

DSCI-560

Lecture 4: Ideals of Data Science Data Science Professional Practicum

Young Cho

Department of Electrical Engineering
University of Southern California

Part 1

- Motivation and overview: open science, reproducible publications, digital scholarship
- Making data accessible

Part 2

- Making software accessible
- Documenting methods and workflows

**Making science more
reproducible**

Traditional reports

- Reports can be in the form of:
 - Technical brief
 - Manager brief
 - Scientific article
 - Long-form

Traditional Reports

Text:

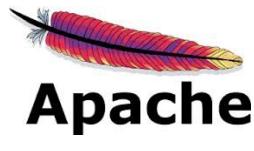
Narrative of method,
the data is in tables, figures/plots,
the software used is mentioned

Open Science

Open data



Open source



Open access



Open publications



Not only in academia...

Survey across 7 Big Data Companies

Importance of ML provenance tracking	Participants agreement
Tracking provenance between data and ML models can be very useful in multiple scenarios	82%
Model debugging	88%
Model sharing	69%
Compliance	56%
Fairness	56%

VASMA, KDD2020

Scientific Methodology

U.S. HOUSE OF REPRESENTATIVES

U.S. HOUSE OF REPRESENTATIVES
COMMITTEE REPOSITORY



COMMITTEE ON
**SCIENCE, SPACE, &
TECHNOLOGY**
Lamar Smith, Chairman

Hearing: Climate Science: Assumptions, Policy Implications, and the Scientific Method

Wednesday, March 29, 2017 (10:00 AM)

Today we will examine the scientific method as it relates to climate change. We must ensure that the underlying science that informs policy decisions is based on credible scientific methodology.

I believe the climate is changing and that humans play a role. However, I also believe significant questions remain as to the extent.

Our actions must be based on sound science. This is the only way we will be able to better address climate change.

Before we impose costly government regulations, we should evaluate scientific uncertainties and ascertain the extent to which they make it difficult to quantify humans' contribution to climate change.

Far too often, alarmist theories on climate science originate with scientists who operate outside of the principles of the scientific method.

The scientific method is a simple process that has been used for centuries. It involves identifying a question, developing a hypothesis, constructing an experiment, and analyzing the results. If the results do not align with the original hypothesis, the hypothesis must be re-examined.

The scientific method welcomes critiques so theories can be refined. And it avoids speculation about distant events for which there is no hard proof.

The scientific method is regarded as the "foundation of modern science." It ensures that scientific experimentation is neither arbitrary nor subjective, and that results can be replicated.

The scientific method also requires that for a hypothesis to become a theory, a repeated validation of the results – called **reproducibility** – should be possible. However, a recent survey found that 70% of scientific researchers have tried and failed to reproduce the experiments conducted by other scientists.

The lack of reproducibility is a warning that the scientific method is not being followed and that the theory may lack **credibility**. To restore faith in science, we must uphold the principles of scientific integrity.

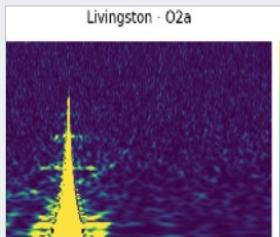
Only when scientists follow the scientific method can policy-makers be confident that they are making the right decisions.

Public Interest



WHAT WILL YOU DISCOVER?

Participate in research of all kinds, from classifying galaxies to counting penguins to transcribing manuscripts. Whatever your interest, there's a Zooniverse project for you.



GRAVITY SPY

Help scientists at LIGO search for gravitational waves, the elusive ripples of spacetime.

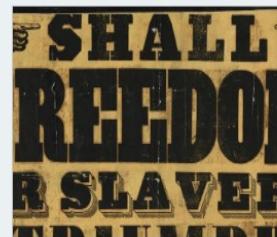
[View Project](#)



BASH THE BUG

Help us fight resistance to antibiotics!

[View Project](#)



ANTI-SLAVERY MANUSCRIPTS

We need your help to turn our collection of handwritten correspondence between anti-...

[View Project](#)



PENGUIN WATCH

Count penguins in remote regions to help us understand their lives and environment.

[View Project](#)

[See All Projects](#)



eBird

oldWeather

Modern Reports

Traditional Published Reports

Text:

Narrative of method,
the data is in tables, figures/plots,
the software used is mentioned



Modern Published Reports

Text:

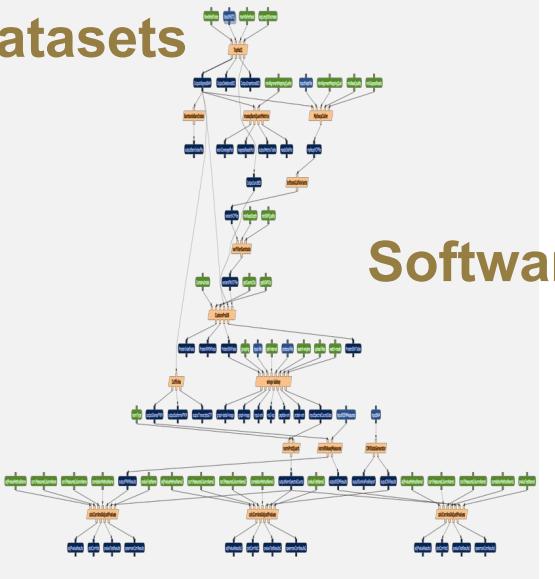
Narrative of method,
the data is in tables, figures/plots,
the software used is mentioned

Data:

Supplementary materials,
pointers to data repositories

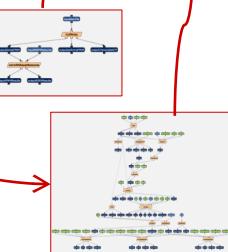
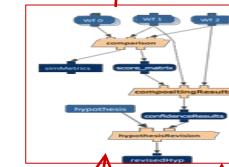
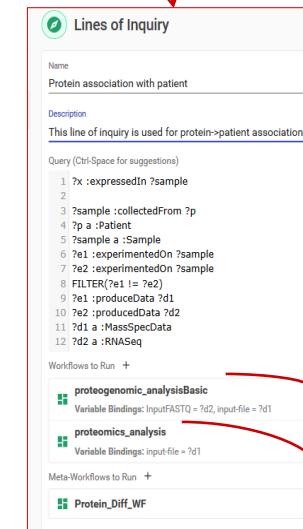
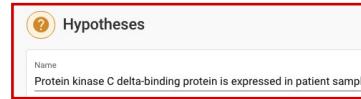
Reproducible Research

Datasets



Software

Workflow



Experimental Design

Open Science

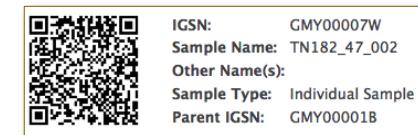
Shared
repositories



Persistent
unique
identifiers



DOI



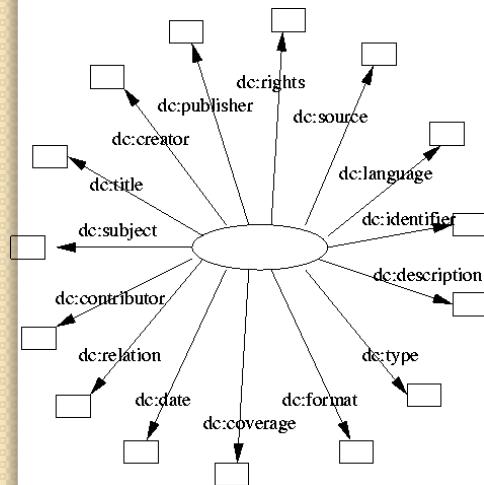
Licenses



Digital scholarship

Data and software citation

Metadata



Electronic Notebooks

- ▶ Record data, software, results, notes, etc.
 - ▶ Records what code was run when generating a result
 - ▶ Can re-run code with new data



<https://jupyter.org>

https://github.com/LinkedEarth/Pyleoclim_util/blob/master/Example/PyleoclimQuickstart.ipynb

Plotting

plotTs()

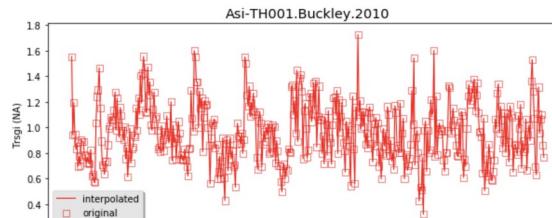
Plot a time series.

Syntax: `fig = pyleoclim.plotTs(timeseries = "", x_axis = "", markersize = 50, marker = "default", figsize = [10,4], saveFig = False, dir = "figures", format="eps")`

Optional arguments:

- `timeseries`: A timeseries object as defined in LiPD. If left blank, you'll be prompted to choose the record from a list.
- `x-axis`: The representation against which to plot the paleo-data. Options are "age", "year", and "depth". Default is to let the system choose if only one available or prompt the user.
- `markersize`: Default is 50
- `marker`: Shape and color. Default uses the Pyleoclim color palette. If you wish to change the default marker, enter the color and shape ([in this order](#)). For instance to use a red square, use '`rs`'.
- `figsize`: The size of the figure
- `saveFig`: If `True`, saves the plot into the `dir` folder in the current working directory.
- `dir`: a folder in the current directory, where the various figures can be saved. If left blank, a folder named `figures` will be automatically created
- `format`: One of the file extensions supported by the active backend. Default is `eps`. Most backends support `png`, `pdf`, `ps`, and `svg`.

In [8]: `%matplotlib inline`
`fig = pyleo.plotTs(timeseries = ts_list[12], marker = 'rs')`



Electronic Notebooks

- ▶ Record data, software, results, notes, etc.
 - ▶ Records what code was run when generating a result
 - ▶ Can re-run code with new data

R Markdown

from  R Studio

<https://rmarkdown.rstudio.com>

```
1. Isabel decides to search for the Basin Pond data geographically, and creates a box around the site.

```{r}
locpoly <- matrix(c(-70,-71,-71,-70,45,45,44,44),ncol=2,byrow = F)
```

2. She then converts that array to GeoJSON FeatureCollection, so she can query GeoDex

```{r}
#install.packages("geojson")
library(geojson)

poly <- geojson::polygon(paste0('{"type": "Polygon", "coordinates": [[[' , "[" , locpoly[1,1]," , ",locpoly[1,2],""], " [",locpoly

feature <- geojson::feature(poly)
feature_collection <- geojson::featurecollection(feature)
feature_collection_str <- geo_pretty(feature_collection)
```

1. Isabel decides to search for the Basin Pond data geographically, and creates a box around the site.
locpoly <- matrix(c(-70,-71,-71,-70,45,45,44,44),ncol=2,byrow = F)

2. She then converts that array to GeoJSON FeatureCollection, so she can query GeoDex
#install.packages("geojson")
library(geojson)

##
## Attaching package: 'geojson'
## The following object is masked from 'package:graphics':
##
##     polygon
poly <- geojson::polygon(paste0('{"type": "Polygon", "coordinates": [[[' , "[" , locpoly[1,1]," , ",locpoly

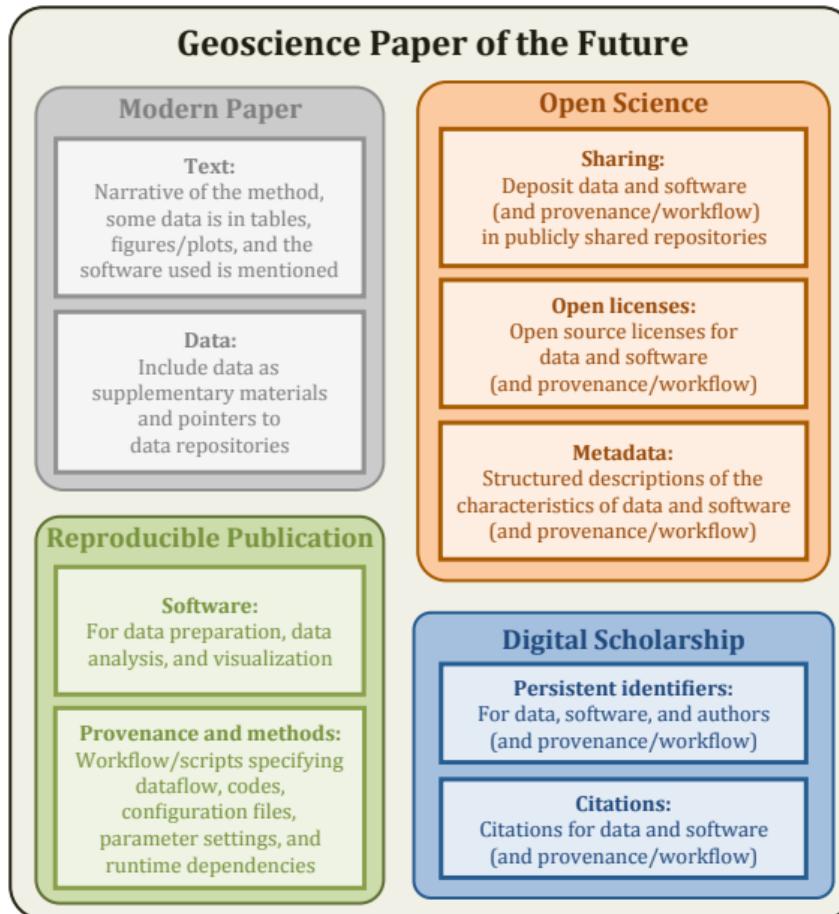
feature <- geojson::feature(poly)
feature_collection <- geojson::featurecollection(feature)
feature_collection_str <- geo_pretty(feature_collection)
```

Workflow Systems

- Many choices
 - Academic prototypes
 - Operational open source
 - Commercial
- Different capabilities
 - Workflow validation
 - Scalable computations



Scientific Paper of the Future



Gil, Y., & . (Ed .) .. (2016, April 17). The Geoscience Paper of the Future: OntoSoft Training (Version 9).
<https://doi.org/10.6084/m9.figshare.1586773.v9>

What is a Future scientific paper?

- **Data:** Available in a public repository, including documentation (metadata), a clear license specifying conditions of use, and citable using a unique and persistent identifier.
- **Software:** Available in a public repository, with documentation (metadata), a license for reuse, and citable using a unique persistent identifier.
 - Not only major software used, but also other ancillary software for data reformatting, data conversions, data filtering, and data visualization.
- **Provenance:** Documented for all results by explicitly describing the series of computations and their outcome with a provenance record of the execution traces and a workflow sketch (or formal workflow)
 - Possibly in a shared repository and with a unique and persistent identifier.

Problems with current practices

- Data is often not made available in publications
 - Limited reproducibility
- Data made available through investigator's URL
 - URL does not resolve (i.e., "rotten")

Nature Genetics 41, 149 - 155 (2009)
Published online: 28 January 2008 | doi:10.1038/ng.295

Repeatability of published microarray gene expression analyses

scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis.

PLOS ONE | DOI:10.1371/journal.pone.0115253 December 26, 2014

RESEARCH ARTICLE

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein^{1*}, Herbert Van de Sompel¹, Robert Sanderson¹, Harihar Shankar¹, Lyudmila Balakireva¹, Ke Zhou², Richard Tobin²

We analyze a vast collection of articles from three corpora that span publication years 1997 to 2012. For over one million references to web resources extracted from over 3.5 million articles, we observe that the fraction of articles containing references to web resources is growing steadily over time. We find one out of five STM articles suffering from reference rot, meaning it is impossible to revisit the web context that surrounds them some time after their publication. When only considering STM articles that contain references to web resources, this fraction increases to seven out of ten.

Including Data with the Paper

- Data paper

Ecological Research
July 2013, Volume 28, Issue 4, p 541
Date: 10 May 2013

Monitoring records of plant species in the Hakone region of Fuji-Hakone-Izu National Park, Japan, 2001–2010

Takeshi Osawa



Abstract

The monitoring of species occurrences is a crucial aspect of biodiversity conservation, and regional volunteerism can serve as a powerful tool in such endeavors. The Fuji-Hakone-Izu National Park in the Hakone region of Kanagawa Prefecture, Japan, boasts a volunteer association of approximately 100 members. These volunteers have monitored plant species occurrences from 2001 to the present along several hiking trails in the region. In this paper, I present the annual observation records of plant occurrences in Hakone from 2001 to 2010. This data set includes 1,071 species of plants from 151 families. Scientific names follow the Y List, and this data set includes several threatened plant species. Data files are formatted based on the Darwin Core and Darwin Core Archives, which are defined by the Biodiversity Information Standards (BIS) or Biodiversity Information Standards Taxonomic Databases Working Group (TDWG). Data files filled on required and some additional item on Darwin Core. The data set can download from the author's personal Web site as of July 2012. These data will soon be published for the Global Biodiversity Information Facility (GBIF) through GBIF Japan. All users can then access the data from the GBIF portal site.

- The complete data set for this abstract published in the Data Paper section of the journal is available in electronic format in Ecological Research Data Paper Archives at http://db.cger.nies.go.jp/JaLTER/ER_DataPapers/archives/2013/ERDP-2013-01.

- Data published in a repository

The US
Long Term Ecological Research
Network

LTER

NTL LTER "WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998" - Lathrop

LTER Identifier:

knb-lter-ntl.279.1

Abstract:

These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin – NTL-LTER program took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...

Owners/Creators:

Lathrop

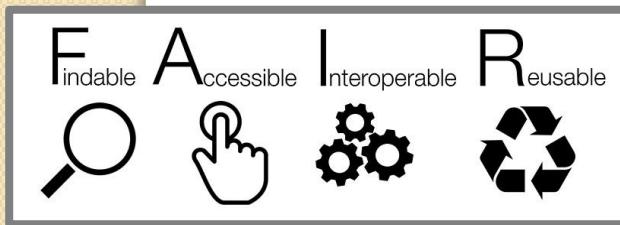
Metadata:

Select [here](#) for full metadata

Data File(s):

- [wdnr_fyke_minifyke_seine_lengths_weights.csv](#)
- [wdnr_boomshock_lengths_weights.csv](#)
- [wdnr_gillnet_lengths_weights_93.csv](#)
- [wdnr_walleye_age_lengths_weights_87.csv](#)
- [wdnr_creel_survey_lengths_weights.csv](#)
- [wdnr_creel_survey_angler_counts.csv](#)

The FAIR Principles



G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

"We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles."

To be 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.

TO BE 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.

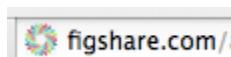
TO BE INTEROPERABLE:

- 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.

TO BE RE-USABLE:

- 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.

Making Data Accessible: Overview of Best Practices



Highly connected drug file

| | | | |
|--------------------|-----|----|---|
| Tretinoin | 257 | 46 | Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi
ici, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR |
| Levothyroxine | 173 | 36 | Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, b
25, cyp130, Rv1264, lppX, gpm1, ligA, nirA |
| Methotrexate | 156 | 32 | TB31.7, Rv1264, mscL, lppX, coxA, pcaA, Rv3676,
fabG1, |
| 4-Hydroxytamoxifen | 115 | 20 | mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pts
TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA |
| Estradiol | 98 | 20 | pth, ethR, clpP, glbN, inhA, |
| Amantadine | 79 | 1 | pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX |
| Rifampine | 78 | 13 | mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, R
5 |
| Raloxifene | 75 | 18 | mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c |
| Propofol | 54 | 5 | pepD, Rv1264, thyX, ethR, trx2B, |
| Indinavir | 51 | 14 | pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp121 |
| Penicillamine | 44 | 10 | |
| Daunorubicin | 44 | 12 | |
| Tricosan | 42 | 5 | |
| Darunavir | 40 | 15 | |

Enlarge to see the rest of the document

Enlarge Download

Cite this: Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.
<http://dx.doi.org/10.6084/m9.figshare.776887>
Retrieved 08:56, Feb 20, 2015 (GMT)

Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

Links

- <http://purl.org/net/tb-drugome-run>

License (what's this?)

CC-BY



1 Publication in a shared repository

2 General & domain metadata

3 Accessibility of Data (manual & machine)

4 Unique persistent identifier (PID)

5 Citation preference



<http://www.thestaffingstream.com/2012/08/06/the-buzz-about-talent-communities/>

Directories of Research Data Repositories

- <http://www.re3data.org>
- http://databib.org/index_subjects.php
- http://oad.simmons.edu/oadwiki/Data_repositories
- <http://www.force11.org>
- <http://www.nature.com/sdata/data-policies/repositories>

General

- Dataset name/title
- Description
- Creator(s)
- Publication date
- License
- Publisher/contact
- Version
- Resource type
- Location of the data

Domain Specific

- Categories
 - Keywords/tags
 - Related links
- ◆ A data repository in a given discipline may request metadata using accepted standards

Recommended Metadata

Choose a License: Creative Commons

Screenshot of the Creative Commons "Choose a License" interface. The interface is divided into several panels:

- License Features:** Options to allow adaptations (Yes or No) and commercial use (Yes or No). The "Yes" option for both is selected.
- Selected License:** Attribution 4.0 International. It includes icons for CC and BY, and a badge stating "This is a Free Culture License!" and "Creative Commons APPROVED FOR FREE CULTURE WORK".
- Help others attribute you!**: Fields for Title of work, Attribute work to name, Attribute work to URL, Source work URL, More permissions URL, Format of work (Other / Multiple formats), and License mark (HTML+RDFa). The "Other / Multiple formats" and "HTML+RDFa" options are selected.
- Have a web page?**: A summary stating "This work is licensed under a Creative Commons Attribution 4.0 International License." and provides a code snippet for linking to the license. It also offers options for "Normal Icon" or "Compact Icon".

Recommended: CC-BY and CC0



Attribution
CC BY

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

CC0 (datasets) "No rights reserved"



CC0 can be particularly important for the sharing of data and databases, since it otherwise may be unclear whether highly factual data and databases are restricted by copyright or other rights.

Databases may contain facts that, in and of themselves, are not protected by copyright law.

<http://creativecommons.org/licenses/>

Manual Accessibility

UNIQUE ID & METADATA

- http://figshare.com/articles/Highly_connected_drug_file/77683

The screenshot shows a Figshare article page. The title is 'Highly connected drug file'. Below the title is a table of highly connected drugs. The table has three columns: Drug Name, Degree, and Node ID. The first few rows are:

| Drug Name | Degree | Node ID |
|--------------------|--------|---------|
| Tretinoin | 257 | 46 |
| Levothyroxine | 173 | 36 |
| Methotrexate | 156 | 32 |
| 4-Hydroxytamoxifen | 115 | 25 |
| Estradiol | 98 | 20 |
| Amantadine | 79 | 1 |
| Rifampin | 78 | 13 |
| Propofol | 54 | 5 |
| Indinavir | 53 | 14 |
| Penicillamine | 44 | 12 |
| Darunavir | 40 | 15 |

Below the table, there is a note: "The embedded functionality can only be used for non-commercial purposes. more". There is also a "Description" section and a "Links" section at the bottom.

DATA

- <http://files.figshare.com/1175525/highConnectedDrugs.txt>

The screenshot shows a text file with the same data as the Figshare page. The table is identical to the one above, listing drug names, degrees, and node IDs.

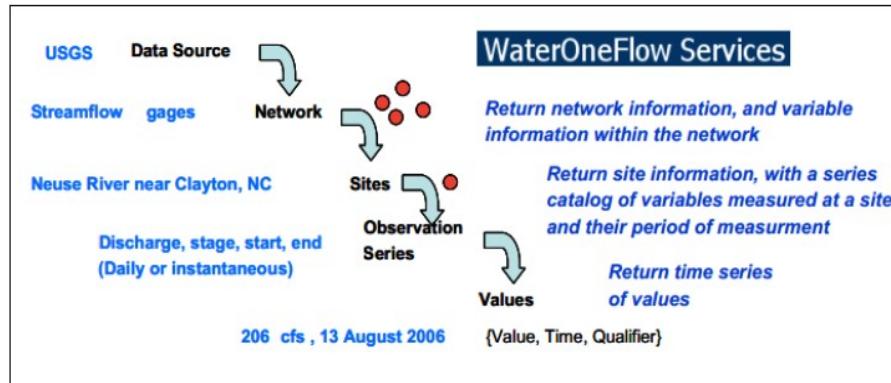
| Drug Name | Degree | Node ID |
|--------------------|--------|---------|
| Tretinoin | 257 | 46 |
| Levothyroxine | 173 | 36 |
| Methotrexate | 156 | 32 |
| 4-Hydroxytamoxifen | 115 | 25 |
| Estradiol | 98 | 20 |
| Amantadine | 79 | 1 |
| Rifampin | 78 | 13 |
| Propofol | 54 | 5 |
| Indinavir | 51 | 14 |
| Penicillamine | 44 | 10 |
| Daunorubicin | 44 | 12 |
| Tricosan | 42 | 5 |
| Darunavir | 40 | 15 |

Metadata is a Necessity!



WaterOneFlow Web Services

Standard mechanism for the TRANSFER of hydrologic data between hydrologic data servers (databases) and users computers, streamlining the tasks of extracting data from a data source, transforming it into a usable format and loading it into an analysis environment.



<https://www.cuahsi.org/Standards>

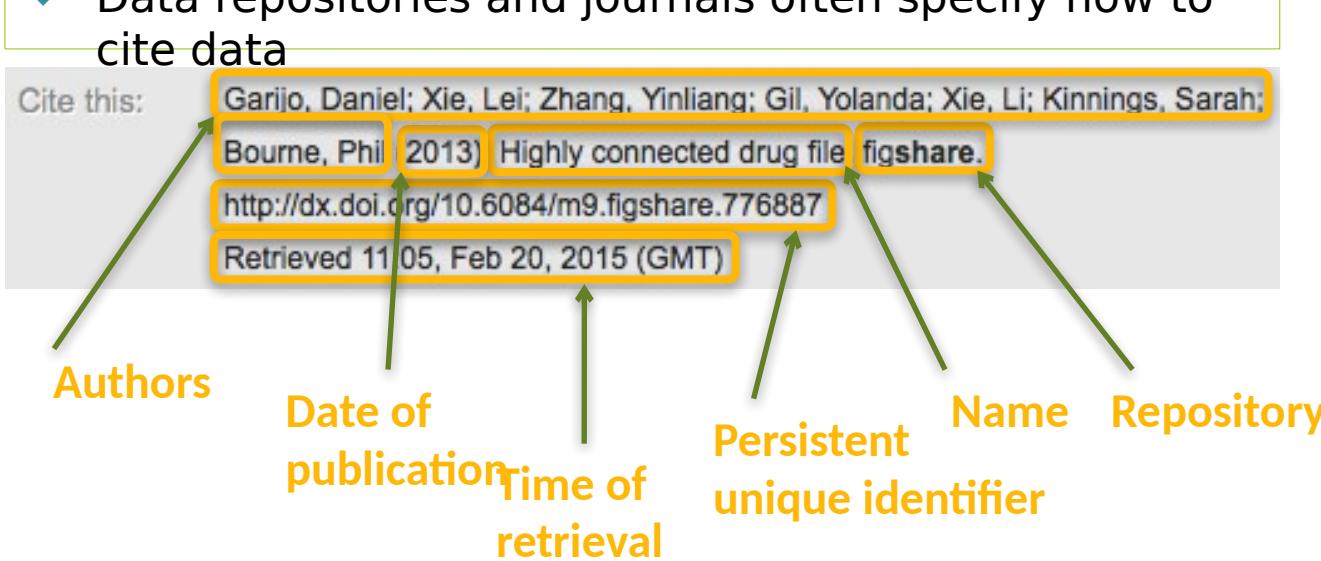


Main Types of Unique Identifiers

- Uniform Resource Locator (URL)
 - Minimal effort to create
 - No guarantee of persistence
 - i.e., almost guaranteed it will not have persistence
 - e.g., <http://www.greatuniversity.edu/gradstudents/joesmith/awesomedata/>
 - DO NOT USE FOR PAPER
- Persistent URL (PURL)
 - The same PURL can be resolved to different Web address over time
 - You always refer to your data with the same PURL:
<http://purl.org/mydataandme/awesomedata.html>
 - Today you are in grad school and tell purl.org to resolve it to:
<http://www.wisc.edu/myadvisorsgroup/awesomedata.html>
 - Tomorrow you have graduated and tell purl.org to resolve it to:
<http://www.stanford.edu/myowngroup/awesomedata.html>
 - It is easy to create your own PURLs, just remember to update whenever you move the data
 - Go to <https://w3id.org> (run by W3C), <http://www.purl.org> (run by OCLC), or other PURL services
- Digital Object Identifier
 - DOIs can only be issued by a DOI authority (e.g., a journal publisher)
 - Data repositories can issue DOIs for data

Data Citation

- ◆ Data repositories and journals often specify how to cite data



Share this:

Share 0

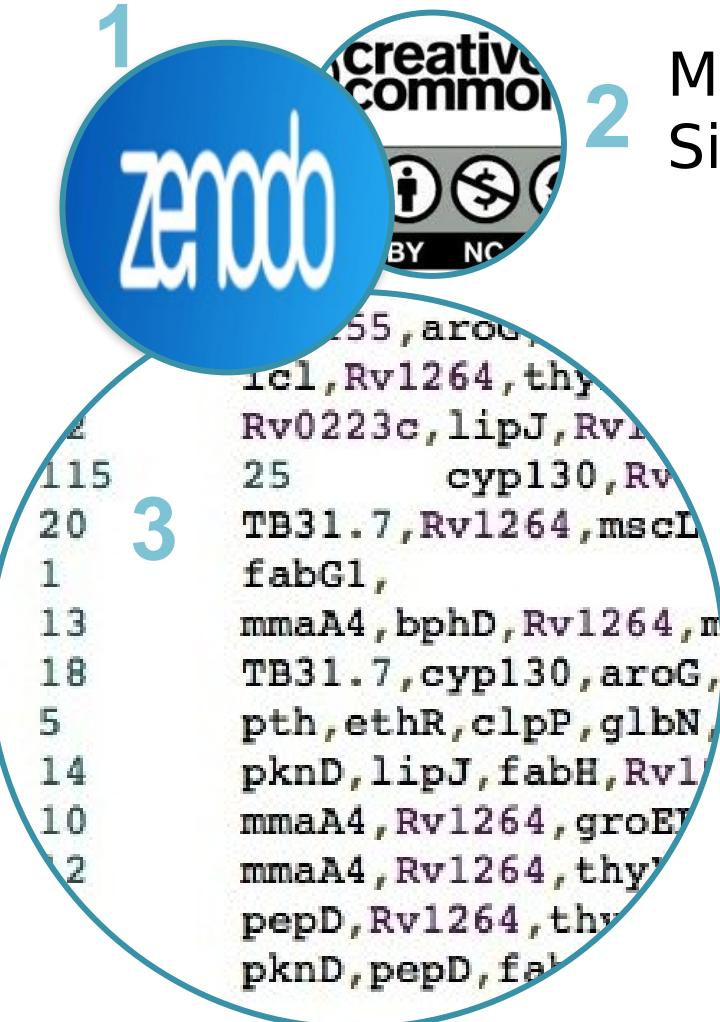
Tweet 0

g+1 0

Embed*

What if?

- **... there are several datasets in several files?**
 - Create a DOI for each file and a DOI for the whole set
- **... the data is from a public repository?**
 - Publish the query, create a DOI + metadata for it, mention the original source in the metadata, point to the original data source
- **... the data is from a colleague?**
 - Get permission in advance and make an agreement, then do as with the data from a public repository
- **... the data comes from many sources?**
 - Credit each source, create URIs as needed
 - Can create a table with “microattribution” that summarize each data source
- **... the data comes from a database?**
 - Create a file (or files) from it
- **... the data has many versions?**
 - Create a DOI either for each slice or for each snapshot



Making Data Accessible: Simplest Approach

- 1. Create a public entry for your dataset with a persistent unique identifier**
 - Go to a domain repository (use a general repository, e.g., figshare, if you cannot find one), create an account
 - Create an entry for your dataset
- 2. Specify the metadata**
 - Including license -- choose from <http://www.creativecommons.org/licenses>
- 3. Upload/point to the data**

Making Data Accessible: Ideal Approach

1. **Find a repository that your community uses, if there is not one then organize one!**
2. **Create a public entry for your dataset with a persistent unique identifier**
 - Create an entry for your dataset
3. **Specify the metadata**
 - Including license -- choose from <http://www.creativecommons.org/licenses>
4. **Upload/point to the data**
5. **Get a data citation from the repository**



2

Rv1155, aroG, aroI, Rv1264, thyR, Rv0223c, lipJ, Rv125, cyp130, Rv122, TB31.7, Rv1264, mscL, fabG1, mmaA4, bphD, Rv1264, mscL, TB31.7, cyp130, aroG, pth, ethR, clpP, glbN, pknD, lipJ, fabH, Rv125, mmaA4, Rv1264, groEL, mmaA4, Rv1264, thyR, pepD, Rv1264, thyR, pknD, pepD, fabH

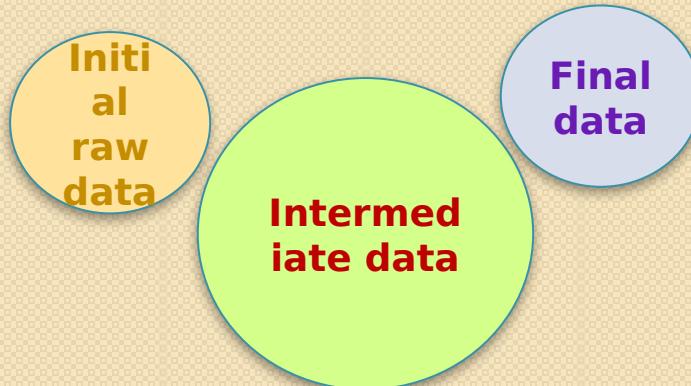
3

115
20
1
13
18
5
14
10
12

- **Citation goes in the References section**
- **How to cite the data? You choose:**
 - With an in-text pointer as you would cite any other paper (recommended)
 - With an in-text pointer in a special “Data Resources” section
 - With an in-text pointer in the “Acknowledgments” section

MAKING DATA ACCESSIBLE:

CITE THE DATA IN YOUR PAPER



Recommended Exercises

- Browse through the following repositories:
 - Figshare (Generic)
 - Pangea (Curated)
 - CUAHSI (Domain-specific)
- Answer the following questions:
 - Is the repository sustainable?
 - Is there a curation process?
 - How are persistent identifiers assigned?
 - Are the citation and license clearly shown?
- Create an account on Figshare
 - Using Figshare to find data
 - Find a dataset on Figshare that is of interest to you.
 - Can you identify the 5 best practices for this dataset?

Reading Assignments

- Please submit 4-5 Slide PowerPoint presentation that summarizes the paper.
- Come up with one Question and Answer that shows that you have read and understood the paper.