Università Ca'Foscari Venezia

e-infrastructures austria

ELSEVIER

A publisher point of view on open research data

Wouter Haak – VP Research Data Management

Nov, 2016

New roles in Open Science and Data Stewardship

As a researcher, when I wake up in the morning the first thing I think of is

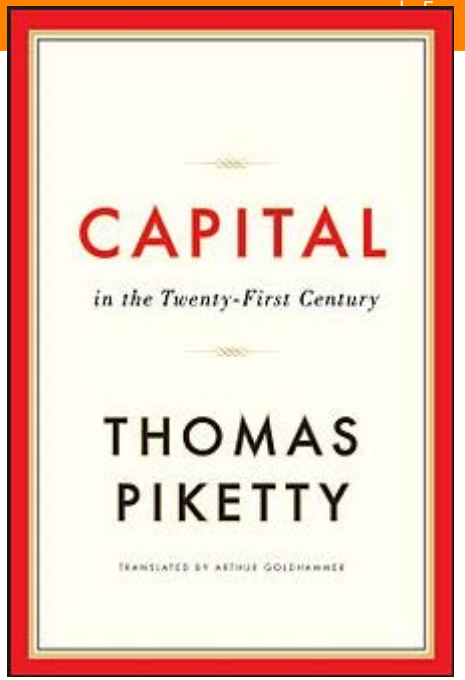
So let’s talk a bit about data

What are we really after: malaria

WWARN, the first malaria data sharing network, has used pooled analysis of shared data to provide evidence to **help improve dosing regimens** of malaria treatments

- 260 institutions globally have worked with WWARN, and over 120,000 individual patient records have been contributed
- Based on the results, the World Health Organization has revised the recommended dose of DP, a commonly used antimalarial for young children

What are we really after: social sciences



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

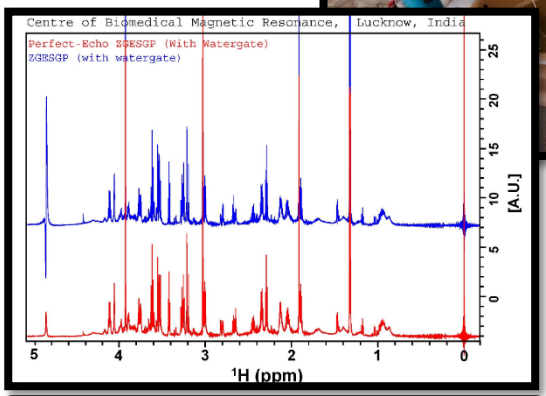
- It focuses on wealth and income inequality in Europe and the United States since the 18th century
- Central thesis is that when the rate of return on capital (r) is greater than the rate of economic growth (g) over the long term, the result is concentration of wealth, and this unequal distribution 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 https://www.quandl.com/data/PIKETTY

Piketty's Capital: An Economist's Inequality Ideas Are All the Rage" by Megan McArdle, *Bloomberg Businessweek*, May 29, 2014
https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century

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

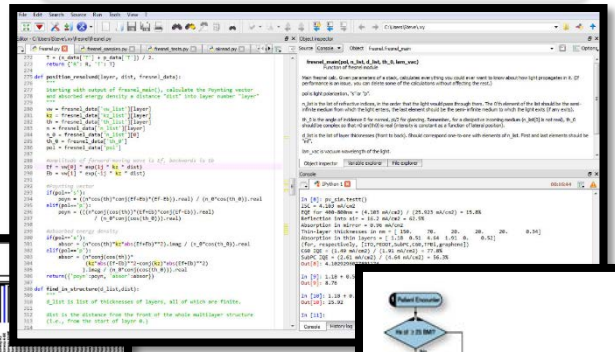
Machine & environment settings

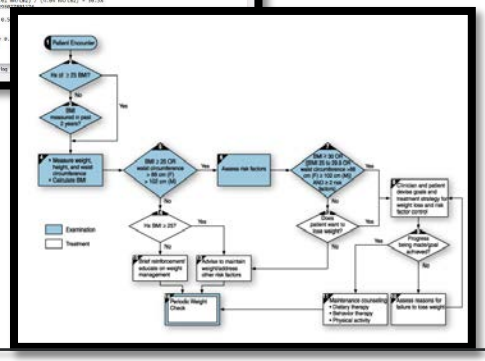
Scripts, analyses, algorithms



Raw data

Processed data



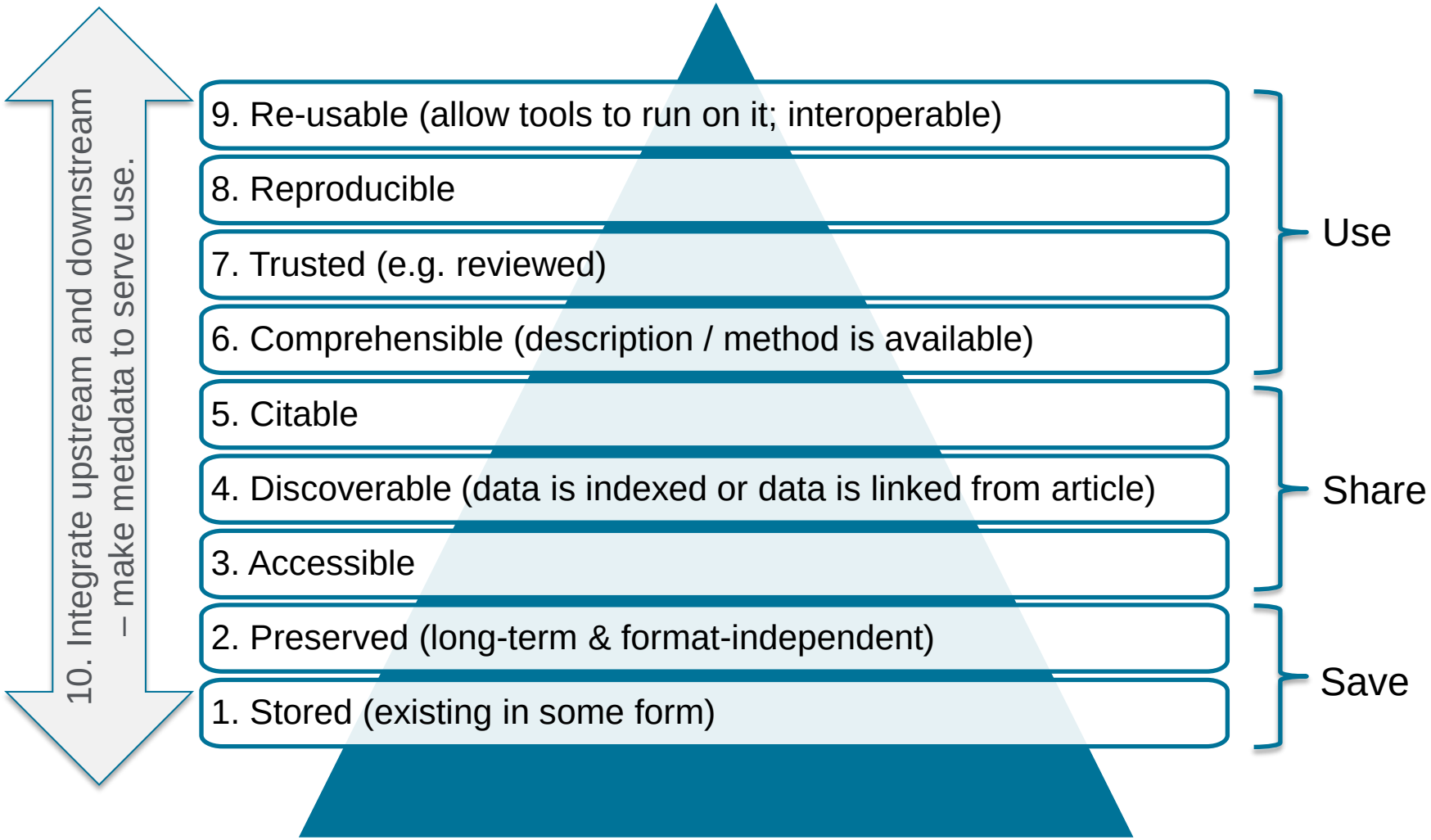


Protocols, methods, workflows

Full provenance needed

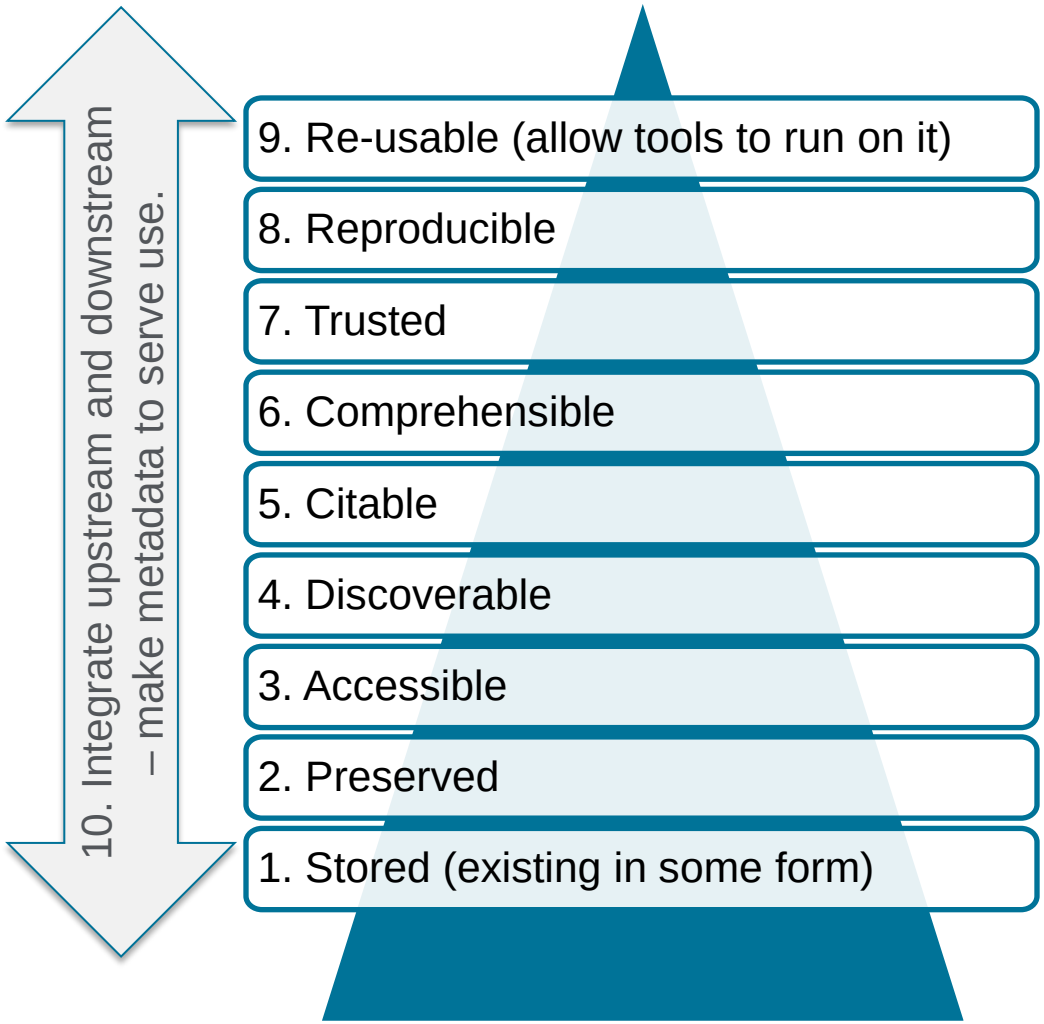
Note: images for illustrative purpose only

The 10 components of effective research data



https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data
De Waard, Cousijn, Aalbersberg

Just another way of saying: “Data needs to be FAIR”



Findable

Accessible

Interoperable

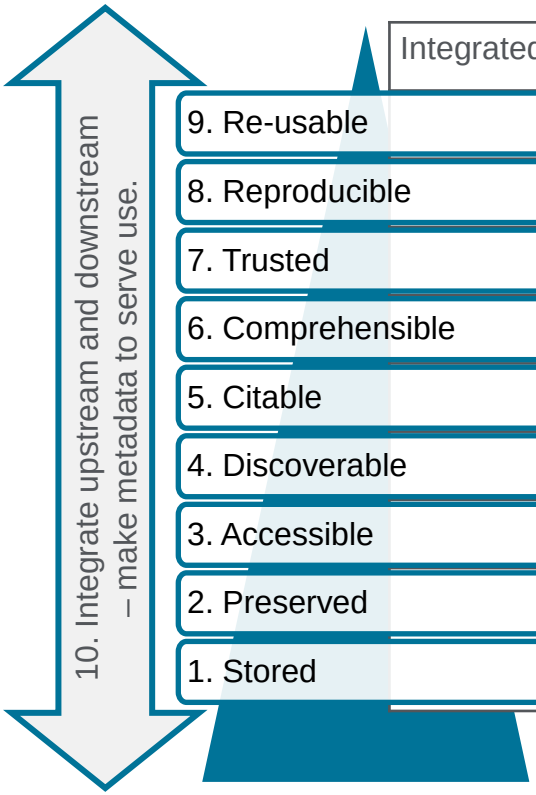
Reusable

https://www.force11.org/group/fairgroup/fairprinciples

Compliance increasingly important for research data

Some funder examples

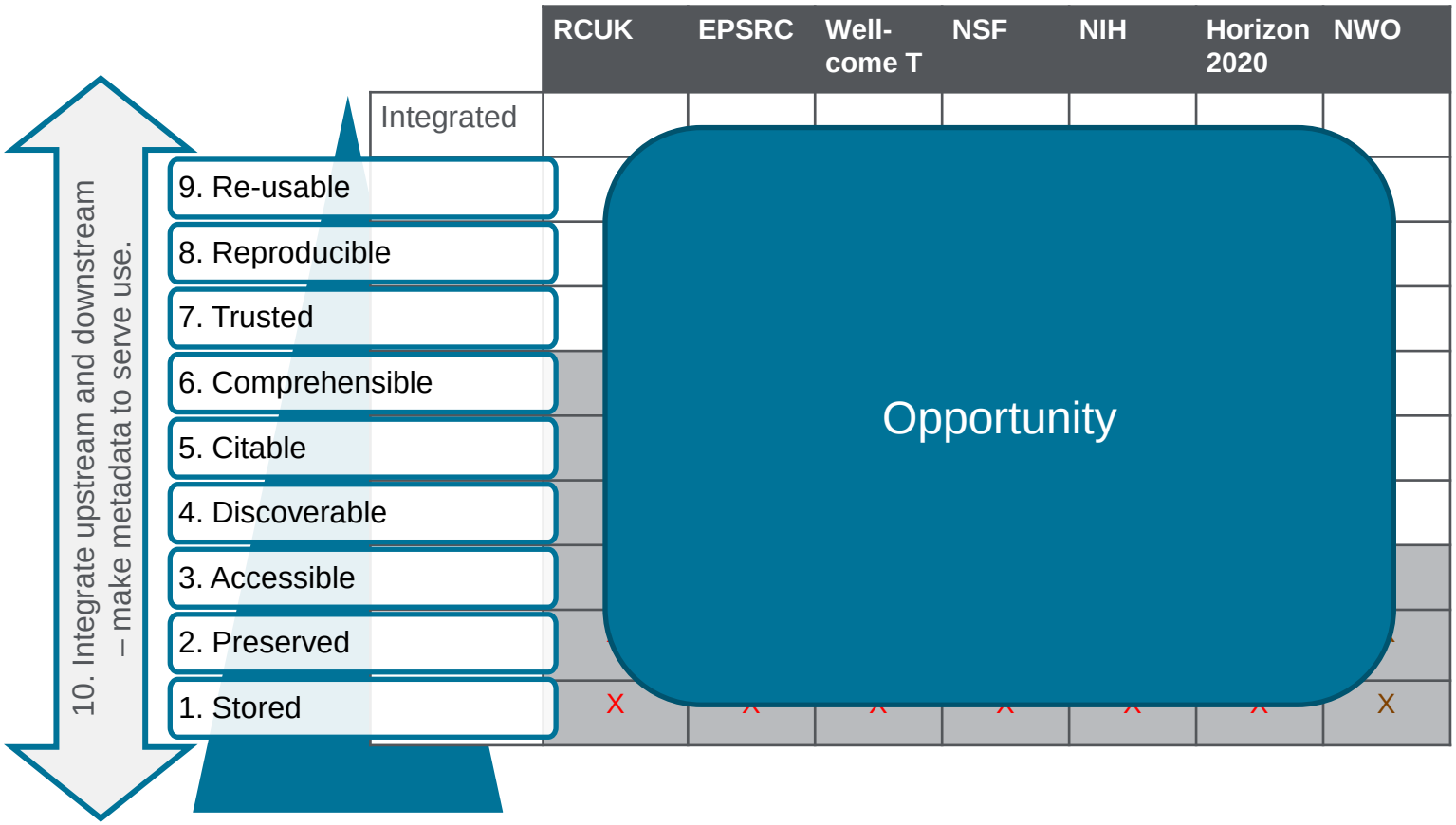
X) Required
X) Recommended



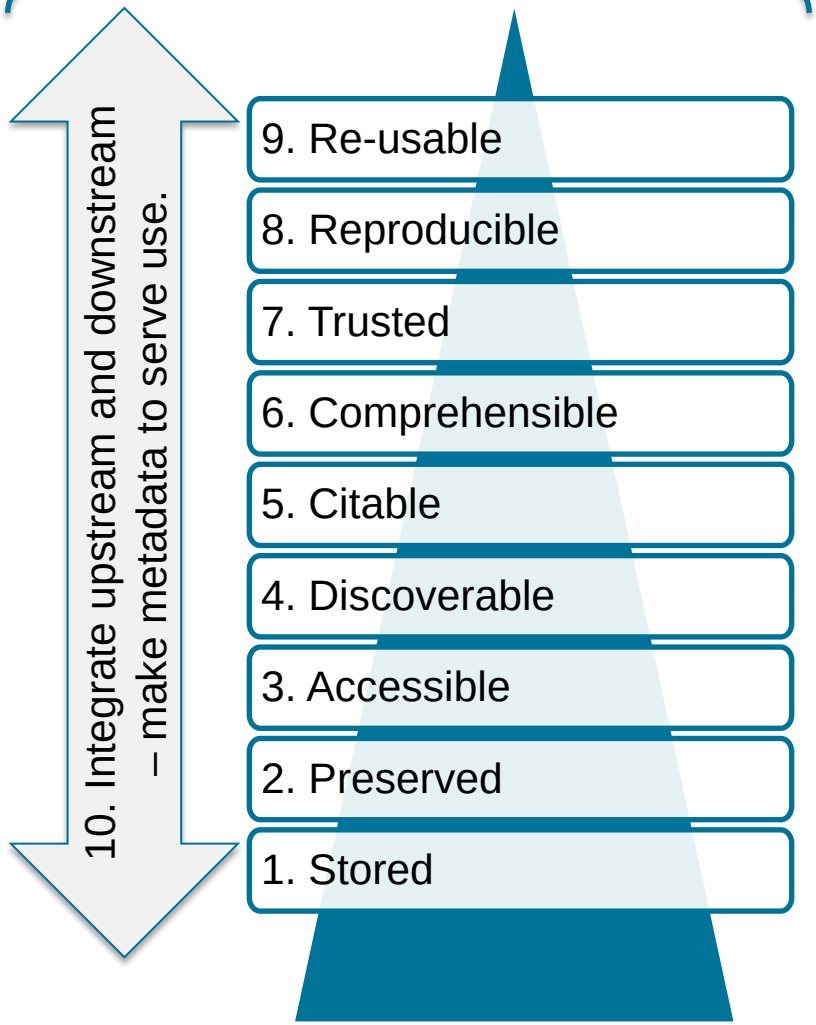
	RCUK	EPSRC	Well-come T	NSF	NIH	Horizon 2020	NWO
Integrated							
9. Re-usable							
8. Reproducible							
7. Trusted							
6. Comprehensible	X	X			X	X	
5. Citable	X	X			X		
4. Discoverable	X	X			X	X	
3. Accessible	X	X	X	X	X	X	X
2. Preserved	X	10 Years Minimum	10 Years Minimum	Duration depends	Duration depends	X	X
1. Stored	X	X	X	X	X	X	X

Based on analysis of policies; august 2016
https://docs.google.com/document/d/1ODK1D4gs5KI59o-ADo-Ps5_zhzdOlc_E0ur8EEB1hIY/

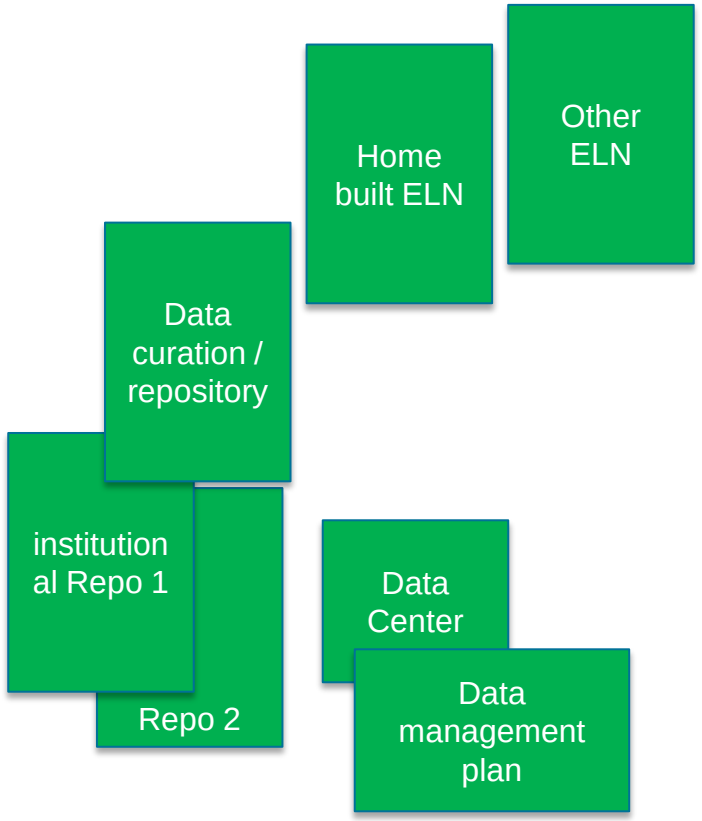
Opportunity to move beyond compliance!



The 10 components for effective research data

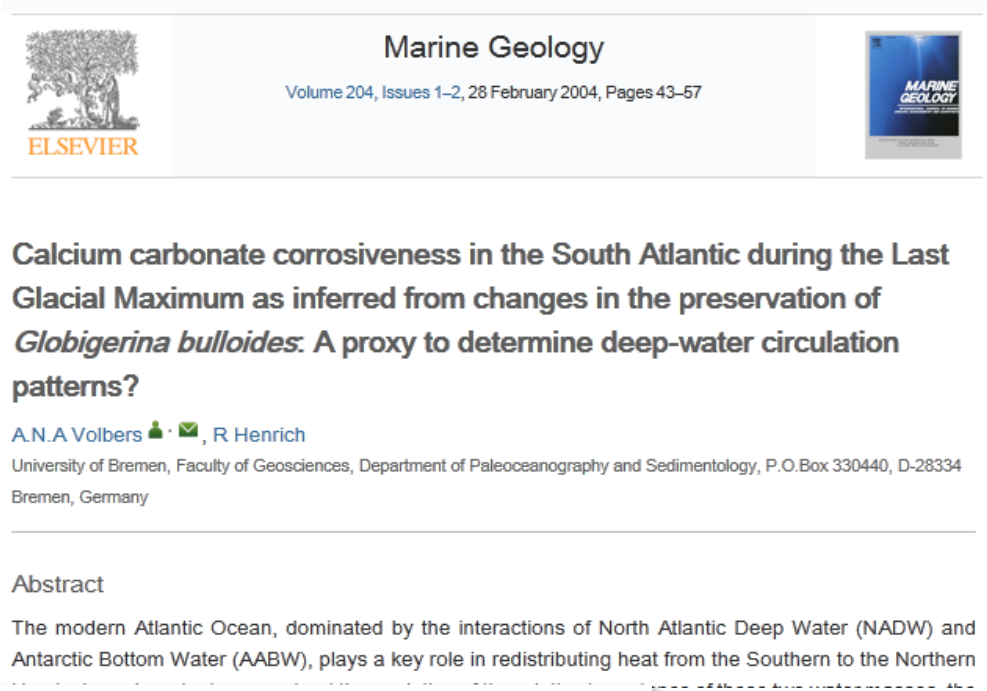


Most researchers already work with several data tools

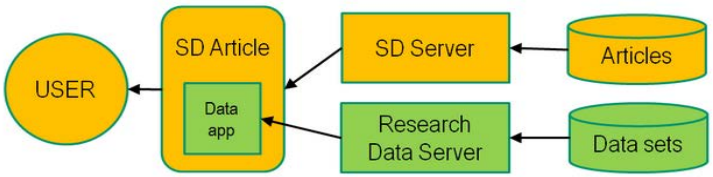


What can Elsevier do?

Linking papers to data: phase 1
Elsevier Database linking program – example Pangaea



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

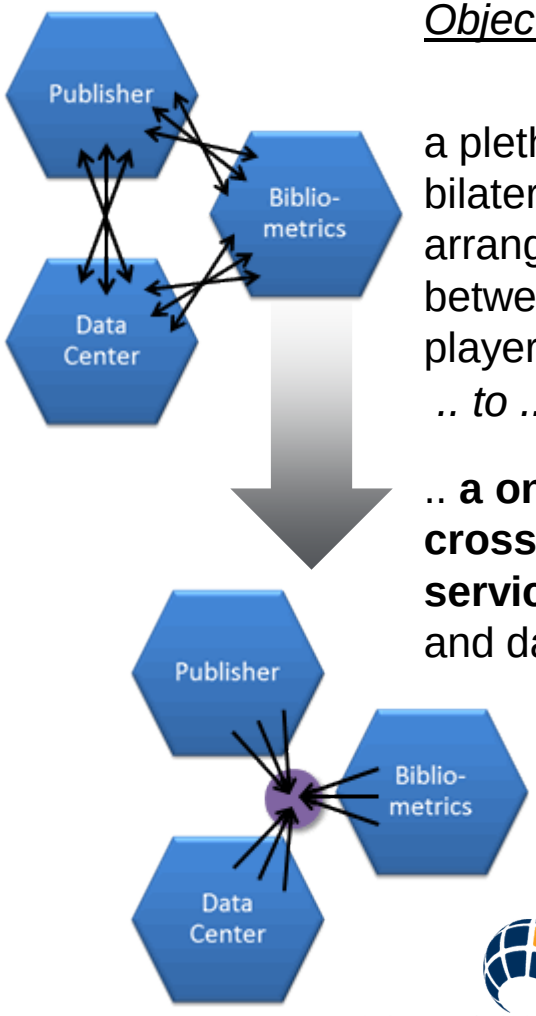


http://www.elsevier.com/databaselinking

Linking papers to data: phase 2
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 ..

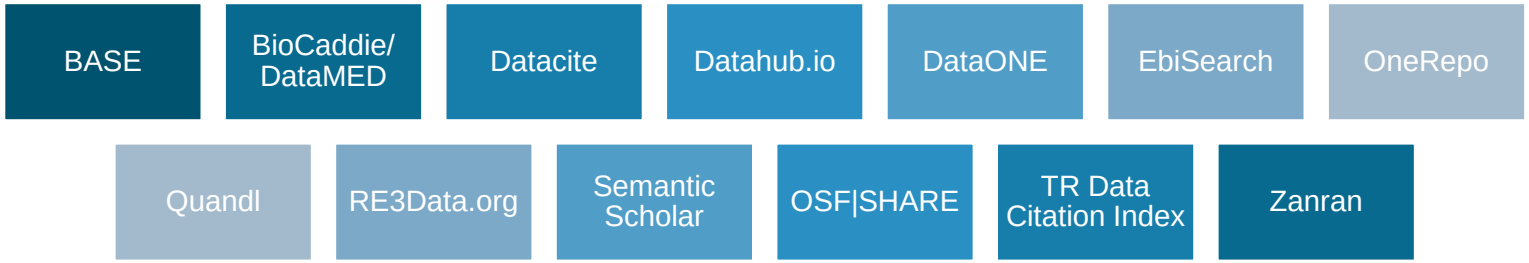
.. **a one-for-all cross-referencing service** for articles and data

ICSU WORLD DATA SYSTEM

RDA RESEARCH DATA ALLIANCE

Phase 3: we need a research datasearch engine!

- Many (broad) datasearch examples already available

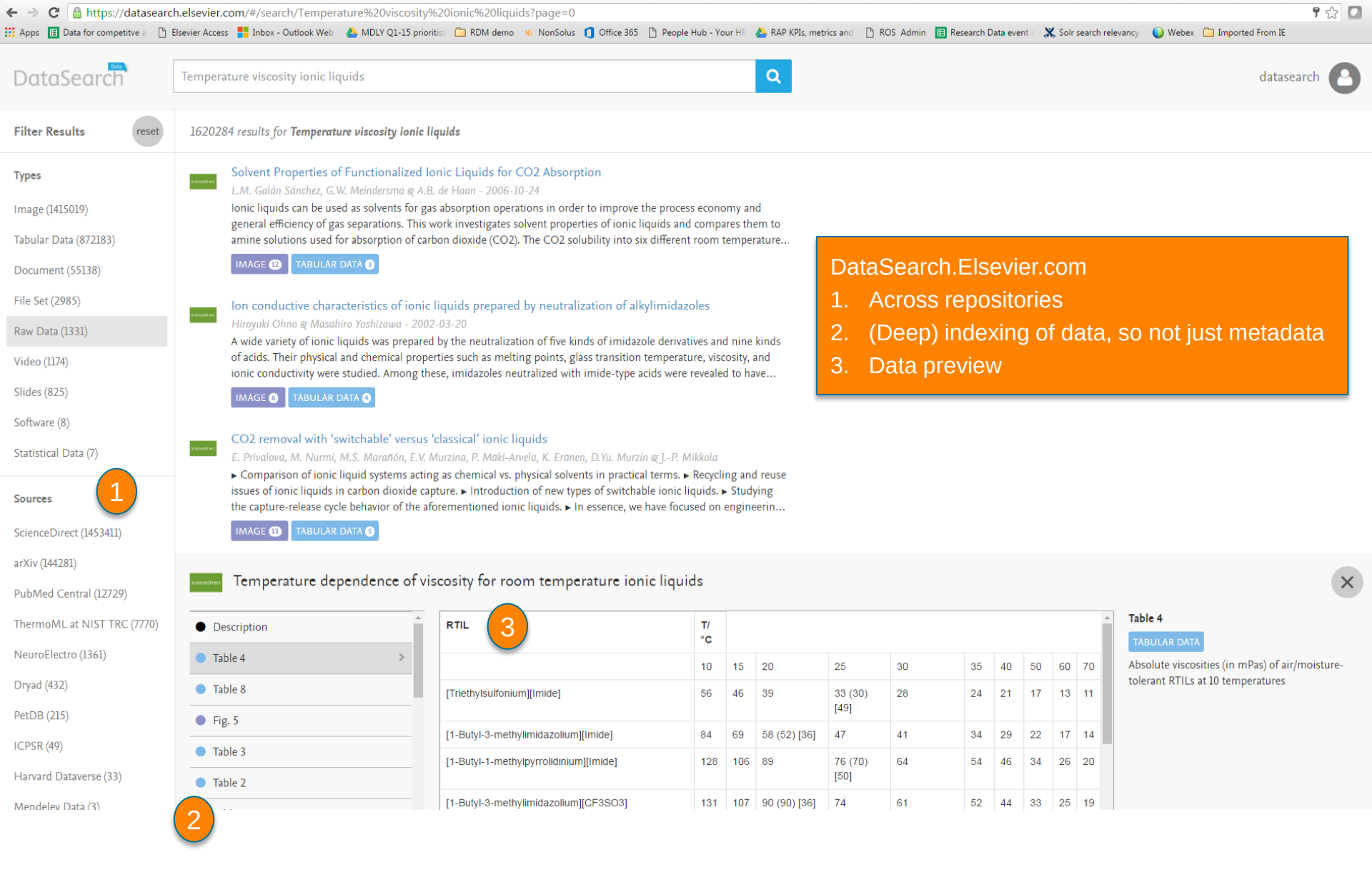


- Some common themes:
  - search of metadata only (i.e. ranking based on metadata only)
  - And/or federated search (i.e. no ranking)
  - And/or focused on giving credit (citation) rather than on discoverability

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

Phase 3: Elsevier Data Search

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



Hivebench Electronic Lab Notebook

www.hivebench.com

hivebench | Search for results | Notebook | Protocol | Inventory | Data

Notebook

Team Notebook 15

experiments schizophrenia risk genes 13

My Notebook 1

Lab notebook

helena 05/21/2016
Schizophrenia risk gene x

helena 04/15/2016
schizophrenia risk gene mir137

helena 03/07/2016
ZNF804A expression

helena 02/05/2016
expression of tcf4

helena 01/21/2016
expression of znf804a

helena 01/12/2016
fetal brain expression

helena 12/31/2015

Experiments

schizophrenia risk gene mir137

helena created on April 15th 2016

-Wash in PBS with 0.1%Tween20, 3 X, and 10 min of each time

-Incubate with secondary antibody (peroxidase-conjugated goat anti-mouse IgG, etc.) following the manufacturer's instruction, usually 1:200 to 1:2000; 1 hour at room temperature

- Wash in PBS with 0.1%Tween20, 3 X, and 10 min of each time

5. Developing: (Use of chromomeric substrate- 3,3'-diaminobenzidine, DAB)

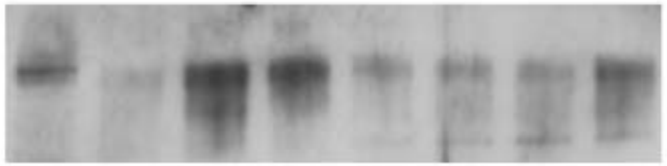
-Transfer the membrane to a shallow tray.

-Add 10ul H_2O_2 (30%) to 10 ml of 0.05%DAB in PBS, mix well immediately;

-Pour the DAB to the membrane, incubate at room temperature with gentle shaking in the dark if possible.

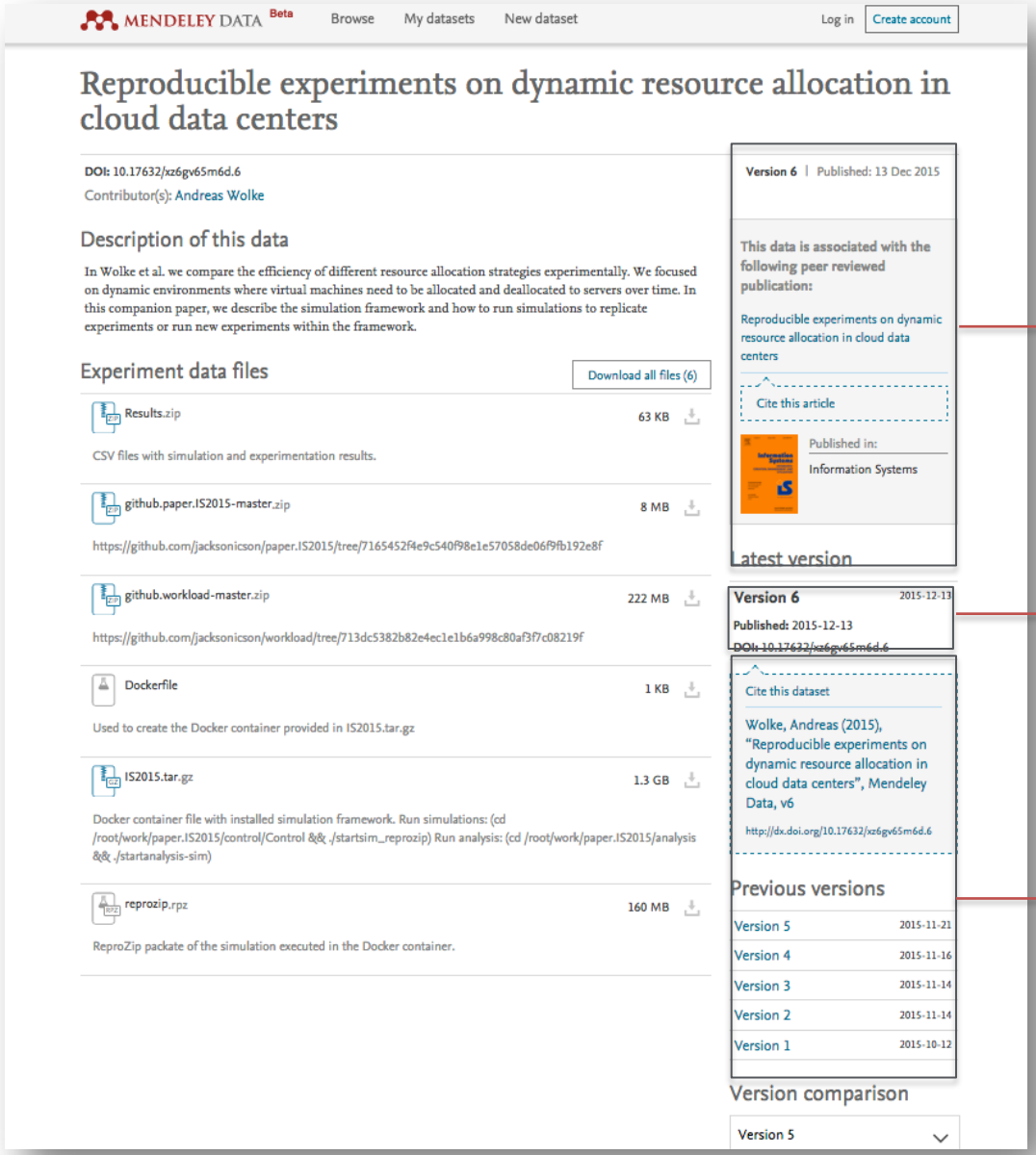
-Monitor the progress of the reaction carefully. When the bands are of the desired intensity (2-5 min), wash the filter briefly in water, and in PBS.

-Dry the membrane and photograph it to provide a permanent record of the experiment.



Protocols and results

Manage, Store: Mendeley Data repository



Linked to published papers – or not

Researcher in control: how/when to share

Versioning and provenance

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

Publish data and software: e.g. SoftwareX and DataInBrief have top 25% most read OA articles on ScienceDirect

SoftwareX

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

Open Access

ISSN: 2352-7110

AWARD FOR INNOVATION IN JOURNAL PUBLISHING

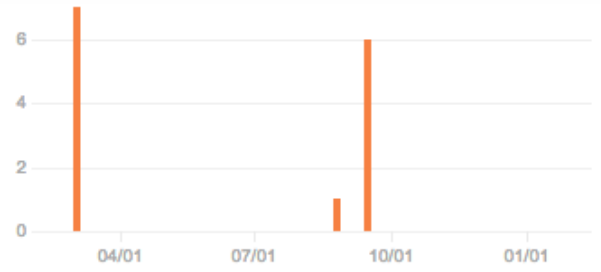
Elsevier
SoftwareX
Editors Dr. Kate Keahey, Dr. Frank Seinstra and Professor David Wallom

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)

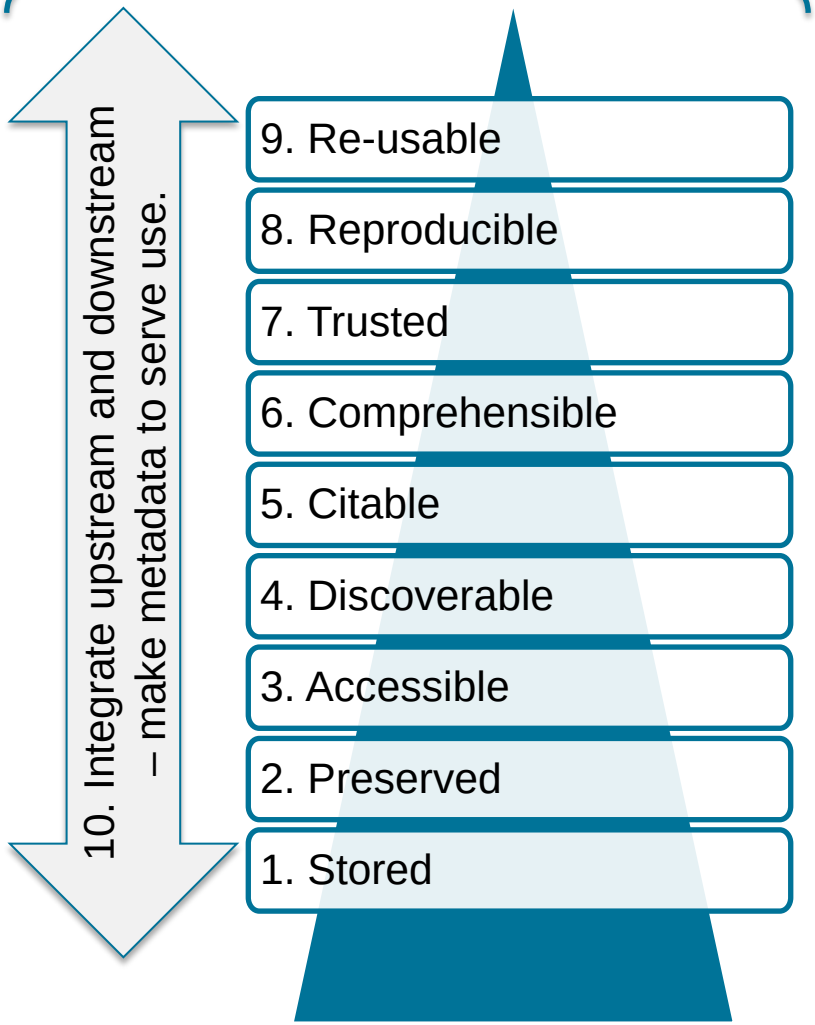
Table options

Commits 14



Powered by GitHub and Scopus

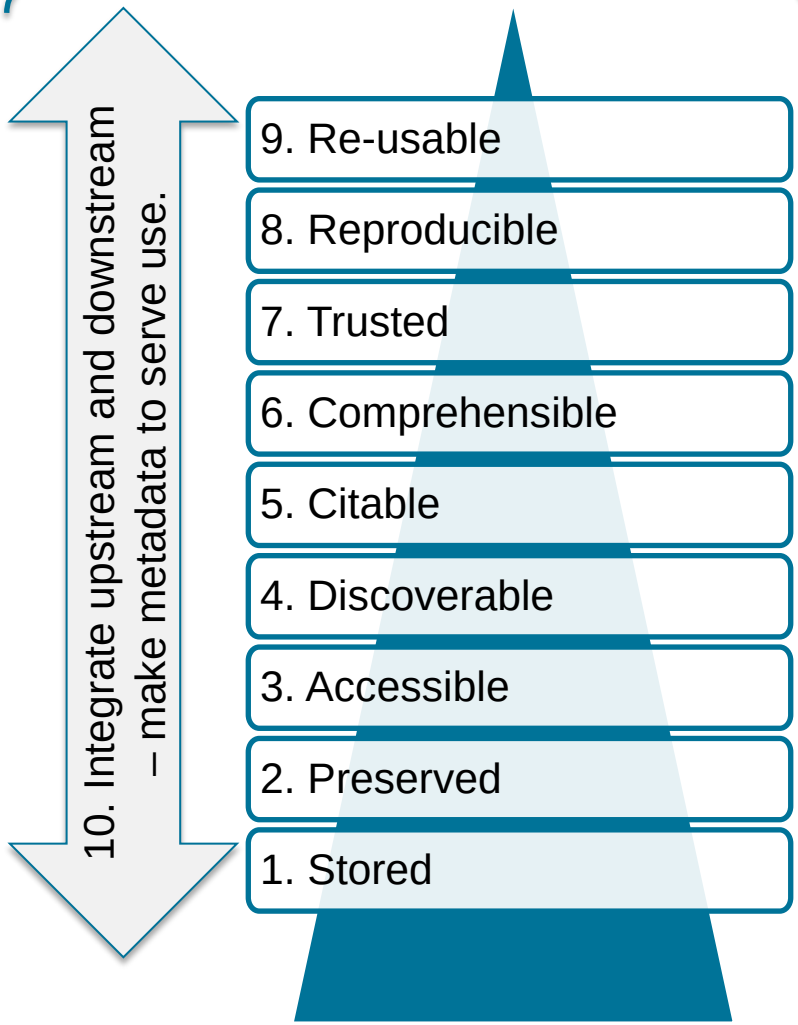
The 10 components for effective research data



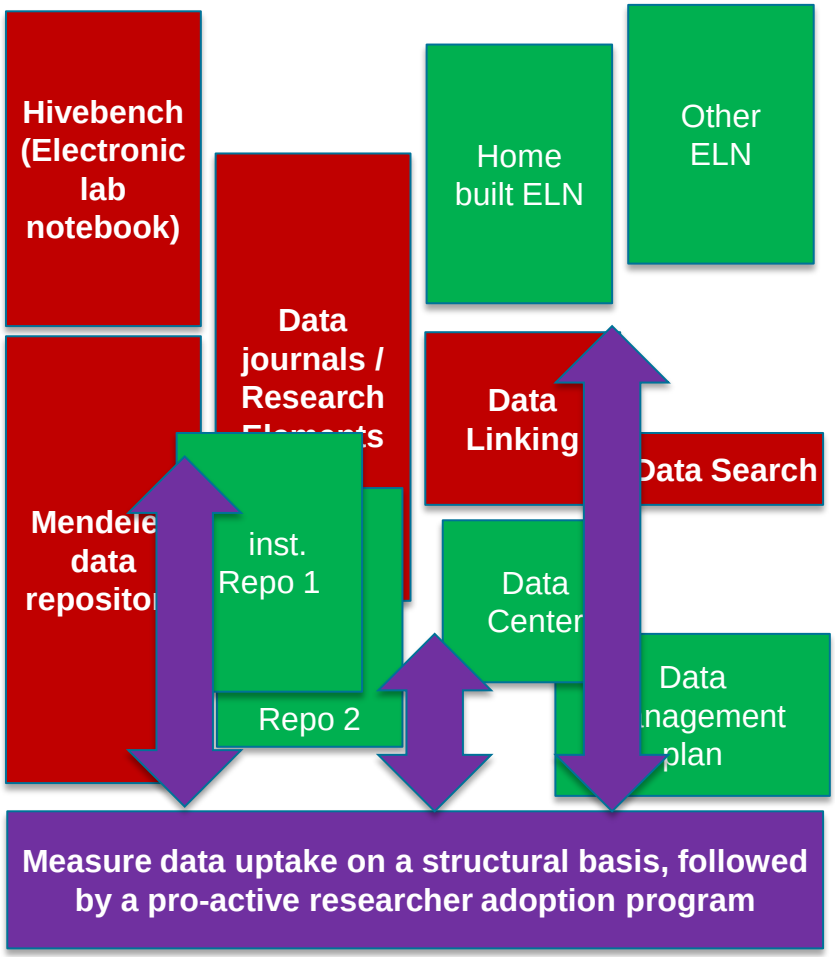
Summary of initiatives discussed

- Hivebench (Electronic lab notebook)
- Mendeley data repository
- Integrated with: Data journals / Research Elements
- Integrated with: Data Linking
- Integrated with :Data Search

The 10 components for effective research data



Moving towards an integrated ecosystem to help the researcher and the institution



Grazie mille!