Automated Observatory Data Curator's Handbook

Daniel Antal, CFA

2021-06-03

Contents

List of Figures

Introduction

Mission statement

We want to win an EU Datathon prize by processing the vast, already-available governmental and scientific open data made usable for policy-makers, scientific researchers, and business researcher end-users.

"*To take part, you should propose the development of an application that links and uses open datasets. Your application should showcase opportunities for concrete business models or social enterprises. It is also expected to find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of open data.*"

We want to win at least one first prize in the EU Datathon 2021.

- Challenge 1: A European Green Deal, with a particular focus on the The European Climate Pact, the Organic Action Plan, and the New European Bauhaus, i.e. mitigation strategies.
- Challenge 2: An economy that works for people, with a particular focus 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.
- Challenge 3: A Europe fit for the digital age, with a particular focus Artificial Intelligence, the European Data Strategy, the Digital Services Act, Digital Skills and Connectivity. We will showcase these horizontal topics with our Digital Music Observatory.

For Challenge 1, we are preparing the Green Deal Data Observatory. For Challenge 2, the Economy Data Observatory – any better name is welcome. For Challenge 3, our Digital Music Observatory to highlight our efforts in trustworthy, ethical AI, and to find a new balance between the interests of artists and music audiences.

Problem Statement

The EU has a 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 5000 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: data.europa.eu/data is the new open data portal of the EU that replaces two previous versions (one for the common institutions, and technically harvesting all EU member states' national open data portals).

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 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 was to modernize the concept of 'data observatories'.

Our solution

Recruit data curators who know how to put important policy data (according to the EU challenges) into a useful, processed format. Help them with open-source statistical software solutions, and open-source data services to make it available for end-use in policy research (NGOs, public entities), scientific research, or business research.

Find a for-profit business model or a non-profit social enterprise model to make our service sustainable. Contest at least 2 of the 3 Datathon challenges to show that our solution is general and not domain specific: minimally, we plan to contest the Green Deal Challenge with environmental and climate policy data, and we plan to convert our Demo Music Observatory into a Digital Music Observatory that showcases important policy issues in the Digital Life challenge.

If we were to find reliable partners in the time prior to the deadline, we would also consider submitting a bid to the third challenge, which mainly deals with economic and social policy.

Our R packages offer a professionally sound version of the data that renders it usable and reliable. In this project, we want so scale up their productivity by embedding them (and other similar packages, and even Python libraries if we can) to services.

- regions corrects inconsistent geographical coding.
- iotables puts extremely complex national accounts data into actually useful environmental and economic impact indicators.
- retroharmonize connects cross-sectional surveys with non-European countries, puts pan-European surveys into time series, and corrects regional subsamples.
- indicator, in its early stage, attempts to bring to a common, tidy format the diverse and untidy indicators of European governmental open data.

The results are new statistical products, which are, in a way, a subjective interpretation of the data that is far more useful than leaving it in its original state. The usefulness of our data products is linked to our reputation, the peer-reviewed processes of our packages, and eventually, the peer-reviewed uses of the datasets created.

This means that all of our data products must possess an authentic, authored, and accountable version; therefore, all versions of our data assets are assigned Digital Object Identifiers (DOI).

For instance, if we recreate a Eurostat statistical product with corrected geocoding (member states have no mandate to correct historical data, which often results in badly-coded data), such as a new version of, say, regional CO2 emission or GDP, our version must be traceable, and eventually be available for rigorous peer-review.

Chapter 1

Application for the Datathlon Prizes: Automated Data Observatories

This year's Datathlon has three challenges that we can contest with an application. We are contesting all three challenges with the same technology, knowledge management, but partly different data used.

1.1 How We Add Value to Data

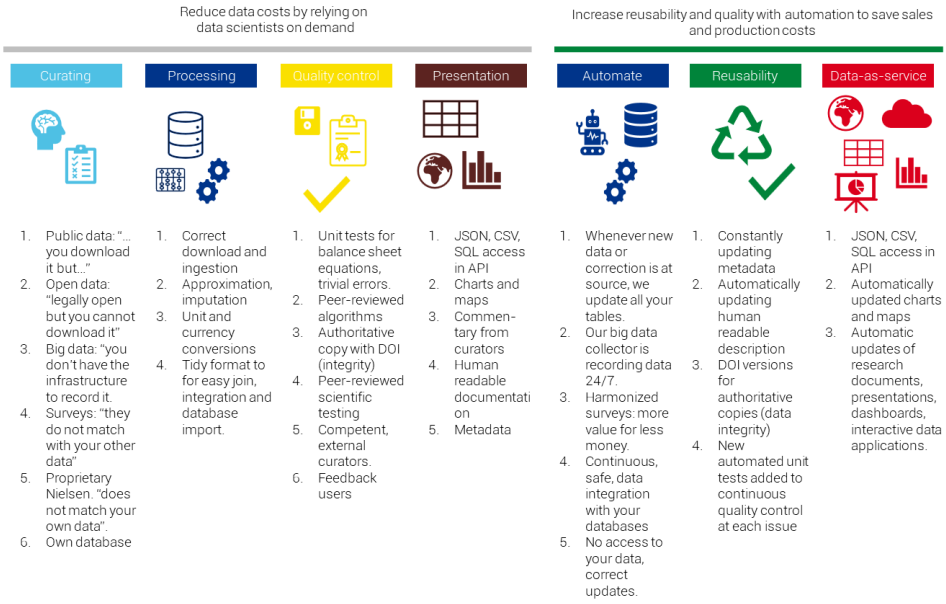
1.1.1 Reproducible Research

1. **Our curatotrs**: help finding the best available information source. This is often open data, which does not equal 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 change million euros to euros in a spreadshit, tons to kilograms, maybe even ounces to grams, but many unit conversation are error-prone if done by humans. Not everybody can make valid currency translation (*When do I use year-end USD/EUR rate? Today's EUR/GBP? Or GBP/EUR? Annual average rate?*) We do it in a way that it confirms the tidy data definition, which allows its easy integration, joining and importing into your database.

While the unit conversions can be automated in Excel or PowerBI, the data tidying requires a more programatic approach.

3. **Quality control**: Our data goes through dozens of computer logical checks (*Do assets and liabilities match? 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 allows 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 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.



1.1.2 Research Automation

Because every step of our data acquisition, processing, strucutring, testing is going through machines, it can be replicated any given year, month, day, or hour. The 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 them in big quantities to find signficiant, meaningful information in them. Most small enterprises and research teams do not have the engineering capactity to organize this. We do continous data collection on all sources to capture the latest updates, or corrections.

2. **Focus on reusability**: In our experience, the resuability of research and consulting work is greatly reduced by two factors, which we resolve with continous data collection and documentation:

   - poor documentation (the bibliography updates and file descriptions are the least priortized tasks)
   - data tables, charts, visualizations became 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, maps, are created by open source and reproducible software, the software can run continously. Instead of providing our users with data tables, charts and maps, we provide them with a subsription to the latest data and its latest visualizations.

1.2 Service Flow

1.2.1 Data Acquisition and Processing

We are accessing 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 more synergies between the more statistically oriented R language and the more genereal application-oriented Python. We are welcoming 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 an open collaboration with rOpenGov and other developers.

We entered into the annual EU Datathon competition in all three challenges with our applications to not only provide open-source software.

1.2.2 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 endorsed. Manual deposition is working fine, and we can very easily retrieve our own data in various versions. It is also free data storage.

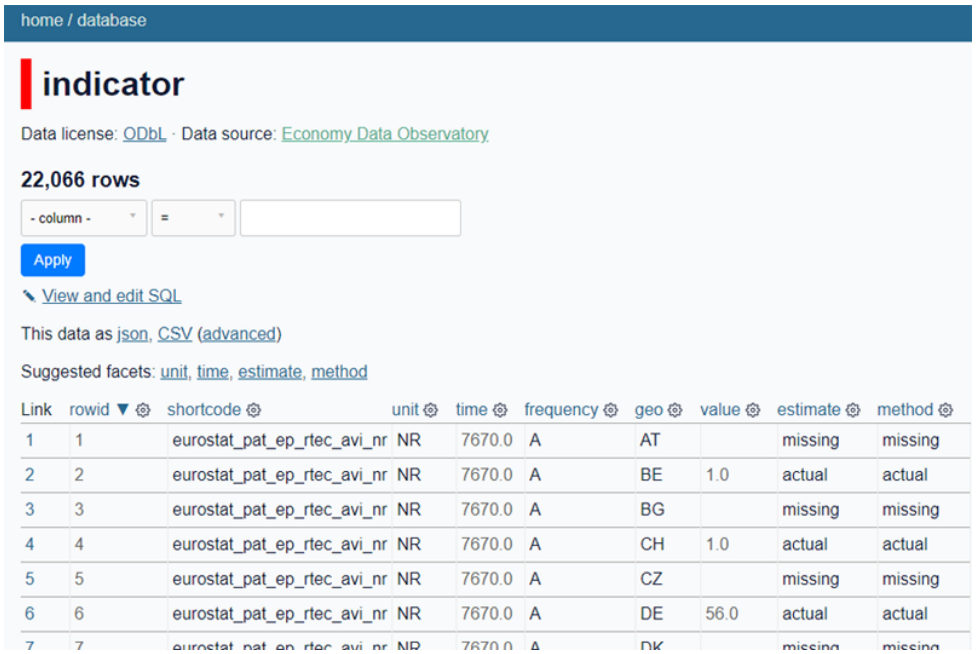
The Zenodo API is not very well documented, particularly for R. But it is supported both in Python and R. We have a tutorial and a code how to deposit our assets programatically via the Zen4R package. It is a bit difficult to use - it mimics "true" object oriented programming with relying on R6 classes,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

1.2.3 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 is taking care of our APIs.

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



home / database

indicator

Data license: ODbL · Data source: Economy Data Observatory

22,066 rows

- column - = Apply

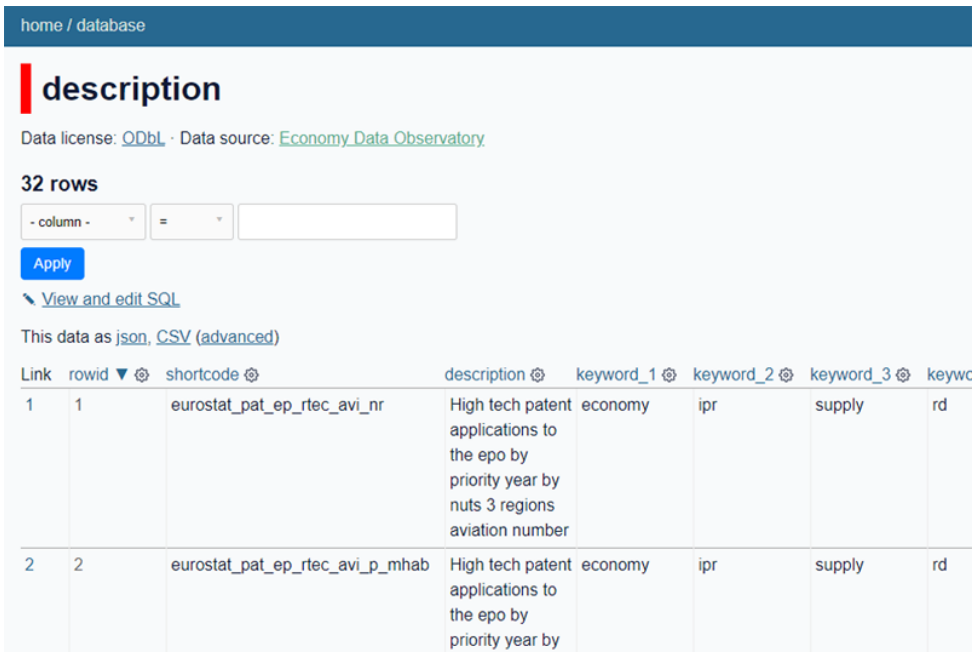
View and edit SQL

This data as json, CSV (advanced)

Suggested facets: unit, time, estimate, method

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.



home / database

description

Data license: ODbL · Data source: Economy Data Observatory

32 rows

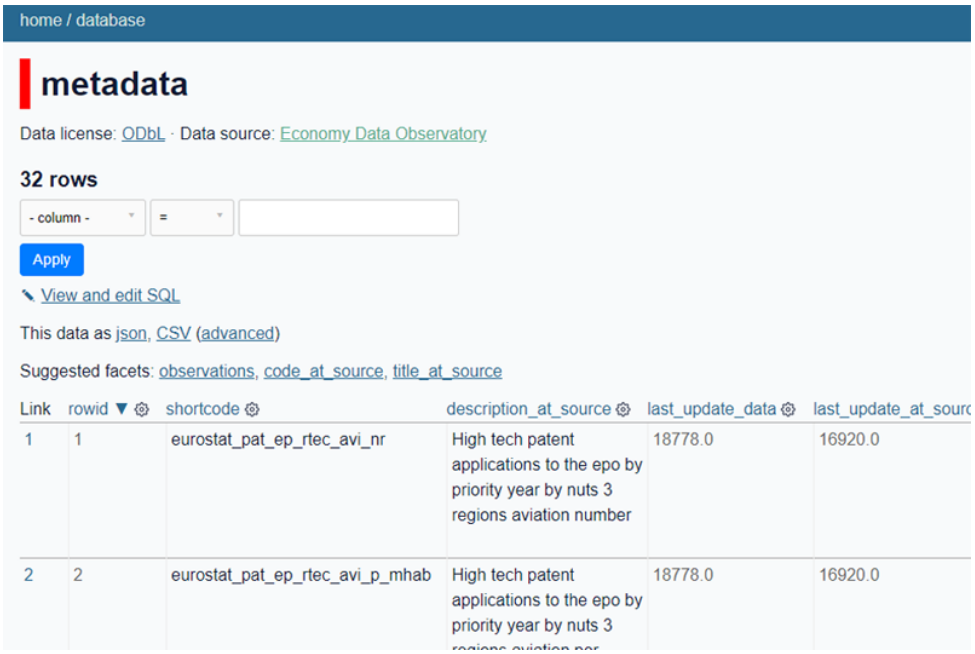
- column - = Apply

View and edit SQL

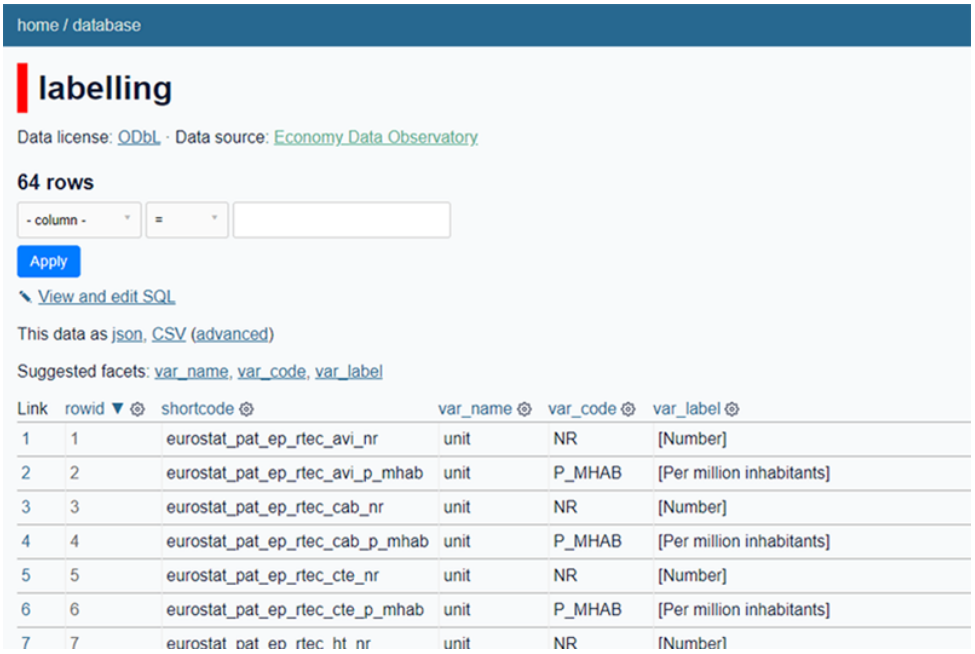
This data as json, CSV (advanced)

Link	rowid ▼	shortcode	description	keyword_1	keyword_2	keyword_3	keywo
1	1	eurostat_pat_ep_rtec_avi_nr	High tech patent applications to the epo by priority year by nuts 3 regions aviation number	economy	ipr	supply	rd
2	2	eurostat_pat_ep_rtec_avi_p_mhab	High tech patent applications to the epo by priority year by	economy	ipr	supply	rd

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



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



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

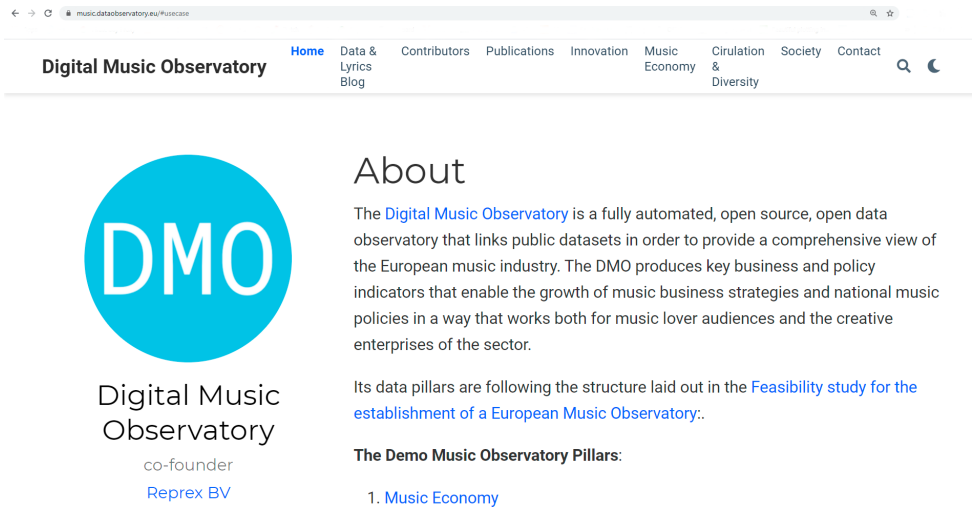
1.2.4 Long form documentation

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

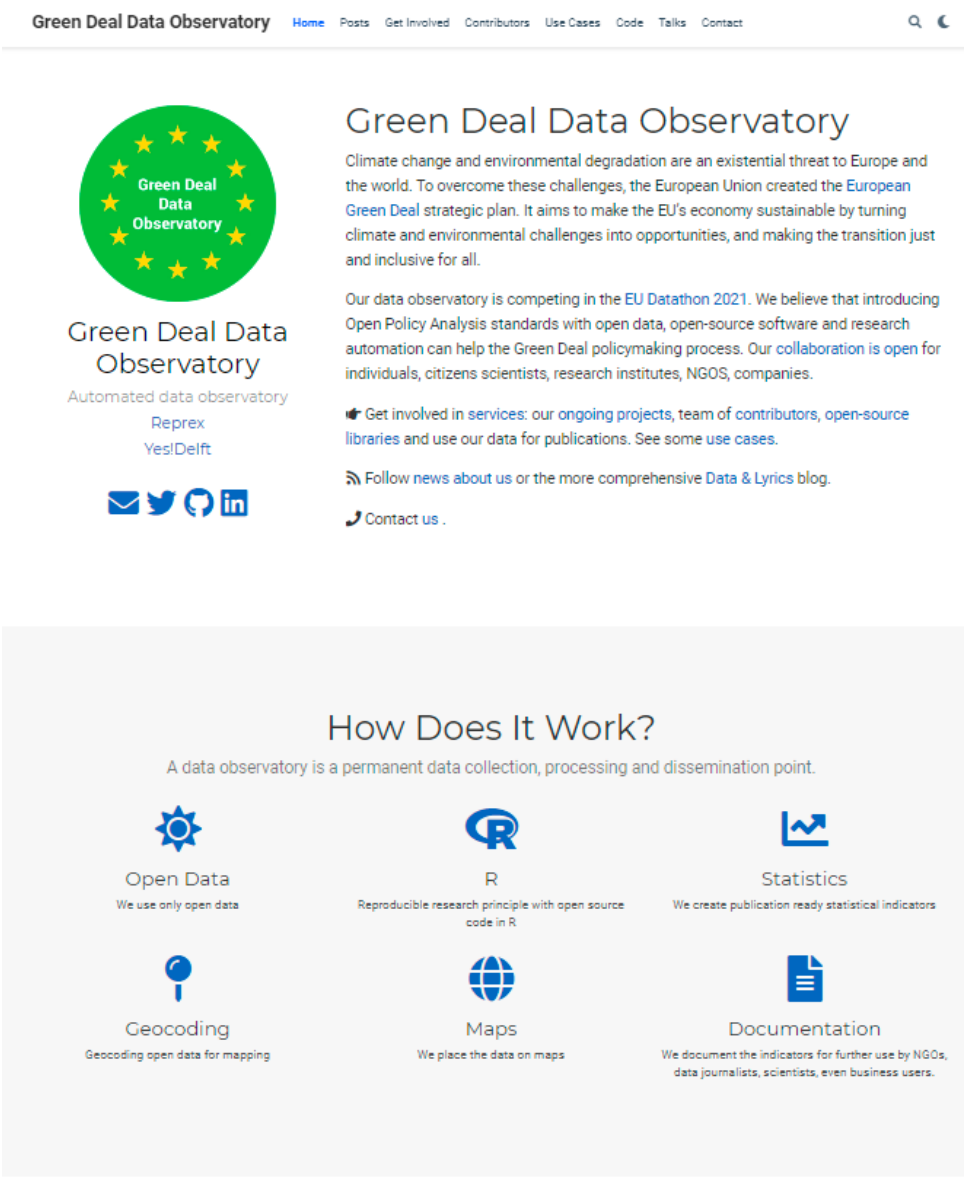
It is a very mature workflow, it produces a long-form website, and PDF, ePUB or Word versions from the API. The current automation is not operational, as we have recently included the API.

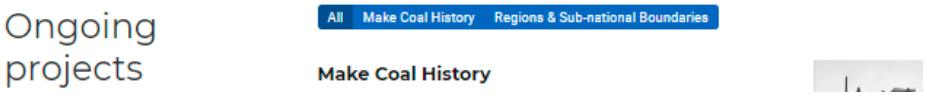
1.2.5 Front-End Websites

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 will win a price, we'll certainly offer them a share!



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 that it can be managed in the same workflow as the indicator generation, package documentation, the long-form documentation is a bit gone.


Green Deal Data Observatory
Home Posts Get Involved Contributors Use Cases Code Talks Contact
Green Deal Data Observatory
Green Deal Data Observatory
Automated data observatory
Reprex
Yes!Delft
Green Deal Data Observatory
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.
Get involved in services: our ongoing projects, team of contributors, open-source libraries and use our data for publications. See some use cases.
Follow news about us or the more comprehensive Data & Lyrics blog.
Contact us .
How Does It Work?
A data observatory is a permanent data collection, processing and dissemination point.
Open Data
We use only open data
R
Reproducible research principle with open source code in R
Statistics
We create publication ready statistical indicators
Geocoding
Geocoding open data for mapping
Maps
We place the data on maps
Documentation
We document the indicators for further use by NGOs, data journalists, scientists, even business users.


Ongoing projects
All
Make Coal History
Regions & Sub-national Boundaries
Make Coal History

Chapter 2

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 subchapter7.1.1 of Tools.)

2.1 Get Inspired

2.1.1 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.

- 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."

2.1.2 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/](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 pdocast *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.

2.1.3 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 7.3.1 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. 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.

6. 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.)

7. 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.

2.2 What is 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.

2.3 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.

2.3.1 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.

2.3.2 Professional Standards in Business

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.

2.4 Indicator Design

We are committing ourselves in the final deliverable to follow the indicator design principles set out by Eurostat: (Eurostat, 2014; kot, 2017; Eurostat, 2014) 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 (PwC, 2008), for both author's and neighbouring rights, and fullfil the IFRS fair value standards, incorporated in EU law and the recent EU jurisprudence (InfoCuria, 2014, 2017).
- Indicators that can be used for calculating damages, or calculating the value of the value gap (Antal, 2019b,a).
- Indicators that quantify the development needs of musicians, and can set objective granting aims and grant evaluations (Antal, 2015).
- 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 (Antal, 2019c).
- Providing detailed comparison of the differences of music audience among countries.
- Measuring exporting success on streaming platforms, and preparing better targeting tools.

2.4.1 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 oour 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.

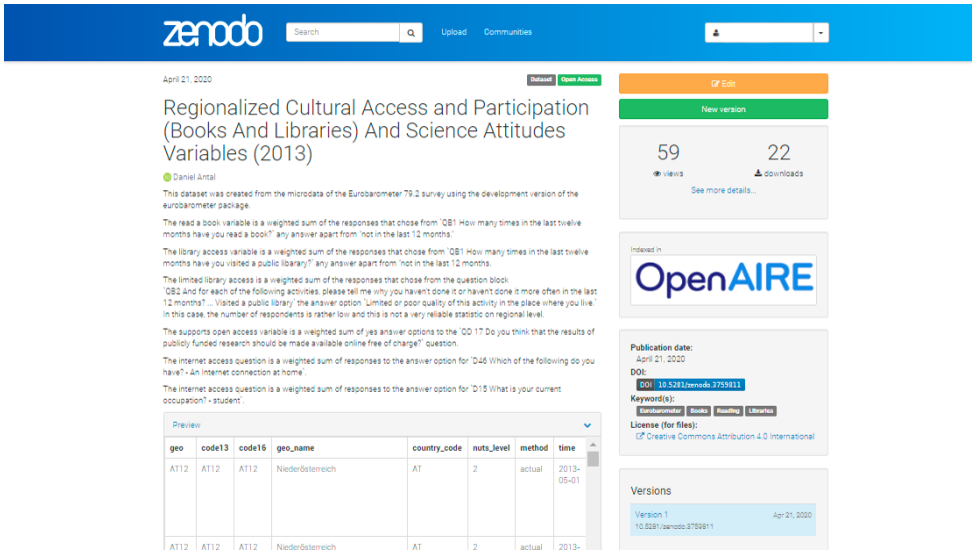


Figure 2.1: Zenodo Deposition Example

You can see this dataset here, which was used in this high-profile scientific publication.

2.5 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 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.

2.5.1 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`.

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.)

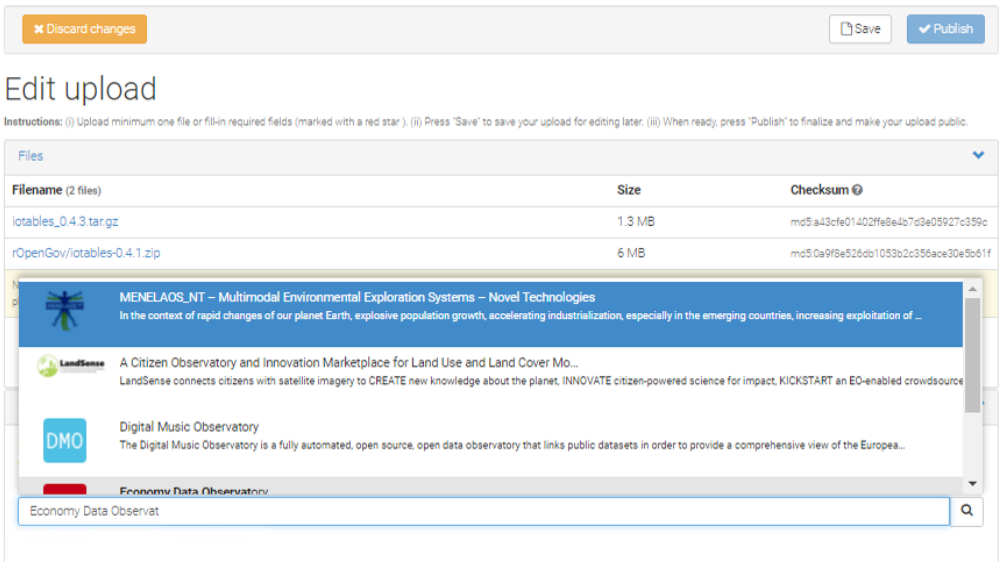


Figure 2.2: How to Add your Existing Zenodo Depositions to Our Observatory?

2.5.1.1 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.

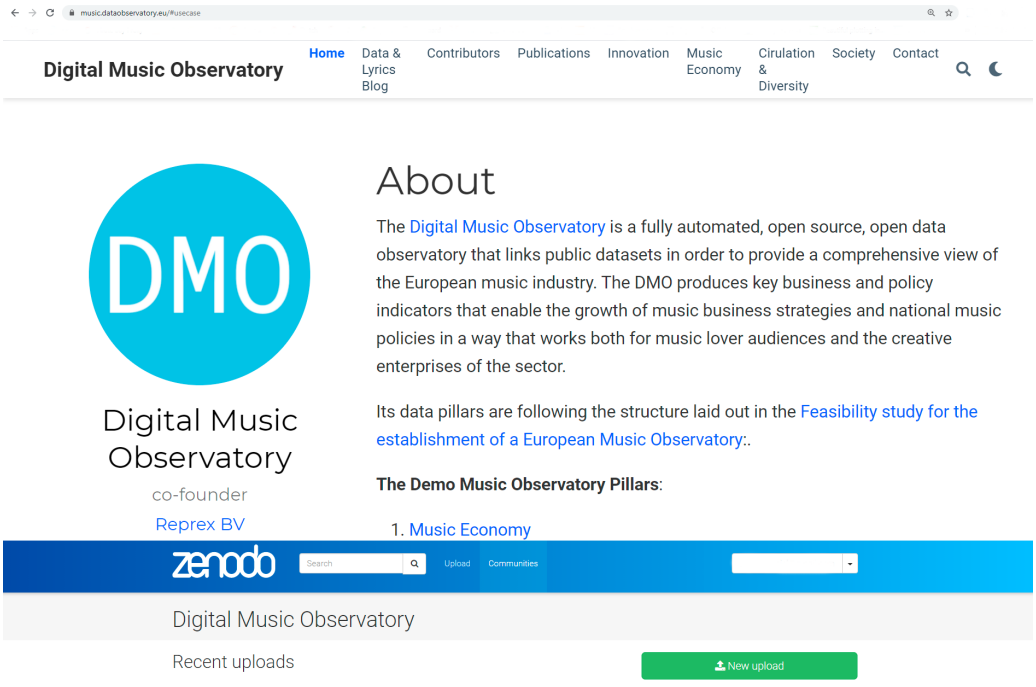


Figure 2.3: Deposit Data, Curate Data on Zenodo for the Digital Music Observatory

2.5.2 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.

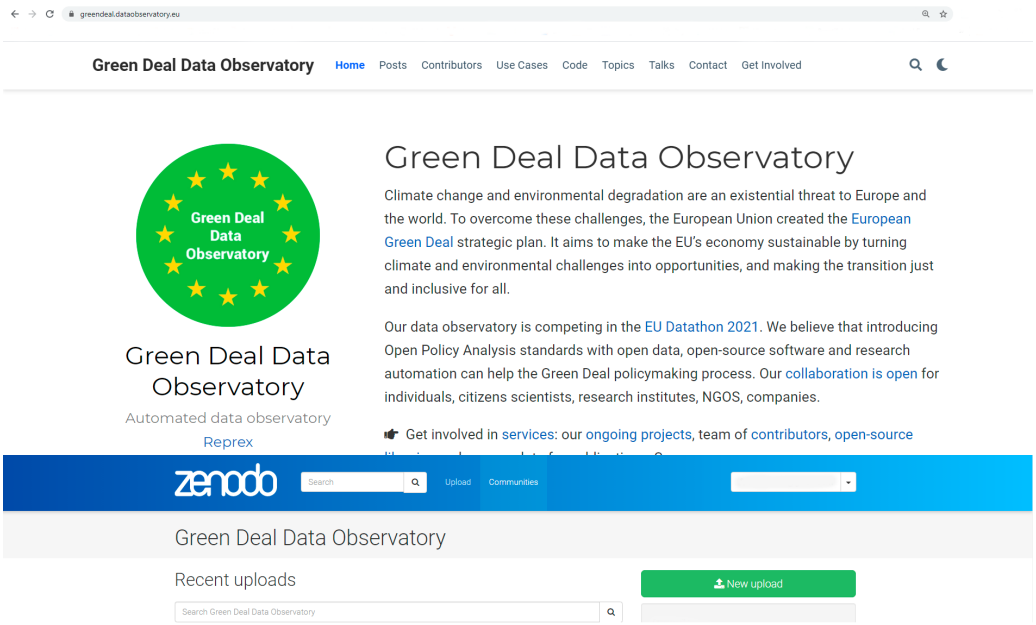


Figure 2.4: Deposit Data, Curate Data on Zenodo for the Green Deal Data Observatory

2.5.2.1 Economy Data Observatory

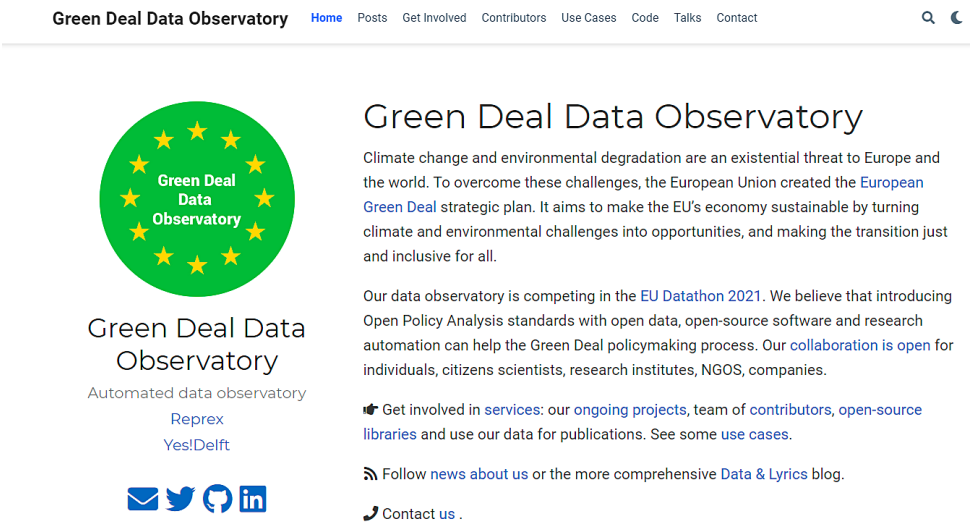
You can deposit your data, or search for new, exciting data on Zenodo itself to our green deal data observatory on zenodo.org/communities/economy_observatory/.

Chapter 3

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.



3.1 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.

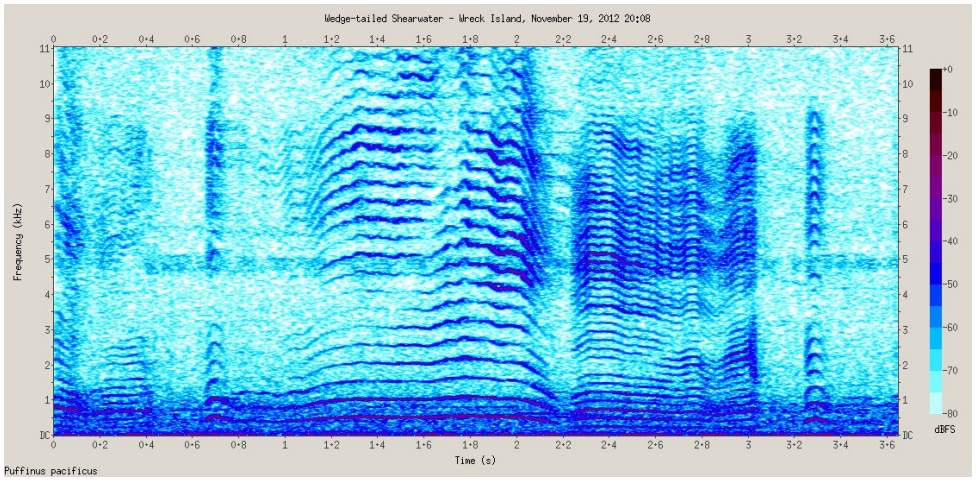


Figure 3.1: Soundscape of the Wedge-Tailed Shearwater from the Acoustic Metrics database

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 Attituted to Climate Change

3.2 Social Attituted 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 suvey harmonization and regional coding efforts.

See our tutorial:

Regional Geocoding Harmonization Case Study - Regional Climate Change Awareness Datasets

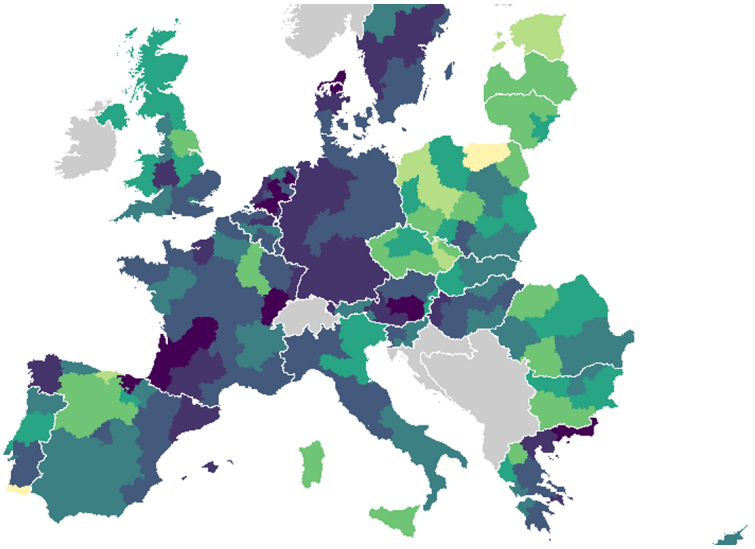


Figure 3.2: Regional attituteds to climate change, from our survey regionalization tutorial

3.3 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.)

3.4 Sensory Data Measuring Climate Change Physical Affects

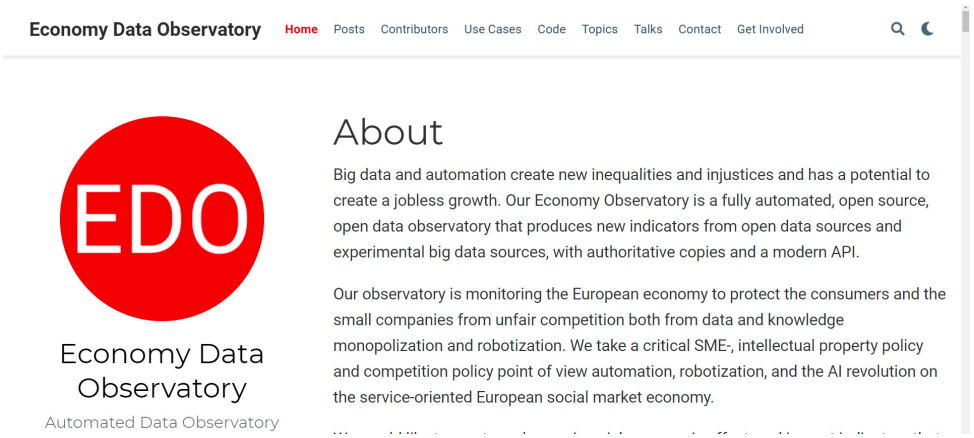
Chapter 4

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.

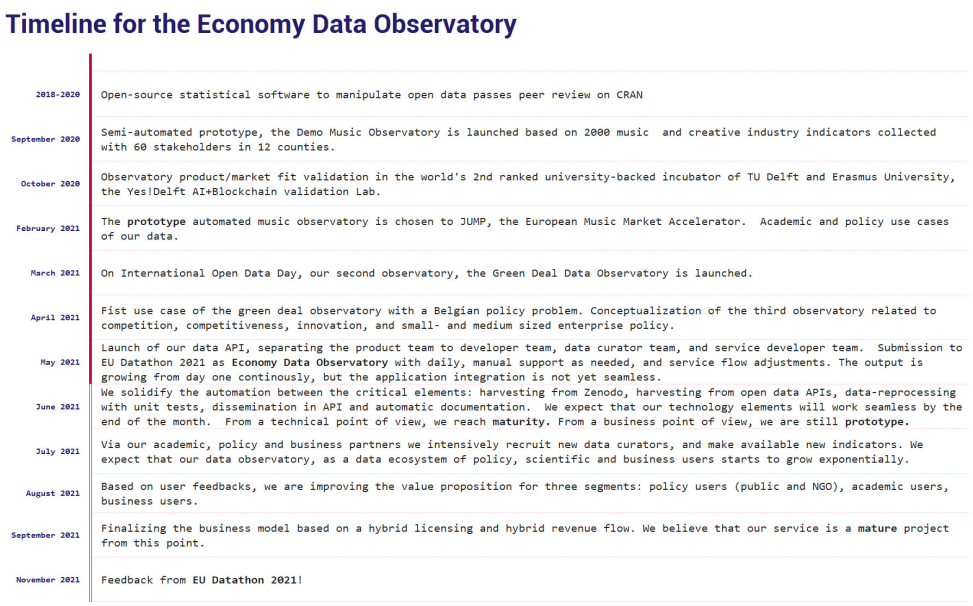
Our observatory is monitoring the European economy to protect the consumers and the small companies from unfair competition both from data and knowledge monopolization and robotization. We take a critical SME-, intellectual property policy and competition policy point of view automation, robotization,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



> 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.



4.1 Social Attitutes 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 suvey harmonization and regional coding efforts.

See our tutorial:

Regional Geocoding Harmonization Case Study - Regional Climate Change Awareness Datasets

4.2 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.

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)

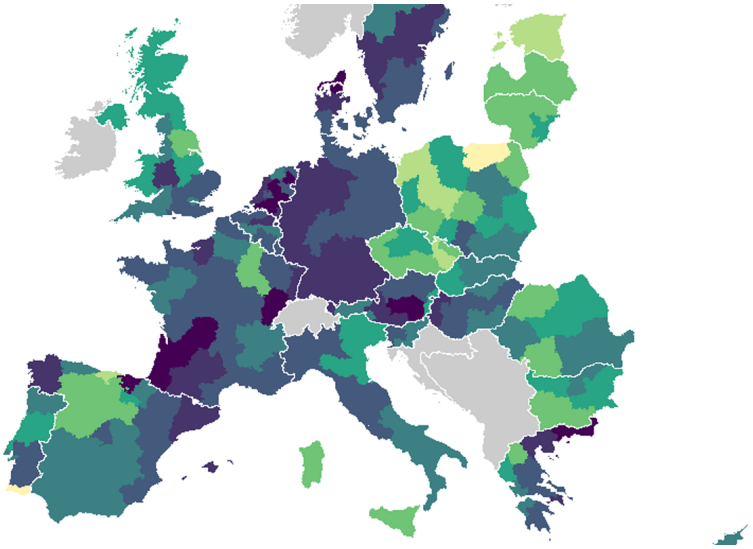


Figure 4.1: Regional attituteds to climate change, from our survey regionalization tutorial

- top-down indicators that show the structural (concentration) changes in the European economy
- connect them to patent databases

These indicators are particularly interestign, 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.

4.2.1 Knowledge Monopolizations, Killer Acquisitions

In killer acquisitions, a large company, for example, in pharmaceutical or technology fiedls, 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 describe by NACE) are more and less effected?
- How is patent concentration changing?

4.3 SME Activity Indicators

4.4 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.)

4.5 Sensory Data Measuring Changes in Economic Activy

Chapter 5

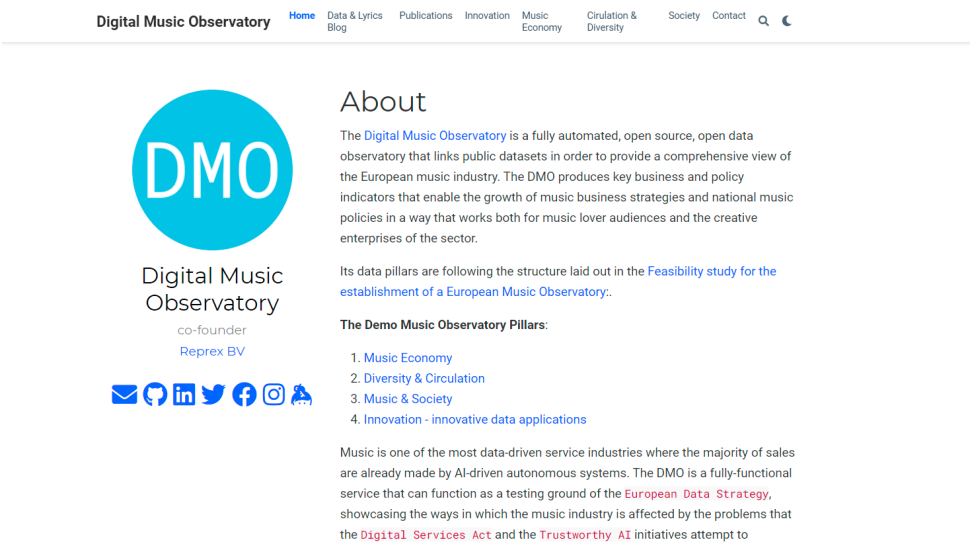
Music Data Indicators

The Digital Music Observatory (DMO) is a fully automated, open source, open data observatory that links public datasets in order to provide a comprehensive view of the European music industry. While the current music ecosystem provides European consumers with ease of access, its effects on artists and creative businesses have been devastating. The DMO produces key business and policy indicators that enable the growth of music business strategies and national music policies in a way that works both for music lover audiences and the creative enterprises of the sector.

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.



5.1 Music Economy

5.1.1 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.

5.1.2 Supply Indicators

Our Music Economy / Supply indicators are related to the supply of new music.

5.1.3 Price Indicators

5.2 Music Diversity

5.2.1 Gender, Language, Ethnic and Other Inclusion Attributes

5.3 Music Circulation

5.3.1 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 straightforward, 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 recodings 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.

5.4 Music & Society

5.4.1 Social Attitutes to Music

5.4.2 Participation in Music

Chapter 6

Partnerships

We are looking for policy partners to win the EU Datathon Challenges.

"To take part, you should propose the development of an application that links and uses open datasets. Your application should showcase opportunities for concrete business models or social enterprises. It is also expected to find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of open data."

We are planning to contest 2 or 3 of the challenges with our automated data observatories, and we are particularly looking for public and NGO policy partners who have an interest in these topics: European Green Deal, and Digital Age, particularly with the European Data Governance Act, Digital Services Act, and their impact on creative industries and their audiences.

Data observatories are recognized terms by the EU, OECD, and UNESCO to create long-term data collection, processing, and dissemination programs. The EU alone currently (co-)finances about 60 such programs which are often managed by large consultancies or research universities. Having reviewed more than 70 of these publicly funded data programs, we realized that none of them utilizes open data.

Our understanding of the problem is that both governmental open data (released under the EU Open Data Directive) and scientific open data is processed for an original use. It does not follow any processing standards, and usually comes without proper documentation. We have been developing research automation solutions to overcome these problems: peer-reviewed, open-source software that re-processes, documents, and validates various forms of open data. Our two demo applications are processing many previously unused open data sources – one is a demo for the planned European Music Observatory (with a projected budget of around 9 million euro) and the other is the future Green Deal Data Observatory.

The EU Datathon prize is a very prestigious prize that requires showcasing a concrete business case. We want to demonstrate, along with our partners, that we can reduce the data acquisition costs of publicly-funded research projects, as well as the private research costs of a big consultancy, partly through a higher use of open data, and partly via automated, not-billed or not-credited hours of manual data validation, documentation, and error-prone manual processing.

6.1 Policy Partners

We are looking for policy partners who want to use open data from various governmental (including publicly funded surveys), scientific, and big data sources, processed and validated to meet scientific standards; or who want to build use cases for trustworthy AI and data governance policy papers. We provide peer-reviewed statistical software solutions, daily data harvesting, and re-processing to meet our partners' research agenda.

In return, we ask academic partners to: 1. Help us curate data – tell us what sort of information is missing for their research agenda, and select what is valuable and what is not; 2. Use our validated,

correctly documented, open datasets, identified with DOIs (possibly with an embargo period) in their peer-reviewed scientific output; 3. Support the growth of our data observatories by including them in their data acquisitions plans; 4. Publicly endorse with testimonials the observatory they use for the EU Datathon Prize.

6.2 Academic Partners

We are looking for academic partners who want to use open data from various governmental (including publicly funded surveys), scientific, and big data sources, processed and validated to meet scientific standards. We provide peer-reviewed statistical software solutions, daily data harvesting, and re-processing to meet our partners research agenda.

In return, we ask academic partners to

1. Help us curate data – tell us what sort of information is missing for their research agenda, and select what is valuable and what is not;
2. Use our validated, correctly documented, open datasets, identified with DOIs (possibly with an embargo period) in their peer-reviewed scientific output;
3. Support the growth of our data observatories by including them in their data acquisitions plans;
4. Publicly endorse with testimonials the observatory they use for the EU Datathlon Prize.

6.3 Business Partners

We are looking for business partners to win the EU Datathon Challenges.

”To take part, you should propose the development of an application that links and uses open datasets. Your application should showcase

We are looking for a consulting partner to form a joint project with our open collaboration team of data scientists and submit proposals to at least 2 of the 3 challenges of the EU Datathlon with the objective of winning the first prize. We are particularly looking for a first class consultancy to help build a “showcase … for concrete business models [… and] find suitable new approaches and solutions to help Europe achieve important goals set by the European Commission through the use of [our observatories’] open data.”

Because of the way the challenge is formulated, winning the prize gives a natural advantage for our consulting partner to win future policy consulting work in the three strategic objective areas of the European Commission. Furthermore, we believe that our research automation technology and know-how can significantly reduce non-billable hours in research and validation, as well as in sales preparation and re-sale.

The partners of the Datathon Challenge are important EU and member state organizations, including the World Bank Group, FAO of the United Nations, the European Central Bank, the EU IP Office, and EFTA. We believe that winning this prize could give a competitive edge for our partners, given the high profile and rigor of the competition.

Chapter 7

Tools

We do reproducible products. If it works once, it must work all the time, on Windows, Mac, Linux, in Word, PDF, html, using a laptop, a cloud server, a smartphone or a table.

- We use Keybase for internal communication
- We are slowly bringing everybody on board with Git for online collaboration and file syncing. For a short time, we'll still use Google Docs.

If you work with us, you must adhere to the Contributor Covenant

The pledge starts with these paragraphs:

We as members, contributors, and leaders 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, religion, or sexual identity and orientation.

We pledge to act and interact in ways that contribute to an open, welcoming, diverse, inclusive, and healthy community. Please, take the time to read once the entire pledge.

7.1 Internal communication & collaboration

7.1.1 Keybase

No more long emails. Nobody's left out. No more misunderstandings. No more waiting for feedback.

What do we need?

- A simple-to-use communication tool where you can opt-in to be informed on a real time basis, or completely ignore us for days – similar to Whatsapp Group, Microsoft Teams, Google Hangout, Slack;
- A single, secure storage for shared documents – similar to Google Drive, Microsoft One Drive, Dropbox;
- An integration with Github – a critical feature to at least the part of our team that is working on software code or long-form technical documentations;
- Ability to oversee communications with more than 50 people with an option to streamline per topic, group, etc.;
- To avoid using emails boxes, Whatsapp, Viber, etc. – part-timers are not flooded with messages at all times, while full-timers can reach out to each other at any time;

- A space where we can communicate all the time without interrupting calls and meetings with clients on other platforms; An integration to a project management tool that we will use for larger projects;
- A tool as simple as possible, light touch;
- Since we work in the open source, open data, open collaboration community, we would like to use something that is open source – but we are willing to pay for solutions.

And the winner is....KEYBASE.IO

Keybase is a very neat, simple, lightweight team management / chat / social networking application that is extremely focused on privacy, security and encryption.

Keybase Key features

- Secure instant messaging, even with a timed self-destruction feature (e.g. for sharing passwords); Starts a Google Meet or Zoom video call natively with a single command;
- Brings your Whatsapp chat to the more private and secure keybase chat on the fly;
- Team chat rooms in real time. You can filter where you want to be involved, and you can always opt-out;

-K-Drive (similar to OneDrive, Google Drive, Dropbox) – only for our team, and fully encrypted; Works with Github, and it even offers a more private version of Private Github Repos, encrypted gits; An integration with other platforms; It is neat, open source, simple, clean, and usually appreciated more in the open source community than Slack, its big corporation rival.

Practical steps you need to follow to use Keybase

1. Download & install Keybase from https://keybase.io/ on your computer.

An easy procedure. Create yourself a professional login name – similarly to a professional github account, a professional email, etc. (you cannot change the name afterwards)

2. Once you log in to the computer, go to *Devices*, and *Create a paper key.* Write this on paper, or print it, and store it somewhere very safe (not near your computer). This to recover the access in case you lose access to all your devices.
3. You can use Keybase simultaneously on multiple devices – Install Keybase on your smartphone, tablet or any other device. You will be guided through installation & paired with your computer.
4. Shall you need them, you have *two recovery options*: the paper key and your smartphone.
5. If your smartphone breaks down and needs a replacement, you can add from your computer your new phone and deactivate the old one.
6. Once you are in, look up `antaldaniel`.
7. After a handshake Daniel will assist your smooth transition, help you find ways to our shared files, your project's files, and set up filters, so you are not flooded with information, while never left out, unless you choose to.
8. Initially, we set up the following "Big teams", as Keybase calls them, and we will send an invitation to join:

- `reprexfriends` for prospective team members, friends, and hoped-for-cooperation partners – partly for people we are discreetly asking to join us, or who want to know more about some of our work and cooperate with us;
- reprexcommunity is an open landing page for anybody, it is a public interface. If you every land there `antaldaniel` will take you to the appropriate, otherwise invisible team room.

Each big team has four special members for a smooth transition: Daniel and Zuzana to assist you with getting familiar with Keybase, zoombot (just type `!zoom` to create a Zoom call with the team members

present) and meetbot (that does the same with Google Meet, `!meet`). Daniel will gradually withdraw from some of the teams, once their support is not needed, though each team will have at least one Reprex co-founder present. We invite everybody to at least one team, but you can sign up to as many as you like, shall you find that convenient.

9. Whenever you are in a situation you want to ignore us (e.g. because you sit in your dayjob), just do it. If you have a smartphone, we are there, separated from your Whatsapp friends, work emails, and you can always check on us. We can always send you a secure (and even encrypted) message to get in touch, if needed. However, we will never ever bother you with long emails, Whatsapp messages and other annoying things.

Let's keep things short, give access to the full picture when needed, and let you find out what mix of response time, details and filters works best for you.

7.1.2 Git & Github

Git is a simultaneous collaboration for for any distributed team work - writing, programming, design work. Git is an open source software which makes sure that your teamwork files are always synchronized, clashes are avoided (you modify the same part of a file at the same time with Daniel.) The only hard part to move to Git is to make sure that Git properly works on your computer - it needs to be installed differently on all Linux distros, Mac OSX version all Windows versions. On Windows, you must make sure that Git is on the startup path. Once you are there, you'll life will be much easier.

- Keybase allows the group work on encrypted documents, like business proposals simultaneously using Git synch.
- RStudio allows us to work simultaneously on business proposals, blog posts, templates via Git.
- Github is allows us to use shared folders (`repositories` or simply `repos`) where we can track changes, modify the same thing at the same time, avoid or resolve conflicting edits, assign tasks, and much more.

If you do not have a github account yet, please, sign up now on github.com. Create a very professional profile. It is likely that you will use this profile for future works for decades, as Git is really becoming the norm of digital nomads, freelancers, and tech teams to work together.

Github is not the only service platform that allows distributed, collaborative teamwork. It has many alternatives, for example, GitLab – don't confuse them. We use Github.

- dataobservatory-eu is our private repo collection and private github collaboration platform.

7.2 Basic file management

If you work across different systems, like Windows 10, Mac, Linux, cloud, than you will easily get a problem with certain file names. Windows uses its own extensions, and does not like space in certain parts of the full filename.

What is a filename:

For example:

`C:\data\my-file.jpg`

- has a path `C:\data\`
- a filename `my-file`
- and an extension: `.jpg`.

Paths work differently on all systems, so we try to work with `relative path` instead of `full path`. For example, all pictures used to illustrate this book are stored in the `dataobservatory-eu/teambook` repo (folder) and the `images` subfolder.

1. Make sure that you use only `relative path` when you work collaboratively.
2. The reason why we prefer `markdown` files is that they are platform independent, simple, clean text files, that can be translated to Windows, Mac, Linux specific files.
3. Use filenames with `snake_case_formatting_without_space.txt` or `snake-case-formatting-with-dash.jpg`. Never use space in a filename, and if possible, avoid the use of uppercase.
4. The extension is just information for the computer how to handle the file. An if you use an `.txt` extension for a markdown file, it will be still editable, but misleading. The `.md` extension just says that preferable open it with a Markdown Editor, like RStudio or stackedit.io.

If you create new content, try to put it into one, single `markdown.md` or `markdown.Rmd` file.

For complex documents, like this Teambook itself, we use separate files per chapter. To make sure that they can be 'kintted' together into a single document, they must follow specifications.

- They must have the right filename
- Must be placed to the right folder
- May have a `YAML` header with further options (like adding a table of contents)

Let the repo editor put your single file to the right place.

1. Put your file to `drafts/draft.md`
2. Make sure that the image references are now `../images/draftimages/draft_01.jpg` format, where `../` tells the computer to go to the folder above drafts.

7.3 Writing, blogging

7.3.1 Markdown

Markdown is a simple "language", or rather a writing notation system that lets the word processor know that `*italics*` means *italics*, or `**bold**` means **bold** or `## Writing, blogging` becomes a level 2 heading, and `### Markdown {#markdown}` is a level 3 heading that can be referenced in the table of contents or internal hypertext links with `[Jump to markdown introduction](#markdown)`.

Markdown makes it possible that we can work on documents that will render fine in `html`, `docs`, `pptx`, `pdf`, `google docs`, or `md` for markdown files.

Markdown is a "markup language". But that is a too big word. It is more of a notation system for writing clear text that later can be automatically formatted.

For example `[Introduction to RMarkdown](https://rmarkdown.rstudio.com/lesson-1.html)` creates the hypertext link Introduction to RMarkdown in `html`, `docs`, `pptx`, `pdf`, `iWork Keynote`, `xlsx` or any file format that we need to create.

Markdown is very important for reproducible research and automation. It makes sure that the content and the form is fully separated.

- It is always the computer's task to make the formatting work perfectly and beautifully.
- It is sometimes the computer's task to fill out the document with text and numbers.
- It is a human task to desnbng beautiful documents, like blogposts, business proposals, research reports that are very easy to replicate automatically.

Markdown is not strictly a language, and it has many 'flavors' or notation version. The differences are usually related to the programming interface that allows the file conversion. If you are new to markdown, you can start with any flavor, the main text functions are the same.

- StackEdit is a wonderful tool, and we would recommend it, if it would not have been suspended from the Google Drive integration. You can try it out online. It is probably the cleanest interface to get you started.
- Heroku Markdown Editor integrates reasonably easily to your Google Drive. This means that you can edit our blogposts in your Google Documents.
- Docs to Markdown translates your Google Docs to Markdown. However, you must link your own images via a valid path.
- RStudio is our preferred offline application. It integrates seamlessly with Github. It is an integrated programming environment with four panels. If you use it as a markdown text editor, you can just minimize or close the programming tools.

RStudio uses the Rmarkdown, a special version of markdown, where you can insert little programs in `R`, `Python`, `C++`, `SQL`, `D3` or `Bash` scripts. For example, you can write a blogpost that retrieves data with a little program written by our musicology team from Spotify, or embeds a YouTube video, etc. The `code chunks` are visually separated from the proposal or blog post text, and you can ignore it if you do not *yet* write code.

- RMarkdown Cheat Sheet
- Rmarkdown Reference

Bibliography

(2017). Towards a harmonised methodology for statistical indicators — part 2: Communicating through indicators.

Antal, D. (2015). Javaslatok a Cseh Tamás Program pályázatainak fejlesztésére. A magyar könnyűzene tartós jogdíjnövelésének lehetőségei. [Proposals for the development of the Cseh Tamas Program grants. The possibilities of long-term royalty growth in Hungarian popular music].

Antal, D. (2019a). The competition of unlicensed, licensed and illegal uses on the markets of music and audiovisual works [A szabad felhasználások, a jogosított tartalmak és az illegális felhasználások versenye a zenék és audiovizuális alkotások hazai piacán].

Antal, D. (2019b). Private Copying in Croatia.

Antal, D. (2019c). Správa o slovenskom hudobnom priemysle.

Eurostat (2014). *Towards a harmonised methodology for statistical indicators — Part 1: Indicator typologies and terminologies*, volume 1 of *Towards a harmonised methodology for statistical indicators*. Publications Office of the European Union, 2014 edition.

InfoCuria (2014). OSA – Ochranný svaz autorský pro práva k dílům hudebním o.s. v Léčebné lázně Mariánské Lázně a.s. Case C-351/12.

InfoCuria (2017). Autortiesību un komunicēšanās konsultāciju aģentūra /Latvijas Autoru apvienība v Konkurences padome.

PwC (2008). Valuing the use of recorded music.