

# **Data-Hacking with Wikimedia Projects:**

Learn by Example, Including  
Wikipedia, Wikidata, and beyond!

@notconfusing (Max Klein)  
@wrought (Matt Senate)

# Who is in the room?

- Data hackers?
- Programmers?
- Artists/designers?
- Open Access folks?
- Academics?
- Wikipedians?
- Wikimedians?
- Wiki-people?

# Wikimedia Movement

- [wikipedia.org](https://wikipedia.org)
- [wikisource.org](https://wikisource.org)
- [commons.wikimedia.org](https://commons.wikimedia.org)
- [wikidata.org](https://wikidata.org)



# Wiki Context

- Wikipedia: far-largest in size and user base.
- Projects often organized by language.
- Each language-project has an independent user community.

# Wiki Context

- See Wikimedia projects as a form of: “curated database”
- Web's *Least Common Denominator* for data.
- Wiki Paradox:
  - Low-barrier-to-entry
  - High-barrier-to-entry

Buneman, et al. *Curated Databases*.

<https://peerlibrary.org/p/rxQ6WBd89XviMF4Tk>

# How do Wikimedia projects work?

- Community
- Opt-in
- Reputation
- Cultural Protocol
- Bureaucracy
- Adhocracy
- Coordination
  - WikiProjects

# History of Wiki Data-Hacking

- Rambot
  - First “bot”
  - 2002
- US Census Data
- Created 200,000 articles
- More than doubled Wikipedia's size at the time
- No permissions

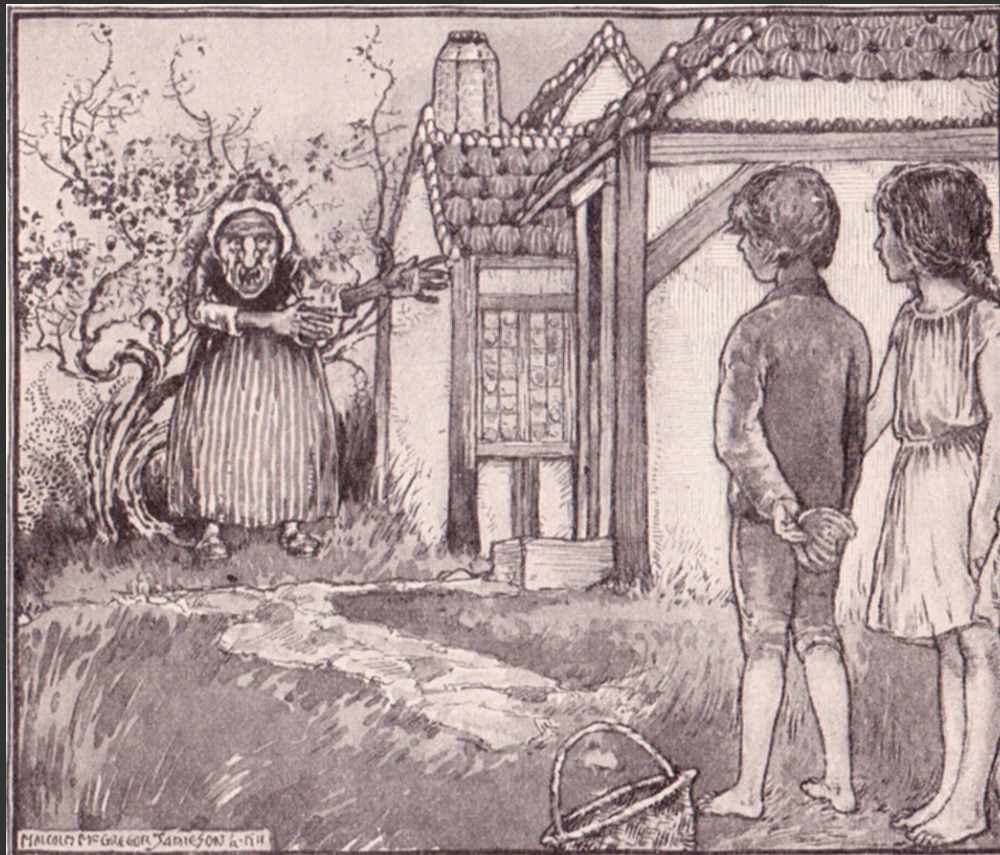
# **“Ignore All Rules” --> Calcification**

- To get things done:
  - Need to issue a “Request for Comment” or RFC
- Everybody, regardless of expertise has trouble with this.
  - Sometimes, not everyone is acting in good faith, but try to assume it, it will help.



# Häskell und Grepl

- Hänsel und Gretel
- Wikidata



[https://commons.wikimedia.org/wiki/File:Hansel\\_and\\_Gretel.jpg](https://commons.wikimedia.org/wiki/File:Hansel_and_Gretel.jpg)

- Rural Hunger Problems

- The Wikimedia datascrape having cross-language data sharing problems.



- Häskell and Grepl are sent to the Forest alone.

- Data is sent to live in Templates, living alone.

```
{{Infobox writer
|birth_name = Jacob Ludwig Carl Grimm
| image = JacobGrimm.jpg
| birth_date = {{birth date|1785|01|04|df=y}}
| birth_place = [[Hanau]], [[Landgraviate of Hesse-Kassel|Hesse-Kassel]]
| death_date = {{death date and age|1863|09|20|1785|01|04|df=y}}
| death_place = [[Berlin]], [[Kingdom of Prussia|Prussia]]
}}
```

- Häskell and Grepl first invent a successful breadcrumb system
- Wikimedia commons allows images to be shared across Wikis.

### File usage on other wikis

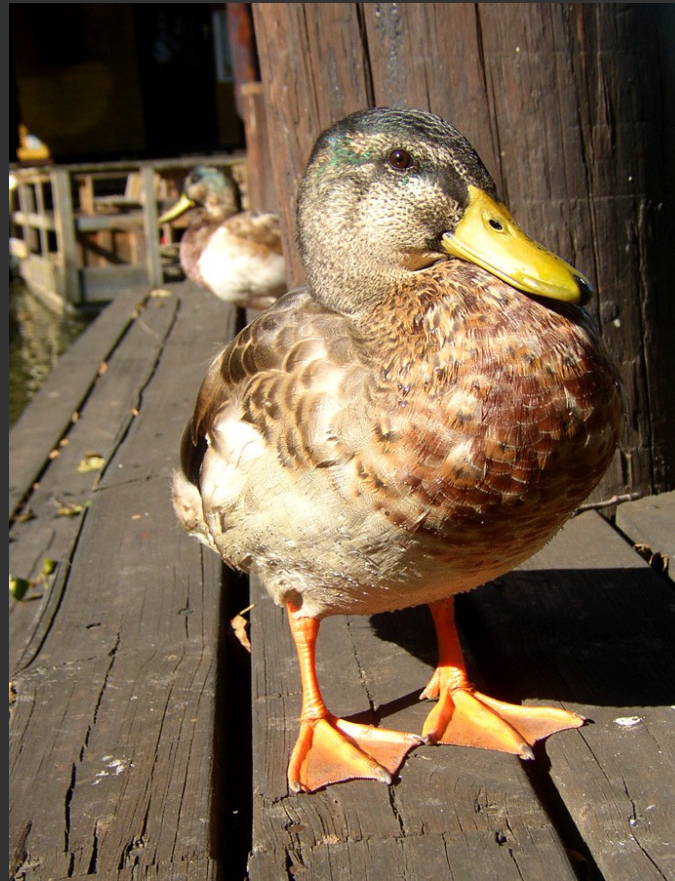
---

The following other wikis use this file:

- Usage on fi.wikipedia.org
  - [Lunar Reconnaissance Orbiter](#) 
- Usage on fr.wikipedia.org
  - [Lunar Reconnaissance Orbiter](#) 
- Usage on pt.wikipedia.org
  - [Lunar Reconnaissance Orbiter](#) 



- Lucky if their breadcrumbs are not eaten / or whatever put.
- Lucky if we knew on which pages the Data was stored.



[http://commons.wikimedia.org/wiki/File:Flickr\\_-\\_Per\\_Ola\\_Wiberg\\_-\\_mostly\\_away\\_-\\_\\_%22No\\_bread,\\_just\\_a\\_camera...huh,\\_quack\\_%22.jpg](http://commons.wikimedia.org/wiki/File:Flickr_-_Per_Ola_Wiberg_-_mostly_away_-__%22No_bread,_just_a_camera...huh,_quack_%22.jpg)

- A magical gingerbread house



- A magical data store
  - Called “Wikidata”
  - Interwiki data sharing
  - Plus with extra sweeties

[http://commons.wikimedia.org/wiki/Gingerbread\\_house#mediaviewer/File:Pepparkakshus.JPG](http://commons.wikimedia.org/wiki/Gingerbread_house#mediaviewer/File:Pepparkakshus.JPG)

- And the house includes many different sweeties.
- Semantic Triples
- Qualifiers
- Ranks



[http://commons.wikimedia.org/wiki/Category:Liquorice\\_candy#mediaviewer/File:Flickr\\_-\\_cyclonebill\\_-\\_Slik\\_%281%29.jpg](http://commons.wikimedia.org/wiki/Category:Liquorice_candy#mediaviewer/File:Flickr_-_cyclonebill_-_Slik_%281%29.jpg)

- Häskell and Grepl eat the roof hungrily.
- In this story, the User's start adding to Wikidata.
  - Importing Wikipedia
  - Foreign Database
  - Manual adds.





- The evil witch trap
- The evil data witches, normally keep the information as a silo.



- Grepl's cunning defeat of the witch



- Identifiers
  - Think of this as a foreign database key (Brian Jacobs)
  - Max imported 400,000 biographical identifiers.
  - This started an Identifier craze on Wikidata
    - Tennis Player
    - Swiss Parliament
    - Danish Companies

- Grepl's cunning defeat of the witch
- All Wikis can transclude arbitrary data (eventually).
- And Citations can be represented as semantic properties.

See how sources are handled with “FRBR” format:  
[http://www.wikidata.org/wiki/Help:Sources#Scientific.2C\\_newspaper\\_or\\_magazine\\_article](http://www.wikidata.org/wiki/Help:Sources#Scientific.2C_newspaper_or_magazine_article)

# Signalling Open Access

- WikiProject Open Access
  - On English Wikipedia
- *(Data) Problem:* Signalling “Open Access” is hard!
- *Solution:* Use clear signals directly to the relevant data.
  - Copyright license
  - Source content
  - Metadata

# What?

In a nutshell [\[edit\]](#)












This image of *Xanthichthys ringens* is sourced from an open-access scholarly article licensed for  re-use.

How can we make that reusability explicit when citing this source in Wikipedia articles?<sup>[1]</sup>  
For further details, see [this Signpost op-ed](#).

**Reference** [\[edit\]](#)

*Note the icons and links complementing the bibliographic information.*

- <sup>^</sup> Williams, J. T.; Carpenter, K. E.; Van Tassell, J. L.; Hoetjes, P.; Toller, W.; Etnoyer, P.; Smith, M. (2010). "Biodiversity Assessment of the Fishes of Saba Bank Atoll, Netherlands Antilles" . In Gratwicke, Brian. *PLoS ONE* **5** (5): e10676. doi:10.1371/journal.pone.0010676 . PMC 2873961 . PMID 20505760 .
-    full text  media  metadata

# How?

- Text → WikiSource
- Media → Commons
- Metadata → Wikidata
- Signals → Wikipedia
- RFC → RecitationBot
- RFC → RecitationBot
- RFC → RecitationBot
- RFC → RecitationBot
- Including license!
  - Public domain, CC0, CC-BY, CC-BY-SA, etc

# In the Wikimedia Universe

- Where does “Signalling Open Access” fit in the Wikimedia narrative?
- *(Data) Problem:* Managing citations & references is hard!
- *Possible Solutions:*
  - Templates (Many)
  - Categories (Many)
  - Namespace (FR)
  - WikiScholar (Dead)
  - VisualEditor (Zotero?)

# **“Signalling OA” Opportunities**

- **Take pass at citation management**
  - Quality & experience
- **Integrate metadata with Wikidata**
  - Where it belongs
  - Sensitivity and rigor
- **Forge deep knowledge resources on Wikipedia**
  - Snapshots of sources, deeper linking
- **Automate, but use human judgment**
  - Save time and energy, improve accuracy.



# Paths for Data Hackers

- Reputation
  - Create a user account
  - Contribute in good faith to wikimedia projects
- Cultural protocol
  - Identify the scope, concerns, and nature of a given project
  - Learn to navigate
- History and Context
  - Form a narrative for your hack based on past endeavors
  - Seek consent and build consensus
- Community
  - Reach out, on IRC and mailing lists!

Of course we're Open Source  
[github.com/wpoa/OA-signalling](https://github.com/wpoa/OA-signalling)  
[github.com/wpoa/recitation-bot](https://github.com/wpoa/recitation-bot)

Thanks!

@notconfusing (Max)  
@wrought (Matt)