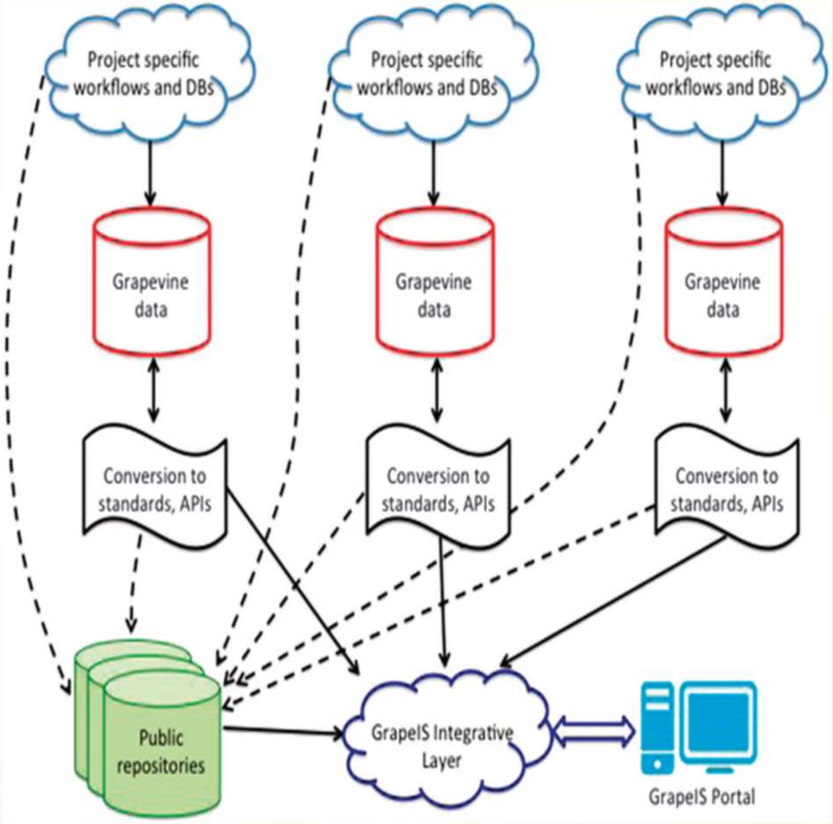
Introduction to F.A.I.R. data management in metabolomics and actual situations in viticulture and wine science

Fulvio Mattivi

UNIVERSITÀ DI TRENTO

*To improve the findability, accessibility, interoperability and reusability (**FAIR**), for more **transparent** dissemination of these data*

Conceptual scheme of the grapevine distributed information system (GrapeIS)


Project specific workflows and DBs
Project specific workflows and DBs
Project specific workflows and DBs
Grapevine data
Grapevine data
Grapevine data
Conversion to standards, APIs
Conversion to standards, APIs
Conversion to standards, APIs
Public repositories
GrapeIS Integrative Layer
GrapeIS Portal

Adam-Blondon et al., Horticulture Research (2016) 3, 16056; doi:10.1038/hortres.2016.56

Interoperability

Interoperability is the ability of different information systems, devices or applications to connect, in a coordinated manner, within and across organizational boundaries to access, exchange and cooperatively use data amongst stakeholders (https://www.himss.org)

WHAT IS INTEROPERABILITY?

Why?

Interoperability and scientific research

- ✓ Knowledge is additive
- ✓ No one can measure everything
- ✓ Standing on the shoulder of giants
- ✓ Good science is reproducible
- ✓ Context is continuously changing
- ✓ …..

Standardization

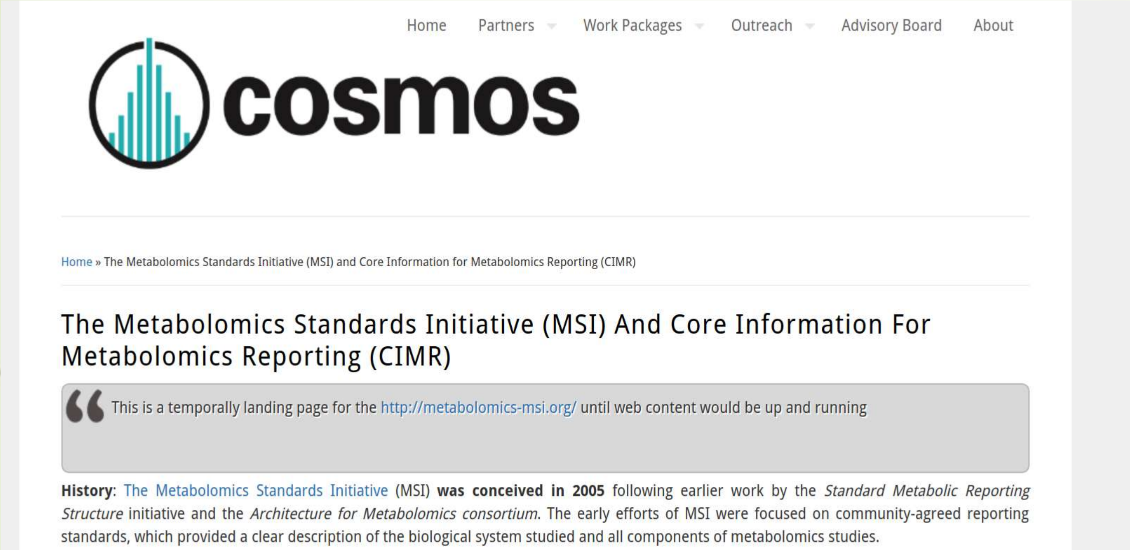
Interoperability requires standardization. Standardization is required if things should be automatically done by a computer

- Experimental designs
- Sample treatment and analysis
- Terms and languages
- Data storage
- Data analysis strategies
- Reporting
- ...

STANDARDIZATION IS ...
THE CONSISTENCY
OF THE WORK SEQUENCE.

+ Metadata!

The starting point (2005)

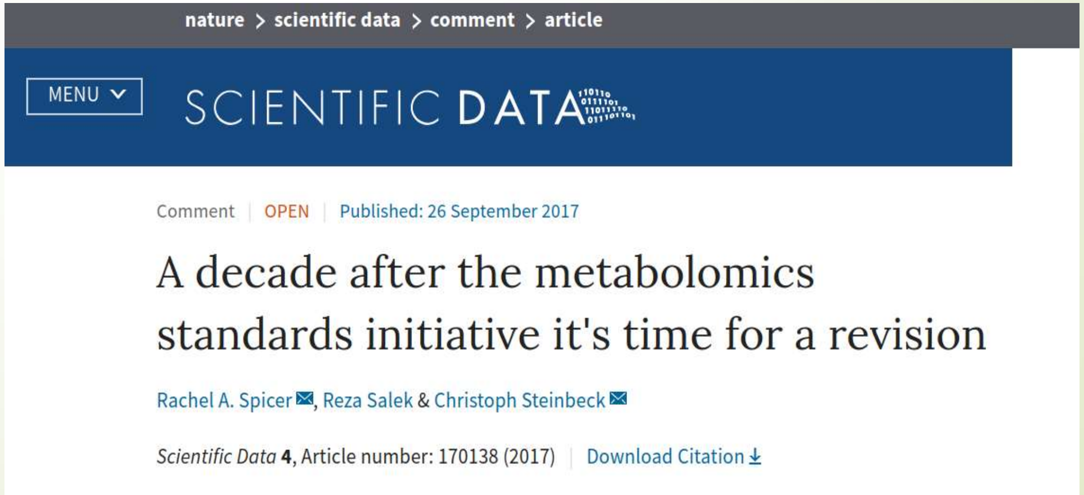


Home » The Metabolomics Standards Initiative (MSI) and Core Information for Metabolomics Reporting (CIMR)

The Metabolomics Standards Initiative (MSI) And Core Information For Metabolomics Reporting (CIMR)

> This is a temporally landing page for the http://metabolomics-msi.org/ until web content would be up and running

History: The Metabolomics Standards Initiative (MSI) **was conceived in 2005** following earlier work by the *Standard Metabolic Reporting Structure* initiative and the *Architecture for Metabolomics consortium*. The early efforts of MSI were focused on community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies.



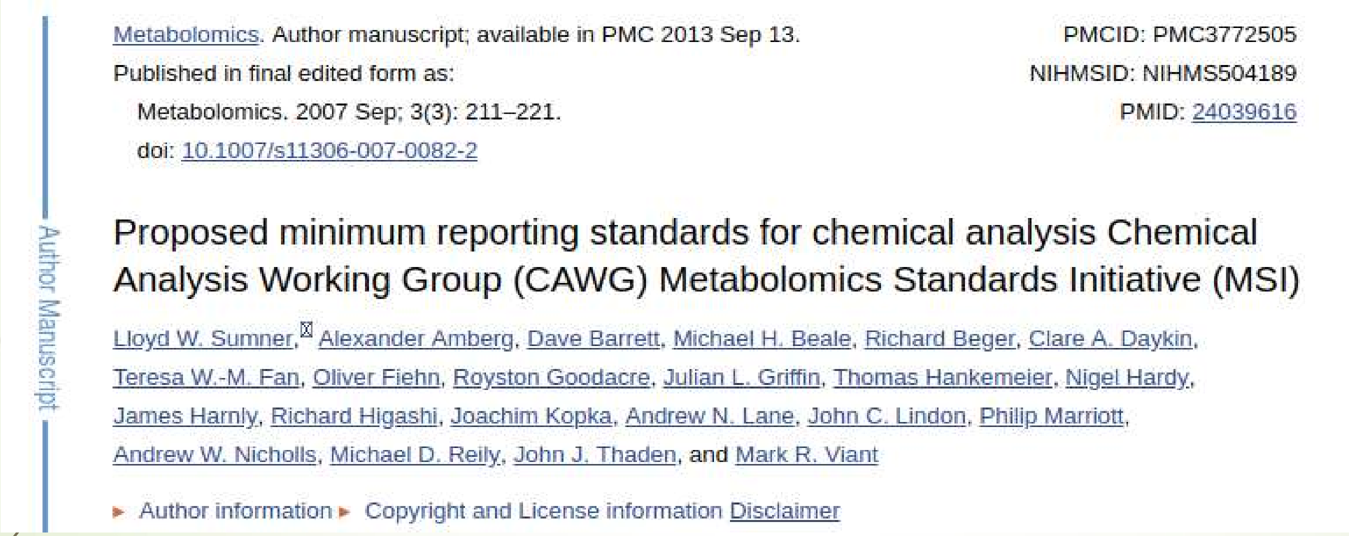
Comment | OPEN | Published: 26 September 2017

A decade after the metabolomics standards initiative it's time for a revision

Rachel A. Spicer, Reza Salek & Christoph Steinbeck

Scientific Data **4**, Article number: 170138 (2017) | Download Citation

… the reference paper for chemical analysis



Metabolomics. Author manuscript; available in PMC 2013 Sep 13.
Published in final edited form as:
Metabolomics. 2007 Sep; 3(3): 211–221.
doi: 10.1007/s11306-007-0082-2

PMCID: PMC3772505
NIHMSID: NIHMS504189
PMID: 24039616

Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)

Lloyd W. Sumner,☒ Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, Teresa W.-M. Fan, Oliver Fiehn, Royston Goodacre, Julian L. Griffin, Thomas Hankemeier, Nigel Hardy, James Harnly, Richard Higashi, Joachim Kopka, Andrew N. Lane, John C. Lindon, Philip Marriott, Andrew W. Nicholls, Michael D. Reily, John J. Thaden, and Mark R. Viant

▸ Author information ▸ Copyright and License information Disclaimer

Proposed

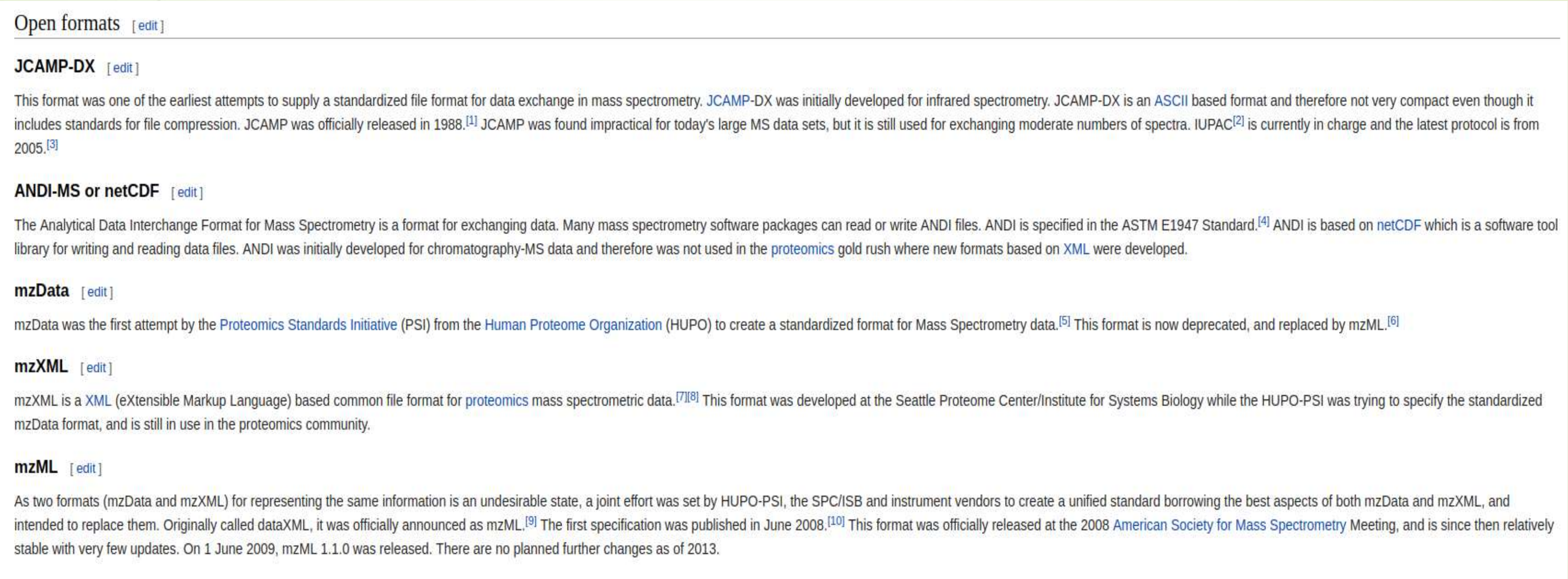
- ✓ minimum metadata for **sample preparation**
- ✓ minimum metadata relative to **chromatography**
- ✓ minimum metadata relative to **mass spectrometry** and **NMR**
- ✓ Minimum metadata relative to **instrumental performance** and **method validation**
- ✓ minimum metadata relative to **data preprocessing**
- ✓ minimum metadata relative to **metabolite identification**

Data formats

For targeted analyses concentrations are sufficient,

for untargeted metabolomics raw data **have to be converted in open formats**

MS



Open formats [edit]

JCAMP-DX [edit]

This format was one of the earliest attempts to supply a standardized file format for data exchange in mass spectrometry. JCAMP-DX was initially developed for infrared spectrometry. JCAMP-DX is an ASCII based format and therefore not very compact even though it includes standards for file compression. JCAMP was officially released in 1988.[1] JCAMP was found impractical for today's large MS data sets, but it is still used for exchanging moderate numbers of spectra. IUPAC[2] is currently in charge and the latest protocol is from 2005.[3]

ANDI-MS or netCDF [edit]

The Analytical Data Interchange Format for Mass Spectrometry is a format for exchanging data. Many mass spectrometry software packages can read or write ANDI files. ANDI is specified in the ASTM E1947 Standard.[4] ANDI is based on netCDF which is a software tool library for writing and reading data files. ANDI was initially developed for chromatography-MS data and therefore was not used in the proteomics gold rush where new formats based on XML were developed.

mzData [edit]

mzData was the first attempt by the Proteomics Standards Initiative (PSI) from the Human Proteome Organization (HUPO) to create a standardized format for Mass Spectrometry data.[5] This format is now deprecated, and replaced by mzML.[6]

mzXML [edit]

mzXML is a XML (eXtensible Markup Language) based common file format for proteomics mass spectrometric data.[7][8] This format was developed at the Seattle Proteome Center/Institute for Systems Biology while the HUPO-PSI was trying to specify the standardized mzData format, and is still in use in the proteomics community.

mzML [edit]

As two formats (mzData and mzXML) for representing the same information is an undesirable state, a joint effort was set by HUPO-PSI, the SPC/ISB and instrument vendors to create a unified standard borrowing the best aspects of both mzData and mzXML, and intended to replace them. Originally called dataXML, it was officially announced as mzML.[9] The first specification was published in June 2008.[10] This format was officially released at the 2008 American Society for Mass Spectrometry Meeting, and is since then relatively stable with very few updates. On 1 June 2009, mzML 1.1.0 was released. There are no planned further changes as of 2013.

- ✓ Fast and efficient **storage of big data**
- ✓ **Extensible**
- ✓ **Vendor agnostic**
- ✓ …

Data Repositories

Raw (and open!) data have to be made available in data repositories, where the experiments **should be** consistently documented ...

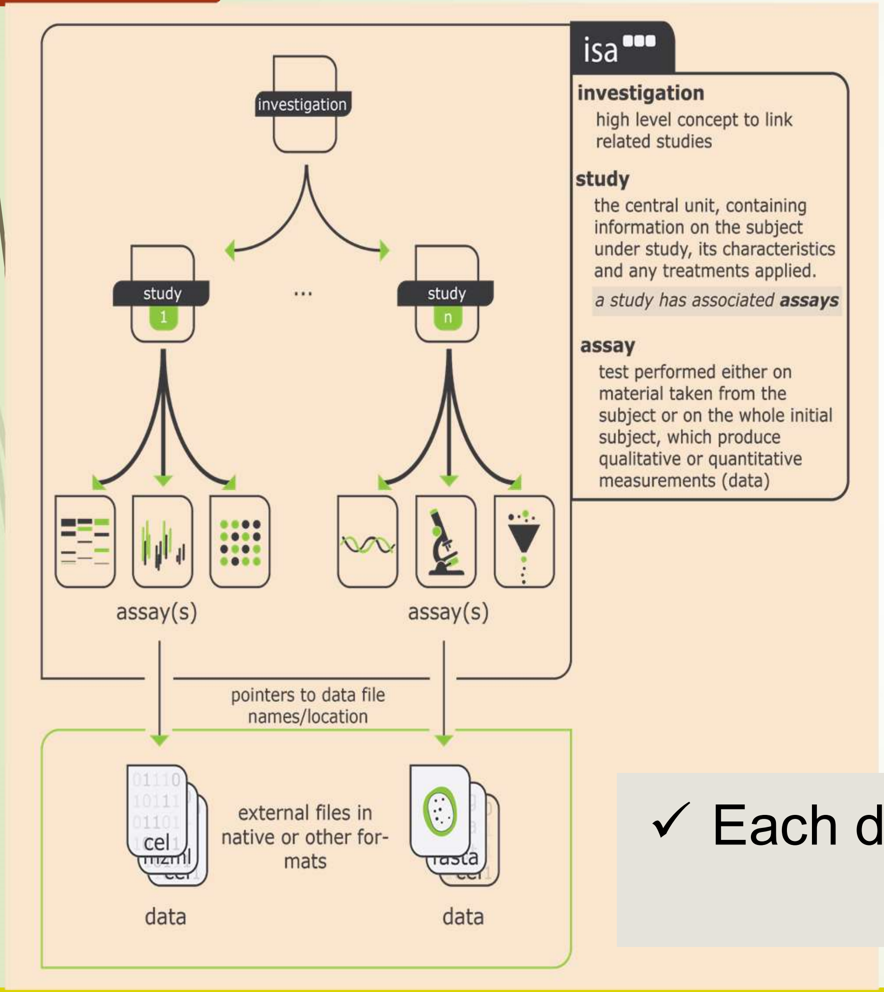
- ✓ Metabolights (EBI)
- ✓ Metabolomics Workbench (NIH)
- ✓ Metabolonote
- ✓ ...

MetabolomeXchange

Connecting metabolomics repositories!

Study metadata

https://isa-tools.org/


isa
investigation
high level concept to link related studies
study
the central unit, containing information on the subject under study, its characteristics and any treatments applied.
a study has associated assays
assay
test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)
investigation
study 1
...
study n
assay(s)
assay(s)
pointers to data file names/location
external files in native or other formats
cel
mzml
fasta
data
data

Built around the '**Investigation**' (the project context), '**Study**' (a unit of research) and '**Assay**' (analytical measurement) data model and serializations (tabular, JSON and RDF), the **ISA framework** helps you to provide rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are **reproducible** and **reusable**.

✓ Each description is performed relying as much as possible on **common dictionaries** and **ontologies**

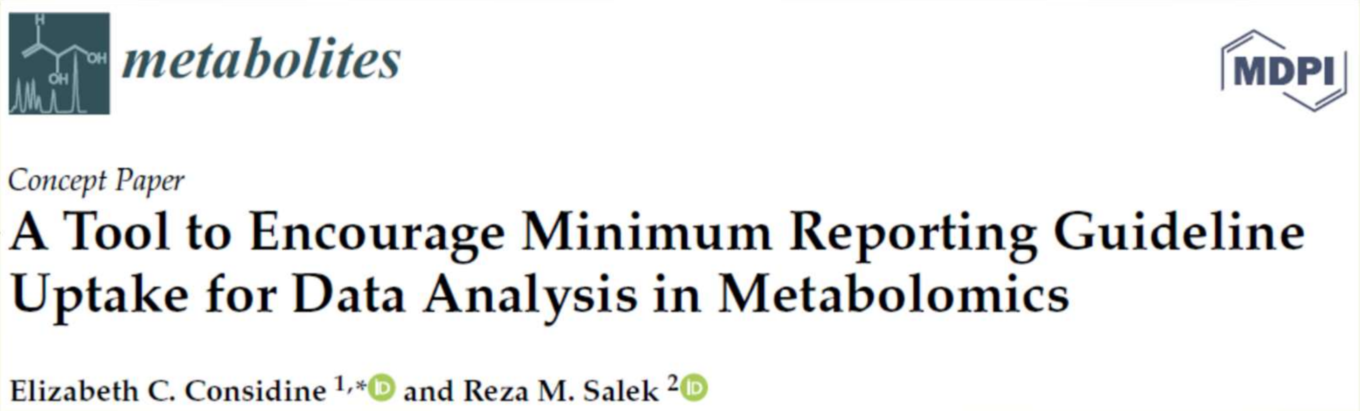
Data analysis

Not unexpectedly also data analysis should be made **reproducible**. The general trend is to promote “cloud” solutions, where “data mining” is performed as close as possible to “data storage” or to rely to open source solutions (mainly in R)

- ✓ Phenomenal (http://phenomenal-h2020.eu/home/)
- ✓ Metaboanalyst (https://www.metaboanalyst.ca/) - MetaboAnalystR
- ✓ Workflow4metabolomics (https://workflow4metabolomics.org/)
- ✓ XCMS online (https://xcmsonline.scripps.edu)

Data analysis reporting

Standardization should ultimately reach the articles published in scientific journals ... (Data Analysis Reporting Using **R Markdown**)



metabolites

MDPI

Concept Paper

A Tool to Encourage Minimum Reporting Guideline Uptake for Data Analysis in Metabolomics

Elizabeth C. Considine [1,*] and Reza M. Salek [2]

“R Markdown file is proposed as a starting point to encourage the data analysis section of metabolomics papers to have a more logical and stepwise presentation and to contain enough information to be understandable. So, even though R Markdown file only attends to the authoring and not the analysis of metabolomics data, we hope that it will coax data analysts into the environment of R Markdown (and GitHub), and therefore be a nudge along the road towards readable, and ultimately, reproducible, metabolomics research”

https://github.com/MSI-Metabolomics-Standards-Initiative/MIDAS

Metabolites **2019**, 9, 43; doi:10.3390/metabo9030043 http://metabolomicssociety.org/index.php/resources/metabolomics-standards

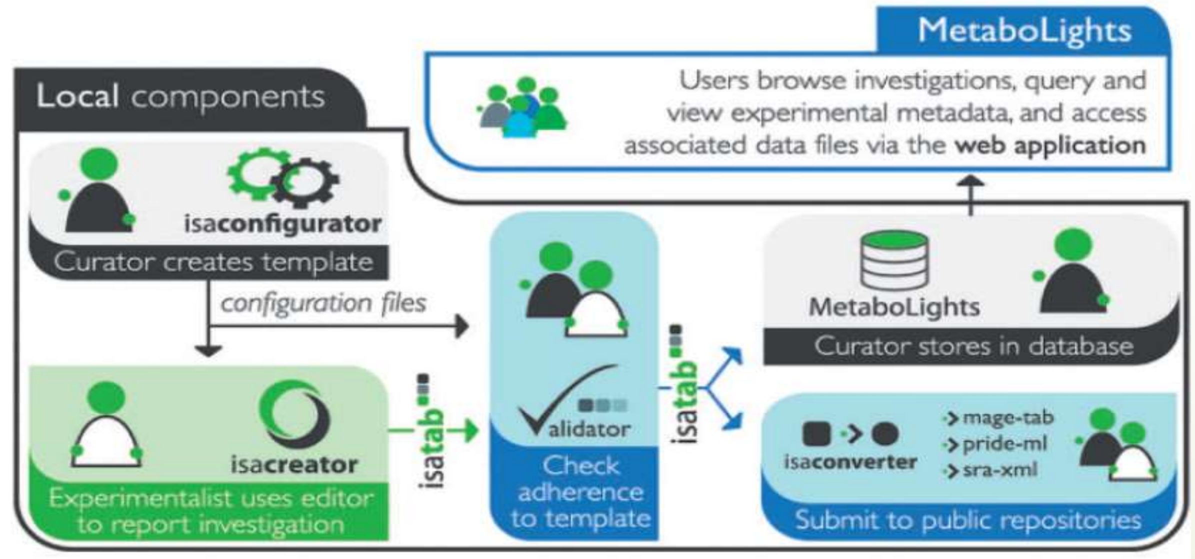
TWO CRUCIAL STEPS

- ✓ **STANDARDIZATION OF DATA AND METADATA**
- ✓ **SUBMISSION TO PUBLIC REPOSITORIES**

METABOLOMICS SOCIETY

MetaboLights

http://www.ebi.ac.uk/metabolights/

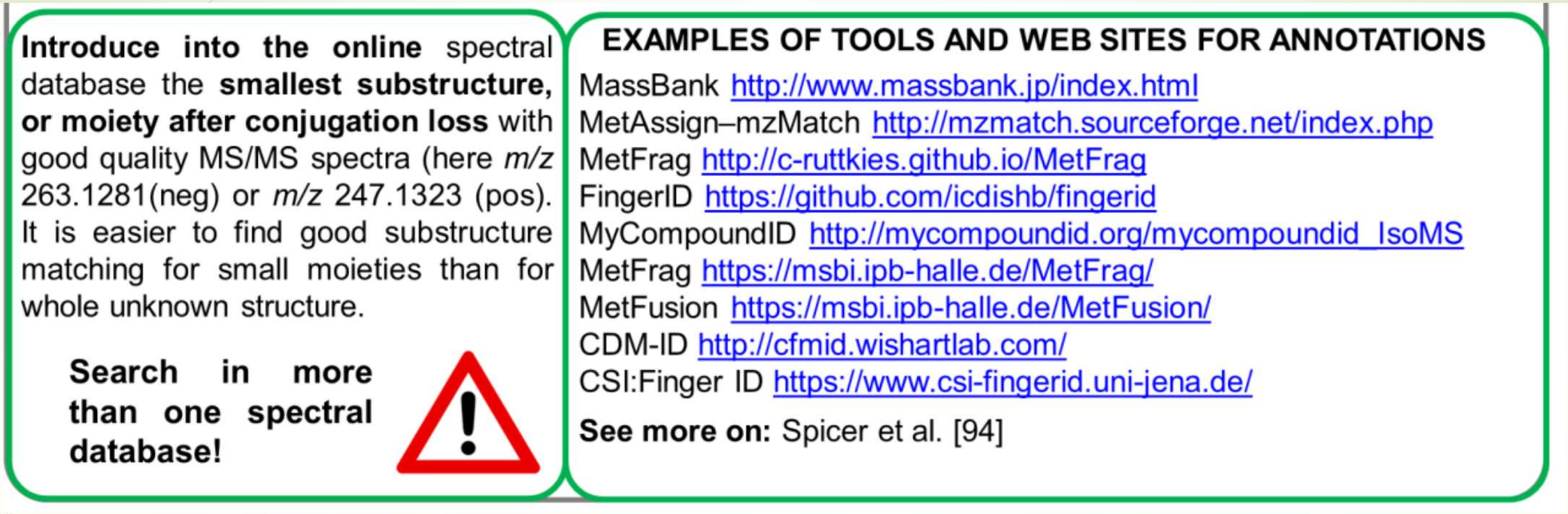


Users browse investigations, query and view experimental metadata, and access associated data files via the **web application**

- ✓ Haug K. et al., Nucl. Acids Res.2013 41(D1): doi: 10.1093/nar/gks1004
- ✓ Salek R.M. et al., Database, Vol. 2013, Article ID bat029, doi:10.1093/database/bat029
- ✓ Haug K. et. Al. Nucleic Acids Research, 2020, 48, doi: 10.1093/nar/gkz1019

Databases of Standards

Annotation of untargeted studies can be performed relying on community databases of spectra



Introduce into the online spectral database the **smallest substructure, or moiety after conjugation loss** with good quality MS/MS spectra (here *m/z* 263.1281(neg) or *m/z* 247.1323 (pos). It is easier to find good substructure matching for small moieties than for whole unknown structure.

Search in more than one spectral database!

EXAMPLES OF TOOLS AND WEB SITES FOR ANNOTATIONS

MassBank http://www.massbank.jp/index.html
MetAssign–mzMatch http://mzmatch.sourceforge.net/index.php
MetFrag http://c-ruttkies.github.io/MetFrag
FingerID https://github.com/icdishb/fingerid
MyCompoundID http://mycompoundid.org/mycompoundid_IsoMS
MetFrag https://msbi.ipb-halle.de/MetFrag/
MetFusion https://msbi.ipb-halle.de/MetFusion/
CDM-ID http://cfmid.wishartlab.com/
CSI:Finger ID https://www.csi-fingerid.uni-jena.de/

See more on: Spicer et al. [94]

hmdb The Human Metabolome Database

METLIN

MassBank

LIPID MAPS

m/z CLOUD Advanced Mass Spectral Database

On-line Databases: from elemental composition to reasonable compound proposal

❖HMDB (hmdb.ca)

❖My compound ID (www.mycompoundid.org)

❖METLIN (metlin.scripps.edu)

❖Lipidmaps (www.lipidmaps.org/data/structure/LMSDSearch.php?Mode=SetupTextOntologySearch)

❖Phenol-explorer (phenol-explorer.eu)

❖ChEBI (www.ebi.ac.uk/chebi/)

❖MassBank (www.massbank.jp)

❖mzCloud (https://www.mzcloud.org)

❖Chemspider (www.chemspider.com)

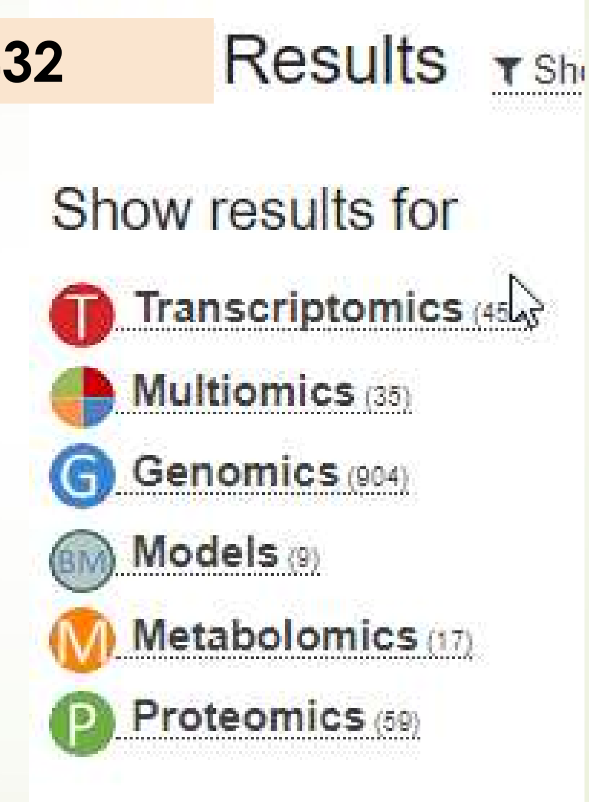
❖PubChem (pubchem.ncbi.nlm.nih.gov)

https://www.omicsdi.org/
search for (**grape* OR wine***)

26-03-2019

03-10-2021

1432



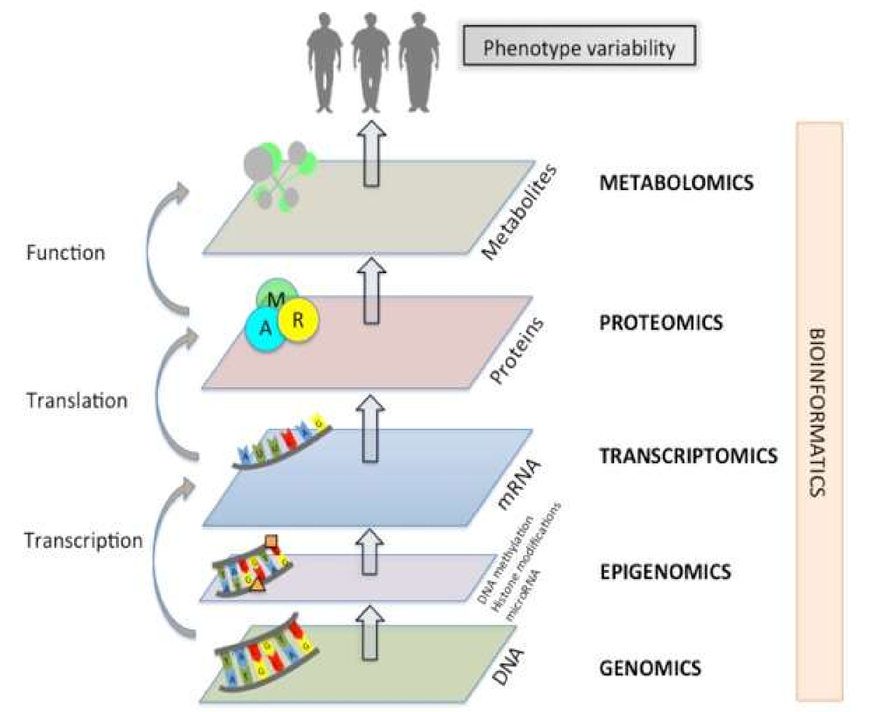
Results ▼ Sh

Show results for

- Transcriptomics (45
- Multiomics (35)
- Genomics (904)
- Models (9)
- Metabolomics (17)
- Proteomics (59)

5413

+ 278%



From D. Albuquerque British Medical Bulletin 123(1):1-15

Omics Discovery Index EMBL-EBI

Where are our data???

https://www.omicsdi.org/

search for omics_type: "**Metabolomics" AND (grape* OR wine*)**

26-03-2019 only 17 results

After manual curation, **only 10 were retained** (including also those focused on yeasts or on human metabolism);

1. **Metabolights (7)**
2. **MetabolomicsWorkbench (2)**
3. **GNPS (1)**

Submitters: Tobias Kind, Silvia dal Santo, Panagiotis Arapitsas, Ron Wehrens, Margaret Whitener, Luca Narduzzi, Alessia Trimigno, Devjanee Swain Lenz, Irene Stefanini, Justin van der Hooft

Institutions: University of California, Davis (1); University of Verona (1); University of Bologna (1); Washington University, St. Louis (1); Wageningen University (1), Fondazione Edmund Mach (5)

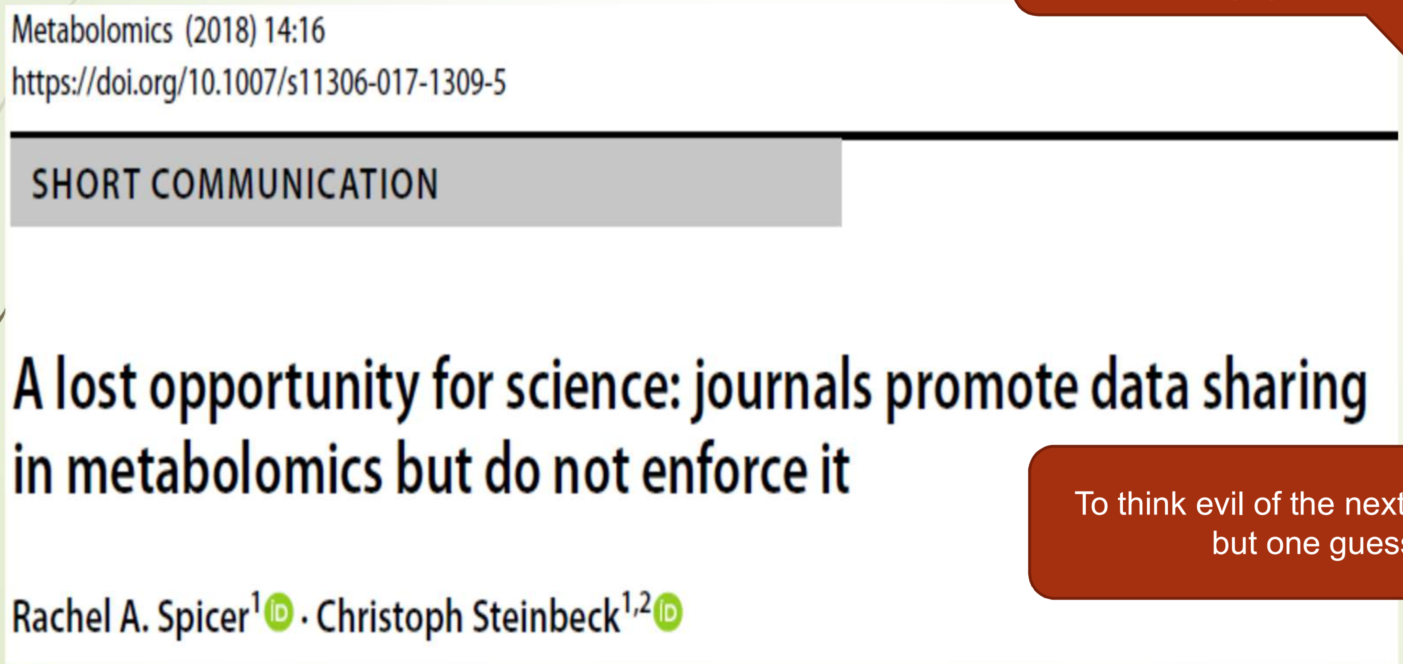
03-10-2021 only 89 results (67 retained)

1. **Biostudies (32)**
2. **MetaboLigths (25)**
3. **MetabolomicsWorkbench (9)**
4. **ENA (7)**
5. **GNPS (6)**
6. **BioModels (5)**
7. **All others (7)**

BioStudies database (at EMBL-EBI), gives a home to all of the data supporting a study providing one package for all the data. This database holds descriptions of biological studies, links to data from these studies in other databases at EMBL-EBI or outside, as well as data that do not fit in the structured archives at EMBL-EBI. The database accepts submissions via an online tool, or in a simple tab-delimited format. It also enables authors to submit supplementary information and link to it from the publication.

Everybody concede that there is a need to improve the findability, accessibility, interoperability and reusability (**FAIR**), for more transparent dissemination of these data. So….

Where are the data?



Metabolomics (2018) 14:16
https://doi.org/10.1007/s11306-017-1309-5

SHORT COMMUNICATION

A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it

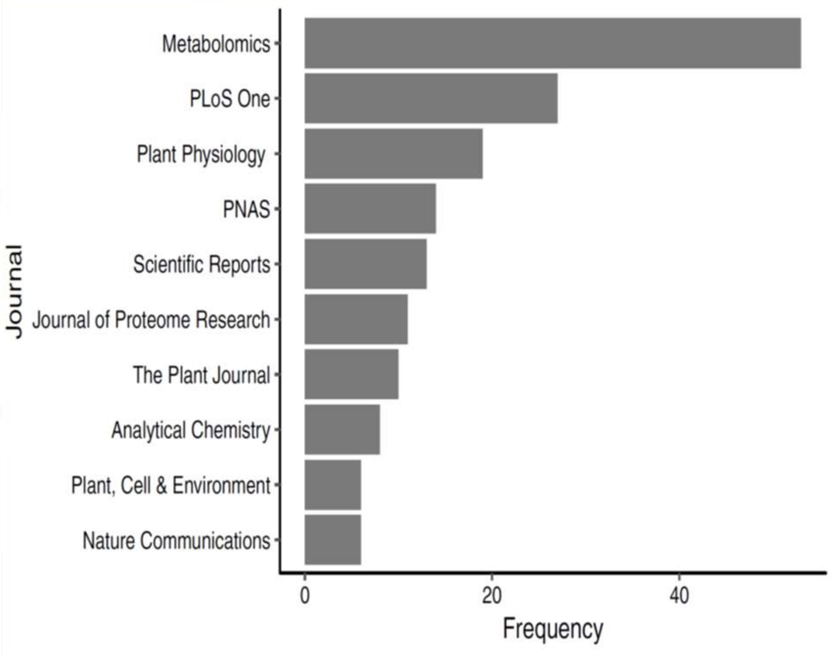
Rachel A. Spicer[1] · Christoph Steinbeck[1,2]

“A pensar male del prossimo si fa peccato ma si indovina” (Achille Ratti) papa Pio XI

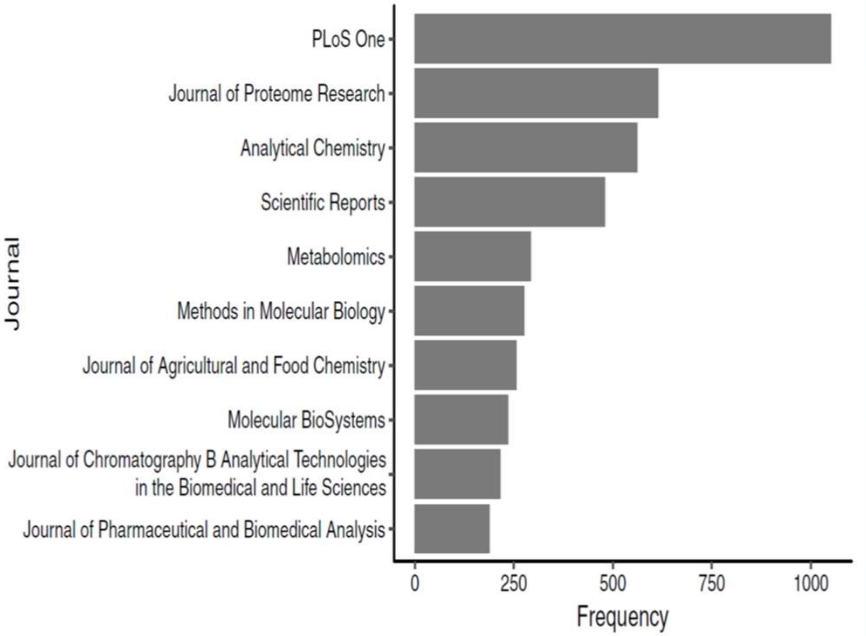
To think evil of the next one is a sin, but one guesses

Spicer RA & Steinbeck C. Metabolomics (2018) 14:16, doi:10.1007/s11306-017-1309-5

In metabolomics, journals that most support data sharing are not necessarily those with the highest number of papers associated to open metabolomics data. In more mature communities such as genomics, it has now become **the absolute default** that data **must be shared**. There must also be greater effort to improve the **linking of data to publications (and vice-versa).**



The ten journals with the highest frequency of publications directly linked from a publicly available metabolomics study, in a dedicated repository (MetaboLights, Metabolomics Workbench, MetaPhen, MeRy-B and GNPS)



The ten journals with the highest frequency of publications when searching PubMed for "metabolome" OR "metabolomics"

Spicer RA & Steinbeck C. Metabolomics (2018) 14:16, doi:10.1007/s11306-017-1309-5

Where are the metabolomics data? And why?

(possibly) **FAIR** data

(certainly) **NON-FAIR** data

Journals That Issue Open Science Badges

none in food & plant science

CENTER FOR OPEN SCIENCE

Addiction Research & Theory | Taylor & Francis
Advances in Archaeological Practice | Cambridge University Press
Advances in Methods and Practices in Psychological Science | SAG
AIS Transactions on Replication Research | Elsevier
American Journal of Orthopsychiatry | APA
American Journal of Political Science | Wiley
American Journal of Primatology | Wiley
Analyses of Social Issues and Public Policy (ASAP) | Wiley
Annual Review of Applied Linguistics | Cambridge University Press
Archive for the Psychology of Religion | SAGE
Asian American Journal of Psychology | APA
Big Earth Data | Taylor & Francis
BMC Microbiology (uses modified badge criteria) | BMC
BMJ Open Science | BMJ
Brain and Neuroscience Advances | SAGE
Canadian Journal of Experimental Psychology (CJEP) | APA
Clinical Psychological Science | APS
Cognitive Science | Wiley
Communication Studies | Taylor & Francis
Communication Research Reports | Taylor & Francis
Cortex | Elsevier
Cultural Diversity & Ethnic Minority Psychology | APA
Decision | APA
Ear and Hearing | Wolters Kluwer
Emerging Adulthood | SAGE
Environmental Toxicology and Chemistry | Wiley
European Journal of Personality | Wiley

Evolution and Human Behavior | Elsevier
Exceptional Children | SAGE
Geoscience Data Journal | Wiley
Gifted Child Quarterly | SAGE
International Gambling Studies | Taylor & Francis
International Journal for the Psychology of Religion | Taylor &
International Journal of Primatology | Springer Nature
Internet Archaeology | University of York
Journal of Behavioral Public Administration (JBPA)
Journal of Cognition and Development | Taylor & Francis
Journal of Comparative Psychology | APA
Journal of Experimental Psychology: Learning, Memory, and Cognition | APA
Journal of Experimental Social Psychology | Elsevier
Journal of International Crisis and Risk Communication Research | Nicholson School of Communication and Media
Journal of Maps | Taylor & Francis
Journal of Neuroendocrinology | Wiley
Journal of Neurochemistry | Wiley
Journal of Neuroscience Research (JNR) | Wiley
Journal of Personality Assessment | Taylor & Francis
Journal of Psychiatric and Mental Health Nursing | Wiley
Journal of Social Psychology | Taylor & Francis
Journal of Research in Personality | Elsevier
Journal of Research on Educational Effectiveness | Taylor & Fr
Journal of Threat Assessment and Management | APA
Language Awareness | Taylor & Francis

Language Learning | Wiley
Language Testing | SAGE
Law and Human Behavior | APA
Management and Organization Review | Cambridge University
Media Psychology | Taylor & Francis
Meta-Psychology | Linnaeus University Press
Neuropsychology | APA
Neuroscience of Consciousness | Oxford University Press
Phosphorus, Sulfur, and Silicon and the Related Elements | Tay Francis
Political Communication | Taylor & Francis
Psi Chi Journal of Psychological Research | Psi Chi
Psychological Science | SAGE
Psychological Methods | APA
Psychology of Addictive Behaviors | APA
Psychology of Men & Masculinity | APA
Psychology of Popular Media Culture | APA
Public Administration Review | ASPA
Quantitative Finance | Taylor & Francis
Quarterly Journal of Experimental Psychology | SAGE
Sexual Abuse | SAGE
Social Psychology | Hogrefe
Strategic Management Journal | Wiley
Studies in Second Language Acquisition | Cambridge Universit
Teaching of Psychology | SAGE
The Modern Language Journal | Wiley
The Photogrammetric Record | Wiley

It is certainly NOT in the top priorities of an early stage researcher!

🙂

1) Acquire skills in the specific fields (continuous process); 🙂

2) Perform (at least part of) the experiment and collect the data; 🙂

3) Organize the data into tables and figures, and participate to the interpretation;

4) Publish papers (as many as you can!); 😕

5) Gain visibility and recognition (internal and external); 😕

6) Get grants (at least try!); 🙄

7) Comply with (multiple) deadlines (not mentioning academic bureaucracy!) 🤐

…..

56) **Upload data and metadata on public repositories** (since unfortunately my professor wants.. 😤

Good news! It is not that difficult!

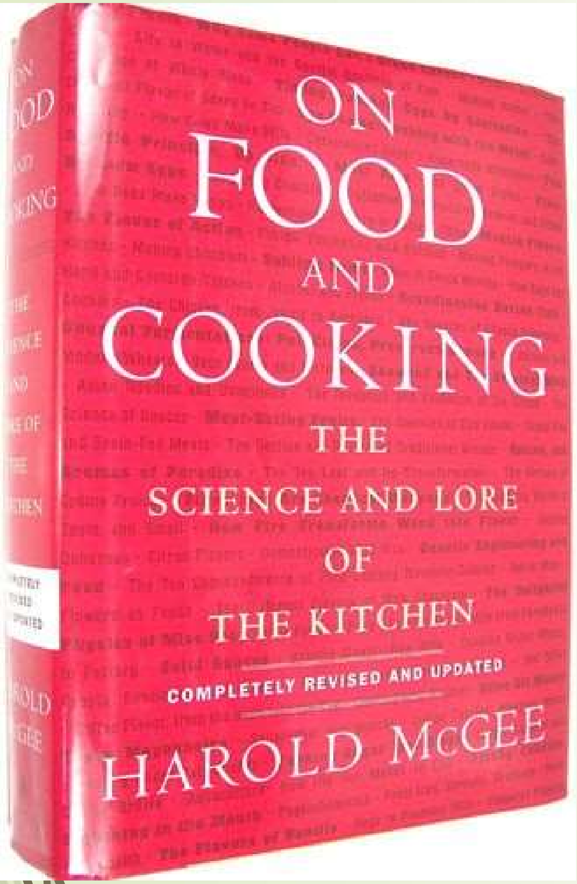
https://www.**wikihow**.com/Make-an-Apple-Pie

PARTS

1. Making the Pastry Dough
2. Mixing the Apple Filling
3. Assembling the Pie
4. Baking the Apple Pie

OTHER SECTIONS

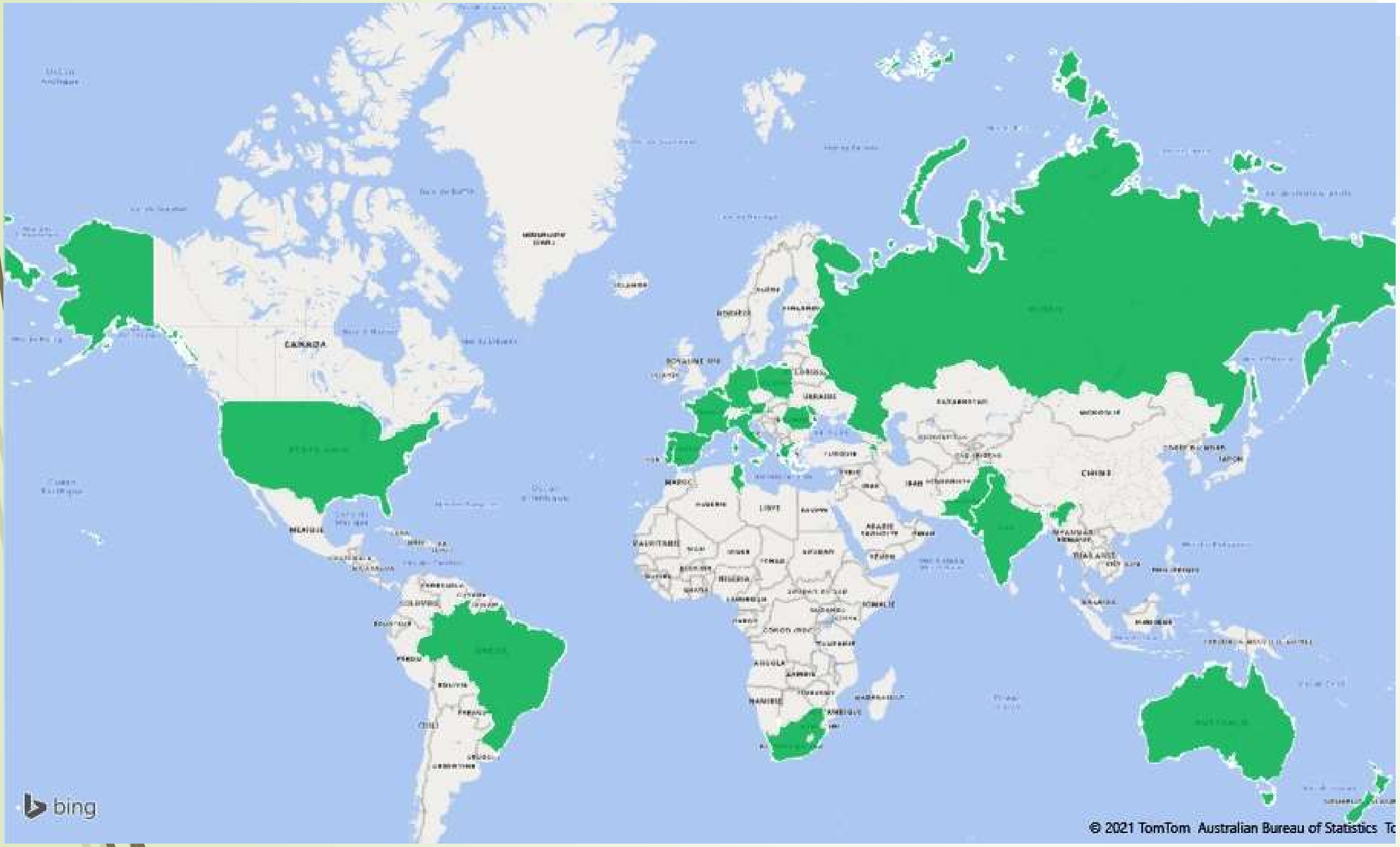
- Video
- References
- Article Summary


ON FOOD AND COOKING
THE SCIENCE AND LORE OF THE KITCHEN
COMPLETELY REVISED AND UPDATED
HAROLD McGEE

A really great book!

Registration from all the continents !!

COST Action
CA17111
INTEGRAPE



© 2021 TomTom Australian Bureau of Statistics To

97 registered for the introductory session, majority female, 54 (4th Oct 2021, today)
23 persons will attend the practical sessions (5th & 6th Oct 2021)

Welcome!

People are engaged at every step in the data value chain in collecting, analyzing, interpreting, and using data. In many cases, people themselves are data points…… **If we want data to work together, we need people to work together. We need human interoperability.**

Steven Ramage & Jenna Slotin
August 25, 2021
(https://www.data4sdgs.org/news/why-people-are-essential-data-interoperability)

COUNTRY	N°
Spain	29
Italy	12
France	9
South Africa	7
Portugal	5
United States	5
India	4
Poland	4
Australia	3
Belgium	3
Germany	2
Greece	2
Israel	2
Armenia	1
Austria	1
Brasil	1
Cyprus	1
New Zealand	1
Pakistan	1
Republic of Moldova	1
Romania	1

AFFILIATIONS (the most representative)	N°
Universidad Zaragoza	10
University of Barcelona	9
Stellenbosch University	6
Fondazione Edmund Mach	4
Pomeranian Medical University in Szczecin	4
BIOISI, Faculdade de Ciências da Universidade de Lisboa	3
INRAE	3
University of Kentucky	3
Australian Wine Research Institute	2
Ben-Gurion University of the Negev	2
CREA - Research Center for Viticulture and Enology	2
CSIC	2
E. & J. Gallo	2
Enveda Biosciences	2
Instituto de Ciencias de la Vid y del Vino	2
University of the Balearic Islands	2
University of Verona	2
URCA	2
ICAR National Research Centre for Grapes	2

Thanks!

COST Action
CA17111
INTEGRAPE

https://integrape.eu/resources/data-management/

- ✓ how-to-describe-a-grapevine-experiment
- ✓ How to submit sequence data to ENA
- ✓ How to submit metabolomic data to MetaboLights
- ✓ How to standardize JBrowse’s tracks (under construction)
- ✓ Apollo Manual Curation Guide for the PN40024.v4 assembly (u.c.)