Data policy and skills in a rapidly changing world

A CODATA workshop as part of the ISC Global Knowledge Dialogue, Muscat, Oman

Programme

1. Major developments and challenges for data and science in a rapidly changing world, Mercè Crosas, Head of Computational Social Sciences, Barcelona Supercomputing Center and CODATA President (30 mins)
2. Discussion in groups and feedback (30 mins)
3. Break (15 mins)
4. The framework for a policy response: Open Science, FAIR, CARE, WorldFAIR, Simon Hodson, Executive Director, CODATA (15 mins)
5. The challenges in research disciplines: example from Chemistry, Richard Hartshorn, Professor, School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand and CODATA Vice-President (15 mins)
6. The challenges in research disciplines: example from Social Sciences, Steve McEachern, Director, UK Data Service and CODATA Officer (15 mins)
7. What skills do researchers need? Shaily Gandhi, Senior Post-Doctoral Researcher, Geosocial Artificial Intelligence Research Group, Interdisciplinary Transformational University, Linz, Austria and ISC Fellow (15 mins)
8. Discussion in groups. (30 mins)
9. Feedback and wrap up (15 mins)

Major developments and challenges for data and science in a rapidly changing world

Mercè Crosas, Ph.D.

Director of Computational Social Science and Humanities at BSC

President of CODATA

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

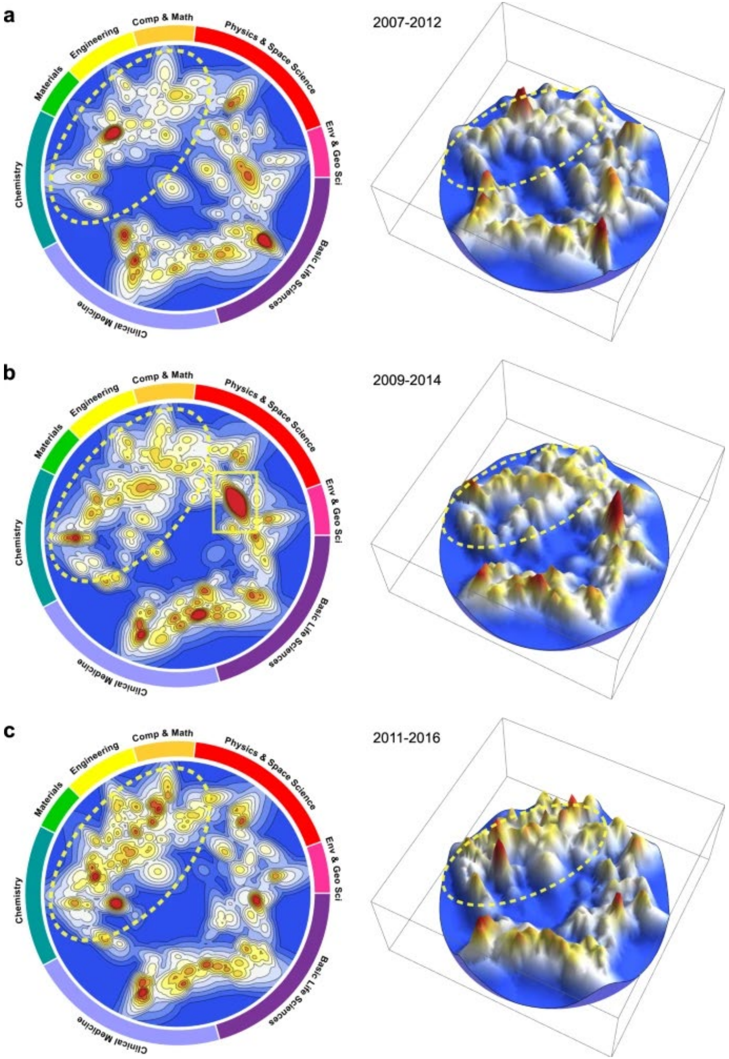
COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Science in a Rapidly Changing world

- More interdisciplinarity, More sectors
  - Interdisciplinary team science for today’s world challenges
  - Partnerships across sectors: industry, government, academia
- More AI, More Data
  - Rapid increase of AI for science
  - Data Challenges

Science in a Rapidly Changing world

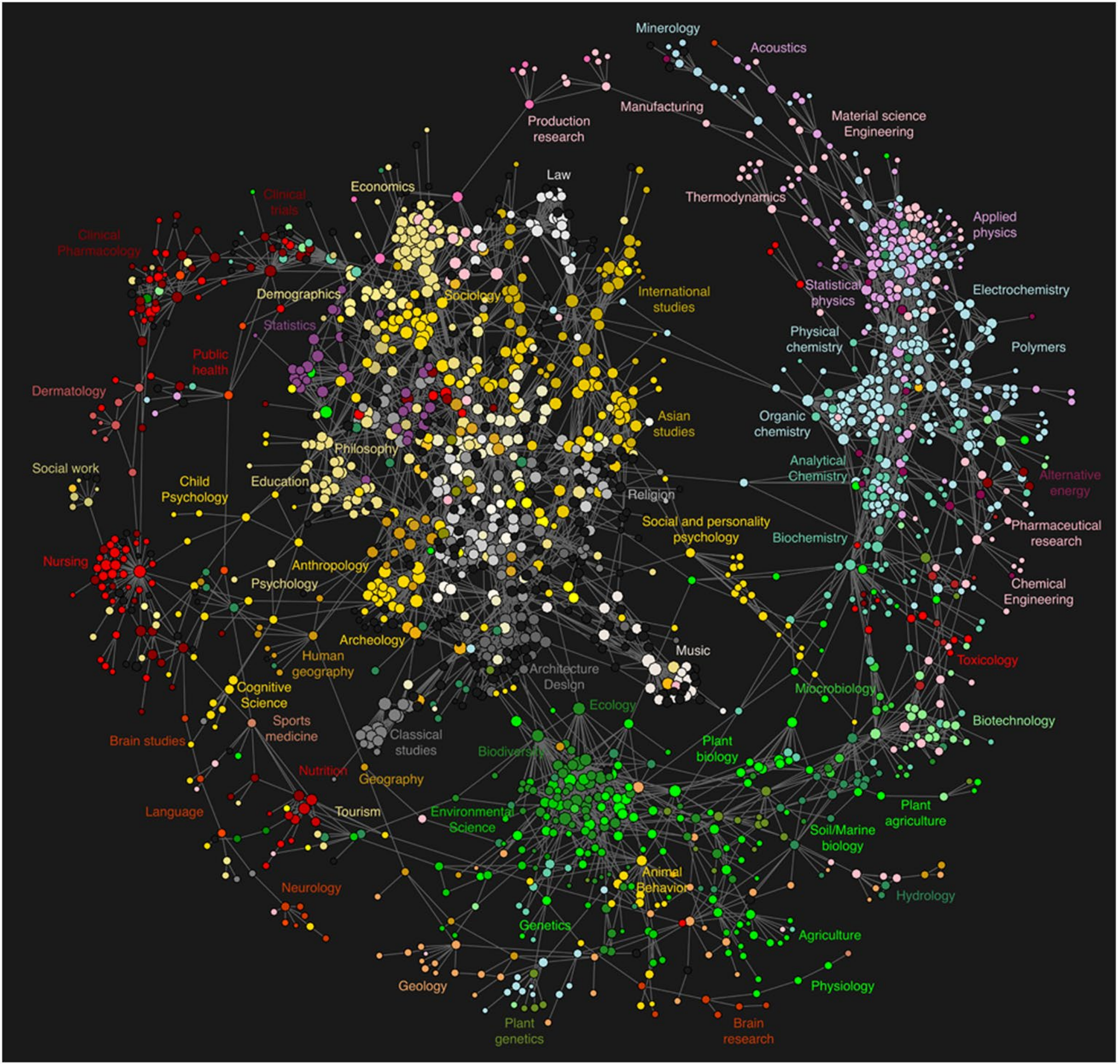
- More interdisciplinarity, More sectors
  - Interdisciplinary team science for today’s world challenges
  - Partnerships across sectors: industry, government, academia
- More AI, More Data
  - Rapid increase of AI for science
  - Data Challenges



The Science Landscape: More interdisciplinarity, more research impact

*"It was found that **an increase by one in the effective number of distinct disciplines** involved in an RF (Research Fronts in natural sciences) was statistically highly significantly associated with an approximately **20% increase in the research impact"***

Okamura, K. (2019). Interdisciplinarity revisited: Evidence for research impact and dynamism. *Palgrave Communications*, *5*(1), 1-9. https://doi.org/10.1057/s41599-019-0352-4



"**Maps of science** (relationship across disciplines) resulting from large-scale **clickstream data** provide a detailed, contemporary view of scientific activity and **correct the underrepresentation of the social sciences and humanities that is commonly found in citation data.**"

Bollen, J., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. Clickstream Data Yields High-Resolution Maps of Science. *PLOS ONE*, *4*(3), e4803. https://doi.org/10.1371/journal.pone.0004803

ISC Science Missions for Sustainability

Beyond interdisciplinarity:

“The myriad issues stemming from unsustainable development necessitate collaboration among the scientific community, funders, policymakers, tech firms, and philanthropists.” *Salvatore Aricó*

STATISTICAL PARTNERSHIPS

When academia meets industry meets government

John Kolassa interviews Ridha Ben Mrad, Nancy Reid and Dave Campbell about the Canadian experience of building statistical partnerships between academia, industry and government

“Twenty years ago, **collaborations between academia and companies** were largely relegated to business schools and some applied sciences projects. Today, demand for these collaborations is so strong that we are on track to deliver 10,000 internships per year – **in every discipline**, **ranging from social sciences and humanities to medicine and applied sciences.** “

Mathematics of Information Technology and Complex Systems (Mitacs)

Canadian Statistical Sciences Institute (CANSSI)

Kolassa, J. (2020). When academia meets industry meets government. *Significance*, *17*(5), 44-45. https://doi.org/10.1111/1740-9713.01453

SOCIAL SCIENCE ONE

Hosted by Harvard's Institute for Quantitative Social Science

Building Industry-Academic Partnerships

Academic social scientists have more data than ever before to study human society, but a smaller proportion of existing data than at any time in history because most of it is now tied up inside companies.

Through information sharing, demonstration projects, collaborations, and other activities, we seek to reduce barriers to industry collaboration, unlock commercial information for public good in privacy protective ways, and enable both academics and companies to advance their separate and joint goals.

“Academic social scientists have more data than ever before to study human society, but a smaller proportion of existing data than at any time in history because most of it is now tied up inside companies.”

https://socialscience.one/

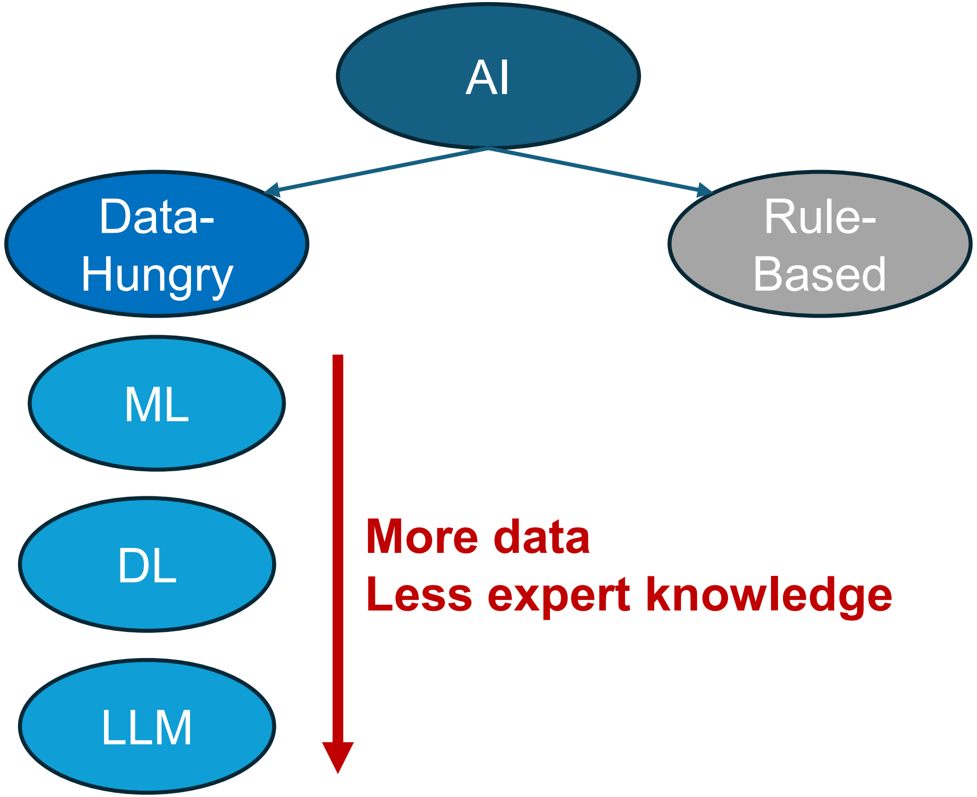
Accessing private data for Science: On-going opportunities, difficult in practice

- **Partnerships:** Industry-Academy partnerships, collaboration with governments
- **Policies**: European Regulations open possibilities:
  - Data access for researchers with Digital Service Act (DSA), in theory
    - “Providers of very large online platforms or of very large online search engines shall give access without undue delay to data, including, where technically possible, to real-time data, provided that the data is publicly accessible in their online interface by researchers”
  - Data donations (“altruism”) observed in Data Governance Act (DGA)
    - “Article 16 (1) DGA allows Member States to establish national strategies and organisational and/or technical provisions to facilitate data altruism.”
- **Automation:** standards and tools that make more efficient repeatable tasks

Science in a Rapidly Changing world

- More interdisciplinarity, More sectors
  - Interdisciplinary team science for today´s world challenges
  - Partnerships across sectors: industry, government, academia
- More AI, More Data
  - Rapid increase of AI for science
  - Data Challenges

The growth of data -hungry AI



AI

Data-Hungry

Rule-Based

Bottom-up

Machine Learning

ML

Deep Learning
Neural Networks

DL

Transformers
Large Language Models
Gen AI

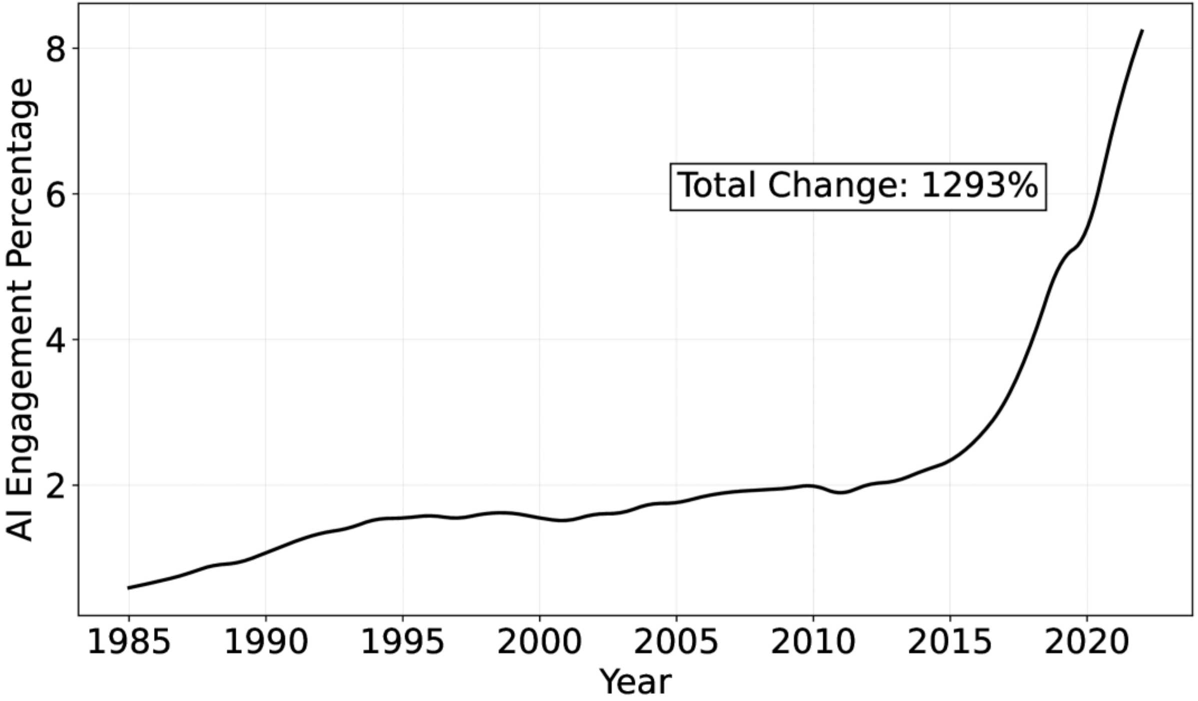
LLM

More data
Less expert knowledge

Top-down

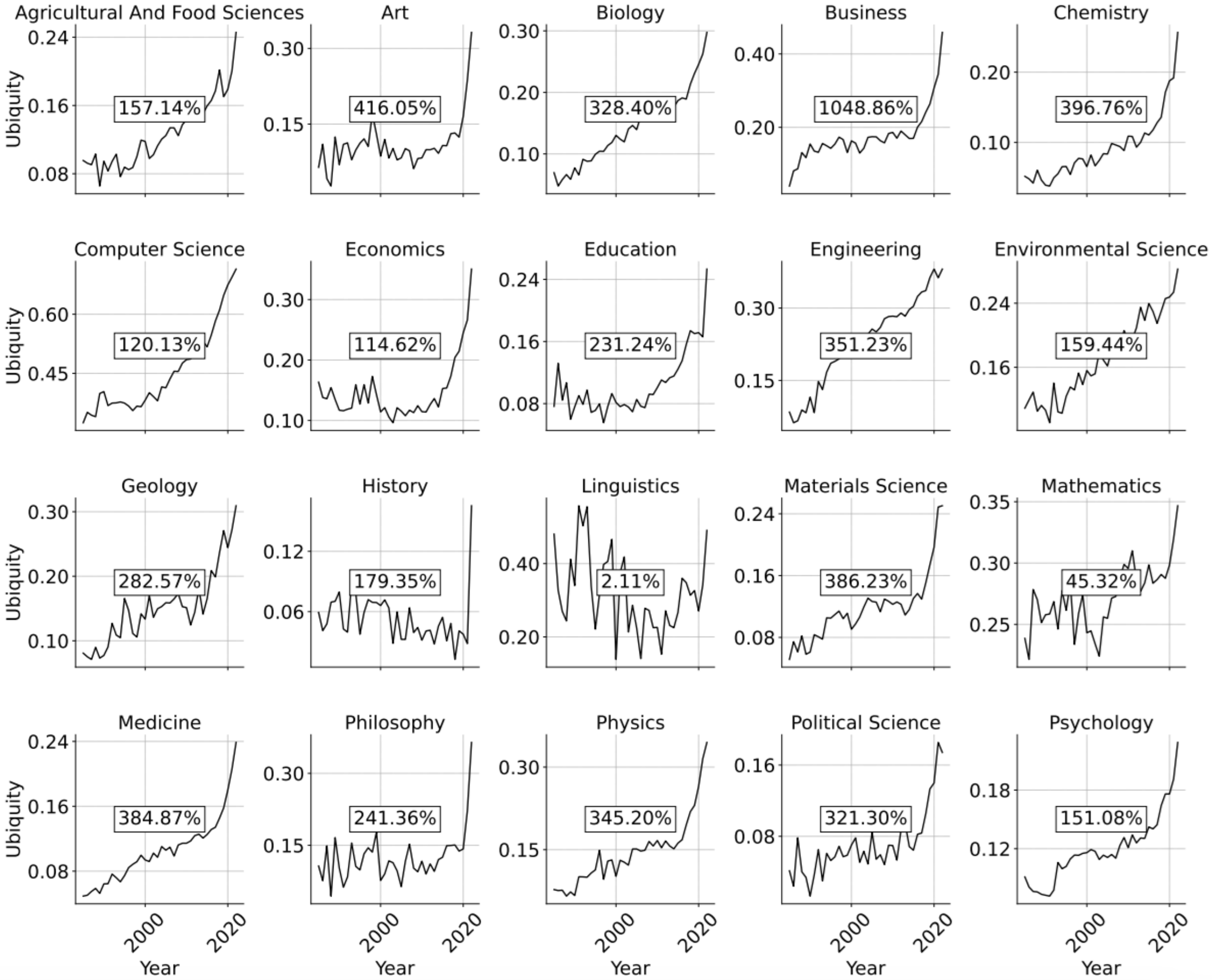
Rules from expert knowledge

Oil or water? Diffusion of AI Within and Across Scientific Fields



“Increase of AI within 80 million research publications across 20 diverse scientific fields, by examining the change in scholarly engagement with AI from 1985 through 2022”

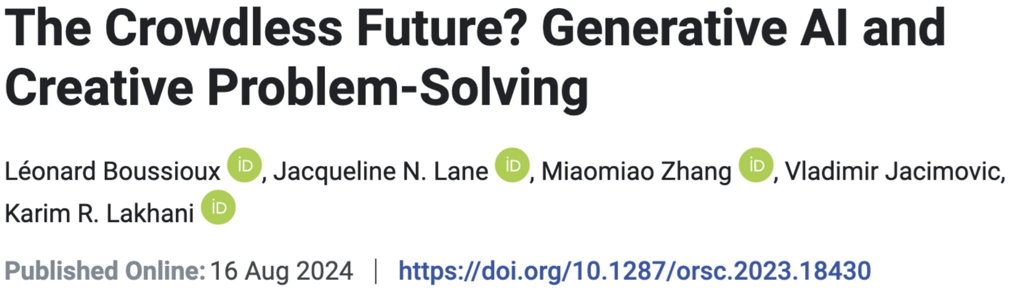
Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). Oil & Water? Diffusion of AI Within and Across Scientific Fields. *ArXiv*. /abs/2405.15828



“the trajectory of the ***rate at which AI engagement is becoming more ubiquitous within fields in the last decade is striking*** *(see Figure) with every field in our corpus experiencing a rapid increase in the diffusion of AI-engaged research across their publication venues.”*

Future of AI in Science?

- More “AlphaFolds” :
  - AlphaFold was possible thanks to the Protein Data Bank (PDB, 1971)
  - New foundation models for science require specialized, good-quality, curated data (e.g., foundation model for astronomy data, climate data)
- AI in the loop:
  - Gen AI for creative problem solving
  - Gen AI for social science experiments



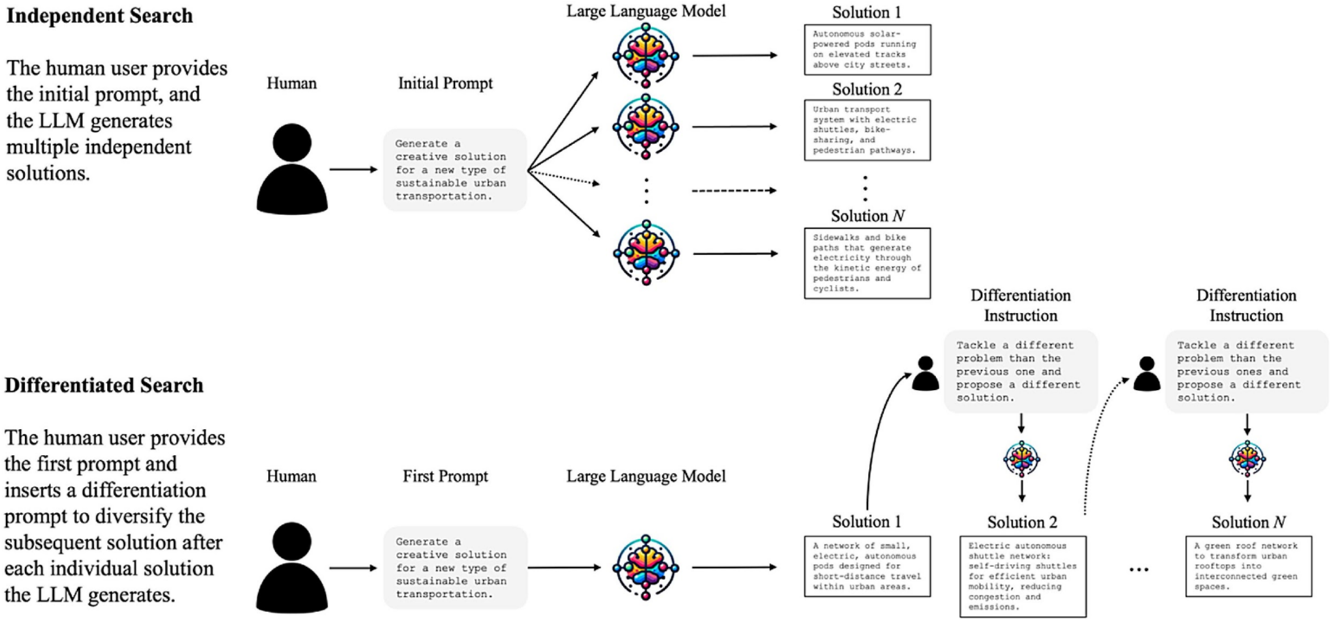
The Crowdless Future? Generative AI and Creative Problem-Solving

Léonard Boussioux, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, Karim R. Lakhani

Published Online: 16 Aug 2024 | https://doi.org/10.1287/orsc.2023.18430

“Human in the loop”

“AI in the loop”



Independent Search

The human user provides the initial prompt, and the LLM generates multiple independent solutions.

Generate a creative solution for a new type of sustainable urban transportation.

Solution 1: Autonomous solar-powered pods running on elevated tracks above city streets.

Solution 2: Urban transport system with electric shuttles, bike-sharing, and pedestrian pathways.

Solution *N*: Sidewalks and bike paths that generate electricity through the kinetic energy of pedestrians and cyclists.

Differentiated Search

The human user provides the first prompt and inserts a differentiation prompt to diversify the subsequent solution after each individual solution the LLM generates.

Generate a creative solution for a new type of sustainable urban transportation.

Tackle a different problem than the previous one and propose a different solution.

Tackle a different problem than the previous ones and propose a different solution.

Solution 1: A network of small, electric, autonomous pods designed for short-distance travel within urban areas.

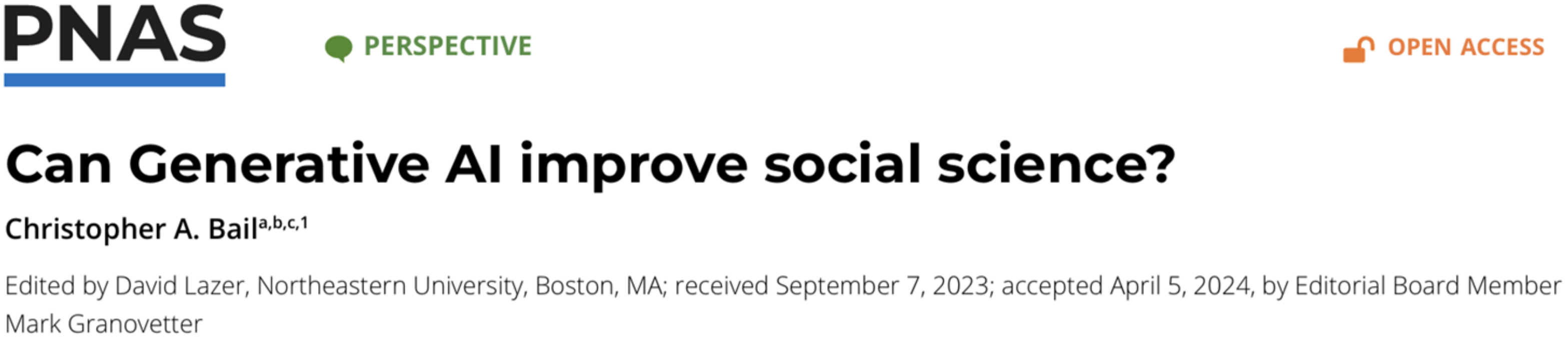
Solution 2: Electric autonomous shuttle network: self-driving shuttles for efficient urban mobility, reducing congestion and emissions.

...

Solution *N*: A green roof network to transform urban rooftops into interconnected green spaces.

“... paradigm shifts underscores the **importance of thoughtfully delineating responsibilities between human and AI collaborators** while recognizing them as complementary components within an interactive system.”

Gen AI for social science experiments



PNAS PERSPECTIVE OPEN ACCESS

Can Generative AI improve social science?

Christopher A. Bail[a,b,c,1]

Edited by David Lazer, Northeastern University, Boston, MA; received September 7, 2023; accepted April 5, 2024, by Editorial Board Member Mark Granovetter

“Generative AI has the potential to improve survey research, online experiments, automated content analyses, agent-based models, and other techniques commonly used to study human behavior.”

BUT “bias in the data used to train these tools can negatively impact social science research”

Data Challenges in a Rapidly Changing World (1)

- **Challenges and Opportunities from More interdisciplinarity, More sectors:**
  - Describe datasets for others not in your field:
    - Standard metadata for cross-domain interoperability (e.g., FAIR, CDIF)
  - Data from industry and governments could be very useful for scientific research:
    - Collaborations, efficient processes and policies, standards to use data (e.g., ODRL)
  - In social and health sciences, data harmonization require high-resolution, sensitive data:
    - Tools to preserve privacy without losing utility (e.g., differential privacy)

Data Challenges in a Rapidly Changing World (2)

- **Challenges and Opportunities from More AI, More Data:**
  - New foundation models for scientific purpose
    - Specialized, well-curated data for pre-training and training or fine-tuning models
    - AI-ready data with metadata standards (e.g., CROISSANT, FAIR)
  - The use of Gen AI in science ought to be scientific and responsible:
    - Transparent, reproducible, rigorous
    - Open, explainable algorithms
    - Open data and/or metadata
    - Proper use of copyright, personal data

What we need to be ready

- **Policy** interventions to make it work
- **Skills** to be ready for the increased use of data and AI
- **Cross-disciplinary** tools and standards
- Specialized adaptation for change in **each scientific domain**

Discussion:
Major developments in data and science

The framework for a policy response: Open Science, FAIR, CARE, WorldFAIR

Simon Hodson, CODATA Executive Director

Open Science

- Open science maximises the access to, the participation in science and its benefits.
- **Open:** “Access to scientific knowledge [including data, code etc] should be **as open as** possible. … Access restrictions need to be **proportionate** and **justified**.”
- **Protection:** Essential to protect sensitive data and information, including: the “protection of human rights, national security, confidentiality, the right to privacy and respect for human subjects of study, legal process and public order, the protection of intellectual property rights, personal information, sacred and secret indigenous knowledge, and rare, threatened or endangered species.”
- Sets out core **values** and guiding **principles**.
- Describes an action plan, including creating a common understanding, developing policy, building capacity, providing enabling infrastructure.

UNESCO Recommendation on Open Science

Simon Hodson, CODATA Executive Director, was co-chair of the Expert Advisory Group

UNESCO Recommendation, 2021: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

The Rs of Open Science

- **Rigorous and reproducible:** sharing data, code, methodologies, protocols to ensure rigour and reproducibility.
- **Responsible and respectful:** responsible in the conduct of science, respectful of subjects.
- **Relevant:** maximise public access to the outputs of science; maximise stakeholder participation in science.
- **Return on Investment:** maximise the economic benefit and the societal impact of science.

COMMITTEE ON DATA

CODATA

INTERNATIONAL SCIENCE COUNCIL

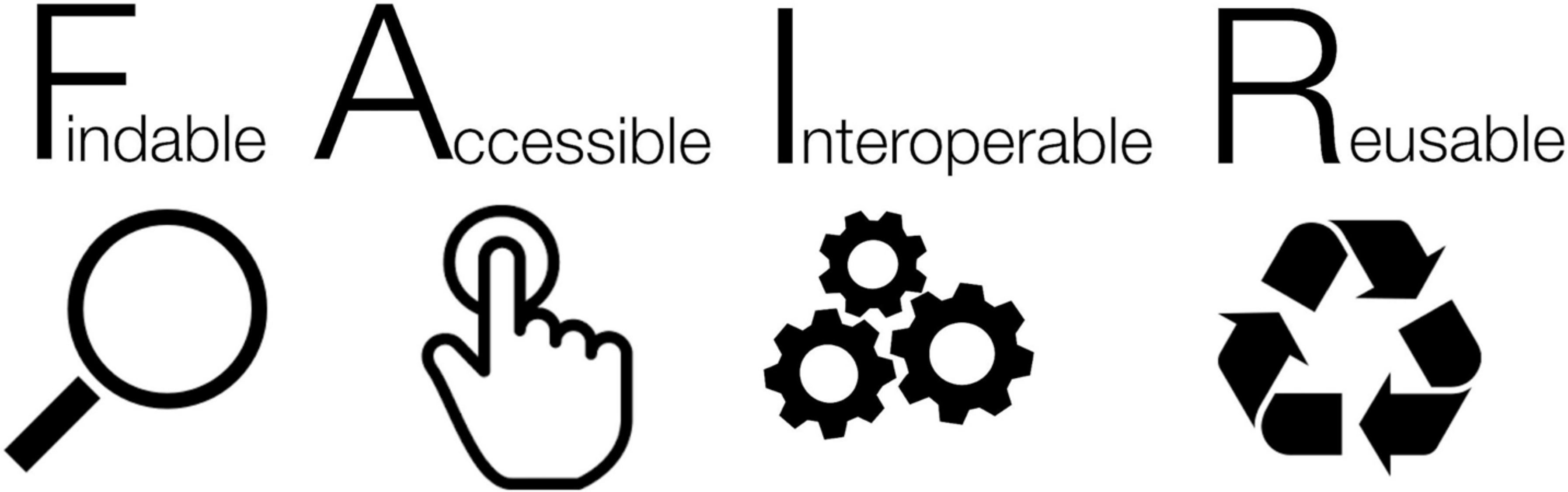
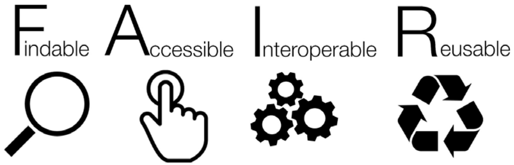


Image CC-BY-SA by SangyaPundir

(Wilkinson, M., et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, http://dx.doi.org/10.1038/sdata.2016.18)

FAIR Principles

- **FAIR: encompasses in an easy communicable acronym, high level principles of good data stewardship**
  - Increases the usability and utility of data, metadata, code.
  - Extremely influential (6519 citations Nature; 15595 citations Google Scholar).
- **Emphasis of the benefits of machine-actionability: network of FAIR data**
  - FAIR principles designed to support the use of data at scale, by machines, harnessing technological potential, better enabling AI.
  - **Vision of harnessing the technologies of the web, to improve querying of vast, dispersed and heterogenous data.**
- **Increases the value of data for science and the economy**
  - PWC report, 2019: **Opportunity cost to the European science system of NOT having FAIR data: 8.2 Bn Euros.**
  - (at least) **80% of project effort goes into downstream ‘data wrangling’, rather than upstream ‘data stewardship’.**



Wilkinson, Mons, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, http://dx.doi.org/10.1038/sdata.2016.18

Barend Mons and Mercè Crosas, past and current CODATA Presidents, both authors of the FAIR Principles.

Final Report and Action Plan from the European Commission Expert Group on FAIR Data
TURNING FAIR INTO REALITY
2018

Turning FAIR Into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018, Hodson (chair of working group/lead author), et al., https://doi.org/10.2777/1524

The CARE principles

- CARE Principles of Indigenous Data Governance https://www.gida-global.org/care (Global Indigenous Data Alliance).
- Carroll, et al. (2020) ‘The CARE Principles for Indigenous Data Governance’, CODATA Data Science Journal, 19(1), https://doi.org/10.5334/dsj-2020-043
    - **Collective Benefit:** Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.
    - **Authority to Control:** Indigenous Peoples’ rights and interests in Indigenous data must be recognised and their authority to control such data be empowered. Indigenous data governance enables Indigenous Peoples and governing bodies to determine how Indigenous Peoples, as well as Indigenous lands, territories, resources, knowledges and geographical indicators, are represented and identified within data
    - **Responsibility:** Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples’ self determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples.
    - **Ethics:** Indigenous Peoples’ rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.
- Specific principles for engaging with data creation with indigenous communities.
- Very prominent in the UNESCO Recommendation on Open Science.

CARE Principles of Indigenous Data Governance
https://www.gida-global.org/care

Data Policy in Times of Crisis

Open Science and FAIR in Specific Circumstances

Joint UNESCO-CODATA Working Group 'Data Policy in Times of Crisis'.

- Prepared a factsheet, guidance and checklist for developing data policies for times of crisis.
- Extensive consultation held from Oct 2024-Jan 2025.
- Webinar held 16 Jan attracted over 200 participants.
- Toolkit components can be accessed from the webinar page: https://bit.ly/UNESCO-CODATA-DPTC-Webinar
- Working Group co-chaired by Ana Persic, UNESCO; Virginia Murray, UKHSA and CODATA; Francis Crawley, chair of CODATA IDPC.

CODATA-WorldFAIR Policy Recommendations: Enabling Global FAIR Data

We need to enable Research Infrastructures to support 21st century science.

There is an urgent need for a shift from a 'bibliographic' data stewardship practice to a data engineering practice!
https://doi.org/10.5281/zenodo.11242702

- We need (international) research infrastructures to aggregate and integrate data for key research areas.
- This task is hampered by the persistence of the bibliographic, publication model for thinking about data.
  - Parsons, M.A. and Fox, P.A., 2013. Is Data Publication the Right Metaphor?. Data Science Journal, 12(0), p.WDS32-WDS46.DOI: https://doi.org/10.2481/dsj.WDS-042
  - The use case in which researchers deposit data in support of their publication is not the only, or the most important, use case.
- Need to enable Research Infrastructure to furnish integrated **data products** for researchers. This is how we reduce the cost of data wrangling and support cross-domain research.

Enabling Global FAIR Data: WorldFAIR Policy Recommendations for Research Infrastructures

This document presents the policy recommendations from the WorldFAIR project. It synthesises the project's findings and presents recommendations for specific stakeholders, for the European Open Science Cloud and other Research Infrastructures globally.

To meet the challenges and opportunities confronting 21st century science, including the need to support interdisciplinary research and the impact of AI, it calls for a shift to a data engineering approach and for investment in metadata uplift and the implementation of the FAIR principles to enable this.

Photo: Mirexon | Shutterstock

Recommendation 1: Data Engineering

Recommendation 1: Data Engineering:

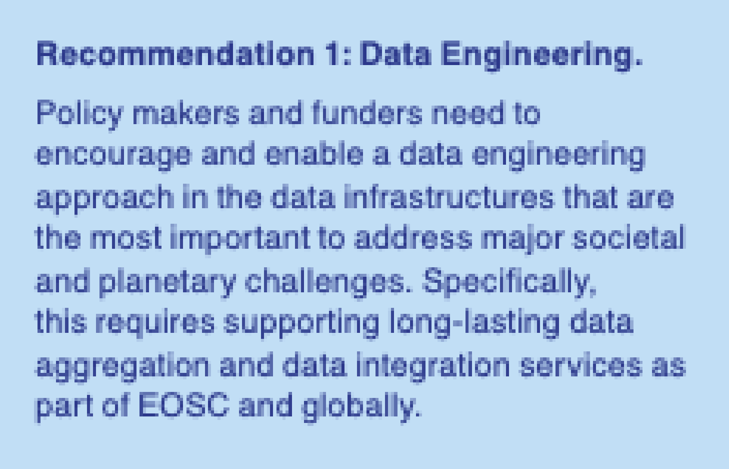
Policy makers and funders need to encourage and enable a data engineering approach in the data infrastructures that are the most important to address major societal and planetary challenges. Specifically, this requires supporting long lasting data aggregation and data integration services as part of EOSC and globally.

Drivers:

- Data integration and integration, particularly for Interdisciplinary research.
- Provenance, processing and data lineage > transparency, reliability, reproducibility.
- Fine-grained data access, automating responsible access to sensitive data.
- Responsible AI.

Good examples…

- Examples of data integration: INSPIRE (WF, WP07) and EOSC Future SP9 (ESS and Climate Data).
- Examples of data federation: ODIS (agreed architecture and KG approach), UN Stats SDGs, SDMX, GBIF new data model.



Recommendation 1: Data Engineering.

Policy makers and funders need to encourage and enable a data engineering approach in the data infrastructures that are the most important to address major societal and planetary challenges. Specifically, this requires supporting long-lasting data aggregation and data integration services as part of EOSC and globally.

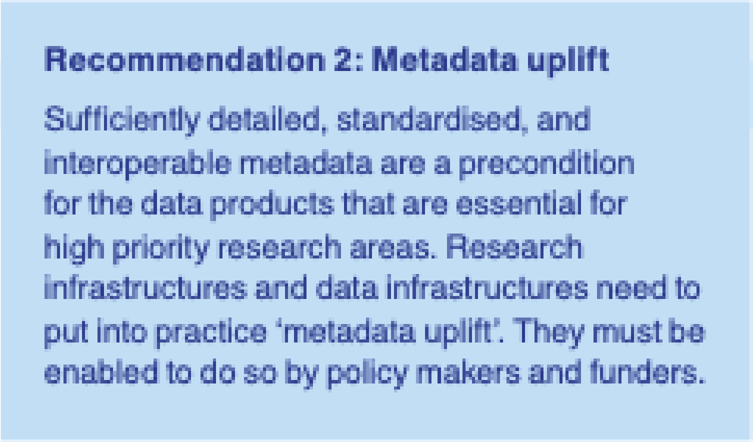
Recommendation 2: Metadata uplift

Recommendation 2: Metadata uplift

Policy makers and funders need to encourage and enable a data engineering approach in the data infrastructures that are the most important to address major societal and planetary challenges. Specifically, this requires supporting long lasting data aggregation and data integration services as part of EOSC and globally.

Drivers:

- An essential component of data engineering is the addition of sufficiently detailed, standardised, and interoperable metadata.
- For data to be (re-)used in research (i.e. compared, combined, analysed) the information about the data needs to be very detailed.
    - Includes what ***may*** and what ***can*** be done; what is scientifically necessary to use the data.
- The increasing scale and complexity of research questions require an increasing scale of data stewardship and 'metadata uplift'.
- **AI can assist; but AI needs good data AND good metadata.**



Recommendation 2: Metadata uplift

Sufficiently detailed, standardised, and interoperable metadata are a precondition for the data products that are essential for high priority research areas. Research infrastructures and data infrastructures need to put into practice 'metadata uplift'. They must be enabled to do so by policy makers and funders.

Recommendation 3-6: Four key drivers

Four key drivers for data engineering and metadata uplift

3: Interdisciplinary research for global challenges.

- Requires data aggregation and integration.

4: Reproducible and transparent research.

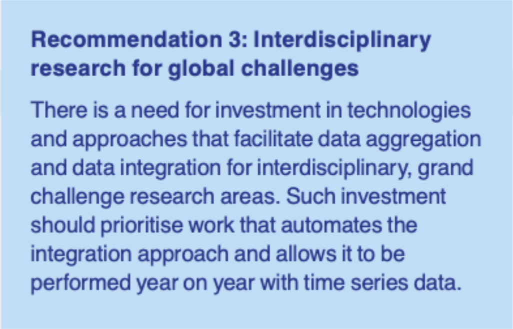
- Requires detailed provenance and processing information.

5: Automated and fine-grained access to sensitive data.

- Essential to increase access, to reduce costs and to increase data security. Requires data engineering.

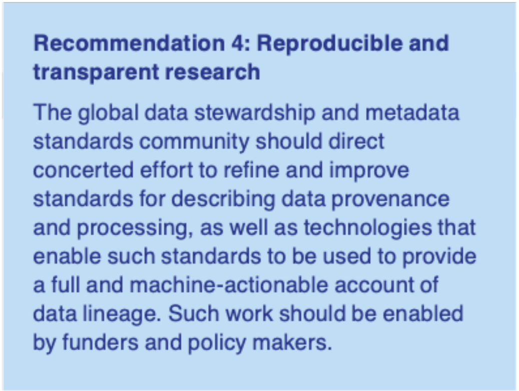
6: Responsible AI (for science)

- Requires high quality data and detailed metadata for training data sets.
- “in AI today, data is the new code… but the AI-ready data infrastructure is missing” (Elena Simperl, ML Commons “Croissant” WG co-chair at BSC-CODATA Conference on Computational Social Science).
- Increasing collaboration and alignment between the ML Commons Croissant initiative and CDIF.



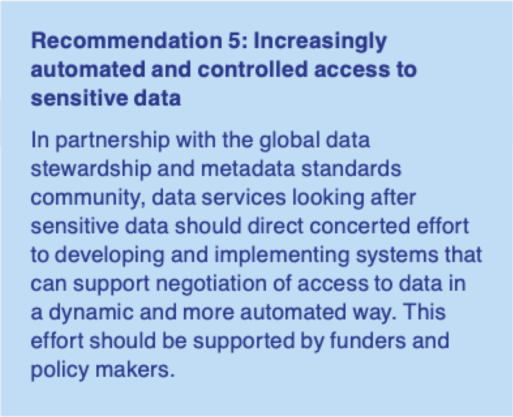
Recommendation 3: Interdisciplinary research for global challenges

There is a need for investment in technologies and approaches that facilitate data aggregation and data integration for interdisciplinary, grand challenge research areas. Such investment should prioritise work that automates the integration approach and allows it to be performed year on year with time series data.



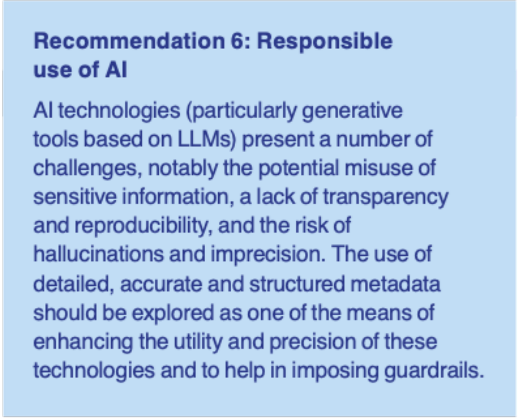
Recommendation 4: Reproducible and transparent research

The global data stewardship and metadata standards community should direct concerted effort to refine and improve standards for describing data provenance and processing, as well as technologies that enable such standards to be used to provide a full and machine-actionable account of data lineage. Such work should be enabled by funders and policy makers.



Recommendation 5: Increasingly automated and controlled access to sensitive data

In partnership with the global data stewardship and metadata standards community, data services looking after sensitive data should direct concerted effort to developing and implementing systems that can support negotiation of access to data in a dynamic and more automated way. This effort should be supported by funders and policy makers.



Recommendation 6: Responsible use of AI

AI technologies (particularly generative tools based on LLMs) present a number of challenges, notably the potential misuse of sensitive information, a lack of transparency and reproducibility, and the risk of hallucinations and imprecision. The use of detailed, accurate and structured metadata should be explored as one of the means of enhancing the utility and precision of these technologies and to help in imposing guardrails.

Recommendation 7-11: Enablers

7: Support the further development of the Cross-Domain Interoperability Framework (CDIF)

8: Invest in Research Infrastructures to enable the transition to a data engineering approach.

9.1: Enable standards bodies and international and representative organisations to develop, maintain and sustain essential data and metadata standards.

9.2: Support WorldFAIR+ as a means of enabling research communities to develop and implement good practice.

9.3: Support the FIPs approach and infrastructure

9.4: Support CODATA and RDA to enable research communities

10: Support the sustainability of semantic artefacts including by means of a survey and analysis of business models, governance and juridical status, and make recommendations to improve the sustainability and effectiveness of the organisations that maintain them.

11: Strengthen international partnerships to develop a global system for data and metadata exchange.

CODATA

WorldFAIR

Enabling Global FAIR Data: WorldFAIR Policy Recommendations for Research Infrastructures

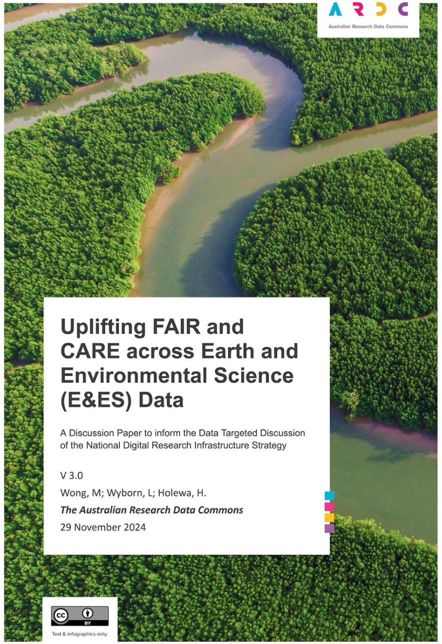
This document presents the policy recommendations from the WorldFAIR project. It synthesises the project's findings and presents recommendations for specific stakeholders, for the European Open Science Cloud and other Research Infrastructures globally.

To meet the challenges and opportunities confronting 21st century science, including the need to support interdisciplinary research and the impact of AI, it calls for a shift to a data engineering approach and for investment in metadata uplift and the implementation of the FAIR principles to enable this.

Photo: Mirexon | Shutterstock

Uplifting FAIR and CARE: CODATA WorldFAIR Policy and Technical Recommendations promoted by ARDC

- Uplifting FAIR and CARE recommends implementing the WorldFAIR policy and technical recommendations for the Australian Planet Data Commons.
- Endorses the WorldFAIR Policy Recommendations.
- Recommends the implementation and further development of CDIF, the Cross-Domain Interoperability Framework.
- Argues that FAIR and CDIF can assist with the implementation of decisions made by indigenous communities in line with the CARE principles.
- Uplifting FAIR and CARE: https://doi.org/10.5281/zenodo.14241825



ARDC

Australian Research Data Commons

Uplifting FAIR and CARE across Earth and Environmental Science (E&ES) Data

A Discussion Paper to inform the Data Targeted Discussion of the National Digital Research Infrastructure Strategy

V 3.0

Wong, M; Wyborn, L; Holewa, H.

The Australian Research Data Commons

29 November 2024

Text & infographics only.

The Challenges in Research Disciplines Examples from Chemistry

Professor Richard Hartshorn
School of Physical and Chemical Sciences
University of Canterbury, Christchurch, New Zealand

Digitally Enabled Workflows

Data Management Plans – coming to a grant application near you

Electronic Lab Notebooks – routinely used in industry

Fully linked spectra and characterisation data

Extended through into publication – being FAIR

Is Your Data FAIR?

Findable

Accessible

Interoperable

Reusable

Findability in Chemistry?

copper acetate

Findability in Chemistry?

copper acetate

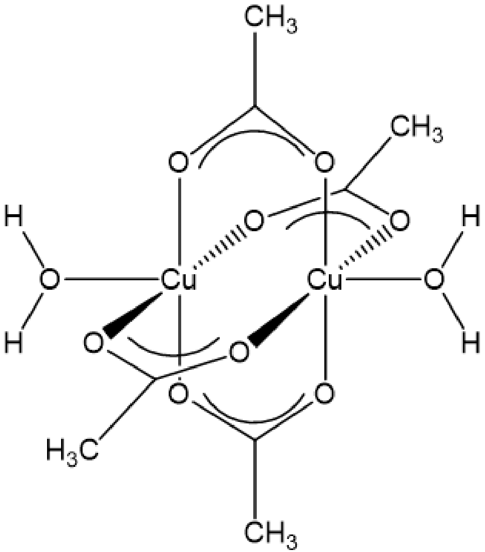
copper(II) acetate monohydrate

Findability in Chemistry?

copper acetate

copper(II) acetate monohydrate

dicopper(II) tetraacetate dihydrate



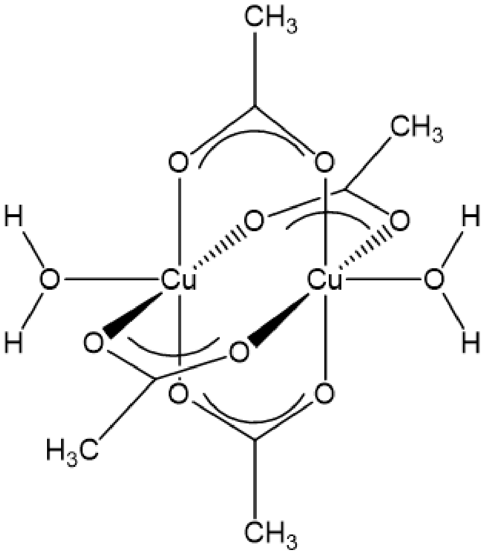
Findability in Chemistry?

copper acetate

copper(II) acetate monohydrate

dicopper(II) tetraacetate dihydrate

tetrakis(μ-acetato-κ^2*O*,*O′*)bis[(aqua)copper(II)]



Findability in Chemistry?

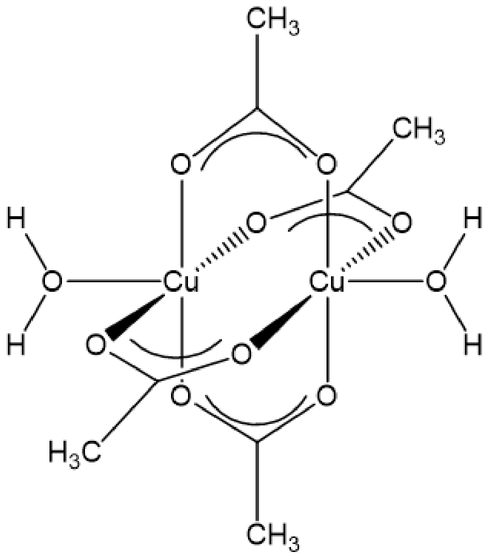
copper acetate

copper(II) acetate monohydrate

dicopper(II) tetraacetate dihydrate

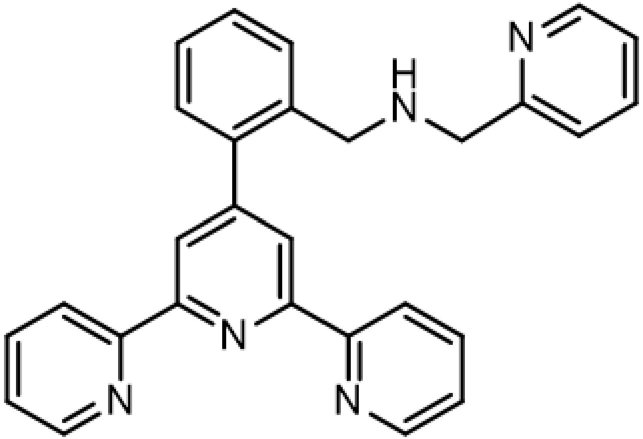
tetrakis(μ-acetato-κ^2*O*,*O′*)bis[(aqua)copper(II)]

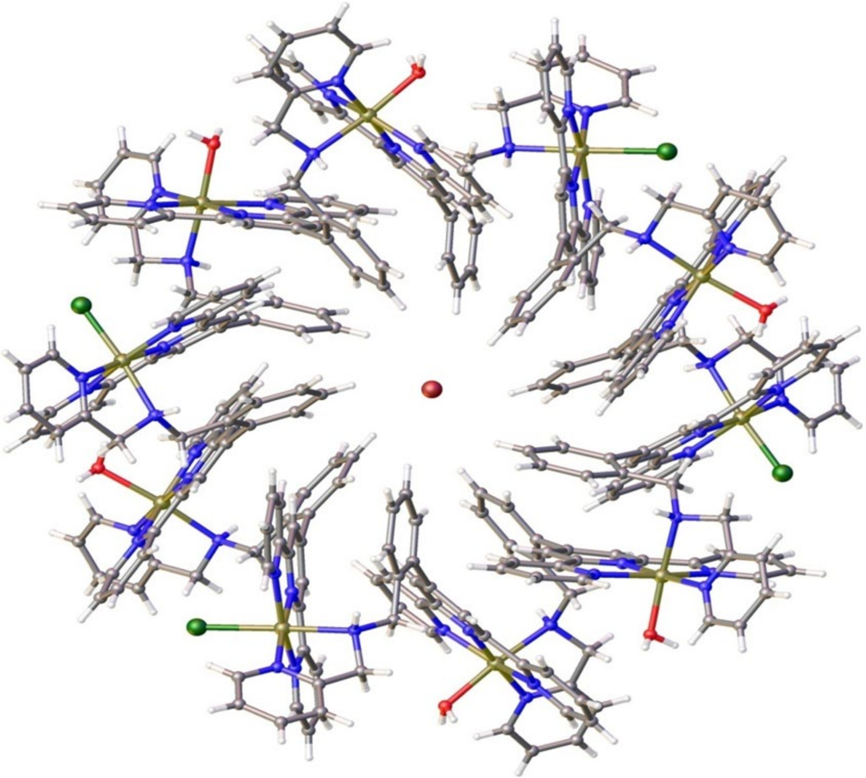
(*SPY*-5-21)(*SPY*-5-21)-tetrakis(μ-acetato-κ^2*O*,*O′*)bis[(aqua)copper(II)]



And if it is More Complicated?

COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

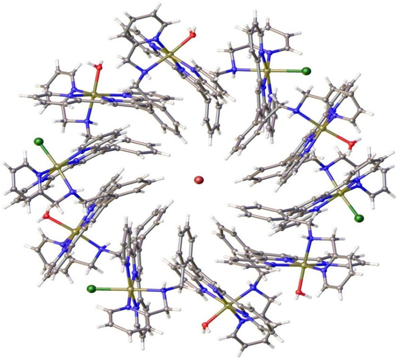




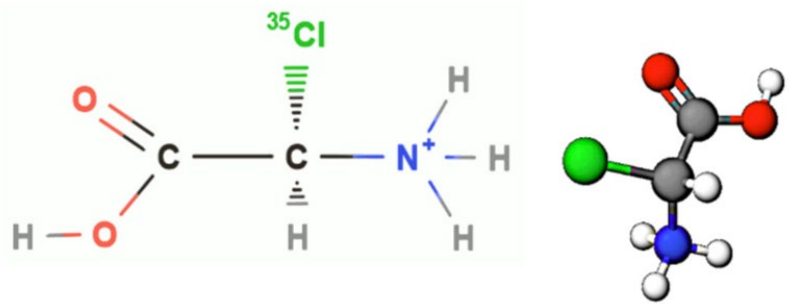
And if it is More Complicated?

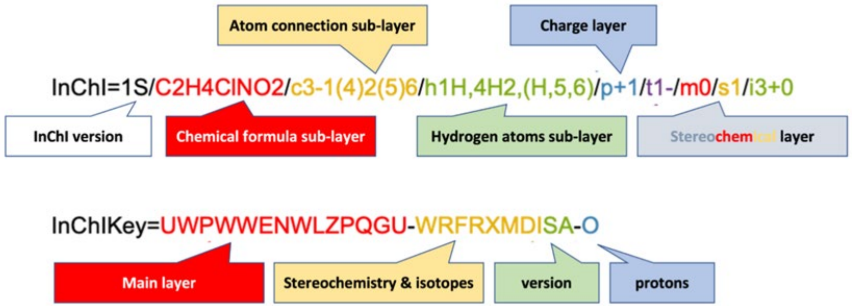
hexaaqua-1κ*O*,2κ*O*,3κ*O*,4κ*O*,5κ*O*,6κ*O*-tetrachlorido-7κ*Cl*,8κ*Cl*,9κ*Cl*,10κ*Cl*-μ-4′-{2-(2-pyridyl-1κ*N*-methylamino-1κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-7κ^{3}*N*-μ-4′-{2-(2-pyridyl-2κ*N*-methylamino-2κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-5κ^{3}*N*-μ-4′-{2-(2-pyridyl-3κ*N*-methylamino-3κ*N*-methyl) phenyl}-2,2′:6′,2″-terpyridine-9κ^{3}*N*-μ-4′-{2-(2-pyridyl-4κ*N*-methylamino-4κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-6κ^{3}*N*-μ-4′-{2-(2-pyridyl-5κ*N*-methyl

amino-5κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-1κ^{3}*N*-μ-4′-{2-(2-pyridyl-6κ*N*-methylamino-6κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-3κ^{3}*N*-μ-4′-{2-(2-pyridyl-7κ*N*-methylamino-7κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-10κ^{3}*N*-μ-4′-{2-(2-pyridyl-8κ*N*-methylamino-8κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-2κ^{3}*N*-μ-4′-{2-(2-pyridyl-9κ*N*-methylamino-9κ*N*-methyl)phenyl}-2,2′:6′,2″-terpyridine-8κ^{3}*N*-μ-4′-{2-(2-pyridyl-10κ*N*]-methylamino-10κ*N*-



InChI - a Useful Chemical Identifier





Atom connection sub-layer

Charge layer

InChI=1S/C2H4ClNO2/c3-1(4)2(5)6/h1H,4H2,(H,5,6)/p+1/t1-/m0/s1/i3+0

InChI version

Chemical formula sub-layer

Hydrogen atoms sub-layer

Stereochemical layer

InChIKey=UWPWWENWLZPQGU-WRFRXMDISA-O

Main layer

Stereochemistry & isotopes

version

protons

An illustration of the layered structure of the InChI string

The InChI and AI

Ethanol	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3	LFQSCWFLJHTTHZ-UHFFFAOYSA-N
ChatGPT	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3	LFQSCWFLJHTTHZ-UHFFFAOYSA-N
Acetamide	InChI=1S/C2H5NO/c1-2(3)4/h1H3,(H2,3,4)	DLFVBJFMPXGRIB-UHFFFAOYSA-N
ChatGPT	InChI=1S/C2H5NO/c1-2(3)4/h1H3,(H2,3,4)	DMDWJBKTFDJGJT-UHFFFAOYSA-N
2,2’-Bipyridine	InChI=1S/C10H8N2/c1-3-7-11-9(5-1)10-6-2-4-8-12-10/h1-8H	ROFVEXUMMXZLPA-UHFFFAOYSA-N
ChatGPT	InChI=1S/C10H8N2/c1-3-7-11-9(5-1)6-2-4-8-12-10(7)6/h1-8H	XOARYHGCYWUOKB-UHFFFAOYSA-N

Take Home Lesson #1

The introduction of generative AI makes it even more important that we build and use data validation protocols.

What can we do with the Data we Collect?

COMMITTEE ON DATA

CODATA

INTERNATIONAL SCIENCE COUNCIL

What can we do with the Data we Collect?

What does it look like?

How is it stored?

In what format?

What can we do with the Data we Collect?

What does it look like?

How is it stored?

In what format?

Units!!!

What can we do with the Data we Collect?

What does it look like?
How is it stored?
In what format?

Units!!!

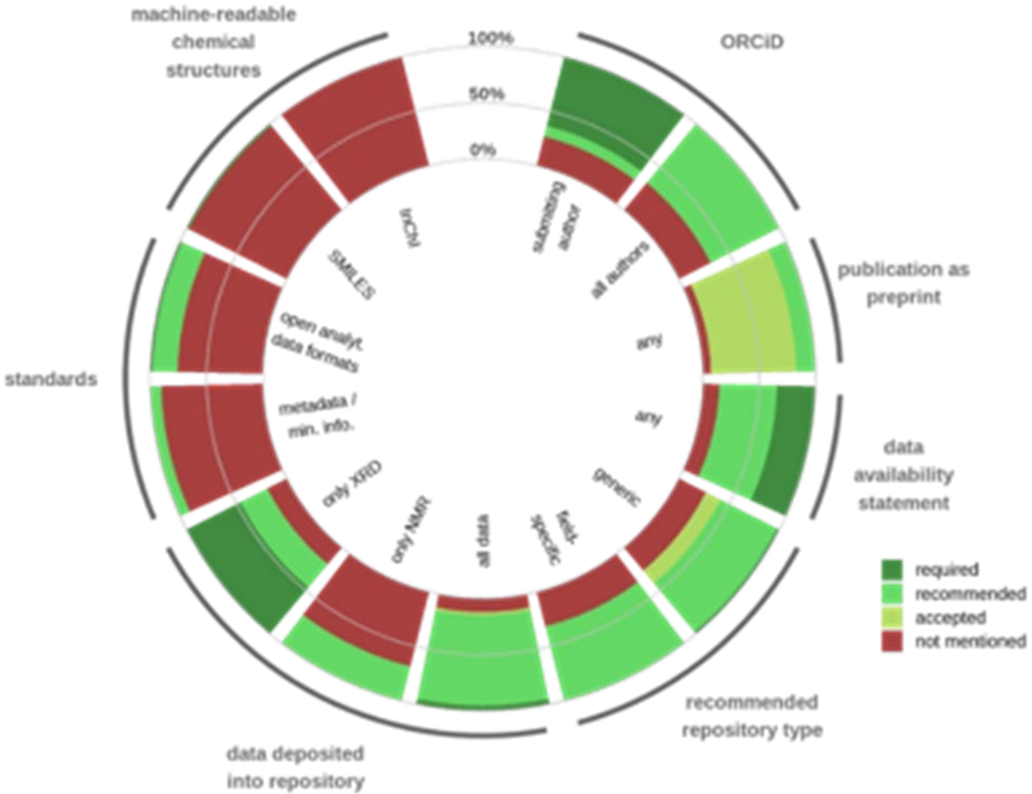
When and how was it collected?
Who by and under what conditions?

Take Home Lesson #2

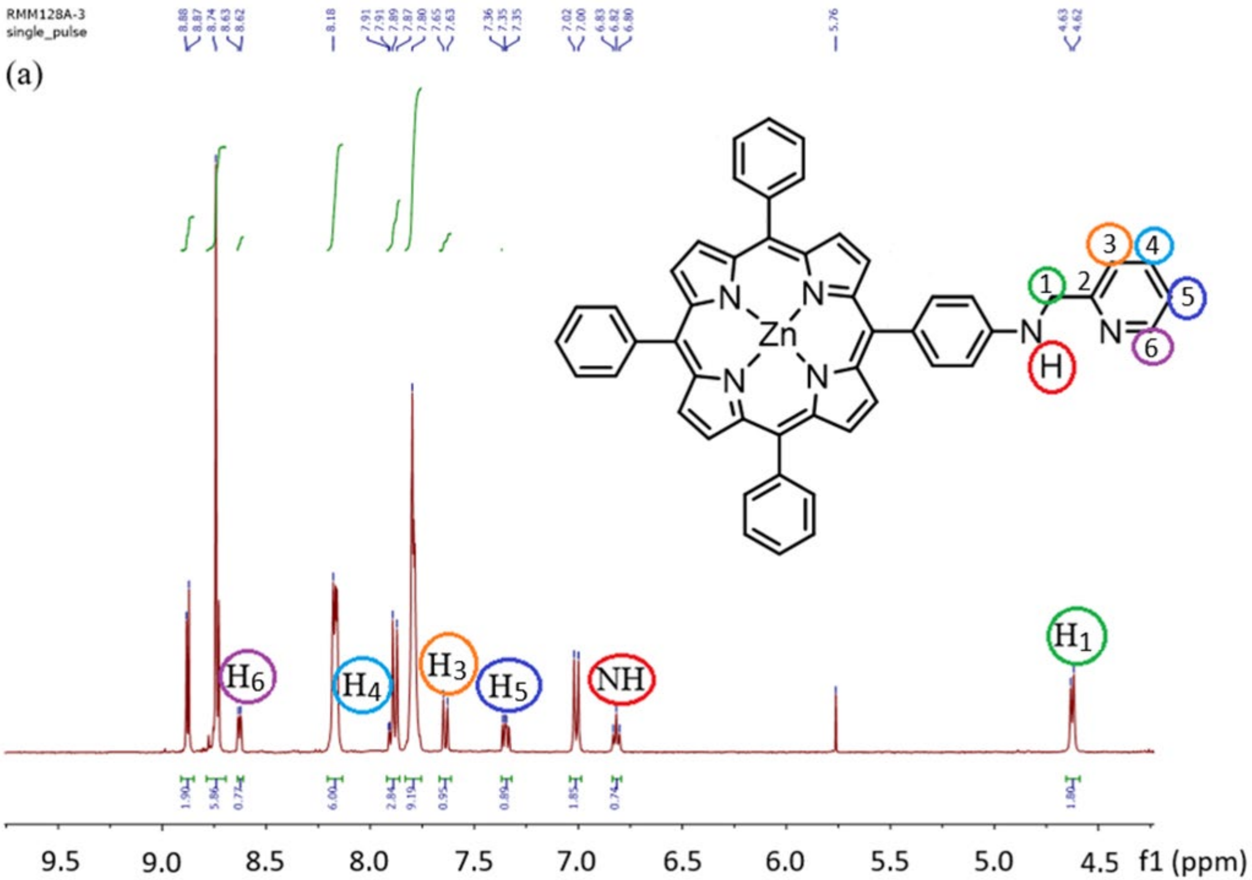
Metadata is just as important as the data itself.

So How Is Chemistry Doing?

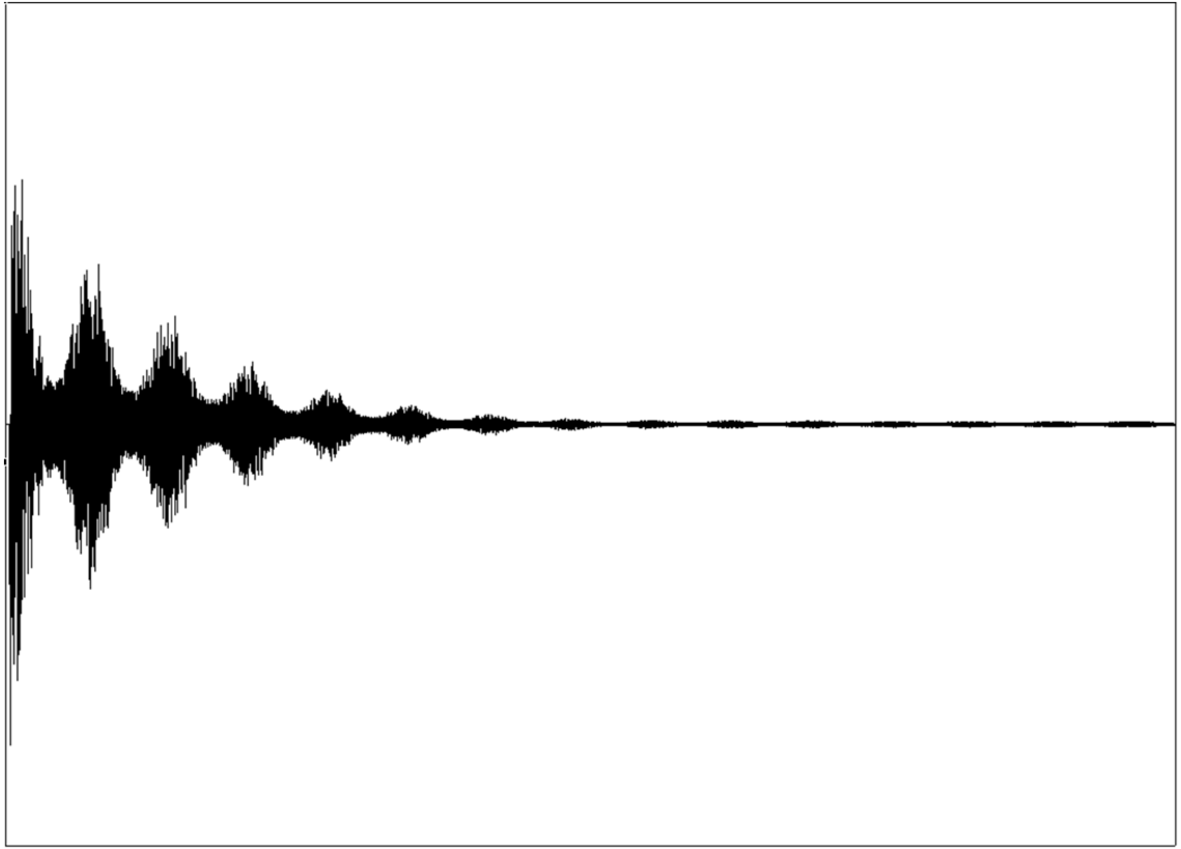
COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL



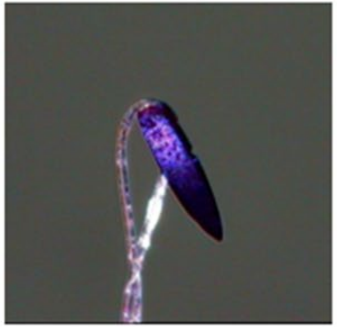
The Current State of Play

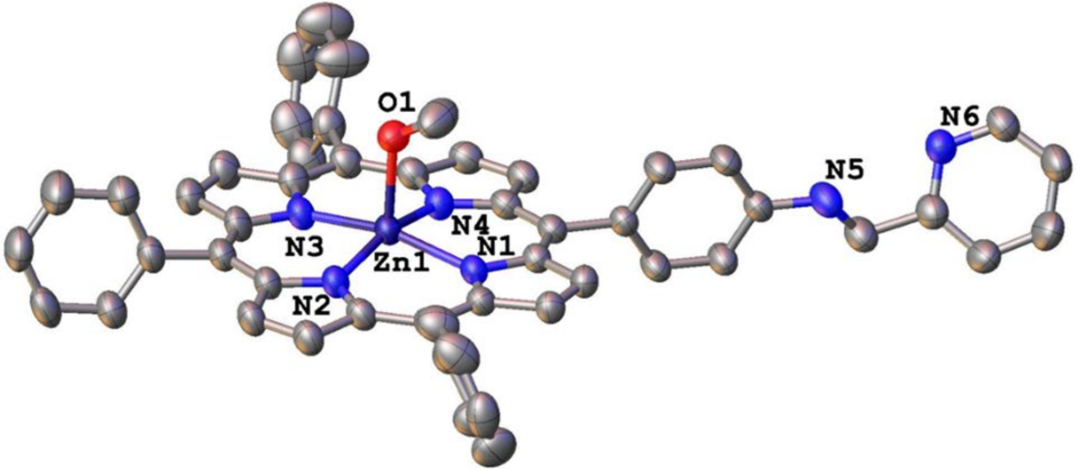


The Current State of Play



The Current State of Play





Take Home Lesson #3

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

We need to publish actual data (and not pictures of it) and make it available to people.

Garbage In - Garbage Out

We train AI on data…so that it can be a predictive tool…

Garbage In - Garbage Out

We train AI on data…so that it can be a predictive tool…

How representative of reality are those training data sets?

Garbage In - Garbage Out

We train AI on data…so that it can be a predictive tool…

How representative of reality are those training data sets?

Reproducibility?

Garbage In - Garbage Out

We train AI on data…so that it can be a predictive tool…

How representative of reality are those training data sets?

Reproducibility?

Best obtained vs Typical?

Failed experiments?

Take Home Lesson #4

We need to be more explicit about the experimental outcomes when we publish, and share the failed experiments as well!

Take Home Lesson #5

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Beware of auto-correct functions and other automated systems – especially if you don’t know they are there!

Acknowledgements

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Dr Rajika Munasinghe
Dr Gurpreet Kaur

Leah McEwen
Professor Jonathan Goodman
InChI Trust

A legion of volunteers…

Challenges from the social sciences

Dr. Steven McEachern
Director, UK Data Service
University of Essex

Challenges for data in the social sciences

Governance

Preservation and protection of data and metadata

Protection versus utility

Access and sharing

Interoperability

Governance responsibilities

Legal and ethical considerations

Federal legislation

Office of the National Data Commissioner – best practice principles and the 5 Safes

Data management and data quality

FAIR principles

Legal and ethical responsibilities

Australian legal responsibilities:

Privacy Act 1988 - 13 Australian Privacy Principles

Data Access and Transparency Act 2022: 5 Data Sharing Principles

Agency specific legislation

State, territory and international legislation

Codes of ethics (e.g. National Health and Medical Research Council, 2018)

Documentation and metadata

Core element of both data utility and information security

Enables the use of data – data without documentation can’t be transformed to information

Documentation can be at multiple levels, depending on user requirements:

Collections of content (e.g. a project or database)

File/table (e.g. a specific table in a database)

Variable and item (the column of information, or even a specific cell in a table)

Metadata

“Metadata are a specific subset of data documentation, which provide standardised, structured information explaining the purpose, origin, time references, geographic location, creating author, access conditions and terms of use of a data collection” (Corti et al., 2018, p.71)

Often based on standards for interoperability purposes, e.g.

International Standards Organisation – ISO19113 for spatial data (basis of NationalMap)

World Wide Web Consortium – DCAT for dataset descriptions (basis of data.gov.au)

UN Economic Council for Europe – SDMX (basis for ABS TableBuilder and ABS.Stat services)

Data access and data sharing

Access, sharing and releasing data

Protection versus utility

What is needed in order to share data?

Managing data access

Human access (downloads, remote access systems, secure rooms)

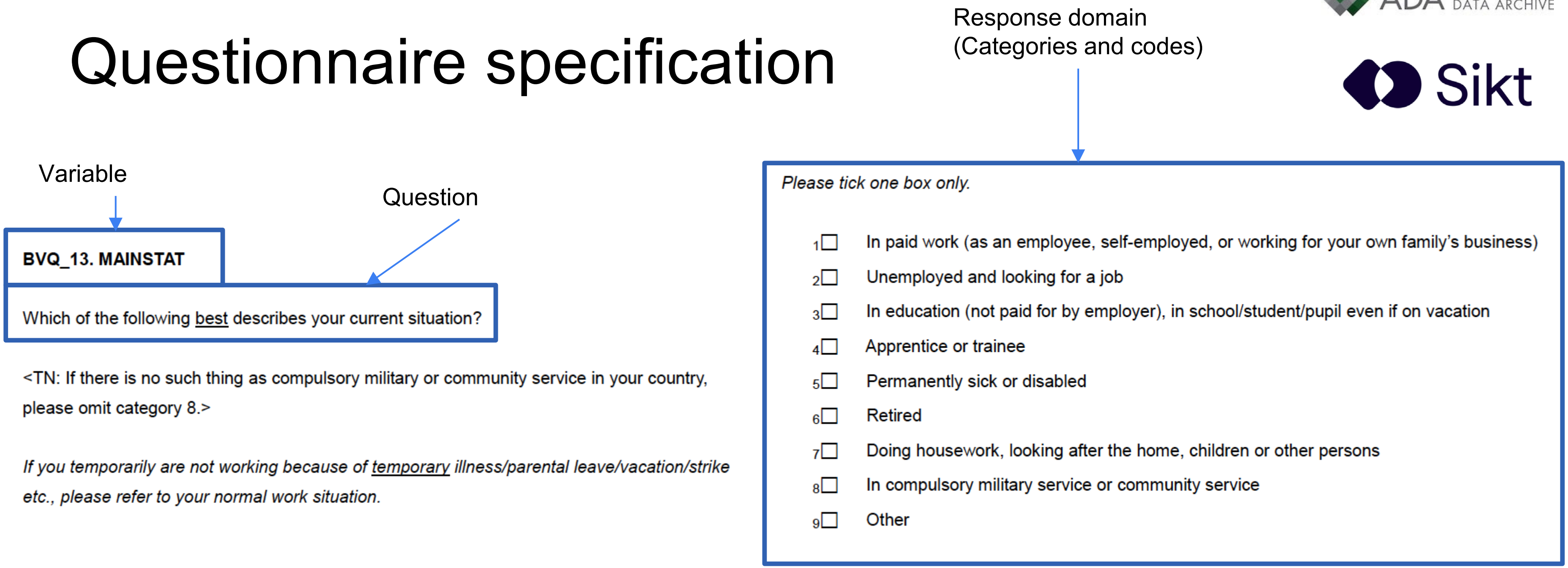
Machine access (APIs, database pipelines)

Requires:

- Technical services: systems for delivering data and communications
- Administrative services: people managing access requests
- Governance: policy and procedures governing the above

Data interoperability

Questionnaire specification



Variable

Question

Response domain
(Categories and codes)

BVQ_13. MAINSTAT

Which of the following best describes your current situation?

<TN: If there is no such thing as compulsory military or community service in your country, please omit category 8.>

If you temporarily are not working because of temporary illness/parental leave/vacation/strike etc., please refer to your normal work situation.

Please tick one box only.

1□ In paid work (as an employee, self-employed, or working for your own family's business)

2□ Unemployed and looking for a job

3□ In education (not paid for by employer), in school/student/pupil even if on vacation

4□ Apprentice or trainee

5□ Permanently sick or disabled

6□ Retired

7□ Doing housework, looking after the home, children or other persons

8□ In compulsory military service or community service

9□ Other

Plus?
Concept: Main status

https://ess-search.nsd.no/en/variable/query/mainact/1

https://www.europeansocialsurvey.org/

https://issp.org/

https://cses.org/data-download/download-data-documentation/variable-table/

Concepts, concepts everywhere

Research question / measurement concept

- Employment status

But how to connect the dots?

We can use DDI – but it needs more

Response domain (code lists and categories)

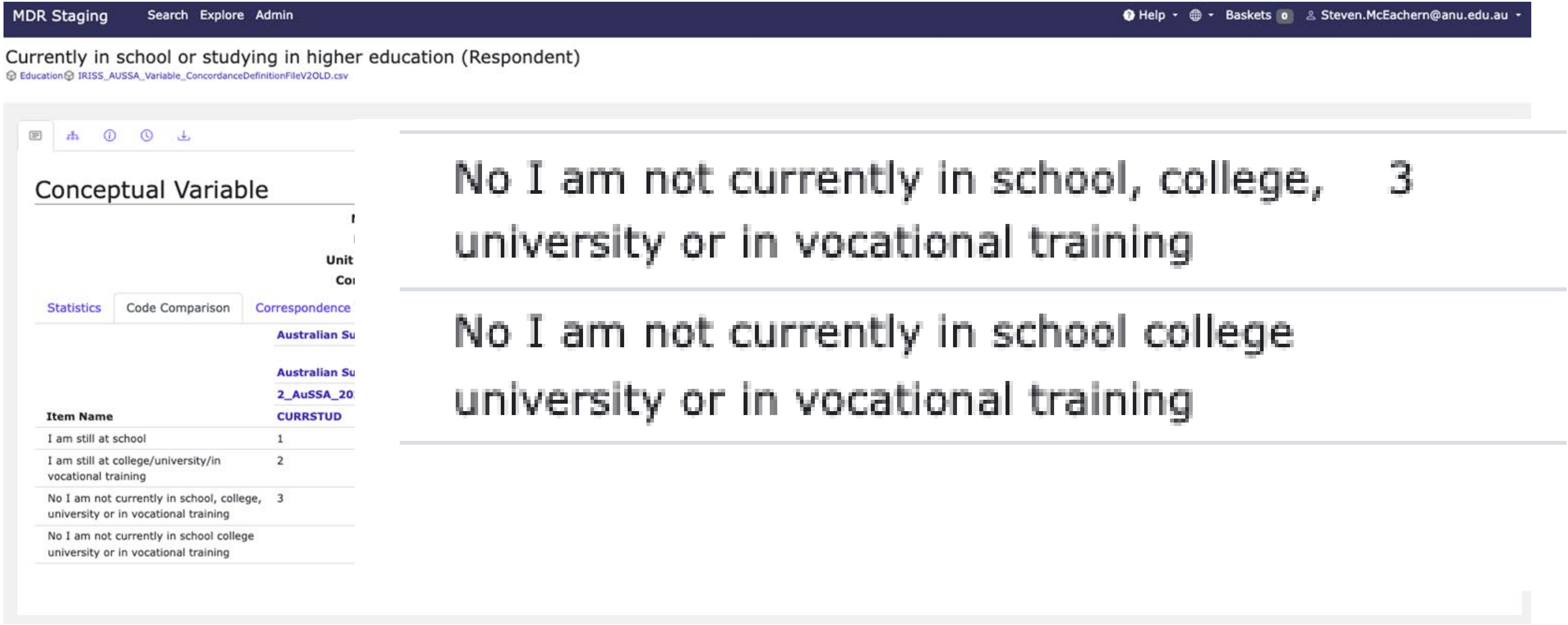
- Paid work
- Unemployed
- In education
- Apprenticeship/traineeship
- Permanently sick or disabled
- Retired
- Doing housework, looking after the home
- Military service
- Community service
- ….

Need to connect humans to machines, and machines to each other

ADA AUSTRALIAN DATA ARCHIVE

Sikt

- Humans make errors in process



Currently in school or studying in higher education (Respondent)

Education IRISS_AUSSA_Variable_ConcordanceDefinitionFileV2OLD.csv

Conceptual Variable

Unit

Co

Statistics | Code Comparison | Correspondence

Australian Su

Australian Su

2_AuSSA_20

Item Name	CURRSTUD
I am still at school	1
I am still at college/university/in vocational training	2
No I am not currently in school, college, university or in vocational training	3
No I am not currently in school college university or in vocational training	

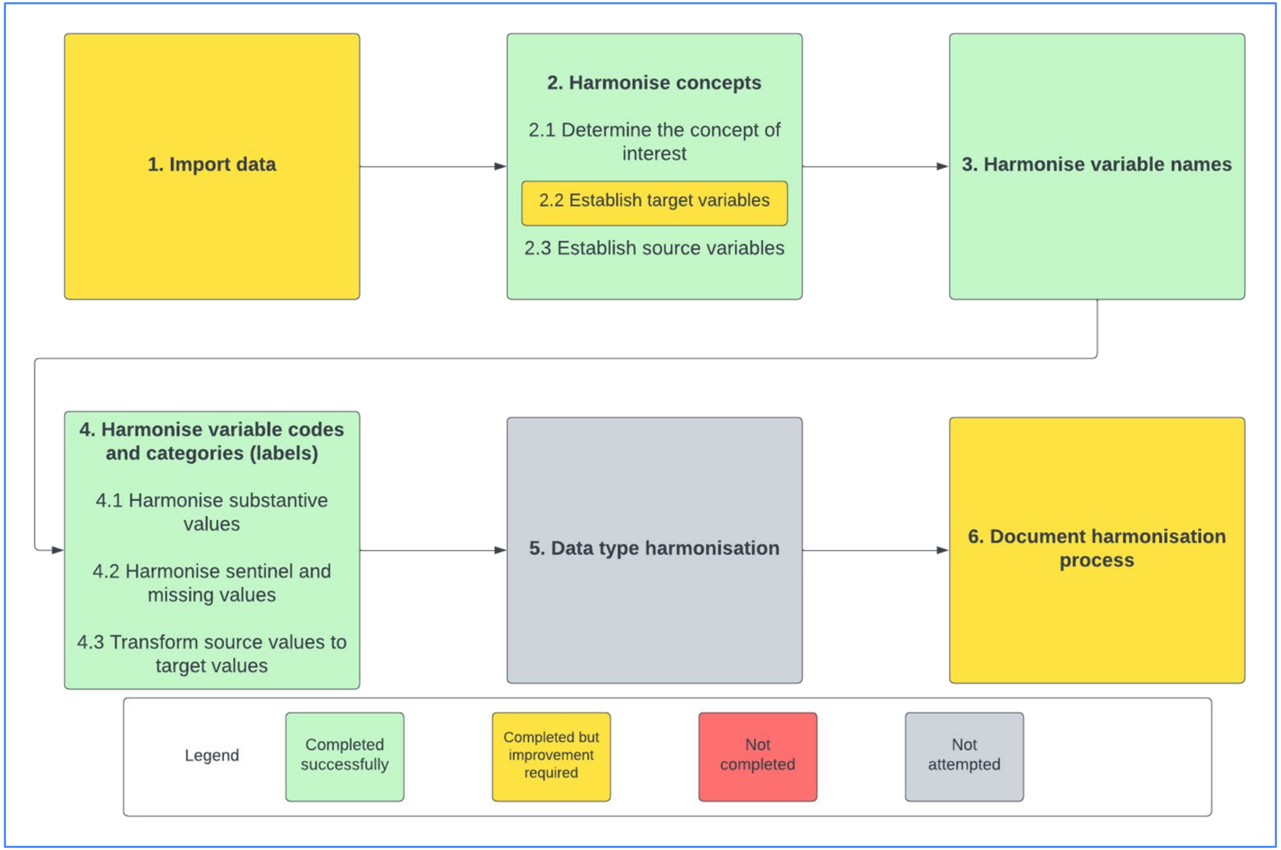
No I am not currently in school, college, university or in vocational training 3

No I am not currently in school college university or in vocational training

Content is not accessible in machine-processable form

- PDF data dictionaries: easy to read, printable
  - Great for humans
  - Useless for machines
- ABS Census Data Dictionary: HTML
  - nicely formatted for humans, easily readable, content-rich
  - HTML is not consistent - for machines – NOT REUSEABLE

Possible approach



1. Import data

2. Harmonise concepts

2.1 Determine the concept of interest

2.2 Establish target variables

2.3 Establish source variables

3. Harmonise variable names

4. Harmonise variable codes and categories (labels)

4.1 Harmonise substantive values

4.2 Harmonise sentinel and missing values

4.3 Transform source values to target values

5. Data type harmonisation

6. Document harmonisation process

Legend

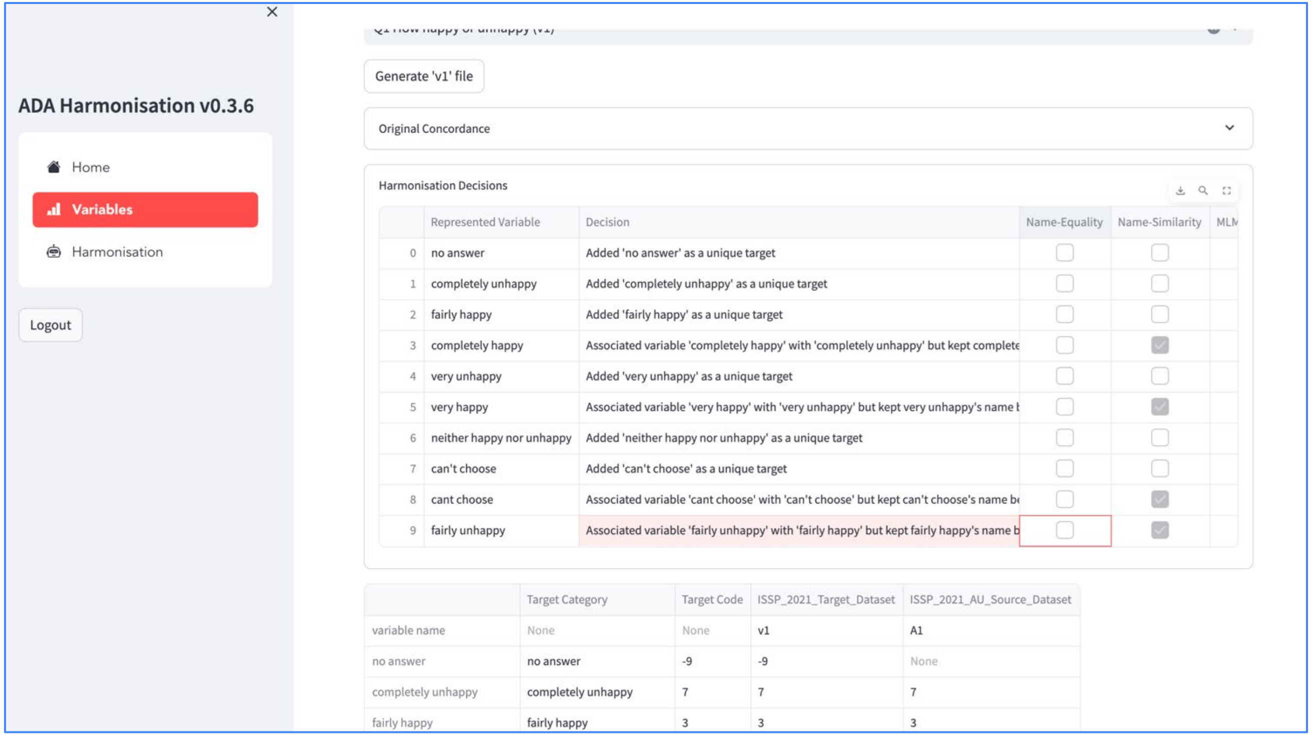
🟩 Completed successfully

🟨 Completed but improvement required

🟥 Not completed

⬜ Not attempted

Automating documentation



Harmonisation Decisions

	Represented Variable	Decision	Name-Equality	Name-Similarity	MLM
0	no answer	Added 'no answer' as a unique target	□	□	
1	completely unhappy	Added 'completely unhappy' as a unique target	□	□	
2	fairly happy	Added 'fairly happy' as a unique target	□	□	
3	completely happy	Associated variable 'completely happy' with 'completely unhappy' but kept complete	□	☑	
4	very unhappy	Added 'very unhappy' as a unique target	□	□	
5	very happy	Associated variable 'very happy' with 'very unhappy' but kept very unhappy's name b	□	☑	
6	neither happy nor unhappy	Added 'neither happy nor unhappy' as a unique target	□	□	
7	can't choose	Added 'can't choose' as a unique target	□	□	
8	cant choose	Associated variable 'cant choose' with 'can't choose' but kept can't choose's name be	□	☑	
9	fairly unhappy	Associated variable 'fairly unhappy' with 'fairly happy' but kept fairly happy's name b	□	☑	

	Target Category	Target Code	ISSP_2021_Target_Dataset	ISSP_2021_AU_Source_Dataset
variable name	None	None	v1	A1
no answer	no answer	-9	-9	None
completely unhappy	completely unhappy	7	7	7
fairly happy	fairly happy	3	3	3

Recommendations

1. Establishment of standardised access controls both to data and metadata registries, to limit the need for less technical users to navigate access control systems
2. Establishment of a code repository for interaction with social science metadata repositories.
3. Establishment of mechanisms for reuse of conceptual variable and other reference metadata across the DDI standards ecosystem.
4. Standardised practices and code libraries for the creation of DDI resource packages for external reuse (to facilitate the reuse in Recommendation 3)

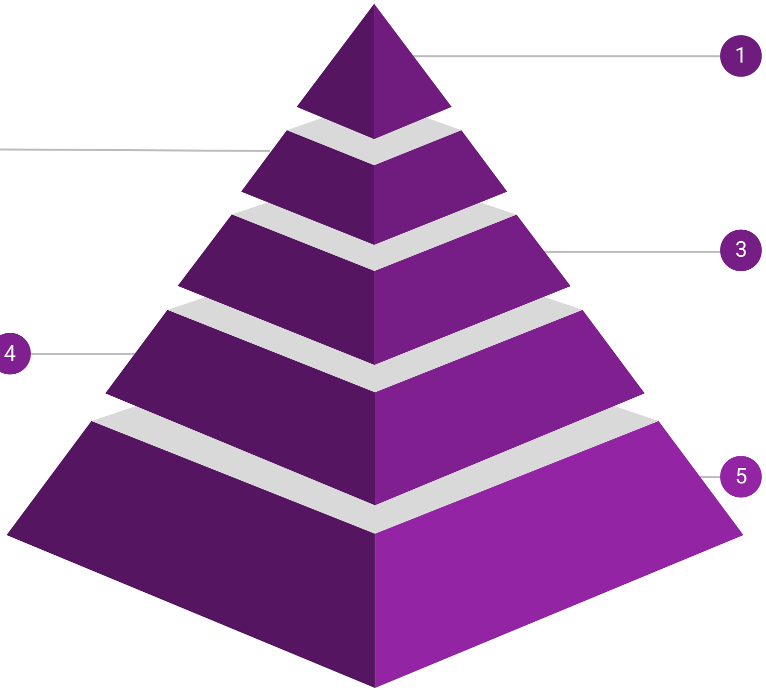
Research Data Skills for the 21st Century

Essential Skills for Researchers in a Rapidly Changing World

Dr. Shaily Gandhi
Senior Post-Doctoral Researcher
Geosocial AI Research Group
Interdisciplinary Transformation University
Austria

What skills do researchers need?

COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL



1

Technical and Methodological Skills

Data Analysis, Experimental Design, Coding and Computational Skills, Lab Techniques and Research Ethics. Open Science Practices and Interdisciplinary Research

2

Analytical and Critical Thinking

Experimental design, Problem-Solving, Critical Evaluation, Quantitative and Qualitative Analysis

3

Communication Skills

Scientific Writing, Public Speaking, Grant Writing, Knowledge Translation

4

Collaboration and Interpersonal Skills

Teamwork, Networking, Mentorship and Leadership, Flexibility, Learning Agility, Resilience

5

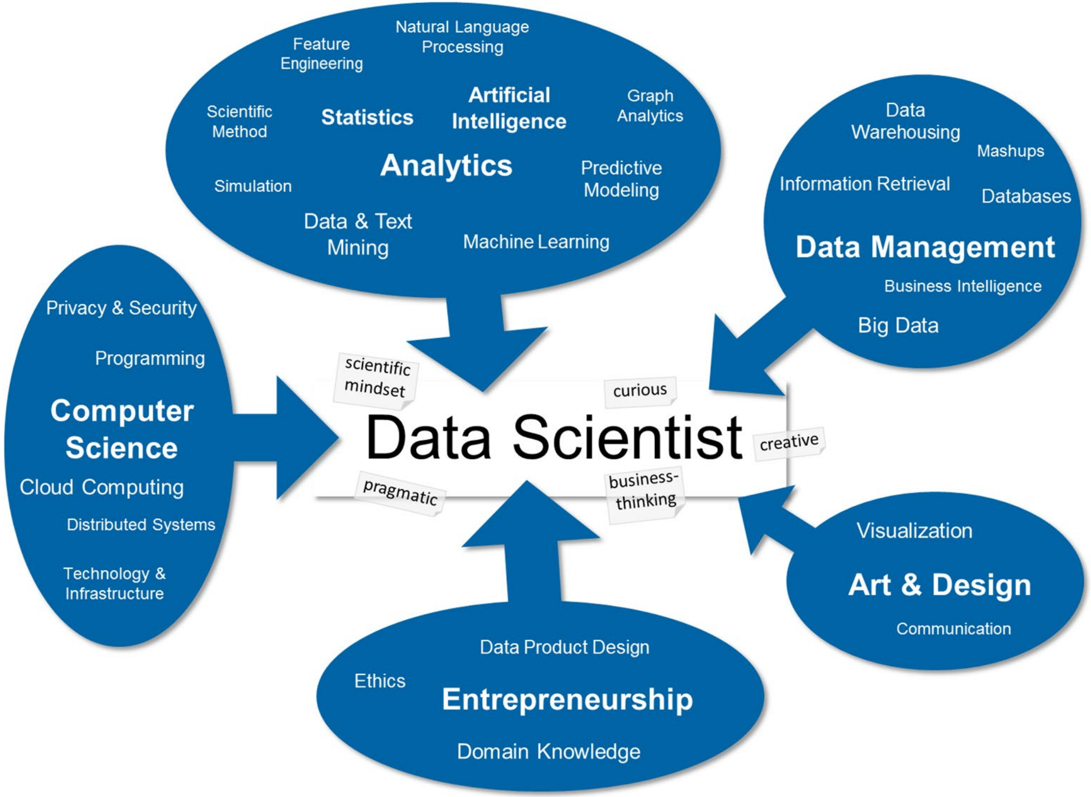
Project and Time Management

Organization, Resource Management, Goal Setting

Why Data Skills Matter ?

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

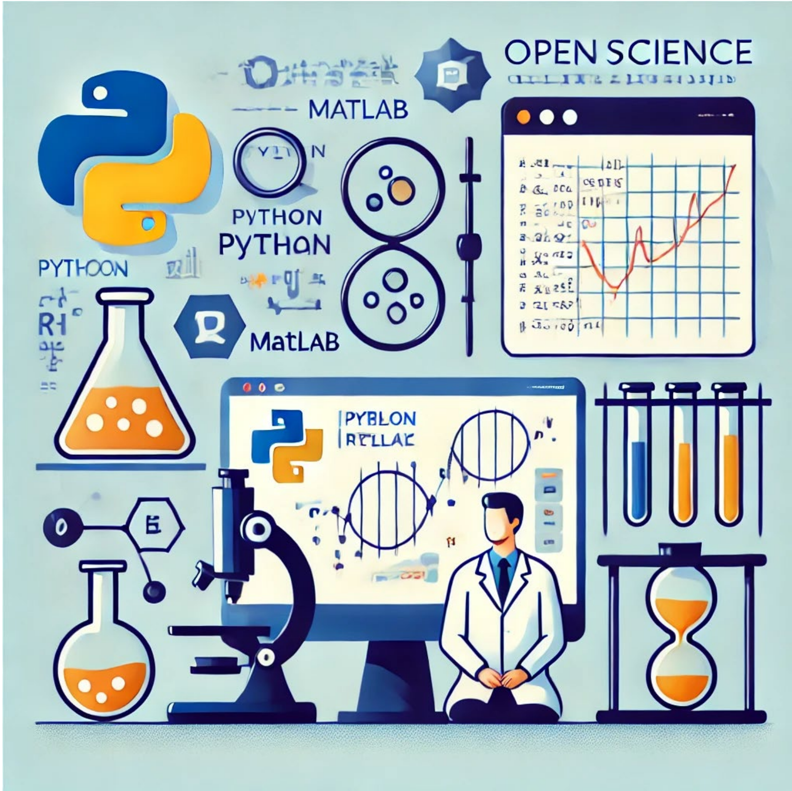
- The explosion of data in research and society
- Bridging the gap between data collection and actionable insights
- Supporting interdisciplinary research and Open Science



Source: https://blog.zhaw.ch/datascience/the-data-science-skill-set/

Foundations of Research Data Skills

- Proficiency in data analysis, experimental design, and computational tools.
- Coding languages like Python, R, or MATLAB.
- Emphasis on Open Science practices and reproducibility.
- Awareness of lab techniques and ethical considerations.
- Ethical use of the data.
- Presenting the research



Source: ChatGPT

Making Sense of Complex Data

- Designing experiments to solve problems.
- Interdisciplinary aspect for data collection.
- Performing quantitative and qualitative analyses.
- Critically evaluating results.
- Considering ethical implications in data interpretation.

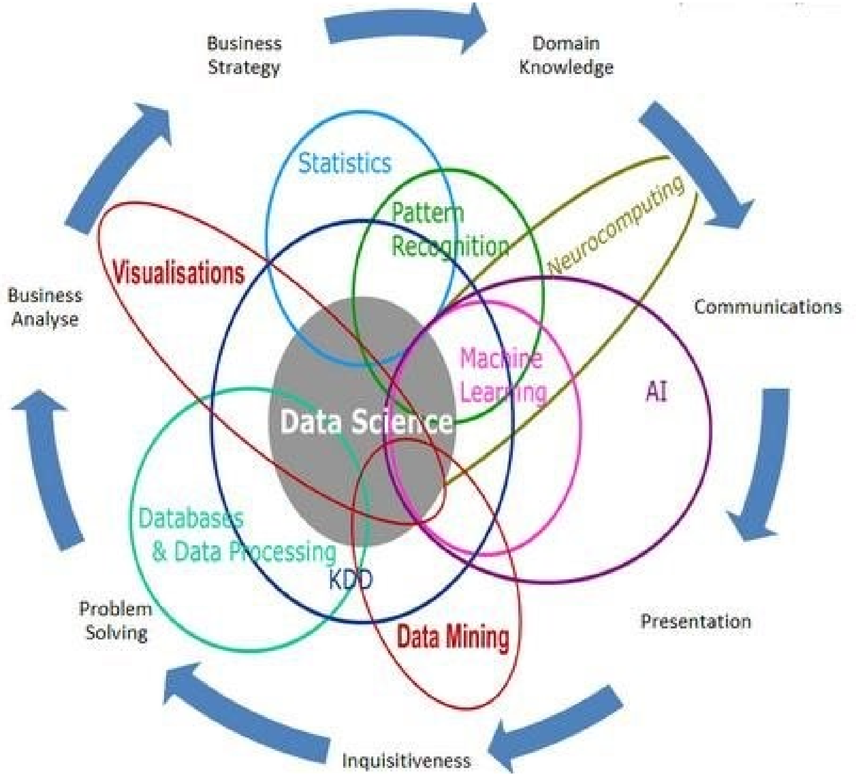
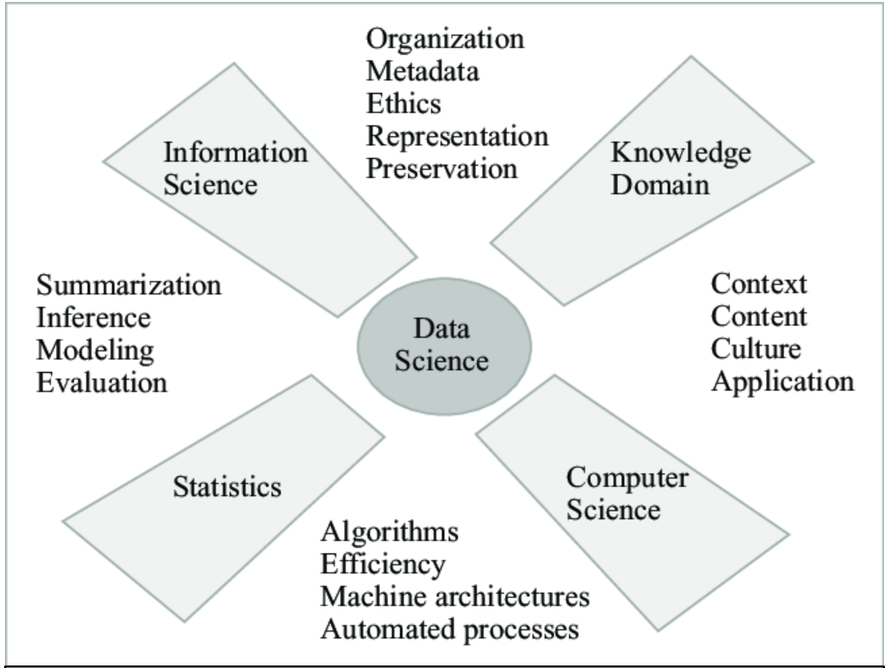


Image Source: Tan, Jamaluddin & Tan, Kim Hua & Nadiah, Anis & Che Abdul Rahman, Anis Nadiah & Mohamed Zaki, Filzah Zahilah & Yaakub, Mohd Ridzwan. (2019). PROSIDING SIMPOSIUM KEMANUSIAAN DIGITAL 2019 (Proceeding of Digital Humanities Symposium 2019)- COMPUTATIONAL THINKING FOR FUTURE DATA SCIENTISTS (eISBN 978-883-9391-71-8).

Teamwork in Interdisciplinary Research

- FAIR Data
- Building effective research teams for better collaborations
- Developing flexibility and resilience platforms for data sharing

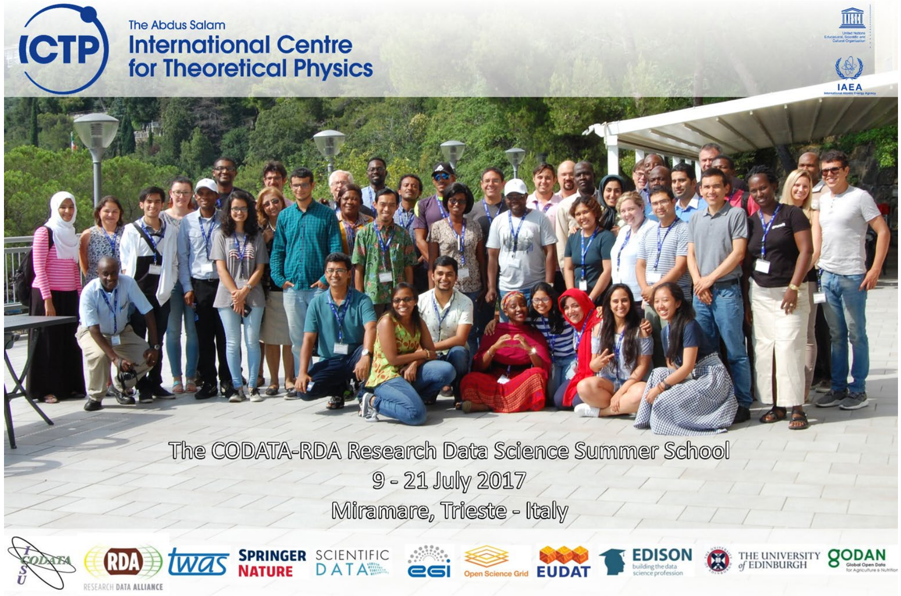


Source: Marchionini, Gary. (2016). Information Science Roles in the Emerging Field of Data Science. Journal of Data and Information Science. 1. 1-6. 10.20309/jdis.201609. DOI:10.20309/jdis.201609 | License: CC BY-NC-ND 4.0

Data Stewardship and FAIR Principles

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

- Implementing FAIR (Findable, Accessible, Interoperable, Reusable) data principles.
- Creating and managing metadata.
- Automating data validation and workflows.
- Ethical considerations in data handling.



CODATA – RDA
Data Schools
SCHOOLS OF RESEARCH DATA SCIENCE

software carpentry

DATA CARPENTRY
BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

D C C

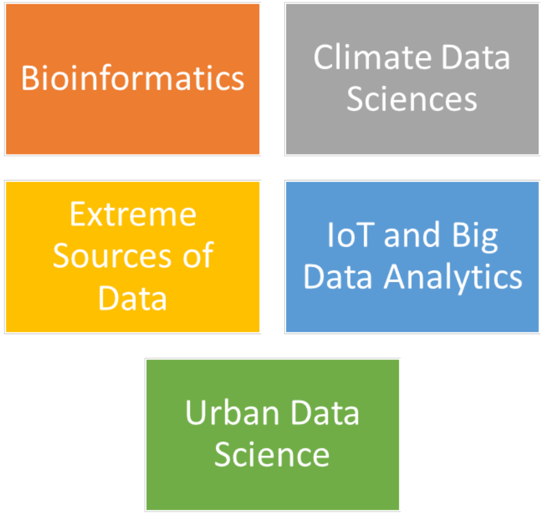
DATA-RDA Schools of Research Data Science

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

The CODATA-RDA Schools of Research Data Science aim to equip early-career researchers with essential data science skills applicable across various disciplines. Key areas of focus include:

- **Open Science Principles**: Emphasizing transparency and accessibility in research.
- **FAIR Data Practices**: Ensuring data is Findable, Accessible, Interoperable, and Reusable.
- **Research Data Management**: Covering the data lifecycle, management plans, persistent identifiers, licensing, and efficient data discovery and publication.
- **Software and Data Carpentry**: Teaching practical skills in tools like GitHub, programming languages (e.g., R, SQL), and the Unix Shell.
- **Data Visualization and Machine Learning**: Introducing visualization techniques and foundational machine learning concepts

Several applied/thematic workshops on Research Data Science



Bioinformatics

Climate Data Sciences

Extreme Sources of Data

IoT and Big Data Analytics

Urban Data Science

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Discussion and Feedback

Group Discussion

- Discussion in groups
- How is your discipline responding to these challenges?
- Where does your discipline need help and collaboration on these issues?

COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL