

Automated Observatory Contributors' Handbook

Daniel Antal, CFA

2021-07-01

Contents

List of Figures

Big Data for All

We want to make sure that every individual research, artists, professional, NGO, small and large organization can take equal benefit in the era of big data from artificial intelligence.

Big data creates injustices, because it is the big corporations, big government agencies and the biggest, best endowed universities that can finance long-lasting, comprehensive data collection. Big data, and large, well-processed, tidy and imputed datasets allow them to unleash the power of machine learning and AI. They create algorithms that decide the commercial success of your product, your artwork, gives them a competitive edge against smaller competitors and helps them evade regulations.

We are looking for partners to develop our [technological solution](#) in [financially sustainable way](#) with bringing more and more relevant, [curated](#) open data [to light](#). Our Product/Market Fit was validated in the world's 2nd ranked university-backed incubator program, the [Yes!Delft AI Validation Lab](#). We are currently developing this project with the help of the [JUMP European Music Market Accelerator](#) program.

How We Add Value To Your Data?

Many countries in the world allow access to a vast array of information, such as documents under freedom of information requests, statistics, datasets. In the European Union, most taxpayer financed data in government administration, transport, or meteorology, for example, can be usually re-used. More and more scientific output is expected to be reviewable and reproducible, which implies open access.

From Data to InformationWe create high value key business indicators, policy evaluation indicators, scientific proofs requires the combination of matching, well formatted, and verified, controled pieces of data. It comes from a verified and legal source, with information about use rights and its complete history. We do not deal in blood diamonds.

Adding metadata exponentially increases the value of data Did your region added a new town to its boundaries – how do you adjust old data? Can I practically combine satellite sensory data with my organization records? And do I have the right? Metadata logs the history of the data, gives instructions who to reuse it in indicators, sets the terms of use. We automate this boring and labor intensive process applying the FAIR data concept.

Data is only potential information, raw and unprocessed.How I translate dollars to euros correctly? Are billioins adjusted for billions? Some of our indicators go through more than 10,000 processing steps. If your team does this in a spreadsheet or statistical software, there is no way it will be faultless - or that senior staff can verify it.

Data curationData sits in every government data warehouse (you can reuse it!), scientific journal, library, your sales records, in sensors. Not having access to it due to budgetary or legal constraints is an absolute barrier, but not being able to correctly assemble it into reliable information can keep its value low.

- In the [??](#). [Open Data](#) chapterwe investigate why not even those organizations, like the European Commission, use open data in their own data dissemination practices, who make them at least legally available.

- In the [??.](#) [FAIR Data and the Added Value of Rich Metadata](#) we introduce how we apply the concept of [FAIR](#) (findable, accessible, interoperable, and reusable of digital assets) in our APIs. Our data comes with metadata that meets the requirements of two metadata standards, the more general [Dublin Core](#) and the more specific, mandatory and recommended values of [DataCite](#) for datasets. We add rich processing metadata beyond these requirements. These are not only nice to have: from 2021, if you apply for Horizon Europe funding, your data must meet this criteria. Our solution can automate this process, and besides making you compliant, it adds a lot of value to your own data management.
- In the [??.](#) [Application: Automated Data Observatories](#) chapter we provide further technical information about our application. We use open source software and open data. The applications are hosted on the cloud resources of [Reprex](#), an early-stage technology startup currently building a viable, open-source, open-data business model to create reproducible research products. Our development team works on an open collaboration basis. Our indicator R packages, and our services are developed together with [rOpenGov](#).
- See [??.](#) [Service Design and Business Case Development](#) we lay out our ideas about finding a suitable business model for data sharing, and shared research activities that maintains the fairly shares the exponential value added from data integration across various business, policy and academic partners.
- The [??.](#) [Data Curators](#) chapter we give information for prospective curators. See also our [get inspired](#) and [your first contribution](#) subchapters.

Big Data For All

Machine learning and AI gives a competitive edge for large companies and governments that can exploit it. But training algorithms requires much, uniformly formatted, high quality data, and algorithms come with many potential side effects.

Trustworthy AIWe help deploying reliable AI that is under human supervision, and algorithms that will not turn against your organization, or engage in discriminative, unlawful or counterproductive behavior. We automate data and metadata management, documentation, verification, because computers are much better than humans in these boring tasks; humans must increasingly focus on oversight of too much data.

Open collaboration for data treasuresWe use the agile open collaboration project methodology of open source software development to make sure that large universities, consultancies, citizen scientists, individual artists and small NGOs can share the research budget, data assets and research of big data, and remain competitive against big tech and large organizations.

Data-as-Service.Most organizations cannot afford to build an in-house data science and data engineer team, or they do not even have in-house market research or IT. Instead of burdening your team with manual data downloads and ad hoc data manipulations, we offer you subscription for curated open and proprietary processed data. We keep all your data assets tidy, documented, and easy to use.

Automated Data ObservatoriesData sits in every government data warehouse (you can [reuse](#) it!), scientific journal, library, your sales records, in sensors. Not having access to it due to budgetary or legal constraints is an absolute barrier, but not being able to correctly assemble it into reliable information can keep its value low. Our observatories are built around open collaborations of [scientific](#), [business](#), public and NGO [policy](#) partners.

Automated Data Observatories

We are working around the open collaboration concept, which is well-known in open source software development and reproducible science, but we try to make this agile project management methodology more inclusive, and include data curators, and various institutional partners into this approach. Based around our early-stage startup, Reprex, and the open-source developer community rOpenGov, we are

working together with other developers, data scientists, and domain specific data experts in climate change and mitigation, antitrust and innovation policies, and various aspects of the music and film industry.

Green Deal Data Observatory is a modern reimagination of existing ‘data observatories’; currently, there are over 70 permanent international data collection and dissemination points. One of our objectives is to understand why the dozens of the EU’s observatories do not use open data and reproducible research. We want to show that open governmental data, open science, and reproducible research can lead to a higher quality and faster data ecosystem that fosters growth for policy, business, and academic data users. [Green Deal Data Observatory on LinkedIn](#) and the [Green Deal Data Observatory on Twitter](#), and join our [contributor team](#).

Digital Music Observatory is a fully automated, open source, open data observatory that creates public datasets to provide a comprehensive view of the European music industry. It provides high-quality and timely indicators in all four pillars of the planned official European Music Observatory as a modern, open source and largely open data-based, automated, API-supported alternative solution for this planned observatory. The insight and methodologies we are refining in the DMO are applicable and transferable to about 60 other data observatories funded by the EU which do not currently employ governmental or scientific open data. The [Digital Music Data Observatory on LinkedIn](#) and the [Digital Music Data Observatory on Twitter](#) and join our [contributor team](#).

is the first offspring of the [Economy Data Observatory](#) incubator. See further details in [?? Competition Data Observatory Chapter](#). We would like to create early-warning, risk, economic effect, and impact indicators that can be used in scientific, business and policy contexts for professionals who are working on re-setting the European economy after a devastating pandemic and in the age of AI. We would like to map data between economic activities (NACE), antitrust markets, and sub-national, regional, metropolitan area data. See the prototype on [web](#).

The Economy Data Observatory works now as an incubator for economy-focused data observatories. Find it on the [web](#) and on social media: [Economy Data Observatory on LinkedIn](#); [Economy Data Observatory on Twitter](#). Join our [contributor team!](#)

Chapter 1

Open Data

In the EU, open data is governed by the [Directive on open data and the re-use of public sector information - in short: Open Data Directive \(EU\) 2019 / 1024](#). It entered into force on 16 July 2019. It replaces the [Public Sector Information Directive](#), also known as the *PSI Directive* which dated from 2003 and was subsequently amended in 2013.

- [Open Data - The New Gold Without the Rush](#)
- [Open Data is Like Gold in the Mud Below the Chilly Waves of Mountain Rivers](#)

The EU has an 18-year-old open data regime and it makes public taxpayer-funded data in the values of tens of billions of euros per year; the Eurostat program alone handles 20,000 international data products, including at least 5,000 pan-European environmental indicators.

As open science principles gain increased acceptance, scientific researchers are making hundreds of thousands of valuable datasets public and available for replication every year.

The EU, the OECD, and UN institutions run around 100 data collection programs, so-called ‘data observatories’ that more or less avoid touching this data, and buy proprietary data instead. Annually, each observatory spends between 50 thousand and 3 million EUR on collecting untidy and proprietary data of inconsistent quality, while never even considering open data.

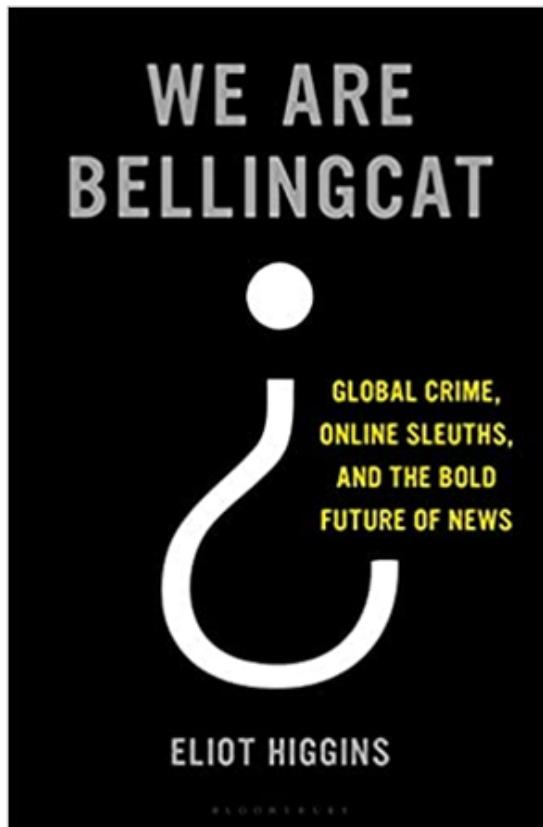


The problem with the current EU data strategy is that while it produces enormous quantities of valuable open data, in the absence of common basic data science and documentation principles, it seems often cheaper to create new data than to put the existing open data into shape.

This is an absolute waste of resources and efforts. With a few R packages and our deep understanding of advanced data science techniques, we can create valuable datasets from unprocessed open data. In most domains, we are able to repurpose data originally created for other purposes at a historical cost of several billions of euros, converting these unused data assets into valuable datasets that can replace tens of millions’ worth of proprietary data.

What we want to achieve with this project – and we believe such an accomplishment would merit one of the first prizes - is to add value to a significant portion of pre-existing EU open data (for example, available on [data.europa.eu/data](#)) by re-processing and integrating them into a modern, tidy database with an API access, and to find a business model that emphasises a triangular use of data in 1. business, 2. science and 3. policy-making. Our mission is to modernize the concept of **data observatories**.

OSINT & Open-Sour



MH17 Russia
Stunt Geolocation – Ver



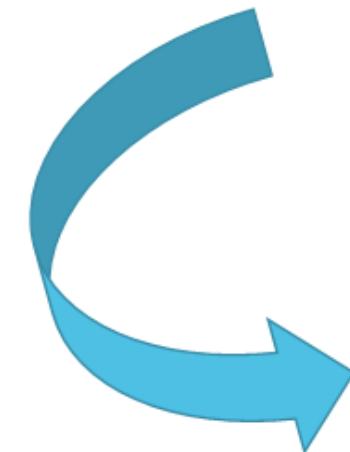
MH17 Russia
Examining the MH17 La



Investigation MH17
Investigation – War and
Investigations



MH17 Russia
Geolocating the June Ru
Millerovo



Creating b datasets

Problem definition

In the past decade interest for the "Cultural and Creative Industries" has increased significantly. This is reflected in the growth of the sector and the recognition that CCIs have a bigger and more strategic employment capacity than previously thought.

The technological changes, digital culture, and tourism, traditional areas, are driving the need for creative value creation. The demand for new data maps, survey harmonization and creation of common standards, business models, and business processes in the cultural and audiovisual industries require industry experts to create knowledge and innovation. Policy makers desire to amend the national GSI and employment aspects with Culture Satellite Accounts and tools.

Introduction and timeline

The historic setting of the United Nations General Assembly in New York City, USA, in September 2015, was the starting point for the UN Sustainable Development Goals (SDGs). The SDGs are a set of 17 global goals designed to end poverty, protect the planet, and ensure prosperity for all. The SDGs are integrated into the 2030 Agenda for Sustainable Development, which also calls for action by all countries, in a global partnership.

The European Commission has developed a set of indicators to monitor progress towards the SDGs. These indicators are based on the UN Sustainable Development Goals and the European Union's own priorities. They are used to track progress across all EU member states and regions, and to inform policy making at both the European and national levels.

Contact

Daniel Añel, CIIA [el.iisweb.es/ciiabiblioteca/](http://iisweb.es/ciiabiblioteca/)
Visgraat Investments - Candidate Partners - CIIA@IIS
Email: daniel.anel@iisweb.com
Mobile: +34 653 100 000
Phone: +34 91 891 2271, +34 91 510 0504

1.1 How We Add Value to Open Data

While the EU announces every year that again billions and billions of worth data became “open” again, this is not gold. At least not in the form of nicely minted gold coins, but in gold dust and nuggets found in the muddy banks of chilly rivers. There is no rush for it, because panning out its value requires plenty of hard work. We are trying to automate this work to make open data useable at scale, even in trustworthy AI solutions.



- retroharmonize
- regions
- lotables
- R/Phython



Our retroharmonize, regions and lotables software has been developed by the community.
The users are potential collaborators to pan out more opportunities.

Most open data is not public, it is not downloadable from the Internet – in the EU parlance, “open” only means a legal entitlement to get access to it. And even in the rare cases when data is open and public, often it is mired by data quality issues. We are working on the prototypes of a data-as-service and research-as-service built with open-source statistical software that taps into various and often neglected open data sources.

We are in a prototype phase in June, but we hope that we will have a well-functioning service by the time of the conference, because we are working only with open-source software elements; our technological readiness level is already very high. The novelty of our process is that we are trying to further develop and integrate a few open-source technology items into a technologically and financially sustainable data-as-service and even research-as-service solutions.

We decided to take a new, modern approach to the ‘data observatory’ concept, and modernize it with the application of 21st century data and metadata standards, the new results of reproducible research and data science. See [??](#). [Service Design and Business Case Development](#)

1.2 Reproducible Research

1. **Our curatotrs** help finding the best available information source. This is often open data, which is not equal to public data. Open data is a governmental or scientific data source which you can legally access. It is almost never available for direct download, and requires much processing. You could probably do this in Excel – if the data was not in various `sql`, `pdf`, `sav`, `csv`, `tsv`, `xls` and various other file formats.
2. **We process the data:** Yes, anybody can convert from millions of euros to euros in a spreadsheet, tons to kilograms, maybe even ounces to grams, but many unit conversions are error-prone if done by humans. Not everybody can make valid currency translations (*When do I use year-end USD/EUR rate? Today's EUR/GBP? Or GBP/EUR? Annual average rate?*) We do this processing in a way that conforms to the tidy data definition, which allows easy integration, joining, and importing of data into your database. While the unit conversions can be automated in Excel or PowerBI, the data tidying requires a more programmatic approach.
3. **Quality control:** Our data goes through dozens of computer logical checks (*Do assets and liabilities match? Do dollar and euro amounts lead to the same result?*) Our algorithms go through automated and human statistical peer-review, and are open to your experts for checking, because transparency and openness allow for the best quality control. This unit testing is not really possible in Excel or Power BI, not to mention the senior supervision of such tasks. To maintain data integrity, we place an authoritative copy with a digital object identifier in the Zenodo scientific data repository. We place both our algorithms and our methods into peer-reviewed scientific publications, and our data products are checked and improved by competent experts in the field.
4. **We produce** the data and its visualization in easy to reuse, machine-readable, platform-independent formats. Just like that, `csv`, `json`, `jpg`, `png`, `doxc`, `epub`, `pdf`, `pptx`, `odt`, `sav`, you name it, we do it.

Reproducible research is a scientific concept that can be applied to a wide range of professional designations, such as accounting, finance or the legal profession. We are applying this concept to [Evidence-based](#), [Open Policy Analysis](#) and [Professional Standards in Business](#), including, for example, reproducible finance in the investment process or reproducible impact assessment in policy consulting. Based on computational reproducibility we believe that the following principles should be followed: - **Reviewability** means that our application’s results are can be assessed and judged by our user’s experts, or experts they trust. - **Reproducibility** means that we provide data products and tools that allow the exact duplication of our results during assessments. This ensures that all logical steps can be verified. - **Confirmability** means that using our applications findings leads to the same professional results as other available software and information. Our data products use the open-source statistical programming language R. We provide details about our algorithms and methodology to confirm our results in SPSS or Stata or sometimes even

in Excel. - **Auditability** means that our data and software is archived in a way that external auditors can later review, reproduce and confirm our findings. This is a *stricter form of data retention* than most organizations apply, because we do not only archive results step-by-step but all computational steps – as if your colleagues would not only save every step in Excel but also their keystrokes. [Read more about this topic here.](#)

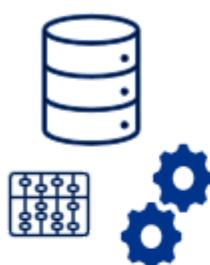
Reduce data costs by relying on
data scientists on demand

Curating



1. Public data: "... you download it but..."
2. Open data: "legally open but you cannot download it"
3. Big data: "you don't have the infrastructure to record it."
4. Surveys: "they do not match with your other data"
5. Proprietary Nielsen. "does not match your own data".
6. Own database

Processing



1. Correct download and ingestion
2. Approximation, imputation
3. Unit and currency conversions
4. Tidy format to for easy join, integration and database import.

Quality control



1. Unit tests for balance sheet equations, trivial errors.
2. Peer-reviewed algorithms
3. Authoritative copy with DOI (integrity)
4. Peer-reviewed scientific testing
5. Competent, external curators.
6. Feedback users

Presentation



1. JSON, CSV, SQL accessible in API
2. Charts and maps
3. Commentary from curators
4. Human readable documentation
5. Metadata

1.2.1 Reproducible Research

Reproducible research is a scientific concept that can be applied to a wide range of professional designations. We are applying this concept to [Evidence-based, Open Policy Analysis](#) and [Professional Standards in Business](#), for example, reproducible finance in the investment process or reproducible impact assessment in policy consulting. Based on the computational reproducibility we believe that the following principles should be followed.

- **Reviewability** means that our application's results are can be assessed and judged by our user's experts, or experts they trust. We help reviewability with full transparency: we publish the software code that created the indicators, our methodology, and an automatically refreshing statistical description of the indicator each day when it receives new data or corrections from the original source.
- **Reproducibility** means that we are providing data products and tools that allow the exact duplication of our results during assessments. This ensures that all logical steps can be verified. Reproducibility ensures that there is no lock-in to our applications. You can always chose a different data and software vendor, or compare our results with them.
- **Confirmability** means that using our applications findings leads to the same professional results as other available software and information. Our data products use the open-source statistical programming language R. We provide details about our algorithms and methodology to confirm our results in SPSS or Stata or sometimes even in Excel.
- **Auditability** means that our data and software is archived in a way that external auditors can later review, reproduce and confirm our findings. This is a *stricter form of data retention* that most organizations apply, because we do not only archive results step-by-step but all computational steps – as if your colleagues would not only save every step in Excel but also their keystrokes. While auditability is a requirement in accounting, we are extending this approach to all the quantitative work of a professional organization in an advisory or consulting capacity.
- **Reviewable findings:** The descriptions of the methods can be independently assessed, and the results judged credible. In our view, this is a fundamental requirement for all professional applications. CEEMID's music data is used to settle royalty disputes in judicial procedures, or in grant and policy design. We believe that the future European Music Observatory should aim at the same bar, making its data & research products open for challenges in the publicity of science, courts, and professional peers.
- **Replicable findings:** We are presenting our findings and provide tools so that our users or auditors or external authorities can duplicate our results.
- **Confirmable findings:** The main conclusions of the research can be obtained independently without our software, because we describe in detail the algorithms and methodology in supplementary materials. We believe that other organizations, analysts, statisticians must come to the same findings with their own methods and software. This avoids lock-in and allows independent cross-examination.
- **Auditable findings:** Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks. See [Open collaboration](#) with academia, auditors, and industry.

These computational requirements require a data workflow that relies on further principles.

- **Record retention:** all aspects of reproducibility require a high level of standardized documentation. The standardization of documentation requires the use of standardized metadata, metadata structures, taxonomies, vocabularies.
- **Best available information / data universe:** the quality of the findings, their confirmation

and auditing success will improve with better data and facts used.

- **Data validations:** The quality of the findings will greatly depend on the factual inputs. While the reproducible findings may have many problems, inputting erroneous data or faulty information will likely lead to wrong conclusions, and in all cases will make confirmation and auditing impossible. Especially when organizations use large and heterogeneous data sources, even small errors, such as erroneous currency translations or accidental misuse of decimals, units can cause results that will not pass confirmation or auditing.

Chapter 2

FAIR Data and the Added Value of Rich Metadata

FAIR data is far more valuable than non-fair data.

We are making our API service FAIR by automatically adding to our data standardized metadata that fulfills the mandatory requirements of the Public Core metadata standards and at the same time the [mandatory requirements](#), and most of the [recommended requirements](#) of DataCite.

2.1 FAIR

In 2016, the [FAIR Guiding Principles for scientific data management and stewardship](#) were published in *Scientific Data*. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

A practical “how to” guidance to go FAIR can be found in the [Three-point FAIRification Framework](#).

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the [FAIRification process](#).

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he/they need to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

2.2 The Dublin Core

Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Description	An account of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Language	A language of the resource.
Publisher	An entity responsible for making the resource available.
Relation	A related resource.
Rights	Information about rights held in and over the resource.
Source	A related resource from which the described resource is derived.

Subject	The topic of the resource.
Title	A name given to the resource.
Type	The nature or genre of the resource.

2.3 DataCite

We use all [mandatory](#) DataCite metadata fields, and many of [the recommended and optional](#) ones.

Identifier	An unambiguous reference to the resource within a given context. (Dublin Core item), but several identifiers allowed, and we will use several of them.
Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order. To supply multiple creators, repeat this property. (Extends the Dublin Core with multiple authors, and legal persons, and adds affiliation data.)
Title	A name given to the resource. Extends Dublin Core with alternative title subtitle, translated Title, and other title(s)
Publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role. For software, use Publisher for the code repository. (Dublin Core item.)
Publication Year	The year when the data was or will be made publicly available.
Resource Type	We publish Datasets, Images, Report, and Data Papers. (Dublin Core item with controlled vocabulary.)

2.3.1 Recommended

Subject	The topic of the resource. (Dublin Core item.)
Contributor	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource. (Extends the Dublin Core with multiple authors, and legal persons, and adds affiliation data.) When applicable, we add Distributor (of the datasets and images), Contact Person, Data Collector, Data Curator, Data Manager, Hosting Institution, Producer (for images), Project Manager, Researcher, Research Group, Rightsholder, Sponsor, Supervisor

Date	A point or period of time associated with an event in the lifecycle of the resource, besides the Dublin Core minimum we add Collected, Created, Issued, Updated, and if necessary, Withdrawn dates to our datasets.
Language	A language of the resource. (Dublin Core item.)
Alternative Identifier	An identifier or identifiers other than the primary Identifier applied to the resource being registered.
Related Identifier	An identifier or identifiers other than the primary Identifier applied to the resource being registered.
Size	We give the CSV, downloadable dataset size in bytes.
Format	We give file format information. We mainly use CSV and JSON, and occasionally rds and SPSS types. (Dublin Core item.)
Version	The version number of the resource.
Rights	We give standards rights description with URLs to the actual license. (Dublin Core item.)
Description	Recommended for discovery.(Dublin Core item.)
GeoLocation	Similar to Dublin Core item Coverage
Funding Reference	We provide the funding reference information when applicable. This is usually mandatory with public funds.
Related Item	We give information about our observatory partners' related research products, awards, grants (also Dublin Core item as Relation.) We particularly include source information when the dataset is derived from another resource (which is a Dublin Core item.)

2.4 Processing Metadata

Value Type	We observe actual, missing, imputed, estimated values.
Method	If the value is estimated, we provide modelling information.
Unit	We provide the measurement unit of the data (when applicable.)
Frequency	We provide the measurement frequency of the data in a more practical format than the Dublin Core and DataCite descriptive dates.
Related Item	We give information about the software code that processed the data (both Dublin Core and DataCite compliant.)

Chapter 3

Trustworthy AI

Human agency and oversight

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches

Technical Robustness and safety

AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Privacy and data governance

besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

Transparency

the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

Diversity, non-discrimination and fairness

Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination.

Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

Societal and environmental well-being

AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.

Chapter 4

Application: Automated Data Observatories

This year's EU Datathlon includes three challenges. We are contesting all three of them with the same technology and knowledge management, but with different data used in a wide range of knowledge domains.

4.1 Data Acquisition and Processing

We access various open governmental and open scientific sources programatically. Our programs are mainly written in the R language, but we have a growing body of software written in Python, too. We thrive to be open for both R and Python developers, and as much as possible, exploit the synergies between the more statistically oriented R language and the more generel application-oriented Python. We welcome [curators](#) and [developers](#) in both languages.

An important aspect of our quality control is that our processing code is open and peer-reviewed. Our observatories are turning the peer-reviewed statistical software of the [rOpenGov](#) community into a continous data-as-service product. This means that we are creating collector software that is making reproducible data assets, mainly business and policy indicators. Then we are running this software daily in the cloud, and place the new data acquisitions' [authoritative copies](#) into a scientific repository for data integrity purposes, then upload it to our [API](#), describe it in our [long form documentation](#), and eventually blog about the newsworthy finds on our [Front-end Websites](#).

The entire research automation system is maintained by [Reprex](#), a Dutch research automation startup, in open collaboration with [rOpenGov](#) and other [developers](#).

4.2 Data Integrity: Authoritative Copies

We rely on a data repository to keep a final, authoritative copy of our data assets and visualizations. This repository is independent from us.

Zenodo is a semi-endorsed EU solution, originating from CERN. In the last EU budget period all EU-funded research was supposed to deposit data there, although this requirement was not often enforced. Manual deposition is in working order, and we can very easily retrieve our own data in various versions. It is also free data storage.

The Zenodo API is not very well documented, particularly for R. But it is supported both in Python and R. We have a tutorial and code on how to deposit our assets programatically via the [Zen4R](#) package. It is a bit difficult to use - it mimics "true" object oriented programming relying on R6 classes, which is extremely rarely used by R programmers.

The Dataverse is much better served, the API is better documented, and technically we could even set up our own instance (new dataverses can be installed.) The best instance is of course the original Harvard Dataverse. Currently Dataverse has no support on CRAN - the R package was just kicked out of CRAN, and it is buggy, but it can be fixed. Should there be a need, we can make a connector to Dataverse, too.

4.3 Automated Data Observatory API

Our observatories APIs are [Datasette](#) instances. It is a lightweight, Python-based application that turns a SQLite database into a powerful API. Our developer, [Botond Vitos](#) manages our APIs.

The indicator table contains the actual values, and the various estimated/imputed values of the indicator, clearly marking missing values, too.

Indicator																																																																																								
Data license: ODbL · Data source: Economy Data Observatory																																																																																								
22,066 rows																																																																																								
<input type="button" value="- column -"/> <input type="button" value="="/> <input type="text"/>																																																																																								
<input type="button" value="Apply"/>																																																																																								
View and edit SQL																																																																																								
This data as json, CSV (advanced)																																																																																								
Suggested facets: unit , time , estimate , method																																																																																								
<table border="1"> <thead> <tr><th>Link</th><th>rowid</th><th>shortcode</th><th>unit</th><th>time</th><th>frequency</th><th>geo</th><th>value</th><th>estimate</th><th>method</th></tr> </thead> <tbody> <tr><td>1</td><td>1</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>AT</td><td></td><td>missing</td><td>missing</td></tr> <tr><td>2</td><td>2</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>BE</td><td>1.0</td><td>actual</td><td>actual</td></tr> <tr><td>3</td><td>3</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>BG</td><td></td><td>missing</td><td>missing</td></tr> <tr><td>4</td><td>4</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>CH</td><td>1.0</td><td>actual</td><td>actual</td></tr> <tr><td>5</td><td>5</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>CZ</td><td></td><td>missing</td><td>missing</td></tr> <tr><td>6</td><td>6</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>DE</td><td>56.0</td><td>actual</td><td>actual</td></tr> <tr><td>7</td><td>7</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>NR</td><td>7670.0</td><td>A</td><td>DK</td><td></td><td>missing</td><td>missing</td></tr> </tbody> </table>									Link	rowid	shortcode	unit	time	frequency	geo	value	estimate	method	1	1	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	AT		missing	missing	2	2	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	BE	1.0	actual	actual	3	3	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	BG		missing	missing	4	4	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	CH	1.0	actual	actual	5	5	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	CZ		missing	missing	6	6	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	DE	56.0	actual	actual	7	7	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	DK		missing	missing
Link	rowid	shortcode	unit	time	frequency	geo	value	estimate	method																																																																															
1	1	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	AT		missing	missing																																																																															
2	2	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	BE	1.0	actual	actual																																																																															
3	3	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	BG		missing	missing																																																																															
4	4	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	CH	1.0	actual	actual																																																																															
5	5	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	CZ		missing	missing																																																																															
6	6	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	DE	56.0	actual	actual																																																																															
7	7	eurostat_pat_ep_rtec_avi_nr	NR	7670.0	A	DK		missing	missing																																																																															

The descriptive metadata is contained in the **description** tables.

Description																											
Data license: ODbL · Data source: Economy Data Observatory																											
32 rows																											
<input type="button" value="- column -"/> <input type="button" value="="/> <input type="text"/>																											
<input type="button" value="Apply"/>																											
View and edit SQL																											
This data as json, CSV (advanced)																											
<table border="1"> <thead> <tr><th>Link</th><th>rowid</th><th>shortcode</th><th>description</th><th>keyword_1</th><th>keyword_2</th><th>keyword_3</th></tr> </thead> <tbody> <tr><td>1</td><td>1</td><td>eurostat_pat_ep_rtec_avi_nr</td><td>High tech patent applications to the epo by priority year by nuts 3 regions aviation number</td><td>ipr</td><td>supply</td><td>rd</td></tr> <tr><td>2</td><td>2</td><td>eurostat_pat_ep_rtec_avi_p_mhab</td><td>High tech patent applications to the epo by priority year by</td><td>ipr</td><td>supply</td><td>rd</td></tr> </tbody> </table>							Link	rowid	shortcode	description	keyword_1	keyword_2	keyword_3	1	1	eurostat_pat_ep_rtec_avi_nr	High tech patent applications to the epo by priority year by nuts 3 regions aviation number	ipr	supply	rd	2	2	eurostat_pat_ep_rtec_avi_p_mhab	High tech patent applications to the epo by priority year by	ipr	supply	rd
Link	rowid	shortcode	description	keyword_1	keyword_2	keyword_3																					
1	1	eurostat_pat_ep_rtec_avi_nr	High tech patent applications to the epo by priority year by nuts 3 regions aviation number	ipr	supply	rd																					
2	2	eurostat_pat_ep_rtec_avi_p_mhab	High tech patent applications to the epo by priority year by	ipr	supply	rd																					

The data transactional and processing metadata is contained in the **metadata** tables.

home / database

metadata

Data license: [ODbL](#) · Data source: [Economy Data Observatory](#)

32 rows

Link	rowid	shortcode	description_at_source	last_update_data	last_update_at_source
1	1	eurostat_pat_ep_rtec_avi_nr	High tech patent applications to the epo by priority year by nuts 3 regions aviation number	18778.0	16920.0
2	2	eurostat_pat_ep_rtec_avi_p_mhab	High tech patent applications to the epo by priority year by nuts 3 regions aviation number	18778.0	16920.0

The variable labelling and unit labelling information is stored in the **labelling** tables.

home / database

labelling

Data license: [ODbL](#) · Data source: [Economy Data Observatory](#)

64 rows

Link	rowid	shortcode	var_name	var_code	var_label
1	1	eurostat_pat_ep_rtec_avi_nr	unit	NR	[Number]
2	2	eurostat_pat_ep_rtec_avi_p_mhab	unit	P_MHAB	[Per million inhabitants]
3	3	eurostat_pat_ep_rtec_cab_nr	unit	NR	[Number]
4	4	eurostat_pat_ep_rtec_cab_p_mhab	unit	P_MHAB	[Per million inhabitants]
5	5	eurostat_pat_ep_rtec_cte_nr	unit	NR	[Number]
6	6	eurostat_pat_ep_rtec_cte_p_mhab	unit	P_MHAB	[Per million inhabitants]
7	7	eurostat_pat_ep_rtec_ht_nr	unit	NR	[Number]

Currently our APIs are re-freshed by an R code. We will soon add a Python collector, too.

4.4 Long form documentation

Our long-form documentation is based on [bookdown](#), which relies on pandoc, rmarkdown and knitr. This handbook is also created in bookdown.



It is a very mature workflow, it produces a long-form website, and PDF, ePUB or Word versions from the API. The current automation is not operational, as we have recently included the API.

4.5 Front-End Websites

Our observatory's client-facing front end is made by the static website generator hugo, which is programmed in the go language. We use the open-source [Starter Hugo Academic](#) template of [wowchemy](#). If we win a price, we'll certainly offer them a share!

The screenshot shows a web browser window with the URL music.dataobservatory.eu/#usecase. The page title is "Digital Music Observatory". On the right side, there is a vertical sidebar with links: "Home" (highlighted in blue), "Data & Lyrics", and "Blog". The main content area features a large teal circular logo with the letters "DMO" in white. Below the logo, the text "Digital Music Observatory" is displayed in a large, black, serif font. Underneath that, the word "co-founder" is followed by a link to "Reprex BV". To the right of the main content, there is a large block of text that is partially cut off on the right edge.

A
The
obs
the
indi
poli
ente
Its
esta
The
1.

Home Data &
Lyrics
Blog

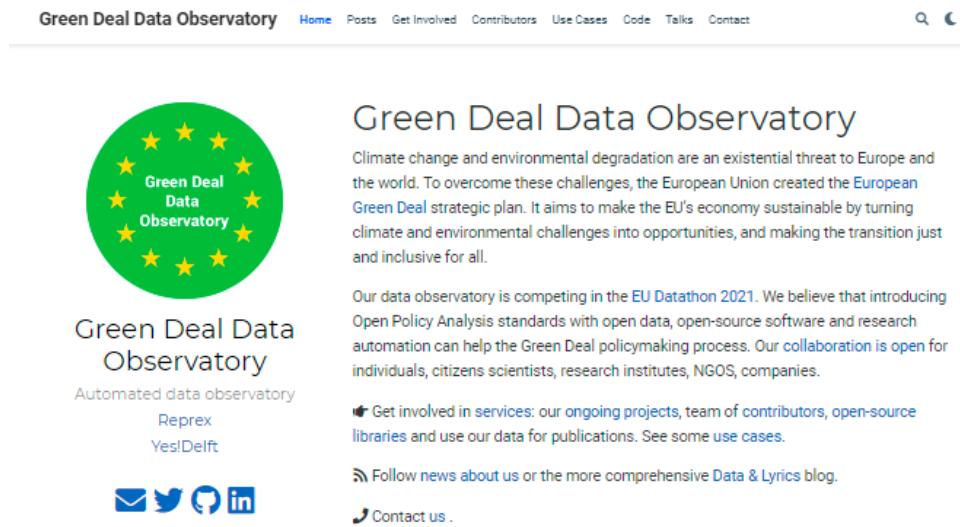
DMO

Digital Music Observatory

co-founder

Reprex BV

The hugo-bookdown integration is partly supported by the R package [blogdown](#). This is a semi-success, and while academic is a super-popular template, it is getting further and further away from blogdown. The original advantage is that it can be managed in the same workflow as the indicator generation, package documentation, the long-form documentation is a bit gone.



The screenshot shows the homepage of the Green Deal Data Observatory. At the top, there's a navigation bar with links to Home, Posts, Get Involved, Contributors, Use Cases, Code, Talks, and Contact. To the right of the navigation are a search icon and a user icon. Below the navigation is a large green circular logo with yellow stars containing the text "Green Deal Data Observatory". To the right of the logo, the page title "Green Deal Data Observatory" is displayed in a large font. A brief introduction follows: "Climate change and environmental degradation are an existential threat to Europe and the world. To overcome these challenges, the European Union created the European Green Deal strategic plan. It aims to make the EU's economy sustainable by turning climate and environmental challenges into opportunities, and making the transition just and inclusive for all." Below this, a section titled "How Does It Work?" is shown, featuring six icons representing different processes: Open Data (sun icon), R (R logo icon), Statistics (line graph icon), Geocoding (keyhole icon), Maps (globe icon), and Documentation (file icon). Each process has a brief description below its icon. At the bottom left, there's a section for "Ongoing projects" with a "Make Coal History" button. At the bottom right, there's a small navigation bar with icons for back, forward, and search.

4.6 Research Automation

Because every step of our data acquisition, processing, structuring, and testing is going through machines, it can be replicated any given year, month, day, or hour. Research automation means that we update our data every day (Is there new data at the source? Corrections? Known issues?) and place the current version in an API.

- 1. Continous data collection:** Big data sources usually provide the user with a large quantity of insignificant data. Because of the large quantity, the data is usually not available historically:

you capture it today or it is gone. You need to process data in big quantities in order to find significant and meaningful information. Most small enterprises and research teams do not have the engineering capacity to organize this. We do continuous data collection on all sources to capture the latest updates, or corrections.

2. **Focus on reusability:** In our experience, the reusability of research and consulting work is greatly reduced by two factors, which we resolve with continuous data collection and documentation:
 - poor documentation (the bibliography updates and file descriptions are the least prioritized tasks)
 - data tables, charts, visualizations become dated then obsolete.
3. From tidy and open data to **data-as-service**: Because all our assets, including key business indicators, policy indicators, scientific replication sets, and their visualizations, as well as maps, are created by open source and reproducible software, the software can run continuously. Instead of providing our users with data tables, charts and maps, we provide them with a subscription to the latest data and its latest visualizations.

Chapter 5

Service Design and Business Case Development

We are taking a new and modern approach to the **data observatory** concept, and modernizing it with the application of 21st century data and metadata standards, the new results of reproducible research and data science. 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, but even these do not use these organizations and their members open data. We are building open-source data observatories, which run open-source statistical software that automatically processes and documents reusable public sector data (from public transport, meteorology, tax offices, taxpayer funded satellite systems, etc.) and reusable scientific data (from EU taxpayer funded research) into new, high quality statistical indicators.

We are building open-source data observatories, which run open-source statistical software that automatically processes and documents reusable public sector data (from public transport, meteorology, tax offices, taxpayer funded satellite systems, etc.) and reusable scientific data (from EU taxpayer funded research) into new, high quality statistical indicators.

We are building various open-source data collection tools in R and Python to bring up data from big data APIs and legally open, but not public, and not well served data sources. For example, we are working on capturing representative data from the Spotify API or creating harmonized datasets from the Eurobarometer and Afrobarometer survey programs.

- Open data is usually not public; whatever is legally accessible is usually not ready to use for commercial or scientific purposes. In Europe, almost all taxpayer funded data is legally open for reuse, but it is usually stored in heterogeneous formats, processed into an original government or scientific need, and with various and low documentation standards. Our expert data curators are looking for new data sources that should be (re-) processed and re-documented to be usable for a wider community. We would like to introduce our service flow, which touches upon many important aspects of data scientist, data engineer and data curatorial work.
- We believe that even such generally trusted data sources as Eurostat often need to be reprocessed, because various legal and political constraints do not allow the common European statistical services to provide optimal quality data – for example, on the regional and city levels.
- With [rOpenGov](#) and other partners, we are creating open-source statistical software in R to reprocess these heterogenous and low-quality data into tidy statistical indicators to automatically validate and document it.
- We are carefully documenting and releasing administrative, processing, and descriptive metadata, following international metadata standards, to make our data easy to find and easy to use for data analysts.
- We are automatically creating depositions and authoritative copies marked with an individual digital object identifier (DOI) to maintain data integrity.

- We are building simple databases and supporting APIs that release the data without restrictions, in a tidy format that is easy to join with other data, or easy to join into databases, together with standardized metadata.
- We maintain observatory websites (see: [Digital Music Observatory](#), [Green Deal Data Observatory](#), [Economy Data Observatory](#)) where not only the data is available, but we provide tutorials and use cases to make it easier to use them. Our mission is to show a modern, 21st century reimagination of the data observatory concept developed and supported by the UN, EU and OECD, and we want to show that modern reproducible research and open data could make the existing 60 data observatories and the planned new ones grow faster into data ecosystems.

We are working around the open collaboration concept, which is well-known in open source software development and reproducible science, but we try to make this agile project management methodology more inclusive, and include data curators, and various institutional partners into this approach. Based around our early-stage startup, Reprex, and the open-source developer community rOpenGov, we are working together with other developers, data scientists, and domain specific data experts in climate change and mitigation, antitrust and innovation policies, and various aspects of the music and film industry.

Our open collaboration is truly open: new [data curators](#), data scientists and data engineers are welcome to join. We develop open-source software in an agile way, so you can join in with an intermediate programming skill to build unit tests or add new functionality, and if you are a beginner, you can start with documentation and testing our tutorials. For business, policy, and scientific data analysts, we provide unexploited, exciting new datasets. Advanced developers can [join](#) our development team: the statistical data creation is mainly made in the R language, and the service infrastructure in Python and Go components.

See or share this introduction to the service plans in a [blogpost](#).

5.1 Professional Standards in Business

5.1.1 Key Business Indicators

5.1.2 Business Record Retention

Some of the requirements of reproducible research are usually required by professional standards. For example, various accounting, finance, legal or consulting professional standards call for appropriate documentation and record retention.

5.2 Professional Standards in Policy

5.2.1 Evidence-based, Open Policy Analysis

In the last two decades, governments and researchers have placed a growing emphasis on the value of evidence-based policy. However, while the evidence generated through research to inform policy has become more rigorous and transparent, policy analysis—the process of contextualizing evidence to inform specific policy decisions—remains opaque.

We believe that a modern data observatory must improve how evidence is created and used in policy reports, and pass on the efficiency gains from increasing reproducibility and automation. Therefore, we pledge that the [music.dataobservatory.eu](#) will comply with the [Open Policy Analysis](#) standards developed by the [Berkeley Initiative for Transparency in the Social Sciences](#) & [Center for Effective Global Action](#). These standards are applied by the World Bank.

5.3 Contributors

We are looking for business associates to help our service design into our

Contributors of the Digital Music Observatory

Join our open collaboration team as a [data curator](#), [developer](#) or [business developer](#)! More about contributing:
[Automated Observatory Contributors' Handbook](#).

developers



Daniel Antal
Data Scientist & Founder of
the Digital Music
Observatory

Pyry Kantanen
R package testing and data
curator

Leo Lahti
OpenGov coordinator

Botond Vitos
Data scientist and
developer

**Andrés García
Molina, PhD**
Data Scientist &
Ethnomusicologist



Kasia Kulma
Contributor, data science
and software engineering

data curators



**Dominika
Šemánková**
Musicologist

Eszter Kabai
Data curator for cultural
diversity and data pooling

Caterina Sganga
Data curator for cultural
diversity and data pooling

Hyojung Sun
Data curator for music
creators' earnings

Katie Long
Music futures and social
equity data curator



Peter Ormosi
Music Economy &
Information Studies researcher

**Prof David
Hesmondhalgh**
Professor of Physics

Stef Koenis
Data curator for classical
music and data engineer

Stephan Okhuijsen
Data visualization and
information design researcher

Figure 5.1: Our open collaboration is truly open: new data curators, data scientists and data engineers are welcome to join in.

- [Green Deal Data Observatory](#)
 - [Economy Data Observatory](#)
 - [Digital Music Observatory](#)
1. Work with governmental or scientific or otherwise [open data](#).
 2. Committed to high policy or business professional standards, and by making their work [reproducible](#), they adhere to reviewability, reproducibility, confirmability and auditability, regardless if they work, or study for various professional roles in business, academia, public or non-governmental policy, and data journalism.

An important aspect of the EU Datathon Challenges is “.. to propose the development of an application that links and uses open datasets [...] to find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of open data.”

Where to find us: - [dataobservatory-eu](#) is our private repo collection and private github collaboration platform, but many of our repos are open. Like this one.

- [Creative Data Observatories LinkedIn Page](#). Make sure you follow us, and spread our messages.
- [twitter.com/dataandlyrics](#) is our twitter handle for our music-oriented blog. If you are on twitter, please follow us, and retweet our blogposts.
- [keybase.io/team/reprexcommunity](#) is our landing page to our otherwise private and invisible internal communication platform. (See more in subchapter?? of [Tools](#).)

You find more information in the ?? chapter about our topics: [What is Open Data?](#), [Reproducible Research](#), and of course, [Get Inspired About Data](#).

5.4 Passion About Our Topic

- You are passionate about one of our topics, but you do not feel that you have the skills yet to become a data curator or a developer.
- You have a curiosity in the field of economic policies, particularly in computational antitrust, innovation research, and understanding the statistically under-represented micro- and small enterprises, join our [Economy Data Observatory](#) as a volunteer.
- You are passionate about environmental research of climate change, designing urban, social and economic mitigation strategies, or understanding how people think about climate change, join our [Green Deal Data Observatory](#) team as a volunteer.
- You want to know how musicians can make a living after the pandemic? How can we make sure that music made by womxn, small country artists or artists of color gets an equal chance? Are you interested in the future of ethical AI and data governance? Join our [Digital Music Observatory](#) team as a volunteer.

5.5 Passion For Trustworthy AI, Open Data and Open Source Projects

- You want to learn to write scientifically valid, open source code in R or Python, but you are a beginner. We help you anywhere - you can even start copyediting or technical documentation (it is a must in open source development) and create tutorials for you to interact with our data and products (if it helps you, it will help others.)

- As a business economist or legal professional, you are interested how open data, open source software, research automation and ethical, trustworthy AI products can find their market.
- As a blogger, data journalist or marketeer you would like to help to make open data, and transparent, ethical, open AI more widely known and used.

5.6 Technical Requirements

- Make sure that you read the [Contributors Covenant](#). You must make this [pledge](#) to make participation in our community a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, caste, color, religion, or sexual identity and orientation. Participating in our data observatories requires everybody to act and interact in ways that contribute to an open, welcoming, diverse, inclusive, and healthy community. It's better this way for you and for us!
- Give users at least one social media account where they can get in touch with you (any of LinkedIn, Twitter, Academia, SSRN, Google Scholar, or even Facebook.)
- Please, follow us on social media, it really helps us finding new users and showing that we are able to grow our ecosystem.
 - [Green Deal Data Observatory on Linkedin](#) and [Green Deal Data Observatory on Twitter](#)
 - [Economy Data Observatory on Linkedin](#) and [Economy Data Observatory on Twitter](#)
 - [Digital Music Data Observatory on Linkedin](#) and [Digital Music Data Observatory on Twitter](#)
- Please send us back this [md file](#) with your data. You can open it with any text editor, but Notepad,TextEdit, Vim and similar clean text editors are the best.
- If you feel you need chatting on onboarding, contact us on [Keybase](#) - it's lightweight, discrete, encrypted, your mother, partner and friends are not there, it is free, open source, and can share/exchange files, too. Otherwise in email.

Chapter 6

Data Curators

We are looking for data curators into our

- [Green Deal Data Observatory](#)
 - [Economy Data Observatory](#)
 - [Digital Music Observatory](#)
1. Work with governmental or scientific or otherwise [open data](#).
 2. Committed to high policy or business professional standards, and by making their work [reproducible](#), they adhere to reviewability, reproducability, confirmability and auditability, regardless if they work, or study for various professsional roles in business, academia, public or non-governmental policy, and data journalism.
 3. They are interested in helping us with [indicator design](#).
 4. Make the authoritative copy of their indicator available on the Zenodo data repository, and keep it up-to-date with our automated observatory's help.

An important aspect of the EU Datathon Challenges is “.. to propose the development of an application that links and uses open datasets [...] to find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of open data.”

Where to find us: - [dataobservatory-eu](#) is our private repo collection and private github collaboration platform, but many of our repos are open. Like this one.

- [Creative Data Observatories LinkedIn Page](#). Make sure you follow us, and spread our messages.
- [twitter.com/dataandlyrics](#) is our twitter handle for our music-oriented blog. If you are on twitter, please follow us, and retweet our blogposts.
- [keybase.io/team/reprexcommunity](#) is our landing page to our otherwise private and invisible internal communication platform. (See more in subchapter?? of [Tools](#).)

6.1 Get Inspired

See our first curatorial interviews:

- Economy Data Observatory: [New Indicators for Computational Antitrust](#)
- Digital Music Observatory: [New Indicators for Royalty Pricing and Music Antitrust](#)

6.1.1 Create New Datasets

Our mission is to create standardized data about a social, economic or environmental process that does not have standardized, well-processed open data. To be a data curator, you must have a passion for a topic where evidence is hard to find, and you must have the knowledge to assess that the evidence we find in hidden data sources is valid or not.

- [This Scientist Stung Himself With Dozens Of Insects Because No One Else Would:](#) The Schmidt Pain Index, as its informally known, runs from 1-4. The common honey bee serves as its anchor point, a solid 2. At the top end of the scale lie the bullet ant and the tarantula hawk (which is neither a tarantula nor a hawk; it's a wasp). Watch the video with [Dr. Schmidt](#), and listen to the whole interview [here](#).
- [Big Data Is Saving This Little Bird](#) “We need to improve conservation by improving wildlife monitoring. Counting plants and animals is really tricky business.”

6.1.2 Remain Critical

Sometimes we put our hands on data that looks like a unique starting point to create a new indicator. But our indicator will be flawed, if the original dataset is flawed. And it can be flawed in many ways, most likely that some important aspect of the information was omitted, or the data is autoselected, for example, under-sampling women, people of color, or observations from small or less developed countries.

- Cathy O’Neil: [Weapons of math destruction](#), which O’Neil are mathematical models or algorithms that claim to quantify important traits: teacher quality, recidivism risk, creditworthiness but have harmful outcomes and often reinforce inequality, keeping the poor poor and the rich rich. They have three things in common: opacity, scale, and damage. <https://blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/>
- Catherine D’Ignazio and Lauren F. Klein: [Data Feminism](#). This is a much celebrated book, and with a good reason. It views AI and data problems with a feminist point of view, but the examples and the toolbox can be easily imagined for small-country biases, racial, ethnic, or small enterprise problems. A very good introduction to the injustice of big data and the fight for a more fair use of data, and how bad data collection practices through garbe in garbe out lead to misleading information, or even misinformation.
- [Why The Bronx Burned](#). Between 1970 and 1980, seven census tracts in the Bronx lost more than 97 percent of their buildings to fire and abandonment. In his book [The Fires](#), Joe Flood lays the blame on misguided “best and brightest” effort by New York City to increase government efficiency. With the help of the Rand Corp., the city tried to measure fire response times, identify redundancies in service, and close or re-allocate fire stations accordingly. What resulted, though, was a perfect storm of bad data: The methodology was flawed, the analysis was rife with biases, and the results were interpreted in a way that stacked the deck against poorer neighborhoods. The slower response times allowed smaller fires to rage uncontrolled in the city’s most vulnerable communities. Listen to the podcast [here](#)
- [Bad Incentives Are Blocking Better Science](#) “There’s a difference between an answer and a result. But all the incentives are pointing toward telling you that as soon as you get a result, you stop.” After the deluge of retractions, the stories of fraudsters, the false positives, and the high-profile failures to replicate landmark studies, some people have begun to ask: “[Is science broken?](#)”. Listen to the pdodcast [Science is Hard](#) [here](#)
- In [Algorithms of Oppression](#), Safiya Umoja Noble challenges the idea that search engines like Google offer an equal playing field for all forms of ideas, identities, and activities. Data discrimination is a real social problem; Noble argues that the combination of private interests in promoting certain sites, along with the monopoly status of a relatively small number of Internet search engines, leads

to a biased set of search algorithms that privilege whiteness and discriminate against people of color, specifically women of color.

- Christopher Ingraham wrote [a quick blog post](#) for The Washington Post about an obscure USDA data set called the **natural amenities index**, which attempts to quantify the natural beauty of different parts of the country. He described the rankings, noted the counties at the top and bottom, hit publish and didn't think much of it. Almost immediately he started to hear from the residents of northern Minnesota, who were not very happy that Chris had written, "the absolute worst place to live in America is (drumroll, please) ... Red Lake County, Minn." He could not have been more wrong ... a year later [he moved](#) to Red Lake County with his family.

6.1.3 Your First Data Contribution

Your first contribution can be made without writing a single program code – but if you are experienced in reproducible science, than you can also submit a code that creates your data.

1. Make sure that you read the [Contributors Covenant](#). You must make this [pledge](#) to make participation in our community a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, caste, color, religion, or sexual identity and orientation. Participating in our data observatories requires everybody to act and interact in ways that contribute to an open, welcoming, diverse, inclusive, and healthy community. It's better this way for you and for us!
2. Send us a plain language document, preferably in any flavor of markdown (See subchapter ?? in the [Tools](#)), or even in a clear text email about the indicator. What should the indicator be used for, how it should be measured, with what frequency, and what could be the open data source to acquire the observations. Experienced data scientists can send us a Jupiter Notebook or an Rmarkdown file with code, but this submission can be a simple plain language document without numbers.
3. Make sure that you have and [ORCID ID](#). This is a standard identification for scientific publications. We need your numeric ORCID ID.
4. Make sure that you have a [Zenodo account](#) which is connected to your [ORCID ID](#). This enables you to publish data under your name. If you curate data for our observatories, you will be the indicator's first author, and depending on what processes help you, the author of the (scientific) code that helps you calculate the values will be your co-author.
5. Please, follow us on social media, it really helps us finding new users and showing that we are able to grow our ecosystem.
 - [Green Deal Data Observatory on Linkedin](#) and [Green Deal Data Observatory on Twitter](#)
 - [Economy Data Observatory on Linkedin](#) and [Economy Data Observatory on Twitter](#)
 - [Digital Music Data Observatory on Linkedin](#) and [Digital Music Data Observatory on Twitter](#)
 - – Please send us back this [md file](#) with your data. You can open it with any text editor, but Notepad,TextEdit, Vim and similar clean text editors are the best.

You can stop here if you do not have programming experience. If you do, please go on with the next steps:

6. Without programming experience your first indicator should be uploaded manually to Zenodo, and we will help automating the new versions. This will mean, for example, the upload of a simple, csv version of an Excel table, and filling in some important information about the contents of the table.
7. With some level of R or Python programming experience, we ask you to create a Github repo where you store your indicator. We will help you with tutorials, program codes, or applications to automate your data publication on Zenodo. In this case, make sure that you also have a [Sandbox Zenodo](#)

account. There is no undo button on Zenodo. If you are tinkering with automatically publishing data, practice first in the sandbox, which is a practicing clone of Zenodo with undo button. (To avoid accidents, you need to have a completely different account with different credential on the real and the sandbox practice repository.)

8. Experienced programmers are welcome to participate in [our developer team](#), and become contributors, or eventually co-authors of the (scientific) software codes that we make to continuously improve our data observatories. All our data code is open source. At this level, you are expected to be able to raise and/or pick up and solve an issue in our observatory's Github repository, or its connecting statistical repositories.

Our data is mainly processed in R language software, and sometimes in Python language software. If you are experienced with R bookdown, R Shiny or working in the hugo language, then you are welcome to join our developer team in non-curatorial roles.

6.2 Indicator Design

We are committing ourselves in the final deliverable to follow the indicator design principles set out by Eurostat: (???) to create high-quality, validated indicators that receive appropriate feedback from users, i.e. music businesses, their trade associations and policy-makers.

What are the characteristics of a good indicator? Based on the above mentioned Eurostat expectation, we formulated it for our observatories in this way.

- **Relevance:** Indicators must ‘meet the users’ needs’; if they do not measure anything useful to policymakers, the public or researchers, they will probably not be widely used. Indicators should also be unambiguous in showing which direction is ‘desirable’.
- **Accuracy and reliability:** Indicators must ‘accurately and reliably portray reality’; an inaccurate indicator can lead to erroneous conclusions, steer the business or policy making process in the wrong direction or let negative effects go undetected.
- **Timeliness and punctuality:** Indicators must be released at a time that is relevant to the end user. If we cannot produce an accurate indicator in a timely manner, we should aim to create a leading indicator that is sooner available and with relatively high accuracy correlates with the indicator that is not available on time.
- **Coherence and comparability:** Indicators should be ‘consistent internally, over time and comparable between regions and countries. This is particularly relevant for indicators used for policy monitoring and assessment, and in international business planning and assessment.
- **Accessibility and clarity**

Examples for indicators in our [Digital Music Observatory](#):

- Indicators that were used with all known royalty valuation methods (?), for both author’s and neighbouring rights, and fullfil the [IFRS fair value](#) standards, incorporated in [EU law](#) and the recent EU jurisprudence (??).
- Indicators that can be used for calculating damages, or calculating the value of the value gap (??).
- Indicators that quantify the development needs of musicians, and can set objective granting aims and grant evaluations (?).
- Understanding how music is taxed, how music contributes to the local and national GDP, and how music creates jobs directly, indirectly and with induced effects (?).
- Providing detailed comparison of the differences of music audience among countries.
- Measuring exporting success on streaming platforms, and preparing better targeting tools.

6.2.1 Creation and Quality Control of Indicators

An indicator values are created if we the data curator has some, preferably at least 20 observation values available in data table that confirms the tidy data principles, i.e. each variable is in exactly one column of the table, and each observation is in one row of the table.

Each indicators should be described in a clear, English language text, describing the meaning of the variables, the source of the observations, and other important information about the processing, refreshing, extending of the dataset.

We are safeguarding the quality of the indicators with various reproducible research methods. Depending on the data scientific level of the curator, we either take over the quality control mechanism, or cooperate with the curator. But the main inputs for quality control should be described by the data curator.

- **Unit testing:** Unit tests are simple, numerical test that avoid logical errors in an indicator. Shall we exclude zero values? Negative values? Do percentages must add up to 100? Some of our indicators go through more than 60 unit tests. We ask your help to get us going, and we will take care of the usual suspects: wrong currency translations, wrong decimal places (thousand, million units), etc.
- **Missing data treatment:** No real life dataset is complete, but many statistical and AI methods cannot handle missing values. Therefore, we make an effort to *impute* with an estimated value the missing values. Imputation is sometimes self-understanding, but sometimes it is a very tricky business, particularly when the data has several dimensions (particularly time or geographical dimension.) We want to agree with the curator why some data may be missing, and how best to handle it. For simple, two dimensional datasets, by default, we use linear approximation, forecasting and backcasting of the values, and in small datasets the last observation carry forward or the next observation carry backwards methods. May compromise the data? Let us know.
- **Testing against peer-reviewed results:** Often we know that after making various computations with a data, we must achieve an already known value. For example, the various components of the GDP in economics must add up with a pre-defined precision. Certain inputs must match a scientifically valid result. If you know of such tests, let us know, and let's include them in the unit-testing processes.
- **Peer-reviewed data manipulation code:** Whenever we re-organize, impute, or otherwise change the original data, we do it only with algorithms that went through scientific peer-review as algorithms. If there is a bug or something to improve in the way we handle the data, our code transparency makes it likely to come out.
- **Peer-reviewed data application:** We encourage our curators, particularly academics, to send the indicators created with the help of our research automation to various forms of scientific peer review, to make sure that the data is valid, useful... and to bring credits to the curators.
- **Authentic copies:** We are placing each new version of the indicators values into [Zenodo](#), a data repository that keeps authentic copies, versions, and assignes them digital object identifiers (DOIs). This makes sure that whenever our curators data is re-used, and incorrectly manipulated by a business, scientific or policy user, we can detect such manipulation.

You can see this dataset [here](#), which was used in [this](#) high-profile scientific publication.

6.3 Authentic Depositions of Indicators

We designed a workflow that helps our curators to put their indicator tables to [Zenodo](#). In many cases, particularly if they do EU-funded research, this is also usually a grant requirement. At the same time, we place the indicator to our database, and make it available on our data observatory's API.

With low-frequency data, such as annual data tables, we place all copies to Zenodo first, and then to the data API. In these cases, each new version of the indicator values (containing a new year, a new

The screenshot shows a Zenodo dataset page. At the top, there is a blue header bar with the Zenodo logo, a search bar, and an upload button. Below the header, the date "April 21, 2020" is displayed. The main title of the dataset is "Regionalized Cultural Access and (Books And Libraries) And Science Variables (2013)". Below the title, the author is listed as "Daniel Antal". A brief description follows: "This dataset was created from the microdata of the Eurobarometer 79.2 survey using the eurobarometer package." Several paragraphs of explanatory text are present, detailing the variables: "The read a book variable is a weighted sum of the responses that chose from 'QB1 How many months have you read a book?' any answer apart from 'not in the last 12 months.'", "The library access variable is a weighted sum of the responses that chose from 'QB1 How many months have you visited a public library?' any answer apart from 'not in the last 12 months.'", "The limited library access is a weighted sum of the responses that chose from the question 'QB2 And for each of the following activities, please tell me why you haven't done it over the last 12 months? ... Visited a public library' the answer option 'Limited or poor quality of the library in this case, the number of respondents is rather low and this is not a very reliable statistic'.", "The supports open access variable is a weighted sum of yes answer options to the 'QB3 Do you think publicly funded research should be made available online free of charge?' question.", "The internet access question is a weighted sum of responses to the answer option 'Yes, I have an Internet connection at home'.", and "The internet access question is a weighted sum of responses to the answer option 'Yes, I have a student occupation? - student'." At the bottom, there is a preview table with columns: geo, code13, code16, geo_name, and country_code. The first row shows "AT12" in all columns. The second row shows "AT12" in the first three columns, "Niederösterreich" in the fourth column, and "AT" in the fifth column.

Preview				
geo	code13	code16	geo_name	country_code
AT12	AT12	AT12	Niederösterreich	AT
AT12	AT12	AT12	Niederösterreich	AT

Figure 6.1: Zenodo Deposition Example

estimation, or a new country, a new observation unit) will have a new DOI version.

With high-frequency data, such as data tables that are refreshing daily or several times a day, we do not think that authentic versioning is useful. In such cases, we create an authentic version at a pre-agreed time frequency, for example, monthly.

6.3.1 How to Add your Existing Zenodo Depositions to Our Observatory

If you have a relevant dataset on Zenodo that should be featured in one of our observatories, or you are just uploading a new dataset, you should send it to our observatory **communities**. Communities are just collections that make your data easier to find and cite.

On your new or existing deposition, go to Edit, and you will find **Communities** right after **Files** and above **Upload Type**.

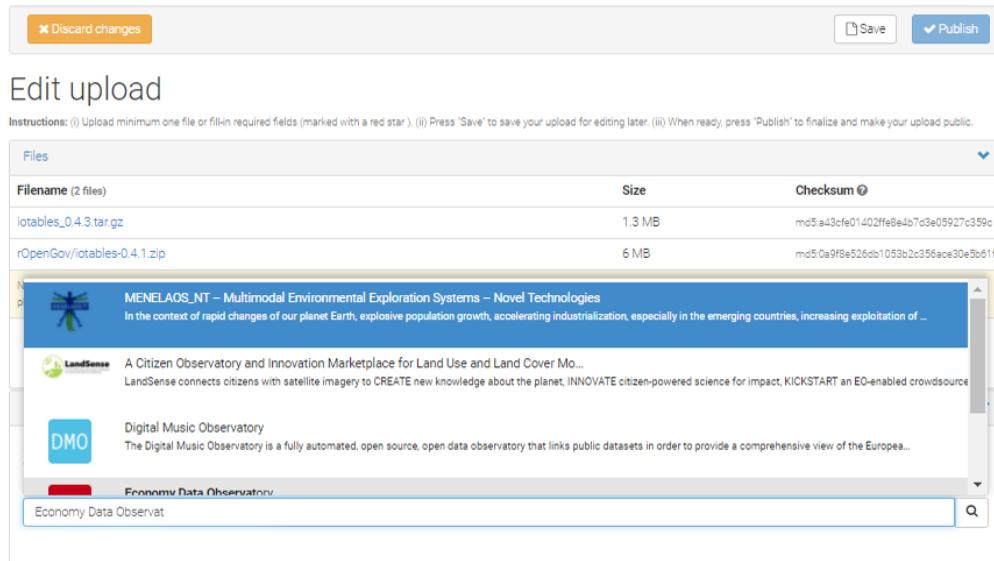


Figure 6.2: How to Add your Existing Zenodo Depositions to Our Observatory?

If you want to be featured regularly in our observatories, your data should conform our database schema. In this case, we will help you maintaining the timeliness of your data – basically we will together keep your dataset growing, expanding, and be available via our API, too. (See an example [here](#). We will add a tutorial on this shortly to our blog.)

6.3.1.1 Digital Music Observatory

You can deposit your data, or search for new, exciting data on Zenodo itself to our music observatory on zenodo.org/communities/music_observatory.

6.3.2 Green Deal Data Observatory

You can deposit your data, or search for new, exciting data on Zenodo itself to our green deal data observatory on zenodo.org/communities/greendeal_observatory/.

6.3.2.1 Economy Data Observatory

You can deposit your data, or search for new, exciting data on Zenodo itself to our green deal data observatory on zenodo.org/communities/economy_observatory/.

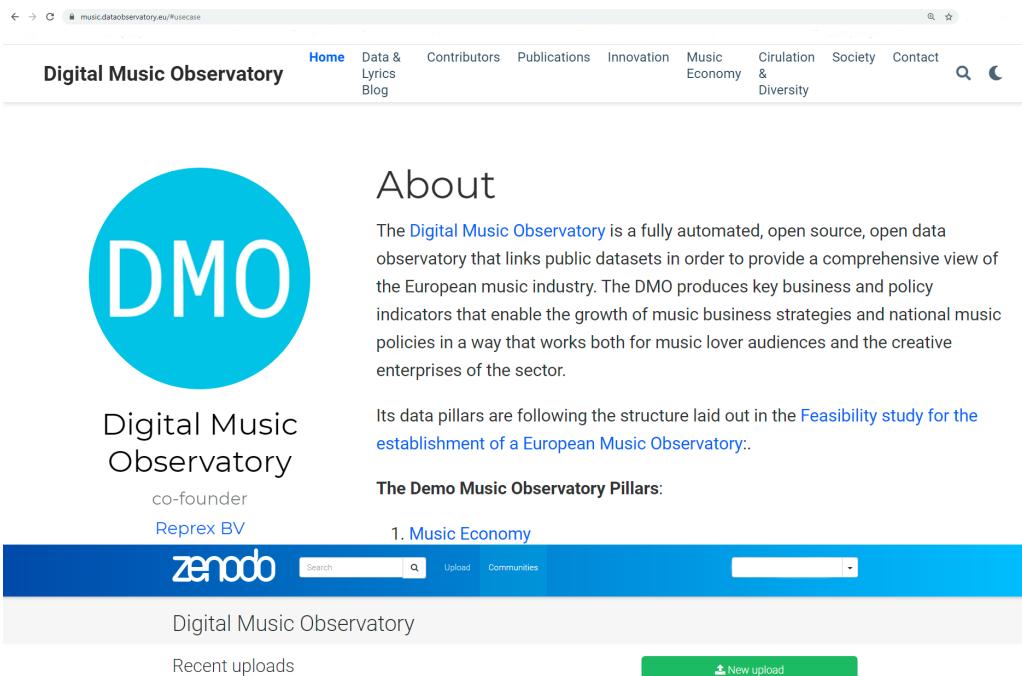


Figure 6.3: Deposit Data, Curate Data on Zenodo for the Digital Music Observatory

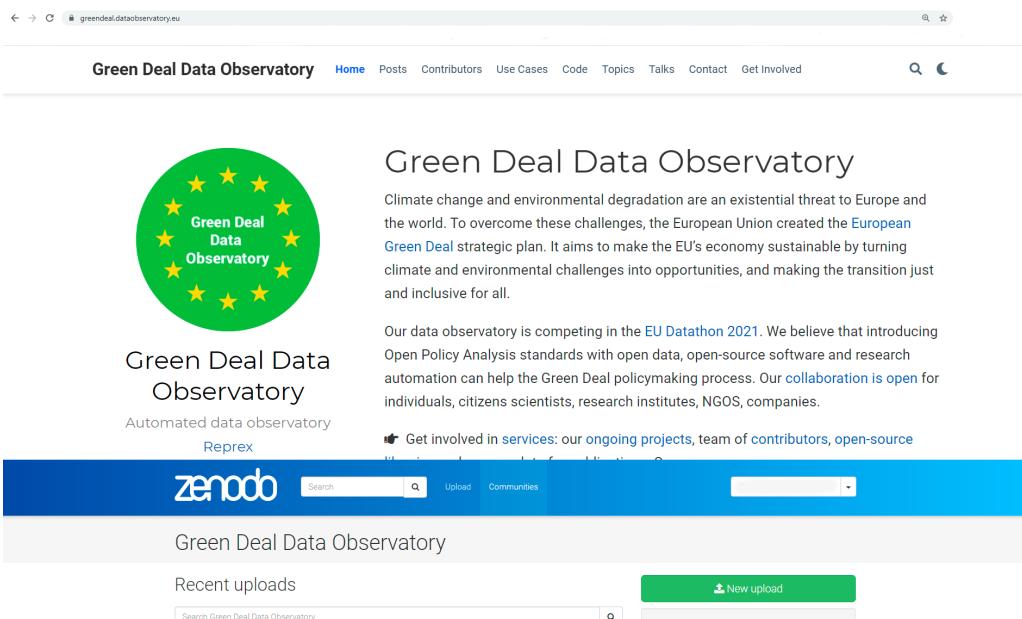


Figure 6.4: Deposit Data, Curate Data on Zenodo for the Green Deal Data Observatory

Chapter 7

Green Deal Indicators

Climate change and environmental degradation are an existential threat to Europe and the world. To overcome these challenges, the European Union created the European Green Deal strategic plan. It aims to make the EU's economy sustainable by turning climate and environmental challenges into opportunities, and making the transition just and inclusive for all.

Our data observatory is competing in the EU Datathon 2021. We believe that introducing Open Policy Analysis standards with open data, open-source software and research automation can help the Green Deal policymaking process. Our collaboration is open for individuals, citizens scientists, research institutes, NGOS, companies.

The screenshot shows two side-by-side web pages. On the left is the 'Green Deal Data Observatory' website, featuring a green circular logo with yellow stars and the text 'Green Deal Data Observatory'. Below it is the text 'Automated data observatory' and 'Reprex'. On the right is the Zenodo interface, showing a search bar and a list of recent uploads for 'Green Deal Data Observatory'. A green button labeled 'New upload' is visible.

Please, follow us on social media, it really helps us finding new users and showing that we are able to grow our ecosystem: the [Green Deal Data Observatory on Linkedin](#) and the [Green Deal Data Observatory on Twitter](#).

7.1 Curators

- Karel Volckaert: Credibility is Enhanced Through Cross Links Between Different Data from Different Domains
- Suzan Sidal: We Need More Reliable Datasets on the Urban Heat Resilience and Disaster Risk Reduction

See our [inspirational examples](#) and [Your First Data Contribution](#) in the ?? chapter.

7.2 Aggregating Count Data

We need to improve conservation by improving wildlife monitoring. Counting plants and animals is really tricky business.

The marbled murrelet is an enigma. It wasn't until the 1970s that biologists discovered where the chunky brown-and-white bird made its home, and even then it was by accident: A tree-climber found a murrelet chick at the top of a redwood. Most other bird habitats had been mapped for centuries. But who would have thought to look for a sea bird's nest miles away in the middle of an old-growth forest? And it's elusive. California birders can go a lifetime without seeing one. Every day at the break of dawn, the murrelet zips down from the redwood forest hills, where it lives, to the ocean, where it feeds. It then returns under the cover of darkness. Using remote acoustic sensors and machine learning to analyze the audio, biologists are now better able to track populations of species that were previously hard to monitor. With a [threatened species](#) like the marbled murrelet, that can make a huge difference. The better the data on its population and nesting patterns, the better our understanding of how its habitat is threatened, and the more effective conservation efforts can be.

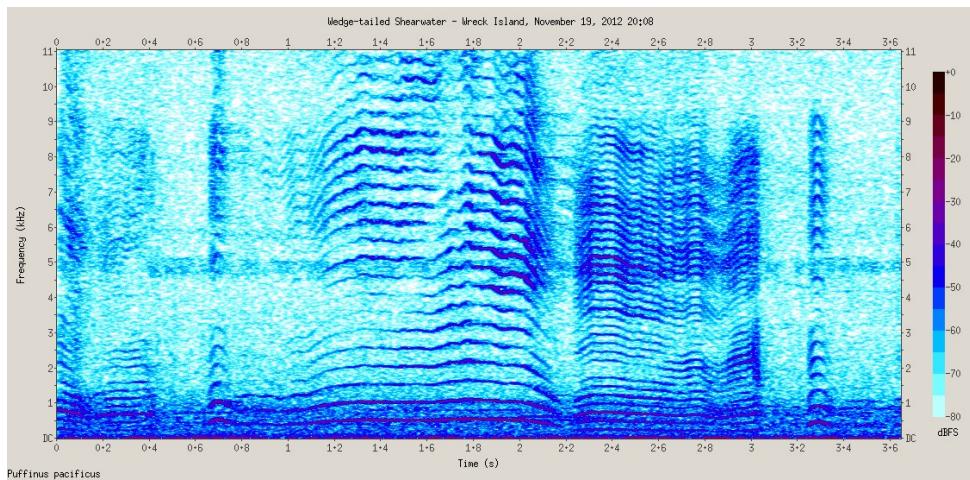


Figure 7.1: Soundscape of the Wedge-Tailed Shearwater from the Acoustic Metrics database

[Big Data Is Saving This Little Bird](#) -listen to the interview [here](#). *The illustration is taken from Jody Avirgan's blogpost.*

To analyze governmental, social data with ecological data, we need to place them on the same map. Biostatisticians, ecologists often work with count data – counting trees, birds, various species. We need to aggregate count data over the same maps that statisticians use to count people, measure the GDP or make environmental and urban planning.

This knowledge is also very important for small area statistics that we apply in [Social Attituded to Climate Change](#)

7.3 Social Attitudes to Climate Change

what do people think about climate change in Europe and other parts of the world? Do they believe that the climate is changing? How? What they think about the causes? Do they report that they change their behavior? Teach their children to do so?

Please take a look at our blogpost [Is Drought Risk Uninsurable?](#) as an example.

As a data curator:

- You identify openly accessible surveys that are harmonized (use standardized questions.) In our tutorial we projected the public opinion data from Eurobarometer 90.2 (fieldwork: October-November 2018.) on the municipal map of Belgium
- Tell us which question items would be good candidates to report. We used the answers to the multiple choice question QB1 **Do you think that the following extreme weather events are due to climate change?** We highlighted areas where people find it more likely to be exposed to **Droughts and wildfires**
- How we should calculate the indicator? Take a certain answer and average it over a region? Weight the answers? How?
- You write at least 1-2 unit tests: what must we check when the calculation is over. No negative numbers? Number of regions must up to 265?

If you write R code, you can get involved in our survey harmonization and regional coding efforts.

See our tutorial:

[Regional Geocoding Harmonization Case Study - Regional Climate Change Awareness Datasets](#)

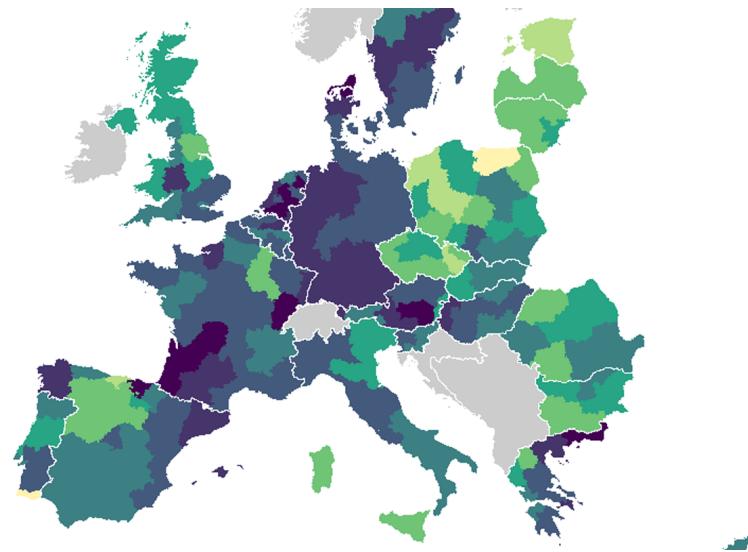


Figure 7.2: Regional attitudes to climate change, from our survey regionalization tutorial

7.4 Environmental Impact Indicator for Economic Activities

Our [iotables](#) package practically implements The [Eurostat Manual of Supply, Use and Input-Output Tables](#) with real life data and in R, and it is checked against the published results from [Jörg Beutel](#) (the author of this excellent manual), the Spicosa Project Report, and official UK statistical tables.

We used it to calculate the effects of cultural activities on various economies, but this methodology is particularly well-suited to measure the effects, or predict the effects of policy changes on greenhouse gas or pollutant emissions.

As a data curator:

- You identify openly accessible surveys data that can contains environmental effects for industries (Eurostat publishes them for many pollutants and greenhouse gases from the European Environmental Data.)
- Tell us which particular data table would be good candidate to report. Give us ideas how to bridge various problems. (The SIOT matrix must be 60x60 or 64x64), sometimes industries must be added together.
- If you write R code, we help you make the calculation yourself, if not, we'll take over.
- Please assess the results with us, and let's publish them regularly. (Some EU member states update their SIOTs every year, others every 5 years, but pollutant data may be available annually.)

7.5 Sensory Data Measuring Climate Change Physical Affects

Chapter 8

Digital Music Observatory

The [Digital Music Observatory](#) is a fully automated, open source, open data observatory that creates public datasets to provide a comprehensive view of the European music industry. It provides high-quality and timely indicators in all four pillars of the planned official European Music Observatory as a modern, open source and largely open data-based, automated, API-supported alternative solution for this planned observatory. The insight and methodologies we are refining in the DMO are applicable and transferable to about 60 other data observatories funded by the EU which do not currently employ governmental or scientific open data.

The screenshot shows the homepage of the Digital Music Observatory. At the top, there is a navigation bar with links for Home, Data & Lyrics Blog, Contributors, Publications, Innovation, Music Economy, Circulation & Diversity, Society, and Contact. Below the navigation bar is a large blue circular logo with the letters 'DMO' in white. To the right of the logo, the text 'Digital Music Observatory' is displayed. Underneath this, it says 'co-founder Reprex BV'. A banner at the bottom of the page features the word 'zenodo' and a search bar. The main content area contains an 'About' section with a summary of the DMO's mission and its pillars, followed by a link to a feasibility study. There is also a 'Recent uploads' section and a 'New upload' button.

Music is one of the most data-driven service industries where the majority of sales are already made by AI-driven autonomous systems. We provide a template that enables making these AI-driven systems accountable and trustworthy, with the goal of re-balancing the legitimate interests of creators and consumers. Music, like all creative industries, can create high-value, human jobs in the future that utilize digital skills and human creativity. Within Europe, this new balance will be an important use case of the European Data Strategy and the Digital Services Act.

The DMO places the values of the European Data Strategy at its center: our observatory model allows

the seamless flow of data within the EU and across sectors; it abides by European rules concerning privacy, access, and use; it pools data from a wide range of industries and sectors and makes it available for further research. The music industry is a global industry, and the best known European music scene in the world is the British. Our observatory aims to support a new relationship between the European and the UK music industries while offering international open data products from global sources.

The DMO is a fully-functional service that can function as a testing ground of the European Data Strategy, showcasing the ways in which the music industry is affected by the problems that the Digital Services Act and European Trustworthy AI initiatives attempt to regulate. Our observatory's policy insights also shed new light on important aspects of the Digital Skills and Connectivity agenda of the European Union. As a user-friendly one-stop shop for all things concerning data and the music industry, our DMO provides the foundations for a healthier and accountable European music ecosystem.

Please, follow us on social media, it really helps us finding new users and showing that we are able to grow our ecosystem: the [Digital Music Data Observatory on Linkedin](#) and the [Digital Music Data Observatory on Twitter](#).

8.1 Curators

- [Eszter Kabai: New Indicators for Royalty Pricing and Music Antitrust](#)
- [Dominika Semaňáková: We Want Machine Learning Algorithms to Learn More About Slovak Music](#)

See our [inspirational examples](#) and [Your First Data Contribution](#) in the ?? chapter.

8.2 Music Economy

8.2.1 Demand Drivers

Our music economy [demand drivers](#) are data that are known to be leading indicators to an increase in mechanical copies, streaming subscriptions, public performance use, private copying or illegal use of music.

8.2.2 Supply Indicators

Our Music Economy / [Supply](#) indicators are related to the supply of new music.

8.2.3 Price Indicators

8.3 Music Diversity

8.3.1 Gender, Language, Ethnic and Other Inclusion Attributes

8.4 Music Circulation

8.4.1 Market shares

We are developing [market share](#) indicators for streaming and broadcast music.

For national, gender or other market share, we need to label both music works and recorded fixations. We use various open source databases, and machine learning algorithms to do prepare the data, but eventually our data goes through human musicology or music journalist curators.

For example, in our case study we were interested in the various definitions of [Slovak market share](#), and representation of [female artists](#). Both problems require rather challenging labeling.

- a) we tried to find a location to the artist / band [you can describe why this is not always straight-forward, for example, in the case of dead artists, etc.]
- b) our algorithm tried to guess the language of the 10 most popular song titles
- c) we checked if the person is on the Wikipedia list of “Slovak male singers”, “Slovak female singers”, “Slovak bands”, or their Czechoslovak versions [who did you decide when somebody was Czechoslovak if they were Slovak]
- d) check if there was a Slovak placename mentioned on their bandcamp site
- e) check if they are associated with Slovakia on the Musicbrainz open source database
- f) if any of the artists released recordings was released in Slovakia
- g) if the majority of the artist's released recordings was released in Slovakia

....

Until we got to the human curation.

and eventually we either `considered_slovak` or not somebody in our `write-in database`. We are also developing an `opt-in database`, where artists can give us their own ethnic, local and gender identity, if they wish to, and of course, they can opt-out from our labeling.

We are using Monte Carlo simulation and non-parametric sampling of various broadcast and music streams to get a representation of the music listened to in various cities, regions, countries, and then we apply `national`, `language`, `gender`, `sex`, `locality` and `folksonomy` tags to measure female, Slovak, Estonian, Amsterdam or Welsh market share, recommendation probability, etc.

8.5 Music & Society

8.5.1 Social Attitudes to Music

8.5.2 Participation in Music

Chapter 9

Economy Data Observatory

Big data and automation create new inequalities and injustices and has a potential to create a jobless growth. Our Economy Observatory is a fully automated, open source, open data observatory that produces new indicators from open data sources and experimental big data sources, with authoritative copies and a modern API.

Our observatory is monitoring the European economy to protect the consumers and the small companies from unfair competition both from data and knowledge monopolization and robotization. We take a critical SME-, intellectual property policy and competition policy point of view automation, robotization, and the AI revolution on the service-oriented European social market economy.

We would like to create early-warning, risk, economic effect, and impact indicators that can be used in scientific, business and policy contexts for professionals who are working on re-setting the European economy after a devastating pandemic and in the age of AI. We are particularly interested in designing indicators that can be early warnings for killer acquisitions, algorithmic and offline discrimination against consumers based on nationality or place of residence, signs of undermining key economic and competition policy goals, and generally help small and medium-sized enterprises and start-ups to grow, and the financial sector to provide loanable and equity funds for their growth.

Economy Data Observatory

[Home](#)

Economy Data
Observatory

Automated Data Observatory

An important aspect of the EU Datathon Challenges is “.. to propose the development of an application that links and uses open datasets [...] to find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of open data.”

In the [An economy that works for people](#) challenge we are focusing on the [Single market strategy](#), and particular attention to the strategy’s goals of 1. Modernising our standards system, 2. Consolidating Europe’s intellectual property framework, and 3. Enabling the balanced development of the collaborative economy strategic goals.

Timeline for the Economy Data Observatory

2018-2020	Open-source statistical software to manipulate open data.
September 2020	Semi-automated prototype, the Demo Music Observatory is developed with 60 stakeholders in 12 counties.
October 2020	Observatory product/market fit validation in the world's first AI+Blockchain validation Lab.
February 2021	The prototype automated music observatory is chosen to be part of our data.
March 2021	On International Open Data Day, our second observatory, the green deal observatory, is announced.
April 2021	First use case of the green deal observatory with a Belgian competition, competitiveness, innovation, and small- and medium-sized enterprises.
May 2021	Launch of our data API, separating the product team to support the EU Datathon 2021 as Economy Data Observatory with daily growing from day one continuously, but the application is still in beta.
June 2021	We solidify the automation between the critical elements with unit tests, dissemination in API and automatic documentation end of the month. From a technical point of view, we move to the next level.
July 2021	Via our academic, policy and business partners we intend to expect that our data observatory, as a data ecosystem of data, will be adopted by the public and private sectors.
August 2021	Based on user feedbacks, we are improving the value proposition for business users.
September 2021	Finalizing the business model based on a hybrid licensing from this point.
November 2021	Feedback from EU Datathon 2021!

9.1 Curators

[Peter Ormosi: New Indicators for Computational Antitrust](#)

See our [inspirational examples](#) and [Your First Data Contribution](#) in the ?? chapter.

9.2 Social Attitudes to Economic Change

what do people think about climate change in Europe and other parts of the world? Do they believe that the climate is changing? How? What they think about the causes? Do they report that they change their behavior? Teach their children to do so?

Please take a look at our blogpost [Is Drought Risk Uninsurable?](#) as an example.

As a data curator:

- You identify openly accessible surveys that are harmonized (use standardized questions.) In our tutorial we projected the public opinion data from Eurobarometer 90.2 (fieldwork: October-November 2018.) on the municipal map of Belgium
- Tell us which question items would be good candidates to report. We used the answers to the multiple choice question QB1 **Do you think that the following extreme weather events are due to climate change?** We highlighted areas where people find it more likely to be exposed to **Droughts and wildfires**
- How we should calculate the indicator? Take a certain answer and average it over a region? Weight the answers? How?
- You write at least 1-2 unit tests: what must we check when the calculation is over. No negative numbers? Number of regions must up to 265?

If you write R code, you can get involved in our survey harmonization and regional coding efforts.

See our tutorial:

[Regional Geocoding Harmonization Case Study - Regional Climate Change Awareness Datasets](#)

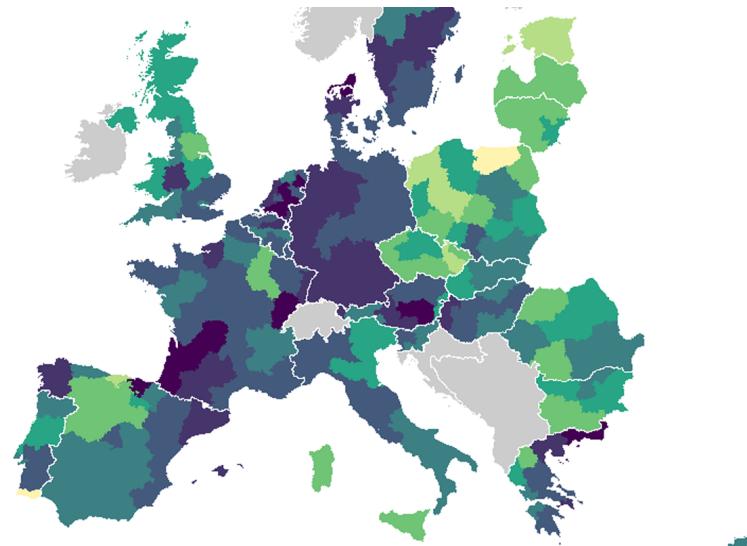


Figure 9.1: Regional attitudes to climate change, from our survey regionalization tutorial

9.3 SME Activity Indicators

9.4 Economic Impact Indicators

Our [iotables](#) package practically implements The [Eurostat Manual of Supply, Use and Input-Output Tables](#) with real life data and in R, and it is checked against the published results from [Jörg Beutel](#) (the author of this excellent manual), the Spicosa Project Report, and official UK statistical tables.

We used it to calculate the effects of cultural activities on various economies, but this methodology is particularly well-suited to measure the effects, or predict the effects of policy changes on greenhouse gas or pollutant emissions.

As a data curator:

- You identify openly accessible surveys data that can contains environmental effects for industries (Eurostat publishes them for many pollutants and greenhouse gases from the European Environmental Data.)
- Tell us which particular data table would be good candidate to report. Give us ideas how to bridge various problems. (The SIOT matrix must be 60x60 or 64x64), sometimes industries must be added together.
- If you write R code, we help you make the calculation yourself, if not, we'll take over.
- Please assess the results with us, and let's publish them regularly. (Some EU member states update their SIOTs every year, others every 5 years, but pollutant data may be available annually.)

9.5 Sensory Data Measuring Changes in Economic Activy

Chapter 10

Competition Data Observatory

Big data and automation create new inequalities and injustices and has a potential to create a jobless growth. Our Economy Observatory is a fully automated, open source, open data observatory that produces new indicators from open data sources and experimental big data sources, with authoritative copies and a modern API.

Our observatory is monitoring the European economy to protect the consumers and the small companies from unfair competition both from data and knowledge monopolization and robotization. We take a critical SME-, intellectual property policy and competition policy point of view automation, robotization, and the AI revolution on the service-oriented European social market economy.

We would like to create early-warning, risk, economic effect, and impact indicators that can be used in scientific, business and policy contexts for professionals who are working on re-setting the European economy after a devastating pandemic and in the age of AI. We are particularly interested in designing indicators that can be early warnings for killer acquisitions, algorithmic and offline discrimination against consumers based on nationality or place of residence, signs of undermining key economic and competition policy goals, and generally help small and medium-sized enterprises and start-ups to grow, and the financial sector to provide loanable and equity funds for their growth.

Competition Data Observatory

[Home](#) [Blog](#) [Data Coverage](#)



Competition Data Observatory

Automated Data Observatory

Reprex

rOpenGov

Yes!Delft AI+Blockchain Validation
Lab



Competitor Observatory

Our Competition Data Observatory is an open-source observatory that provides access to various data sources, with a focus on competition law.

Our observatory is not just a passive data store; it also develops tools for competition analysis, such as property policy and market structure analysis, revolution on the scale of the European Union.

We would like to create a platform where data can be used in scientific research and policy making, on re-setting the European Union's competition rules. We would like to make this platform available to sub-national, regional, and local governments.

👉 Get involved in setting up the observatory, contribute to our open source libraries and use our tools to analyze your data.

RSS Follow news about the observatory and our work.

📞 Contact us if you have any questions or suggestions.

10.1 Curators

Peter Ormosi: New Indicators for Computational Antitrust

See our [inspirational examples](#) and [Your First Data Contribution](#) in the ?? chapter.

10.2 Competition

We are seeking API level access to the European Commissions Mergers database, and monitor approved and declined merger requests programatically. These mergers are important cases enough to have a potential impact on the structure of the European economy.

Policy Area	Case Number	Member State	Last Decision Date	Title	Action
Merge	M_10			CONAGRA / IDEA	Show details
Merge	M_10000			PREZERO INTERNATIONAL / SUEZ NORDIC	Show details
Merge	M_10001			MICROSOFT / ZENIMAX	Show details
Merge	M_10002			HOYER / RHEINUS / JV	Show details
Merge	M_10003			DWS / VERTIX BIODENERGY	Show details
Merge	M_10004			EQT / ZENTRUM CITY / CANOIL / REICHHARH	Show details
Merge	M_10005			CPPIB / SIXTH STREET / CLARA	Show details
Merge	M_10006			COVESTRO / KONINKLIJKE DSM (NEBINS & FUNCTIONAL MATERIALS BUSINESS) / INNOVATION POLYMERS	Show details
Merge	M_10007			TELEFONICA / BANCO BILBAO VIZCAYA ARGENTARIO / MOVISTAR MONEY COLOMBIA JV	Show details
Merge	M_10008			EBERHA / PARCON / WOOD HOLDINGCO JV	Show details
Merge	M_10009			PREUSAG / HAGAS-LLOYD	Show details
Merge	M_10010			INVESTINDUSTRIAL GROUP / CSM INGREDIENTS	Show details

As a data curator, you help us designing datasets

- created from Commission and member state merger decision text databases (we will use NLP extraction from the text of the decisions)
- top-down indicators that show the structural (concentration) changes in the European economy
- connect them to patent databases

These indicators are particularly interesting, because we are trying to connect to databases that fall under the [Directive on open data and the re-use of public sector information - in short: Open Data Directive \(EU\) 2019 / 1024](#), but programmatic access appears to be problematic. We need to secure reproducible, programmatic access to these important open data sources.

10.2.1 Knowledge Monopolizations, Killer Acquisitions

In killer acquisitions, a large company, for example, in pharmaceutical or technology fields, buys a small company, or even a start-up, to avoid disruptive innovation. We are building several types of indicators in this field.

- What type of patents companies hold in smaller entities of mergers and acquisitions? How can we characterize potentially disruptive technology?
- Which economic activities (industries as described by NACE) are more and less effected?
- How is patent concentration changing?