European Pollinator Observatory

2023-07-06

An **observatory** is a location historically used for observing terrestrial, marine, or celestial events. In the last 30 years, many social, humanities and economic observatories were established to provide a consistent and permanent data and knowledge recording point. We have counted more than 80 EU, OECD or UN institution recognised, international data observatories; every year, about 3-5 new ones go functional and 1-2 become discontinued. Their services, data quality and quantity varies.

`Reprex` is a Dutch startup that grew out of the prestigious `Yes!Delft AI+Blockchain Lab`[^introduce-reprex]. With the support of the EU's Horizon Europe research and innovation program, we give the data observatory concept a modern and consistent format. We use the latest innovations of statistics, data science, data governance and open science to help knowledge triangles (academic, business and public policy partners) to build new observatories fast with a very high data quality. Because we employ the open collaboration method to use open data, and build on open source modular components, you can start recording systematic data in days.

In the `Open Music Europe` project, we build a data-to-policy pipeline for the fragmented European music sector. Based on the European Commission's analysis of data gaps, we collect and integrate data about the diversity and circulation of music, the economic, societal and sustainability aspects of music, and we are fueling innovative applications with data[1]. A data pipeline is a method in which raw data is ingested from various data sources and then ported to a data store for further analysis, in this case, to an open, shared, collaborative music observatory. We extend this pipeline with the help of an open-source statistical software ecosystem that processes the data into reproducible, self-refreshing datasets, visualisations, report components, or entire internal or public reports. We are automatically those research jobs that are usually not rewarded and neglected: repeated input and download, proper documentation, unit-testing, and formatting so that our human users can focus on where they beat the computers: interpreting the information our observatory provides.

[1]Reprex initiated the Open Music Europe Consortium which is building a data-to-policy pipeline for the European music sector on the basis of its Digital Music Observatory minimum viable product following the Feasibility study for the establishment of a European Music Observatory.

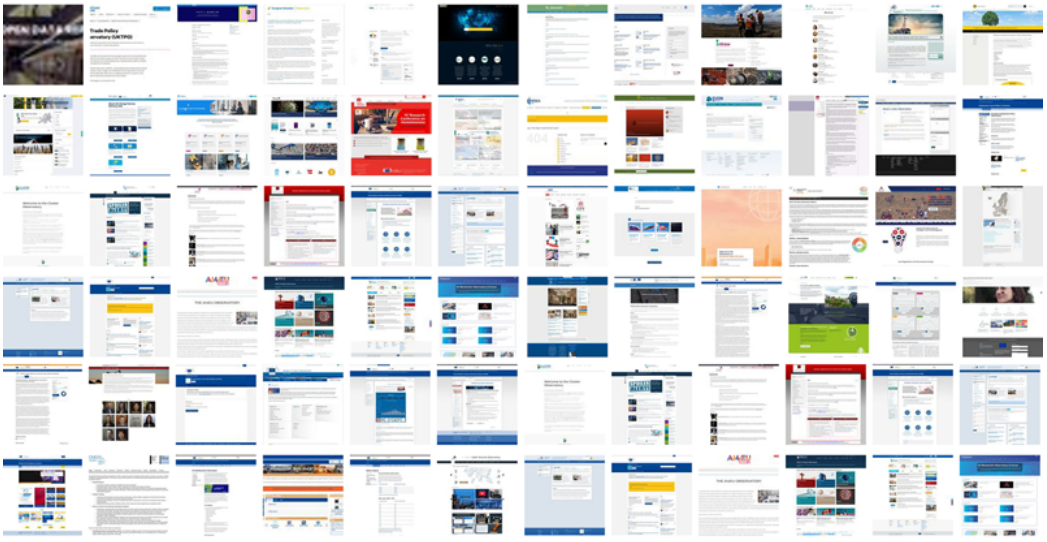


Figure 1: Various UN and OECD bodies, and particularly the European Union support or maintain more than 60 data observatories, or permanent data collection and dissemination points.

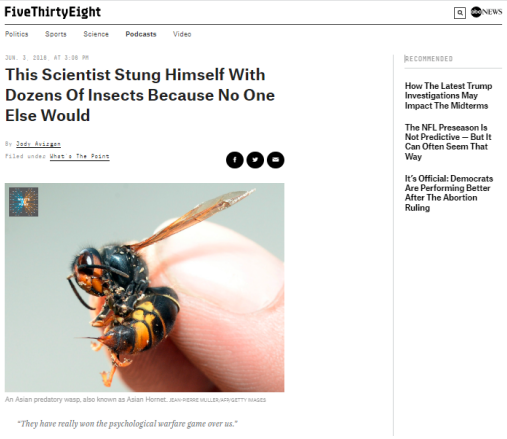
> **Open Call for Observatory**
>
> BeeSage & Reprex are planning to apply for the Horizon Europe Research and Innovation Action call: *Biodiversity and ecosystem services (HORIZON-CL6-2024-BIODIV-01)* (Horizon Europe Framework Programme (HORIZON) 2023) as *innovators*. This call requires to "build a platform that will serve one-stop shop for information on animal pollination ecology. A database with systematised information on plan-pollinator interactions, including the spatial dimension of plant-pollinator networks, should be part of the platform. The platform should build on what already exists and should be devised in close collaboration with researchers and other potential users."

Data coordination

We work with data curators and experts in their domain to tell us what data would be required to create evidence-based public or business policies [to make film production more sustainable/to understand more the role of animal pollinators in ecosystems and their services to agriproduce-based business supply chains]. Our data curators do not need to be data scientists or statisticians: they must know, first and foremost, the needs of their professional domain and, when we support them with data, to tell usable data from non-usable.

Recruit data curators

Data curators may be researchers, industry analysts, citizen scientists or data users. They should understand the desirability and usability of data that would help create better scientific articles; be used in public policy processes, which could form the basis of internal or external key performance indicators or benchmarks to set targets for the desired changes KPIs (for example reduced greenhouse gas emissions, or increasingly more equal pay among women and men in the supply chain.)


FiveThirtyEight

Politics Sports Science Podcasts Video

This Scientist Stung Himself With Dozens Of Insects Because No One Else Would

RECOMMENDED

How The Latest Trump Investigations May Impact The Midterms

The NFL Preseason Is Not Predictive — But It Can Often Seem That Way

It's Official: Democrats Are Performing Better After The Abortion Ruling

"They have really won the psychological warfare game over us."

"For the implementation of the eligibility condition on the 'multi-actor approach', proposals should ensure adequate involvement of researchers, farmers and other land managers, businesses involved in the food, medicine, energy and/or materials sectors, decision-makers at local and/or regional level, civil society organisations and other relevant actors."

- BeeSage is reviewing what already exists;
- Reprex is setting up a demo observatory as the backbone of our proposed research infrastructure;
- We are reviewing what already exists with farmers, land managers, decision makers at local/regional level and civil society organisations. We invite people from these organisations as `data curators` to tell us where shall we initiate data and knowledge gathering on our research infrastructure.

Get inspired: this scientist stung himself with dozens of insects because no-one else would volunteer to provide pain data - big data is saving this little bird

Reprex supports the data curators by mapping existing open and proprietary data and harmonising statistical processes to bridge the data gaps. The European Open Data Directive makes almost all publicly funded data collection accessible for free or at a very low cost in the EU; since the introduction of this directive, European public institutions and enterprises have made more than €200 billion euros worth of data available. This treasure trove does not only contain data about Europe: the EU's research programs, such as the Copernicus environmental satellite, provide sensory information about the entire surface of the Earth. Open data is second-hand data: it was collected originally for the public policy

needs of an EU or member state body, for example, a regulatory agency or a tax authority. Reprex helps the data curators to navigate these muddy waters and process the uncut free gems into usable data for [the film and media industry/for researching animal pollination].

Reprex is an expert on statistical data and metadata exchange and the business of data exchange. Because the needs identified by the data curators sometimes cannot be supported by open data, we connect with differently licensed private data sources, too. We help our observatory community to pay with open data and engage in exchanges instead of data purchases. Because open data comes from government agencies with special permissions and capabilities to collect data (for example, with satellites or with the powers of tax authorities), we can complement privately collected data with open data and create win-win situations.

> **i** Initiate data exchanges
>
> Using the new data layer of the world wide web facilitates automated connections between your databases, the observatory, and other databases where the observatory collects secondary data. With the help of your data curators, we can set up the first data ingestion pipelines from machine-actionable public sources, such as international organisations and statistical agencies' APIs, and initiate a few automated data collection processes. In our experience, we can produce unexpectedly high-quality datasets, visualisations and dashboards with very little investment. By being more competitive than in-house, not shared data units, this often encourages proprietary data owners to start using our observatory data and even to exchange their valuable proprietary data with us. We can offer win-win situations when the proprietary data owner and the observatory can enrich each other's datasets and visualisations.
> To put this into motion, we start our `European Pollinator Demo Observatory` with connecting data from `BeeSage` and its user and via `Reprex` the European Environmental Agency data warehouse.

Our data coordination eventually helps to organise more valuable and cheaper data collection plans. Using modern data harmonisation both on the level of microdata (with standardisation) and on the level of processed statistical data (with advanced data science like data fusion), we design so-called statistical processes, i.e., surveys or queries to the administrative records of industry, that minimise the data collection effort and focus on missing information. We use harmonised surveys to complement existing open and proprietary data.

Data management

We use the R environment for statistical computing to manage the data according to modern statistical and data science principles. We only use transparent, peer-reviewed, and open algorithms to harmonise, process, test, and improve the data: industry experts, auditors, and scientists can always check what goes into our datasets, visualisations and the machine learning models they fuel.

We strive for a very high level of interoperability. R has a high-level, interpreted language, which we use for our open and peer-reviewed code. A modern R environment can incorporate code snippets from C++, Python, SQL and R, the most likely used languages to work with our observatory's data. Our data processing and visualisation algorithms are peer-reviewed and tested on major Linux, Mac and Windows operating system distributions; because R is an interpreted language, it poses very few security issues when you want to review our products.

Our observatory's data products are uploaded regularly to Zenodo, the open science permanent repository of the CERN funded by the European Union. Zenodo provides every dataset version with a digital object identifier (DOI) and tampering-free independent data access for at least 20 years. This fosters data integrity and security: even observatory managers cannot change their datasets that are not logged for perpetuity. The data storage is independent of the observatory; in case the observatory or its manager is dissolved, the data will be available for reuse.

Our documentation is written in markdown and automatically disseminated with the datasets and visualisation on the open-source Hugo platform. Our Hugo websites are written in Google's open-source Go language to optimise web presence. We use the World Wide Web consortium's standards to accompany our data with metadata that help "to ensure that data management planning becomes standard scientific practice and to support the dissemination of research data that are findable, accessible, interoperable and re-usable (the FAIR principle)", therefore complying with the European Union's open science and statistical standards and recommendations. This makes the data observatory's products easily accessible for humans and machines, too[2].

i A cost-effective research infrastructure

Reprex's mission is not research but to support business, scientific and policy research with open-source data infrastructure. Our work can easily be financed from policy and scientific grants or tendered services because such services always budget for data management, documentation and dissemination. This is where our automated data observatory concept excels: it is reproducible due to its modularity, flexibility, and cost-effectiveness. This means that, in many cases, we can support you without cutting into your real research funding.
`BeeSage` and `Reprex` are not research entities, we are data innovators who work with scientific partners. Our role in the planned consortium is to be partners in charge of data management, dissemination and communication; an outreach to farmers, local government and civil society actors; and data providers about pollinators. We will leave the research to our university and other researcher partners.

Reprex's data observatory concept is no longer a prototype. In Open Music Europe, we are scaling up this data-to-policy pipeline for Pan-European use; we are not only contracted and reviewed by the European Commission, but we have signed a Memorandum

[2]Reprex and the Open Music Europe Consortium signed the *Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors' public policies* (Open Music Europe 2023b).

of Understanding to use and live-test our observatory in monitoring and further developing the Slovak Republic's cultural and creative industry policies on the national level. We also emphasise creating data products that not only support public policy but business policy, particularly on the transition of enterprises from financial controlling to a wider ESG (environmental, social and governance) controlling, reporting and audit requirement.

Reprex is a Dutch-American startup, and we follow the best practices of reproducible data science on both sides of the Atlantic. Apart from complying with the mandatory and recommended requirements of the Horizon Europe programme in the FAIR framework, and following, when applicable, the same quality assurance that European national statistical authorities do, we are early adopters of the U.S. *Open Policy Analysis Guidelines*[3] aiming for more transparent use of social sciences in providing policy evidence. In terms of our observatories, this means a higher reviewability of the assumptions we make when we prepare data for later policy analysis. When applying these standards to programming decisions, when possible, we follow global metadata standards, like the SDMX or W3C standards, which are applied in the EU and the US and the rest of the world similarly. Located in The Hague's *ImpactCity* incubator system for SDG impact startups, we are ideally located because the Netherlands is a global leader in statistics, and many European and international standards on measuring cultural participation or environmental and social impacts are developed and piloted in our Randstad region.

Data usability

Data productivity experience shows that data analysts spend 80-95% of their time searching for and reorganising data instead of analysing it. Because most [ecological/sustainable film] data analysts are not trained data engineers or statisticians, they tend to make the most mistakes during these unproductive steps. Human analysts make the most mistakes when they process the data: they move a number to another cell in their spreadsheet application; when they convert a dollar figure using the euro exchange data from a wrong date; when they multiply a number taken from the wrong cell. Another common source of the problem is misunderstanding the data semantics: taking a column in a table that means something else than the analyst thinks.

Our observatory uses the 'tidy data principle' and the 'datacube' organisation with the international Statistical Data and Metadata Exchange (SDMX) standard coding and documentation. These principles mix good data organisation with good semantics and minimise the need to move data cells when downloading, importing or using the data. They follow decades of experience to avoid as many misunderstandings in the analytical process as possible.

[3]See: Hoces de la Guardia et al.: *A framework for open policy analysis,* and the concise guidelines (BITSS 2019; Hoces de la Guardia, Grant, and Miguel 2020); our observatories are fulfilling the nine requirements on the highest level 3.

Encourage free trials

Tidy data follows the most fundamental principles of relational algebra, and by definition, it is very easy to import into an in-house organisational database: it is optimised to be imported. Using the new data layer of the world wide web facilitates automated connections between your databases, the observatory, and other databases where the observatory collects secondary data. It is also an optimal format to start working in your favourite spreadsheet software like Excel or statistical software packages like SPSS or Stata (our observatories into their file formats automatically.) We can set up demonstration datasets that will speak for themselves.
`Reprex` and `BeeSage` is aiming to create a European Demo Pollinator Observatory ready with farmers, civil society actors and local government entities by the 17 October 2023 when the `Biodiversity and ecosystem services (HORIZON-CL6-2024-BIODIV-01)` will be open for applications. By this time, we would like to find scientific partners to form a consortium into getting this grant.

When our data curators tell us that our data needs to be reorganised in a spreadsheet application like Excel, we rather redefine the dataset to make sure that analysts do analyse the data and not re-process them. This way not only improves the quality of the analysis but saves countless unproductive hours.

Embrace open

According to the study published by the European Commission on the impact of open-source software (OSS) and open-source hardware (OSH) on the European economy, conducted by *Fraunhofer ISI* and *Open Forum Europe* on 6 September 2021, open-source software contributed between €65 to €95 billion to the European Union's GDP. It promises significant growth opportunities for the region's digital economy. The 2020 report on the *Economic Value of Open Data* estimated the value of open data available in the European economy at €184 billion and is forecasted to reach between €199.51 and €334.21 billion in 2025. To unlock this potential, the study makes similar but less specific recommendations as the OSS/OSH study described above[4].

Open is not cheap

We agree with the EU policies: OSSH and open data are far more effective in value creation than proprietary software, hardware or data. It offers a better price-to-value, and in many cases, open data has no proprietary competitor. For example, the only affordable way to receive data from satellites from space (very useful for market research and sustainability measurement) is via the open data regime of the EU; space exploration is way too costly and regulated for private data vendors. Because

[4]See *The Impact of Open Source Software and Hardware on Technological Independence, Competitiveness and Innovation in the EU Economy* (European Commission et al. 2021)and *Economic Value of Open Data* (Huyer and van Knippenberg 2020).

open solutions are easier to externally audit, and they are open to open scientific peer review, the quality is often higher.
Open data is free from a collection point of view, but it needs to be (re)processed for the particular use your users have in mind. Open data curators and open source developers are highly trained professionals employing very thought-after skills. They need to be remunerated for their work. Because OSSH often offers a better value for money, using open solutions is not a matter of budgeting but changing your procurement and acquisition strategy. You are not buying data or databases, but you pay for data integration services or you provide grants to developers. Liability requirements may need to be adjusted, too.
`Reprex` and `BeeSage` will critically review pre-existing data assets and sources that can be used for the biodiversity and ecosystem services monitoring from a pollinator point of view.

Increase impact

Our open data observatories are about increasing impact with a better division of tasks between computers, data scientists, and domain-area researchers. We employ an open-source software ecosystem to automate repetitive, boring tasks where humans fail: data documentation, logging, data processing and wrangling, and visualisations. Often delegated to junior consultants, postdocs, or PhD students, this seemingly non-essential task is not credited in science, not billable in consultancy, and not well compensated. Our observatories take over these tasks to allow humans to focus on their research insights, business and policy control, reporting or audit.

Even though publicly funded research and services are often required to disseminate their data and results to the public, most researchers, analysts, and scientists are not trained for these roles. Our observatories employ the latest developments in machine-to-machine communication, standardisation and data exchange to fuel reports, websites, newsletters, and policy documents with data, citations, visualisation, and tables. We place your research output on knowledge graphs and the Internet of Things to create the most competitive technological dissemination platform. We enrich research products with very rich metadata that help other researchers' computers, search engines, library services, and databases to find and use your output. We break them down so that a single dataset or visualisation can earn credited use.

As a result of this, you can expect a far greater scientific impact and also a much bigger business and policy impact. Your analysis, experts, and decision-makers can focus on what they know best and make sure that the data as evidence is available for them in the most usable format with a very favourable value-for-money ratio. Our observatories allow your team to utilise a shared, open collaboration of data engineers and data scientists that would be very costly to develop and maintain in-house in your organisation.

BITSS. 2019. 'Guidelines for Open Policy Analysis'. Berkeley Initiative for Transparency in the Social Sciences. http://www.bitss.org/wp-content/uploads/2019/03/OPA-Guidelines.pdf.

European Commission, Directorate-General for Communications Networks, Content, Technology, K Blind, S Pätsch, S Muto, M Böhm, T Schubert, P Grzegorzewska, and A Katz. 2021. *The Impact of Open Source Software and Hardware on Technological Independence, Competitiveness and Innovation in the EU Economy : Final Study Report*. Publications Office. https://doi.org/10.2759/430161.

Hoces de la Guardia, Fernando, Sean Grant, and Edward Miguel. 2020. 'A framework for open policy analysis'. *Science and Public Policy* 48 (2): 154–63. https://doi.org/10.1093/scipol/scaa067.

Horizon Europe Framework Programme (HORIZON). 2023. 'Call: Biodiversity and ecosystem services (HORIZON-CL6-2024-BIODIV-01)'. European Commission. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl6-2024-biodiv-01-3.

Huyer, Esther, and Laura van Knippenberg. 2020. *The Economic Impact of Open Data. Opportunities for Value Creation in Europe*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2830/63132.

Open Music Europe. 2023a. 'Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems'. https://doi.org/10.3030/101095295.

Open Music Europe, Ministerstvo kultúry SR AND. 2023b. 'Memorandum o porozumení o využití výsledkov analýz otvorených politík v kontexte slovenského kultúrneho a kreatívneho priemyslu a sektorových verejných politík v spolupráci s konzorciom pre výskum a inovácie s názvom OpenMuse. [Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors' public policies]'. https://www.crz.gov.sk/zmluva/7645338/.