All's FAIR in Love and War: What About Research Data?

Stuart J. Chalk
Department of Chemistry
schalk@unf.edu

UNF UNIVERSITY of NORTH FLORIDA.

COAS Scholars Lecture – Spring 2021

"All's FAIR in Love and War"

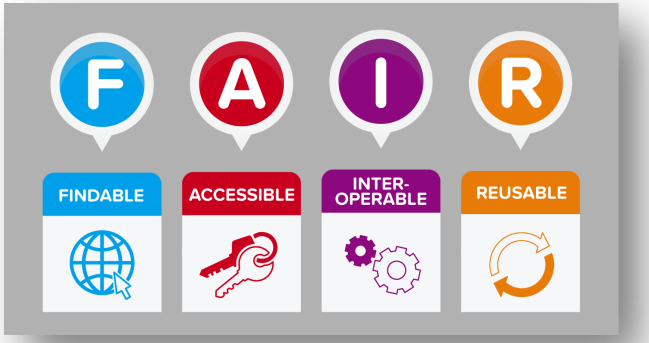
* Attributed to John Lyly, from his work 'Euphues: The Anatomy of Wit'.

Love, *subs.* **1. All's fair in love and war.** *c.*1578: Lyly, *Euphues*, I 236, Anye impietie may lawfully be committed in love, which is lawlesse. 1620: T. Shelton, tr. *Don Quixote*, ii xxi, Love and warre are all one … It is lawfull to use sleights and stratagems to … attaine the wished end. *c.*1630: B.&F., *Lovers' Progress*, V ii, All stratagems In love, and that the sharpest war, are lawful. 1687: A. Behn, *Emp. of the Moon*, I iii, Advantages are lawful in love and war. 1710: Centlivre, *Man's Bewitch'd*, V i, Stratagems ever were allow'd of in love and war. 1850: Smedley, *Frank Fairlegh*, ch 1. 1906: Lucas, *Listener's Lure*, 196.

https://www.google.com/books/edition/Dictionary_of_Proverbs/7PMZJqSR4sAC?hl=en&gbpv=1&dq=%22Love+and+warre+are+all+one%22&pg=PA355&printsec=frontcover

Outline

* NSF FAIR Research Data Project
* Current Practices in Research
* What is Open Science, and what is it not?
* The State of Research Data Today
* What does FAIR Mean?
* Example of FAIR Data
* Final Thoughts



https://kidsfirstdrc.org/news/fair-data/

NSF FAIR Research Data Project

Award:
#1835643

Project Description

RUI: Framework: Data - An Open Semantic Data Framework for Data-Driven Discovery

Introduction and Significance

The paradigm of open data has now become a global movement that promises to forever change the way in which science is communicated and ultimately proffer a new workflow for scientific discovery. As a result, there are open access journals, data journals, reports on the importance/value of open sharing of data[1], and new organizations launched specifically to pioneer studies on and recommendations for this new discipline[2]. In addition, the development of the findability, accessibility, interoperability, and reusability (FAIR) principles[3-5] give a tangible perspective on open data and provide a way to describe what open data means, and metrics to describe how open it is[6].

https://www.nsf.gov/awardsearch/showAward?AWD_ID=1835643

NSF FAIR Research Data Project

Award:
#1835643

* Convert heterogeneous publicly available research data and convert it to a FAIR format for semantic analysis
* Teaching PostDoc – Chemical Informatics
* Funding for 30 undergraduate research students

https://www.nsf.gov/awardsearch/showAward?AWD_ID=1835643

FAIR Research?

The Traditional Research Approach

* Write a grant and get funding to do some research (optional)
* Develop and run experiments and collect the raw data
* Work up the raw data into results
* Analyze the results relative to the hypothesis -> findings/conclusions
* Write up a paper detailing what you did and what you found
* Publish the paper in a ***non open-access peer-reviewed journal***
  * …and pay for (subscription) access to the journal to read your own paper!
    -OR-
* Publish the paper in an ***open-access peer-reviewed journal***
  * …and pay page charges ($1-10K) to publish so everyone can read it

FAIR Research?

The Traditional Research Approach

* What are the problems with this approach?
  * Research papers are an incomplete record of the research
  * Research that is published is the 'best data'
  * Typically, you can't get actual research data even from publication
  * Supplemental Information is in most cases difficult to use…
  * … and thus reproduce
  * It costs researchers/institutions money!

FAIR Research?
The Open Science (Open Research) Approach

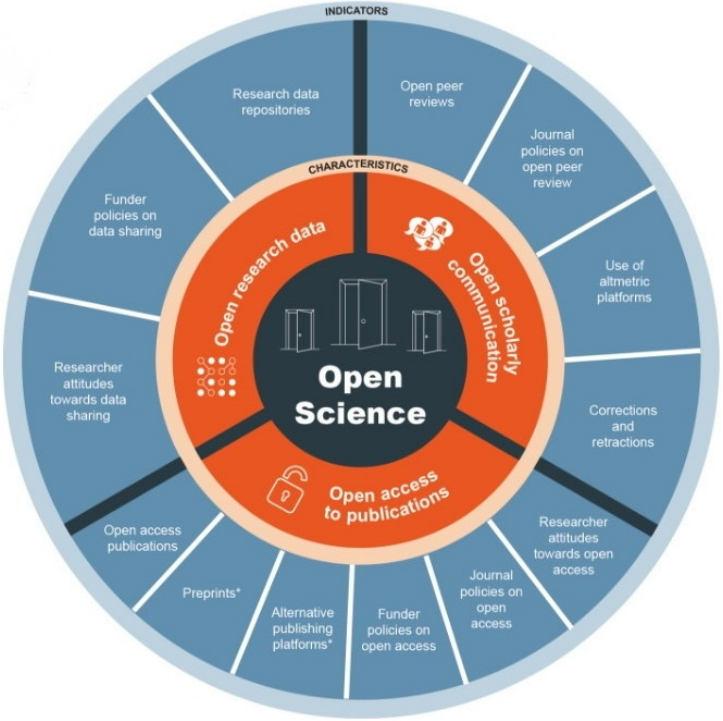
'*Journal papers should be supplements to data*' – Barend Mons (GO FAIR)

* Based on the ideals of doing science:
  * Do research for the greater good (not personal gain)
  * Ethically perform unbiased research (1)
  * Be honest in the analysis of research data
  * Transparently communicate the approach, methodologies, data, and results of research to the scientific community (2)
* Communicate about research data via mechanisms that allow others to locate, access, understand, and reuse it
* Open science ≠ Free science (in terms of cost)

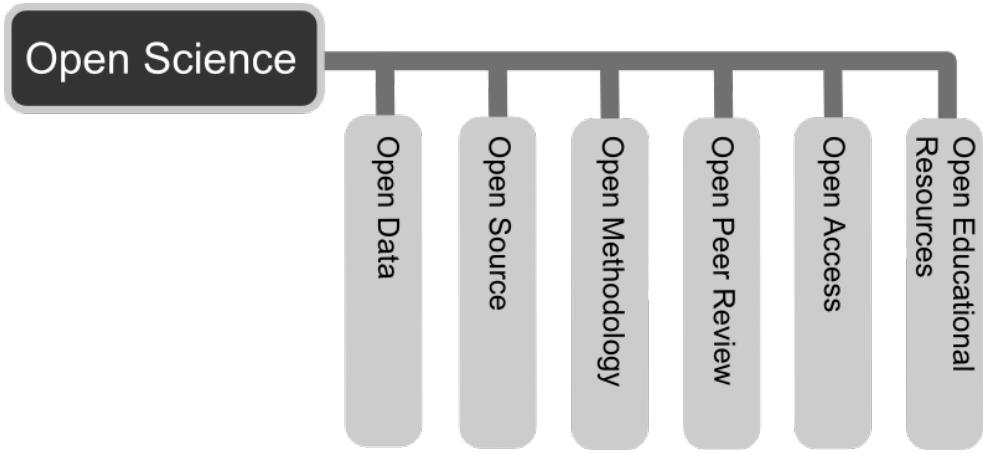
1) https://www.aaas.org/programs/scientific-responsibility-human-rights-law/aaas-statement-scientific-freedom
2) NASEM. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: NAP

FAIR Research?

The Open Science (Open Research) Approach



* These indicators are for both open access to publications and open scholarly communication.



https://upload.wikimedia.org/wikipedia/commons/9/9c/Open_Science_-_Prinzipien.png
Andreas E. Neuhold, CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, via Wikimedia Commons

http://aims.fao.org/activity/blog/open-science-monitor-access-data-and-trends-open-science

FAIR Research Data?
The State of Research Data Today

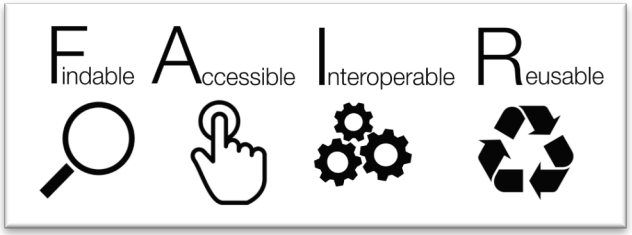
* Volume, Velocity, Variety, Veracity, Value
  * How do you find the specific data(set) that you need, easily, completely?
* Heterogeneous
  * What format is it in? Can I read the format? Am I allowed to read the file?
* Trusted?
  * Has the data been 'edited'? -> data integrity (Blockchain?)
  * Where does come from? -> provenance
* Incomplete
  * Only the 'best' data available
  * Not described with enough context to be used with other data
  * Reported without enough specificity
* Paywalled data, Sensitive data, Legacy data, Locked data (in format)

General Terms and Acronyms

* Metadata – data about data
* PID – Persistent Identifier (e.g., ORCID)
* GUPRI – Globally Unique Persistent Resolvable Identifier
* XML – Extensible Markup Language
* JSON – JavaScript Object Notation
* RDF – Resource Description Framework (semantic data)
  * Encoded in a variety of formats – XML, JSON-LD, TTL (Turtle)

What is FAIR?

* Findable, Accessible, Interoperable and Reusable
  * *Findable* – data is well identified so that it can be found in searches
  * *Accessible* – information about where it is and who can access it
  * *Interoperable* – if the data is available it is in a format where the meaning of the data can be understood
  * *Reusable* – has a license that tells the user what they can and cannot do with that data
* FAIR Data ≠ Open Data



The Evolution of FAIR

* Developed in 2014 as an initiative to support scientific data (1)
* Expanded as a set of Guiding Principles in 2016 (2, 3)
* Now a major initiative in Europe coordinated under the GO-FAIR Office in Leiden, NL (4)
* Activities now in Europe (EOSC), US (NIH, NSF), UNESCO, CODATA, US NAS, CESSDA, etc.

1) Data FAIRport: Find Access, Interoperate, and Re-use Data https://www.datafairport.org/
2) FORCE11: The FAIR Data Principles https://www.force11.org/group/fairgroup/fairprinciples
3) The FAIR Guiding Principles for scientific data management and stewardship, Mons et. al. https://doi.org/10.1038/sdata.2016.18
4) The Global Open (GO) FAIR Initiative https://www.go-fair.org/

The Components of FAIR: *F*

* ***Findability***
  * A unique identifier to make it easy to find and reference
  * Metadata to describe the data so that it can be found
  * Metadata are indexed in a searchable database
  * Example (non-research data) unique identifiers
    * ORCID – for researchers (https://orcid.org) (Stuart J. Chalk)
    * Digital Object Identifiers – for publications (https://doi.org) (paper)
    * ISNI – for people, organizations and more (https://isni.org) (UNF)

The Components of FAIR: *A*

* ***Accessible***
  * Metadata for a resource can be found via a standard method, e.g. Internet protocols like http/https, ftp, ssh
    * The protocol must be open, free, and available on any device
    * The protocol allows for **authentication/authorization** if needed
  * Metadata are accessible even if the data is not
  * FAIR ≠ Open – many companies are making their data FAIR but only for internal use

The Components of FAIR: *I*

* ***Interoperable***
  * (Meta)data use a formal, shared, applicable knowledge representation (i.e. format to describe meaning)
  * (Meta)data use vocabularies, taxonomies, ontologies that are FAIR (i.e. provide the meaning/semantics)
  * (Meta)data use qualified references to other data e.g. give meaning to the relationship(s) to other data
  * Use common available open data formats: XML, JSON, RDF

The Components of FAIR: *R*

* ***Reusable***
  * Have a clear documented using license (e.g. CC BY-NC)
    * Provide attribution when you reuse
    * Adapt, remix, transform or build on as you like
    * For non-commercial use only
  * Include detailed provenance about the data
  * Describe the data using domain-relevant community standards, e.g. Dublin Core for resource metadata

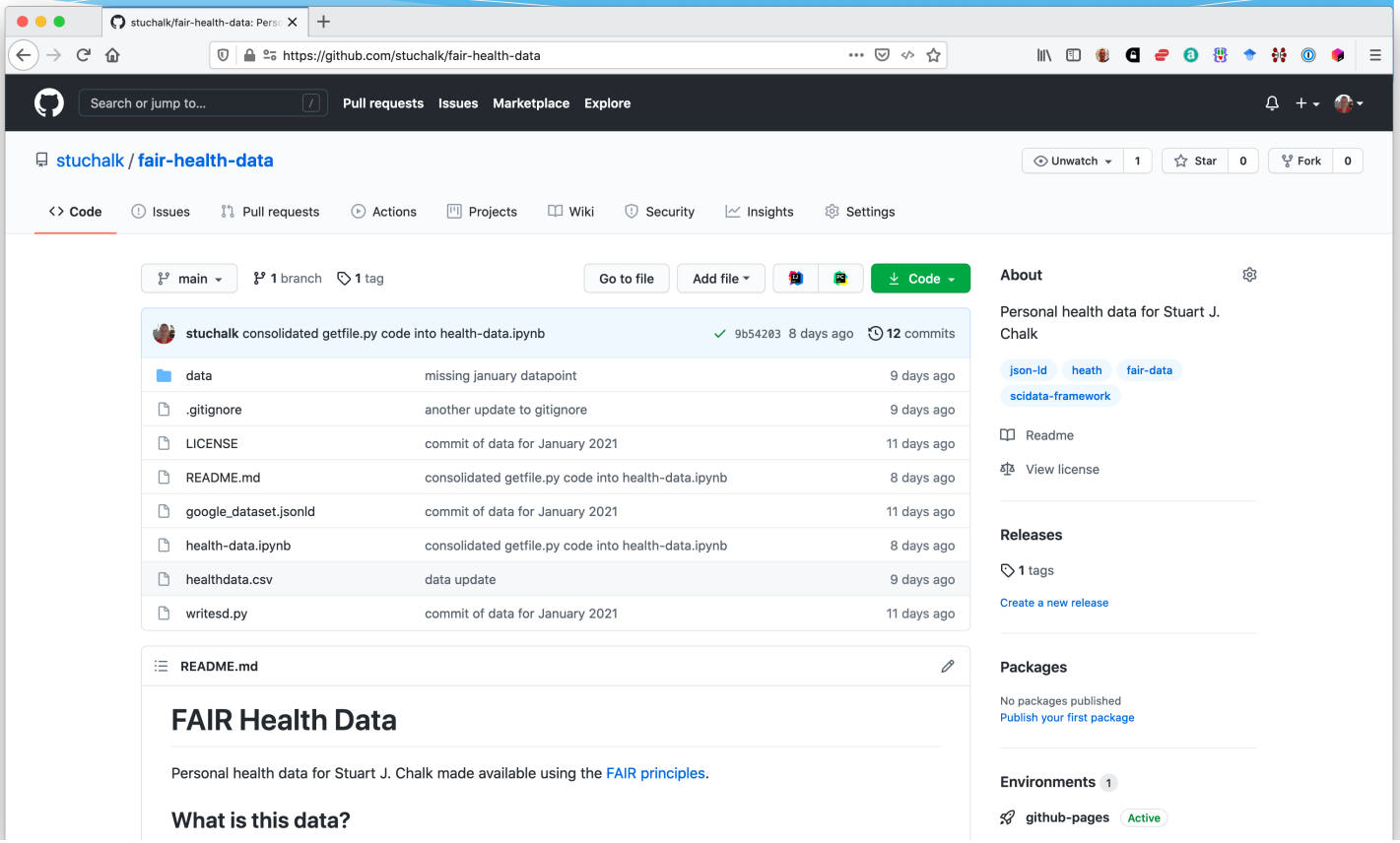
Example - A FAIR ~~Chemistry~~ Chemists Dataset

* Personal Health Data
  * body weight
  * body temperature
  * resting heart rate
  * data from daily activity (running)
    * VO2 max, calories, distance, pace, link to activity

https://github.com/stuchalk/fair-health-data

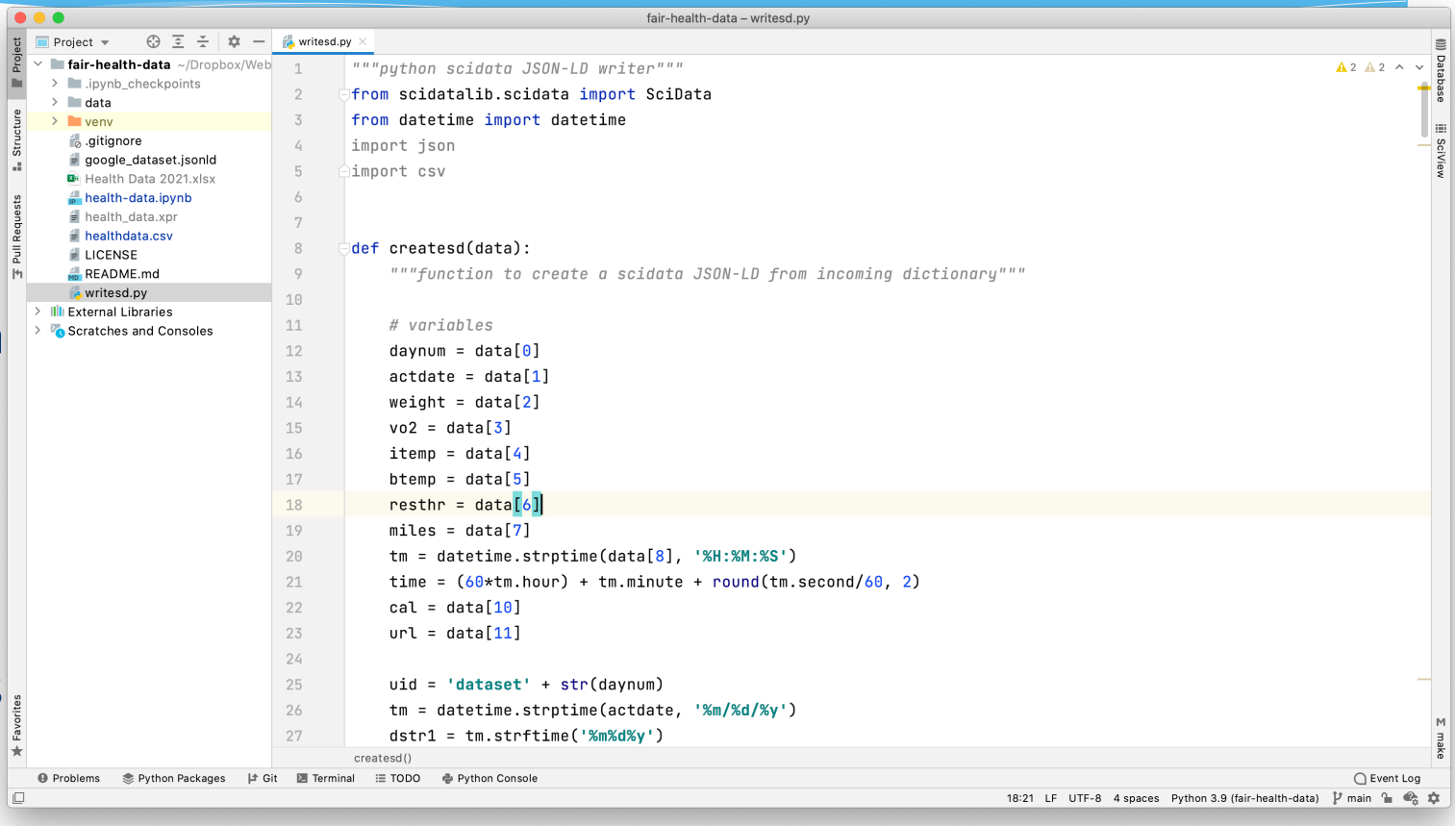
Example - A FAIR ~~Chemistry~~ Chemists Dataset

* GitHub
  * ‘Repo’ Hosting
  * Code/data
  * Open source
  * Based on ‘git’ version control system (VCS)
  * Benefits for academic users



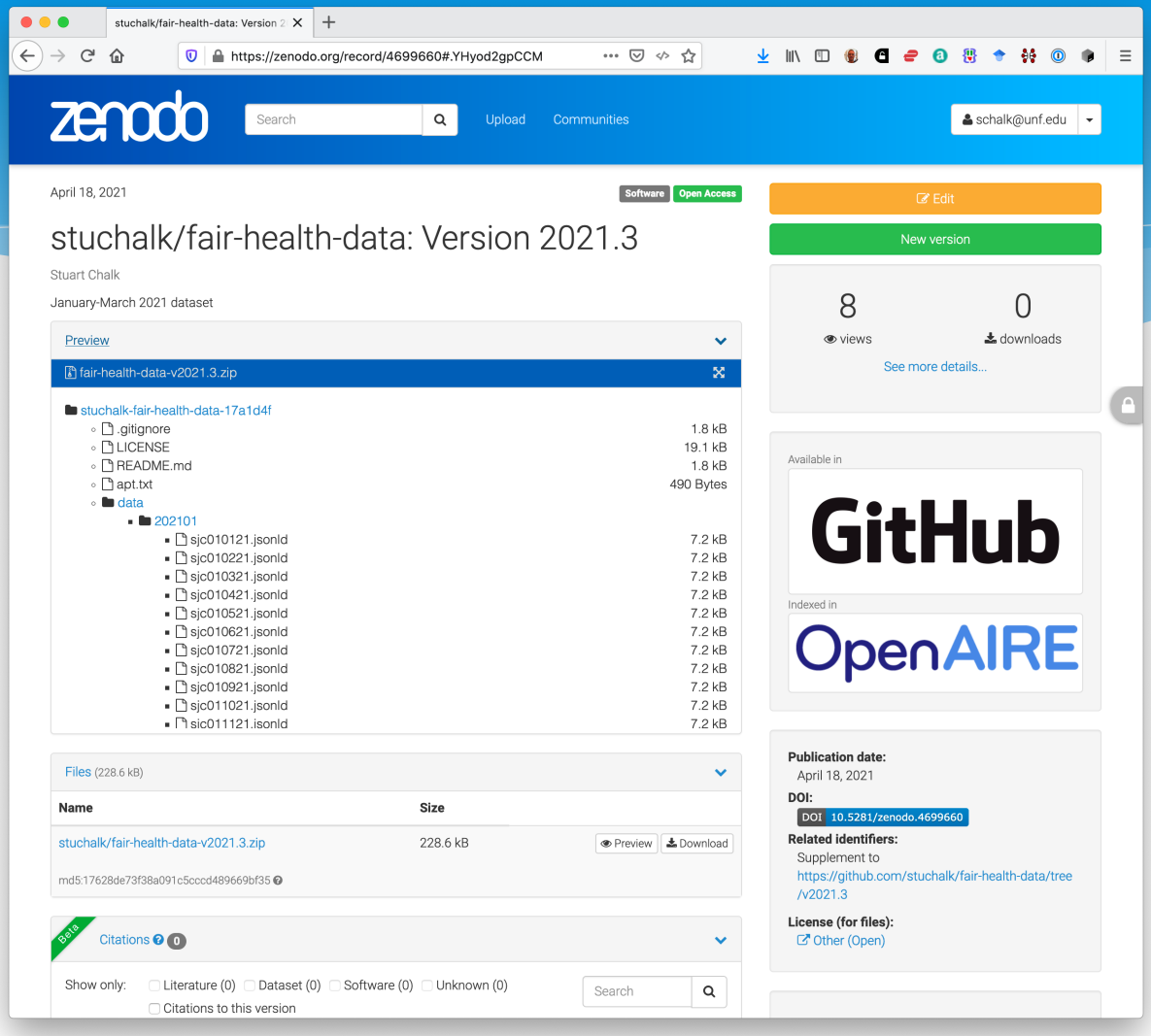
Example - A FAIR ~~Chemistry~~ Chemists Dataset

* PyCharm IDE
  * Python Code development
  * Integrates with GitHub
  * Educational Edition
  * Free for academic users



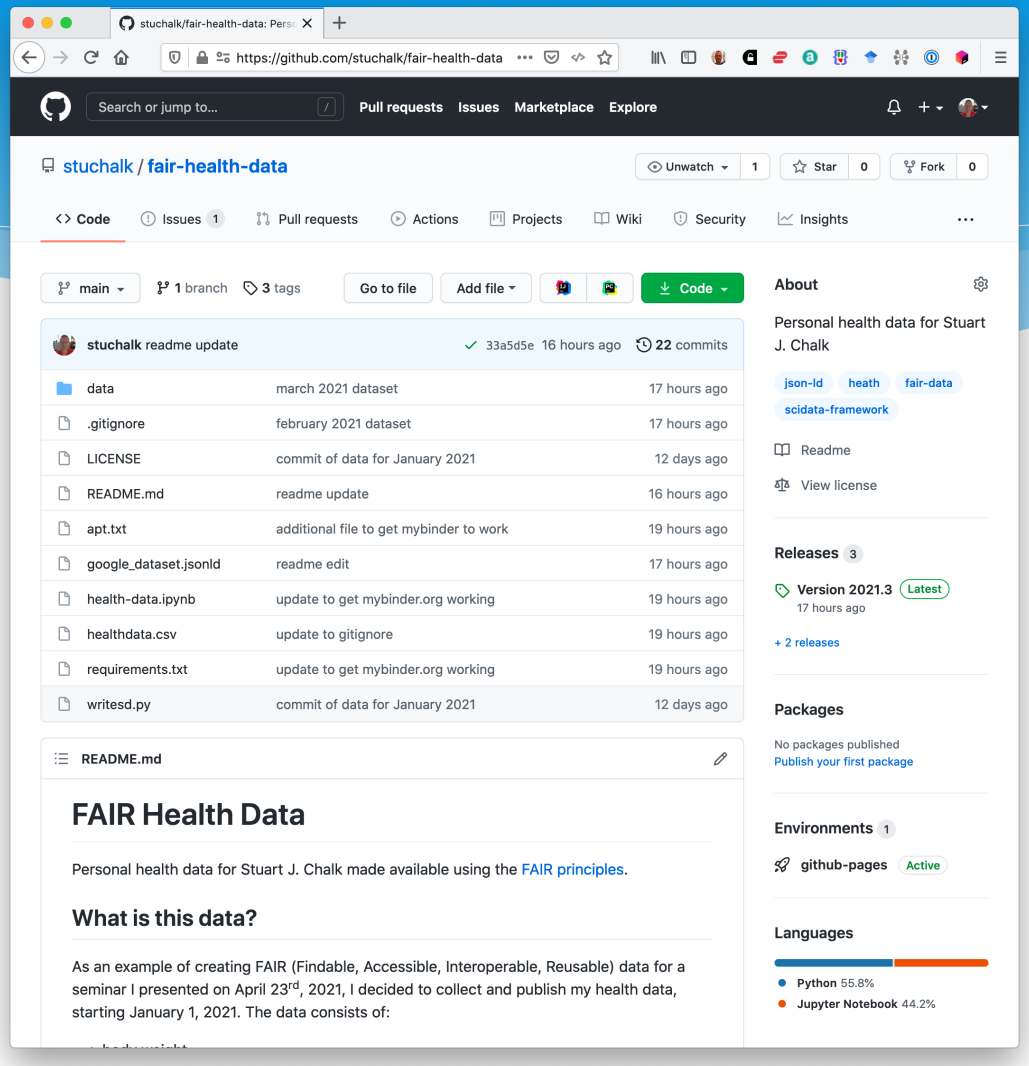
Findable Dataset

* Digital Object Identifiers (DOIs) assigned to each version in the GitHub repository via Zenodo



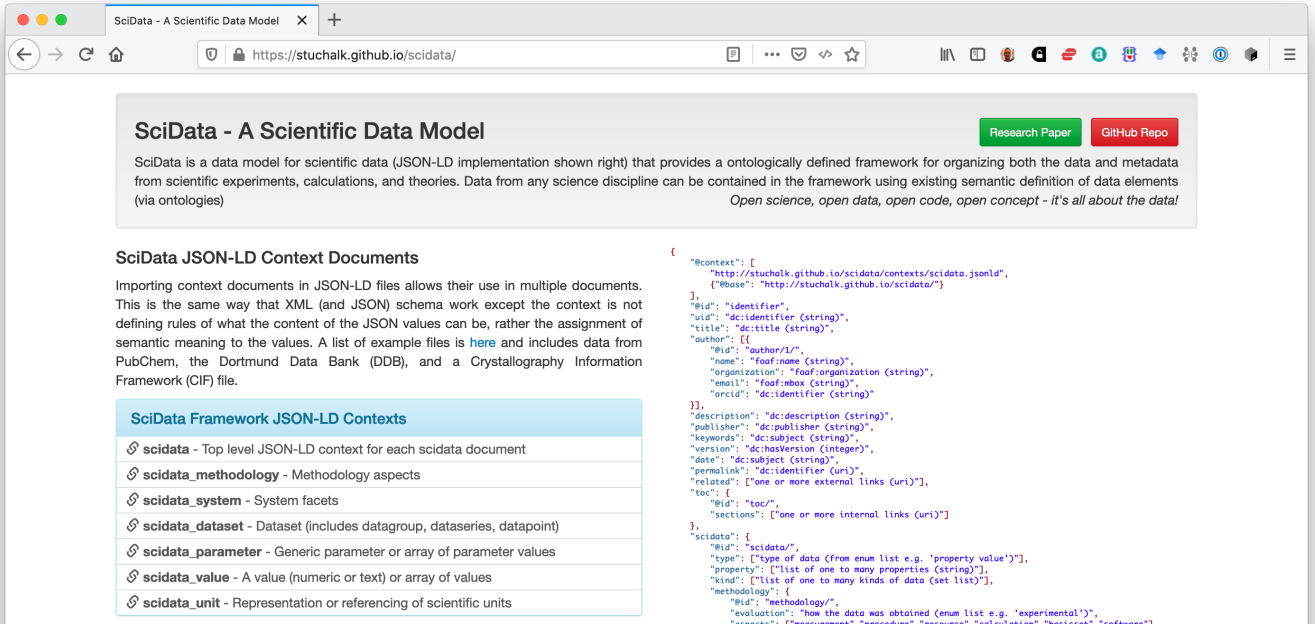
Accessible Dataset

* Data access via GitHub
* Users can report issues
* Collaborators can join
* Discussions
* The ‘ Repo’ can be ‘Forked’ (anyone can take a copy to work on)



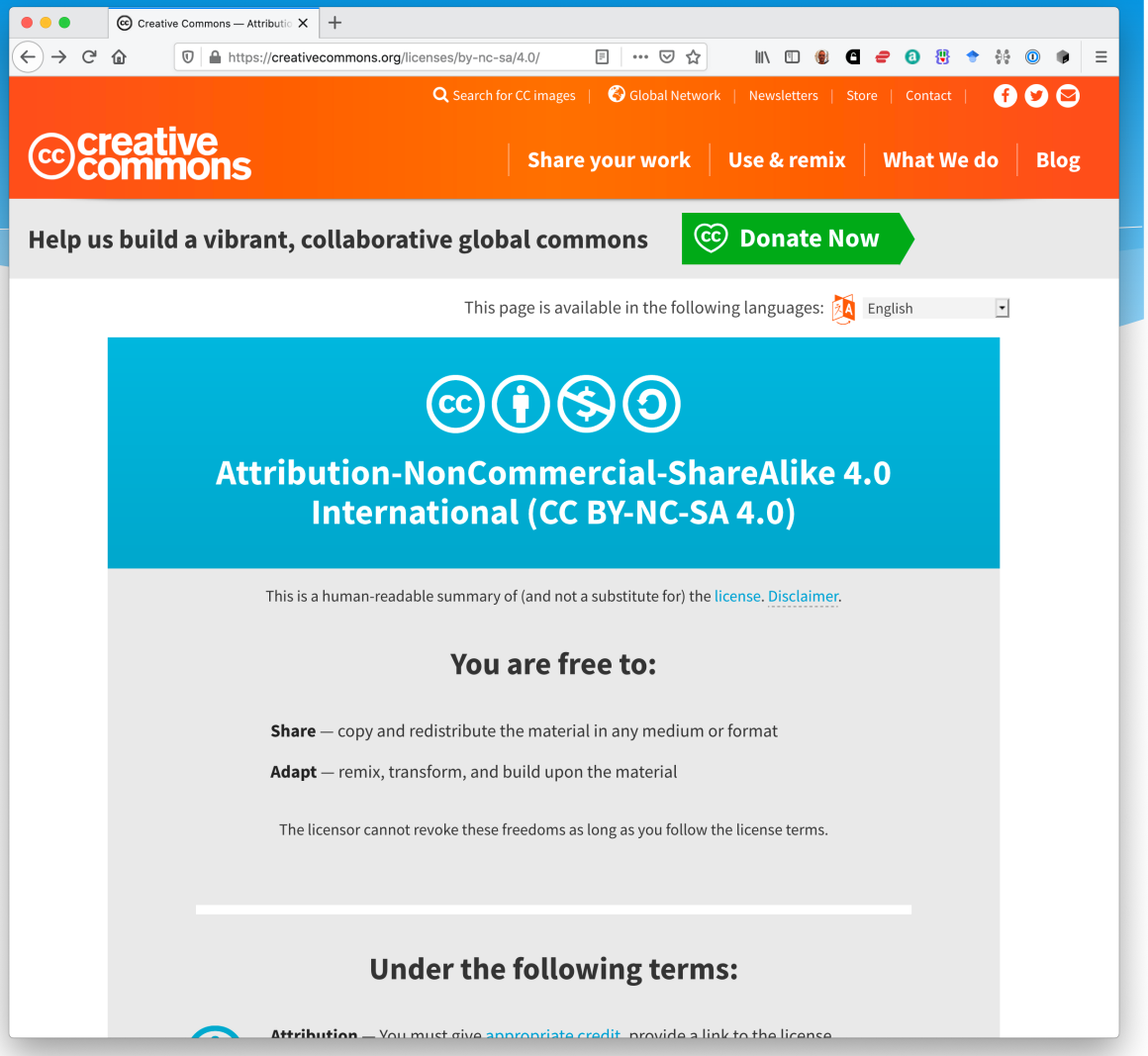
Interoperable Dataset

* JSON-LD - JSON for Linked-Data
* Semantic annotation
* SciData website, paper & ontology



Reusable Dataset

* Tell users what they can do with the data (and any restrictions)
* Creative Commons Open License
* BY – by attribution
* NC – non-commercial
* SA – share alike



Future Thoughts

* In 20 years the way we do research will be different
  * Data first
  * “Born digital” data
  * H-index replaced by “D-Index”? (Tenure requirements?)
  * Web notebooks for data collection, integration, analysis and dissemination
  * Global data services

https://www.rd-alliance.org/trust-principles-rda-community-effort

Questions?

* schalk@unf.edu
* Phone: 904-620-1938
* Skype: stuartchalk
* LinkedIn: https://www.linkedin.com/in/stuchalk
* ORCID: http://orcid.org/0000-0002-0703-7776

FAIR Data Examples (Chemistry)

* Caffeine @ Wikidata – for humans, for machines
* Theobromine @ PubChem – for humans, for machines
* Activities (9822) of 6900 compounds against SARS-COV2 @ ChEMBL – for humans, for machines
* Sulfadiazine @ CCDC – for humans, for machines
* Cyclohexanes @ ZINC – for humans, for machines
* DDT @ USEPA – for humans