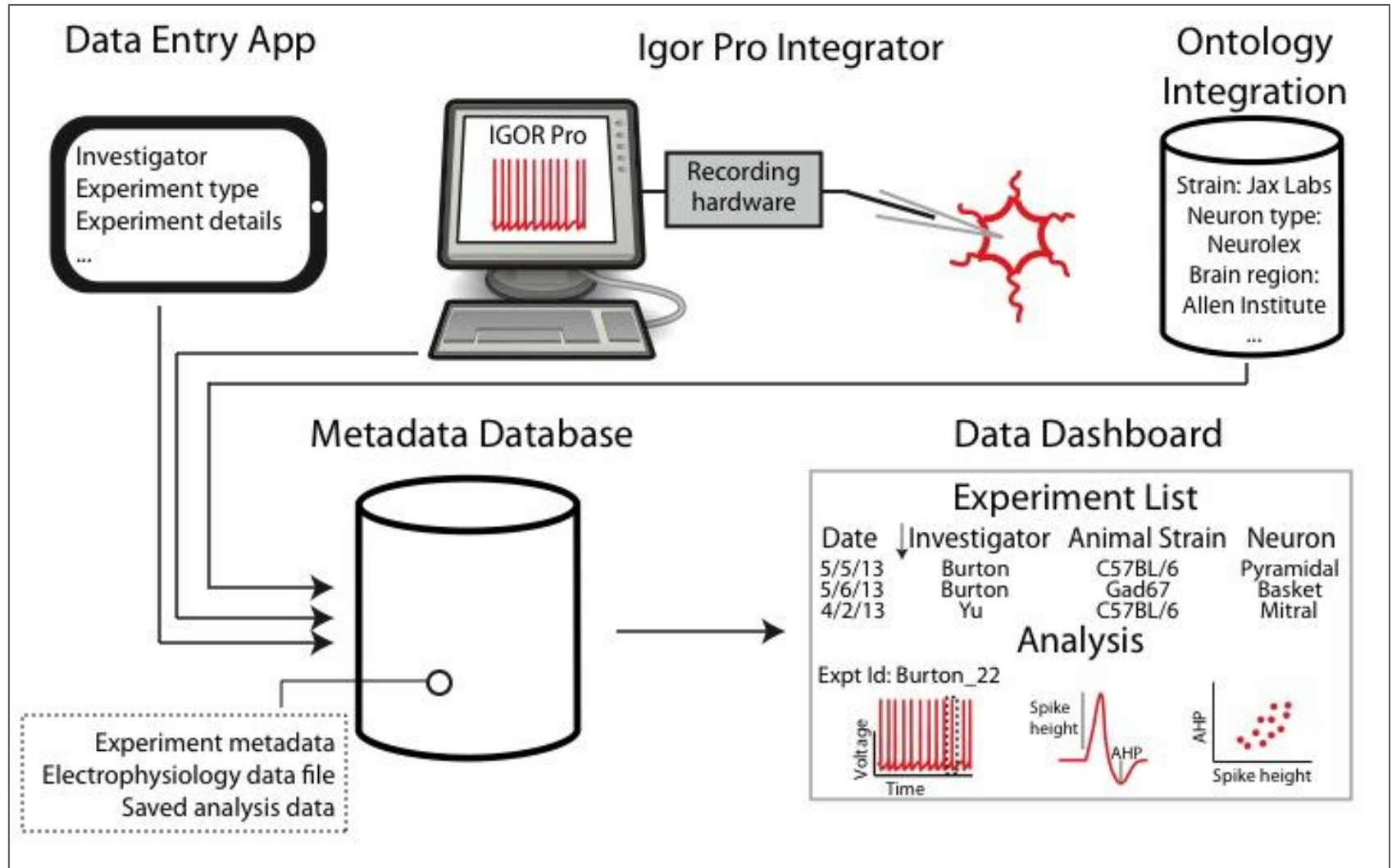
- Executable papers
- Microarticles
- Data articles
- Data linking
- Data integration
- Supplem. files
- Standard Cmt
- Pilot projects:
 - Urban Legend
 - Moonrocks

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









Shawn Burton CMU Urban Labs 

Unspecified Oct 28, 2013 08:11

No Description

mouse (C57BL/6)
unspecified Gender
N/A days
No Drugs
No Anesthetic
Food:
high-fat/unspecified
N/A
olfactory bulb

[← Edit](#)

Slice 1    Slice 2    [+ Add another slice](#)

[Details](#)

Harvesting: Orientation: Thickness (µm):

[Dissection Solution](#)
[Incubation Solution](#)
[Electrodes](#)
[Cells](#)
[Runs](#)

[Sync](#) [Verbose Printout](#)

Urban Legend – Data Dashboard



CMU Urban Labs - Sweep Dashboard

EXPERIMENTS COLLECTIONS CONTENTS Shawn Burton > Jan 8 2014 - Jan 14 2014 > Accessory Olfactory Bulb neuron physiology: Jan 13, 2014 18:49 > Cell 3 > Run 002

Search

ADD TO A COLLECTION INSPECT

Shawn Burton

- Jan 8 2014 - Jan 14 2014
 - Accessory Olfactor - Jan 13, 2014 18:49
 - Cell 1
 - Cell 2
 - Cell 3
 - Run 001
 - Run 002
 - Run 003
 - IV curve of MTC-PGC - Jan 9, 2014 20:19
 - Cell 1
 - Cell 2
 - Jan 1 2014 - Jan 7 2014
 - Dec 25 2013 - Dec 31 2013
 - Dec 18 2013 - Dec 24 2013
 - Dec 11 2013 - Dec 17 2013
 - Dec 4 2013 - Dec 10 2013
 - Nov 27 2013 - Dec 3 2013

FILTER EXPERIMENT

Investigator

- Shawn Burton
- Yiyi Yu
- Adam Large
- Santosh Chand

Date Created

- Shawn and Yiyi Olfactory...
- Untitled 3

Experiment Type

- Unspecified (36)
- Olfactory bulb neuron physiology (136)
- NeuroElectro data validation (94)
- TC-dSAC synaptic connectivity (112)
- Test App Usage (8)
- Spatial GC input (187)
- IV curve of MTC-PGC connection (148)
- Feedforward granule cell inputs (18)

Species

Save filter settings

Untitled 4 Save

Untitled 1

Untitled 2

Untitled 3

SDB spat GC ACSF, 1 mM Ca...

Threshold (20mV/msec)

- 0-10 mV (2)
- 11-20 mV (3)
- 21-30 mV (5)
- 31-40 mV (2)

Height

FWHM

- 4.1-5 ms (1)
- 5.1-6 ms (3)
- 6.1-7 ms (5)
- 7.1-8 ms (2)
- 8.1-9 ms (1)

Rise-time

Fall-time

- 0-10 ms (1)
- 11-20 ms (8)
- 21-30 ms (2)

METADATA PLOT CALCULATED PROPERTIES

FWHM

Threshold

X: Threshold

Y: FWHM

More Properties

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



Hector Borja Moon Rocks

Dataset title Jan 24, 2014 11:58

Dataset title
Dataset Type
Authors
Truncated description
????

Table of Sample Data

Manually add samples and parameters Upload spreadsheet

Edit

- Describe Methods
- Describe Data Quality
- Describe Sample(s)
- Describe Person(s)
- Describe Location/Site
- Describe Time of Measure
- Describe Instrument(s)

Urban Legend – Data Dashboard



Hector Borja Moon Rocks

Dataset title Jan 24, 2014 11:58

Dataset title
Dataset Type
Authors
Truncated descript
???

< Edit

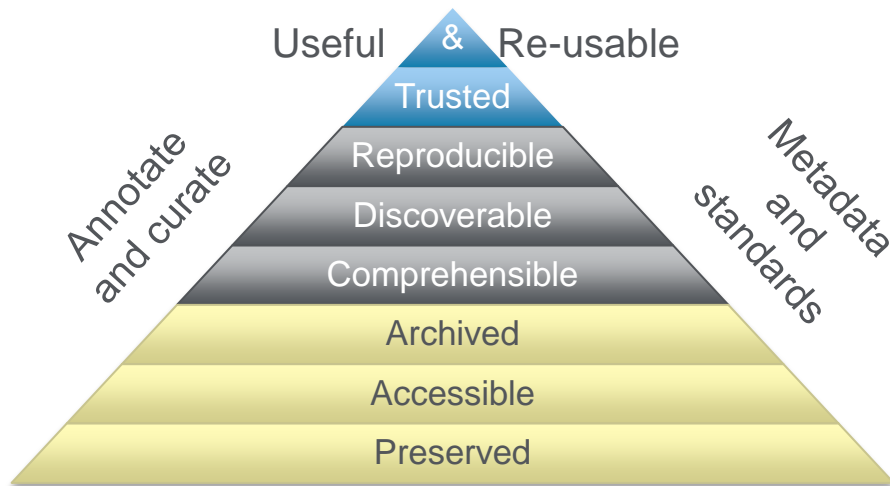
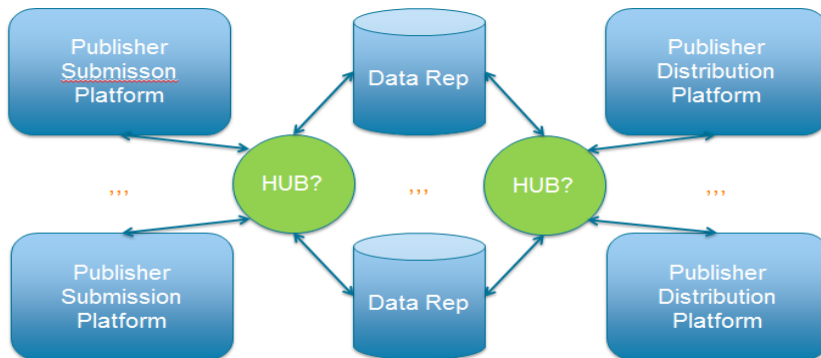
fileuploaded.xls

The following parameters have been mapped to existing parameters. Please indicate if these are correct:

	Use	Don't use
MnO 2 has been substituted with Manganese(IV) oxide	<input checked="" type="radio"/>	<input type="radio"/>
Silica has been substituted with SiO2 (Silicon dioxide)	<input checked="" type="radio"/>	<input type="radio"/>
KO2 has been substituted with K ₂ O	<input checked="" type="radio"/>	<input type="radio"/>
Cr2O3 has been substituted with Chromium(III) oxide	<input type="radio"/>	<input checked="" type="radio"/>
A12O3 has been substituted with Sapphire Crystal (Al2O3)	<input checked="" type="radio"/>	<input type="radio"/>
Dopside has been substituted with Diopside	<input checked="" type="radio"/>	<input type="radio"/>
MnO 2 has been substituted with Manganese(IV) oxide	<input checked="" type="radio"/>	<input type="radio"/>

Cancel Next >

- 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