The Overlap Between FAIR for Research Software and Open Science

Daniel S. Katz (d.katz@ieee.org, @danielskatz)
Chief Scientist, NCSA
Associate Research Professor, CS, ECE, iSchool
University of Illinois at Urbana Champaign

Co-authors: Michelle Barker, Neil P. Chue Hong, Leyla Jael Castro, Morane Gruenpeter, Jennifer Harrow, Carlos Martinez, Paula Andrea Martinez, Fotis E. Psomopoulos

ILLINOIS
NCSA | National Center for Supercomputing Applications

9 March 2022

Open Science Conference

https://doi.org/10.5281/zenodo.6340732

Why do we care about research software?

- Funding
  - ~20% of NSF projects over 11 years topically discuss software in their abstracts ($10b)
  - 2 of 3 main DOE ECP areas are research software (~$4b)
- Publications
  - Software intensive projects are a majority of current publications
  - Most-cited papers are methods and software
- Researchers
  - >90% of US/UK researchers use research software
  - ~65% would not be able to do their research without it
  - ~50% develop software as part of their research

Collected from http://www.dia2.org in 2017

Nangia and Katz; 10.1109/eScience.2017.78
"Top 100-cited papers of all time," Nature, 2014 10.1038/514550a

S. Hettrick; https://www.software.ac.uk/blog/2016-09-12-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers
S.J. Hettrick, et al,; 10.5281/zenodo.14809
U. Nangia and D. S. Katz; 10.6084/m9.figshare.5328442.v1

https://doi.org/10.5281/zenodo.6340732

Research and research software vision

- All research software that can be is open — Open Science
- All research software is high-quality and robust — Software Engineering
- All research software is findable, accessible, and usable & used by others (for their own research) — FAIR
  - And is cited when it is used
- All contributors to research software are recognized for their work — Software Citation, JOSS
  - With good careers — RSE +
- All research software is sustained as long as it is useful — SSI, URSSI, ARDC
- All research is reproducible — Reproducibility

Note overlaps in terms of incentives and policies; all start with recognition of research software

https://doi.org/10.5281/zenodo.6340732

Open Science (Collaboration)

- The free sharing of scientific ideas, methods, and results
- But not just science, rather wissenschaft (knowledge, scholarship, …)
- Initially via hand-written letters and books, mostly for other scientists
- Then more frequently via printed journals, expanding the audience
- Digitalization expanded opportunities for sharing, as well as what could be shared
- Democratization of research (public funding) and information sharing (BBS, WWW) expanded the community (at least the audience)
- Idea of knowledge as a common (societal) good
- *R. Şentürk, "Toward an Open Science and Society: Multiplex Relations in Language, Religion and Society -Revisiting Ottoman Culture-," islôm Araştırmaları Dergisi. 2001 93-129*
  - *"In this paper I introduce a new concept, 'open science,' to denote a pluralist and democratic science culture"*

J. P. Tennant, …, D. S. Katz, …, "A tale of two 'opens': intersections between Free and Open Source Software and Open Scholarship," SocArXiv, 6 Mar. 2020. DOI: 10.31235/osf.io/2kxq8

https://doi.org/10.5281/zenodo.6340732

Economics drives our lives (Competition)

- We live in a capitalistic society
- Economics drives our lives and careers
  - Where we work (hiring)
  - How we support ourselves (promotion)
  - How we get funding to do science (support, recognition)
  - Which science we do (what areas we think will lead to reward)
  - Which students we train or take advantage of (depending on your viewpoint)
- Economics: the science of allocating scarce resources to maximize the achievement of competing ends
  - Sometimes a false argument, some resources can be increased, e.g., digital

Paula Stephan, How Economics Shapes Science, Harvard University Press, 2015.

https://doi.org/10.5281/zenodo.6340732

Why not Open Science

- Sharing takes effort, immediate benefits go to others
- Mechanisms of sharing are new, not the way we work
- Metrics for evaluating products that can be shared are underdeveloped
- Intellectual property laws
- Commercial entities profit from restricting access
- Non-profit scientific societies are dependent on journal subscription fees [to support themselves/work they do]

C. Titus Brown, “What is open science?,”24 October 2016. http://ivory.idyll.org/blog/2016-what-is-open-science.html

https://doi.org/10.5281/zenodo.6340732

Human behavior (Competition & Collaboration)

- Engagement: meaningful and valuable actions that produce a measurable result
- Engagement = Motivation + Support – Friction
  - Intrinsic motivation: self-fulfillment, altruism, satisfaction, accomplishment, pleasure of sharing, curiosity, real contribution to science
  - Extrinsic motivation: job, rewards, recognition, influence, knowledge, relationships, community membership
  - Support: ease, relevance, timeliness, value
  - Friction: technology, time, access, knowledge

Adapted from Joseph Porcelli

https://doi.org/10.5281/zenodo.6340732

Open Science

- Terms
  - Open access
  - Open data
  - Open source
  - Open governance
  - Open use
- Meanings
  - Open to read
  - Open to use/run
  - Open to build
  - Open to change
  - Open to work for a group / collaborate
- Items
  - Preprint/paper
  - Elements: text, figures, citations
  - Notebook
  - Data
  - Software
  - ML model
  - Protocol/method
  - Research plan
  - DMP (SMP)
  - Study
  - Standard

Mapping as future work

https://doi.org/10.5281/zenodo.6340732

FAIR data

- Vision for a scientific commons
  - Context: Open access movement, from 2001 Budapest Open Access Initiative (BOAI)
- Started with 2014 Leiden workshop, “Jointly Designing a Data Fairport”
- Built to change practices from closed to open
- Based on current practices of publishing (often closed)
- And lack of data sharing
- Using data repositories
- But no requirement for openness
- Pragmatic

https://doi.org/10.5281/zenodo.6340732

The FAIR Principles


The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

194k Accesses | **2450** Citations | **1852** Altmetric | Metrics

A set of principles, to ensure that data are shared in a way that enables and enhances reuse by humans and machines

Findable

F1. (Meta)data are assigned a globally unique and eternally persistent identifier.

F2. Data are described with rich metadata.

F3. (Meta)data are registered or indexed in a searchable resource.

F4. Metadata specify the data identifier.

Accessible

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1. The protocol is open, free, and universally implementable.

A1.2. The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the data are no longer available.

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles.

I3. (Meta)data include qualified references to other (meta)data.

Reusable

R1. (Meta)data have a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

R1.2. (Meta)data are associated with their provenance.

R1.3. (Meta)data meet domain-relevant community standards.

https://doi.org/10.5281/zenodo.6340732

FAIR for non-data objects: some context

- FAIR Principles, at a high level, are intended to apply to all research objects; both those used in research and those that are research outputs
- Text in principles often includes "(Meta)data …"
  - Shorthand for "metadata and data …"
- Principles applied via dataset creators and repositories, collectively responsible for creating, annotating, indexing, preserving, sharing the datasets and their metadata
  - Assumes separate and sequential creator/publisher (repository) roles
- What about non-data objects?
  - While they can often be stored as data, they are not just data
- While high level goals (F, A, I, R) are mostly the same, the details and how they are implemented depend on
  - How objects are created and used
  - How/where the objects are stored and shared
  - How/where metadata is stored and indexed
- Work needed to define, then implement, then adopt principles

https://doi.org/10.5281/zenodo.6340732

Need for FAIR for non-data objects

- FAIR Principles, are intended to apply to all digital objects (Wilkinson et al. 2016)

> ***Recommendation n°5*** *:*
>
> *Recognise that FAIR guidelines will require* ***translation for other digital objects*** *and support such efforts.*

2020: 'Six Recommendations for Implementation of FAIR Practice'

(FAIR Practice Task Force EOSC, 2020)

https://doi.org/10.5281/zenodo.6340732

FAIR for non-data objects: some efforts

Ten simple rules for making training materials FAIR

Leyla Garcia, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, Niall Beard, Denise Carvalho-Silva, Alexandros C. Dimopoulos, Victoria Dominguez del Angel, Michel Dumontier, Kim T. Gurwitz, [...], Patricia M. Palagi [view all]

Published: May 21, 2020 • https://doi.org/10.1371/journal.pcbi.1007854

PLOS COMPUTATIONAL BIOLOGY

January 01 2020

FAIR Computational Workflows

Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, Daniel Schober

> Author and Article Information

Data Intelligence (2020) 2 (1-2): 108–121.

https://doi.org/10.1162/dint_a_00033

DATA Intelligence

Steps towards defining FAIR principles for Machine Learning (ML)

Home

28 JUL 2021

Steps towards defining FAIR principles for Machine Learning (ML)

Submitted by Fotis Psomopoulos

Breakout 7 | Data Infrastructures - Organisa… | The FAIR Agenda | WGs Getting started

WG FAIR for Virtual Research Environments: FAIR for VREs - The Path Forward

7:30 AM - 9:00 AM

Room E

RDA RESEARCH DATA ALLIANCE

ILLINOIS NCSA

https://doi.org/10.5281/zenodo.6340732

Software vs. data

- Software is data, but it is not just data
  - Software is executable, data is not
  - Data provides evidence, software provides a tool
  - Software is a creative work, scientific data are facts or observations
    - Different licensing and copyright practices
  - Software suffers from a different type of bit rot than data
    - It is frequently built to use other software, leading to complex dependencies, and these dependent software pack- ages also frequently change
  - The lifetime of software is generally not as long as that of data
  - For open source, no natural sequential creator/publisher process & no natural publisher (repository)

D. S. Katz et al., "Software vs. data in the context of citation," PeerJ Preprints 4:e2630v1, 2016. https://doi.org/10.7287/peerj.preprints.2630v1

https://doi.org/10.5281/zenodo.6340732

Free software (Stallman/Gnu)

- The initial free software freedoms
  - First, the freedom to copy a program and redistribute it to your neighbors, so that they can use it as well as you
  - Second, the freedom to change a program, so that you can control it instead of it controlling you; for this, the source code must be made available to you.
- Currently:
  - The freedom to run the program as you wish, for any purpose (freedom 0).
  - The freedom to study how the program works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.
  - The freedom to redistribute copies so you can help others (freedom 2).
  - The freedom to distribute copies of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

GNU'S BULLETIN, v.1(1), October 1986. https://www.gnu.org/bulletins/bull1.txt

GNU Operating System, "What is Free Software?," 11 October 2021. https://www.gnu.org/philosophy/free-sw.html

https://doi.org/10.5281/zenodo.6340732

FAIR for Research Software (FAIR4RS)

- Working group defining FAIR principles for research software
  - Led by Michelle Barker, Neil Chue Hong, Leyla Garcia, Morane Gruenpeter, Jennifer Harrow, Daniel S. Katz, Carlos Martinez, Paula A. Martinez, Fotis Psomopoulos

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

<ReSA>
Research Software Alliance

https://doi.org/10.5281/zenodo.6340732

FAIR4RS initial subgroups

1. A fresh look at FAIR for Research Software
   - Examined the FAIR principles in the context of research software from scratch, not based on pre-existing work; published: Katz DS, Gruenpeter M, Honeyman T, et al. (2021). A Fresh Look at FAIR for Research Software. arXiv:2101.10883 [cs.SE], https://arxiv.org/abs/2101.10883
2. FAIR work in other contexts
   - Analyzed how FAIR principles are applied to research objects other than data/software – final report
3. Research software definition
   - Reviewing existing definitions and to specify the scope for the WG outputs – final report
4. New research related to FAIR Software
   - Review recent research and studies around FAIR software
   - Via up-to-date identification of approaches that can help structure FAIR4RS work, in form of Zotero reading list and short report on important insights from review and survey – draft report

https://doi.org/10.5281/zenodo.6340732

Working group status

- ~40 webinars and talks overall
- Jan – Feb 2021: Initial analysis of subgroup work led to a set of questions
- March 2021: Working group's input on these questions published
- April 2021: Group leads+ held writing sprint and assembled draft from subgroup products and initial community input
- 17 – 30 May 2021: Working group review of initial draft
- 11 June – 11 July 2021: Official community review (part of the RDA process) of second draft
- Now: WG is drafting final v1.0 FAIR4RS principles for RDA & journal

https://doi.org/10.5281/zenodo.6340732

Defining Research Software

- What is software?
  - Conceptually, software can mean a project or entity; a community around a project; or a software idea, algorithms, solutions
  - A software artifact can be source code, binaries, executables, containers
- What is the role of software in the research process?
  - It can be a tool, a research outcome or result, or the object or the research

- **Research Software** includes source code files, algorithms, scripts, computational workflows and executables that were **created during the research process or for a research purpose**
- Additional software components (e.g., operating systems, libraries, dependencies, packages, scripts, etc.) that are used for research but were **not created during or with a clear research intent should be considered software in research** and not Research Software
- This differentiation may vary between disciplines

https://doi.org/10.5281/zenodo.5504016

https://doi.org/10.5281/zenodo.6340732

FAIR Principles for Research Software

- Available version is 0.x, not yet 1.0
- Concepts won't change in v1.0, but some language will
- Citation and download:
  - Hong, N. P. C., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2021). FAIR Principles for Research Software (FAIR4RS Principles). Research Data Alliance. DOI: 10.15497/RDA00065
  - v1.0 will have DOI: 10.15497/RDA00068 (but isn't available yet)

https://doi.org/10.5281/zenodo.6340732

F Principles

Findable: Software, and its associated metadata, is easy to find for both humans and machines.

- F1. Software is assigned a globally unique and persistent identifier
  - F1.1. Different components of the software are assigned distinct identifiers representing different levels of granularity
  - F1.2. Different versions of the same software are assigned distinct identifiers
- F2. Software is described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the software they describe
- F4. Metadata are FAIR and are searchable and indexable

https://doi.org/10.5281/zenodo.6340732

A Principles

Accessible: Software, and its metadata, is retrievable via standardized protocols.

- A1. Software is retrievable by its identifier using a standardized communications protocol
  - A1.1. The protocol is open, free, and universally implementable
  - A1.2. The protocol allows for an authentication and authorization procedure, where necessary
- A2. Metadata are accessible, even when the software is no longer available

https://doi.org/10.5281/zenodo.6340732

I Principles

Interoperable: Software interoperates with other software through exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.

- I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards
- I2. Software includes qualified references to other objects

https://doi.org/10.5281/zenodo.6340732

R Principles

Reusable: Software is both usable (it can be executed) and reusable (it can be understood, modified, built upon, or incorporated into other software).

- R1. Software is described with a plurality of accurate and relevant attributes
  - R1.1. Software is given a clear and accessible license
  - R1.2. Software is associated with detailed provenance
- R2. Software includes qualified references to other software
- R3. Software meets domain-relevant community standards

https://doi.org/10.5281/zenodo.6340732

FAIR4RS principles

Findable: Software, and its associated metadata, is easy to find for both humans and machines.

F1. Software is assigned a globally unique and persistent identifier

- F1.1. Different components of the software are assigned distinct identifiers representing different levels of granularity
- F1.2. Different versions of the same software are assigned distinct identifiers

F2. Software is described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the software they describe

F4. Metadata are FAIR and are searchable and indexable

Accessible: Software, and its metadata, is retrievable via standardized protocols.

A1. Software is retrievable by its identifier using a standardized communications protocol

- A1.1. The protocol is open, free, and universally implementable
- A1.2. The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the software is no longer available

Interoperable: Software interoperates with other software through exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.

I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards

I2. Software includes qualified references to other objects

Reusable: Software is both usable (it can be executed) and reusable (it can be understood, modified, built upon, or incorporated into other software).

R1. Software is described with a plurality of accurate and relevant attributes

- R1.1. Software is given a clear and accessible license
- R1.2. Software is associated with detailed provenance

R2. Software includes qualified references to other software

R3. Software meets domain-relevant community standards

Unpublished v1.0 language; older version is "FAIR4RS WG. (2021, June). FAIR Principles for Research Software (10.15497/RDA00065)"

https://doi.org/10.5281/zenodo.6340732

Personal view of FAIR4RS status

- Original FAIR principles mixed metadata and data, e.g., “(Meta)data,” too strongly
  - Much of the metadata part translates directly to metadata about software
  - The data part doesn't
- F & A: basically not changed, but gaps appear
- I & R: multiple possible definitions that need to be resolved
- Lots of ecosystem gaps (open questions), particularly related to metadata, archiving, versions
  - Creator/publisher sequence doesn’t typically apply
  - Where is metadata stored? (in code repository for open source?, for closed source?, in archival repository?, in registry?)
  - Where is code archived? (GitHub/Gitlab are not archival, registries are not archival, repositories? Software Heritage?)
  - Different use cases need specific version, latest version, all versions

https://doi.org/10.5281/zenodo.6340732

Current steps

- We've formed subgroups on adoption and future governance, now wrapping up
- Adoption guidelines
  - Developing guidelines and instructions (checklists, how-tos, …) for how to make research software FAIR
- Adoption support
  - Identifying early adopters of the FAIR4RS principles and sharing their lessons & results
  - Currently includes American Geophysical Union (AGU), Digital Research Alliance of Canada (formerly NDRIO), Dutch Research Council (NWO), ESMAValTool, German Aerospace Center (DLR), Karlsruhe Institute of Technology (KIT), National Institute of Standards and Technology (NIST), Netherlands eScience Center, Network for Computational Modeling in the Social and Ecological Sciences (CoMSES Net), Nordic Collaboration on e-Infrastructures for Earth System Modelling (NeIC-NICEST), Universidad Politécnica de Madrid (UPM), ZB MED Information Centre for Life Sciences
- Governance
  - Determining the governance structure of the FAIR4RS principles to the community from the release onward
    - Who provides official answers to questions about FAIR4RS, interpretation of principles?
    - What happens if changes are needed (v1.1, v2.0)?

https://doi.org/10.5281/zenodo.6340732

Future steps

- Work on metrics
  - Measuring the FAIRness of specific research software
  - Measuring the adoption of FAIR for research software in an organization
  - Measuring the FAIRness of the set of all research software
- Incentives and policies
  - Use lessons from early adopters to consider changes in culture
  - Implement changes in policies of research institutions, publishers, funders, repositories, societies, …

https://doi.org/10.5281/zenodo.6340732

What's missing from FAIR

- Replicable evaluation
- Correctness, quality
- Credit, attribution
- Openness as a requirement?
- Sharing via license vs sharing via community/governance
- Open development, not just sharing at the end

https://doi.org/10.5281/zenodo.6340732

Summary

- Open science and open source have an interwoven history; both seem to be moving forward
- FAIR (data) principles set out good goal: ensure that data are shared in a way that enables and enhances reuse by humans and machines
- Work is needed to apply this goal to research software, both open source and not, to fulfil the open science concepts
- Principles have been created, with ~500 people involved & ~60 events
  - Draft: FAIR4RS WG. (2021, June). FAIR Principles for Research Software. 10.15497/RDA00065
- Open science, open software, FAIR4RS all open communities, with significant overlap
- Now finalizing and publishing v1.0 of principles
  - v1.0 will be 10.15497/RDA00068
- Work underway to create guidance, adopt principles, define governance
- Next steps will be to create metrics and widen adoption

https://doi.org/10.5281/zenodo.6340732