MUSINGS ABOUT THE OPEN SCIENCE PRIZE

Authored by Melissa Haendel (/Users/Melissa-Haendel) | December 31, 2016 | 1 Comments

As I was thinking about casting my vote for the **Open Science Prize (https://www.openscienceprize.org/)**, I realized that I would in fact need a rubric for choosing. I was concerned that the public vote would tend towards popularity, familiarity, or bling, rather than the quality of the open science. But what does it mean to be “quality open science?” What should be the most important criteria? The different semi-finalist projects are all very different - on different topics, of varying degrees of maturity (some pre-date the competition and some do not), and targeting different audiences. If successful, each will have different societal impacts. I applaud them all.

Recently, we evaluated over eighty manuscripts from PLOS to determine which ones were most significant, impactful, or otherwise representative to form the core of the **new PLOS Open Data Collection (http://collections.plos.org/open-data)**. In this context, we created a rubric for evaluation and then scored each manuscript objectively. For each manuscript, what was the impact on policy change? Were ethical issues considered? Did the science advance our abilities to share data or use shared data? Did the project utilize shared data (the noble discipline of “**data parasitism (https://www.force11.org/blog/may-force11-be-research-parasites)**”)? Was the community involved? How sexy were the figures? How much did the work foster cross-pollination of ideas and approaches across disciplines? And of course, what did people think about the work? I needed a similar rubric here, but for knowledgebases and not manuscripts.

Knowledge is our collective insights, captured by experts and able to provide an explanatory framework for evaluating new observations. A knowledgebase makes that knowledge findable and computable. A recent NIH **RFI: “Metrics to Assess Value of Biomedical Digital Repositories” (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-133.html)** highlighted the ineffectiveness of current knowledgebase evaluation. Traditional citation and impact factors as a measure of success or value are inadequate. For example, almost everyone in biomedicine relies on PubMed, but almost no one ever cites or mentions it in their publications. In response to this RFI, our group (consisting of the **NCATS Data Translator (https://ncats.nih.gov/translator)** and the **Monarch Initiative (https://monarchinitiative.org/)**) developed a rubric arranged according to the commonly cited **FAIR principles (http://www.nature.com/articles/sdata201618)** -- Findable, Accessible, Interoperable, and Reusable, but with three additional principles: Traceable, Licensed, and Connected. These latter three extensions are, in my opinion, fundamental to “quality open science”, as without them, you do not have computability, legal ability to reuse the data/knowledge, and no ability to navigate the fabric of the data landscape.

Therefore, for evaluation of the open science projects, I applied the rubric we described in **our response to the RFI (https://doi.org/10.5281/zenodo.203295)**, but with additional considerations throughout relating to the PLOS data science collection curation, and trying to take into account advances since the open science prize project began (since some projects were preexisting and backed by other funds/projects, where others were brand new). I note that this is as much an evaluation of the rubric as it is of the projects themselves.

I purposefully did not watch any of the videos explaining the projects on the Open Science Prize website before performing the evaluation. I wanted to determine how well the projects themselves related their goals, content, and functionality. As a potential user of the data, I aimed to evaluate the ease of navigating the data and its access and reuse directly. Most importantly, I wanted to avoid bias where the real distinctions between projects might be obscured by video production quality, rather than highlighting each project’s genuine values and their differences. It would be all too easy to create a great video about a great idea, and then not implement a quality platform based on strong open science principles, such as open code and data access, or the FAIR+ principles: Findable, Accessible, Interoperable, and Reusable,Traceable, Licensed, and Connected.

One might ask, why bother? The first reason was I wanted to determine how well the preliminary rubric we laid out in our response to the RFI might work in the real world, as we plan to write a more thorough proposal for knowledgebase/data repository evaluation in the future. The second reason is that I simply wanted the evaluation of these projects to inform the future development of the open science projects I work most on, such as the **Monarch Initiative (https://monarchinitiative.org/)** (genotype-phenotype data aggregation across species for diagnostics and mechanism discovery), **Phenopackets (http://www.phenopackets.org)** (a new standard for exchanging computable phenotype data for any species in any context), and **OpenRIF (https://github.com/openrif)** (computable representation of scholarly outputs and contribution roles to better credit non-traditional scientists). How can we all do better and learn from the Open Science competition? In other words, such a competition shouldn’t just be about the six finalists, but rather it should inform how we all go about practicing open science in general.

So now you are probably wondering, which project(s) did I vote for? Well, that is for you to infer. As you review the musings below, consider your own values for what constitutes robust open science. Comments and corrections entirely welcome.

FINDABILITY

F1: DISCOVERABLE THROUGH VARIOUS EXTERNAL MECHANISMS (REGISTERED, DISCOVERABLE VIA SEARCH ENGINES, LINKED VIA THIRD PARTY APPLICATIONS)

F2: CONTENTS/COMPONENTS ARE WELL DOCUMENTED AND SEARCHABLE (METADATA DOCUMENTED, INDEXED, SEARCHABLE VIA MULTIPLE MECHANISMS, CONTACTABLE)

While discoverability should be via a variety of mechanisms, here I Googled each project to see if I could easily find them (and their codebases) as a simple measure of discoverability. I could have included text mentions in Google Scholar, for example, but did not as many of these projects are new and therefore this may not be a good indicator; similarly, registration within data indices such as **Re3Data (http://www.re3data.org/)** or **BioCaddie (https://biocaddie.org/)**. I also evaluated whether I could determine what the content of the projects were about, and how easy they were to search.

Fruitfly Brain observatory:

Discoverability: The top hit was the project website, the second one the open science prize, the third a preprint, and the fourth the website Github repository.

Content description: The homepage has a very nice summary of content:

> "..stores and processes data related to the neural circuits of the fly brain including location, morphology, connectivity and biophysical properties of every neuron, seamlessly integrates the structural and genetic data from multiple sources that can be queried, visualized and interpreted, automatically generates models of the fly brain that can be simulated efficiently using multiple Graphics Processing Units (GPUs) to help elucidate the mechanisms of human neurological disorders and identify drug targets."

Documentation: Described on homepage, with links to more detail and the Github sites:

Supporting the NeuroNLP, NeuroGFX and NeuroAPPs is a highly sophisticated software architecture. The two key components at the back-end are: a database, called NeuroArch a brain emulation platform, called Neurokernel.

The relationships between the different components were confusing, probably an architecture/component/process diagram would be helpful.

Searchability: I found the home page extremely complicated and couldn't figure out how to get started. I finally clicked on NeuroNLP. Once I got there, there were helpful tutorials about the sorts of things I could search, and beautiful imagery of the neurons in the fly brains.

I also tried NeuroGFX. I couldn't figure out what to do here, nothing I clicked on seemed to do anything. There was a tutorial video, but I didn't watch it as I prefer inline documentation or example queries/operations (and in the interest of time for these evaluations, as well as above video considerations).

Open Neuroimaging Laboratory:

Discoverability: The top hit was a RIO publication of the Open Science Proposal (I liked that - I believe it is the only one of the six to make their proposal available), the second the Open Science prize site, the third the project website. The Github site was on the second page of results, after a number of other open content sites.

Content description: The homepage states the following:

> "The Open Neuroimaging Laboratory is a collaborative platform that facilitates finding, improving, and reusing the massive amount of neuroimaging data available on the Web. This data represents an enormous funding effort and the time and goodwill of thousands of participants: researchers, clinicians, patients, and their families. We want to maximize its use in brain research, medicine, education, and to advance open science.
>
> Our prototype applications transform these static data into "living" matter for collaborative curation and analysis using only a Web browser. No data are downloaded, no software installed. Users can find data, collaborate online, and incrementally improve each other's work. This increases scientific efficiency, improves public data quality, and reduces redundant effort. Our platform lowers the barriers for researchers, students, and citizen scientists to fuel scientific discovery."

While these goals are to be applauded, the text doesn't really tell me exactly what to expect with regard to the specific data types and formats I might find there, or the kinds of operations I might be able to perform on the data. Further down, we read about the tools that have been implemented:

BrainBox – a Web app to work with shared brain imaging data directly online. Progress becomes incremental: each researcher's work improves that of the whole community. The community can then tackle projects that would be impossible for individual research groups. BrainBox enables a growing catalogue of all MRI data available online, currently indexing more than 6000 links.

MetaSearch, another Web app, complements BrainBox by providing access to additional metadata about participants (e.g., age, sex, IQ), but also other imaging information (e.g., functional MRI, diffusion fMRI). It harmonizes the disparate sources of information and aggregates these metadata in one place, provides an interactive interface to the user to filter this rich repository of public data, and find data of relevance.

Searchability: I used the metadata search, which had a nice inline tutorial built in, making it easy to understand the different tabs and operations. I also opened the monkey brain and some other brains in Brainbox. The monkey brain was much harder, I found an annotation "Roberto for prez," which while humorous for demo purposes, didn't get me much further in understanding how to view, edit, or download the annotations.

MyGene2: Accelerating Gene Discovery with Radically Open Data Sharing:

Discoverability: The top hit was the project website, followed by Facebook, and the Open Science website. I did not find the project's codebase in the top five pages of results, nor linked from the project website.

Content description: The homepage doesn't really have a statement up front about what the site does, but lower down it states the following:

> Benefits of sharing on MyGene2:
>
> Find a genetic diagnosis for your family
>
> Help other families find a genetic diagnosis
>
> Match with other families with the same condition
>
> Contact other families with the same condition
>
> Share health and genetic data (variants of unknown significance, candidate genes, VCF files, etc.) directly with other families, clinicians, researchers
>
> Free reanalysis of shared exome data
>
> Receive matches to your profile automatically
>
> Directly engage doctors and researchers as a citizen scientist

From this summary, one can surmise that this is a patient data sharing platform and there is a link to the families that have shared information. There isn't a description of what data or tools to operate on it are available, though.

Searchability: There is a clear search box on the homepage, front and center, where you can search for a gene, family, or clinical finding (such as a disease or phenotype). Results show candidate variants in families where they exist, or suggest to the user to login and submit new ones where they do not.

OpenAQ:

Discoverability: The top hit was the project website, the second Medium, and the third their Github codebase.

Content description: The homepage is quite clear in its goal:

> "...aggregates and shares open air quality data from around the world"

And the data description:

Our community has collected 33,529,796 air quality measurements from 4,694 locations in 42 countries. Data are aggregated from 56 government level and research-grade sources.

A "where does the data come from" **link (https://medium.com/@openaq/where-does-openaq-data-come-from-a5cf9f3a5c85#.84ah4zkhf)** takes you to a high quality description of the data and how it is collected and aggregated, with examples.

Searchability: There wasn't an obvious way to get started and/or search, but there were obvious filtering mechanisms to help navigate the data in both the map tab and the data/locations and data/countries tab. After that, it was easy to get to see the data and to compare contaminants in different parts of the world over time.

Nextstrain: Real-Time Evolutionary Tracking for Pathogen Surveillance and Epidemiological Investigation:

Discoverability: The top hit was the project website, the third and fourth the project Github.

Content description: The landing page redirects to a zika-specific application, and states: "Real-time tracking of Zika virus evolution", but there is no help or about page, and no other description.

Searchability: Searching for strains using autocomplete was easy. The filtering options were nice and display visually appealing. There was not much explanation though, of what the axes or components meant to help a naive user contextualize their results/display.

OpenTrialsFDA:

Discoverability: The Open Science prize site was the first hit, then the project website. The Github site was on the second page of results.

Content description: On the landing page, it states:

> "Find FDA Documents by name, text and other keywords"

Which makes it clear that one can find FDA documents, but not really what those might be. In the USA, we all know what the FDA is; in other parts of the world, perhaps not as much. In the about page, it states:

> "OpenTrialsFDA works on making clinical trial data from the FDA (the US Food and Drug Administration) more easily accessible and searchable. Until now, this information has been hidden in the user-unfriendly Drug Approval Packages that the FDA publishes via its dataportal **Drugs@FDA (mailto:Drugs@FDA)**. OpenTrialsFDA extracts the relevant data from the FDA documents, link it to other clinical trial data and present it through this new user-friendly web interface. Any user can type in a drug name, and see all the places where this drug is mentioned in a FDA document. This information is also available via the OpenTrials API."

This is much more descriptive and makes it clear what kind of data is available and how one might operate on it. I also like that the API is indicated on the About page.

Searchability: The Explore tab has a very clear and configurable search interface, allowing search over a variety of fields: Trial title, identifier, keyword, condition, intervention, sample size, gender, registration dates, publications, and more.

ACCESSIBILITY

A1: DIVERSE DATA ACCESS MECHANISMS (DUMPS, QUERY DOWNLOADS)

A2: WELL STRUCTURED AND PROVISIONED APIS

A3: UNDERSTANDABLE DATA AND SCOPE (AUDIENCE, CONTENT, BROWSING, DOCUMENTATION, TUTORIALS)

I was surprised at the lack of attention that seemed to be paid to data access in the various projects. To me, open data access is a salient feature of open science in general. Also, in retrospect, I think that the A3 content could be merged with the Findability/content evaluation above and so I don't discuss again here.

Fruitfly Brain observatory:

NeuroKernal seems to have API access **here (http://neurokernel.readthedocs.io/en/latest/reference_model.html)**, but I could not find much else in terms of ability to grab data. A tutorial video was available (see above, I didn't watch any videos in the context of this evaluation).

Open Neuroimaging Laboratory:

You can download annotations from any given image, but I could not find any API or other bulk download options. Very nice in-line tutorial information.

MyGene2:

I could not find information on computational or other data access. There was a demo version available in one of the tabs.

OpenAQ:

There is a **well documented API (https://docs.openaq.org/)** linked from the homepage, with data types such as cities and contaminant type, dates, etc. clearly indicated. Data downloads are also readily available in the different pages. I didn't really find a site tutorial, but there is a lot of good documentation on the data content and its processing in the Github site.

Nextstrain:

I could not find information on computational data access or other download options from the website or Github sites, or a tutorial for use of the platform.

OpenTrialsFDA:

There is both a data search interface and a reasonably well **documented API (https://api.opentrials.net/v1/docs/)** with data entity types: trials, publications, conditions, organisations, persons, interventions, sources, fda_applications, and documents. I did not immediately see study design as a searchable entity type, which i might have expected. No tutorial, but there are query examples in the search interface.

INTEROPERABILITY

I1: IDENTIFIERS (CREDIT, DESIGN, DOCUMENTATION, ETC.)

I2: VOCABULARIES, ONTOLOGIES, AND EXCHANGE STANDARDS (SEMANTICS/DATA STRUCTURE)

I3: VERSIONING

I was also surprised at the lack of efforts towards interoperability in many cases, though these are largely new projects and so such things may not yet have been implemented. In some cases, they may not apply (e.g. no standards exist or are needed).

Fruitfly Brain observatory:

The documentation for the underlying database is **here (http://neurokernel.github.io/rfc/nk-rfc5.pdf)**, and provides in depth requirements for data types that require representation. However, I did not see any indication of use of any persistent identifiers or vocabularies, even for things as simple as the taxon ID. They do suggest in the paper however, that: "NeuroArch should support loading data from and saving data to several specification formats such as SWC9, CSV, GEXF, NeuroML, NineML, or SpineML…"

However, perhaps this has not yet been done? I could not find any evidence of use of other standards such as the fly anatomy ontology or disease vocabularies. I tried the NLP queries but it was unclear what dictionaries/ontologies it might be using, and so was hard to tell if it was missing things or including too much.

I didn't find any version information for the site, the data, or any components.

Open Neuroimaging Laboratory:

The Github site states:

> "MetaSearch accomplishes this by **extracting (https://github.com/OpenNeuroLab/metasearch/blob/master/crawler/extract/Extract.ipynb)** metadata for these projects from the AWS cloud, **transforming (https://github.com/OpenNeuroLab/metasearch/blob/master/crawler/extract/Transform.ipynb)** it into a common data model, and the **loading (https://github.com/OpenNeuroLab/metasearch/blob/master/crawler/extract/Load.ipynb)** the integrated dataset into the MetaSearch app."

But the links to the extracting and transforming are not public so one cannot tell what modeling has been done and what identifiers/standards are used in that context. I did find some pdf data dictionary files **here (https://github.com/OpenNeuroLab/metasearch/tree/master/docs/data_dictionaries)**, where there are clearly some standard clinical instruments that are used, but no other standards that I could find.

MyGene2:

Uses the **Human Phenotype Ontology (http://human-phenotype-ontology.github.io/)** (HPO) for phenotypic features. Uses the **Matchmaker Exchange (http://www.matchmakerexchange.org/)** for comparing with other families in other platforms. I did not find any version information on the platform, data, or components such as the HPO.

OpenAQ:

Uses ISO date/time standard, lat/long location standard, ISO country codes. Data is versioned daily. Data format information is **here (https://github.com/openaq/openaq-data-format)**.

Nextstrain:

The main website states that the data was last updated 20 Dec 2016. FASTA, Entrez are used for sequence data. I didn't see reference to other standards.

OpenTrialsFDA:

Uses ISO date/time standard. Identifiers for registration in EU or other clinical trials registries are provided for each entry. I didn't see any standards used for drug ids/synonyms, conditions, or other interventions, nor version information, though based on search behavior, I believe some of these are in use.

REUSE

R1: USE (EVIDENCE THAT RESOURCE IS IN DEMAND)

R2: IMPACT (UNIQUENESS, EMBEDDEDNESS, PUBLICATIONS/CASE STUDIES, TESTIMONIALS)

R3: AWARENESS AND RESPONSIVENESS TO KEY USER NEEDS (TRACKERS, ADVISORS, CONTACT INFO, RESPONSIVENESS, RELEASE NOTES)

R4: QUALITY OF DATA CONTENT AND SERVICE (CURRENCY, UPTIME, COMPREHENSIVENESS)

Fruitfly Brain observatory:

I didn't find any issues in the Github site, which was just the website and not the platform. Workshops and hackathons are being scheduled **here (http://fruitflybrain.org/workshops.html)**. I could not find any information about the relationship with the seemingly very similar **Virtual Fly Brain (http://www.virtualflybrain.org/)** project, or with **FlyBase (http://flybase.org/)** for nomenclature and interoperability purposes.

Open Neuroimaging Laboratory:

Active Github (https://github.com/openneurolab) presence linked from homepage, also **issue tracker. (https://github.com/OpenNeuroLab/BrainBox/issues)** Feedback is also provided within the community forum Neurostars. It seems quite clear that there is a need to provide tools to use crowdsourcing to provide annotations on brain imaging data, and to be able to search across these annotations. However, the relationship to projects/platforms such as **NITRC (https://www.nitrc.org/)** and others wasn't clear.

MyGene2:

The clear goal is for very rare disease patients to contribute their genotype-phenotype data to help identify other families with the same disease and/or get a diagnosis. A nice guide for patients is **here (https://www.mygene2.org/MyGene2/getstarted)**. I didn't find any issue tracker or way for the community to contribute to the project, but there is a contact form available. The relationship with other patient-centered rare disease efforts, such as **Orphanet (http://www.orpha.net/)** or **IRDIRC (http://www.irdirc.org/)** wasn't described.

OpenAQ:

Has very helpful contribution guidelines **here (https://github.com/openaq/openaq-fetch/blob/master/CONTRIBUTING.md)**, and an active blog and twitter feed. Responsive **issue tracker (https://github.com/openaq/openaq-fetch/issues)**. Community QA on Reddit. Conducted very nice **requirements analysis (https://medium.com/@openaq/a-report-what-does-the-open-air-quality-community-need-in-order-to-be-awesome-73586bf6f45#.w2xzzr8ou)** of diverse users. Active Slack channel. Active and engaged contributor/user community.

Nextstrain:

The website doesn't provide much information about the project's purpose, for this you have to go to the Open Science Priz website. However, there is an emerging Github community; trackers are available **here (https://github.com/nextstrain)**.

OpenTrialsFDA:

Seems to have a lot of community engagement and contribution events indicated **here (http://opentrials.net/press-resources/)** and a lot of ways for different types of people to contribute indicated **here (http://opentrials.net/contribute/)**. Active **issue tracker (https://github.com/opentrials/opentrials/issues)**. The relationship with ClinicalTrials.gov is not obvious.

TRACEABILITY

T1: PROVENANCE (DATA'S PROVENANCE IS WELL DOCUMENTED AND ATTRIBUTED)

T2: ATTRIBUTION (CONTRIBUTIONS TO THE CONTENT (DATA, TOOLS, ALGORITHMS, SOURCES, ETC.) ARE CLEARLY DECLARED)

Fruitfly Brain observatory:

There is an acknowledgments and credits page **here (http://fruitflybrain.org/licenses.html)** that describes general contributions to the codebase, and citation of the three data sources that are used in the platform. The data in the platform doesn't seem to be individually attributed/contain provenance.

Open Neuroimaging Laboratory:

There is a "who we are" attribution on the About page **here (http://openneu.ro/index.html#four)**. Data sources are linked to source, for example "**http://files.figshare.com/2284784/MRI_n4.nii.gz (http://files.figshare.com/2284784/MRI_n4.nii.gz)**" but oddly don't leverage the persistent ID/DOI provided by Figshare.

MyGene2:

The **About (https://mygene2.org/MyGene2/about)** page lists contributors and the attributes different roles/teams, as well as a general listing of external resources (some are links, some are citations, but not how they are used).

OpenAQ:

The source of each data item is directly linked. Partners and sponsors are attributed on the About page. I did not find mention of team members/contributors on the site (though one can see these in the Github repos).

Nextstrain:

Code and data contributions (by citation) are acknowledged on the **homepage (http://nextstrain.org/zika/)**. Individual data elements also link to the contributing source.

OpenTrialsFDA:

Sponsors are listed at the bottom of the page. Sponsor, publications, and acknowledgements are listed for each entry. External sources are listed **here (https://explorer.opentrials.net/terms-of-use)**, but not how they are used. Contributors to the codebase are not really listed other than the leads.

LICENSURE

L1: DOCUMENTED, CLEAR, STANDARD, MINIMALLY RESTRICTIVE, CONTACTABLE

L2: TRANSPARENT ABOUT FLOWTHROUGH IMPLICATIONS

Data licensing is one of the most important open science principles, as it is one of my biggest time sinks in trying to reuse data.

Fruitfly Brain observatory:

The code and data licensing is information is available **here (http://fruitflybrain.org/licenses.html)**. Code is BSD, Documentation and Website are CC-BY 4.0. However, the code links only to the homepage/website, and I could not find any open code for any of the signfiicant components, such as NeuroNLP or NeuroGFX.

Data is more complicated, with some citations being listed (with no license/use agreement) and one source as follows:

This data cannot be used for biological discovery or biological publications before Fly EM publishes it. For other purposes (general bio data analysis algorithms, etc), this data is made available under the Open Data Commons Attribution License: **http://opendatacommons.org/licenses/by/1.0/ (http://opendatacommons.org/licenses/by/summary/)**. A simple description of the license is provided **here (http://opendatacommons.org/licenses/by/summary/)**.

Open Neuroimaging Laboratory:

I was able to find the data licenses after digging around in the **Github site (https://github.com/OpenNeuroLab/metasearch)**:

> Please refer to and follow the data licenses and use agreements listed on the homepage of each of the datasets in the table below.
>
> **Project Link to License**
> **Brain Genomics Superstruct Project**
> **(https://dataverse.harvard.edu/dataset.xhtml?**
> **persistentId=doi:10.7910/DVN/25833)**
> **IXI – Information eXtraction from Images**
> **(http://brain-development.org/ixi-dataset/)**

For the former, I had to provide my credentials in order to download the data use agreement PDF. The PDF does not provide any specific license, but does state "I may redistribute original GSP Open Access data and any derived data as long as the data are redistributed under these same Data Use Terms."

For the latter, there is a quality description stating that there is a **CC BY-SA 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/legalcode)**. If you use the data please acknowledge the source of the data, e.g. this website.

The code uses an Apache license **here (https://github.com/OpenNeuroLab/BrainBox/blob/master/LICENSE)**.

MyGene2:

Terms of use are **here (http://www.washington.edu/online/terms/)**, but this links to a generic University of Washington website terms of use and does not mention MyGene2. On the disclaimer link, it states (with regard to contributors):

> Ownership of content: You own all of the content (photos and text) and information you post on MyGene2. By sharing content on MyGene2, you specifically grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use it. This License ends when you delete your content or your account, unless your content has been shared with others, and they have not deleted it.

I could not find any further information on either the code or the data license/use agreements.

OpenAQ:

Code: Clear BSD license information is **here (https://github.com/openaq/openaq-design-system/blob/master/LICENSE)**, describing reuse considerations.

Data: The data is described **here (https://medium.com/@openaq/where-does-openaq-data-come-from-a5cf9f3a5c85#.kr7mbq5x2)**. There does not seem to be any data use agreement, but the data that is aggregated in real time is public data to begin with. There is a CC0 license file **here (https://github.com/openaq/awesome-air-quality/blob/master/LICENSE)**, I believe it applies to all the data, but this was difficult to determine.

Nextstrain:

It says directly on the homepage:

Code: All source code is freely available under the terms of the GNU Affero General Public License. Data updated 20 Dec 2016.

Data: There are multiple statements depending on which data.

For data reuse (particularly for publication), please contact the original authors (with many listed following)…. then:

A disclaimer from the ZiBRA team: "Please note that this data is still based on work in progress and should be considered preliminary. If you intend to include any of these data in publications, please let us know – otherwise please feel free to download and use without restrictions. We have shared this data with the hope that people will download and use it, as well as scrutinize it so we can improve our methods and analyses. Please contact us if you have any questions or comments. For ZiBRA sequence data email us at **n.j.loman@bham.ac.uk (mailto:n.j.loman@bham.ac.uk)**."

In all, unrestricted reuse of the data is not that clear, as it would require contacting one of the many the primary creators/contributors to the resource, and redistribution (flowthrough) seems ok with the ZiBRA data but not the rest?

OpenTrialsFDA:

Terms of use are listed **here (https://explorer.opentrials.net/terms-of-use)**.

Code: The website says the code is open under a free license on Github, but I could not find any licenses. The website says CC0 at the bottom.

Data: The data license is strong:

You are free to use the data published on this site for any purpose, subject to the following conditions:

- Content created by third parties is offered to you under their terms. Please see the full list of sources and links to their licensing terms, where available, in the External Sources (https://explorer.opentrials.net/terms-of-use#tou-sources) section on this page. You should refer to the original source if you have questions about your rights to use this content.
- Content we create, or which contributors submit directly to us, is offered under the Creative Commons Public Domain Dedication (https://creativecommons.org/publicdomain/zero/1.0/) (usually called "CC0"). However, we encourage you to give appropriate credit to OpenTrials and OpenTrials contributors. For more details, see the full text of the license (https://creativecommons.org/publicdomain/zero/1.0/).

CONNECTEDNESS

C1: CONNECTEDNESS

I did not find much evidence of links to other resources in most of the resources. Perhaps in part this is because some of them are new, for others there may not be relevant resources to link to.

Fruitfly Brain observatory: I didn't find any link outs.

Open Neuroimaging Laboratory: I didn't find any link outs.

MyGene2: Links to ClinVar, OMIM, GeneCards, and the Monarch Initiative (via Human Phenotype Ontology).

OpenAQ: Links go back to the source providers.

Nextstrain: I didn't find any link outs.

OpenTrialsFDA: I didn't find any link outs.

DISCLAIMERS:

Dr. Erick Turner (member of the OpenFDA Trials project) is employed at Oregon Veteran's Hospital and the Oregon Health & Science University, where I work. Dr. Turner has guest lectured in my ethics class.

Dr. Tudor Groza (member of the MyGene2 project) is a site PI for the Monarch Initiative, a project that I co-lead. MyGene2 leverages the Human Phenotype Ontology, which Monarch co-develops.

I have tried to the best of my abilities to be objective in my evaluation, and have not been directly involved in any of the projects.

(/users/melissa-haendel)

About **Melissa Haendel**

(/users/melissa-haendel)

Melissa is an active participant in the Force11 community; she is on the executive board, chaired the Force2016 conference, was the program chair for Force2015, and currently co-leads the Attribution WG. Melissa also co-leads the Monarch Initiative, which aims to provide open integrated access to model organism and human phenotype-... **More (/users/melissa-haendel)**

View Profile (/users/melissa-haendel)

1 Comments | Melissa Haendel's Blog (/Blogs/Melissa-Haendel) | Sign In Or Join Now! To Post Comments

Post To Twitter (Http://Twitter.Com/?Status=Https%3A//Www.Force11.Org/Blog/Musings-About-Open-Science-Prize%20Musings%20about%20the%20Open%20Science%20Prize%20)

COMMENTS

Posted by Christa Hasenkopf (/Users/Christa-Hasenkopf) | January 4, 2017

Thanks! (/comment/541#comment-541)

Hi Melissa -

I really want to say THANKS for doing this analysis on the OpenScience Finalists (full disclosure, I'm with OpenAQ). It's exactly the kind of user feedback - and really a general, systematic framework of thought - that is SO valuable to groups like ours starting out, yet often very difficult to find.

If it's of interest to you or anyone reading this piece - for our specific project, we've just put together a list of FAQs (only on GitHub, not yet linked to our mainsite) that may help address some of the gaps above. For anyone reading, we'd love to hear what other Q's are missing or gaps you may feel remain.

https://github.com/openaq/openaq-info/blob/master/FAQ.md (https://github.com/openaq/openaq-info/blob/master/FAQ.md)

Thanks again.

Sign In or Join Now! to post comments

ABOUT FORCE11 (/ABOUT)

Overview (/about)

Manifesto (/about/manifesto)

Guiding Principles (/about/mission-and-guiding-principles)

Endorsement Policy (/about/endorsement-policy)

Sustainability Plan (/about/support)

COMMUNITY (/COMMUNITY)

Blog (/blogs)

Members Directory (/community/members-directory)

GROUPS (/GROUPS)

Groups - Active (/groups)

Groups - Completed (/groups/completed)

Start a Group (/groups)

NEWS + EVENTS (/EVENTS)

News (/community-news)

Upcoming Events (/events)

Event Calendar (/calendar/month)

PROMO MATERIALS (/MEDIA/VIDEOS)

Photos & Videos (/media)

CONTACT US (/CONTACT)

FORCE11
La Jolla, CA 92093

Email Us (/contact)