



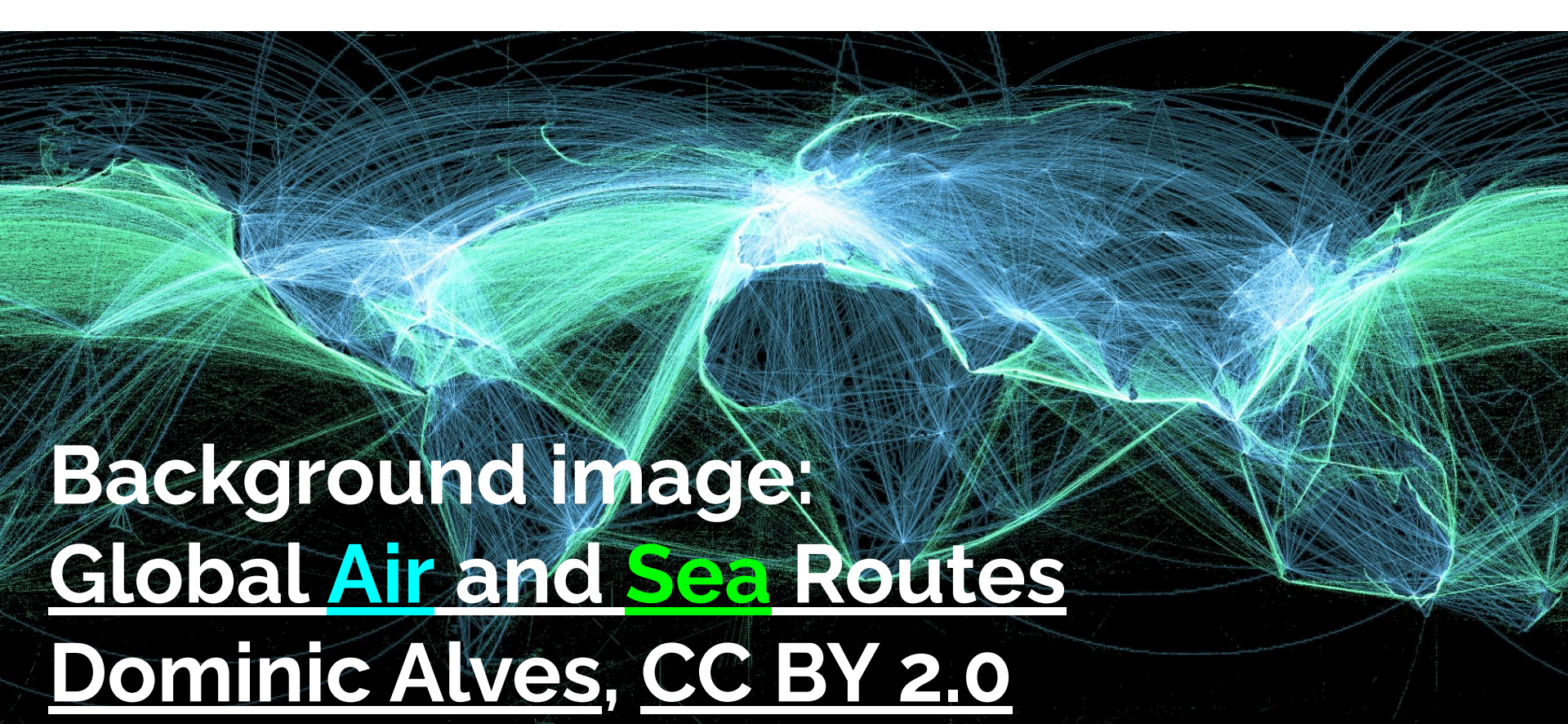
Wikidata as a data collaboration across multiple boundaries

SciDataCon 2022 session “Data Collaborations Across Boundaries”
Daniel Mietchen - 21 June 2022 - DOI: [10.5281/zenodo.6670026](https://doi.org/10.5281/zenodo.6670026)

Structure of the talk



- *Part 1: Overview of Wikidata*
- *Part 2: Quick look at the session's themes*
- *Part 3: Wikidata examples for each of the themes*
 - *boundaries to watch out for:*
 - *linguistic, national, legal, technical, disciplinary, cultural, human/ machine, data type, metadata standard, expert/ layperson, desktop/ mobile, open/ closed licensing ...*
- *Part 4: Distill points for discussion*
- *Part 5: Discussion*

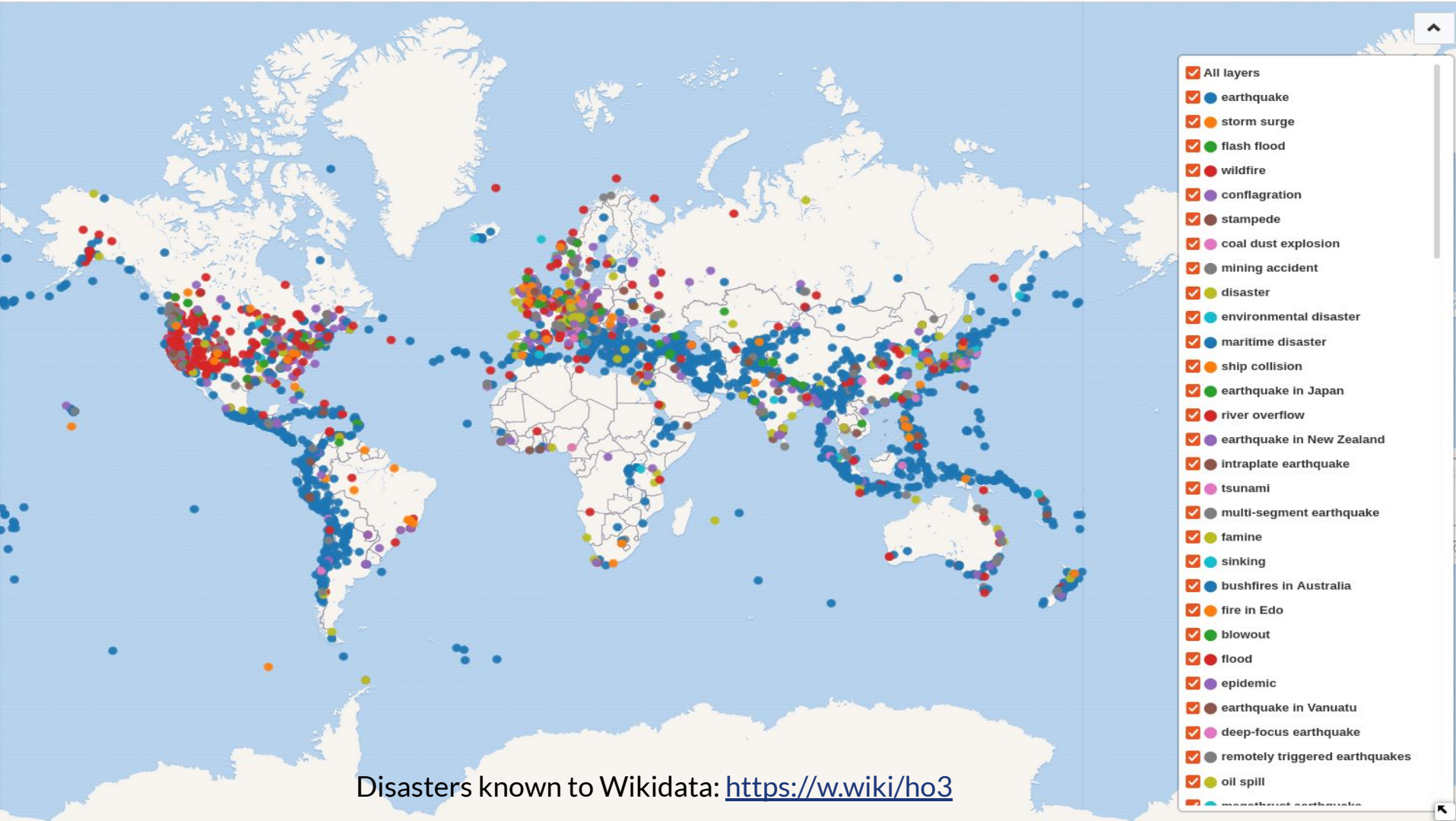


Background image:
Global Air and Sea Routes
Dominic Alves, CC BY 2.0

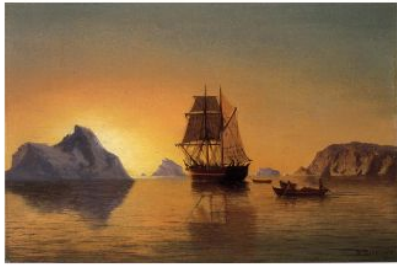
Part 1: Overview of Wikidata

- *Part of the ecosystem around Wikipedia*
- *Open data ([CC0](#)) served via open software*
- *Cross-disciplinary knowledge graph*
- *Collaborative platform*
- *Global community*
- *FAIR (meta)data*
- *Multilingual*
- ***Edit button for the semantic web***
- *20k active contributors, incl. 300 bots*
- *100M items, 10k properties, 700k lexemes*
- *10B triples, 200 queries per second, 100 edits per minute*





Disasters known to Wikidata: <https://w.wiki/ho3>



 commons:Bradford William An Arctic Scene 1881.jpg
Q wd:Q19968171



 commons:William Bradford - Morning on the Arctic Ice Fields (c.1880).jpg
Q wd:Q19968173



 commons:William Bradford - Looking out of Battle Harbor (1877).jpg
Q wd:Q19968175



 commons:William Bradford - 'Abandoned in the Arctic Ice Fields', 1876, High Museum....
Q wd:Q19968178

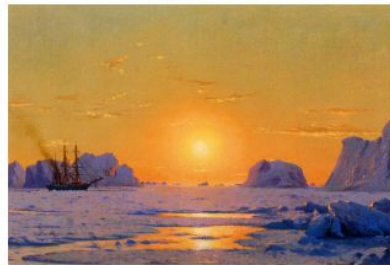


 commons:Near Cape St Johns Coast of Labrador-William Bradford 1874.jpg
Q wd:Q19968181

Paintings depicting icebergs: <https://w.wiki/8jz>



 commons:Bradford William Labrador Fishing Boats near Cape Charles 1862.jpg
Q wd:Q19968183



 commons:William J Bradford, Off the Greenland Coast under the Midnight Sun, 1873...
Q wd:Q20032042








 commons:Ward of-hull fishery.jpg
Q wd:Q20186801



 commons:South Point, Bowdoin Bay, Greenland SAAM-1950.8.27 1.jpg
Q wd:Q20504904

Resources used to study invasive species

Search:

Count ↑↓	Use	Zoom ↑↓	Use description	Example work ↑↓
102	ImageJ		image processing software	Deletion of the Plasmodium falciparum merozoite surface protein 7 gene impairs parasite invasion of erythrocytes
73	ArcGIS		geographic information system maintained by Esri	Geographic selection bias of occurrence data influences transferability of invasive Hydrilla verticillata distribution models
22	ggplot2		data visualization package for the statistical programming language R	Host Plant Use by the Invasive Halyomorpha halys (Stål) on Woody Ornamental Trees and Shrubs.
16	common garden experiment		method in ecology	Positive effects of nonnative invasive Phragmites australis on larval bullfrogs
12	RStudio		integrated development environment for the programming language R	Identifying the physical features of marina infrastructure associated with the presence of non-native species in the UK

<https://scholia.toolforge.org/topic/Q183368#uses>

Incomplete? [Contribute](#)

Part 2: Quick look at the session's themes



1. *Support and governance for data collaborations*
2. *Data policies for data collaborations*
3. *Funding sources for data collaborations*
4. *Data workflows designed for (external) data collaborations*
5. *Collaborative web platforms for data workflows*
6. *Examples and insights from data collaborations*
7. *Debating pros & cons of centralized data services.*
8. *Practical lessons learned*

Support mechanisms and governance structures for data collaborations across organizations/ communities

- [Wikidata Tours](#)
- [Data donation](#) guidelines
- [Assume good faith](#)
- Trust but verify (e.g. [ORES](#))
- [Notability criteria](#)
- Consensus culture
- Open documentation
- Various discussion fora
- Page protection
- [Weekly status updates](#)
- [Events](#)
 - [Wikidata birthday](#)
 - [WikidataCon](#)
 - [Wikidata Quality Days](#)
 - [Wikidata Training](#)
- SWJ [Special Issue on Wikidata](#)
- [Thank you button](#)
- Hundreds of [tools](#)

Support mechanisms and governance structures for data collaborations across organizations/ communities

- Point for discussion:
 - Thank you button
 - What if research infrastructures would allow their users to thank infrastructure providers or content curators for specific contributions?
 - Formal governance
 - To what extent would more formal governance structures be helpful?

Data policies for data collaborations

- Anyone can view & edit, often in the language of their choice
- Contributions are CC0
- Only CC0/public domain data can be imported
- Anything can be exported and remixed in any way
- Strong emphasis on verifiability
- Community processes for
 - coordination (e.g. [Project chat](#), [WikiProjects](#))
 - data modeling (e.g. [Property proposals](#))
 - scaling (e.g. [Bot requests](#))
 - requesting/ offering help (e.g. [with SPARQL queries](#))
 - ...

Data policies for data collaborations



- Point for discussion:
 - Tragedy of the Commons
 - How can we systemically encourage users of Wikidata's resources to “give back” eventually?

Funding sources for data collaborations

- Initial funding:
 - Allen Institute for Artificial Intelligence [AI]²
 - Gordon and Betty Moore Foundation
 - Google, Inc.
- Current funding:
 - 7M individual donors supporting the Wikimedia ecosystem
 - project-based funding, e.g. from
 - Volkswagenstiftung (example)
 - Sloan Foundation (example)
- Other funding sources:
 - U.S. National Institutes of Health (just dried up), ...

Funding sources for data collaborations



- Point for discussion:
 - Sustainable workflows
 - If project-based funding ends, the Wikidata infrastructure & data remain, and some of the community remain active.
 - contrast this with traditional research funding
 - If Wikidata is increasingly used as a research infrastructure, how can it be integrated better with classical research funding workflows?

Data workflows designed for (external) data collaborations

- Data under CC0: importing only CC0, exporting anywhere
- Software under OSI-approved licenses
- Data accessible via
 - UI (anyone can read and edit, *in their language*)
 - API (anyone can read and, many can edit)
 - SPARQL endpoint (anyone can query)
 - inc. federated queries to and from selected endpoints
 - Dumps (anyone can download)
 - including mirrors (anyone is invited to help)
- Rich use of internal and external identifiers
- Community processes for collaboration (≠ one-off dumps)

Data workflows designed for (external) data collaborations



- Points for discussion:
 - Systemic choices
 - How to decide - consistently across the research ecosystem - which data or metadata should go into Wikidata versus another Wikibase versus another database versus a dump etc.?
 - How to inform source databases systemically about relevant curation happening on Wikidata & vice versa?
 - Why are so few ontologies CC0/ public domain?

Collaborative web platforms for data workflows



- [Wikidata wiki](#)
- [API](#)
- Editing frameworks like [PyWikiBot](#), [Wikidata Integrator](#)
- [Phabricator](#)
- [PAWS](#)
- Frontends like [Scholia](#), [Ordia](#), [Reasonator](#), [Inventaire](#)
- [Tools](#) and [Games](#)
- Cross-wiki scripting via [Lua](#)
- Collaborative curation, e.g. via the [Author Disambiguator](#)

Collaborative web platforms for data workflows



- Points for discussion:
 - End-to-end integration
 - As more and more aspects of scholarly workflows are covered by proprietary silos, [Wikimedia platforms are a key element of open alternatives](#). How can we strengthen that?

Debating pros & cons of centralized data services



- Item granularity: [Bonnie & Clyde problem](#)
- Property scope, e.g. [minimal](#)/ [maximal](#) frequency of audible sound
- [Namespaces](#)
- [Wikidata vs. Wikibase](#)
- [Federated queries](#) to and from any set of SPARQL endpoints
- Hosting on shared infrastructure or not
 - e.g. [Toolforge](#) or [Wikibase.Cloud](#)

Debating pros & cons of centralized data services



- Points for discussion:
 - Control vs. community engagement
 - Operating a research infrastructure typically means having a lot of control, less of community. At Wikidata, the situation is reversed. What are the implications?

Practical lessons learned



- Multilinguality [matters](#), especially for crowdsourcing
- Avoid depending on non-open infrastructure
 - e.g. [Blazegraph](#) was bought, not developed any more
- Empower user communities, encourage experimentation
 - [Sum of all Paintings](#), [WikiCite](#)
 - Don't forget some fun elements
- Disambiguate early and often
- Archive early and often

Practical lessons learned



- Points for discussion:
 - Collective learning
 - Given the complexity of Wikidata's content/ community/ infrastructure and their environment, how can learning patterns be shared effectively?



Thank you for your attention!

You can find the slides at <https://doi.org/10.5281/zenodo.6670026>.

Next stop: Discussion

[User:Daniel Mietchen](#) / [@EvoMRI](#) / [ORCID: 0000-0001-9488-1870](#)
[Leibniz Institute of Freshwater Ecology and Inland Fisheries](#)
& [Ronin Institute for Independent Scholarship](#)
& [Institute for Globally Distributed Open Research and Education \(IGDORE\)](#)