

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*, Vi, 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 data¹, and new organizations launched specifically to pioneer studies on and recommendations for this new discipline². In addition, the development of the findability, accessibility, interoperability, and reusability (FAIR) principles³⁻⁵ give a tangible perspective on open data and provide a way to describe what open data means, and metrics to describe how open it is⁶.

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

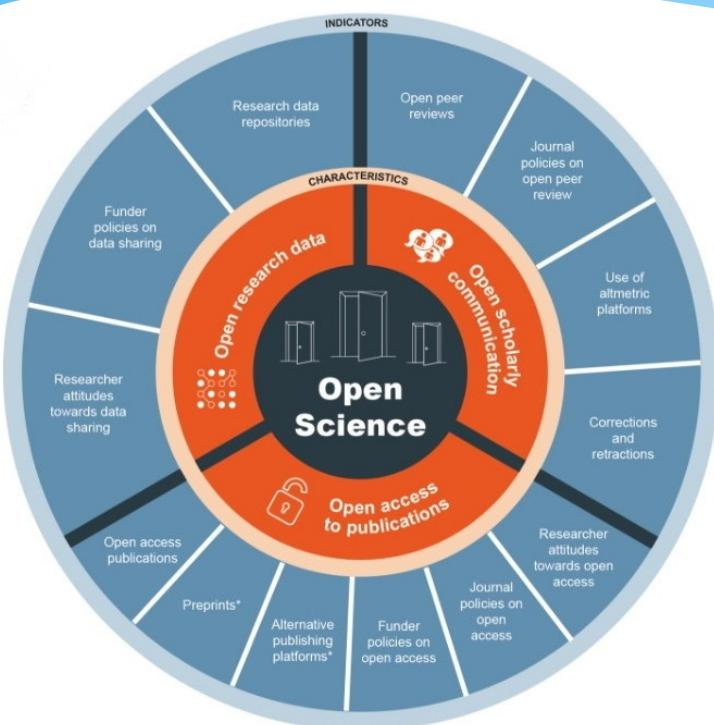
'Journal papers should be supplements to data' – Barend Mons (GO FAIR)

- * 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)

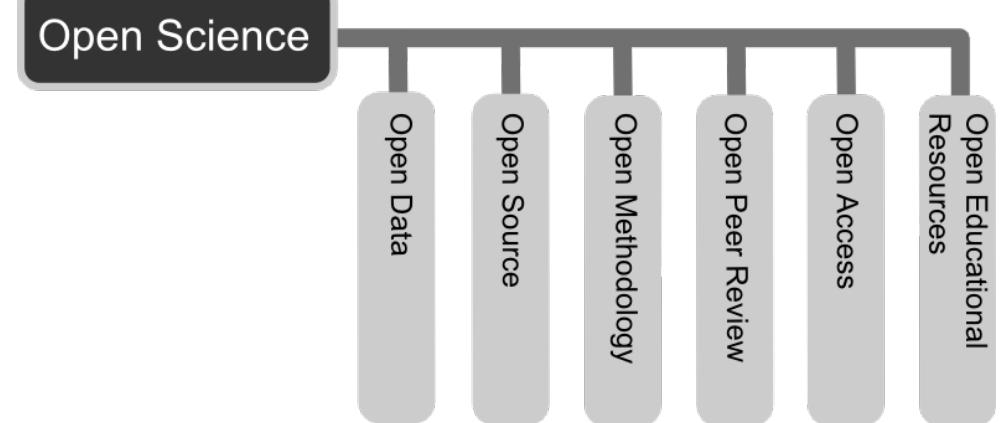
1) <https://www.aaas.org/programs/scientific-responsibility-human-rights-law/aaas-statement-scientific-freedom>
2) NASEM. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: NAP

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/g/gc/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

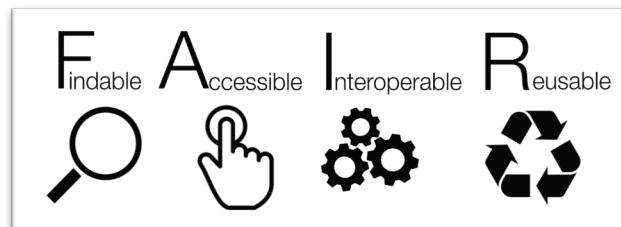
- * Volume, Velocity, Variety, Veracity, Value
 - * How do you find the specific data(set) that you need, easily, completely?
- * Heterogeneous
 - * What format is it in? Can I read the format? Am I allowed to read the file?
- * Trusted?
 - * Has the data been ‘edited’? -> data integrity (Blockchain?)
 - * 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](#))
- * GUPI – 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 Dataset

Chemists

- * 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 Dataset

Chemists

* GitHub

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

The screenshot shows a GitHub repository page for 'stuchalk / fair-health-data'. The repository has 1 branch and 1 tag. The code tab is selected, showing a list of commits:

Commit	Message	Date	Commits
stuchalk consolidated getfile.py code into health-data.ipynb	missing january datapoint	9 days ago	12 commits
.data	another update to gitignore	9 days ago	
.gitignore	commit of data for January 2021	11 days ago	
LICENSE	consolidated getfile.py code into health-data.ipynb	8 days ago	
README.md	commit of data for January 2021	11 days ago	
google_dataset.jsonld	consolidated getfile.py code into health-data.ipynb	8 days ago	
health-data.ipynb	data update	9 days ago	
healthdata.csv	commit of data for January 2021	11 days ago	
writesd.py			

The README.md file contains the following text:

FAIR Health Data

Personal health data for Stuart J. Chalk made available using the [FAIR principles](#).

What is this data?

On the right side of the page, there are sections for About, Releases, Packages, and Environments.

Example - A FAIR Chemistry Dataset

Chemists

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

```
"""python scidata JSON-LD writer"""
from scidatalib.scidata import SciData
from datetime import datetime
import json
import csv

def createsd(data):
    """function to create a scidata JSON-LD from incoming dictionary"""

    # variables
    daynum = data[0]
    actdate = data[1]
    weight = data[2]
    vo2 = data[3]
    itemp = data[4]
    btemp = data[5]
    resthr = data[6]
    miles = data[7]
    tm = datetime.strptime(data[8], '%H:%M:%S')
    time = (60*tm.hour) + tm.minute + round(tm.second/60, 2)
    cal = data[10]
    url = data[11]

    uid = 'dataset' + str(daynum)
    tm = datetime.strptime(actdate, '%m/%d/%y')
    dstr1 = tm.strftime('%m%d%y')

    createsd()
```

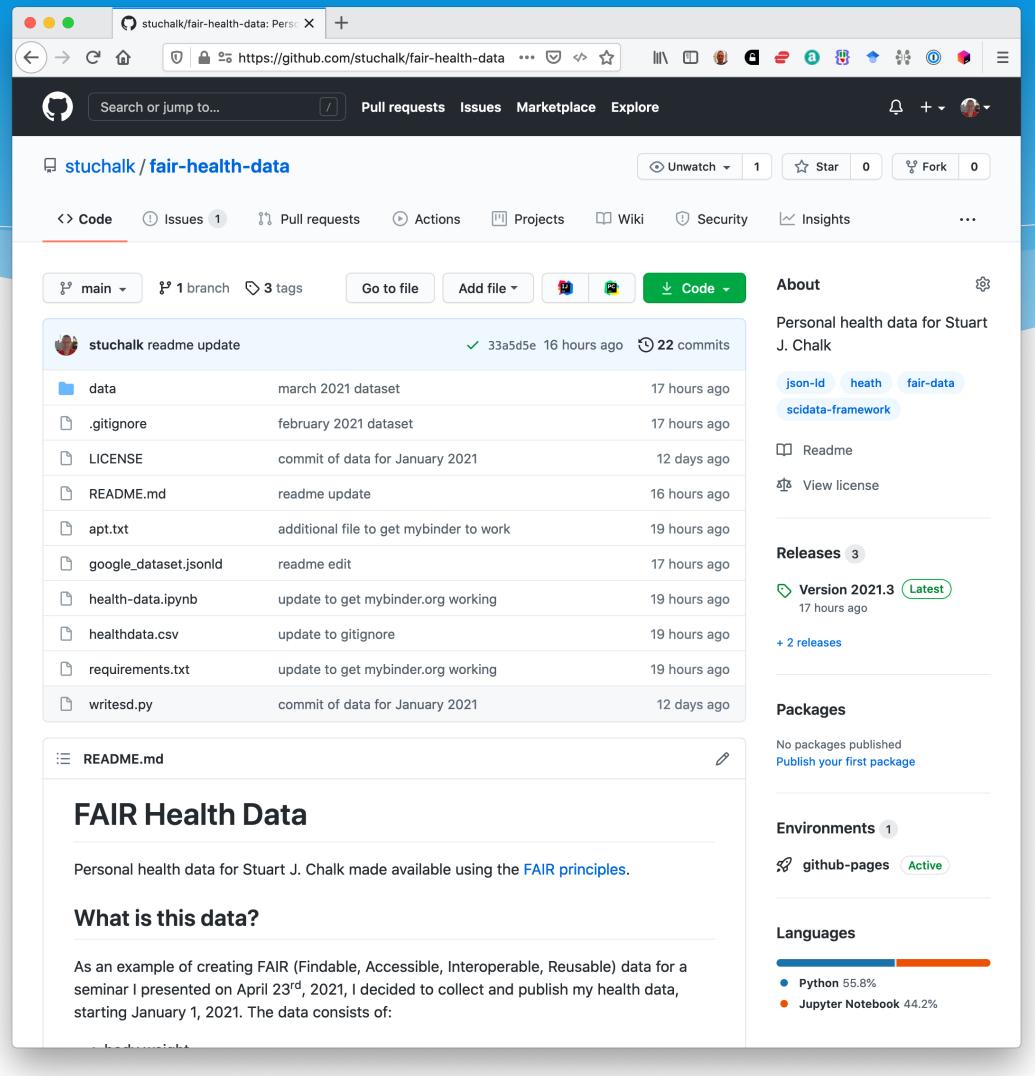
Findable Dataset

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

The screenshot shows a web browser window displaying a Zenodo dataset page. The URL in the address bar is <https://zenodo.org/record/4699660#.YHyod2gpCCM>. The page title is "stuchalk/fair-health-data: Version 2021.3". The dataset was created by "Stuart Chalk" on "January-March 2021 dataset". The file "fair-health-data-v2021.3.zip" is listed, containing several files including ".gitignore", "LICENSE", "README.md", "apt.txt", and a "data" folder. The "data" folder contains a "202101" folder with numerous JSON files. The total size of the files is 228.6 kB. The page also shows statistics: 8 views and 0 downloads. It is available in GitHub and OpenAIRE. The publication date is April 18, 2021, and the DOI is [10.5281/zenodo.4699660](https://doi.org/10.5281/zenodo.4699660). Related identifiers include a GitHub link to the repository. The license is marked as "Other (Open)".

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

The screenshot shows two browser windows side-by-side.

Top Window: W3C Recommendation - JSON-LD 1.1

- Table of Contents:**
 - 1. Introduction
 - 1.1 How to Read this Document
 - 1.2 Contributing
 - 1.3 Typographical conventions
 - 1.4 Terminology
 - 1.5 Design Goals and Rationale
 - 1.6 Data Model Overview
 - 1.7 Syntax Tokens and Keywords
- 2. Conformance
- 3. Basic Concepts
- 3.1 The Context
- 3.2 IIRIs

Bottom Window: SciData - A Scientific Data Model

- SciData - A Scientific Data Model**
- SciData Context Documents**

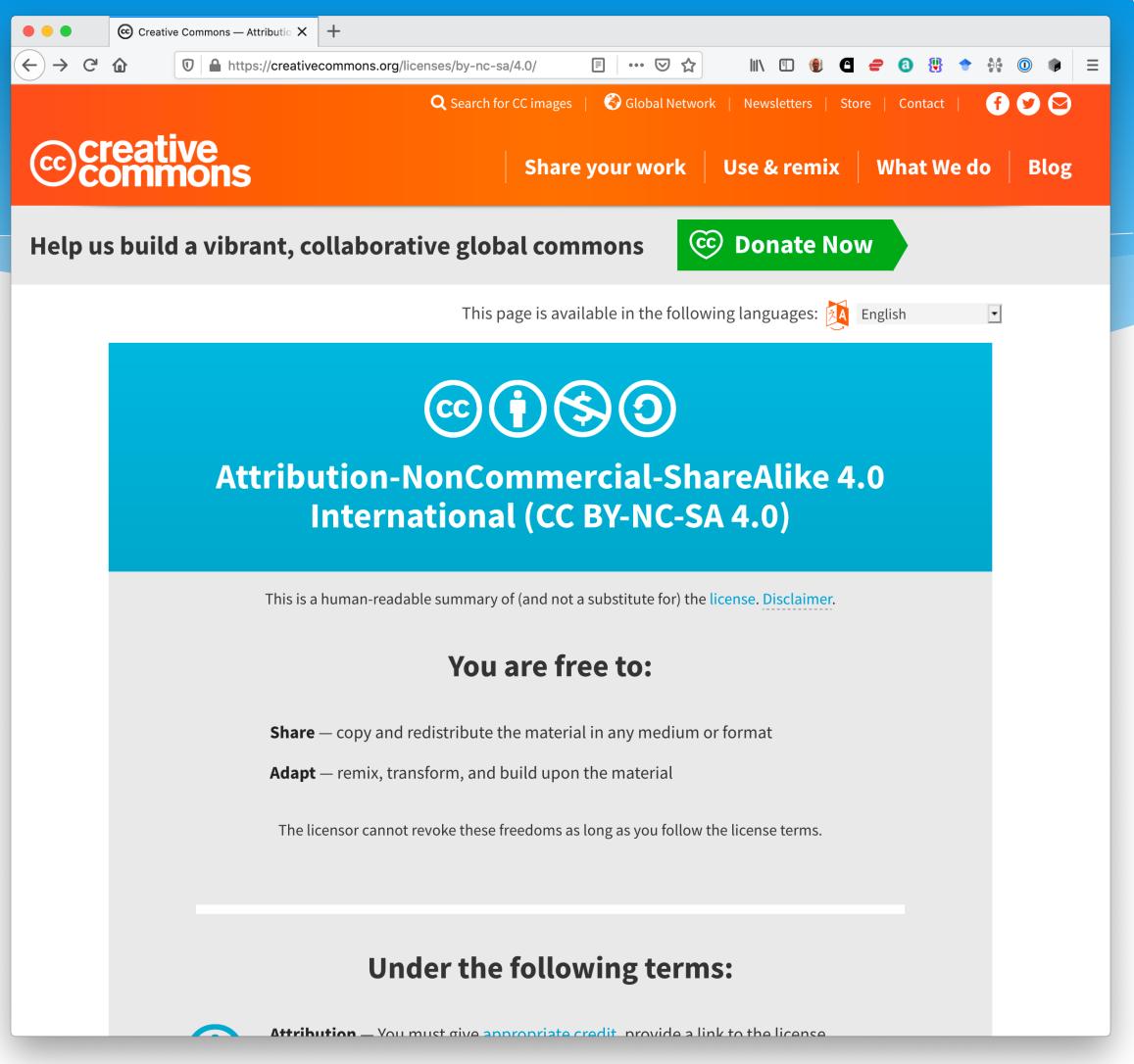
SciData is a data model for scientific data (JSON-LD implementation shown right) that provides an ontologically defined framework for organizing both the data and metadata from scientific experiments, calculations, and theories. Data from any science discipline can be contained in the framework using existing semantic definition of data elements (via ontologies)

Open science, open data, open code, open concept - it's all about the data!
- SciData Framework JSON-LD Contexts**
 - scidata - Top level JSON-LD context for each scidata document
 - scidata_methodology - Methodology aspects
 - scidata_system - System facets
 - scidata_dataset - Dataset (includes datagroup, dataseries, datapoint)
 - scidata_parameter - Generic parameter or array of parameter values
 - scidata_value - A value (numeric or text) or array of values
 - scidata_unit - Representation or referencing of scientific units
- JSON-LD Context (Partial View):**

```
{
  "@context": [
    "http://stuchalk.github.io/scidata/context/scidata.jsonld",
    {"@base": "http://stuchalk.github.io/scidata/"}
  ],
  "id": "identifier",
  "uid": "dc:identifier (string)",
  "title": "dc:title (string)",
  "author": "foaf:author (string)",
  "name": "foaf:name (string)",
  "organization": "foaf:organization (string)",
  "email": "foaf:email (string)",
  "orcid": "dc:identifier (string)",
  "description": "dc:description (string)",
  "publisher": "dc:publisher (string)",
  "keywords": "dc:subject (string)",
  "version": "dc:version (integer)",
  "date": "dc:date (string)",
  "permalink": "dc:identifier (uri)",
  "related": ["one or more external links (uri)"],
  "toc": "#id: toc",
  "sections": ["one or more internal links (uri)"]
},
"scidata": {
  "id": "scidata",
  "type": "type of data (from enum list e.g. 'property value')",
  "property": ["list of one to many properties (string)"],
  "kind": ["list of one to many kinds of data (set list)"],
  "meta": [
    {
      "id": "methodology",
      "evaluation": "how the data was obtained (enum list e.g. 'experimental')",
      "annotation": "annotation"
    }
  ]
}
```

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](#)