

A data journey

From the lab bench to FAIR data reporting

euroCRIS Strategic Members Meeting
Bratislava, 21-22 November 2017

Alberto Zigoni
Elsevier Research Data Management Solutions

Funders, associations, and institutes increasingly require data sharing



RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles



OPEN RESEARCH DATA IN HORIZON 2020

CHALLENGE

Wider access to scientific facts and knowledge helps researchers, innovators and the public find and re-use data, and check research results:

offers better value for EU research funds

encourages research across scientific fields



a public benefit



essential for solving today's complex societal challenges

SOLUTION

Horizon 2020 already mandates open access to all scientific publications



From 2017, research data is **open by default**, with possibilities to opt out

to be made open

VSNU calls for a national RDM strategy

06 MAY 2015

The Association of Universities in the Netherlands (VSNU) has asked SURF to set up a National Research Data Management Coordination Point. The VSNU's Research and Valorisation Steering Group notes that much work in the RDM field is already taking place at universities.



WAGENINGEN
UNIVERSITY & RESEARCH

Home > Expertise & Services > Data Management Support Hub

Data Management Planning

Data management planning is a matter of good research practice. At Wageningen University & Research PhD candidates and Chair Groups are required to have a Data Management Plan.

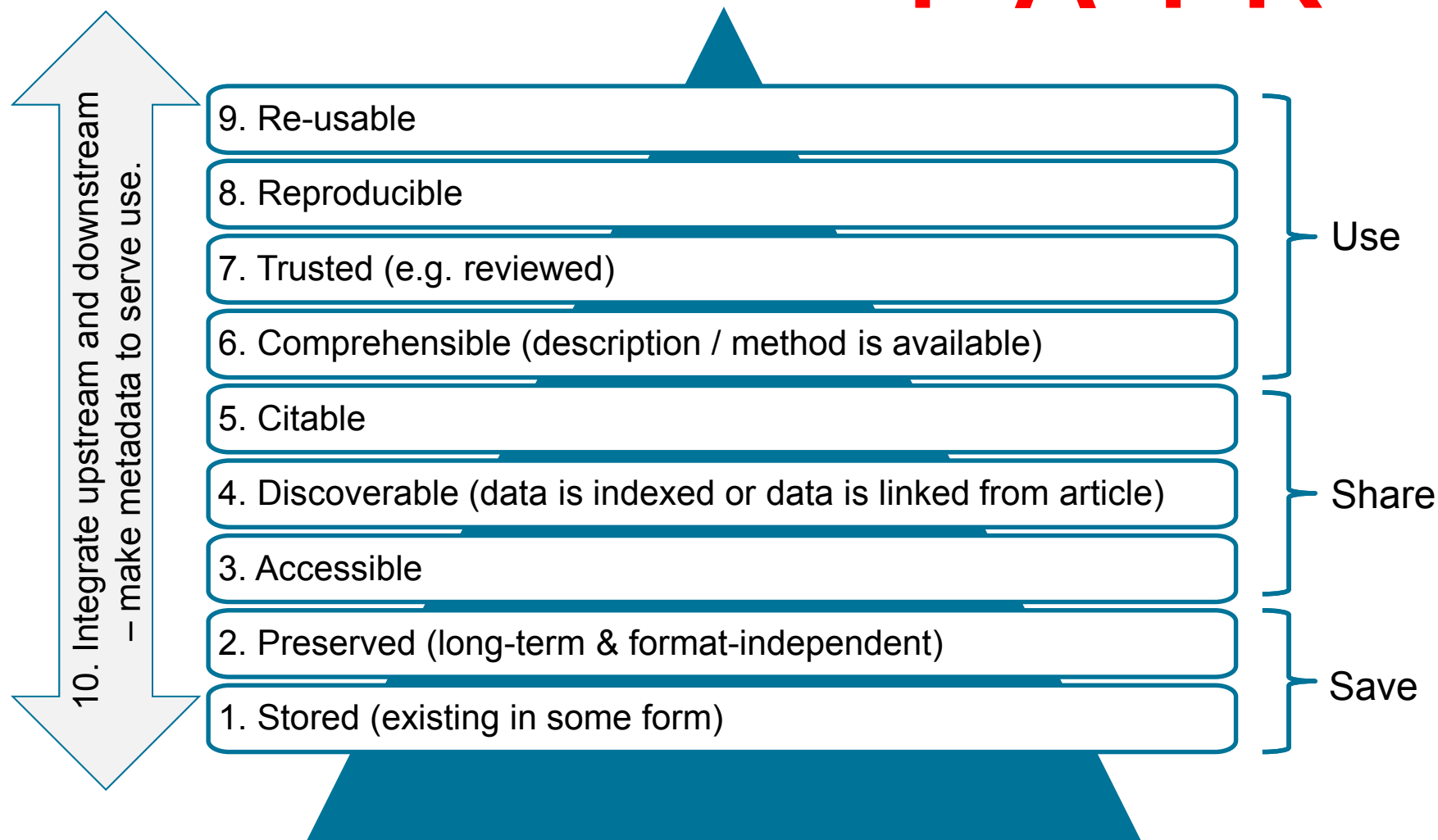


National Coordination Point
Research Data Management

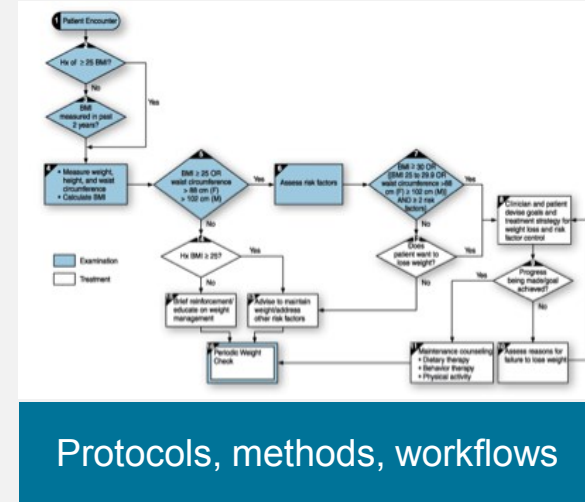
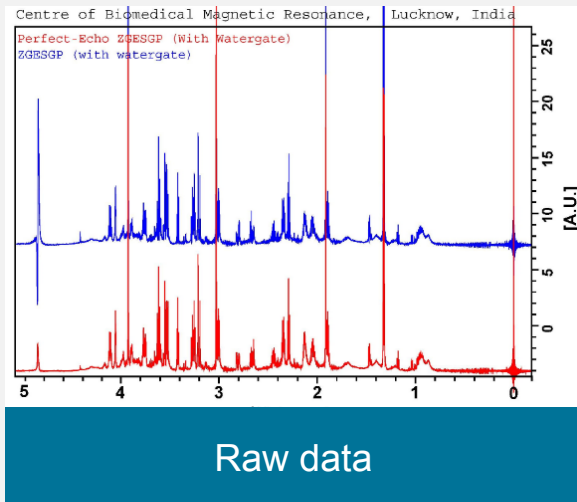


10 components of effective research data

FAIR



When talking about data, we talk about...



```

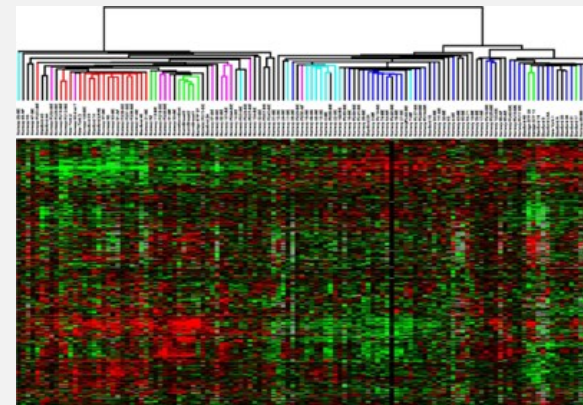
def find_max_value_in_array(arr):
    """Find the maximum value in a 1D array"""
    max_val = arr[0]
    for i in range(1, len(arr)):
        if arr[i] > max_val:
            max_val = arr[i]
    return max_val

def find_max_value_in_2d_array(arr):
    """Find the maximum value in a 2D array"""
    max_val = arr[0][0]
    for i in range(1, len(arr)):
        for j in range(1, len(arr[i])):
            if arr[i][j] > max_val:
                max_val = arr[i][j]
    return max_val

# Example usage
arr1 = [1, 2, 3, 4, 5]
arr2 = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]

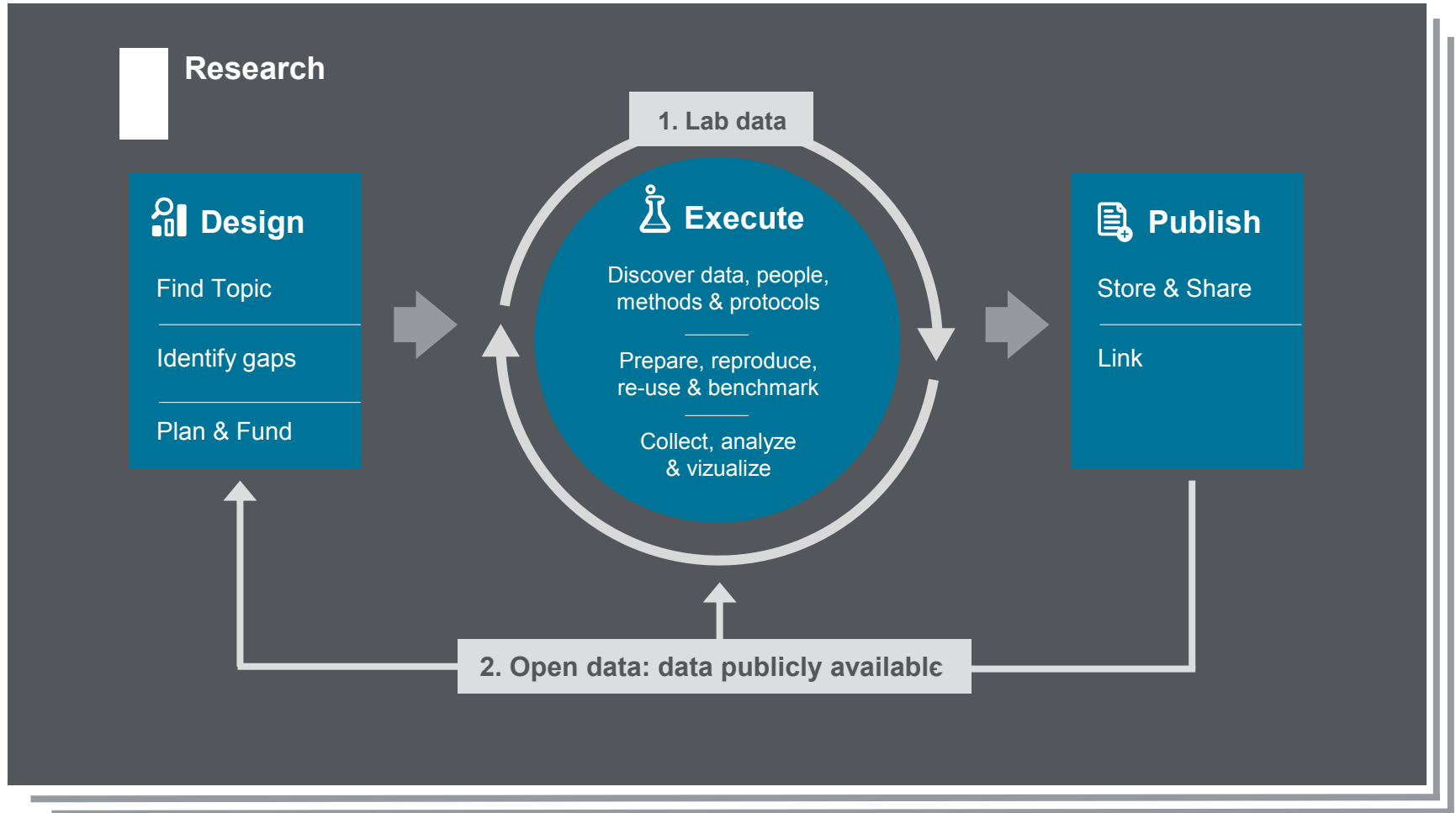
print(find_max_value_in_array(arr1))
print(find_max_value_in_2d_array(arr2))
  
```

Scripts, analyses, algorithms

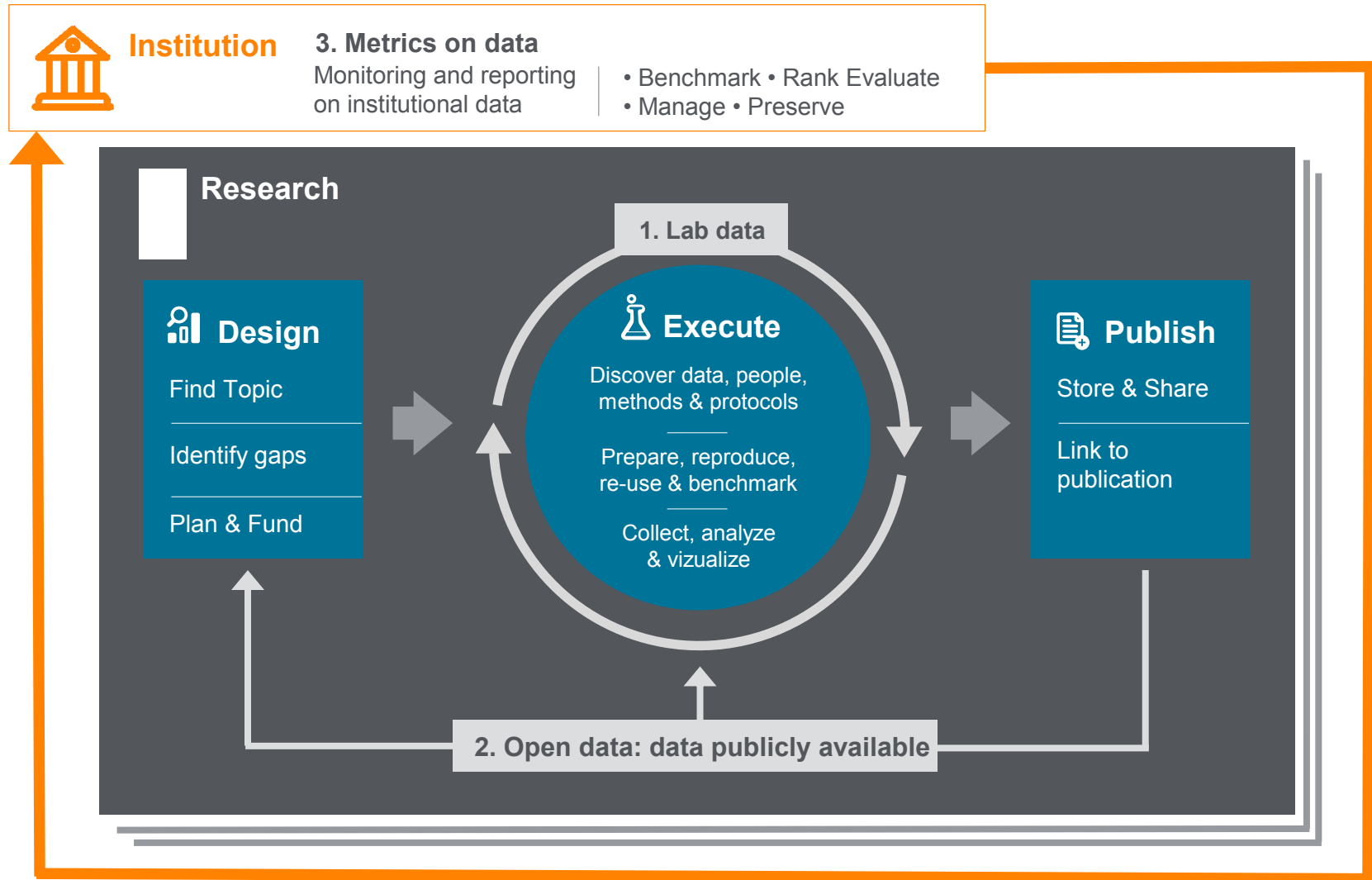


The research life cycle depends heavily on two data life cycles:

1. Within the lab
2. Within the world at large



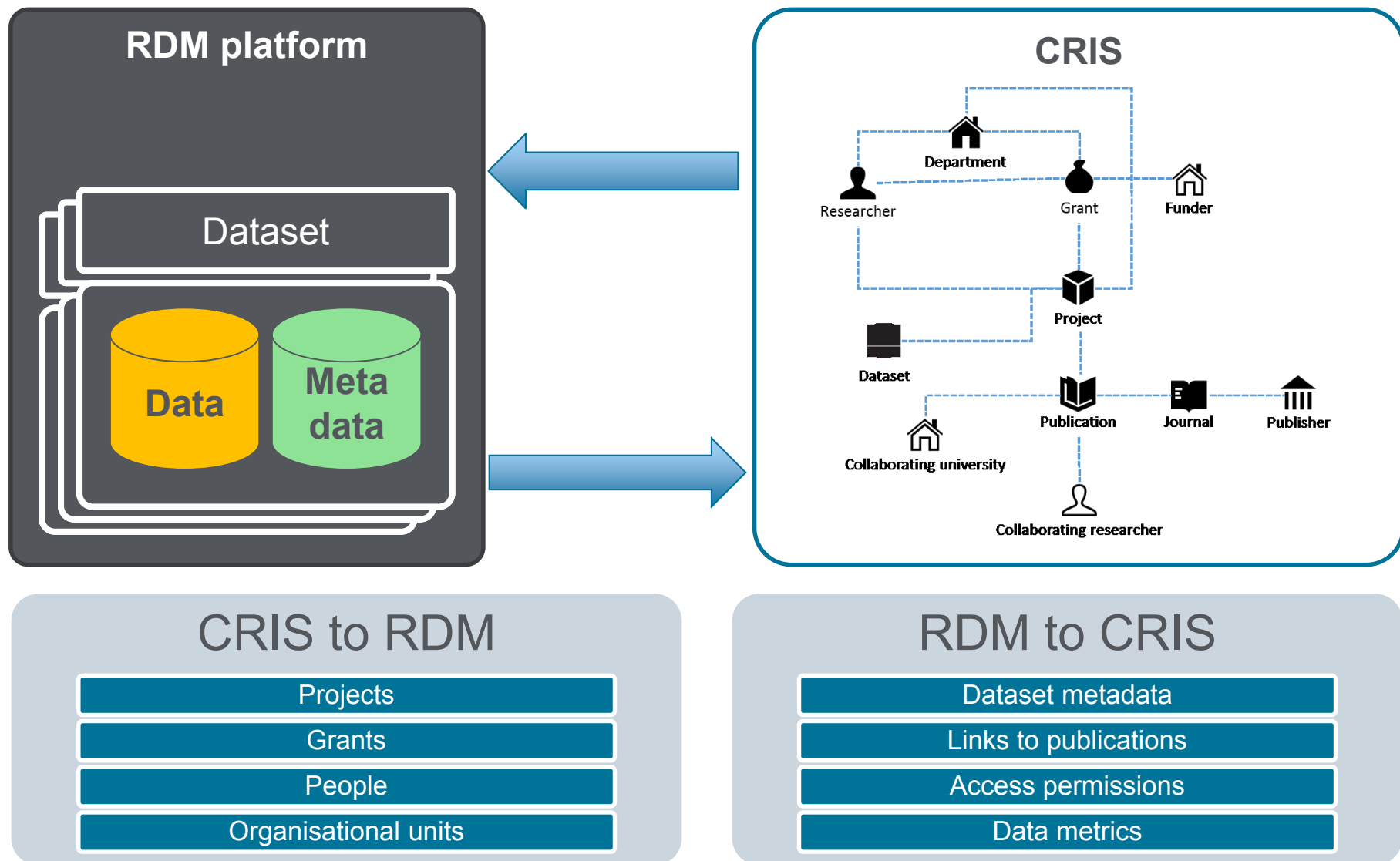
Taking the institutional lens, we can speak of 3 interlocking data life cycles



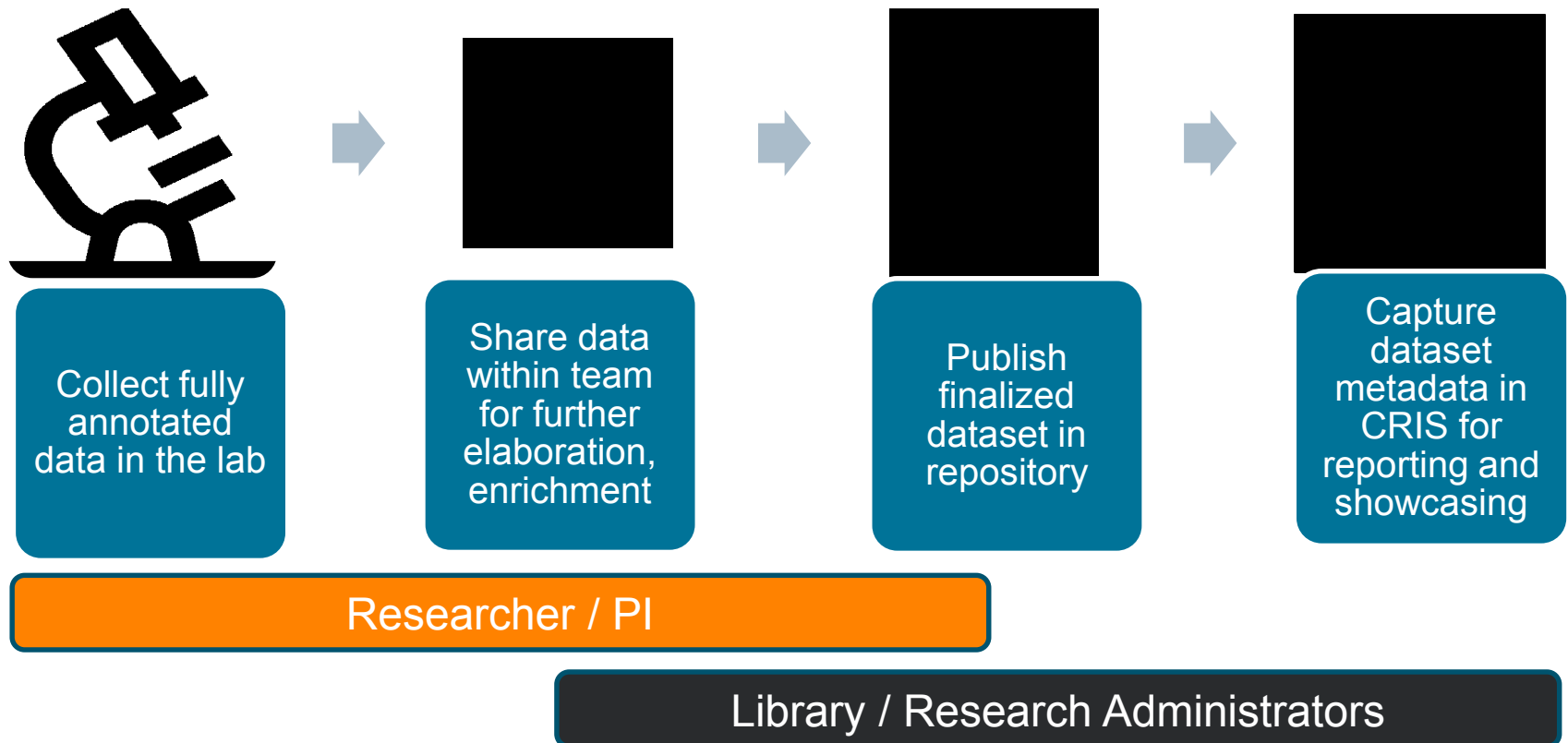
Re-using research data improves outcomes for the research life cycle

- This means improving the research data life-cycles: (1) within the lab and (2) to the world at large
- This also means keeping track of the institutional data lifecycles, and (3) reporting on them

Data should be exchanged between RDM platform and Current Research Information System



The data journey: from the lab bench to a report



A real example: experiment captured in electronic notebook

The screenshot displays the Hivebench electronic notebook interface. The main section is titled "Protocol" and shows a "Wings Drugome Workflow" by anitaawaard123. The workflow is categorized under "General" and includes details such as the author (Sarah L. Kinnings), laboratory (Department of Chemistry and Biochemistry, University of California), and date (18-08-2016). The abstract states: "We use the Wings workflow to capture and analyze experimental data." The workflow components are listed as follows:

- Comparison of ligand binding sites using [MAP](#)
- Comparison of drug chemical similarity using [custom scripts](#)
- Comparison of global protein structures and filtering using [FATCAT](#)
- Molecular docking using [Autodock Vina](#). The original work used [eHiTS](#), which has similar functionality.
- Visualization using [yEd](#), [Circos](#) and [Gephi](#). The original work used [Cytoscape](#)

The interface also shows a list of protocols on the left: Basic PCR protocol, PBS 10X recipe, Test Protocol, PCR protocol, and izMiR: computational ab initio microRNA detection. On the right, the "Inputs" section lists various files and their sizes, including configurationFile, drugKeys, drugBindingSites, pvalue, solvedStructuresFiles, structureForComparison1, tbProteinInfo, template_pdb, structureForComparison, and model_info. The "Outputs" section lists dockingResults1, highlyConnectedDrugs, insignificantresults, networkVisualization, and Conclusion. A blue speech bubble with the text "Protocol, input and output data captured together" is overlaid on the interface.

Draft dataset added to project space for collaboration



Mendeley

Browse Datasets

My Datasets

My Projects

New Dataset

FAQ

[< Back](#)

TB drugome

Members (3): [View list](#)Details: [Show](#)

My Datasets

Private

Project Team

All mem

ZNF804A expression data

Wings Drugome Workflow

Project metadata can be pulled in from CRIS

[+ Create new dataset](#)

Custom, discipline-specific metadata can be added to datasets based on metadata template libraries

Protein Name	Vav
Accession Number	Genbank ↕
HPRD Identifier	
Gene Symbol	
Chromosome Locus	22q13.1
Molecular Class	DNA telomerase AND C See List
PTMs	Myristoylation See List
Cellular Component	Cytoskeleton See List
Domain Name	EGFCA See List
Motif	HLH See List
Expression	Bed nucleus of the Stria See List
Length of Protein Sequence	From : to : in amino acids
Molecular Weight	From : to : in kDa
Diseases	Ovarian cancer

Final version of the dataset published in the data repository



MENDELEY DATA

Browse

My datasets

New dataset

You are previewing this dataset.

TB-Drugome - Run 126

Contributor(s): Sarah L. Kinnings, Li Xie, Kingston H.

Description of this data

A run of the TB-Drugome Workflow.

Experiment data files



IN:run_126_workflow.png



IN:smmap_properties.txt



IN:drug_key.csv



IN:drug_binding_sites_list.csv



IN:MtbStructures_tar_gz.gz



IN:homology_models_list.txt.txt

17 KB



IN:tb_protein_info.csv

42 KB



IN:template_pdb.csv

24 KB



Steps to reproduce

The TB-Drugome Workflow is complicated to use without the proper knowledge in bioninformatics and computer science. This section aims to provide the initial pointers and references to look at when trying to reuse the TB-Drugome workflow.

- It is recommended to first understand the flow of the data by looking at published run schema.
- Then look at the original Drugome paper for further information and the tools being used.
- Once you are familiar with the workflow, have a look at the published workflow run, decompose the workflow in the different modules and test them separately.
- In order to test the different steps of the workflow, you can access <http://wind.isi.edu/marbles/> and import the "Drugome" domain. Different workflows already encode the sub parts of the workflow and are ready to be run. Note that you will need to ask for a user and password (contact Yolanda Gil, Varun Ratnakar or Daniel Garijo).
- You will find suitable inputs to test the workflow in the input data section of this document. If you reuse the inputs, the outputs should look like the ones described here.

Latest version

Draft (of version 2)

You are previewing this dataset. You can share this preview privately or publish it to share it with the world.

Reserved DOI:

doi:10.17632/r69mvkckmn.2

Cite this dataset

Kinnings, Sarah L.; Xie, Li; Fung, Kingston H.; Jackson, Richard M.; Xie, Lei; Bourne, Philip E. (2016), "TB-Drugome - Run 126", Mendeley Data, v2

<http://dx.doi.org/10.17632/r69mvkckmn.2>

DOI is reserved but not active

Previous versions

Version 1

2016-09-09

Link back to
protocols

Create DOI for
Citation

Keep track of
versions of
dataset

Related links

<https://www.hivebench.com/notebooks/8524/experiments/20562>

entity is source of this dataset ⓘ

Copy of the dataset sent to long-term archive

[HOME](#)[REGISTER](#)[LOG IN](#)

EASY

Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal (published by Brill)

For more info: brill.com/rdj

EASY offers sustainable archiving of research data and access to thousands of datasets.

[SEARCH](#)[➤ Search help](#)[➤ Advanced search](#) [➤ Browse](#)

TB-DRUGOME - RUN 126

[⏪ Back to list](#)[Overview](#)[Description](#)[Data files \(13\)](#)

Cite as:

Dataset indexed by data search engine for easy discoverability

DataSearch

drugome

Filter Results

reset

6 results for *drugome*

Data File Types

☐ Tabular Data (4)
☐ Image (4)
☐ Unknown (3)
☐ Software/Code (3)
☐ Document (3)
☐ Text (2)
☐ File Set (1)

Data Source Types

Data Sources

Date


TB-Drugome - Run 126

IN:drug_key.csv

IN:run_126_workflow.png

OUT:3d45443116d0020...

OUT:fb42c612923ef80...

IN:solved_tb_structure...

IN:smap_properties.txt

IN:tb_protein_info.csv

14 of 14

2ZJT	Rv0005	gyrB	DNA gyrase subunit B
3ILW	Rv0006	gyrA	DNA gyrase subunit A
1W74	Rv0009	ppiA	iron-regulated peptidyl-prolyl cis-trans isomerase A
1MRU	Rv0014c	pknB	transmembrane serine/threonine-protein kinase B PKNB (protein kinase B) (STPK B)
1O6Y	Rv0014c	pknB	transmembrane serine/threonine-protein kinase B PKNB (protein kinase B) (STPK B)
2FUM	Rv0014c	pknB	transmembrane serine/threonine-protein kinase B PKNB (protein kinase B) (STPK B)
3F61	Rv0014c	pknB	transmembrane serine/threonine-protein kinase B PKNB (protein kinase B) (STPK B)

Go to data source

IN:tb_protein_info.csv

TABULAR DATA (non-matching)

CRIS configured to import datasets from researchers affiliated to institution and related to projects, equipment, grants,...

Dataset - Pure 5.10.1 - Mozilla Firefox

https://rdm.pure.elsevier.com/admin/editor/dk/atira/pure/modules/datasets/external/model/dataset/editor/dataseteditor.xhtml?id=9268938

Id: 9268938

TB-Drugome - Run 126
Dataset

EDIT

Metadata

OVERVIEW

Relations

Display

HISTORY AND COMMENTS

History and comments

Identification

Title *

TB-Drugome - Run 126

Description

A run of the TB-Drugome Workflow.

Date of data production

☒ Specific date

Year * Month Day

2016

☐ Period of time

People

People *

Sarah L. Kinnings, Contributor
Person

Li Xie, Contributor
Person

Kingston H. Fung, Contributor
Person

Richard M. Jackson, Contributor
Person

Lei Xie, Contributor
Person

Relations to other content

Projects

Tuberculosis Drugome
Project: Research project

Equipment

High Performance Computer Cluster
Equipment/facility: Equipment

Research outputs

The Mycobacterium tuberculosis drugome and its polypharmacological implications
Research output: Research - peer-review > Article

Activities

Last saved: 15/11/17 9:07 PM
Import source: Mendeley Data

Status: Validated

Save

Showcasing outputs in the CRIS web portal

[Home](#) [Profiles](#) [Research Units](#) [Equipment](#) [Research Output](#) **[Datasets](#)** [Projects](#)



TB-Drugome - Run 126

SARAH L. KINNINGS (CONTRIBUTOR), LI XIE (CONTRIBUTOR), KINGSTON H. FUNG (CONTRIBUTOR), RICHARD M. JACKSON (CONTRIBUTOR), LEI XIE (CONTRIBUTOR), PHILIP E. BOURNE (CONTRIBUTOR)

Dataset

Description

A run of the TB-Drugome Workflow.

DATE MADE AVAILABLE

9 Sep 2016

PUBLISHER

University of California San Diego

Related Content

Projects

[Tuberculosis Drugome](#)

Research Output

[The Mycobacterium tuberculosis drugome and its polypharmacological implications](#)

Equipment

[High Performance Computer Cluster](#)

DOI

 [10.17632/r69mvkckmn.1](#)

[Access Dataset](#)

Reporting on projects, grants, datasets and publications

5. Awards

Listing of Awards

Fundings > Funding organisation	Title	Managing organisational unit	Fundings > Awarded amount	Actual start/end date > Start Date
European Commission	TB Drugome grant	Department of Electronics, Information and Bioengineering (DEIB)	€200,000.00	3/4/17

1. Datasets

Analysis of Count on Datasets

Organisations > Organisational unit	Projects > Project	Fundings > Funding organisation	Date made available > Year	Count
Politecnico di Milano	MPH2020: Tuberculosis Drugome	European Commission	2016	1
			 2017	2
Total	-	-		3

3. Research output

Analysis of Count on Research output

Authors > Organisations > Organisational unit	Projects > Project	Fundings > Funding organisation	Current publication status > Date > Year	Count
Politecnico di Milano	MPH2020: Tuberculosis Drugome	European Commission	2016	1
			 2017	3
Total	-	-		4

Reporting: publications and related datasets from funded projects

Projects > Project	Title of the contribution in original language	Current publication status > Date	Authors	Journal > Journal	Electronic version(s) of this work > DOI (Digital Object Identifier)	Datasets > Dataset
MPH2020: Tuberculosis Drugome	CCM: Controlling the change magnitude in high dimensional data	2017	Cesare Alippi, Giacomo Boracchi		10.1007/978-3-319-47898-2_23	Juicebox sample data Dataset
.....						
■ ■ ■ ■ ■ ■ ■						
	The Mycobacterium tuberculosis drugome and its polypharmacological implications	2016	Donatella Sciuto	PLoS Computational Biology	10.1371/journal.pcbi.1000976	TB-Drugome - Run 126 Dataset

Working with the community

Elsevier is:

- Participating in and leading working groups
- Implementing and testing new standards



Conclusions

Data travels across different platforms and users from the point of creation to that of dissemination and recording

Data changes during the process and additional (meta)data (metrics) are generated that are useful for tracking / evaluation, etc.

The integration between the platforms should be as seamless as possible

The admin work needed by researchers should be kept to the minimum

Current Research Information Systems, with their wealth of curated, interconnected data can be used together with RDM platforms to increase the FAIRness of data

The availability of standards and APIs and their adoption by vendors is critical in enabling the creation of a federated system of research information exchange.



Thanks!

a.zigoni@elsevier.com

+39 366 5804425