Citizen Science and Collaborative Data Crosswalks

Mike Linksvayer (@mlinksva)

2016 Workshop on Collaborative Data Projects / Taipei

2016-12-08

Xinzhong Street South View from Fujin Street · 玄史生 · CC0-1.0

1

Mass Collaboration [Policy]: What, Why, and Choices

2

Intellectual Provenance

?

3

easy to get hung up on vulgar policy, eg licensing

think broadly about what, where, ends

4

what is a mass collaboration data project?

a: latent / explicit

y: distributed / centralized

z: crowdsourced / collaborative

--(...

5

consider location of

"science"

"semantic web"

freelance

musicbrainz

openstreetmap

wikidata

dbpedia

zooniverse

...

6

what is a mass collaboration data policy?

7

restrictions & permissions

try to be Open; use terms other projects do; if you have to think, public domain for mass clones, copyleft [look again at nearby projects] for mass recipient, but also possibly assume facts are free...

8

"IP" never the ideal answer

poor substitute for other regulations...

9

...privacy, security, integrity

public licenses dull instruments for these legitimate objectives

when so used overly restrictive, incompatible

10

...transparency, replicability, modifiability

copyleft!

part copyright etc neutralization, part regulatory

wholesome goals

beware FUD, incompatibility

11

explicit

+centralized

+collaborative

= evinces need for public licenses like nothing else

12

governance

of what?

depends on location in each dimension

13

software to facilitate mass collaboration, eg

wikis

--source control

semantic stuff

14

data management

15

consider key ~policy choices (so far) made by a couple of aforementioned projects

16

at what level should "policy" be considered (depends...)

international

jurisdiction

institutional

standards

project

individual contributor

17

some things to desire (~ends) from mass collaboration projects (thus to consider in "policy")

curation

provenance

data improvement, normalization

...

innovation/disruption/not imagined uses

orientation toward public good

18

at least 3 methods of promoting and protecting commons

law&policy

building

ecological

each has reactionary and progressive aspects

19

building

free knowledge & software invalidates assumptions of, mitigates, and builds constituency against bad policy

each contribution to a commons is a strike against bad policy and dystopian future; that's something to celebrate and proclaim!

20

imagine a world in which Wikipedia is not unusual:

where mass collaboration obtains...

disruptive innovation

superior "product"

greater equality

greater freedom

...in all of the most crucial human enterprises

21

links: convey yourself to

22

open data policy for scientists as citizens and for citizen science

23

ACAWIKI

24

in part

Post- thoughts

25

"citizen science"

contributions to scientific processes outside of institutional context and/or by non-scientists?

"scientists as citizens"

scientists [including citizen scientists?] cognizant of the role of their work for society?

26

high priests used to be looked upon for "truth" -- but they controlled it tightly

scientists are looked upon to work towards truth ... and their work is always shared freely and open to scrutiny

at least we have the right ideal now ☺

27

"open data"

anyone, any purpose

--

28

easy to get hung up on vulgar policy, eg licensing

think broadly about what, where, ends

29

what is a mass collaboration data project?

a: latent / explicit

y: distributed / centralized

z: crowdsourced / collaborative

--(...

30

consider location of

"science"

"semantic web"

freelance

musicbrainz

openstreetmap

wikidata

dbpedia

zooniverse

...

31

what is a mass collaboration data policy?

32

restrictions & permissions

try to be Open; use terms other projects do; if you have to think, public domain for mass clones, copyleft [look again at nearby projects] for mass recipient, but also possibly assume facts are free...

33

"IP" never the ideal answer

poor substitute for other regulations...

34

...privacy, security, integrity

public licenses dull instruments for these legitimate objectives

when so used overly restrictive, incompatible

35

...transparency, replicability, modifiability

copyleft!

part copyright etc neutralization, part regulatory

wholesome goals

beware FUD, incompatibility

36

explicit

+centralized

+collaborative

= evinces need for public licenses like nothing else

37

governance

of what?

depends on location in each dimension

38

software to facilitate mass collaboration, eg

wikis

--source control

semantic stuff

39

data management

40

consider key ~policy choices (so far) made by a couple of aforementioned projects

41

at what level should "policy" be considered (depends...)

international

jurisdiction

institutional

standards

project

individual contributor

42

some things to desire (~ends) from mass collaboration projects (thus to consider in "policy")

curation

provenance

data improvement, normalization

...

innovation/disruption/not imagined uses

orientation toward public good...

43

at least 3 methods of promoting and protecting commons

law&policy

building

ecological

each has reactionary and progressive aspects

44

building

free knowledge & software invalidates assumptions of, mitigates, and builds constituency against bad policy

each contribution to a commons is a strike against bad policy and dystopian future; that's something to celebrate and proclaim!

45

imagine a world in which Wikipedia is not unusual:

where mass collaboration obtains...

disruptive innovation

superior "product"

greater equality

greater freedom

...in all of the most crucial human enterprises

46

links: convey yourself to

Disciplining or Empowering the Citizenry Through Citizen Science-
Historical and Normative Perspectives on Knowledge and Power

知識、權力與公民科學
歷史與規範的觀點

Dec 7-9, 2016
@Academia Sinica
Taipei, Taiwan

what is a mass collaboration data project?

- x: latent / explicit
- y: distributed / centralized
- z: crowdsourced / collaborative
- …: …

consider location of

- **“science”**
- **“semantic web”**
- **freebase**
- **musicbrainz**
- **openstreetmap**
- **wikidata**
- **dbpedia**
- **zooniverse**
- **…**

governance

of what?

depends on location in each dimension

explicit

+centralized

+collaborative

= evinces need for public licenses like nothing else

at what level should “policy” be considered (depends…)

international

jurisdiciton

institutional

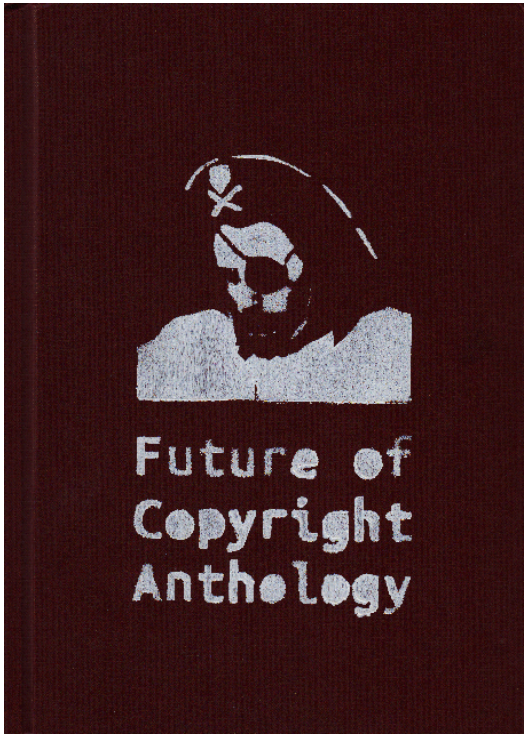
standards

project

individual contributor

“IP” never the ideal answer

poor substitute for other ~~regulation~~disciplining...



Legal Interoperability of Research Data: Principles and Implementation Guidelines

Date: Sep 9, 2016

AddThis

The CODATA-RDA Legal Interoperability Interest Group has studied the issues related to the intellectual property of data: the resulting outcome is a set of principles and practical implementation guidelines. They are offered as high-level guidance to all members of the research community—the funders, managers of data centers, librarians, archivists, publishers, policymakers, university administrators, individual researchers, and their legal counsel—who are engaged in activities that involve the access to and reuse of research data from diverse sources. The Principles are synergistic, so their greatest benefit is realized when they are considered together.

The following Principles on the Legal Interoperability of Research Data focus on all types of data that are used primarily in publicly funded research in government and academia:

- One: Facilitate the lawful access to and reuse of research data.
- Two: Determine the rights to and responsibilities for the data.
- Three: Balance the legal interests.
- Four: State the rights transparently and clearly.
- Five: Promote the harmonization of rights in research data.
- Six: Provide proper attribution and credit for research data.

necessary, but easy to get hung up on vulgar policy, eg licensing

think broadly about what, where, ends

some things to desire (~ends) from mass collaboration projects (thus to consider in “policy”)

curation

provenance

data improvement, normalization

metrics

…

innovation/disruption/not imagined uses

orientation toward public good

imagine a world in which Wikipedia is not unusual:

where mass collaboration obtains...

disruptive innovation

superior “product”

greater equality

greater freedom

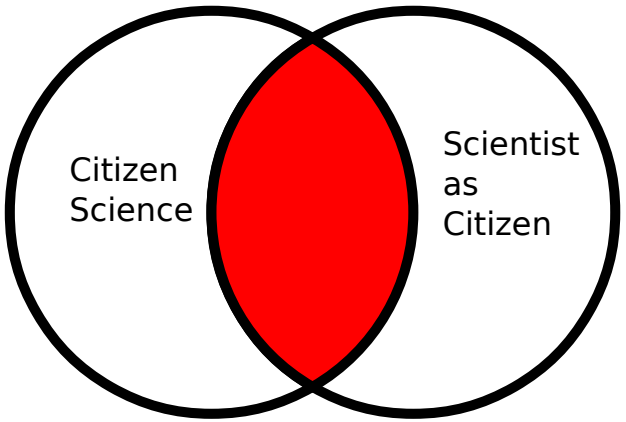
...in all of the most crucial human enterprises

“citizen science”

contributions to scientific processes outside of institutional context and/or by non-scientists?

“scientists as citizens”

scientists [including citizen scientists?] cognizant of the role of their work for society?



high priests used to be looked upon for “truth” … but they controlled it tightly

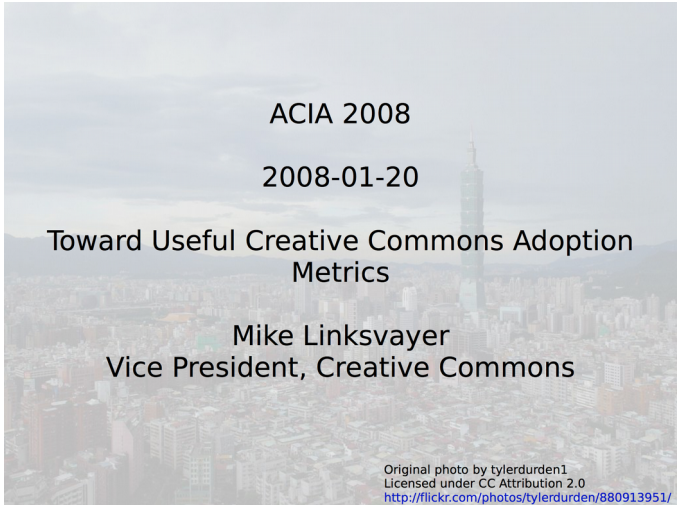
scientists are looked upon to work towards truth … and their work is always shared freely and open to scrutiny

at least we have the right ideal now ☺

Wanted: collaborative data projects about open collaboration...

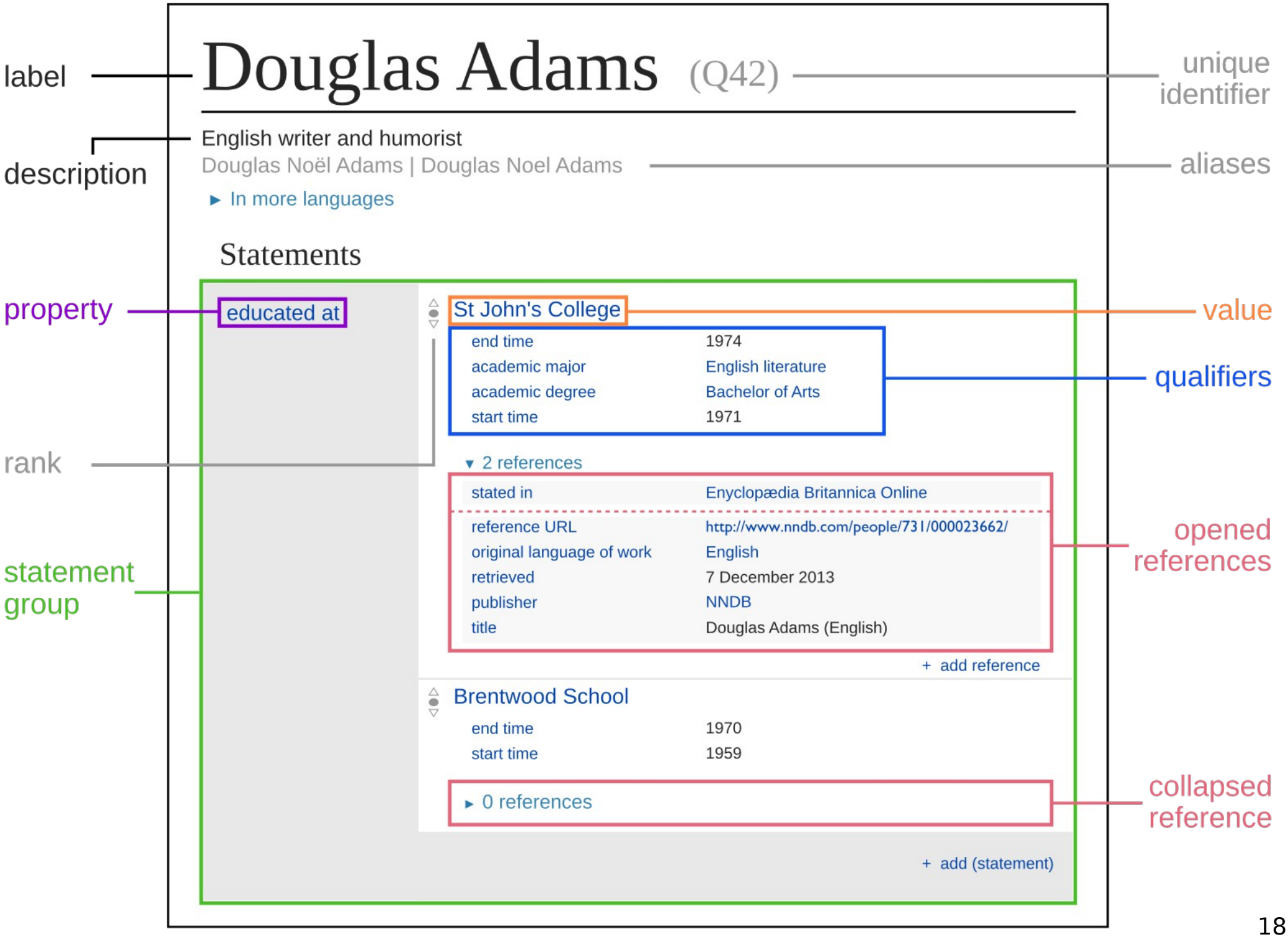
- control data, control arrangements
- surpass counting patents as proxy for innovation
- altmetrics moneyball
- identify and create cultural relevance

...a personal thread



Wikidata is extremely cool and relevant

- Mass collaboration ontology development
- Literal data crosswalks
- Community tech
- Site of data collaboration projects
- Increase value and efficiency of Wikipedias and other projects
- ...


label
Douglas Adams
(Q42)
unique identifier
description
English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
aliases
► In more languages
Statements
property
educated at
St John's College
value
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971
qualifiers
rank
▼ 2 references
stated in Enyclopædia Britannica Online
reference URL http://www.nndb.com/people/731/000023662/
original language of work English
retrieved 7 December 2013
publisher NNDB
title Douglas Adams (English)
opened references
statement group
+ add reference
Brentwood School
end time 1970
start time 1959
► 0 references
collapsed reference
+ add (statement)

Current WikiProjects on Wikidata [edit | edit source]

Some WikiProjects have been created, here is a list of them:

- ▶ Completed WikiProjects
- ▶ Construction WikiProjects
- ▶ Cultural WikiProjects
- ▶ Flemish art collections, Wikidata and Linked Open Data project
- ▶ Geographical WikiProjects
- ▶ GLAM WikiProjects
- ▶ History WikiProjects
- ▶ Lists of WikiProject participants
- ▶ Meta WikiProjects
- ▶ WikiProject Occupations and professions
- ▶ Properties list in a WikiProject
- ▼ Science WikiProjects
  - ▶ WikiProject Astronomy
  - ▶ WikiProject Chemistry
  - ▶ WikiProject Microbiology
  - ▶ WikiProject Molecular biology
  - ▶ WikiProject Open Access
  - ▶ WikiProject Tropical cyclones
  - ▶ WikiProject Wikidata for research
  - ▶ WikiProject Source MetaData
  - ▶ WikiProject Taxonomy
  - *WikiProject Anatomy*
  - *WikiProject Biology*
  - *WikiProject Cat breeds*
  - *WikiProject Chemistry*
  - *WikiProject Dog breeds*
  - *WikiProject Ecology*
  - *WikiProject Geology*
  - *WikiProject Informatics*
  - *WikiProject Mathematics*
  - *WikiProject Medicine*
  - *WikiProject Mineralogy*
  - *WikiProject Molecular biology*
  - *WikiProject Physics*
  - *WikiProject Space*
  - *WikiProject Identification Keys*
- ▶ Social science WikiProjects

Wikidata:Property proposal

Shortcut: WD:PP

Translate this page

Other languages:	العربية • беларуская • বাংলা • bosanski • català • čeština • dansk • Deutsch • Ελληνικά • **English** • Canadian English • British English • Esperanto • español • فارسی • suomi • français • Frysk • ગુજરાતી • עברית • हिन्दी • magyar • Հայերեն • Bahasa Indonesia • Ilokano • italiano • 日本語 • ქართული • 한국어 • Kurdî (latînî) • Lëtzebuergesch • lietuvių • latviešu • македонски • മലയാളം • Bahasa Melayu • norsk bokmål • Nederlands • occitan • polski • پښتو • português • português do Brasil • русский • shqip • српски / srpski • ไทย • Türkçe • українська • Tiếng Việt • 中文

WD:PP redirects here. For the protection policy, please see Page protection policy (shortcut WD:PPP).

Before a new property is created, it has to be discussed here. When after some time there are some supporters, but no or very few opponents, the property is created by a property creator or an administrator.

You can propose a property here or on one of the subject-specific pages listed below.

The property proposals are divided by topic:

Generic (24)	**Authority control** (22)	**Person** (12)	**Organization and politics** (6)	**Event** (12)
Creative work (18)	**Term** (1)	**Space** (0)	**Place** (13)	**Sister projects** (1)
Economics (0)	**Transportation** (5)	**Natural science** (23)	**Property metadata** (0)	**Pending**

2016 Community Wishlist Survey/Categories/Wikidata

< 2016 Community Wishlist Survey

2016 Community Wishlist Survey

Wikidata

26 proposals, 176 contributors, 365 support votes

Cast your votes now! Only support votes are counted (learn more)
Note that proposals are rotated every 6 hours to ensure all get fair visibility.
Please don't post any new proposals; we've got plenty!
Thanks for participating in this year's survey!
Voting closes on 12 December 23:59 UTC.

Watchlists • WikiProjects

Create infobox for books in Wikipedia [edit | edit source]

- **Problem**: There is a lot of bibliographic data and properties in Wikidata, however this cannot be used in Wikipedia because there is no infobox capable of displaying the rich data model (see: d:Wikidata:WikiProject_Books).
- **Who would benefit**: Editors who work with bibliographic data. Readers.
- **Proposed solution**: create a Lua module that can display information

Contents [hide]

1 Create infobox for books in Wikipedia
2 Datatype for Hours:Minutes:Seconds
3 Easily duplicate parts of an item
4 Easy inter-project linking
5 Improve QuickStatements
6 Improved watchlist integration
7 Keep Wikidata in sync with external databases

Science [edit | edit source]

Biology and Medicine [edit | edit source]

Diseases [edit | edit source]

The number of existing translations for diseases in Wikidata [edit | edit source]

Items used: disease (Q12136)

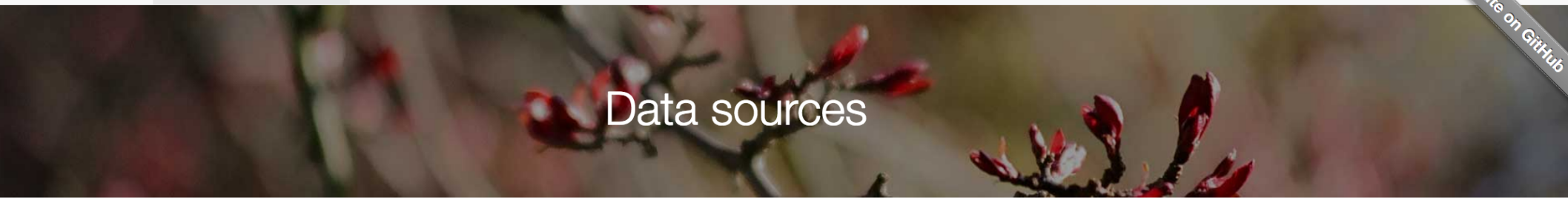
```
#added before 2016-10

SELECT ?disease ?doid ?enLabel (count(?language) as ?languages)
WHERE
{
    ?disease wdt:P699 ?doid ;
             rdfs:label ?label ;
             rdfs:label ?enLabel .
    FILTER (lang(?enLabel) = "en")

    BIND (lang(?label) AS ?language)
}
group by ?disease ?doid ?enLabel
order by desc(?languages)
```

Try it!

A network of Drug-disease interactions on infectious diseases (Source: Disease Ontology, NDF-RT and ChEMBL) [edit | edit source]



Data sources

Mission
Values
Governance
People
Roadmap
F.A.Q.
Technology
Privacy
Data sources
Partners
Credits

For transparency, we are committed to using only **publicly available data** to generate our search results.

Here is an exhaustive list of what we currently use:

- **Common Crawl**: The largest open repository of web crawl data. This is currently our unique source of raw page data.
- **Wikidata**: A free, linked database that acts as central storage for the structured data of many Wikimedia projects like Wikipedia, Wikivoyage, Wikisource, ...
- **UT1 Blacklist**: Maintained by Fabrice Prigent from the Université Toulouse 1 Capitole, this blacklist categorizes domains and URLs into several categories, including "adult" and "phishing".
- **DMOZ**: Also known as the Open Directory Project, it is the oldest and largest web directory still alive. Though their data is not as reliable as it was in the past, we still use it as a signal and metadata source.
- **Web Data Commons Hyperlink Graphs**: Graphs of all hyperlinks from a 2012 Common Crawl archive. We are currently using their Harmonic Centrality file as a temporary ranking signal on domains. We plan to perform our own analysis of the web graph in the near future.
- **Alexa top 1M sites**: Alexa ranks websites based on a combined measure of page views and unique site users. It is known to be demographically biased. We are using it as a temporary ranking signal on domains.

Do you know about another data source that could be useful? Send us an email!

Core values

Mission
Values
Governance
People
Roadmap
F.A.Q.
Technology
Privacy
Data sources
Partners
Credits

Our core values are the DNA of Common Search. They clearly state what we stand for, what makes us different and provide guidance when making hard decisions.

Starting with the most essential:

Radical transparency. Our search results must be explainable and reproducible. All our code is open source and results are generated only using publicly available data. Transparency also extends to our governance, finances and day-to-day operations.

Independence. No single person, company or special interest must be able to influence the order of our search results to their benefit. Our board of trustees is the watchdog of that independence.

Public service. We want to build and operate a free service targeted at a large, mainstream audience. Our impact and ultimate contribution to the Web grows with our size so we should make our service accessible and useful to as many users as we can.

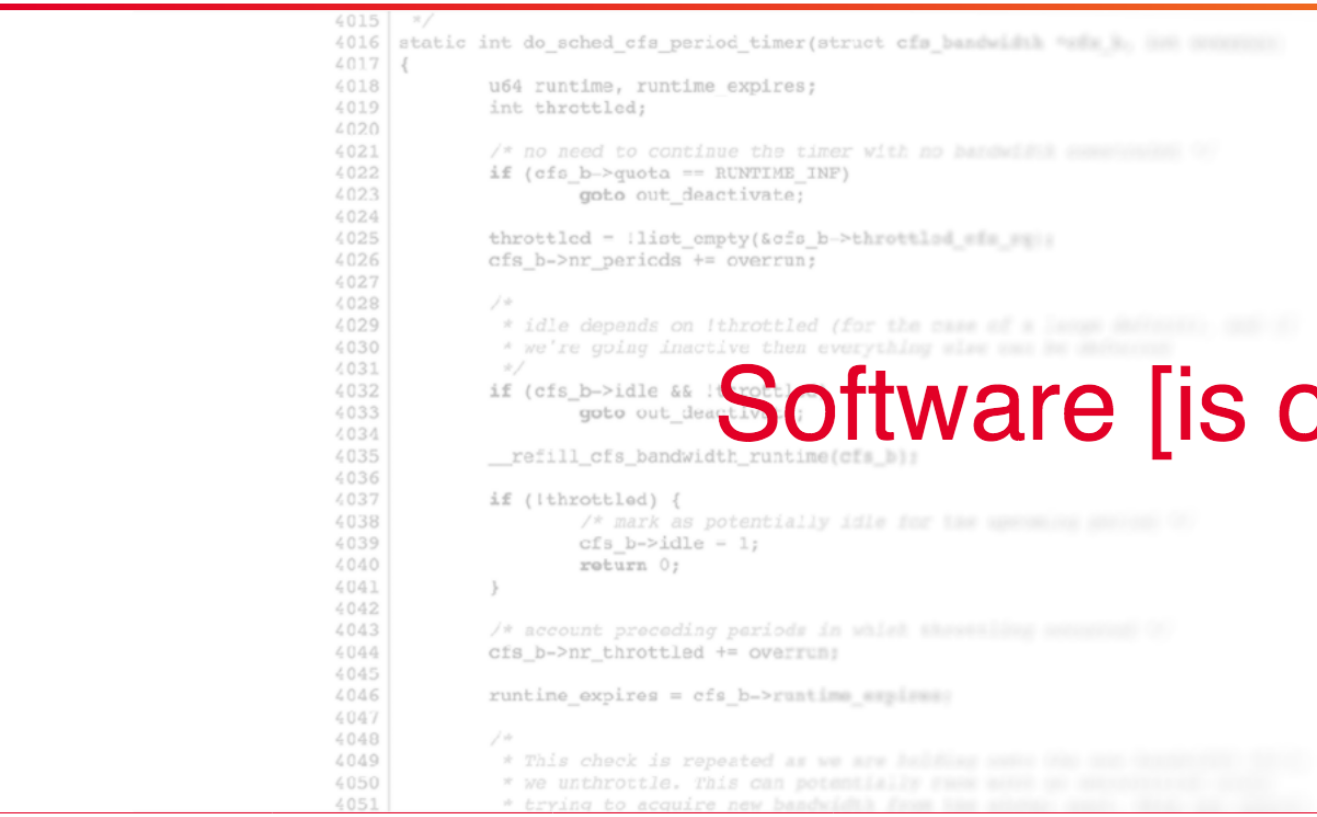
Pragmatism. Recognizing the immensity of the task at hand, we should be willing to accept short-term compromises when necessary, as long as they don't go against the values above.

Privacy. Users should be informed and in full control of the personal information they share with us and with any third parties.

Focus. We are building a search engine. Not a browser, not an operating system, not an encyclopedia. Collaboration with other organizations sharing our core values should be encouraged instead of replicating their efforts.

Frugalism. Lowering costs is easier than increasing revenue. Having a low burn rate reduces the influence of money and guarantees our long-term sustainability.

Participation. Contributing to Common Search should be easy for everyone, developer or otherwise. We must embrace and nurture an open community that will surely lead, in time, to a great search engine.



Software [is our] Heritage

We are building the universal software archive

Collect
Preserve
Share

We **collect** and **preserve** software in source code form, because software embodies our technical and scientific knowledge and humanity cannot afford the risk of losing it.

Software is a precious part of our cultural heritage. We curate and make accessible all the software we collect, because only by **sharing** it we can guarantee its preservation in the very long term.

links: con**vey** yo**ur**self to

gondwanaland.com/mlog

@mlinksva

Segregated cycle facilities of Dunhua South Road in Taipei City · Mike841125 · CC0-1.0

K∗0♡