Data Stewardship & Sustainability

Digital Scholarship Services
University Library System
University of Pittsburgh

Data Stewardship means Making Choices

Source: **Cliff Strutz**
http://www.roadfood.com/photos/10203.jpg

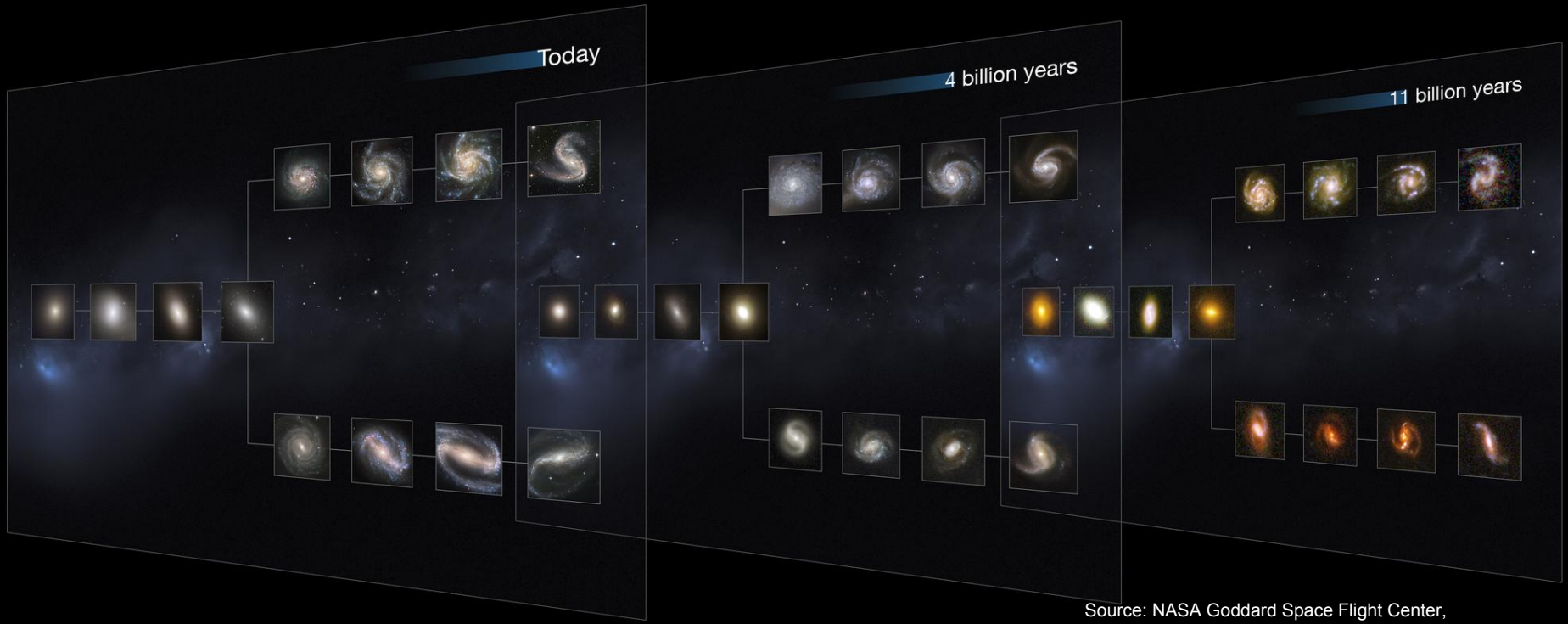
to support

Discoverability
and
Sustainability

but also

Integrity / Trust
Credit / Attribution
Security
Provenance
Usability
Visibility
Compliance
Interoperability
Metrics
Lower cost $$$ more?

...and stewardship means supporting these over time



Source: NASA Goddard Space Flight Center,
https://www.flickr.com/photos/gsfc/9524854754

Digital data is fragile, the web is volatile

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

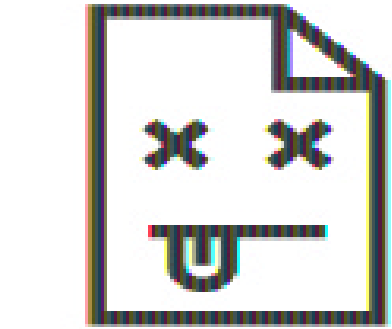
Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

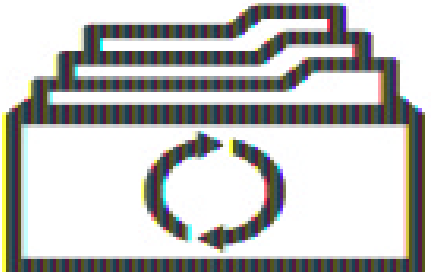
Published: December 26, 2014 • DOI: 10.1371/journal.pone.0115253

Broad open access today...

it's also good for long-term sustainability



Format Obsolecence



Archive (concept)

Another perspective on the same:

Lots of copies keeps stuff safe

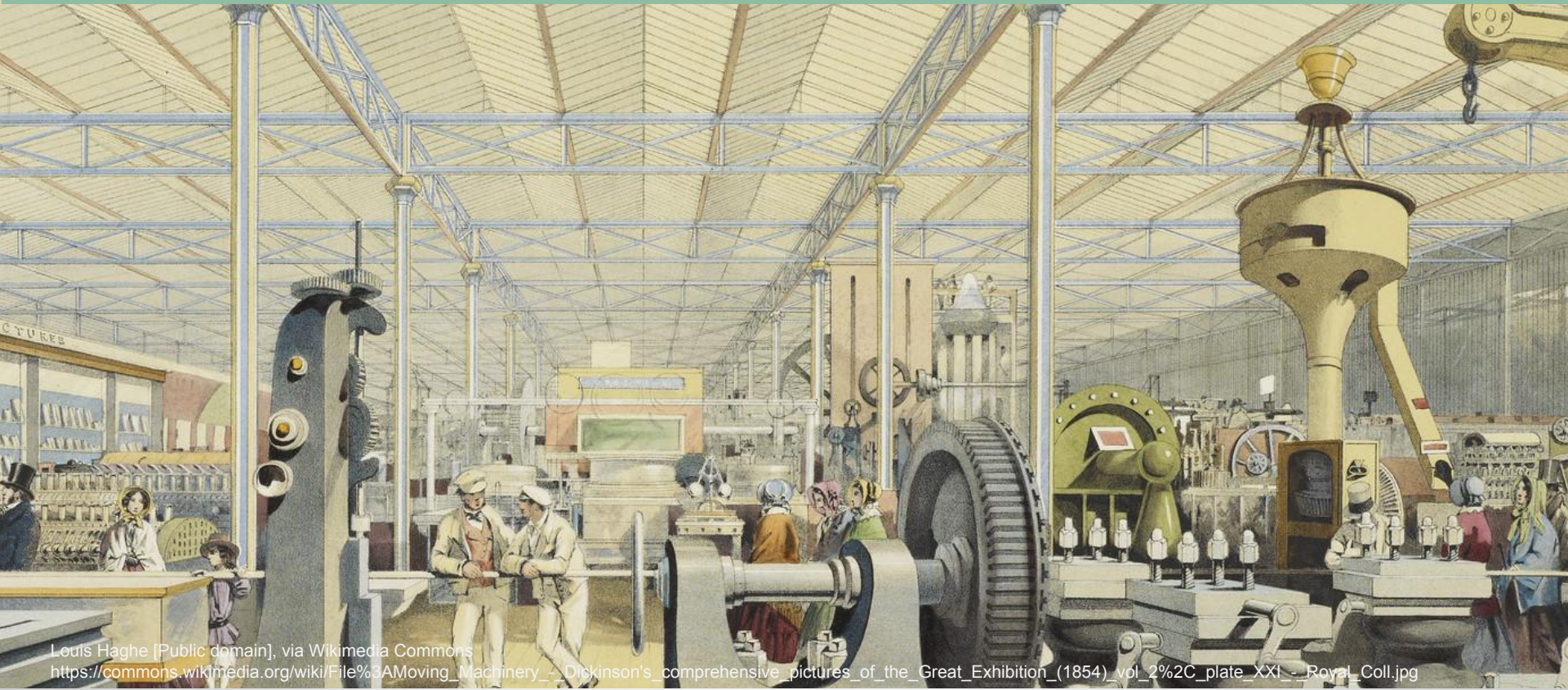
Lots of description keeps stuff meaningful

Lots of services keeps stuff useful

Lots of uses keeps stuff valuable

Abrams, S., Cruse, P., & Kunze, J. (2009, May). *Permanent Objects, Disposable Systems*. Presented at the 4th International Conference on Open Repositories, Georgia Institute of Technology, Atlanta GA. Retrieved from https://smartech.gatech.edu/handle/1853/28490

Though we're at a tech workshop, please note: *there is no digital stewardship machine*



Louis Haghe [Public domain], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3AMoving_Machinery_-_Dickinson's_comprehensive_pictures_of_the_Great_Exhibition_(1854)_vol_2%2C_plate_XXI_-_Royal_Coll.jpg

Structures & *planning* aid the decision-making *process*
in policies:

UK Data Service

Collections Development Policy

1.3. Curation Categories

The policy of the UK Data Service will be to treat collections in four broad categories:

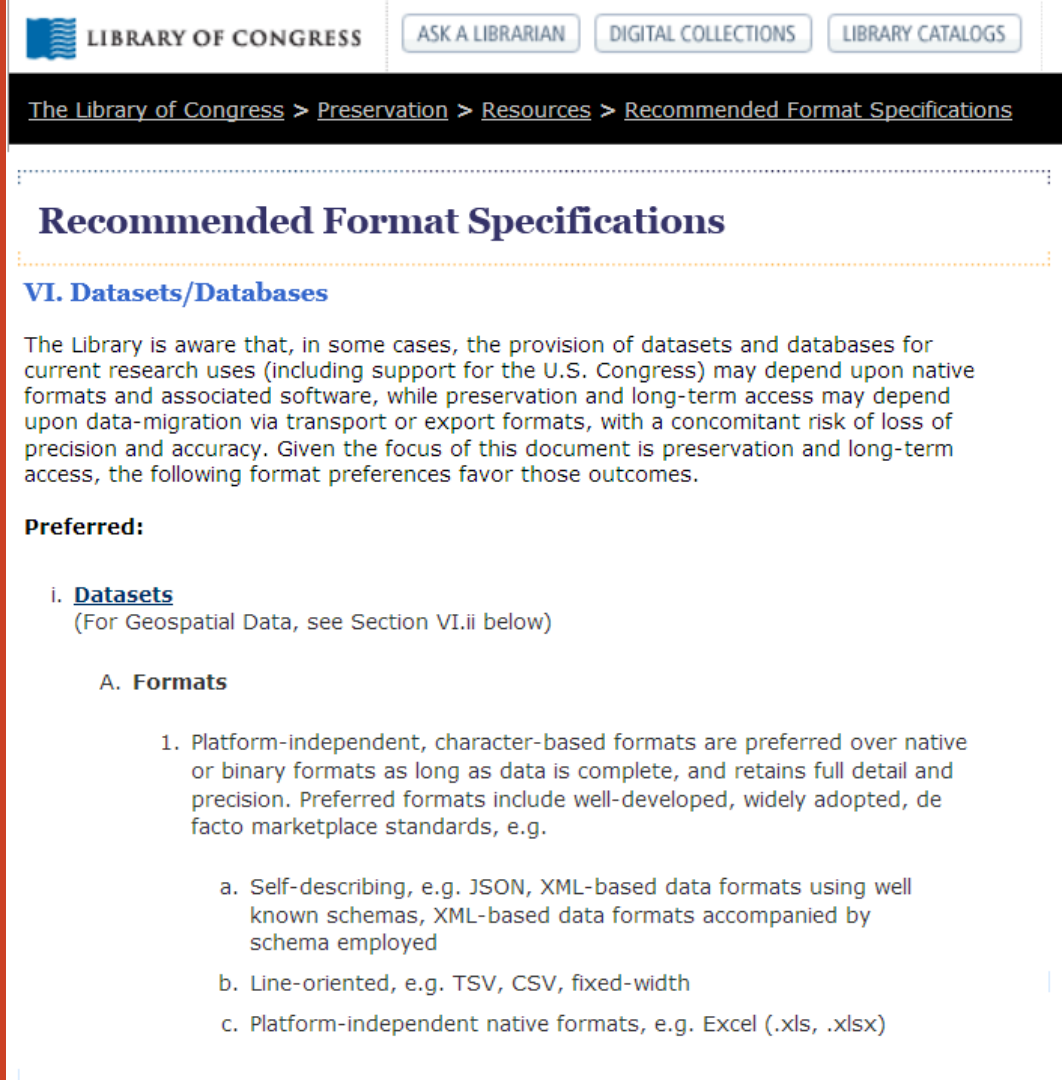
1. *Data collections selected for long-term curation*. These data will have long term secondary analysis potential. These collections are likely to be made available for download, or accessible via online access tools;
2. *Data collections selected for "short-term" management*. These collections will not (initially) be retained for long-term preservation, rather they will be backed-up (i.e., bit-level preservation only), made accessible and discoverable through online access tools[v] or via repository software (ReShare);
3. *Data collections selected for 'delivery' only*, e.g., where data from third parties are accessed via APIs/web services and delivered to end users via a UK Data Service interface. Issues such as level of trust in owner, what documentation/metadata are required, and how rights/registration are handled would need to be agreed;
4. *Data collections selected for "discovery" only*. These collections will not be brought formally into the holdings of the UK Data Service, they will exist only in other (institutional) repositories, but the UK Data Service will harvest (or in exceptional circumstances, create) metadata records to allow these data collections to be found more easily.

Data collections may initially fall into categories two or three, but be moved to category one at a point of time in the future.

http://ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf

...in recommended formats

http://www.loc.gov/preservation/resources/rfs/data.html



LIBRARY OF CONGRESS | ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS

The Library of Congress > Preservation > Resources > Recommended Format Specifications

Recommended Format Specifications

VI. Datasets/Databases

The Library is aware that, in some cases, the provision of datasets and databases for current research uses (including support for the U.S. Congress) may depend upon native formats and associated software, while preservation and long-term access may depend upon data-migration via transport or export formats, with a concomitant risk of loss of precision and accuracy. Given the focus of this document is preservation and long-term access, the following format preferences favor those outcomes.

Preferred:

i. **Datasets**
(For Geospatial Data, see Section VI.ii below)

A. **Formats**

1. Platform-independent, character-based formats are preferred over native or binary formats as long as data is complete, and retains full detail and precision. Preferred formats include well-developed, widely adopted, de facto marketplace standards, e.g.

   a. Self-describing, e.g. JSON, XML-based data formats using well known schemas, XML-based data formats accompanied by schema employed

   b. Line-oriented, e.g. TSV, CSV, fixed-width

   c. Platform-independent native formats, e.g. Excel (.xls, .xlsx)

...in file names

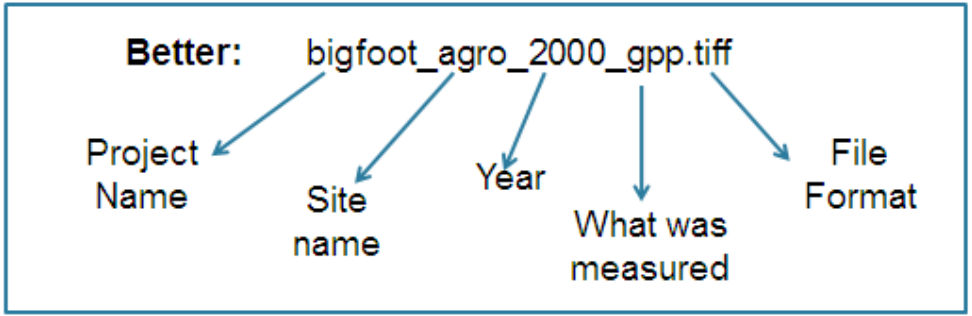
source: Federation of Earth Science Information Partners' Data Management for Scientists Short Course. "Assign Descriptive File Names." authored by Robert Cook from the Oak Ridge National Laboratory. http://commons.esipfed.org/node/702

Assign descriptive file names

- Use descriptive file names
  - Unique
  - Reflect contents
  - ASCII characters only
  - Avoid spaces
- Provide an explanation of the convention used to name files

Bad: Mydata.xls
2001_data.csv
best version.txt

Better: bigfoot_agro_2000_gpp.tiff



Structures & tools are important, but people are crucial for stewardship, too

Data Management in Universities and its Drivers

"There are three primary (and related) motivations for developing a robust data curation infrastructure: enabling new discoveries by exposing data for use in data-driven research, ensuring access to and preservation of scholarly output, and meeting existing or forthcoming requirements of funding agencies or institutions regarding data management, retention, and access."

Cornell University Library Data Working Group, "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library," May 2008, accessed April 30, 2015, http://ecommons.library.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf

Data Management in Universities and its Drivers

Benefits to research & public good

Requirements from stakeholders

Good data management begins with a plan.

A valuable resource for data management planning --

Digital Curation Centre (2013), Checklist for a Data Management Plan. v.4.0, Edinburgh: Digital Curation Centre, accessed April 30, 2015. Available online: http://www.dcc.ac.uk/resources/data-management-plans

Points to consider:

1. “What types of data will be produced in terms of format, file size, and classification?

2. What metadata standards do you need to follow for documentation?

3. Do any considerations need to be make to protect sensitive information, including study participant confidentiality and intellectual property protection?

4. What policies do you need to follow with respect to data sharing and reuse?

5. How will you ensure archiving and preservation of the data you will produce?”

From University of Minnesota Libraries, “Creating a Data Management Plan,” https://www.lib.umn.edu/datamanagement/DMP

Graphic from Bielefeld University https://data.uni-bielefeld.de/en/data-management-plan

Should we plan to save all the data we create?

Not a very practical approach in the long-run --

Finding what you need later can be made more difficult;

Digital content will continue to grow;

There are costs associated with properly preserving data (i.e. backups; human resources; creating metadata)

mkuram's Flickr stream, https://www.flickr.com/photos/mkuram/4872078284/

Considerations for what to keep --

Relevance to organizational mission

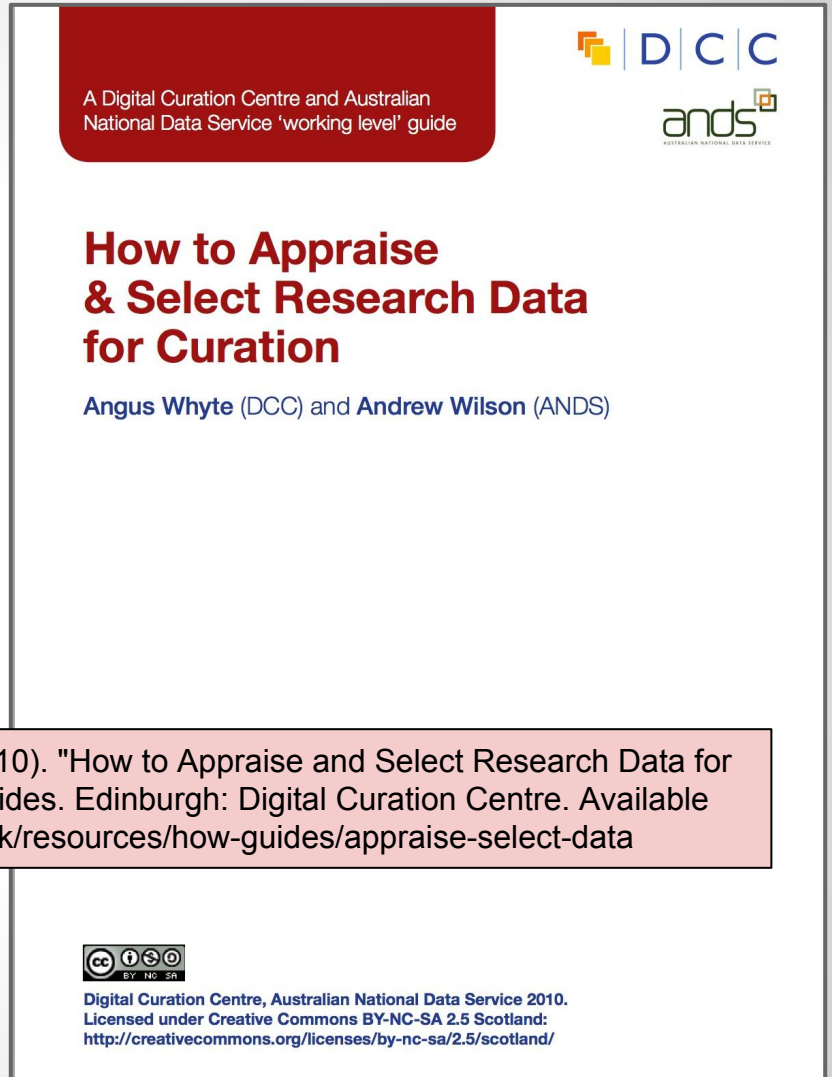
Legal requirements

Uniqueness

Future historical/scientific value

Replicability

Costs



Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

5.0 City Guidelines

5.1 Data Set Selection

Agencies should use the following guidelines to select and prioritize their data sets for publication.

5.1.1 Prioritization Criteria

For purposes of prioritizing public data sets, Agencies should consider whether information embodied in the public data set:

- Increases Agency accountability and responsiveness;
- Improves public knowledge of the Agency and its operations;
- Responds to a need or demand identified by the public;
- Furthers the mission of the Agency;
- Reduces the impact of automated tools which scan the City's website for data;
- Fosters agency/interagency efficiency; or
- Creates economic opportunity.

5.1.2 Public Input and Participation

Public input is essential to selecting and disseminating information. The NYC OpenData portal includes an online forum to solicit feedback from the public and to encourage public discussion on open data policies and public data set availability. Agencies should use this forum to solicit recommendations regarding the presentation of data, data types, and metadata from individuals, groups, and organizations.

Excerpt from NYC's *Open Data Policy and Technical Standards Manual*, September 2012. Available online: http://www.nyc.gov/html/doitt/downloads/pdf/nyc_open_data_tsm.pdf

About licensing

"If information is to be truly public, and maximally re-usable, there should be no license-related barrier to the re-use of public information. To be completely "open," public government information should be released completely into the worldwide public domain and clearly labeled as such."

Sunlight Foundation, "Open Data Policy Guidelines," accessed April 30, 2015. Available online: http://sunlightfoundation.com/opendataguidelines/#license-free

For a brief and useful discussion on IP and data, see: Peter Hirtle, “Introduction to Intellectual Property Rights in Data Management, 2011, https://confluence.cornell.edu/display/rdmsgweb/introduction-intellectual-property-rights-data-management

Available online: http://opendatacommons.org/licenses/

Open Data Commons *Legal tools for Open Data* Home License

Licenses

- Public Domain Dedication and License (PDDL) — “Public Domain for data/databases”
- Attribution License (ODC-By) — “Attribution for data/databases”
- Open Database License (ODC-ODbL) — “Attribution Share-Alike for data/databases”

Applying a License

For instructions on how to apply the licenses to your material please see each license’s home page.

Questions or suggestions? There is:

- A License FAQ (plus the General FAQ).
- A public mailing list, wiki page and contact email. See the contact page for details.

An Open Knowledge Foundation project.
Find out more about open data or read the definition of open data on OpenDefinition.org.

Privacy Considerations

Data and Safety Monitoring Plan

Documentation from the University of Pittsburgh IRB

Definitions:

A data and safety monitoring plan (DSMP) is a specific plan, developed by the local principal investigator (PI), that outlines how study progress will be monitored throughout the course of the research to ensure the safety of subjects as well as the integrity and confidentiality of data.

Overview:

It is required that every research study, with the exception of studies designated as “exempt,” includes a **formal** data and safety monitoring plan.

Privacy Considerations

There is a need to be mindful of both *direct identifiers* and *indirect identifiers* in your data.

The UK Data Archive notes --

"Anonymising research data can be time consuming and therefore costly. Early planning can help reduce the costs...**Anonymisation techniques for quantitative data may involve removing or aggregating variables or reducing the precision or detailed textual meaning of a variable**. Special attention may be needed for relational data, where connections between variables in related datasets can disclose identities, and for geo-referenced data, where identifying spatial references also have a geographical value."

UK Data Archive, "Anonymisation" guidance, accessed May 4, 2015. Available online: http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

UK Information Commissioner's Office. *Anonymisation: Managing Data Protection Risk Code of Practice*, November 2012, accessed May 4, 2015. https://ico.org.uk/media/1061/anonymisation-code.pdf

UK Information Commissioner's Office, *The Guide to Data Protection,* v2.2, March 31, 2015, accessed April 30, 2015. https://ico.org.uk/media/for-organisations/documents/1607/the_guide_to_data_protection.pdf

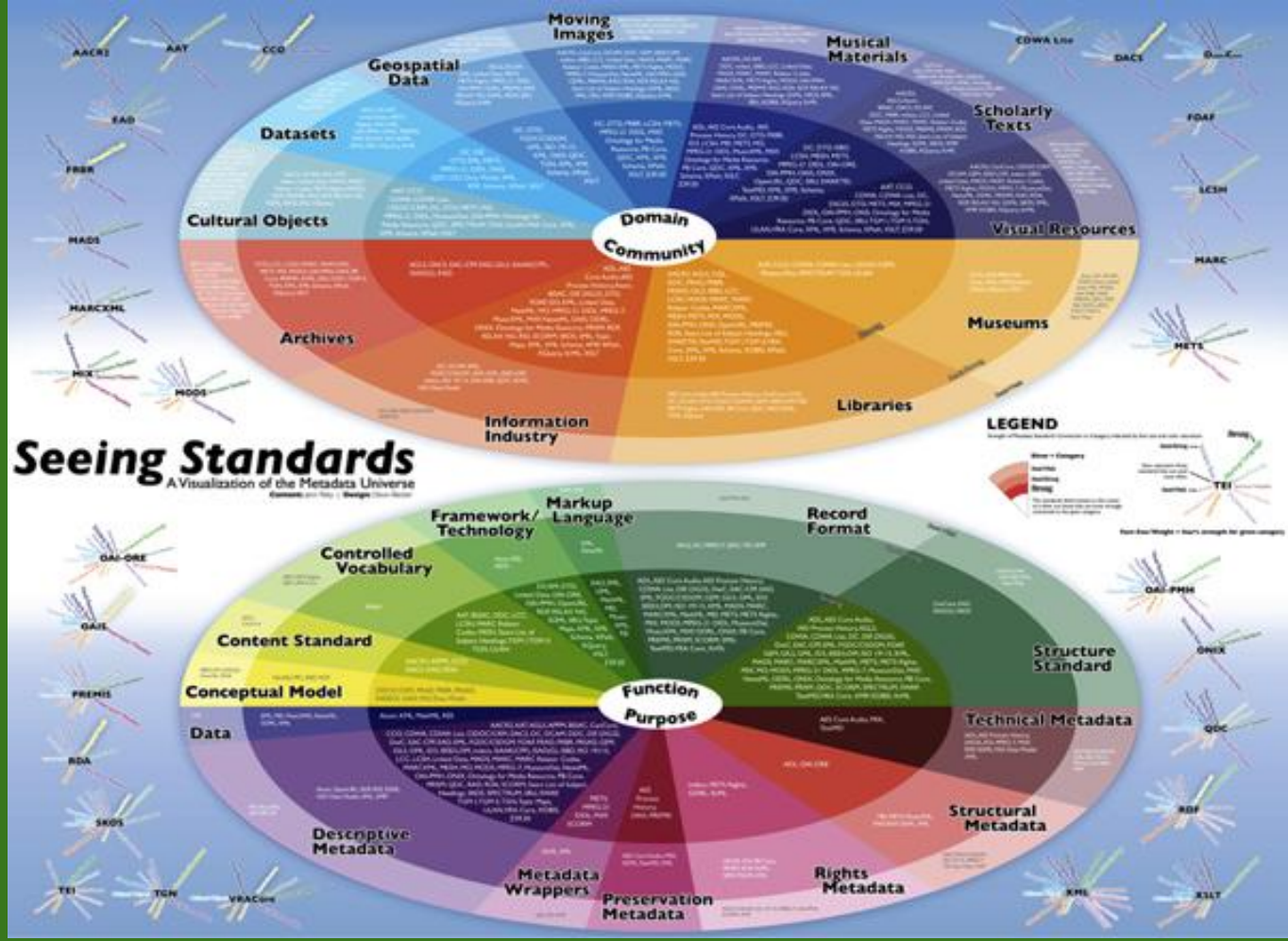
Metadata: Data about Data

A better definition...

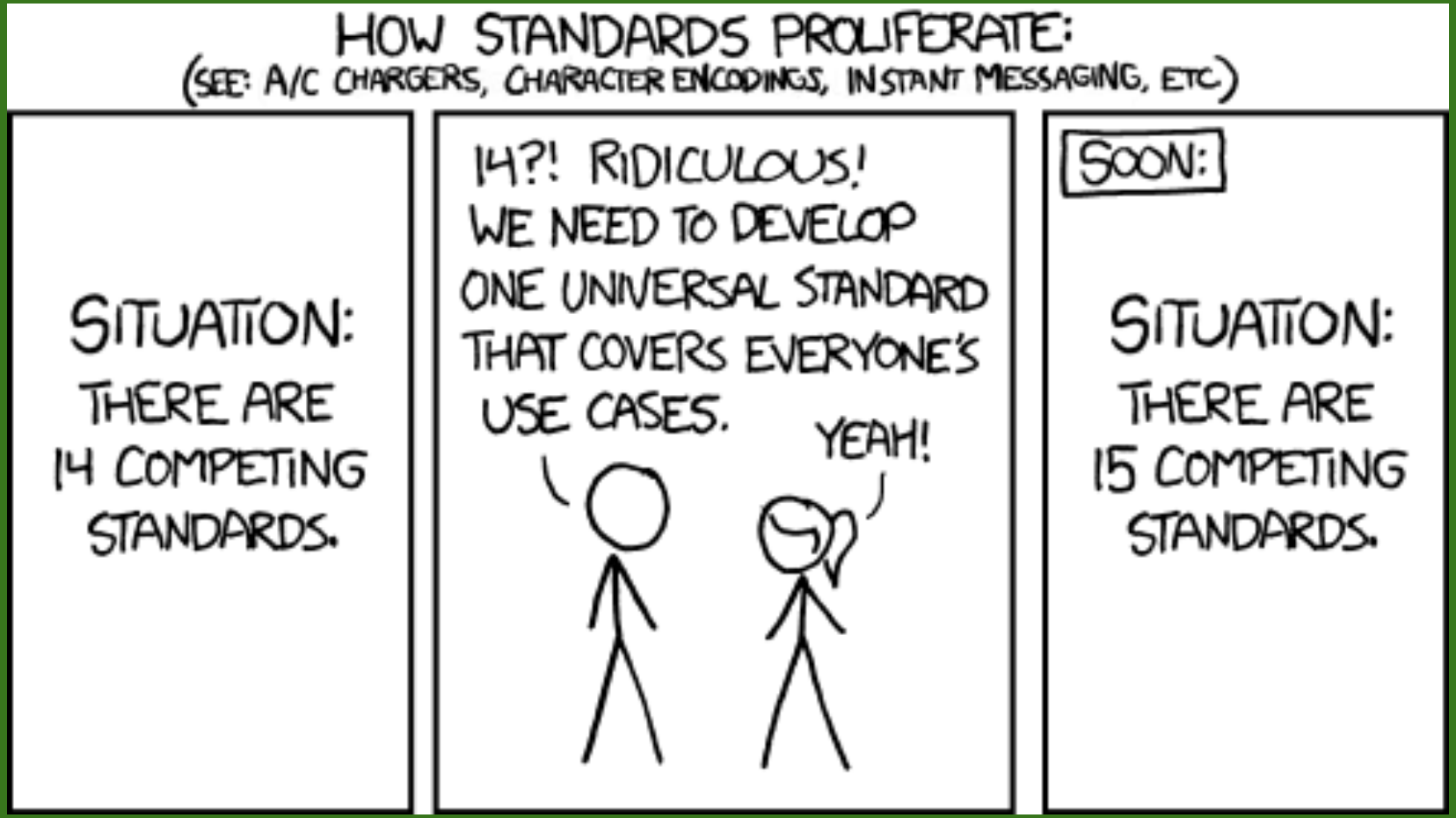
“In essence, metadata is the extra baggage associated with any resource that enables a real or potential user to find that resource; to decide whether or not is valuable to them; to discover where, when, and by whom it was created, as well as for what purpose; to know what tools will be needed to manipulate the resource; to determine whether or not they will actually be allowed to access the resource and how much this will cost them. Metadata is, in short, a means by which largely meaningless data may be transformed into information, interpretable and reusable by those other than the creator or the data resource.”

Paul Miller – “Metadata: What It Means for Memory Institutions” In Metadata Applications and Management (2004)

Consistency & Standards

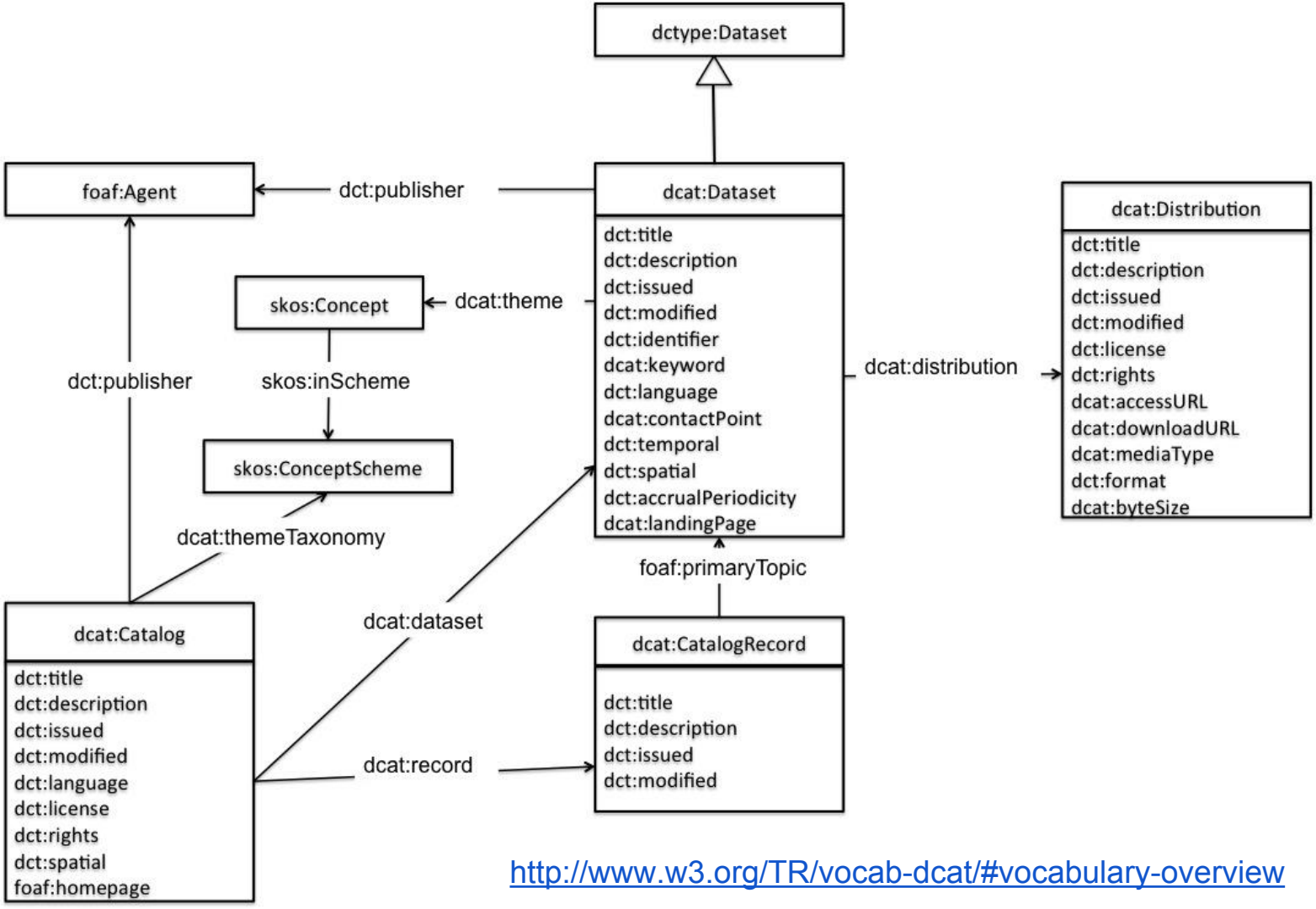


Riley, Jennifer. (2010). Seeing Standards: A Visualization of the Metadata Universe. http://www.dlib.indiana.edu/~jenlrile/metadatamap/



Data Catalog Vocabulary (DCAT)

- An RDF vocabulary designed to describe data sets and facilitate interoperability between data catalogs published on the Web
  - http://www.w3.org/TR/vocab-dcat/
- Built upon existing, well known, and vetted metadata schema -- Dublin Core

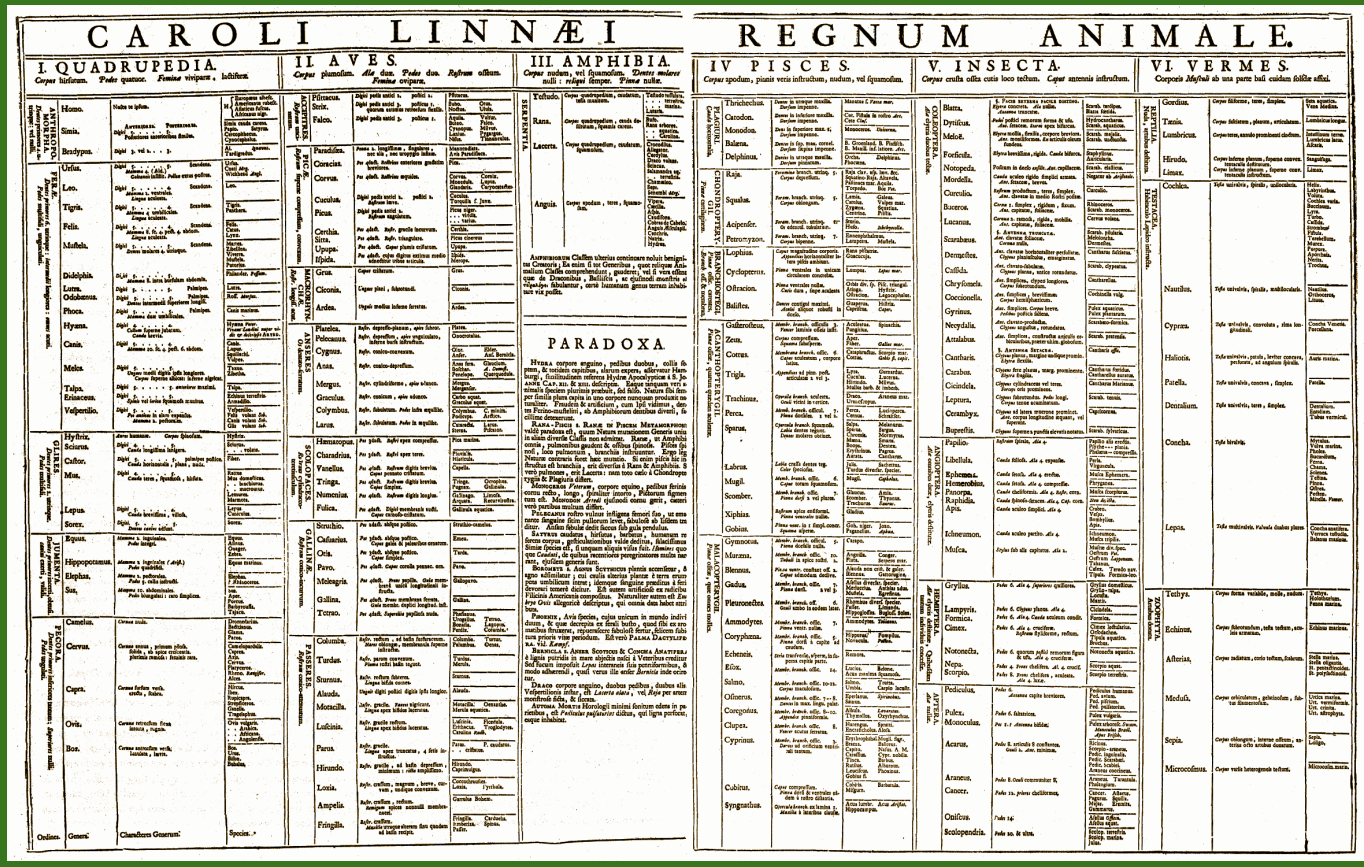


http://www.w3.org/TR/vocab-dcat/#vocabulary-overview

Project Open Data Metadata Schema v1.1

- A set of required fields (title, description, tags, etc.), based on DCAT, for describing data sets in U.S. Government data catalogs.
  - https://project-open-data.cio.gov/v1.1/schema/
- Data.gov follows the schema for every data set displayed on http://data.gov.

Controlled Vocabularies, Thesauri, & Taxonomies



CAROLI LINNÆI

REGNUM ANIMALE.

I. QUADRUPEDIA.

II. AVES.

III. AMPHIBIA.

IV. PISCES.

V. INSECTA.

VI. VERMES.

PARADOXA.

- Geonames
  - Geographical database covers all countries and contains over eight million place names
  - http://www.geonames.org/
- Internet Assigned Number Authority (IANA) Media Types
  - aka MIME Types
  - http://www.iana.org/assignments/media-types/media-types.xhtml

Data Dictionaries

setName	fldName	fldType	fldMin	fldMax	Dec	fldTitle	Order1	Group1	TwoColum	Default	PickL
a_users	useralias	C	3	20		User Name	2	all	☐		
a_users	password	C	4	15		Password	4	all	☐		
a_users	password2	C		15		Verify Password	4.5	all	☐		
a_users	Zodiac	L					5	all	☐	--	Aries, Lib
a_users	fullname	C	2	40		Name	7	all	☐		
a_users	phone1	C	0	40		Phone	9	read	☐		
a_users	email	C	5	80		E-Mail Address	11	all	☐		
a_users	ccn	C	0	20		Credit Card Numbe	13	all	☐		
a_users	othercontact	C	0	60		Other Contact Info	17	all	☐	test defaul	
a_users	whenAdded	D					19	none	☐		
a_users	State	L					23	all	☐	--	
a_users	BirthMonth	L				Birth Month	30	all	☑	Mar.	Jan.,Feb.
a_users	BirthDay	L				Birth Day	31	all	☑		1,2,3,4,5
a_users	BirthYear	N		2999	0	Birth Year	33	all	☑		
a_users	BirthTime	C	0	15		Birth Time	35	all	☑		
a_users	userExpires	D	1/1/1990	12/31/2999		End Date	37	all	☐		
a_users	Notes	M		1000			40	all	☐		
a_users	subscribed	Y				Wants Promo Ema	43	all	☑		
a_users	discount	Y				Discounted	48	all	☑	True	
sample2	orderID	n			0	Order ID	1	R	☑		
sample2	Descript	c	2	30		Description	2	A	☑		
sample2	price1	n	0.00	9000	2	Amount	3	A	☑		
sample2	Tax	n		9000	2		4	A	☑		
sample2	price2	n			2	Alternate Price	6	R	☑		
sample2	nullMe	c		250		Intro	7	A	☑		
sample2	nonNull	c		99		Mfr.	8	A	☑		

Record: 6 of 27

Documentation

Bonanjo - Centre de documentation et information urbanisme (CUD). Photo by Marta Pucciarelli, Douala, 2013. - Bonanjo - Centre de documentation et information urbanisme (CUD) 05.JPG. Licensed under CC BY-SA 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Bonanjo_-_Centre_de_documentation_et_information_urbanisme_%28CUD%29_05.JPG

Application Profile

Field #	accrualPeriodicity
Cardinality	(0,1)
Required	No
Accepted Values	ISO 8601 Repeating Duration (or `irregular`)
Usage Notes	Must be an ISO 8601 repeating duration unless this is not possible because the accrual periodicity is completely irregular, in which case the value should simply be `irregular`. The value should not include a start or end date but rather simply express the duration of time between data publishing. For example, a dataset which is published on an annual basis would be `R/P1Y`; every three months would be `R/P3M`; weekly would be `R/P1W`; and daily would be `R/P1D`. Further examples and documenation can be found here.
Example	`{"accrualPeriodicity":"R/P1Y"}`

Field #	bureauCode
Cardinality	(0,n)
Required	Yes, for United States Federal Government agencies
Accepted Values	Array of Strings
Usage Notes	Represent each bureau responsible for the dataset according to the codes found in OMB Circular A-11, Appendix C (PDF, CSV). Start with the agency code, then a colon, then the bureau code.
Example	The Office of the Solicitor (86) at the Department of the Interior (010) would be: `{"bureauCode":["010:86"]}`. If a second bureau was also responsible, the format like this: `{"bureauCode":["010:86","010:04"]}`.

Field #	conformsTo

Crosswalk

Catalog Fields

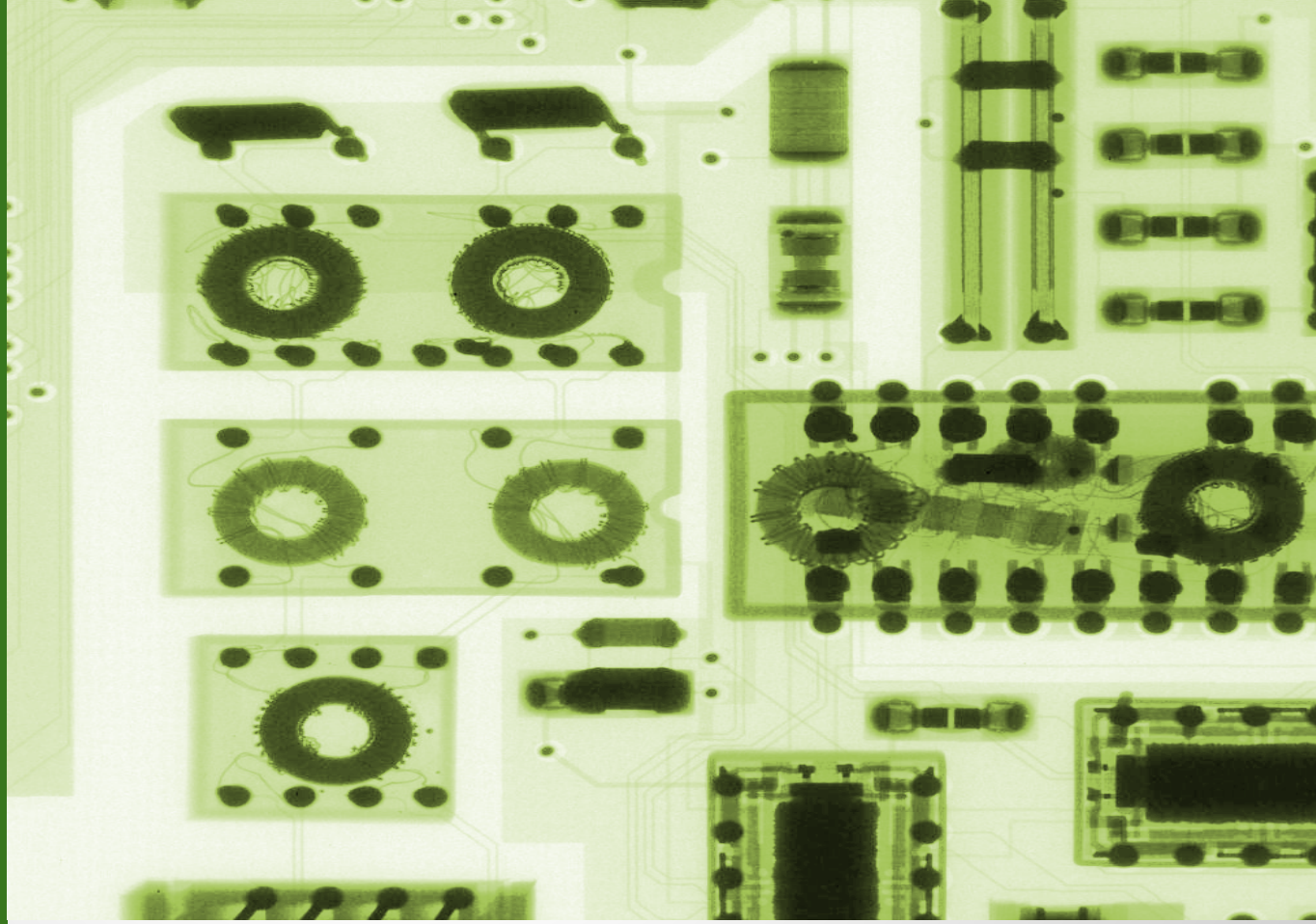
Label	POD v1.1	POD v1.0	CKAN API	DCAT	Schema.org
Metadata Context	*@context*	*n/a*	*n/a*	*n/a*	*n/a*
Metadata Catalog ID	*@id*	*n/a*	*n/a*	*n/a*	*n/a*
Metadata Type	*@type*	*n/a*	*n/a*	*n/a*	*itemtype attribute*
Schema Version	*conformsTo*	*n/a*	*n/a*	*n/a*	*n/a*
Schema URL	*describedBy*	*n/a*	*n/a*	*n/a*	*n/a*
Dataset	*dataset*	*n/a*	*results*	dct:dataset	dataset

Dataset Fields

Note the mapping for `license` and `rights` from Project Open Data to DCAT applies the fields from the Dataset object in Project Open Data to each of the Distribution objects in DCAT.

Label	POD v1.1	POD v1.0	CKAN API	DCAT	Schema.org
Metadata Type	*@type*	*n/a*	*n/a*	*n/a*	*itemtype attribute*
Title	*title*	*title*	*title*	dct:title	name
Description	*description*	*description*	*notes*	dct:description	description
Tags	*keyword*	*keyword*	*tags*	dcat:keyword	keywords
Last Update	*modified*	*modified*	*n/a*	dct:modified	dateModified
Publisher	*publisher → name*	*publisher*	*organization → title*	dct:publisher → foaf:name	publisher → Organization:name
Publisher Parent Organization	*publisher → subOrganizationOf*	*n/a*	*n/a*	dct:publisher → org:subOrganizationOf	publisher → Organization:memberOf
Contact Name	*contactPoint → fn*	*contactPoint*	*maintainer*	dcat:contactPoint → vcard:fn	provider → Person:name
Contact Email	*contactPoint → hasEmail*	*mbox*	*maintainer_email*	dcat:contactPoint → vcard:hasEmail	provider → Person:email
Unique Identifier	*identifier*	*identifier*	*id*	dct:identifier	*n/a*
Public Access Level	*accessLevel*	*accessLevel*	*n/a*	*n/a*	*n/a*

Improving Metadata Quality



"X-Ray Circuit Board Zoom 2" by X-Ray_Circuit_Board_Zoom.jpg: SecretDiscderivative work: Emdee (talk) - X-Ray_Circuit_Board_Zoom.jpg. Licensed under CC BY-SA 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:X-Ray_Circuit_Board_Zoom_2.jpg#/media/File:X-Ray_Circuit_Board_Zoom_2.jpg / green filter

- Documenting the Process
  - Leveraging existing standards
  - Creating application profiles
  - “Requiring” data dictionaries
  - Understanding software & hardware requirements
  - Considering storage, maintenance, & stewardship needs
- Documenting the Policies
  - Collection & retention
  - Privacy & rights
  - Data cleaning/scrubbing
  - Licensing & reuse

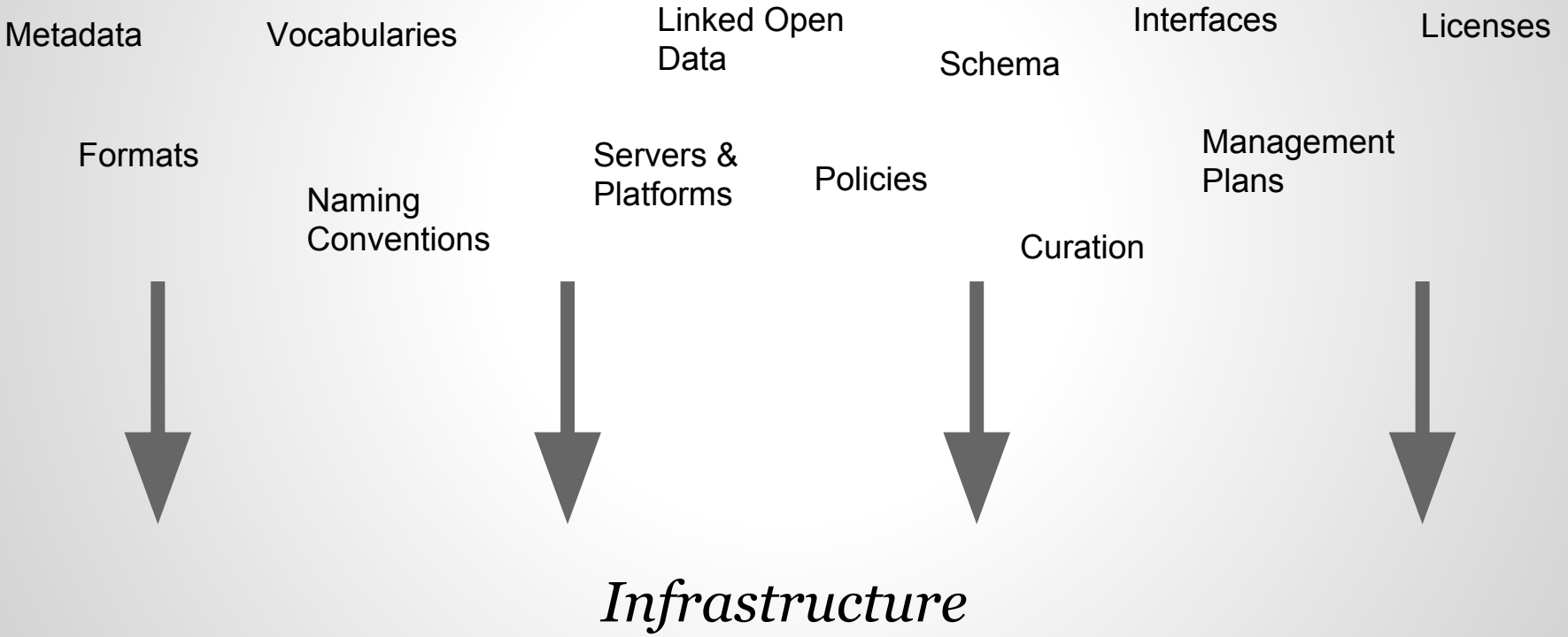
Infrastructure?

Source: Mac Qin
https://www.flickr.com/photos/qin1109/9627495095/

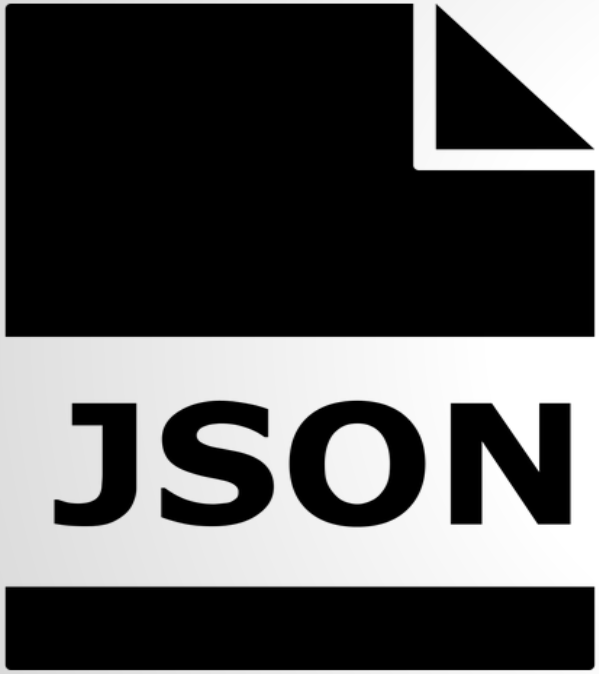
Infrastructure

Source: Kordite
https://www.flickr.com/photos/kordite/6578136503

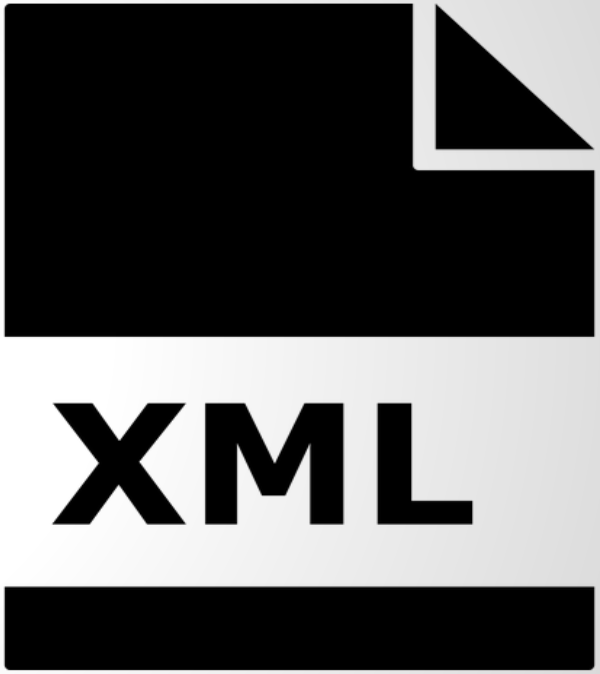
Choices sink into Infrastructure



Standards are Infrastructure



Created by useiconic.com
from the Noun Project



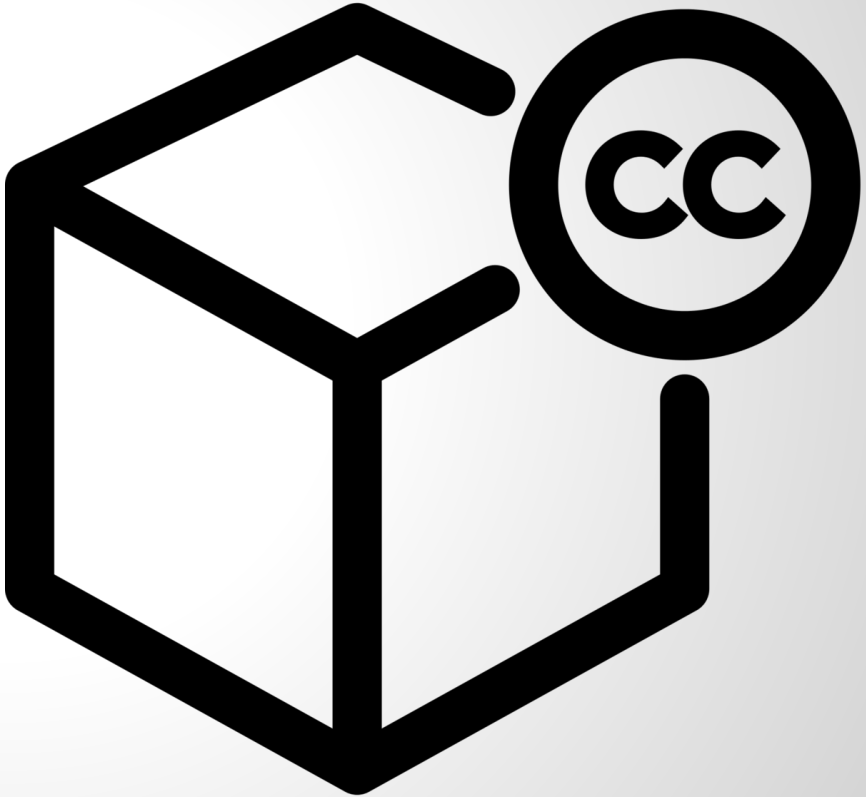
Created by useiconic.com
from the Noun Project

APIs are Infrastructure

- Computational access
  - Computers talking to Computers
- Allows for *generative* access
  - Discoverability
  - Interoperability
  - and more!
- Promotes use
  - Use promotes value

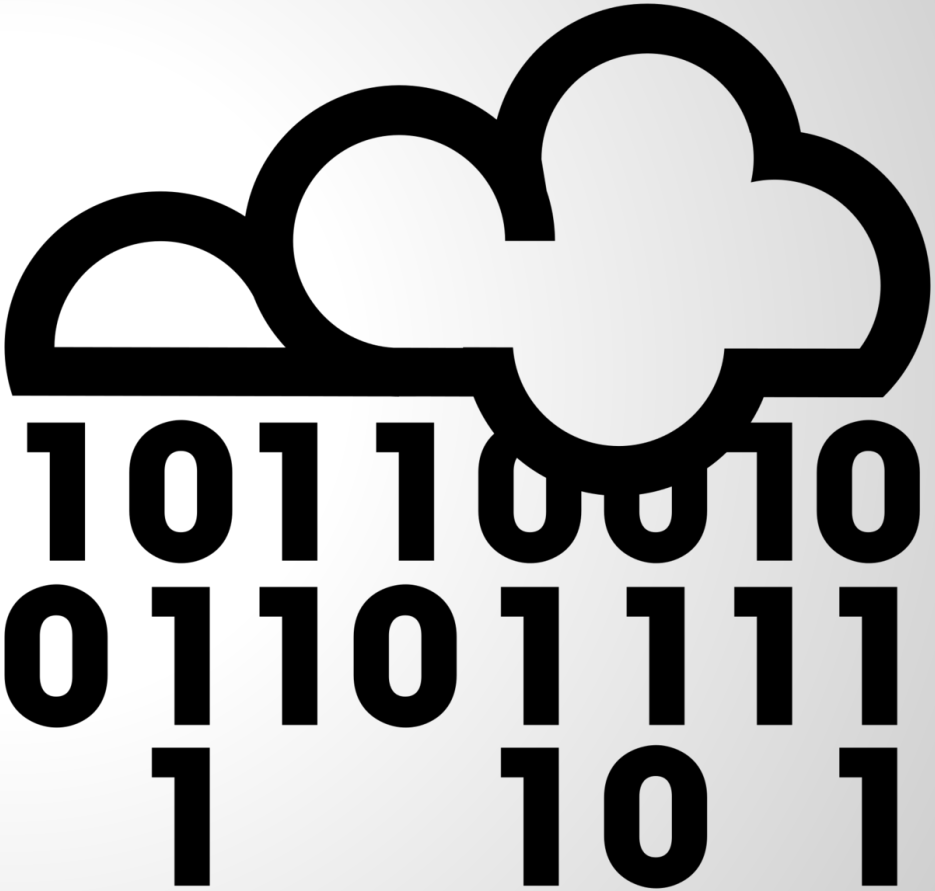
Licenses are Infrastructure

- Legal Access
  - Lawyers talking to Lawyers
- Allows for *generative* use
  - Use keeps stuff valuable
- Creative Commons
  - Free as in Beer
  - Free as in Speech



The Cloud is Infrastructure

- Infrastructure as a Service (IaaS)
- Maintenance becomes *other people's problem*
- But what about stewardship?
  - Dark side of the cloud



INFRASTRUCTURE IS PEOPLE!

One person’s infrastructure...

Source: Zach Frailey
https://www.flickr.com/photos/zrfraileyphotography/6510707249/

...is someone else's job.

Librarians know all this...stuff

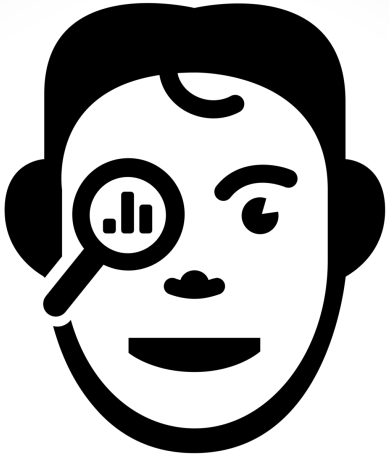
- And we are here to help!
- Social and Technical Expertise
- Find us at:
  - Public & Academic Libraries
  - Schools of Information / Library Science

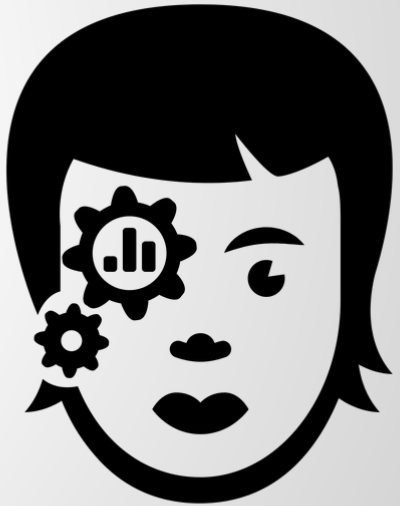
Created by Wynne Nafus Sayer from the Noun Project

Questions?

Cyborgs...er...Data Professionals

Created by Thibault Geffroy
from the Noun Project





Created by Thibault Geffroy
from the Noun Project

Hackers & Security