Green Deal Data Observatory

Finding reliable historic and new data and information about climate change, as well as the impact of various European Green Deal policies is surprisingly hard to find if you are a scientific researcher. And it is even more hopeless if you work as a (data) journalist, a policy researcher in an NGO, or if you are responsible for the corporate social responsibility disclosures of a company that does not provide you with an army of (geo)statisticians, data engineers, and data scientists who can render various data into usable format, i.e. something that you can trust, quote, visualize, import, or copy & paste.

With the help of modern data science, open data, and open science we started to build solutions for these problems. We would like to find partners to build up a Green Deal Data Observatory that serves the needs of all stakeholders of the European Green Deal with far more practical information and data solutions than the Commissions' Taxonomy Compass.

Our aim is to provide hundreds of more comprehensive, timelier, and scientifically better validated data service than Eurostat's intergovernmental service, because putting the European Green Deal into practice requires better data products, more user-friendly service, and quicker response. We aim to provide a far better value-for-money for co-founders than any of the 60 observatories (permanent, thematic data gathering and dissemination programs) recognized by the EU, OECD, or UNESCO.

The 2019/1024 Open Data Directive makes access to taxpayer funded satellite, opinion poll and other data assets in the governmental and scientific field available for reuse but expects the investment to make them usable from private non-profit or for-profit parties. These data assets, collected for literally billions of euros annually, are sitting idle, and not contributing to environmental policies or sustainability targets. We are looking for partners to make this investment, and re-release hundreds of already public, but low-quality datasets, and further hundreds of missing datasets in validated, high-quality format, and with visualizations.

Green Deal Data Observatory

A subjective republication of the data products of the European Environmental Agency and Eurostat

Public data, such the statistical data products of the European Environmental Agency (EEA) or the Eurostat are not easy to use, and they are not really timely. The creation of these indicators requires intergovernmental consensus, and these agencies have no mandate to re-process and improve data provided by various governmental actors. Even though these organizations offer almost 1000 environmental and sustainability data products, their quality has many problems, and this makes their use for data journalists, NGOs, or smaller corporations impractical who do not have in-house data scientist. We want to automate this data science improvement and make it a public good.

The data products of the EEA or the Eurostat are valuable because they are free and because they go through an authoritative quality control process. But we can make them more valuable if they go through further quality improvements, not permitted because of the lacking intergovernmental consensus or lacking institutional capacity. Our data releases are subjective in the sense that they are not gaining authority from an intergovernmental agency, but from a transparent, open scientific process. They have an author, various data contributors (curator, manager, publisher, as defined by open science standards), who follow various scientific peer-review processes for their computing algorithm, codebooks, data validation. We make quality improvements that are based on a statistical and scientific consensus: we follow the same guidelines that Eurostat would use; we use peer-reviewed, statistical software algorithms, and we document each data scientific step how we made a better statistical product. (See more details in Appendix.)

You can see three examples for this work. Biomass Exports by Country has 18% more datapoints, and in a statistical software or machine learning, AI application that cannot handle missing data, it has a 10.4% larger best full partial dataset. The Government Budget Allocations for R&D in Environment dataset 40% more countries, and a 23% larger dataset available for chart making, or supervised and unsupervised learning models. The Environmental Subsidies and Similar Transfers from Europe to the Rest of the World dataset we have increased the largest (congruent) part of data that can be used in machine learning and software, or in data visualization for journalists by 167%.

1. Our co-founder or main sponsor can be the publisher of these improved statistical products, and increase its impact by directing traffic from journalists, researchers, NGOs, corporate social responsibility managers from the websites of the European Environmental Agency or the Eurostat to its own publication.
2. We built a modern API (similarly to Eurostat) which provides JSON access to frequent, programmatic users, for example, who already read into their corporate or research systems automatically Eurostat's data, could do with our 'improved Eurostat' data.
3. To increase impact and credibility, we place each item in the European open science repository Zenodo where almost all EU-funded research ends up from the Horizon Europe program, too.
4. For further impact, data journalists, and researchers looking for ideas, we place free visualizations of each data on Figshare, a similar scientific data visualization repository.

Green Deal Data Observatory

Filling the data gaps left by intergovernmental agencies

Most green NGOs cannot finance an environmental satellite to monitor land cover or cloud patterns. High-quality Pan-European surveys to understand how the public opinion is shifting cost millions of euros to conduct. Only large governmental bureaucracies and universities can afford such research – therefore the EU Open Data Directive, and various open science regulations make it mandatory to share all such taxpayer data sources for non-profit or even commercial reuse.

Data economics is unusual. Some data collection, like sensory data in agriculture and water management is getting so cheap that the bottleneck is in smart use of the data – with big data we are "drowning in numbers". Other data collection, particularly satellite-based data, or high-quality face to face surveys in an era a social media fatigue and social distance due to Covid is becoming extremely expensive. Both extremes lead to data overcollection: because it is so difficult to find interviewees for opinion polls, the surveys of statistical agencies, or the European Commission, ask more questions that they need. But they only publish and process what they immediately need.

The EU Open Data Directive and the open science regulations leave enormously large data assets legally open for reuse in data journalism, NGOs, climate mitigation research or developing corporate sustainability indicators. However, tax authorities, environmental agencies, the Eurobarometer program, and other data collectors are only obliged to hand over the data as is, without changing the file format, providing an easy-to-use English language manual, a codebook. The policymaker envisioned that it will be the private sector and non-governmental sector that will make these further investments – and that is exactly what our data observatories aim to do in an open collaboration with potential non-profit, academic, and even corporate users.

We have released open-source, scientifically validated software to bring to light much of these data assets. In our blogpost 100,000 Opinions on the Most Pressing Global Problem to understand how we do this in more detail. Our retroharmonize software allows the creation of new datasets, maps, charts, model calculations from data left for private investment in many pan-European survey programs. Our iotables software makes automatic economic and environmental impact analysis using open data that we process form every EU member state, and for comparison, from the UK, US, and Japan.

5. Our co-founder or main sponsor can be the publisher of these new statistical products, popular maps (or best used map received 100,000 social media interactions) and increase its policy impact by being a reliable data source of curated information of its interest. We can not only improve the data quality of the Eurostat and EEA, but we can double the quantity available public information, too.
6. With our technology and know-how, we can significantly reduce our observatory partner's research costs. Harmonization with existing, taxpayer funded, open governmental or open science data means that you must collect less information.

Green Deal Data Observatory

Call for open collaboration

Intergovernmental organizations like the EU, OECD, World Bank, UNESCO have been co-financing and supporting the creation of data observatories, i.e., permanent data recording, collection, processing, and dissemination points for decades. We have reviewed the services, budgets, organizational and governance issues of about 80 observatories, including defunct (failed, abandoned, discontinued) ones.

- Our observatory is based on the agile open collaboration method of open-source software development and open knowledge projects like Wikipedia. It is decentralized, and allows that large organizations, such as foundations, corporation, universities, or even the European Commission or the World Bank can team up with individual researchers, research groups, NGOs, data journalists or citizen scientist.
- We utilize peer-reviewed, open-source software for data collection and processing, and we use reproducible research automation for the higher quality production of documents, visualizations, codebooks, and other media assets instead of the mainly manual, error prone and costly working methods of most of the traditional data observatories. We even wrote software to correct the mis formatted data of best funded data observatory that we know of.

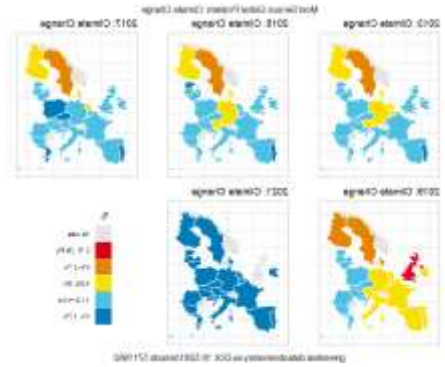
Benefits

1. Better user experience: We employ the best data formatting and documentation standards to make the data as interoperable and as easy to use as possible. We want to make sure that our data is truly plug-and-play in any environment where data journalists, NGOs, or corporations work: the data follows tidy data principles (open in Excel, OpenOffice, SPSS, STATA, import easily to Access, relational databases, can be accessed by computers, too, via our API), follows the FAIR metadata librarian standards (data can be connected with library catalogues via standardized subject headings, keywords, truly meaningful descriptions) and in interoperable file formats (which work on Windows, Mac, Linux systems.)
2. Value for money and scalable budget: We have minimal fixed labour costs, and with automated software adjustments we can increase the offering as more curators or data is present. The reviewed cc 80 observatories have an annual budget of about € 30,000 to € 3 million – we believe that we offer a very good service with an annual budget of €100,000 euros and a top-tier, world class service with € 300,000 per annum. For the minimal budget, we can offer a far better service than the reviewed low-cost observatories which usually are centred around very niche topics.
3. Collaborative data accumulation: Many public and private organizations collect data for different purposes. We aim to carry out own data collection, but also encourage various entities to place their data in our observatory, particularly if it is already public. We will polish their data assets to a much higher quality, ensure that they are constantly refreshed, and we connect them with hundreds of other data tables for more insight.

4. Transparency and integrity: We use only open-source data processing software, and we send our methods and data for various forms of scientific peer review. We keep the data open, and we store the data on the repository managed by CERN (the European Organization for Nuclear Research) to maintain full data integrity and long-term storage.
5. Flexible funding opportunities: We want to keep as much data and as much high-quality visualizations open for reuse as possible. We can receive grants to do this, but we can also sell data-as-service to corporations, provide exclusive service to partners. If necessary, we can charge money for some data or visualization services.
6. Measurable impact: We thrive to be the most used data source in the open scientific repositories of Europe, and we provide full analytics to the use, including scientific publication use, of our data. We want to create a competitive service that has higher traffic than other data observatories and brings data to a similar number of users like Eurostat or the European Environmental Agency website, or the Copernicus Data Service (for Europe's enviromental satellite sensory data) does. We want to make sure that evidence for policy making, sustainable finance activities and other imporant uses has the best possible dissemination via or observaotry.

Green Deal Data Observatory

Annex I: Bigger Better Faster More Data



Novel data products: Official statistics at the national and European levels follow legal regulations, and in the EU, compromises between member states, which means that they create new products with 5 years' delay after a problem arises. Not tied to these official procedures, but using the very same data and methodology, and different but equally thorough data quality procedures, we can produce indicators almost immediately. We only need a short validation period when you can make sure that you and all users are happy with the information content, coverage, timeliness, and data quality of our releases. See our blogpost 100,000 Opinions on the Most Pressing Global Problem to understand how we do this in more detail.

Better data: Statistical agencies, old fashioned observatories, and data providers often do not have the mandate, know-how or resources to improve data quality. Using peer-reviewed statistical software and hundreds of computational tests, we can correct mistakes, impute missing data, generate forecasts, and increase the information content of public data by 20-200% percent. This makes the data usable for NGOs, journalists, and visual artists—among other potential users—who do not have this statistical know-how to make incomplete, mislabelled, or low-quality data usable for their needs and applications. See our example with the indicator Government Budget Allocations for R&D in Environment

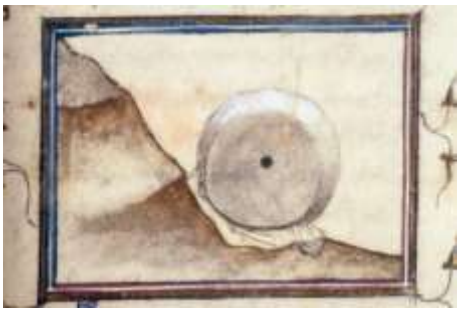
Never seen data: The 2019/1024 directive on open data and the re-use of public sector information of the European Union (which is an extension and modernization of the earlier directives on re-use of public sector information since 2003) makes data gathered in EU institutions, national institutions, and municipalities, as well as state-owned companies legally available. According to the European Data Portal the estimated historical cost of the data released annually is in the billions of euros. But if this data is a gold mine, its full potential can only be unlocked by an experienced data mining partner like Reprex. Here is why: data is not readily downloadable; it sits in various obsolete file formats in disorganized databases; it is documented in various languages, or not documented at all; it is plagued with various processing errors. We make the powerful promise Government Budget Allocations for R&D in Environment of the EU legislation a reality in the field of the Green Deal policy context.

Green Deal Data Observatory

Annex II: Increase Your Impact, Avoid Old Mistakes

Reprex helps its policy, business, and scientific partners by providing efficient solutions for necessary data engineering, data processing and statistical tasks that are as complex as they are tedious to perform. We deploy validated, open-source, peer-reviewed scientific software to create up-to-date, reliable, high-quality, and immediately usable data and visualizations. Our partners can leave the burden of this task, share the cost of data processing, and concentrate on what they do best: disseminating and advocating, researching, or setting sustainable business or underwriting indicators and creating early warning systems.

Increase impact: We publish the data in a way that it is easy to find—as a separate data publication with a DOI, full library metadata, and place it in open science repositories. Our data is more findable than 99% of the open science data, and therefore makes far bigger impact. See our data on the European open science repository Zenodo, managed by CERN (the European Organization for Nuclear Research).



Easy-to-use data: Our data follows the tidy data principle and comes with all the recommended Dublin Core and DataCite metadata. This increases our data compatibility, allowing users to open it in any spreadsheet application or import into their databases. We publish the data in tabular form, and in JSON form through our API, enabling automatic retrieval for frequent users. We not only increase compatibility: our statistical software with hundreds of built-in checks makes those individually simple, but manually error-prone steps, like converting thousand euros to million euros, removing the % sign, converting kilograms to tons, or dollar amounts at the correct exchange rate to euros that manual processing so often gets wrong. See our blogpost on the data Sisyphus.

Green Deal Data Observatory

Appendix III. From Data to Solutions

With our new software-as-service product we want to show that the open, scientific, peer-reviewed software that we develop and the enhanced open data that our Green Deal Data Observatory provides can create significant, tangible user value. Our new Eviota service prepares financial institutions, large companies to report the total sustainability impact on their entire value chain, from suppliers to buyers. According to the impact assessment of the EU sustainable finance pact, preparations to the implementation will cost for large companies on average 25,000 euros, and compliance will cost on average 75,000 euros annually. We want to show that using open science and open data, the sustainability oversight and management of the entire supply chain can be made more reliable and cheaper. In fact, so cheap that SMEs and non-profits, who need not comply with the CSRD Directive and apply the EU Taxonomy Regulation can choose to follow this best practice in a very cost-effective way. See our offering to large corporations that must comply with the CSRD directive, for financial institutions, and for SMEs and nonprofits that want to embrace the total impact assessment of their value chain in a cost-effective way.

Get in touch with us:

- Green Deal Data Observatory on LinkedIn.
- Green Deal Data Observatory on Twitter.
- Daniel Antal, co-founder, LinkedInv or Email.