Embracing research data

Lucie Boudova, Elsevier

- 6 December 2016

Empowering Knowledge

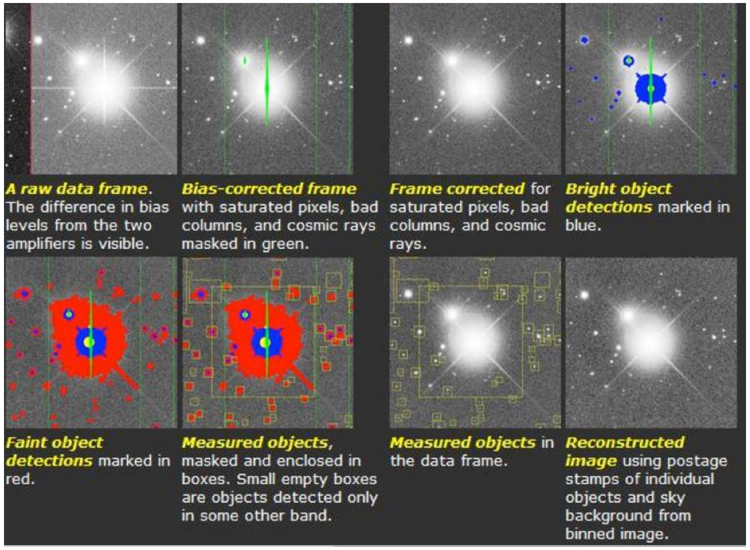
Embracing research data

- **Impact of research data sharing**
- **Fundamentals of research data**
- **Components of effective research data**
- **Tools and programmes supporting research data**
    - Linking-data programme
    - Industry standards
    - Data search
    - Research protocols (HiveBench)
    - Data repository (Mendeley Data)
    - Data journals
- **Research Data Policy**

What are we really after: astronomy

Extracts from "the top 10 benefits of data sharing in astronomy", from Sloan Digital Sky Survey:

- **Early data releases greatly improve the final product, e.g. more people "looking" at the data increases the chance of finding subtle problems, especially important for space missions with finite lifetime, e.g. the ESA's Gaia mission**
- **More science is extracted from the same dataset, e.g. diversity of ideas: many of the most visible SDSS results were unanticipated in the original project proposal**
- **Sometimes the only way to secure scarce resources,** "easy things" (e.g. those that can be put together by a small number of groups/institutions) have been done in the last century; the "road ahead" requires more substantial merging of research resources, like HST Deep Field, UKIDSS, LSST
- **Results in more citations and prestige to the team who produced data;** practically all postdocs from the first phase of SDSS hold faculty-level positions today



A raw data frame. The difference in bias levels from the two amplifiers is visible.

Bias-corrected frame with saturated pixels, bad columns, and cosmic rays masked in green.

Frame corrected for saturated pixels, bad columns, and cosmic rays.

Bright object detections marked in blue.

Faint object detections marked in red.

Measured objects, masked and enclosed in boxes. Small empty boxes are objects detected only in some other band.

Measured objects in the data frame.

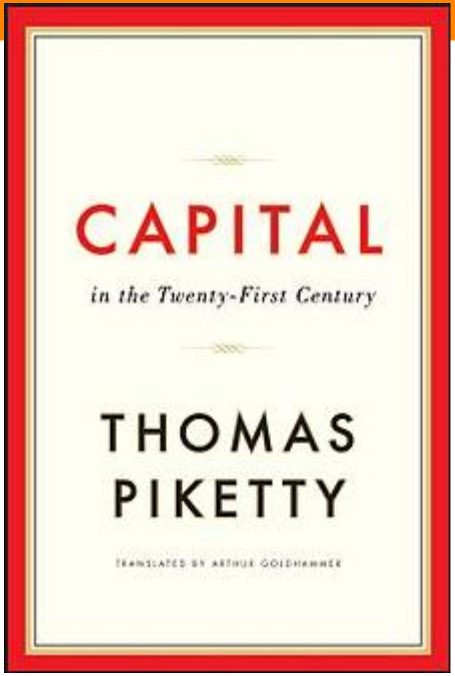
Reconstructed image using postage stamps of individual objects and sky background from binned image.

Željko Ivezić, Department of Astronomy, University of Washington - The Sloan Digital Sky Survey Telescope - Apache Point Observatory, NM
With contributions from: Andy Connolly, Bob Hanisch, David Hogg, Mario Jurić, Andy Lawrence,
Robert Lupton, Mathias Steinmetz, Michael Strauss, Alex Szalay, Tony Tyson, Roy Williams

What are we really after: social sciences

Capital in the Twenty-First Century is a 2013 book by French economist Thomas Piketty.

- **It focuses on wealth and income inequality in Europe and the United States since the 18th century**
- **Central thesis is that when the rate of return on capital (r) is greater than the rate of economic growth (g) over the long term, the result is concentration of wealth, and this unequal distribution of wealth causes social and economic instability**
- **All raw data, normalized data, the analysis, and methods have all been made publicly available on a dedicated website**



CAPITAL

in the Twenty-First Century

THOMAS PIKETTY

TRANSLATED BY ARTHUR GOLDHAMMER

“Here are enormous quantities of information distilled from tax rolls, inheritance records, and various other public data sources, laid out in charts that should be readily accessible to the layest of lay readers. Not all of the information in these sections is novel or startling. Having it together in one place, however, is valuable, and even most of the book’s fiercest critics respect this achievement.” [1]

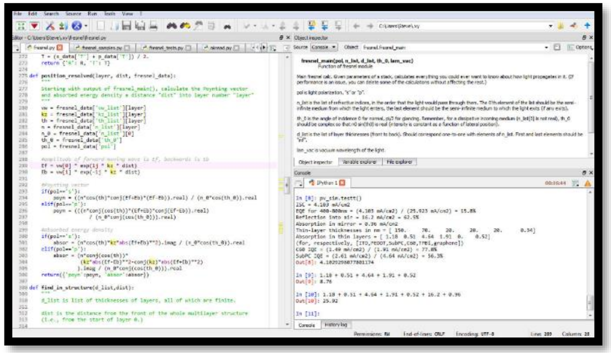
It also shows data sharing can lead to issues [2]:

- Chris Giles, economics editor of the Financial Times (FT), identified what he claims are "unexplained errors" in Piketty's data, in particular regarding wealth inequality increases since the 1970s. “contain a series of errors that skew his findings”
- Subsequently, Piketty wrote a response defending his findings; the accusation and responses received wide press coverage
- E.g. Scott Winship, a sociologist at the MIPR, claims the allegations are not "significant for the fundamental question of whether Piketty's thesis is right or not"

When we talk about data, we really talk about the following:

Machine & environment settings

Scripts & analyses



```
    T = (s_data['T'] + p_data['T']) / 2.
    return {'R': R, 'T': T}

def position_resolved(layer, dist, fresnel_data):
    """
    Starting with output of fresnel_main(), calculate the Poynting vector
    and absorbed energy density a distance "dist" into layer number "layer"
    """
    vw = fresnel_data['vw_list'][layer]
    kz = fresnel_data['kz_list'][layer]
    th = fresnel_data['th_list'][layer]
    n = fresnel_data['n_list'][layer]
    n_0 = fresnel_data['n_list'][0]
    th_0 = fresnel_data['th_0']
    pol = fresnel_data['pol']

    #amplitude of forward-moving wave is Ef, backwards is Eb
    Ef = vw[0] * exp(1j * kz * dist)
    Eb = vw[1] * exp(-1j * kz * dist)

    #Poynting vector
    if(pol=='s'):
        poyn = ((n*cos(th)*conj(Ef+Eb)*(Ef-Eb)).real) / (n_0*cos(th_0)).real
    elif(pol=='p'):
        poyn = (((n*conj(cos(th))*(Ef+Eb)*conj(Ef-Eb)).real)
                    / (n_0*conj(cos(th_0))).real)

    #Absorbed energy density
    if(pol=='s'):
        absor = (n*cos(th)*kz*abs(Ef+Eb)**2).imag / (n_0*cos(th_0)).real
    elif(pol=='p'):
        absor = (n*conj(cos(th))*
                   (kz*abs(Ef-Eb)**2-conj(kz)*abs(Ef+Eb)**2)
                  ).imag / (n_0*conj(cos(th_0))).real
    return({'poyn':poyn, 'absor':absor})

def find_in_structure(d_list,dist):
    """
    d_list is list of thicknesses of layers, all of which are finite.

    dist is the distance from the front of the whole multilayer structure
    (i.e., from the start of layer 0.)
```

fresnel_main(pol, n_list, d_list, th_0, lam_vac)
Function of fresnel module

Main fresnel calc. Given parameters of a stack, calculates everything you could ever want to know about how light propagates in it. (If performance is an issue, you can delete some of the calculations without affecting the rest.)

pol is light polarization, "s" or "p".

n_list is the list of refractive indices, in the order that the light would pass through them. The 0'th element of the list should be the semi-infinite medium from which the light enters, the last element should be the semi-infinite medium to which the light exits (if any exists).

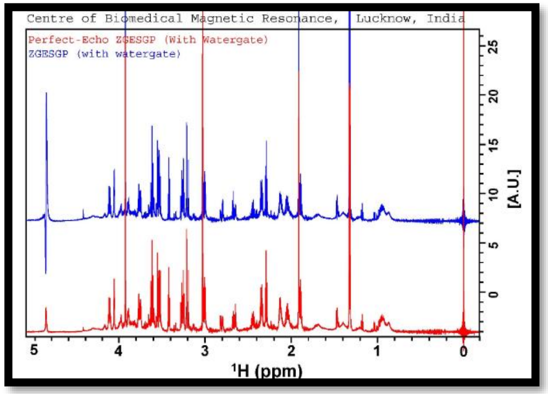
th_0 is the angle of incidence 0 for normal, pi/2 for glancing. Remember, for a dissipative incoming medium (n_list[0] is not real), th_0 should be complex so that n0 sin(th0) is real (intensity is constant as a function of lateral position).

d_list is the list of layer thicknesses (front to back). Should correspond one-to-one with elements of n_list. First and last elements should be "inf".

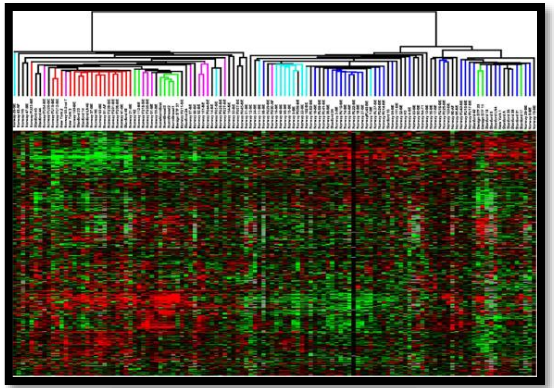
lam_vac is vacuum wavelength of the light.

```
In [8]: pv_sim.testt()
ISC = 4.103 mA/cm2
EQE for 400-800nm = (4.103 mA/cm2) / (25.923 mA/cm2) = 15.8%
Reflection into air = 16.2 mA/cm2 = 62.5%
Absorption in mirror = 0.96 mA/cm2
Thin-layer thicknesses in nm = [ 150.    70.    20.    20.    20.     0.34]
Absorption in thin layers = [ 1.18  0.51  4.64  1.91  0.    0.52]
(for, respectively, [ITO,PEDOT,SubPC,C60,TPBi,graphene])
C60 IQE = (1.49 mA/cm2) / (1.91 mA/cm2) = 77.8%
SubPC IQE = (2.61 mA/cm2) / (4.64 mA/cm2) = 56.3%
Out[8]: 4.1029298077801174

In [9]: 1.18 + 0.51 + 4.64 + 1.91 + 0.52
Out[9]: 8.76

In [10]: 1.18 + 0.51 + 4.64 + 1.91 + 0.52 + 16.2 + 0.96
Out[10]: 25.92

In [11]:
```

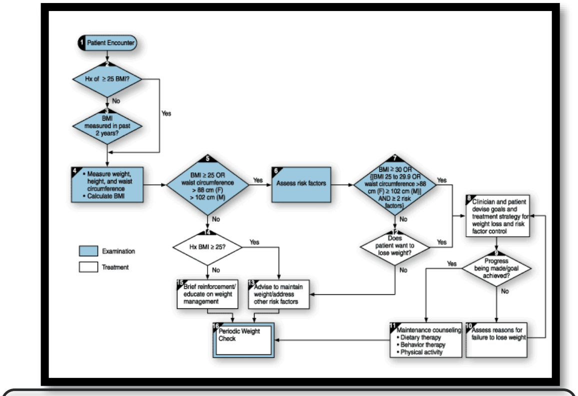
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 289 Columns: 28



Raw data



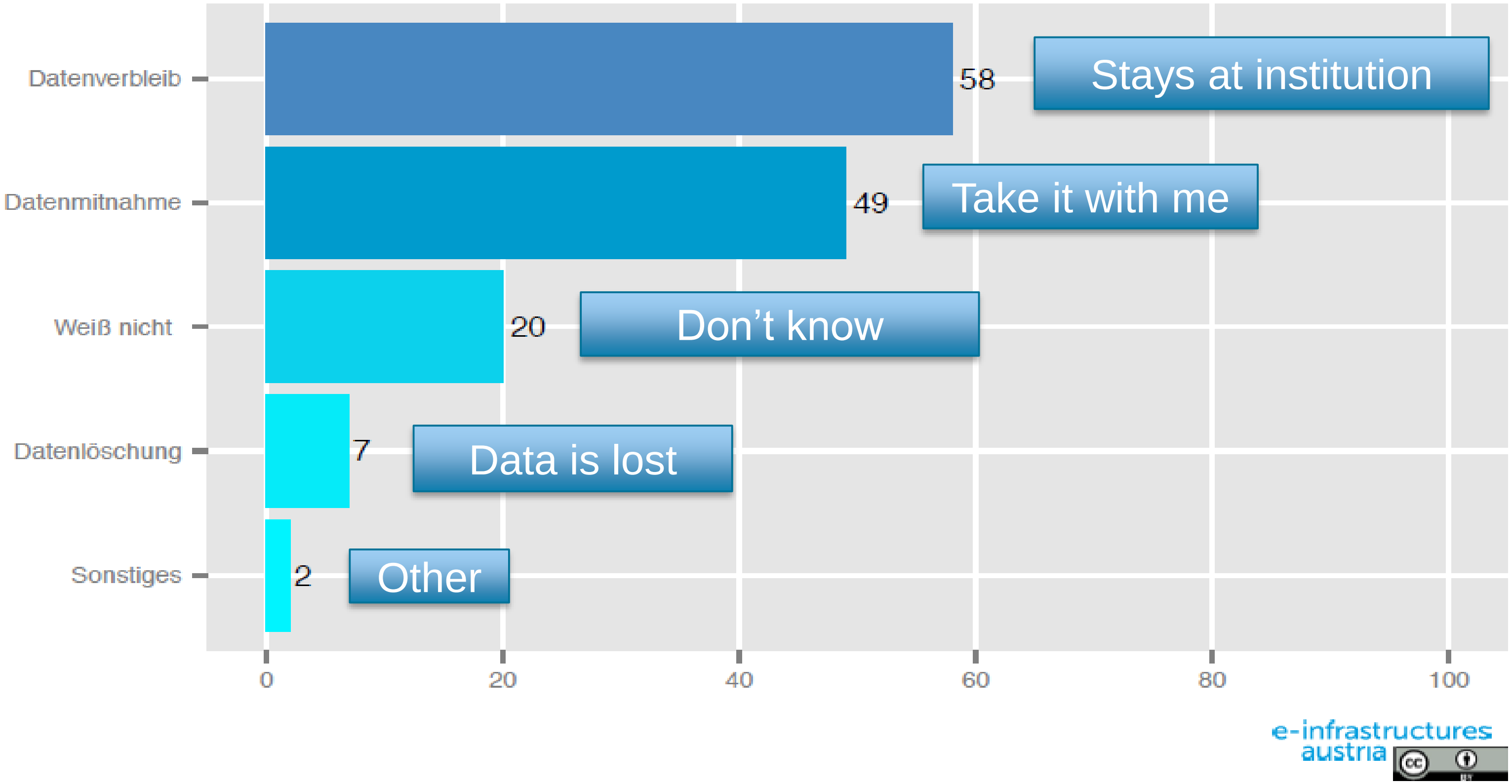
Processed data


1 Patient Encounter
2 Hx of ≥ 25 BMI?
No
Yes
3 BMI measured in past 2 years?
4 • Measure weight, height, and waist circumference
• Calculate BMI
5 BMI ≥ 25 OR waist circumference > 88 cm (F) > 102 cm (M)
Yes
No
6 Assess risk factors
7 BMI ≥ 30 OR [[BMI 25 to 29.9 OR waist circumference >88 cm (F) ≥ 102 cm (M)] AND ≥ 2 risk factors]
Yes
No
8 Clinician and patient devise goals and treatment strategy for weight loss and risk factor control
9 Progress being made/goal achieved?
Yes
No
10 Assess reasons for failure to lose weight
11 Maintenance counseling:
• Dietary therapy
• Behavior therapy
• Physical activity
12 Does patient want to lose weight?
Yes
No
13 Advise to maintain weight/address other risk factors
14 Hx BMI ≥ 25?
Yes
No
15 Brief reinforcement/ educate on weight management
16 Periodic Weight Check
Examination
Treatment

Protocols, methods, algorithms

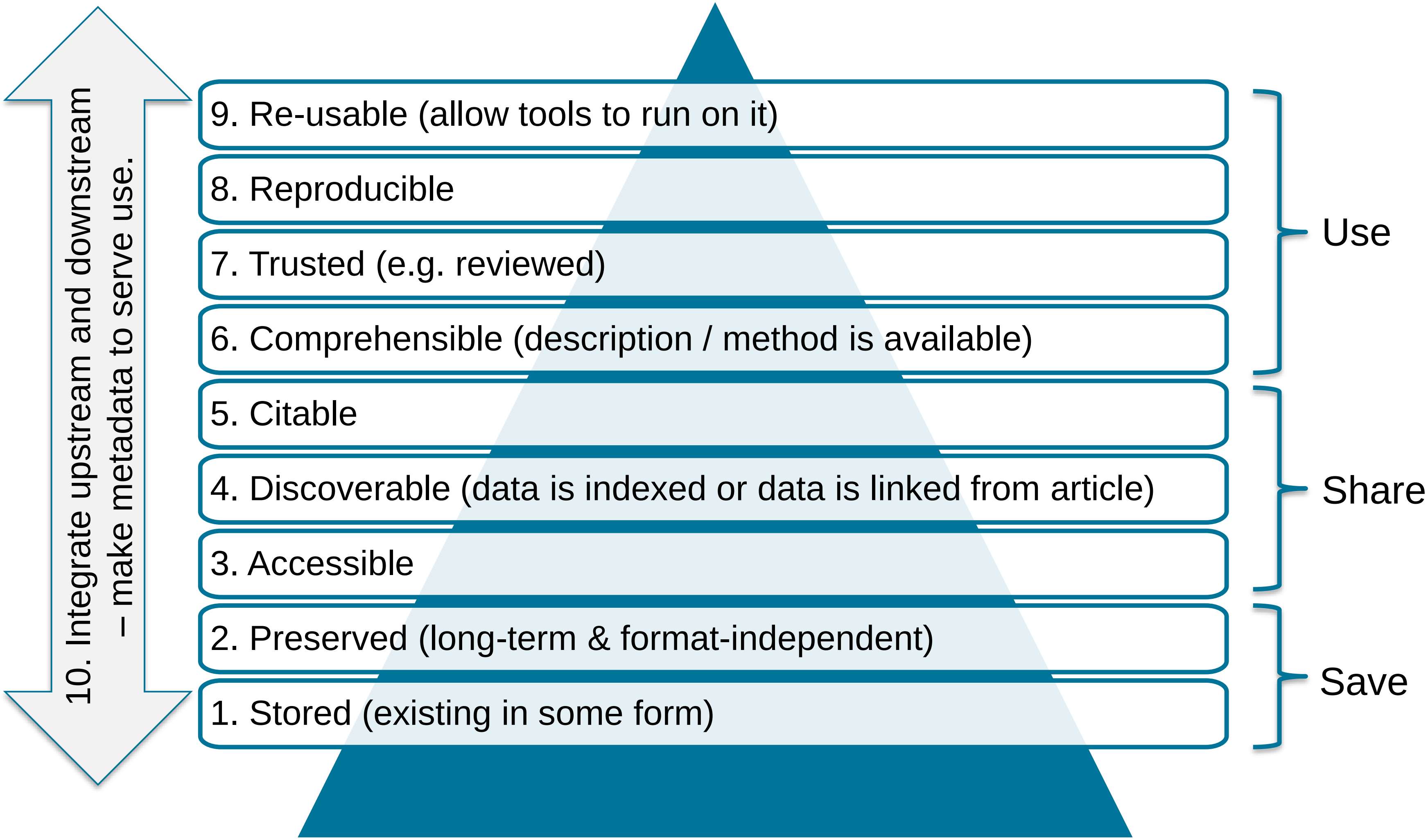
Note: images for illustrative purpose only

When You Leave Your Institution, What Happens To Your Data?



„Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung (eBook)“
E-infrastructures Austria
Bauer, B. (Bruno) et all
Oct 2015
https://phaidra.univie.ac.at/detail_object/o:407736

The 10 components for effective research data



10. Integrate upstream and downstream – make metadata to serve use.

9. Re-usable (allow tools to run on it)

8. Reproducible

7. Trusted (e.g. reviewed)

6. Comprehensible (description / method is available)

5. Citable

4. Discoverable (data is indexed or data is linked from article)

3. Accessible

2. Preserved (long-term & format-independent)

1. Stored (existing in some form)

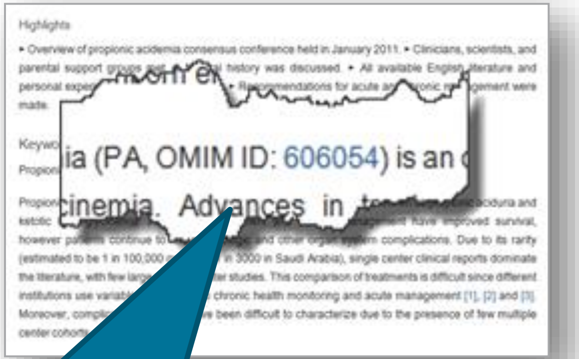
Tools and programmes supporting research data

ELSEVIER

Empowering Knowledge

Data-linking programme

- Elsevier has an extensive programme with 60+ leading domain-specific data repositories to interlink articles and data
- Makes it easier to find relevant data and place data into the right context
- Linking through in-article accession numbers, data DOI's, or data banners



Highlights

• Overview of propionic acidemia consensus conference held in January 2011. • Clinicians, scientists, and parental support groups ... history was discussed. • All available English literature and personal exper... • Recommendations for acute and chronic management were made.

Keywo...

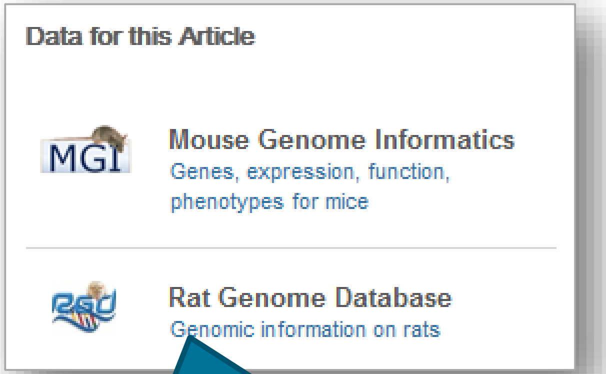
Propion...

Propion... aciduria and ketotic ... have improved survival, however patients continue to ... and other organ system complications. Due to its rarity (estimated to be 1 in 100,000 ... in 3000 in Saudi Arabia), single center clinical reports dominate the literature, with few large ... studies. This comparison of treatments is difficult since different institutions use variable ... chronic health monitoring and acute management [1], [2] and [3]. Moreover, complic... been difficult to characterize due to the presence of few multiple center cohorts.

ia (PA, OMIM ID: 606054) is an

cinemia. Advances in

Linking through **in-article data accession numbers**



Data for this Article

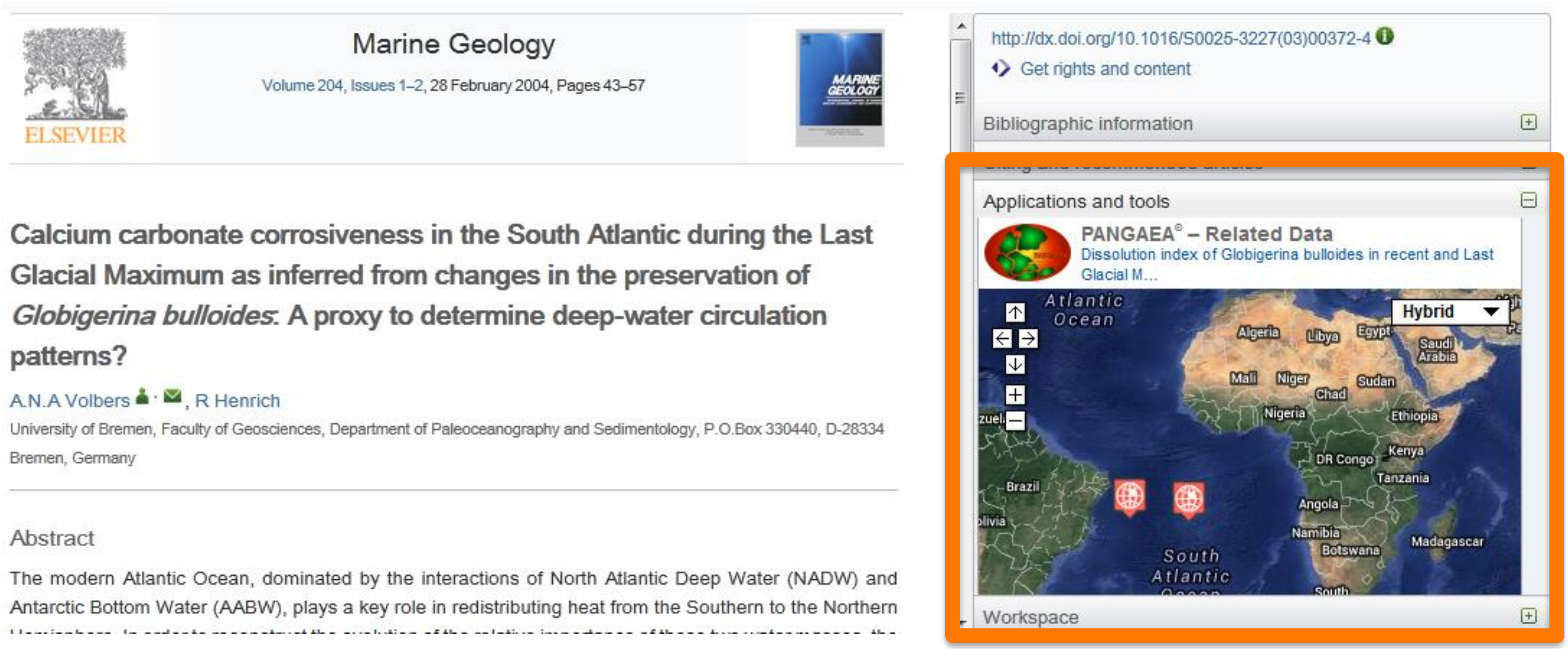
MGI Mouse Genome Informatics
Genes, expression, function, phenotypes for mice

RGD Rat Genome Database
Genomic information on rats

Database banners shown next to the article on ScienceDirect

See http://www.elsevier.com/databaselinking

Data-linking programme – example Pangaea



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, R Henrich

University of Bremen, Faculty of Geosciences, Department of Paleoceanography and Sedimentology, P.O.Box 330440, D-28334 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

http://dx.doi.org/10.1016/S0025-3227(03)00372-4

Get rights and content

Bibliographic information

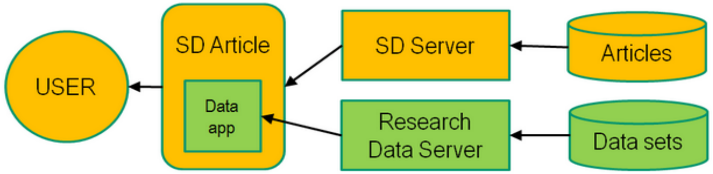
Applications and tools

PANGAEA® – Related Data

Dissolution index of Globigerina bulloides in recent and Last Glacial M...

Workspace

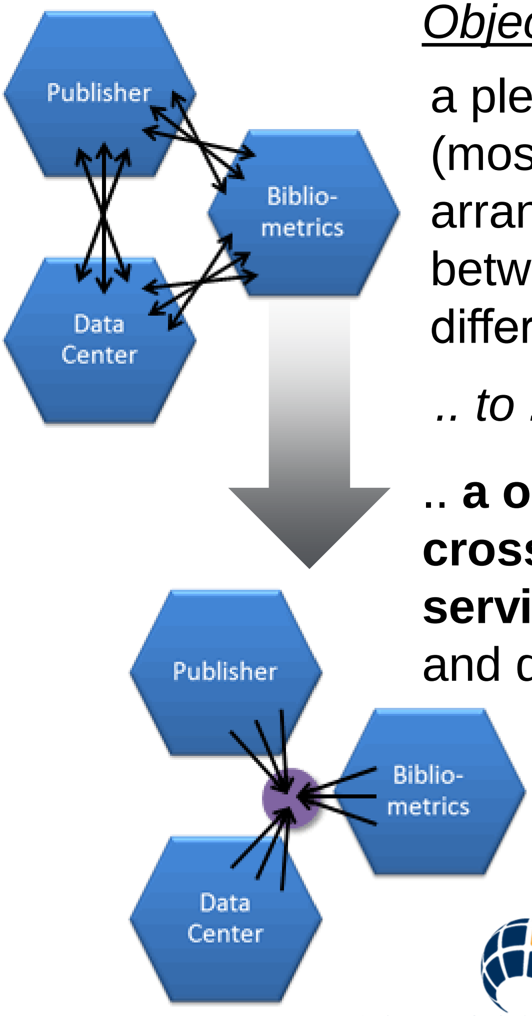
- Supplementary data at PANGAEA
- Bidirectional links between PANGAEA & ScienceDirect
- Data visualized next to the article



Research Data Working Groups and Development of Industry Standards example www.Scholix.org

SCHOLIX

- ICSU/WDS/RDA Publishing Data Service Working group
- Creating linked-data model for exposing DOI to DOI links outside publisher's firewall
- Collaboration between CrossRef, DataCite, Europe PubMed Central, ANDS, Thompson Reuters, Elsevier, OpenAire



Objective: move from

a plethora of (mostly) bilateral arrangements between the different players…

.. to ..

.. **a one-for-all cross-referencing service** for articles and data

ICSU WORLD DATA SYSTEM

RDA RESEARCH DATA ALLIANCE

Datasearch engine!

- Many (broad) datasearch examples already available

BASE | BioCaddie/ DataMED | Datacite | Datahub.io | DataONE | EbiSearch | OneRepo

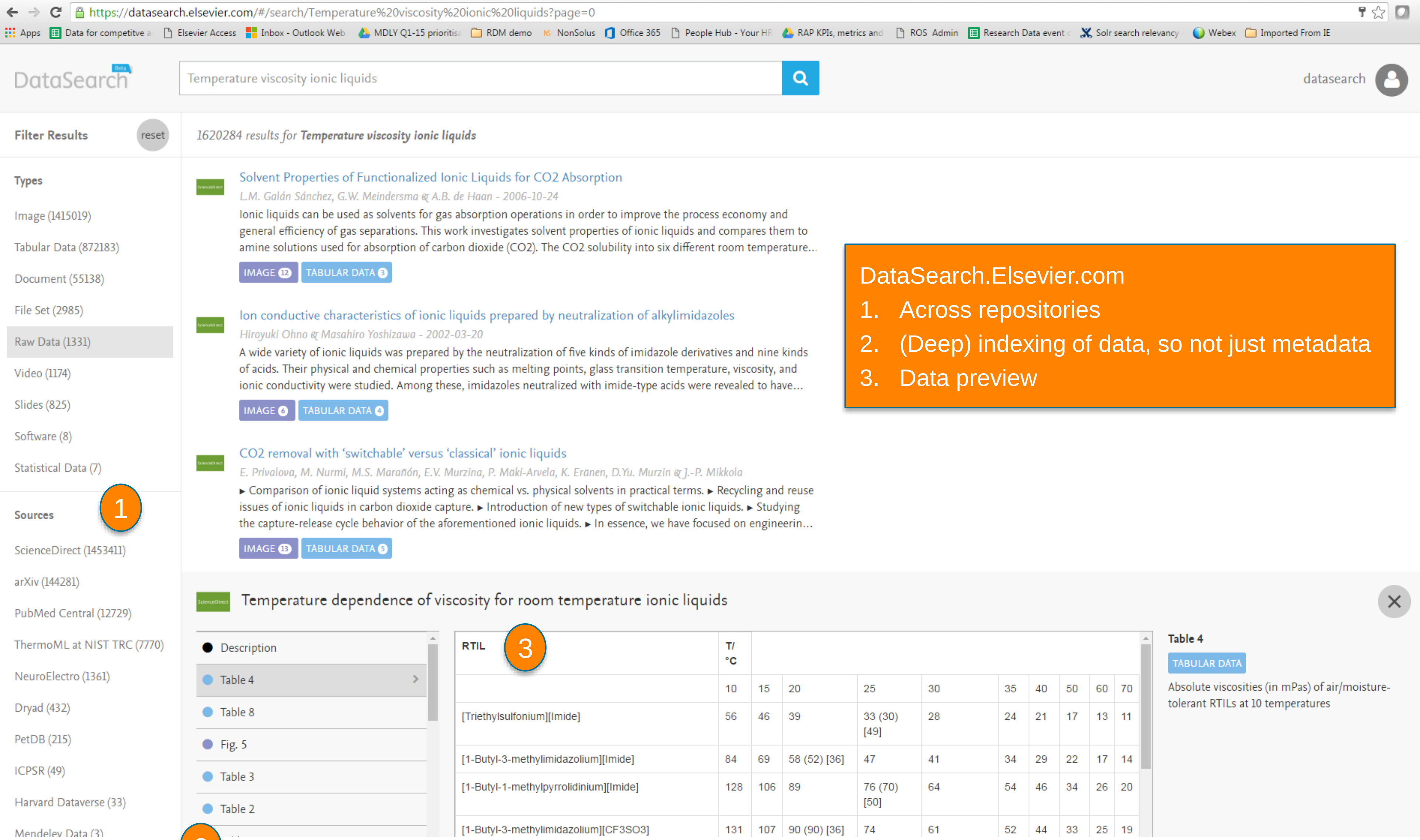
Quandl | RE3Data.org | Semantic Scholar | OSF|SHARE | TR Data Citation Index | Zanran

- Some common themes:
  - search of metadata only (i.e. ranking based on metadata only)
  - And/or federated search (i.e. no ranking)
  - And/or focused on giving credit (citation) rather than on discoverability

- Uncommon (because difficult):
  - Deep indexing of datasets (so real ranking and filtering)
  - Search engine really focused on data discovery

Elsevier Data Search

E.g. search for "Temperature viscosity ionic liquids"



1620284 results for ***Temperature viscosity ionic liquids***

Types

- Image (1415019)
- Tabular Data (872183)
- Document (55138)
- File Set (2985)
- Raw Data (1331)
- Video (1174)
- Slides (825)
- Software (8)
- Statistical Data (7)

Sources

- ScienceDirect (1453411)
- arXiv (144281)
- PubMed Central (12729)
- ThermoML at NIST TRC (7770)
- NeuroElectro (1361)
- Dryad (432)
- PetDB (215)
- ICPSR (49)
- Harvard Dataverse (33)
- Mendeley Data (3)

Solvent Properties of Functionalized Ionic Liquids for CO2 Absorption
L.M. Galán Sánchez, G.W. Meindersma & A.B. de Haan - 2006-10-24
Ionic liquids can be used as solvents for gas absorption operations in order to improve the process economy and general efficiency of gas separations. This work investigates solvent properties of ionic liquids and compares them to amine solutions used for absorption of carbon dioxide (CO2). The CO2 solubility into six different room temperature...
IMAGE 12 TABULAR DATA 3

Ion conductive characteristics of ionic liquids prepared by neutralization of alkylimidazoles
Hiroyuki Ohno & Masahiro Yoshizawa - 2002-03-20
A wide variety of ionic liquids was prepared by the neutralization of five kinds of imidazole derivatives and nine kinds of acids. Their physical and chemical properties such as melting points, glass transition temperature, viscosity, and ionic conductivity were studied. Among these, imidazoles neutralized with imide-type acids were revealed to have...
IMAGE 6 TABULAR DATA 4

CO2 removal with 'switchable' versus 'classical' ionic liquids
E. Privalova, M. Nurmi, M.S. Marañón, E.V. Murzina, P. Maki-Arvela, K. Eranen, D.Yu. Murzin & J.-P. Mikkola
▸ Comparison of ionic liquid systems acting as chemical vs. physical solvents in practical terms. ▸ Recycling and reuse issues of ionic liquids in carbon dioxide capture. ▸ Introduction of new types of switchable ionic liquids. ▸ Studying the capture-release cycle behavior of the aforementioned ionic liquids. ▸ In essence, we have focused on engineerin...
IMAGE 13 TABULAR DATA 5

Temperature dependence of viscosity for room temperature ionic liquids

- ● Description
- ● Table 4
- ● Table 8
- ● Fig. 5
- ● Table 3
- ● Table 2

RTIL	T/ °C								
	10	15	20	25	30	35	40	50	60	70
[Triethylsulfonium][Imide]	56	46	39	33 (30) [49]	28	24	21	17	13	11
[1-Butyl-3-methylimidazolium][Imide]	84	69	58 (52) [36]	47	41	34	29	22	17	14
[1-Butyl-1-methylpyrrolidinium][Imide]	128	106	89	76 (70) [50]	64	54	46	34	26	20
[1-Butyl-3-methylimidazolium][CF3SO3]	131	107	90 (90) [36]	74	61	52	44	33	25	19

Table 4

TABULAR DATA

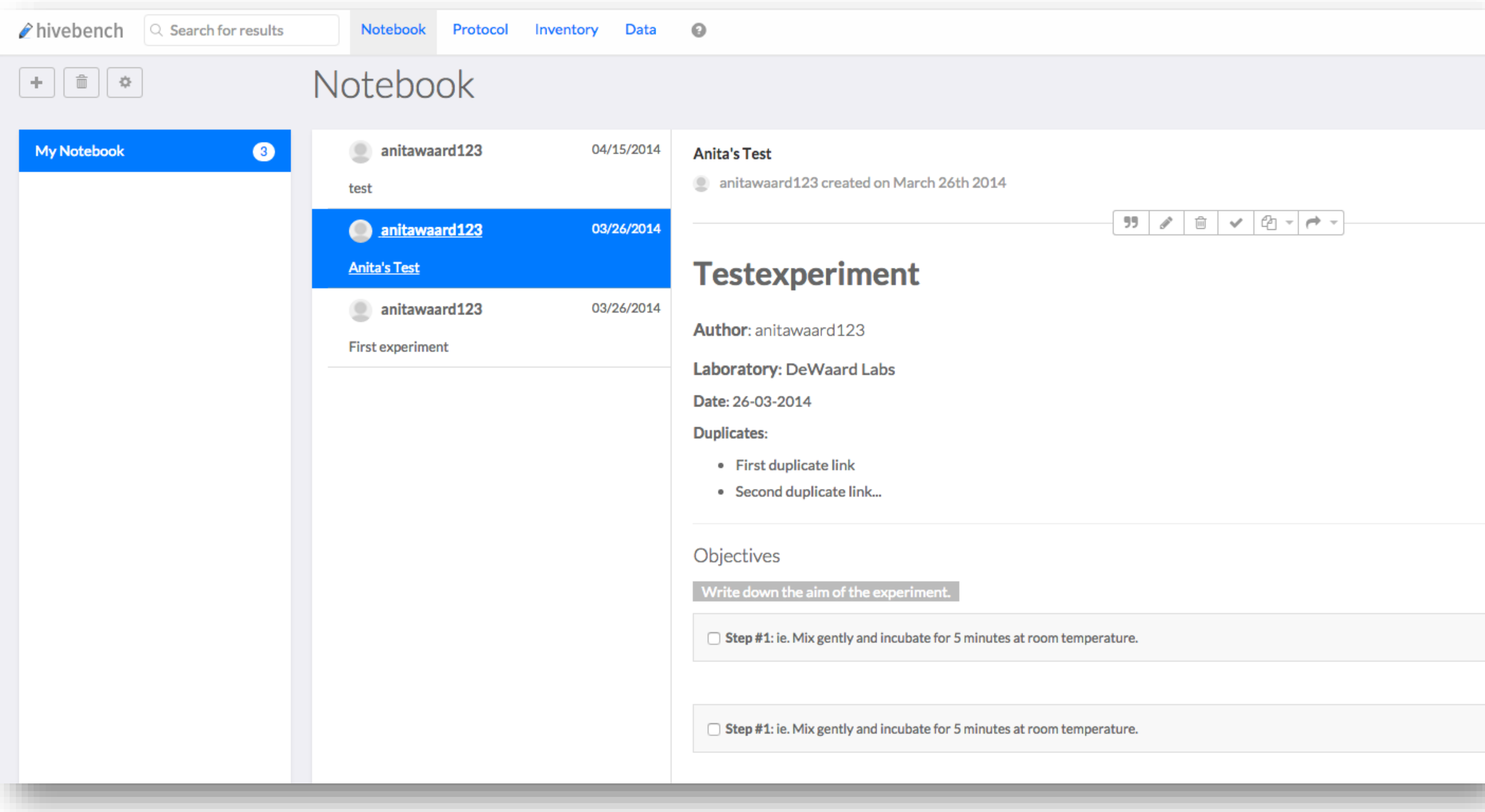
Absolute viscosities (in mPas) of air/moisture-tolerant RTILs at 10 temperatures

DataSearch.Elsevier.com

1. Across repositories
2. (Deep) indexing of data, so not just metadata
3. Data preview

ELSEVIER

Research Protocols – capturing and sharing



Notebook

My Notebook 3

anitawaard123 04/15/2014
test

anitawaard123 03/26/2014
Anita's Test

anitawaard123 03/26/2014
First experiment

Anita's Test

anitawaard123 created on March 26th 2014

Testexperiment

Author: anitawaard123

Laboratory: DeWaard Labs

Date: 26-03-2014

Duplicates:

- First duplicate link
- Second duplicate link...

Objectives

Write down the aim of the experiment.

□ **Step #1**: ie. Mix gently and incubate for 5 minutes at room temperature.

□ **Step #1**: ie. Mix gently and incubate for 5 minutes at room temperature.

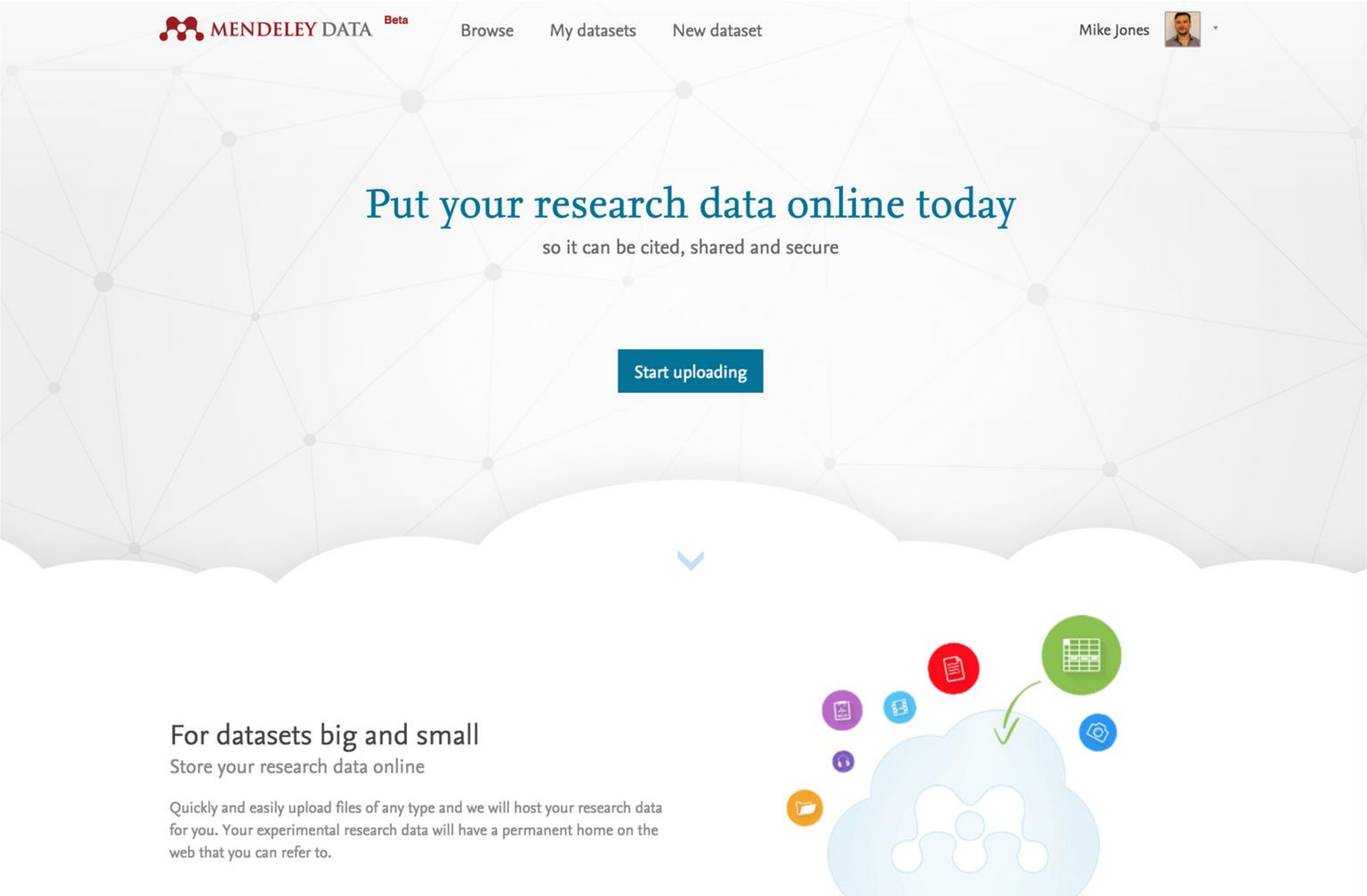
www.hivebench.com

ELSEVIER

http://data.mendeley.com/

Manage, store: Mendeley Data

An open repository for posting & reusing research data



Put your research data online today

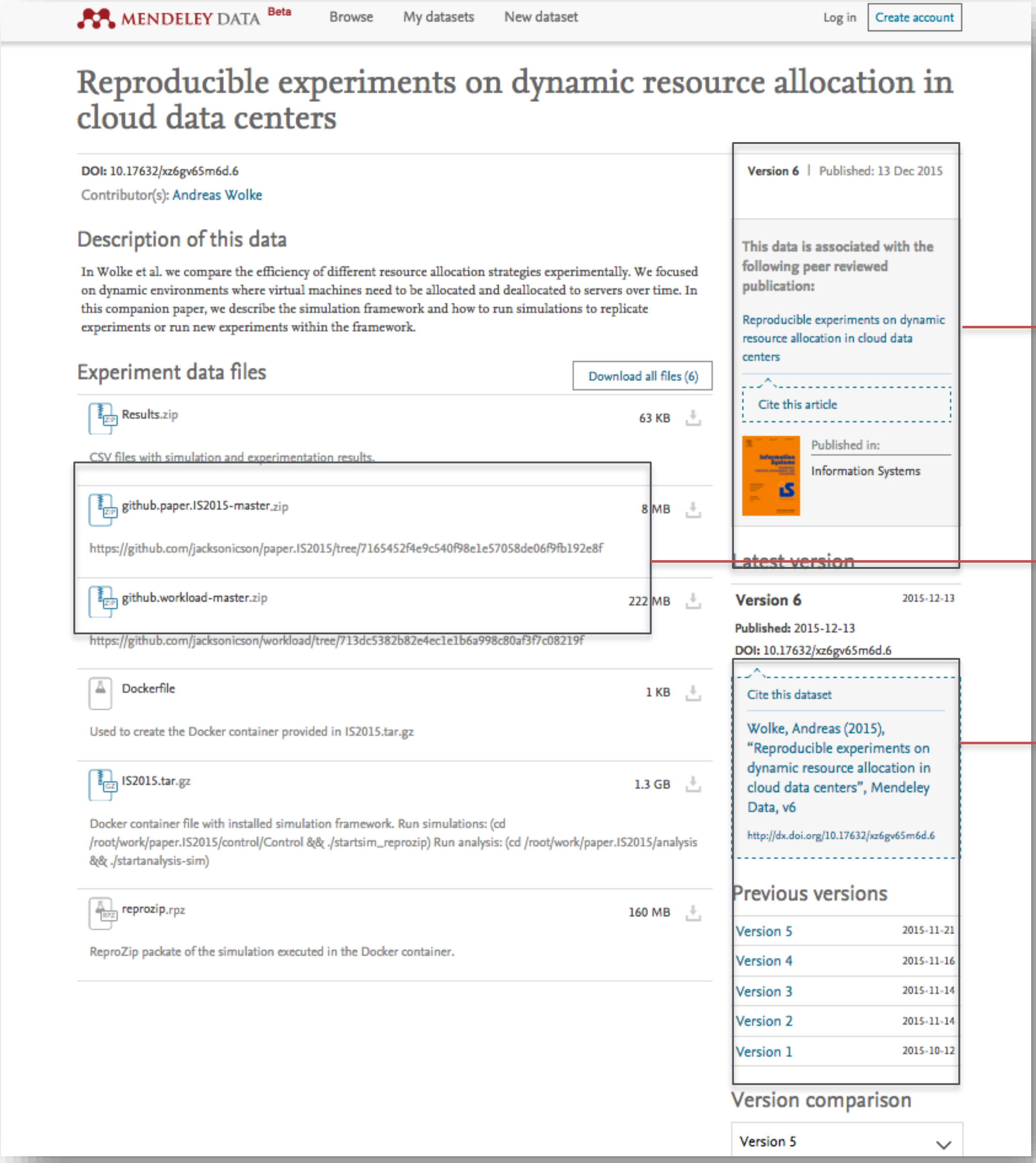
so it can be cited, shared and secure

For datasets big and small

Store your research data online

Quickly and easily upload files of any type and we will host your research data for you. Your experimental research data will have a permanent home on the web that you can refer to.

Manage, Store: Mendeley Data



Reproducible experiments on dynamic resource allocation in cloud data centers

DOI: 10.17632/xz6gv65m6d.6

Contributor(s): Andreas Wolke

Description of this data

In Wolke et al. we compare the efficiency of different resource allocation strategies experimentally. We focused on dynamic environments where virtual machines need to be allocated and deallocated to servers over time. In this companion paper, we describe the simulation framework and how to run simulations to replicate experiments or run new experiments within the framework.

Experiment data files

Results.zip 63 KB

CSV files with simulation and experimentation results.

github.paper.IS2015-master.zip 8 MB

https://github.com/jacksonicson/paper.IS2015/tree/7165452f4e9c540f98e1e57058de06f9fb192e8f

github.workload-master.zip 222 MB

https://github.com/jacksonicson/workload/tree/713dc5382b82e4ec1e1b6a998c80af3f7c08219f

Dockerfile 1 KB

Used to create the Docker container provided in IS2015.tar.gz

IS2015.tar.gz 1.3 GB

Docker container file with installed simulation framework. Run simulations: (cd /root/work/paper.IS2015/control/Control && ./startsim_reprozip) Run analysis: (cd /root/work/paper.IS2015/analysis && ./startanalysis-sim)

reprozip.rpz 160 MB

ReproZip packate of the simulation executed in the Docker container.

Version 6 | Published: 13 Dec 2015

This data is associated with the following peer reviewed publication:

Reproducible experiments on dynamic resource allocation in cloud data centers

Published in:

Information Systems

Latest version

Version 6 2015-12-13

Published: 2015-12-13

DOI: 10.17632/xz6gv65m6d.6

Wolke, Andreas (2015), “Reproducible experiments on dynamic resource allocation in cloud data centers”, Mendeley Data, v6

http://dx.doi.org/10.17632/xz6gv65m6d.6

Previous versions

Version 5	2015-11-21
Version 4	2015-11-16
Version 3	2015-11-14
Version 2	2015-11-14
Version 1	2015-10-12

Version comparison

Linked to published papers – or not

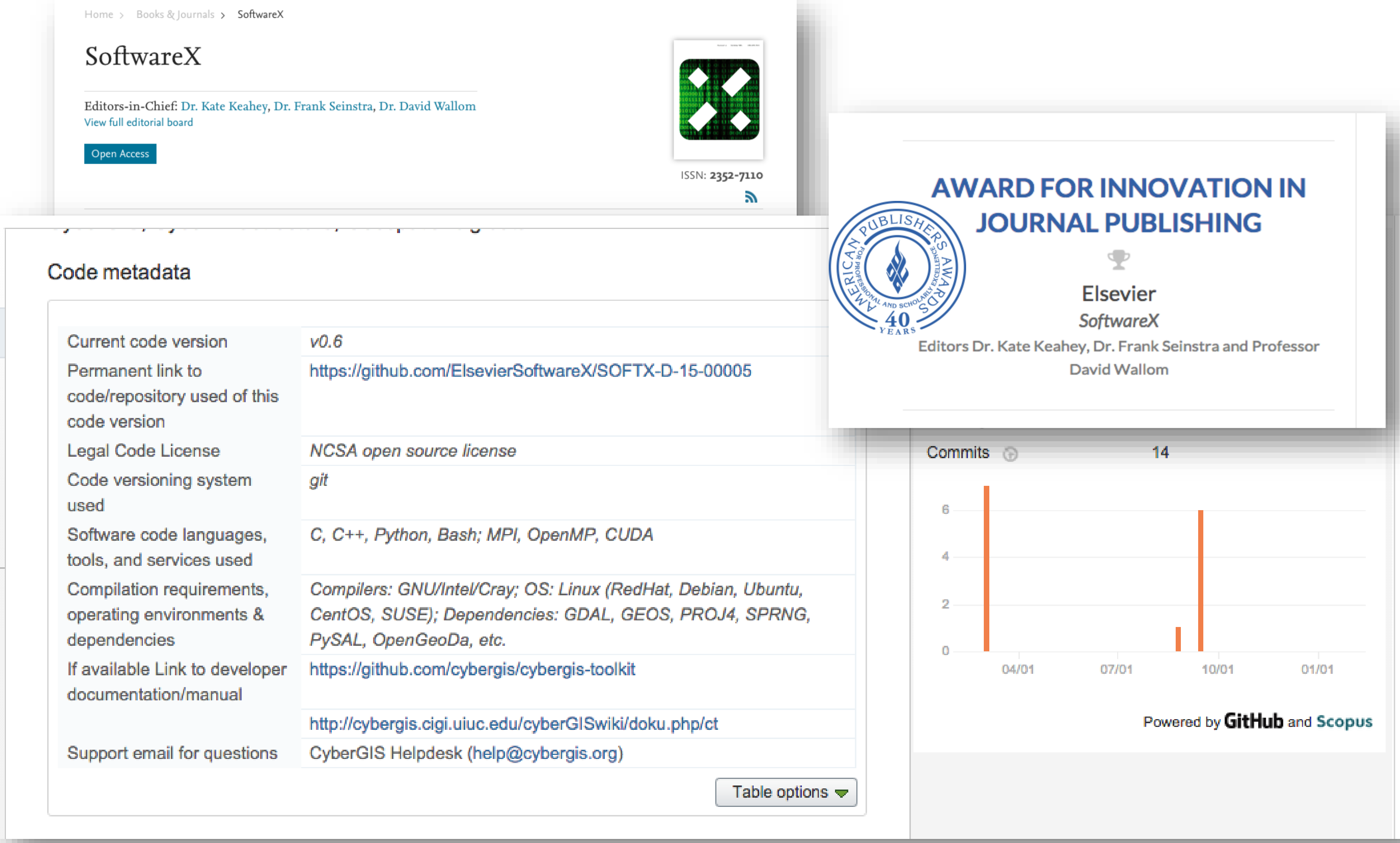
Linked to Github – or not

Versioning and provenance

https://data.mendeley.com/

https://data.mendeley.com/datasets/xz6gv65m6d/6

Data journals: SoftwareX



SoftwareX

Editors-in-Chief: Dr. Kate Keahey, Dr. Frank Seinstra, Dr. David Wallom
View full editorial board

ISSN: 2352-7110

Code metadata

Current code version	*v0.6*
Permanent link to code/repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00005
Legal Code License	*NCSA open source license*
Code versioning system used	*git*
Software code languages, tools, and services used	*C, C++, Python, Bash; MPI, OpenMP, CUDA*
Compilation requirements, operating environments & dependencies	*Compilers: GNU/Intel/Cray; OS: Linux (RedHat, Debian, Ubuntu, CentOS, SUSE); Dependencies: GDAL, GEOS, PROJ4, SPRNG, PySAL, OpenGeoDa, etc.*
If available Link to developer documentation/manual	https://github.com/cybergis/cybergis-toolkit
	http://cybergis.cigi.uiuc.edu/cyberGISwiki/doku.php/ct
Support email for questions	CyberGIS Helpdesk (help@cybergis.org)

AWARD FOR INNOVATION IN JOURNAL PUBLISHING

AMERICAN PUBLISHERS AWARDS FOR PROFESSIONAL AND SCHOLARLY EXCELLENCE
40 YEARS

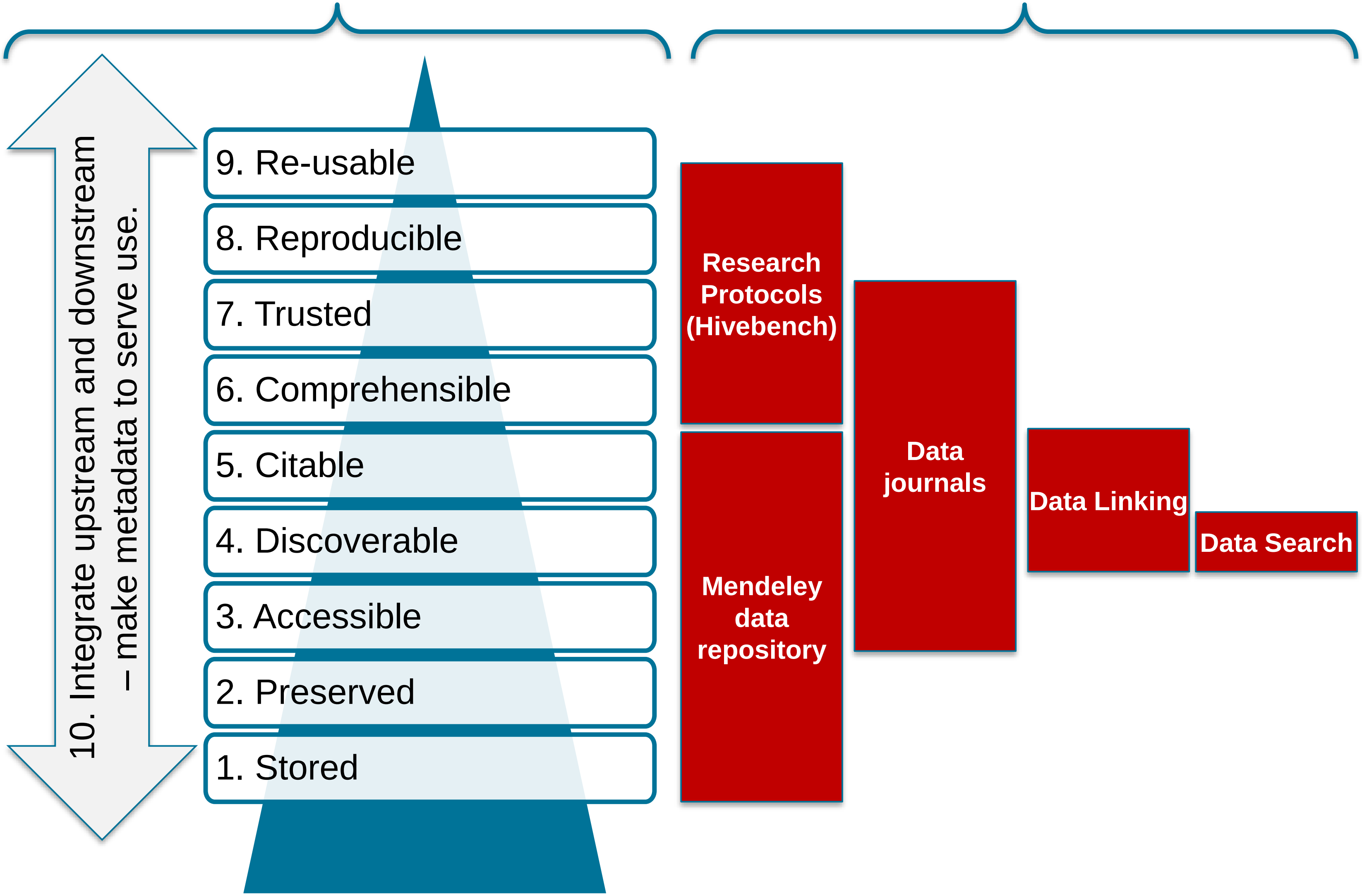
Elsevier
SoftwareX
Editors Dr. Kate Keahey, Dr. Frank Seinstra and Professor David Wallom

Commits 14

Powered by **GitHub** and **Scopus**

The 10 components for effective research data

Elsevier initiatives



10. Integrate upstream and downstream – make metadata to serve use.

9. Re-usable

8. Reproducible

7. Trusted

6. Comprehensible

5. Citable

4. Discoverable

3. Accessible

2. Preserved

1. Stored

Research Protocols (Hivebench)

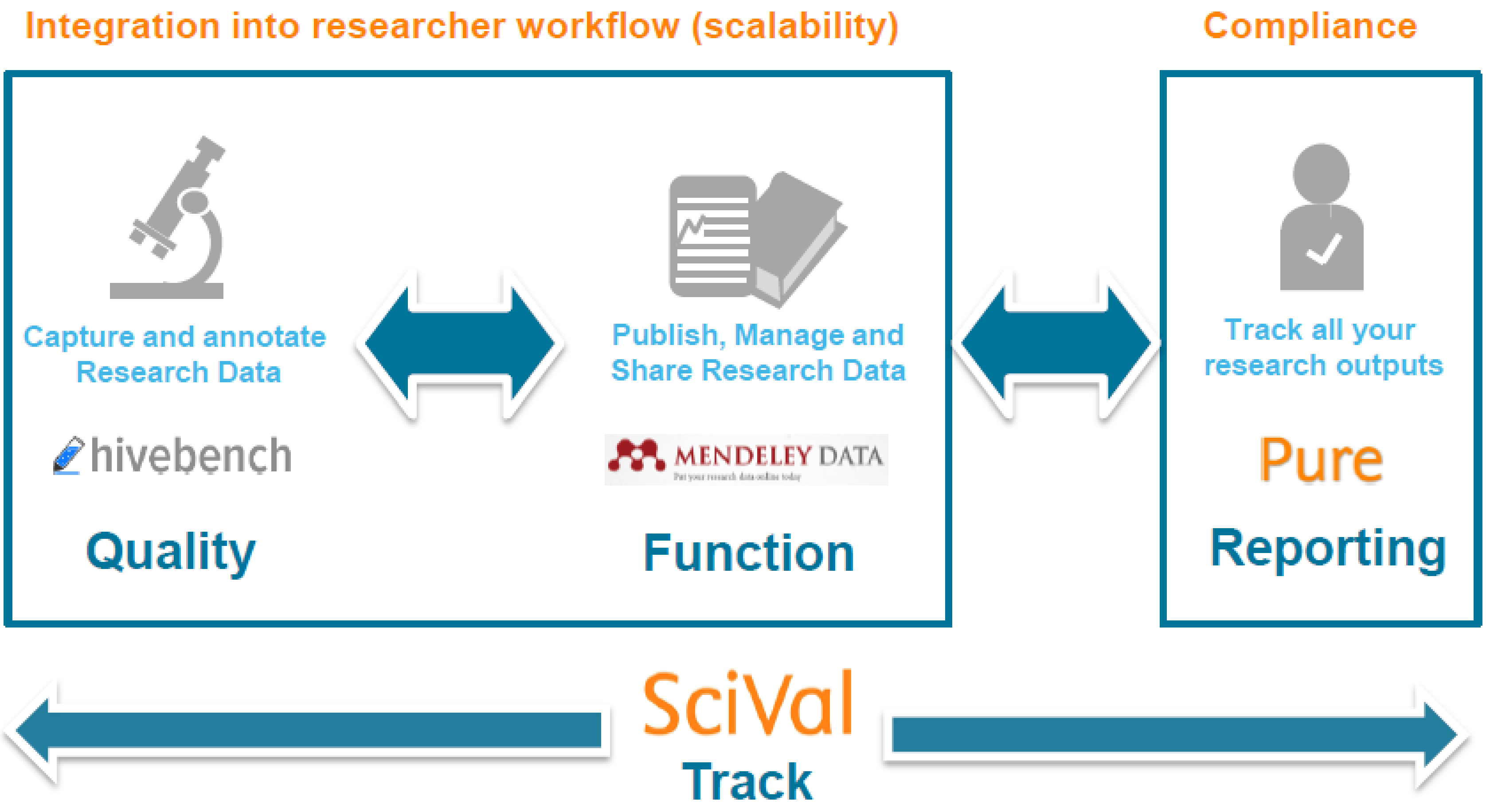
Mendeley data repository

Data journals

Data Linking

Data Search

Efficiency – integration is the building stone



Integration into researcher workflow (scalability)

Compliance

Capture and annotate Research Data

hivebench

Quality

Publish, Manage and Share Research Data

MENDELEY DATA

Function

Track all your research outputs

Pure

Reporting

SciVal

Track

Research Data Policy

Elsevier will:

- Encourage and support researchers and research institutions to **share data where appropriate and at the earliest opportunity**.
- Provide **guidance to authors regarding the deposit and sharing** of data.
- **Encourage and enable two-way linking** of relevant datasets and publications **using permanent standard identifiers.**
- Encourage and **support proper data citation practices** so that researchers can be cited and credited for their work.
- Work closely with the scientific community to **establish data review practices** to ensure that published research data is valid, properly documented and can be re-used.
- **Develop tools and services** to support researchers to **discover, use and reuse** data to further their research.

"Raw research data should be made freely available to all researchers wherever possible" – STM Brussels Declaration 2007

Thank you ! Questions?

Contact me at L.Boudova@Elsevier.com

ELSEVIER

Empowering Knowledge