Hands-on with Research Data Management in Chemistry

Alain Borel

Francesco Varrato

Research Data Team, EPFL Library

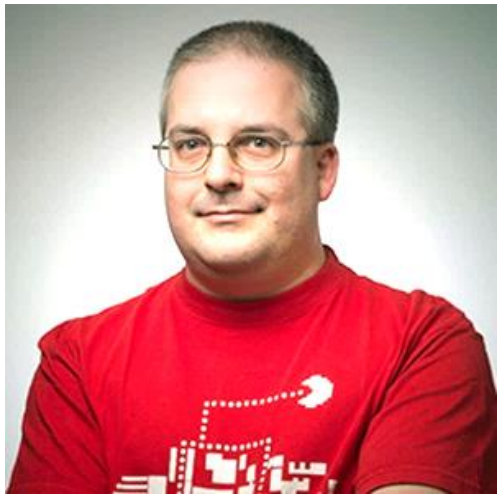
EDCH-RDM, 2025

Program

	2025 – 02 – 25	2025 – 02 – 27	2025 – 03 – 06
Morning 9:00-12:30	1. [1h] **Theory 1**	5. [1h] **Tools 2**	9. [1h] **Theory 3**
	2. [3h] ***Patiny***	6. [3h] ***Dürr***	10. [3h] **Theory 4**
LUNCH BREAK			
Afternoon 13:15-17:00	3. [1h] **Tools 1**	7. [3h] **Hands-on: Data workflow**	11. [3h] **Hands-on: Project / Report**
	4. [3h] **Theory 2**	8. [1h] **Feedback 1**	12. [1h] **Feedback 2**

+ **short breaks** every 1-2 hours

Us



Research Data Management Specialist

RDM Library Team

PhD in Chemistry

https://orcid.org/0000-0003-3268-3195

Research Data Management Specialist

RDM Library Team

PhD in Physics

https://orcid.org/0000-0002-0983-0831

+ External interventions

You?

1. **Name**
2. **When** did you start your PhD?
3. **What** do you expect from this course?

1 minute
live self-introduction

Alain Borel, Francesco Varrato

Main goals and topics

- To provide you with useful information, tools and other resources for your RDM
- To understand the costs and benefits of planning your RDM
- Hands-on a.k.a. exercises on planning, management, processing and publication
- Examples to follow or errors to avoid
  - Open Science, especially FAIR principles
  - Data Management Plan (DMP)
  - Digital formats for collaborations and reproducibility
  - Organize and document datasets
  - Naming conventions and metadata standards
  - Storage solutions and back-ups
  - Data publication
  - Legal and ethical issues
  - Licensing and reuse of datasets

Moodle

go.epfl.ch/ChE-601

Final evaluation

Report (max 3 pages)

- **Analyze the existing data workflow of your research project**
  Describe your current practice (data collection, processing, analysis, storage, sharing, publication, archiving, ...)

- **Propose improvements using principles & tools presented during the course**
  Identify improvements areas and pain points, list and prioritize actions, ...

Workflow diagram (1 page)

- **Create a diagram for your improved data workflow**
  Integrate the proposed changes

Evaluation criteria

A. **Consistency** of the RDM approach
B. **Completeness** of the report
C. **Completeness** of the data workflow diagram
D. **Coherence** of data workflow representation with the report
E. **Compliance** with instructions (deadline, length, diagram attached)

About this training material

Acknowledgements

This course includes content adapted from previous EPFL continuing education workshops, with contributions by Eliane Blumer, Alain Borel, Simon Leonard Dürr, Kevin Jablonka, Antoine Masson, Luc Patiny, Jessica Pidoux, Mathilde Panes, Sitthida Samath and Francesco Varrato.

Course material

Available on Moodle : go.epfl.ch/ChE-601

Acronyms and where to find them

A	**ACS**	**American Chemical Society**
	API	Application Programming Interface
C	CC	Creative Commons
	CYD	Cyber Defence
D	DCC	Digital Curation Centre
	DDI	Data Documentation Initiative
	DLCM	Digital Lifecycle Management
	DMP	Data Management Plan
	DOI	Digital Object Identifier
	DP	Data Protection
E	ELN	Electronic Lab Notebook
	EOSC	European Open Science Cloud
	ERC	European Research Council (Funding agency)
F	FADP	Federal Act on Data Protection
	FAIR	Findable, Accessible, Interoperable, and Reusable
G	GDPR	General Data Protection Regulation
	GIT	Global Information Tracker (when it works properly); or Goddamn Idiotic Truckload of sh*t (when it doesn't work properly)
H	HRA	Human Research Act
I	IPR	Intellectual Property Rights
	ISO	International Standard Organization
	IT	Information Technologies
L	LIMS	Laboratory Information Management System
N	NAS	Network Attached Storage
	NCCR	National Centres of Competence in Research
O	OA	Open Access
	OECD	Organization for Economic Co-operation and Development
	ORCID	Open Researcher and Contributor ID
	ORD	Open Research Data
P	PID	Persistent Identifier
	PPDP	Privacy Preserving Data Publishing
R	RDA	Research Data Alliance
	RDM	Research Data Management
S	SDMS	Scientific Data Management System
	SNSF	Swiss National Science Foundation (Funding agency)
U	URL	Uniform Resource Locator
	URN	Universal Resource Name

A	**AREC**	**Animal Research Ethics Committee**
D	DPO	Data Protection Officer
	DSI	(former VPSI) Information Systems Management Department (Domaine des Systèmes d'Information)
H	HREC	Human Research Ethics Committee
O	OSSC	Open Science Strategic Committee
R	REO	Research Office
S	SDSC	Swiss Data Science Center
	SCITAS	Scientific IT and Application Support
	SISB	(i.e. Library) Scientific information and libraries (Information scientifique et bibliothèques)
T	TTO	Technology Transfer Office

on Moodle

go.epfl.ch/ChE-601

RDM is the new normal: don’t get left behind

Alain Borel, Francesco Varrato

→ DIGITAL LIBRARY

→ RECORD MANAGEMENT

→ ELN

Images:
- commons.wikimedia.org/wiki/File:Lab_Notebook.jpg
- commons.wikimedia.org/wiki/File:Library_Book_Texture.jpg
- commons.wikimedia.org/wiki/File:Meubles_fiches_Institut_de_France.JPG

What is RDM?

DATA

Factual records: numerical scores, textual records, images, sounds, protocols, **source code**, ...

RESEARCH DATA

Data used as primary sources for scientific research, and commonly accepted in the scientific community to validate research findings (OECD)

OPEN RESEARCH DATA (ORD)

Research data that managed during its **lifecycle** to comply with the **FAIR** principles

RESEARCH DATA MANAGEMENT (RDM)

The care and maintenance of research data during the research cycle (UC Berkeley Library)

Technical, legal & ethical aspects are involved.
Documentation is an essential component

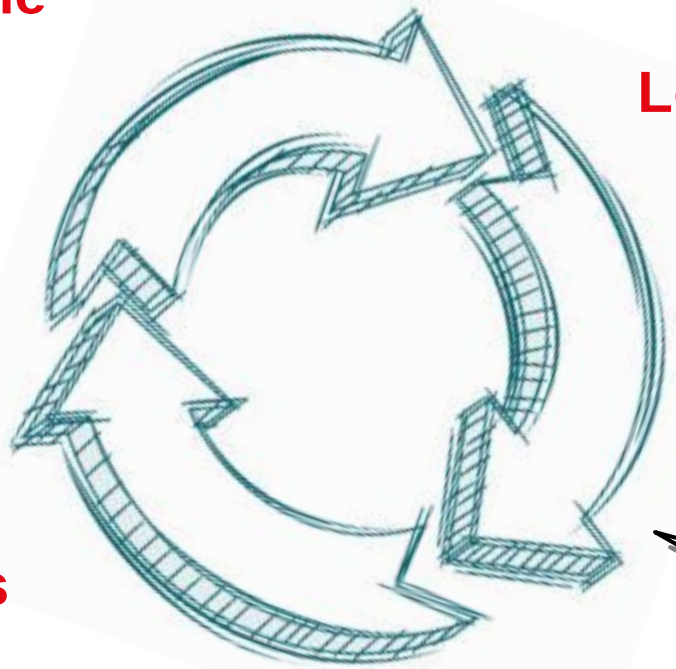
Alain Borel, Francesco Varrato

What is RDM? A risk management perspective

Alain Borel, Francesco Varrato

Wasted time

Legal consequences



Endless updates

Ethics' committee resubmission

Financial costs

Towards sustainability and accountability

Learning and sharing

Who are you and what do you need?

We can't say for sure what each EPFL chemist needs…

Hints from the literature: survey at the Chinese Academy of Science

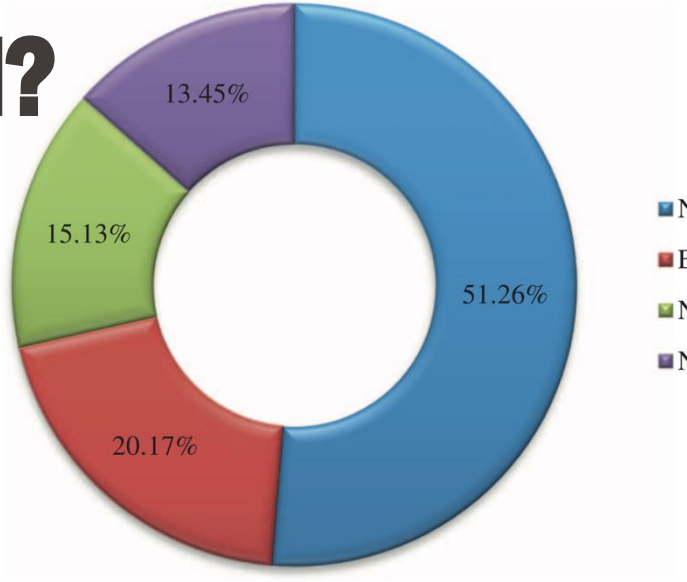


Fig. 16. Needs degree for data management services.

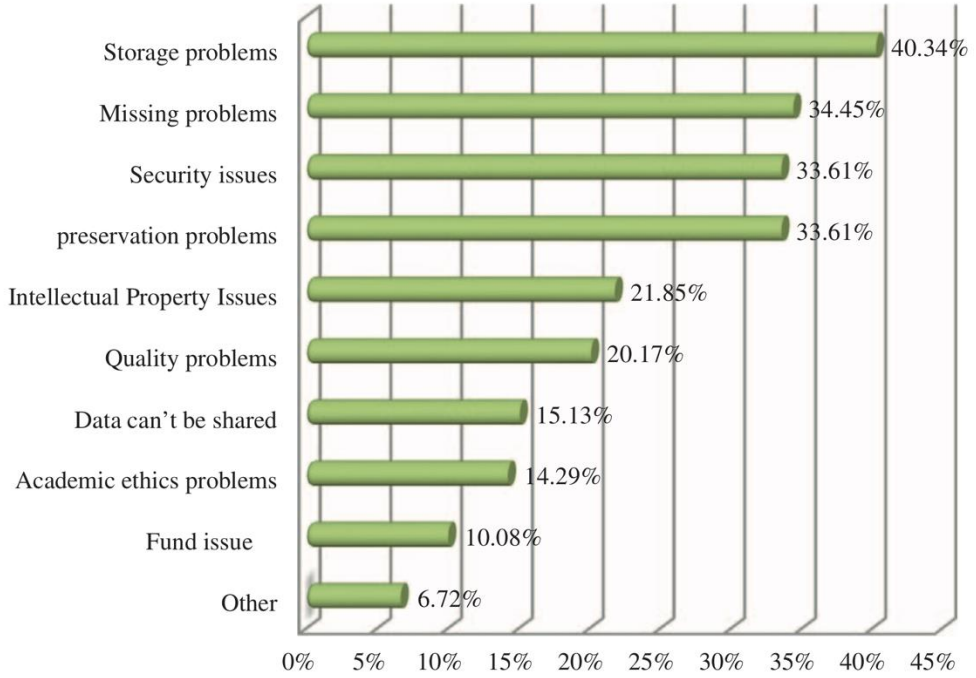


Fig. 15. Main problems of research data management.

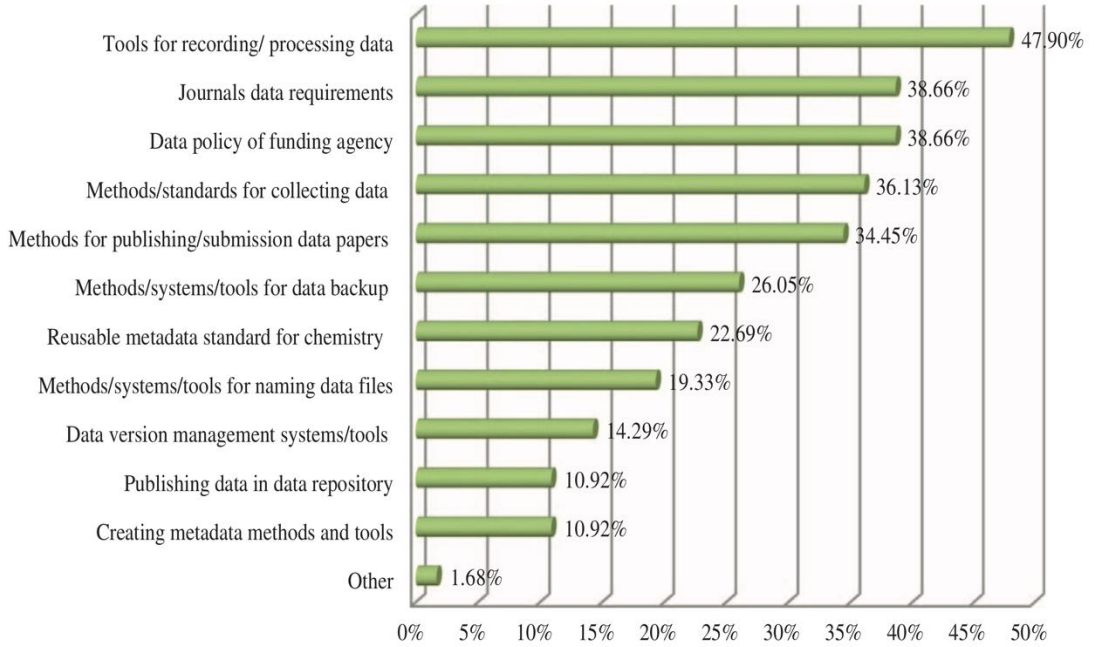


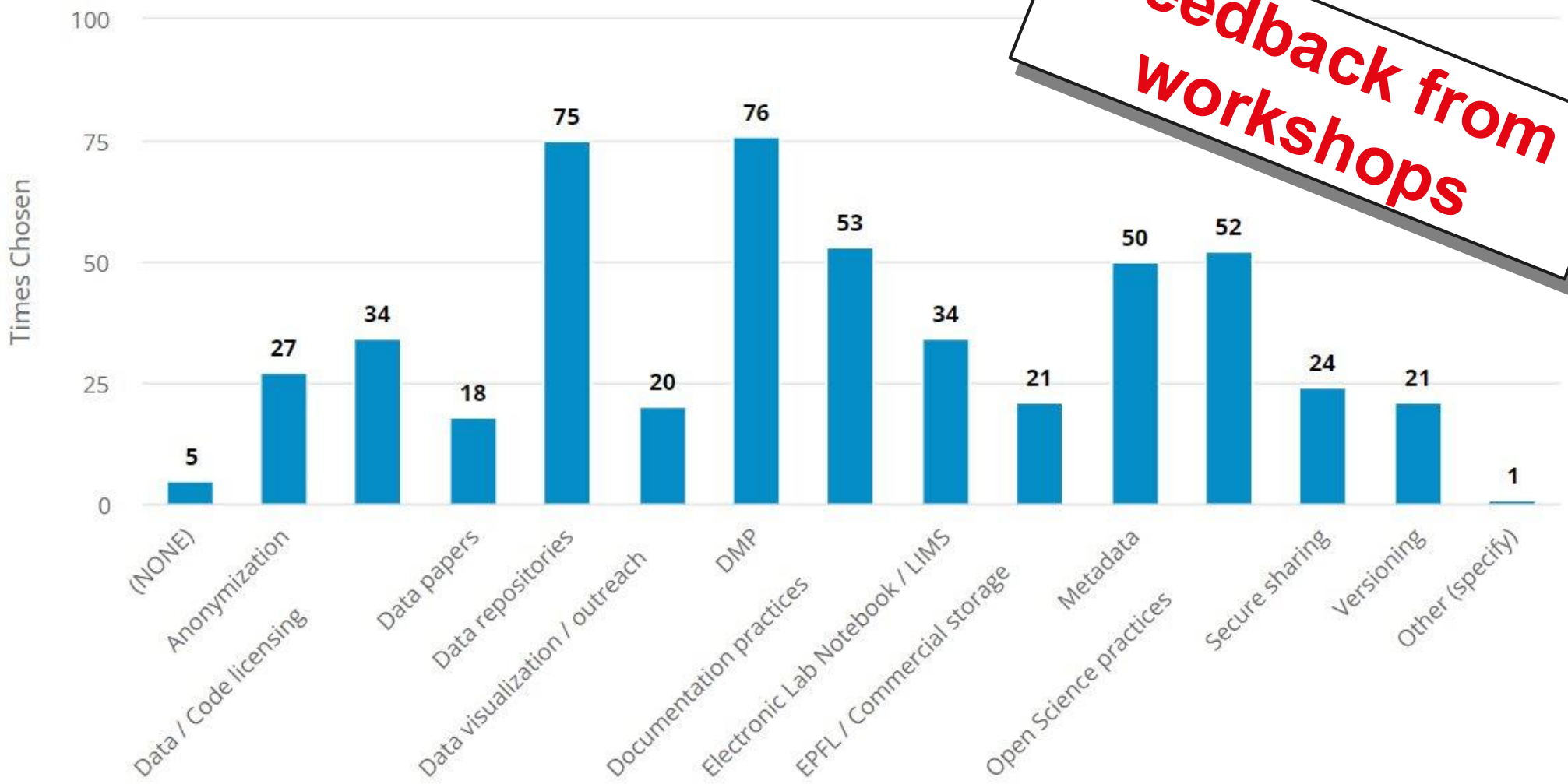
Fig. 17. Services contents respondents want to acquire.

[Chen & Wu, *Journal of Academic Librarianship*, **2017**, DOI 10.1016/j.acalib.2017.06.006]

Alain Borel, Francesco Varrato

Which RDM subject is the most important for you?

Number of responses: 194



Feedback from workshops

Who are you and what do you need?

Data-intensive chemist (usually theoreticians) or “classical” chemist (eg. experimentalists)?

Frequently processing data (experimental or not) or mostly working at the bench?

In charge of infrastructure (IT, instruments, …) or beginner with limited experience?

Etc. …

⇒ Whatever your profile is, talking to you is an **exciting challenge** for us!

Discussion: RDM self-evaluation

Alain Borel, Francesco Varrato

go.epfl.ch/**rdm-self**

Discussion: RDM self-evaluation

Alain Borel, Francesco Varrato

go.epfl.ch/**rdm-self**

Your results

Research Data Lifecycle (experimentalists)

Alain Borel, Francesco Varrato

Creating / Re-using

- Data production
- **Data collection**
- Data sources
- ...

Processing / Analyzing

- Validate data
- **Cleaning data**
- Transform data
- **Analyze data**
- **Interpret data**
- ...

Preserving / Publishing

- Review data
- Convert formats
- Decide IP license
- Depositing data
- Promote data re-use
- …

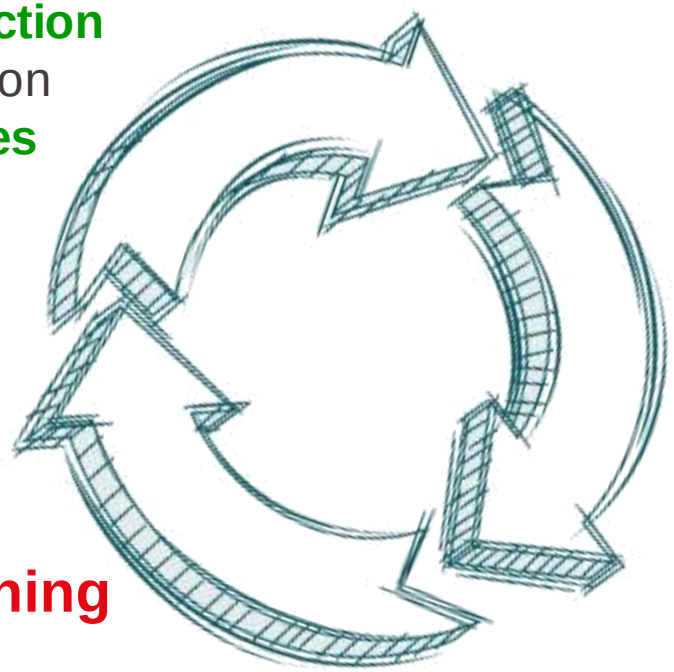
Planning all along

Documenting all along

Research Data Lifecycle (theoreticians)

Creating / Re-using

- **Data production**
- Data collection
- **Data sources**
- ...



Processing / Analyzing

- Validate data
- Cleaning data
- **Transform data**
- **Analyze data**
- **Interpret data**
- ...

Preserving / Publishing

- Review data
- Convert formats
- Decide IP license
- Depositing data
- Promote data re-use
- …

Planning all along

Documenting all along

RDM: why bother? (9h40)

Alain Borel, Francesco Varrato

NYU Health Sciences Library, youtu.be/66oNv_DJuPc

FAIR Principles

[…] the annual cost of not having FAIR research data costs the European economy at least **€10.2bn every year**. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, […], we concluded that these unquantified elements could account for **another €16bn annually on top** of what we estimated.

Intro (pg.4)

[…] at €10.2bn per year in Europe, **the measurable cost of not having FAIR research data makes an overwhelming case in favour of the implementation of the FAIR principles**. […] To top this, figures for the open data economy suggest that the impact on innovation of FAIR could add another €16bn to the minimum cost we estimated.

Conclusions (pg.31)

Source: European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data – Cost of not having FAIR research data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/02999

FAIR Principles

- ***F* indable**
  Data and metadata are easy to find by both humans & computers.
  - Use metadata
  - Deposit (meta)data in repository/registry
  - Assign a persistent identifier (eg. DOI, HANDL, URN)

- ***A* ccessible**
  Machines & humans can readily access or download (meta)data.
  - As-open-as-possible access to your data (licensing, …)
  - Services with user-friendly interfaces
  - Leave the metadata available after data deletion

- ***I* nteroperable**
  Data from different datasets are ready to be exchanged or combined.
  - Use open file format(s), whenever possible
  - Use standardized vocabularies/tags
  - Use cross-linking as much as possible

- ***R* eusable**
  (Meta)data are easily replicated / combined in future research.
  - Attach standardized license to your data (CC, GPL, …)
  - Capture provenance information as precisely as possible

Download our **Fast Guide** on FAIR principles

More from the *GO FAIR Initiative*

F – Examples

Alain Borel, Francesco Varrato

- ***F* indable**
  Data and metadata are easy to find by both humans & computers.

- Provide rich, structured metadata (title, contributors, description, creation year...)
- Get specific DOIs for published datasets
- Cite these DOI in articles that use the data
- Not necessarily the same as for the associated article(s)!

Download our **Fast Guide** on FAIR principles

More from the *GO FAIR Initiative*

A – Examples

- ***A* ccessible**
  Machines & humans can readily access or download (meta)data.

- As-open-as-possible access to your data (licensing, …)
- Use services with user-friendly interfaces
- Leave the metadata available after data deletion

Download our **Fast Guide** on FAIR principles

More from the *GO FAIR Initiative*

I – Examples

- ***I*** **nteroperable**
  Data from different datasets are ready to be exchanged or combined.

- **Use open file format(s), whenever possible**
- Contrary to common wisdom, text is better than images!
- Vocabularies: remember the IUPAC Gold Book?
- Use standard identifiers whenever possible:
  ORCID for contributors, InChi for chemical compounds...

Suggested links:

- Your ORCID iD - your digital name identifier (ORCID)
- InChI and InChIKeys for chemical structures (InChI Trust)
- Compendium of Chemical Terminology (IUPAC Gold Book)

Download our **Fast Guide** on FAIR principles

More from the *GO FAIR Initiative*

R – Examples

- ***R* eusable**
  (Meta)data are easily replicated / combined in future research.

- Standardized license means people will know what they may or may not do with your data
- Data with better provenance information (how/where/when/by whom was it generated?) is easier to verify (perhaps to reproduce), and thus more trustworthy

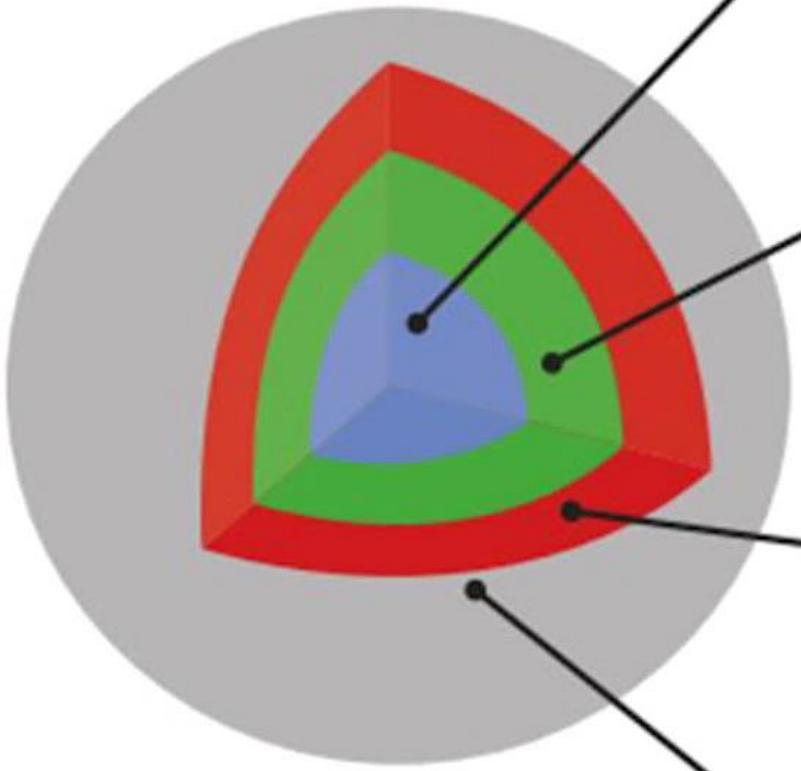
Download our **Fast Guide** on FAIR principles

More from the *GO FAIR Initiative*

Alain Borel, Francesco Varrato

FAIR Principles

Boils down to …



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

Scholarly publications and the FAIR Principles

Source: www.kb.se/openaccess

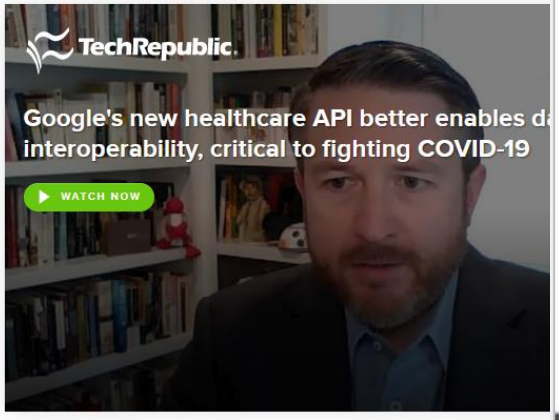
Importance of FAIR principles

Google's new healthcare API better enables data interoperability, critical to fighting COVID-19

www.techrepublic.com/article/googles-new-healthcare-api-better-enables-data-interoperability-critical-to-fighting-covid-19

by Dan Patterson in Big Data on June 29, 2020, 1:32 PM PST

Joe Corkery, Google Cloud product management direct
the new API will help developers scale healthcare solut



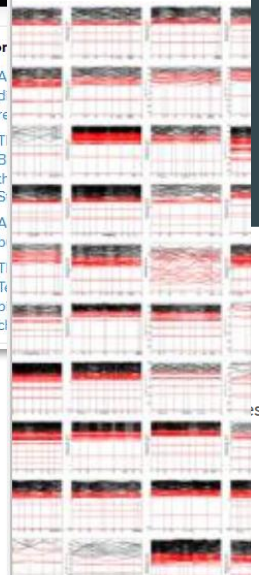
Dan Patterson, senior producer for CNET and CBS News, spoke with Joe Corkery, director of product management, healthcare and life science, Google Cloud, about the use of machine learning in healthcare applications. The following is an edited transcript of their conversation.

Joe Corkery: The Google Cloud Healthcare API is an application, or basically an application layer that we built to enable healthcare data interoperability, to enable healthcare organizations, healthcare application developers, to share a wide variety of different types of healthcare data types. In particular, it's focused on medical record and medical imaging data, supporting DICOM (Digital Imaging and Communications in Medicine) data for medical imaging, as well as

Giovanni Pizzi • 1st
Research Scientist at EPFL (École polytechnique fédérale de Lausanne)
4mo •

Our paper "Automated high-throughput Wannierisation" has been just published! https://lnkd.in/dxGVJPg You can also find a highlight on the NCCR MARVEL website:
https://lnkd.in/dAyRx9Y

I'm proud to have published a fully #FAIR and reproducible paper, with all data shared on the #MaterialsCloud Archive, including also #QuantumMobile virtual machine with #A
materials!) A gre
Jonathan Yates



doi.org/10.1016/S0140-6736(20)31324-6

COMMENT | VOLUME 395, ISSUE 10240, P1820, JUNE 13, 2020

Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Mandeep R Mehra • Frank Ruschitzka • Amit N Patel

Published: June 05, 2020 • DOI: https://doi.org/10.1016/S0140-6736(20)31324-6 • Check for updates

After publication of our *Lancet* Article,[1] several concerns were raised with respect to the veracity of the data and analyses conducted by Surgisphere Corporation and its founder and our co-author, Sapan Desai, in our publication. We launched an independent third-party peer review of Surgisphere with the consent of Sapan Desai to evaluate the origination of the database elements, to confirm the completeness of the database, and to replicate the analyses presented in the paper.

Our independent peer reviewers informed us that Surgisphere would not transfer the full dataset, client contracts, and the full ISO audit report to their servers for analysis as such transfer would violate client agreements and confidentiality requirements. As such, our reviewers were not able to conduct an independent and private peer review and therefore notified us of their withdrawal from the peer-review process.

Discussion: Issues & Services [10’]

ISSUE	WHO
☐ Set-up storage & backup	☐ Data Protection Officer (DPO)
☐ Publishing data / code	☐ Technology Transfer Office (TTO)
☐ Ethical Requirements	☐ Research Office (ReO)
☐ Budgeting for data storage	☐ My own research group
☐ Licensing of data / code	☐ EPFL Ethics Committee
☐ Interpretation of data analysis	☐ EPFL Data Champions
☐ Leak of sensitive data	☐ RDM Library team
☐ Sharing TB of images	☐ Faculty / Central IT

?

[10’]

Alain Borel, Francesco Varrato