

The Ultimate Guide to Open Data, with Examples and Explanations

Open data is crucial to business, government, and social enterprises - this idea isn't new. What has changed in recent years is how open data has become part of most people's everyday lives.

Open data is no longer a niche topic but essential, for example, to completing journeys or keeping us safe from Covid 19. But it's not only tech companies and social enterprises who are pushing boundaries with open data, law-makers are also starting to clock on to its future potential.

Plans recently published by the EU set forth ambitions to capitalize on open data's economic potential by 2025:

- The open data market size will grow from a value of €184.45 billion in 2019 to between €199.51 - €334.20 billion.
- Between 1.12 – 1.97 million people will be employed in open data jobs.
- Open data can help save the environment, e.g., reducing household energy consumption by the equivalent 5.8 Million tonnes of oil. [Reference]

So why is this important to you?

Got a smartphone? If the answer is yes, then chances are you're benefiting from open data, or equally likely you are creating open data. As the statistics above show, open data is increasingly important and will play a significant role in our lives. So it makes sense to know a bit about what is going on.

In this article, we'll start with a definition of open data, then look at the concepts (with examples) that make up open data and point you in the direction of places where you can access and download open data.

If you are a seasoned open data user, the last section is for you as we feature repositories and resources to get your teeth into.

We hope you find the prepared content helpful!

Contents

The Ultimate Guide to Open Data, with Examples and Explanations	1
Open Data definition.....	2
Open data: key concepts defined.....	2
The fundamental concepts of open data broken-down with examples and explanations ..	4
What does "Availability and Access" of open data mean?	4
What does "Re-use and Redistribution" mean?	5
The two main Open License types	5
The three levels of openness	6
Why are there different levels of openness?	7
License and level of openness combined	8
What does "Universal Participation" of open data mean?.....	8
What do "Data Interoperability" and "Technical Openness" mean?	9
Technical Openness	10
Open data repositories and sources	11
Open Data resources and impact case studies	17

Open Data definition

The term open data, it is claimed, first appeared in 1995 in an American scientific agency document relating to cross-border geophysical and environmental data.

Fast forward to today, and the Open Data Handbook's definition of open data is widely used:

"Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike." [Reference 1]

If you are new to the concept of open data, this definition may not mean that much to you. Even when you know your NonDerivatives from your NonCommercial open data, we find it useful to unpack this definition with real-world examples.

Open data: key concepts defined

You'll notice that the open data definition uses the terms:

- freely used,
- re-used and
- redistributed by anyone, and
- attribute and sharealike.

Each of these terms represents an open data concept. For data to be "open" it needs to adhere to these concepts. Each one helps us understand what makes data open.

Let's take a look at concepts provided by The Open Data Handbook:

1. **Availability and Access** - the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
2. **Re-use and Redistribution** - the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
3. **Universal Participation** - everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.
4. **Interoperability** - denotes the ability of diverse systems and organizations to work together (inter-operate). [Reference]

Notice how each concept broadly overlaps with the open data definition terms. For example, the "Availability and Access" relates to "freely used" as it talks about cost, or how "Re-Use and Redistribution" links to how the data can be "intermixed" when it is re-used.

The concepts are summaries of an even more detailed definition of "openness" from the Open Knowledge Foundation. We highly recommend this resource, as it is complete and thorough.

The fundamental concepts of open data broken-down with examples and explanations

Here, we break down the Open Data Handbook key concepts line-by-line and offer real-world examples to illustrate what that looks like.

What does "Availability and Access" of open data mean?

To recap, the Open Data Handbook states:

"The data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form." [reference]

- "The data must be available as a whole"

As it sounds, all the data needs to be there for all to see. This concept is easier to understand in reverse, i.e., when all the data isn't there.

For example, imagine a website that gives out company ratings but does not show the data sources used to calculate the rating. We can view the final ratings, but we can't view the original data to see how the ratings were reached. This doesn't mean that the ratings are wrong; however, it is not open data, as the "whole" data set is not viewable.

- "and at no more than a reasonable reproduction cost,"

Typically, open data is free, for example, like WikiRate's. However, there are some exceptions: The Open Data Barometer - an analysis of 115 countries and jurisdictions' open data offerings - found that "10% of all datasets surveyed were not available free of charge."

The logic behind payment is that you pay a reasonable cost to cover expenses incurred when making the open data available. For example, the price may include; paying a developer, hiring a lawyer to ensure all laws are followed or paying to maintain the system. This [blog from GovEx](#) gives a breakdown of the different areas that may require funding and good examples.

- "preferably by downloading over the internet."

Making open data available for download on the internet may seem so obvious as barely worth mentioning. However, when you consider approximately 35% of the world's population, or 2.76 billion people, are living without the internet, it's important

to remember that not all open data is online. Indeed, two young Indonesians prove that offline open data can have a crucial and positive impact.

- "The data must also be available in a convenient and modifiable form."

Two examples of data formats accepted as both "convenient and modifiable" are CSV and JSON formats. We can consider them convenient as the file types work well with various programs and compress data to sizes that make them easier to transfer. Likewise, these file formats permit users to modify the data using multiple computer applications and programs.

If the data is a photo, then these terms may refer to the available resolution sizes and file types that work well with computer software. A good example here is Flickr Creative Commons content.

What does "Re-use and Redistribution" mean?

To recap, the Open Data Handbook states:

"the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets." [reference]

Essentially, "terms that permit" means the kind of license the data has: the type of data license determines, or permits, how you can and can't use it.

The power and utility of open data rests on the concept that people are free to use data without breaking the law, such as copyright. So the rules, or 'license' attached to the data are critical.

Licensing of data is a vast and complex topic, which we'll only touch on here. But these are two valuable resources about licensing for (re)users and publishers of open data:

- Reuser's Guide to Open Data Licensing
- Publisher's Guide to Open Data Licensing

The two main Open License types

Let's take a look at the two main licenses for different content types that you will likely encounter. And let's go through the license levels, which tell users what they can and can't do with the content.

The Open Data Institute explains there are two main types of license that relate to different types of content:

- "Open Licences for Creative Content: *Creative content, such as text, photographs, slides and so on, may be licensed using a Creative Commons Licence.*"
- "Open Licences for Databases: *You might encounter a similar set of licences which is available for databases from the Open Data Commons.*" [Reference]

The three levels of openness

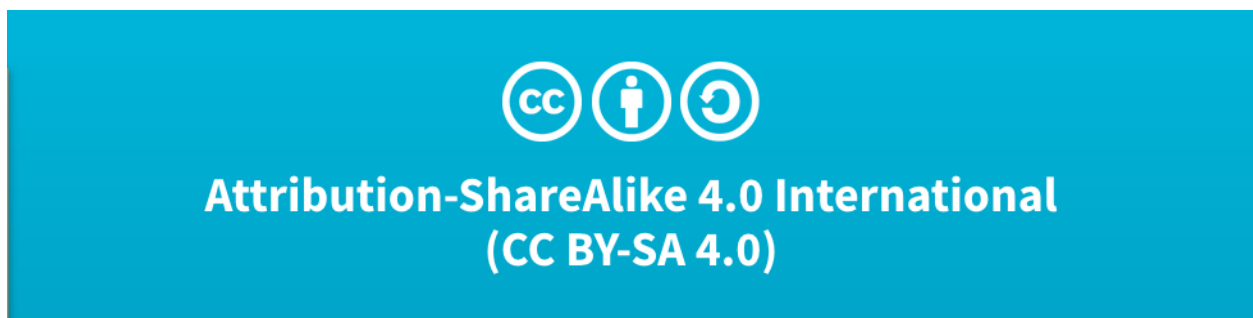
Open Licenses for Creative Content and Databases both have three levels of openness. Each level gives the user progressively more flexibility in *how* they use the content. The publisher sets the level.

Here are the three levels of open licenses with their Creative Commons' labels next to them:

- Public Domain: CCO
- Attribution: CC-by
- Attribution & share-alike: CC-by-sa

The thing is, when you find a piece of content with an open license, the labels will look slightly different from the three above.

You may recognize the titles better as symbols that appear next to or close to the content that has an open license. Below is an example of an attribution & share-alike license:



Here are the three levels of open licenses with their "Open Data Commons Licence" labels next to them:

- Public Domain: PDDL
- Attribution: ODC-by
- Attribution & share-alike: ODbL

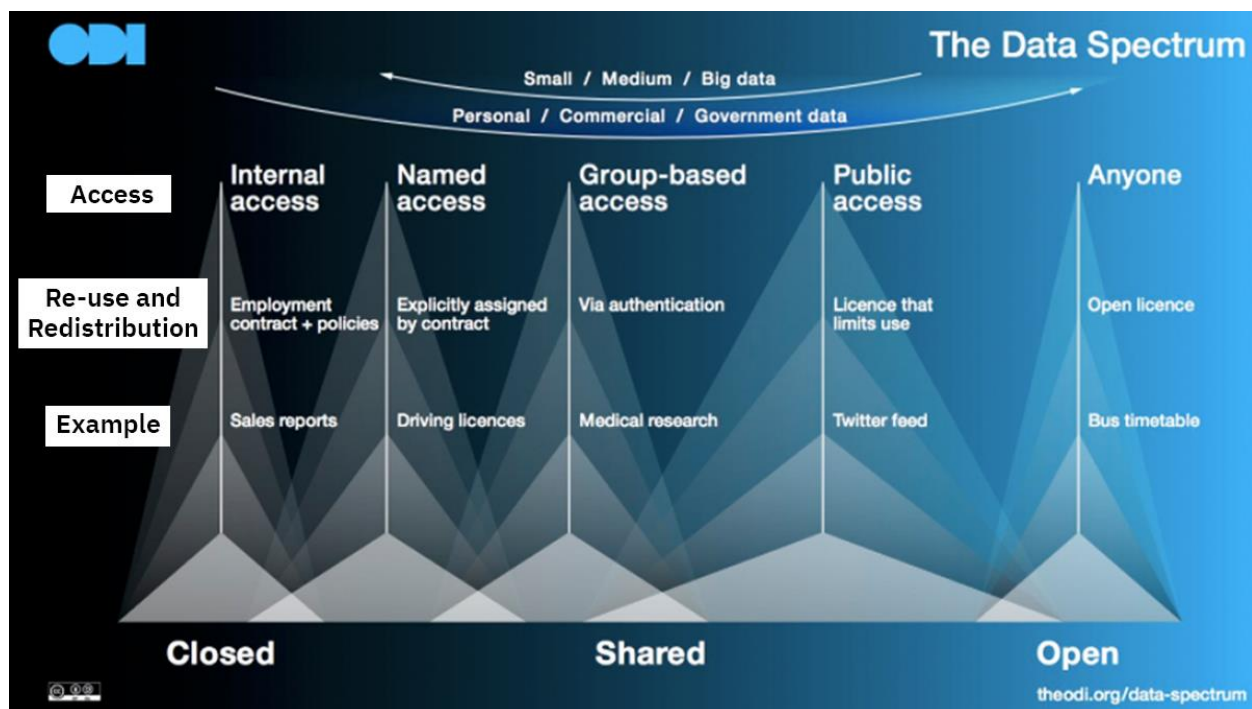
Open licenses are fundamental to the sharing and use of open data. We encourage you to investigate the Reuser and Publisher's guides by the ODI linked above to learn more.

Next, we'll look at why different levels and licenses of openness are necessary and useful.

Why are there different levels of openness?

The simple answer is that all data represents a real-world fact, and sometimes we shouldn't or don't want to share those facts for privacy or security reasons.

Data sits on a spectrum. Again, those helpful folks at ODI have created a graphic (we've re-used, remixed, and reshared - see what we did there) that shows this concept in action.



On the right is open data, the left closed, and the middle shared. We've labeled Access, Re-use and Redistributed, and Examples on the left.

Spikes to the left of "Shared" show inherently closed or partially closed situations; therefore, they don't have a license that allows re-use and redistribution.










On the right of "Shared," there are two different licenses mentioned: "Licence that limits use" and "Open licence". We can see under the respective examples that this could be a Twitter feed and a bus timetable.

It's safe to say that it is ok if "anyone" has access to a bus timetable. However, some people may not want to share their Twitter feed with everyone and limit access to it. We can see the varying degrees of data openness in these two examples.

License and level of openness combined

Now we've covered license types and levels of openness, we can see them in action together.

We produced this handy table in a [report about the future of open data](#) in the labor rights space that shows both the license type and level of openness for Creative Commons.

CREATIVE COMMONS LICENSING							
LICENCE TYPE		OPENNESS					
		SHARE	SHARE WITH ATTRIBUTION	CREATE MODIFIED VERSIONS	REDISTRIBUTE COMMERCIALY	RELEASE UNDER ANY LICENSE	
No Rights Reserved		Yes	No	Yes	Yes	Yes	
Attribution		Yes	Yes	Yes	Yes	Yes	
Attribution ShareAlike		Yes	Yes	Yes	Yes	No	
Attribution-NonCommercial		Yes	Yes	Yes	No	Yes	
Attribution-NonCommercial-ShareAlike		Yes	Yes	Yes	No	No	
Attribution-NoDerivatives		Yes	Yes	No	Yes	-	
Attribution-NonCommercial-NoDerivatives		Yes	Yes	No	No	-	
All Rights Reserved		No	-	No	-	-	

What does "Universal Participation" of open data mean?

To recap, the Open Data Handbook states:

"everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example,

'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed." [reference]

In short, you can't exclude people or groups from using the data. For example, on our own website, we make it clear that any person or group may use the data for their chosen purpose:

"Making Metric data available under a Creative Commons license helps everyone to benefit from data researched and cultivated on [WikiRate.org](https://wikirate.org) for any purpose: including research, advocacy or even commercial purposes."

Notice that for-profit uses are not precluded. In fact, there are many good examples of businesses using open data, an example of which is referenced in the last section of this article.

What do “Data Interoperability” and “Technical Openness” mean?

To recap, the Open Data Handbook states:

"...the ability of diverse systems and organizations to work together (inter-operate). In this case, it is the ability to interoperate - or intermix - different datasets." [reference]

Most people will be familiar with the concept of interoperability, perhaps without even knowing.

Anyone who has changed from an iOS device to an Android one will know that it's complicated, if not impossible, to transfer purchased apps. One of the main reasons for this is the two systems are not interoperable.

Another example comes from the Covid 19 pandemic. Many EU member states launched contact tracing and warning apps to inform people if they had come into contact with Covid 19. The European Union ensured that "20 apps which are based on decentralised systems can be interoperable through the gateway service."



Now we've covered interoperability as a concept, we want to flag a key interoperability challenge common to the process of collecting data: Technical Openness of datasets.

Technical Openness

Technical openness of datasets facilitates interoperability; however, many organizations and companies don't consider this when publishing data.

Companies publish reports with data about their impacts in polished, text heavy PDF documents. These reports may look great on-screen or in-print. However, getting the data out of these reports in this format is highly time-consuming. Imagine going line-by-line and copying and pasting values into a spreadsheet.

Overall, It makes it much harder to compare, for example, how a company is performing on its environmental or social commitments.

So for that reason, the:

"Open Definition has various requirements for "technical openness," such as requiring that data be machine readable and available in bulk." [reference]

Machine Readability is something WikiRate is very keen on. You can read why here. You can read more about WikiRate in open data repositories and sources section

Why is WikiRate Open Data?

Simply put, WikiRate couldn't do what it does without open data.

WikiRate was founded on the idea that:

"seeing how companies' operations impact people and the environment is the first step in improving the world."

We make it our mission to create and open up datasets about corporate impacts so that **everyone** can have access. And when we say everyone, we mean everyone!

To name a few, our datasets have been accessed by students, sustainability experts, companies, investors, consumers, civil society organizations, trade unions and supply chain workers.

But we don't only make data open, we also make it possible for everyone to contribute to creating these datasets.

To create a picture of corporate impacts isn't so easy. There are literally millions of companies in the world. Building that picture with a dedicated team of around ten to fifteen employees would likely take several lifetimes.

But what about a team of several thousand people with a passion for greater company transparency?

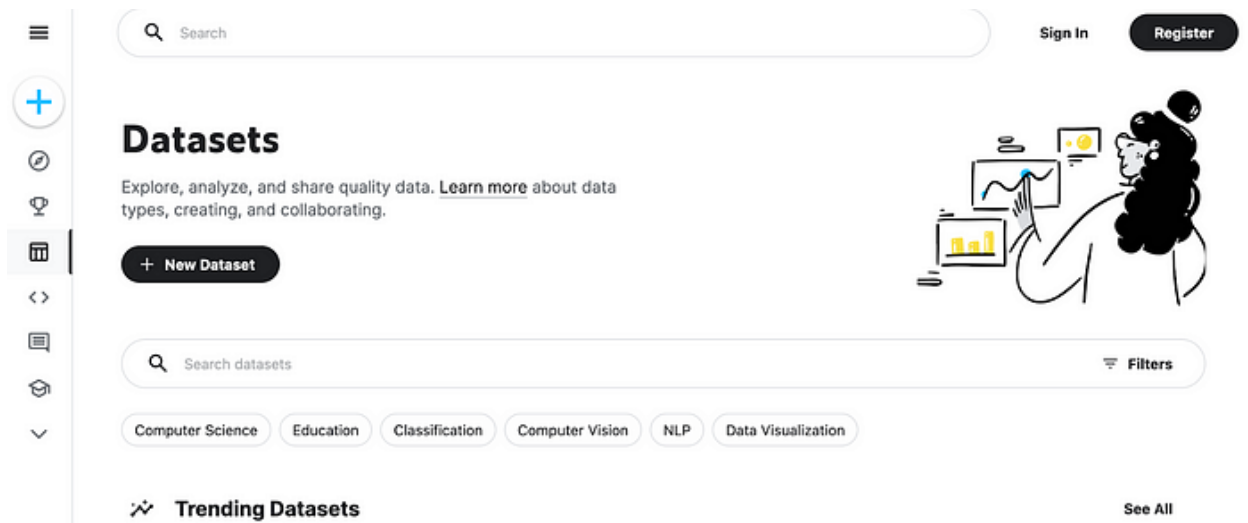
WikiRate embraces the idea that anyone can play a critical role in adding a piece of data to the picture. That's why we are a Wiki:

"[A Wiki is] collaboratively edited and managed by its own audience directly using a web browser." [reference]

This means anyone can add a piece of data to WikiRate to help create the bigger picture. In turn, because WikiRate is open data, other people can utilize the data how they choose, for example, to download it, share it and synthesize new analyses.

Open data repositories and sources

These are a few of our favorite examples of open data platforms and resources. Please let us know if you have an open data repository or source that you think should be added to the list.



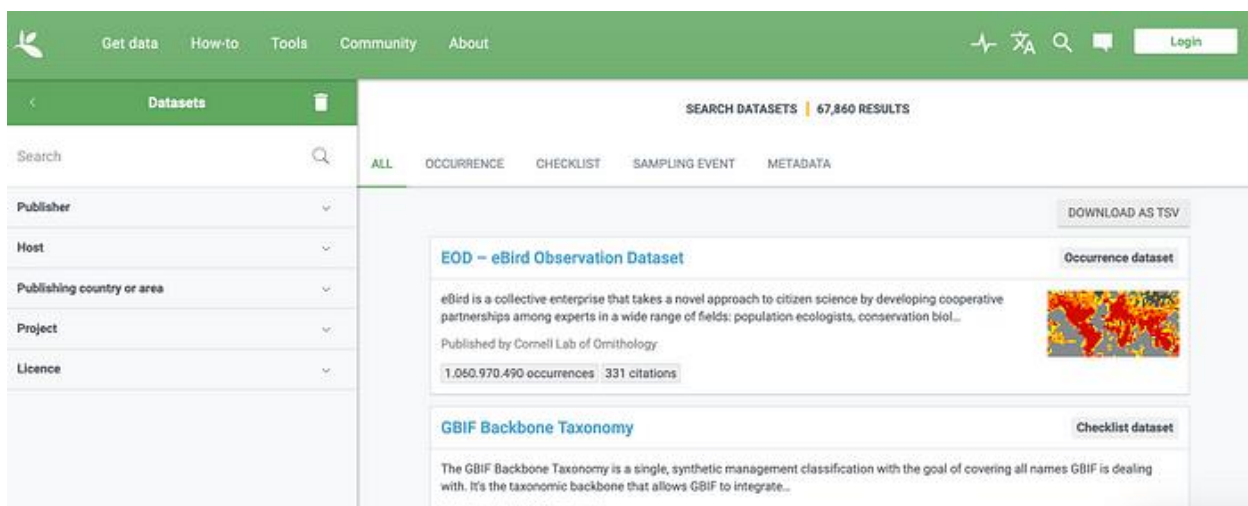
Kaggle

Name: Kaggle

Type: Data Repository

Data set types: Computer science, Education, Classification, Computer Vision, NLP, Data Visualizations

Description: Kaggle is a subsidiary of Google and is described as an "online community of data scientists and machine learning practitioners." As with many other Google products, the impressive things about this repository is its scope and depth, and versatility that allows the user to manipulate the data in many useful and interesting ways.

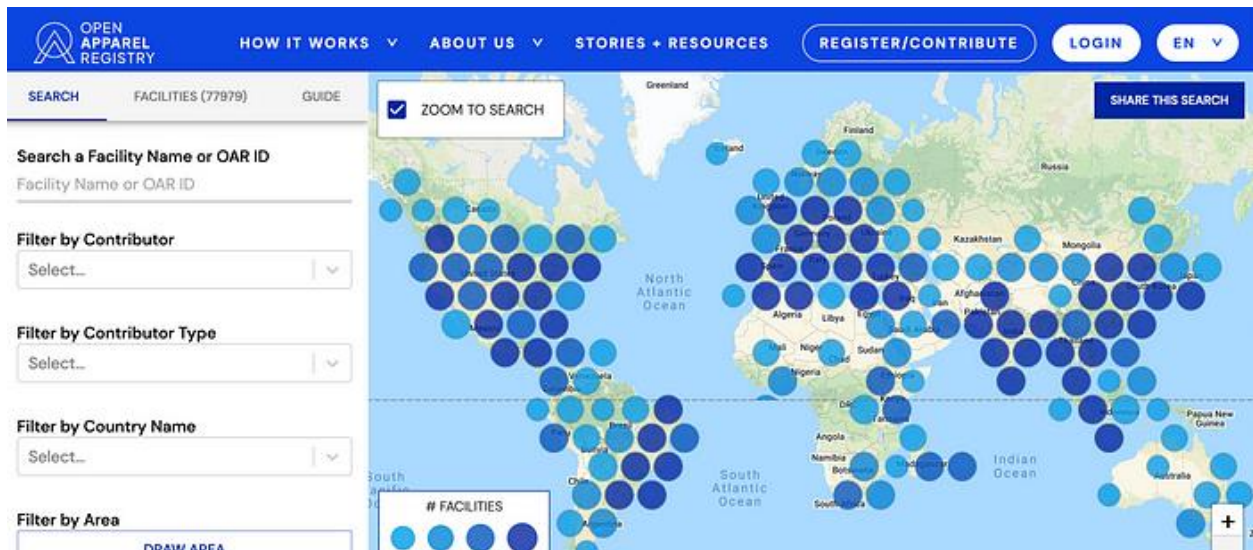


Global Biodiversity Information Facility

Name: Global Biodiversity Information Facility (GBIF)

Data set types: Occurrence, Checklist, Sampling Event, Metadata

Description: GBIF “is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth.”



Open Apparel Registry

Name: Open Apparel Registry

Data set types: Apparel Facilities

Description: “The Open Apparel Registry (OAR) is a free, open data tool mapping garment facilities worldwide and allocating a unique ID to each.”

Emissions Data

Global Carbon Budget

Global carbon cycle and emissions data, including fossil-fuel, cement, and land-use change emissions, atmospheric growth, and ocean and land sinks



CDIAC National Fossil-Fuel CO₂ Emissions

National CO₂ Emissions Estimates primarily derived from energy statistics



PRIMAP-hist

PRIMAP-hist combines several datasets to create comprehensive greenhouse gas emission data covering the years 1850 to 2015 and all UNFCCC (United Nations Framework Convention on Climate Change)

Open Climate Data

Name: Open Climate Data

Data set types: Emissions, Agreements, National Climate Plans

Description: Open Climate Data is a collection of climate related “datasets [that] come from different institutes and organisations”. The climate data sets are mostly hosted on GitHub.



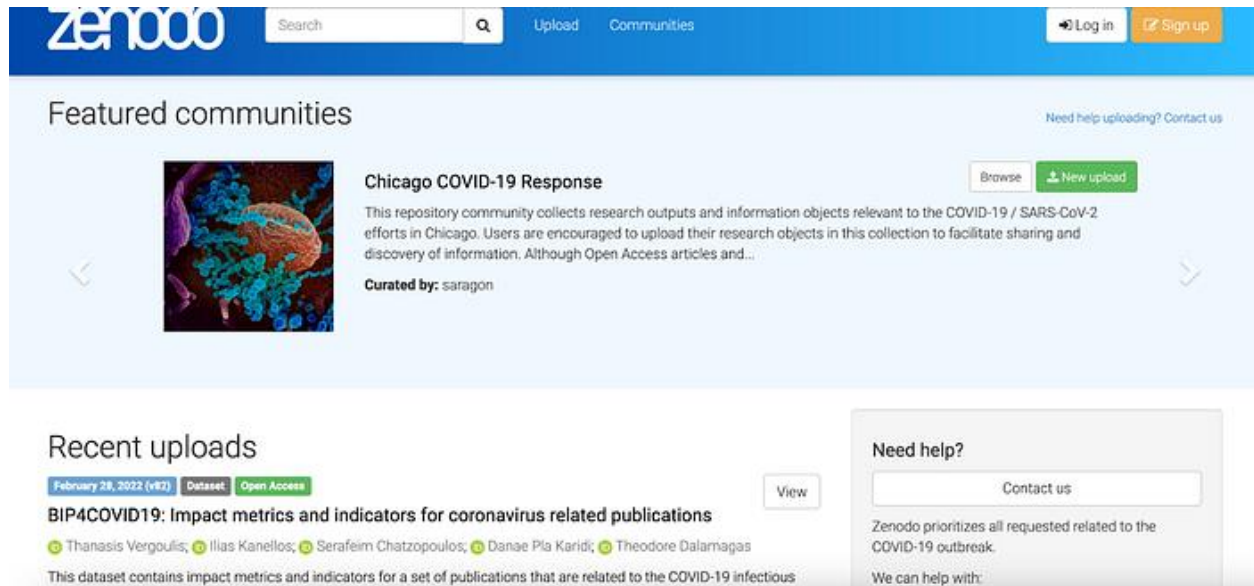
The screenshot shows the MetaBrainz website. The header includes the MetaBrainz logo and navigation links: About, Policies, Supporters, Sponsors, Reports, Donate, Register, and Sign in. The main heading is "The MetaBrainz Datasets". Below this, a paragraph states: "We ask commercial users to support us in order to help fund the creation and maintenance of these datasets. The MetaBrainz Foundation provides free, open access to various types of useful data gathered and verified by volunteers. Here, we list and describe the various datasets we maintain." There are three dataset cards: 1. MusicBrainz Data Dumps: Describes the MusicBrainz Database built on PostgreSQL, containing music metadata like artists, releases, and labels. 2. AcousticBrainz: Describes a project to crowdsource acoustic information for all music globally, including low-level spectral information and genre/mood data. 3. A third card is partially visible but mostly obscured.

MetaBrainz

Name: MetaBrainz

Data set types: music metadata, acoustic information, Creative Commons licensed music reviews, music listening archive:

Description: “The MetaBrainz Foundation provides free, open access to various types of useful data gathered and verified by volunteers.”



Zenodo

Name: Zenodo

Type set types: Data Repository

Data set types: Research Papers, Datasets, Research Software, Reports and any other research related Digital Artefacts

Description: Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN.

Welcome to Wikimedia.

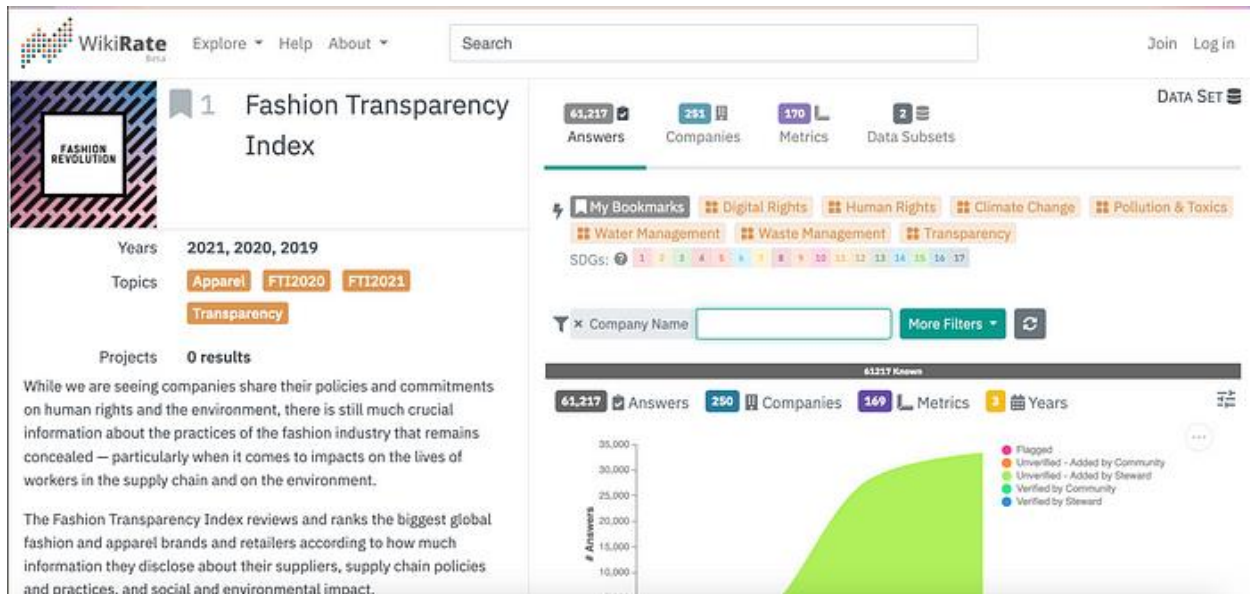


Wikimedia

Name: WikiMedia

Data set type: Multiple and Various!

Description: Probably the best known open data source in the world, it is a huge and complex entity that seems to resist categorization. It's almost like the repository of all repositories.



Wikirate

Name: Wikirate

Data set type: Company ESG data repository, Analysis tool, Communication and Resource creator

Description: Wikirate as our name suggests is a Wiki. Wikiraters research and add data about companies' impacts on the world. Wikirate provides the tools to collect and analyze the data, and communicate impacts to help tell the stories the world needs to hear.

Open Data resources and impact case studies

Please let us know if you have a resource or impact case study added to the list.

1. A Repository of Open Data Repositories: Open Data Impact Case Studies and Examples: Exactly as the name suggests, some of the most extensive repositories for Open Data can be found here.
2. Envisioning an Integrated and Open Labor Data Ecosystem Challenges and Opportunities: Humanity United and Wikirate have scoped out the opportunities and challenges involved in building an open and integrated data ecosystem for labor rights in supply chains.
3. Open Contracting: Impact Stories: Jammed with great examples of what 'openness can do'.

4. Open Data Case Studies: This blog covers some of the largest and most used open data sources. If you are new to open data this is a must read, and if you are a veteran you've probably used several of these.
5. Open Data Handbook Value Stories: Think open data is only useful to government and not-for-profits, think again. This collection includes a link to some fascinating stories of business using open data — including one that utilizes tree census data in Buenos Aires!
6. The Potential Role Of Open Data In Mitigating The COVID-19 Pandemic: Challenges And Opportunities: Good examples of “Use Cases Of Open Data In The Global COVID-19 Response.”

References

1. European Data Portal | The Economic Impact of Open Data Opportunities for value creation in Europe, European Data Portal and European Commission.
2. Open Knowledge Foundation | The Open Data Handbook.
3. Open Data Institute | Publisher's Guide to Open Data Licensing.
4. Open Knowledge Foundation | Defining Open Data by Laura James.
5. Wikipedia | Wiki [Accessed Friday 3 March, 2022].