

Drafting Guidelines for Vocabulary Selection by Data Curators and Repositories

— Presenters:

Margaret O'Brien 0000-0002-1693-8322

Mark Schildhauer 0000-0003-0632-7576

Agenda

Brief Orientation & session info (Erin)

Why Vocabularies? (Margaret, Mark; 20 min)

Crowdsourcing Vocabulary experiences: Everyone (10 min)

Commonalities, Needs, Priorities: Breakouts (5 groups, random, 15 min)

Choosing Vocabularies-- Discussion & synthesis: Plenary (20 min)

Report-back: moderators summarize (5 x 2 mins)

Group Synthesis

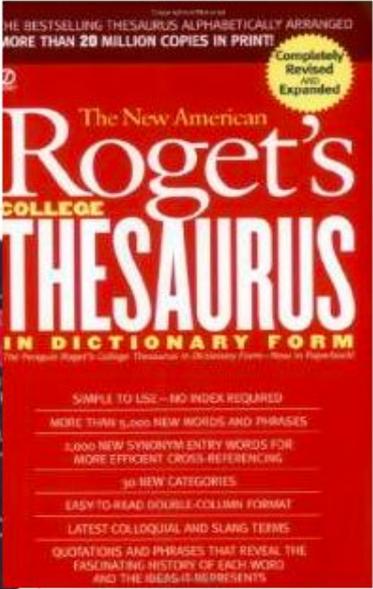
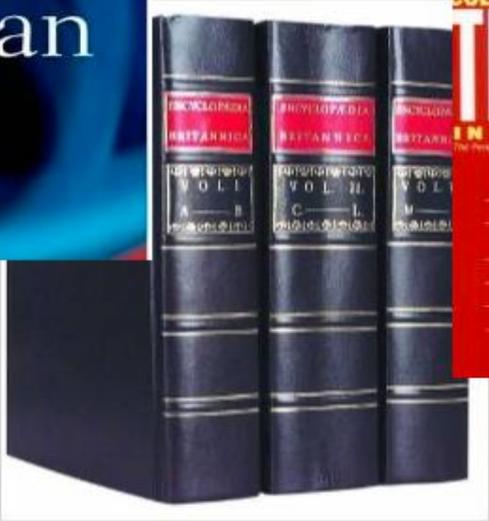
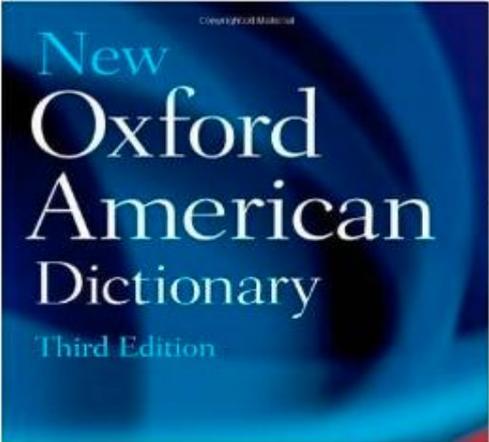
Rebuttals, Omissions and Harrumphs: Breakouts (8-9/room, random, 10 min)

Consensus: Action Items (13 min)

Terminological Clarity makes data more FAIR*

- **Findable:** Enhanced recall/precision of FINDING desired measurements
- **Accessible:** ESIP guidelines can facilitate common Access interfaces and services
- **Interoperable:** W3C-based Semantic Web frameworks and languages
- **Reusable:** detailed descriptions of Dataset contents are key to facilitating proper data re-use

Classical Semantic References



(was) the Platinum *standard*

kil·o·gram (-gram') *n.* [Fr. *kilogramme*: see KILO- & GRAM] a unit of weight and mass, equal to 1,000 grams (2.2046 lb.); also, chiefly Brit., *kil'ogramme'*: abbrev. *kg* (sing & pl)



Today, based on Planck's constant, velocity of light, and atomic transition frequency of Cesium, i.e. the kilogram is now measured in space (meters) and time (seconds)

...and then there are most of the rest

earth (ɜrth) *n.* [ME. *erthe* < OE. *eorthe*, akin to G. *erde* < IE. base **er-*, whence Gr. *era*, earth, Corn. *erw*, field, Du. *oarde*, earth] 1. the planet that we live on; terrestrial globe: it is the fifth largest planet of the solar system and the third in distance from the sun: diameter, 7,927 mi.: symbol, ⊕ 2. this world, as distinguished from heaven and hell 3. all the people on the earth 4. land, as distinguished from sea or sky; the ground 5. the soft, granular or crumbly part of land; soil; ground 6. [Poet.] *a*) the substance of the human body *b*) the human body *c*) the concerns, interests, etc. of human life; worldly matters 7. the hole of a burrowing animal; lair 8. [Obs.] a land or country 9. *Chem.* any of the metallic oxides, formerly classed as elements, which are reduced with difficulty, as alumina, zirconia, strontia, etc. 10. *Elec.* [Brit.] same as GROUND¹ —*vt.* 1. to cover (*up*) with soil for protection, as seeds or plants 2. to chase (an animal) into a hole or burrow —*vi.* to hide in a burrow: said of a fox, etc. —**come back** (or **down**) **to earth** to stop being impractical; return to reality —**down to earth** practical; realistic —**on earth** of all things: an intensive used mainly after interrogative pronouns [what *on earth* do you mean?] —**run to earth** [*< use in fox hunting*] 1. to hunt down 2. to find by search

1 or

3 or

4 or

5 or

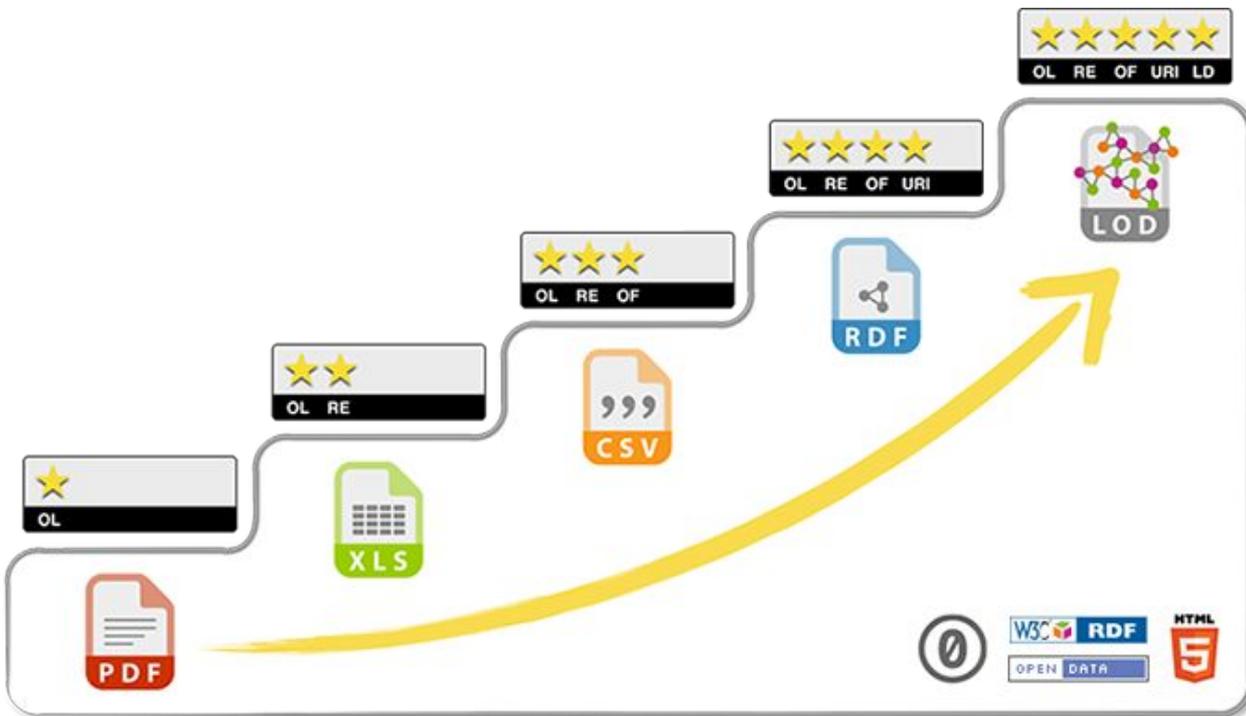
9

?

AND... we've got the Web to Help!!!

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. This scheme contains the five steps that this entails.

1. 1 star - make your stuff available on the web (whatever format) under an open license
2. 2 stars - make it available as structured data (e.g. Excel instead of image scan of a table)
3. 3 stars - make it available in a non-proprietary open format (e.g. CSV instead of Excel)
4. 4 stars - use URIs to denote things, so that people can point at your stuff
5. 5 stars - link your data to other data to provide context (LOD / Linked Open Data)

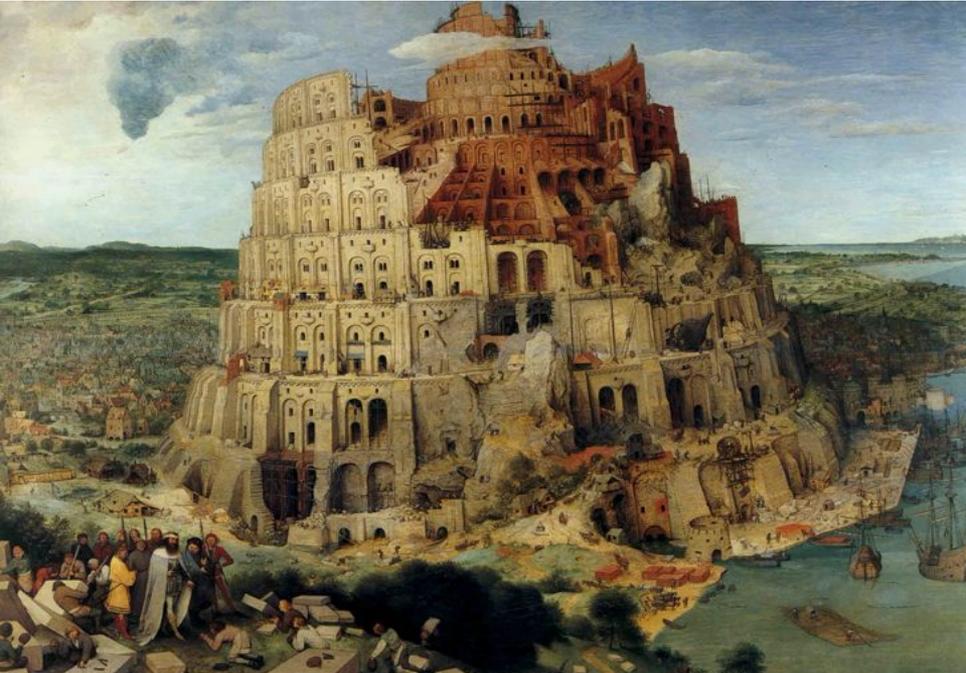


Choosing Vocabularies

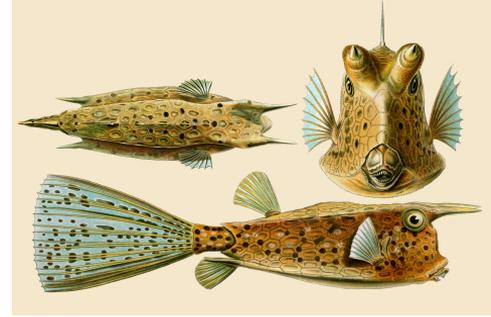
“A Confusion of Tongues”

OR
?

“A Peaceable Kingdom”



Point of View: Vocabulary User



I want to annotate this dataset with formal terms from a vocabulary ...

- Which vocabulary should I use to annotate the whole thing? Part of it, like a single measurement?
- How do I find a term's URI?
- Is this SKOS vocabulary OK? Or should I look for something more complex?
- Is one version of a vocabulary more reliable than another?

Point of View: Repository Engineer



Our community wants us to support vocabulary [XYZ]

- Is [XYZ] in a system with a Web-accessible API?
- Is it stable?
- What format is it stored in?
- For a display about the term on a data set's web page,
 - which attributes?
- We want to build a tool to help users understand (or choose) terms. What term attributes should we filter on? Display to a user?

Features to Consider - Sociocultural

- **Topical Coverage**
 - Topics and terms that fit your data? Are definitions included?
- **Level of community adoption**
 - Wide use and awareness within discipline?
 - Across disciplines?
 - Does it use terms from other vocabularies?
- **Governance process**
 - open or not? (to add or amend terms)
 - responsive? support?
- **Prospects for sustainability**

Basic Features to Consider - Technical

- **API** for programmatic access, machine-processing and reasoning
- Each term has its own GUID
 - **GUIDs are HTTP IRIs dereferenceable *at the term level over the Web***
- **No “naked terms”**-- must have descriptions or definitions, their sources, and ideally examples of use
- **Terms persistent**; not deprecated without redirect or other information
- **Tools** for vocabulary maintenance, revisions
- **Human-readable interface** for exploring and choosing terms

Advanced Technical Features

- Vocabularies are expressed using W3C-recommended frameworks and languages such as **RDF**, **OWL**, and **SHACL**
- **Re-use don't reinvent:** Vocabularies “borrow/import” terms from other well-established vocabularies where possible
- Terms inter-related “vertically” via **Class/SubClass** hierarchies
- Terms inter-related “horizontally” through **logical Predicates** (Object and Datatype properties)
- **Vocabulary is an Ontology**, that affords reasoning and inferencing of new “facts”



Ontology repositories make “Terms” more FAIR*

BioPortal

<http://bioportal.bioontology.org>

ESIP COR

<http://cor.esipfed.org>

Ontoportal Alliance

<http://ontoportal.org>

Ontobee

<http://ontobee.org> ...and others...

- **Findable, Accessible:** FINDING and understanding desired terms
- **Accessible:** ESIP guidelines can facilitate common Access interfaces and services
- **Interoperable:** W3C-based Semantic Web frameworks and languages
- **Reusable:** adoption of common terminologies facilitates interpretation and re-use (think “kilogram”!)

Crowdsourcing

start time: 00:20, duration = 10 minutes

Individually summarize your experience with vocabularies

Choose a best fit for Theme and Role.

Many of you undoubtedly play many Roles, and are involved with several Themes

Use these spreadsheets:

1. A-D: https://docs.google.com/spreadsheets/d/1VEfNae0Zp2BVBaM3ciTIGINyGRhICM8I-51_0Nmqxm8
2. E-I: <https://docs.google.com/spreadsheets/d/1oIMsR56CwcQ0CZc6CPLmd7vSBIDnsGOyV1veYyLhcug>
3. J-N: https://docs.google.com/spreadsheets/d/10sIN8Mgy15spDMSNUbOchN78Fbq_OBpBbqjyrOJteQA
4. O-S: <https://docs.google.com/spreadsheets/d/1Ls4mUxaT2ixzejHRVftOXr9OUG3GWGt42wLW4pFTB6Q>
5. T-Z: https://docs.google.com/spreadsheets/d/1dwIALf-QMe8oJV1KjBrdMMogM3kFAAXTbwce_3n8iG8



Breakout A (5 rooms)

start time: 00:30, duration = 15 min

5 rooms (random)

Charge: Collate the experiences

Notes:

https://docs.google.com/document/d/1wWwcuVeVO3EB9v5B6zWuYXTwtsaPL7xrkZJbgJtrG_U/edit#

Volunteers needed: Moderator (does report-out), and Note-taker in each room



Plenary Discussion

start time: 00:45, duration = 20 min

Breakout Moderators summarize (2 min each)

Group Synthesizes

Add to Notes doc:

https://docs.google.com/document/d/1wWwcuVeVO3EB9v5B6zwuYXTwtsaPL7xrkZJbgJtrG_U/edit#

Breakout B (~10 rooms)

start time: 00:65, duration = 10 minutes

Small breakouts, random (everyone talks!)

What additional questions/concerns are missing from synthesized outcomes?

Add to spreadsheet(s), last column

1. A-D: https://docs.google.com/spreadsheets/d/1VEfNae0Zp2BVBaM3ciTIGINyGRhICM8I-51_0Nmqxm8
2. E-I: <https://docs.google.com/spreadsheets/d/1oIMsR56CwcQ0CZc6CPLmd7vSBIDnsGOyV1veYyLhcug>
3. J-N: https://docs.google.com/spreadsheets/d/10sIN8Mgy15spDMSNUbOchN78Fbq_OBpBbqjyrOJteQA
4. O-S: <https://docs.google.com/spreadsheets/d/1Ls4mUxaT2ixzejHRVftOXr9OUG3GWGt42wLW4pFTB6Q>
5. T-Z: https://docs.google.com/spreadsheets/d/1dwlALf-QMe8oJV1KjBrdMMogM3kFAAXTbwce_3n8iG8



Wrap up

Semantic clusters at ESIP-- (participate!)

Semantic Technologies: https://wiki.esipfed.org/Semantic_Technologies

Semantic Harmonization: <https://wiki.esipfed.org/SemanticHarmonization>

Community Ontology Repository: <http://cor.esipfed.org/>

Schema.org: https://wiki.esipfed.org/Schema.org_Cluster

Work with Cox/Wyborn et al. to develop *ESIP Guidelines for Vocabulary Choice*:

<https://2020esipsummermeeting.sched.com/event/clwD/plenary-proliferation-of-vocabularies-in-solid-earth-space-and-environmental-sciences-which-one-should-i-use-and-which-ones-can-i-trust>

Consider activity for existing (or new?) ***ESIP Cluster on Vocabulary Guidelines?***
“3 other takeaways”?

