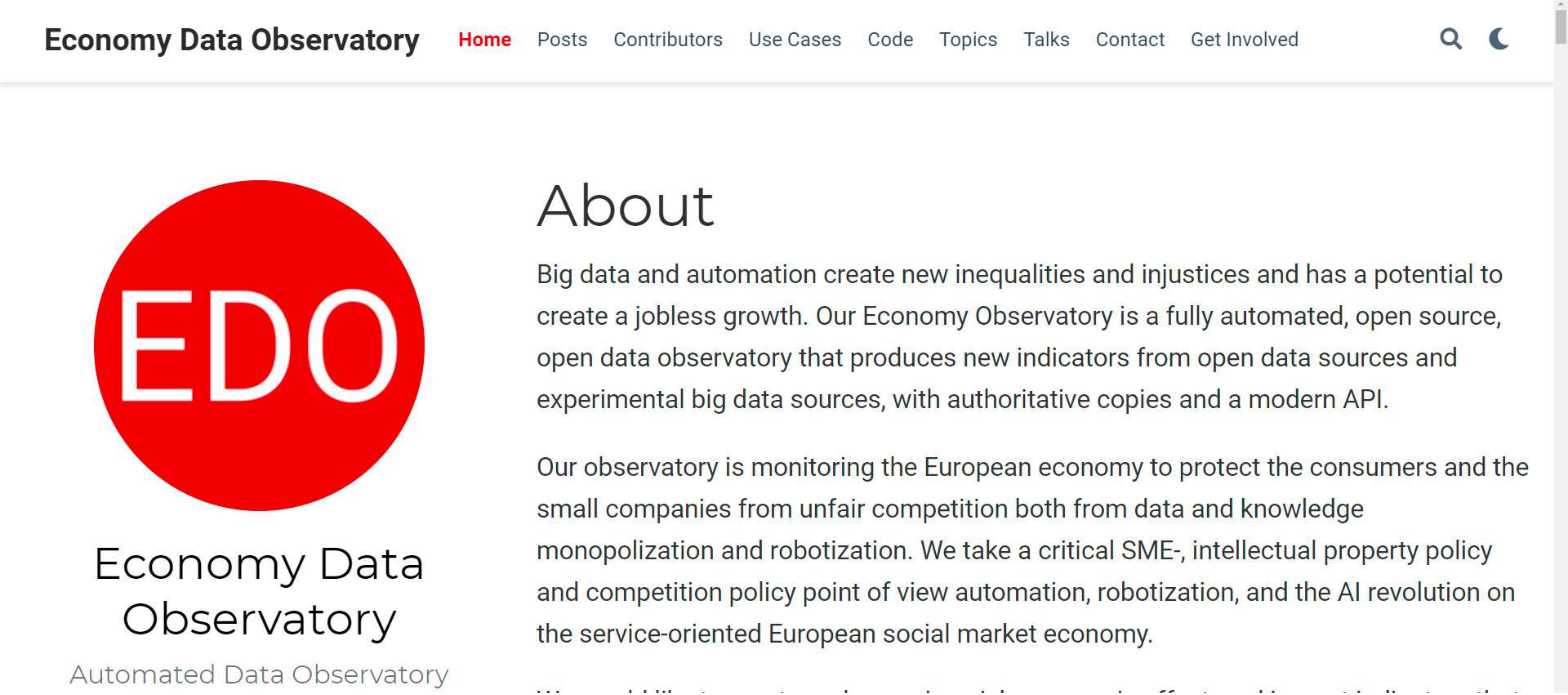


Economy Data Observatory

Automated Data Observatory

About

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.

Target audience of app

Business strategists and planners who work with various key business indicators; Public and non-governmental policy-makers who work with various impact and effect indicators; Academic researchers; Data journalists; Open-source developers.

Tagline of the app (140 characters), which could be published later

Our Future 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.

Description of the app (maximum 250 words)

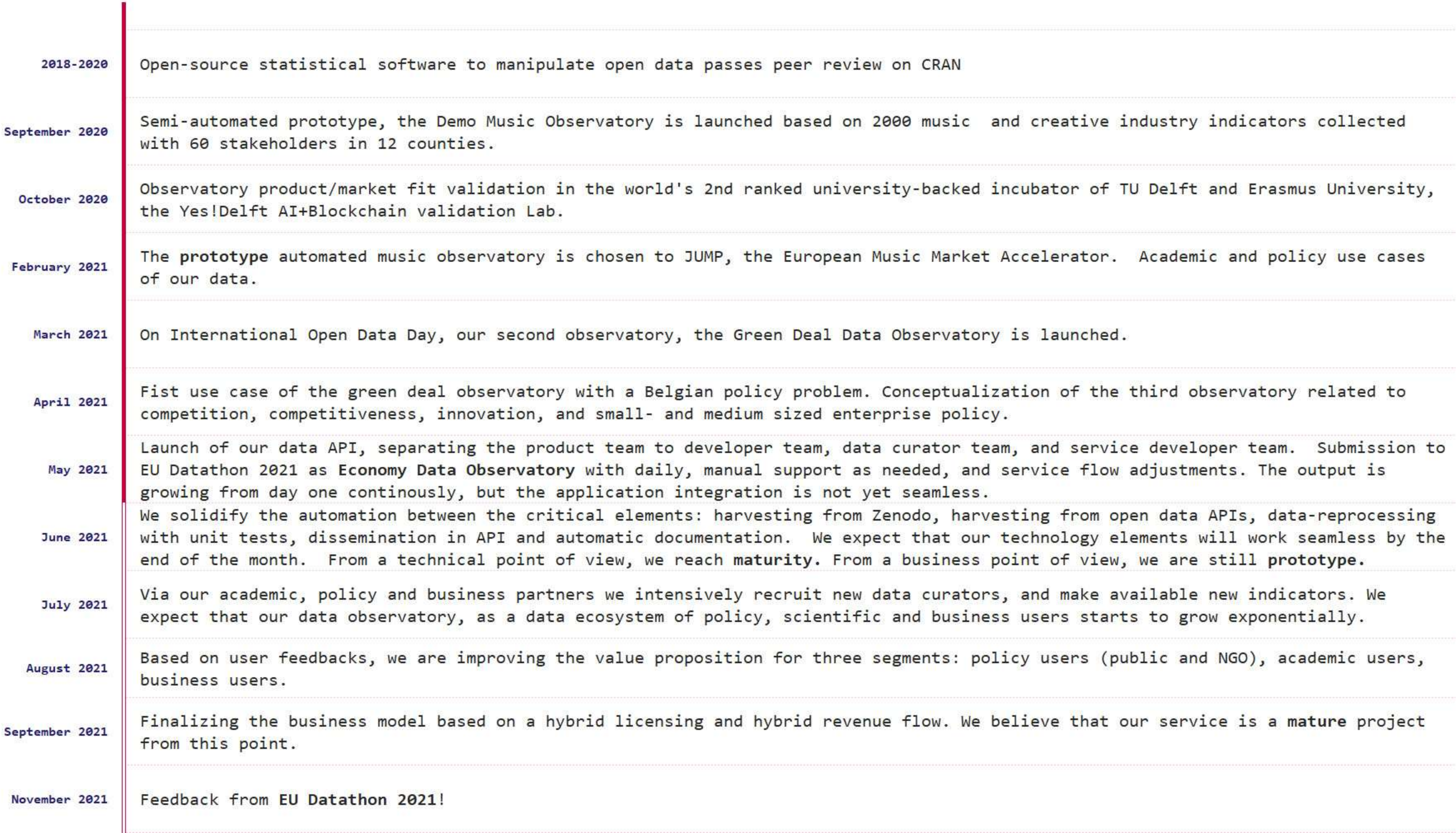
Big data and automation create new inequalities and injustices and have the potential to create a jobless growth economy. Our Economy Observatory is a fully automated, open source, open data observatory that produces new indicators from open data sources and experimental big data sources, with authoritative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

Our observatory monitors the European economy to protect consumers and 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 of 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, and signs of undermining key economic and competition policy goals. Our goal is to help small and medium-sized enterprises and start-ups to grow, and to furnish data that encourages the financial sector to provide loans and equity funds for their growth.

Project Plans & Readiness

Timeline for the Economy Data Observatory



2018-2020	Open-source statistical software to manipulate open data passes peer review on CRAN
September 2020	Semi-automated prototype, the Demo Music Observatory is launched based on 2000 music and creative industry indicators collected with 60 stakeholders in 12 counties.
October 2020	Observatory product/market fit validation in the world's 2nd ranked university-backed incubator of TU Delft and Erasmus University, the Yes!Delft AI+Blockchain validation Lab.
February 2021	The **prototype** automated music observatory is chosen to JUMP, the European Music Market Accelerator. Academic and policy use cases of our data.
March 2021	On International Open Data Day, our second observatory, the Green Deal Data Observatory is launched.
April 2021	Fist use case of the green deal observatory with a Belgian policy problem. Conceptualization of the third observatory related to competition, competitiveness, innovation, and small- and medium sized enterprise policy.
May 2021	Launch of our data API, separating the product team to developer team, data curator team, and service developer team. Submission to EU Datathon 2021 as **Economy Data Observatory** with daily, manual support as needed, and service flow adjustments. The output is growing from day one continously, but the application integration is not yet seamless.
June 2021	We solidify the automation between the critical elements: harvesting from Zenodo, harvesting from open data APIs, data-reprocessing with unit tests, dissemination in API and automatic documentation. We expect that our technology elements will work seamless by the end of the month. From a technical point of view, we reach **maturity.** From a business point of view, we are still **prototype.**
July 2021	Via our academic, policy and business partners we intensively recruit new data curators, and make available new indicators. We expect that our data observatory, as a data ecosystem of policy, scientific and business users starts to grow exponentially.
August 2021	Based on user feedbacks, we are improving the value proposition for three segments: policy users (public and NGO), academic users, business users.
September 2021	Finalizing the business model based on a hybrid licensing and hybrid revenue flow. We believe that our service is a **mature** project from this point.
November 2021	Feedback from **EU Datathon 2021**!

Legal & IP

- Our submission is an open collaboration among private persons, research organizations and an early-stage startup, Reprex BV, which is developing a business model to process open data with open-source software. We do not plan any changes in the timelines of the EU Datathon 2021. Our prototypes currently have no significant income, and we hope to receive little contributions from first users. These will be invoiced by Reprex, and current costs are expensed by Reprex.
- Reprex is supported by rOpenGov, which is a collaboration or R developers who write open source, peer reviewed software to access open data. Reprex's software were released together with rOpenGov. rOpenGov is hosted by the University of Turku, and their contribution to the project is in-kind.
- The development, data curator, and service development team members work in different organizations and their contribution is on a volunteer basis, and in-kind. They are not employed by Reprex, but Reprex pays some expenses (GitHub Action, Amazon AWS hosting.)

<u>Technology and Freedom To Operate:</u>

- We did not apply for a patent.
- We use only open-source technology, and we have complete FTO.
- Our critical components are released under MIT, GPL-2, or GPL-3 and similar licenses, and go through the quality control of peer review in releases, mainly on CRAN.

Business Planning

How do users are going to pay?

- We are aiming at the data acquisition budget of our public policy, NGO, consultancy and research institute clients, which is about 10-15% of their annual R&D spending ranging between €5k-50k.
- We also want to win public tenders of the EU, OECD and UN to run 'data observatories '. We want to keep as much as possible fully open and free.

What is the users ROI (rate on investment)?

- For a research-oriented organization, they get 2-4x times more data with us, but how this translates to a research product / research (wage) cost is being quantified in our pilot projects.
- The ROI is differently defined for a public policy organization, an NGO, a business consultancy or a scientific research organization.

What are your costs?

• Our server costs were won in the Yes!Delft incubator from Amazon and paid from the Reprex account. Github automation costs are paid by Reprex.

•So far, almost all our costs are fixed personnel costs, and all our team members work for free on a volunteer basis.

•In our current competition setup, all contributors are volunteers, and we try to get from users a small budget for small expenses. These will be handled by Reprex.

Technical Readiness Level

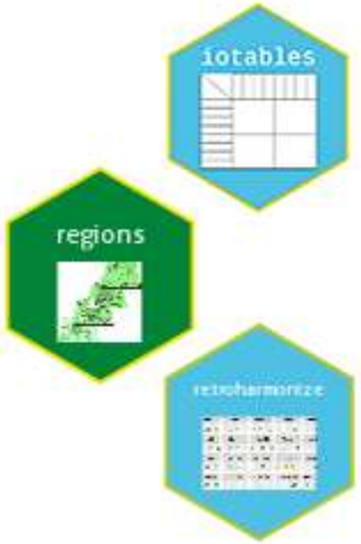
Technical readiness level:

•Different components are on levels 5-8.

•Our technological innovation lies in re-processing already existing public sector data, and mapping data sources, which enables us to create new and affordable features for AI apps.

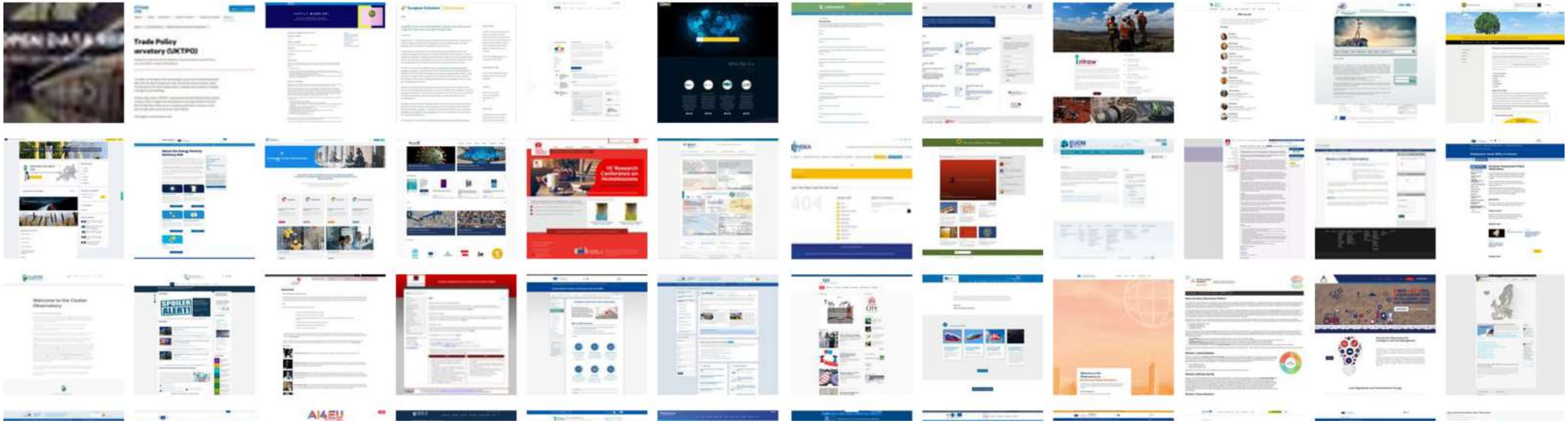
• We have passed TRL Level 5, and our prototypes are working in isolation. They are being tested as a seamless service flow now.

• We only use open-source technology which usually overstates TRL Levels. We use others' work, and contribute, too, our critical components are peer-reviewed statistical software releases. This makes or TLR higher than our business readiness.



Why is open data not trusted?

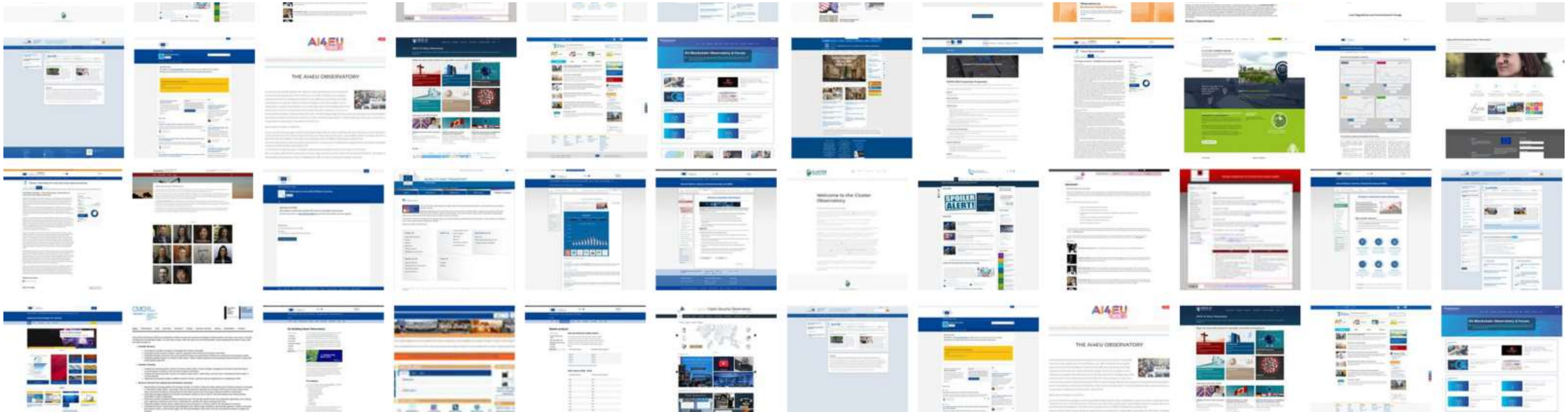
Why is Open Data Not Used?



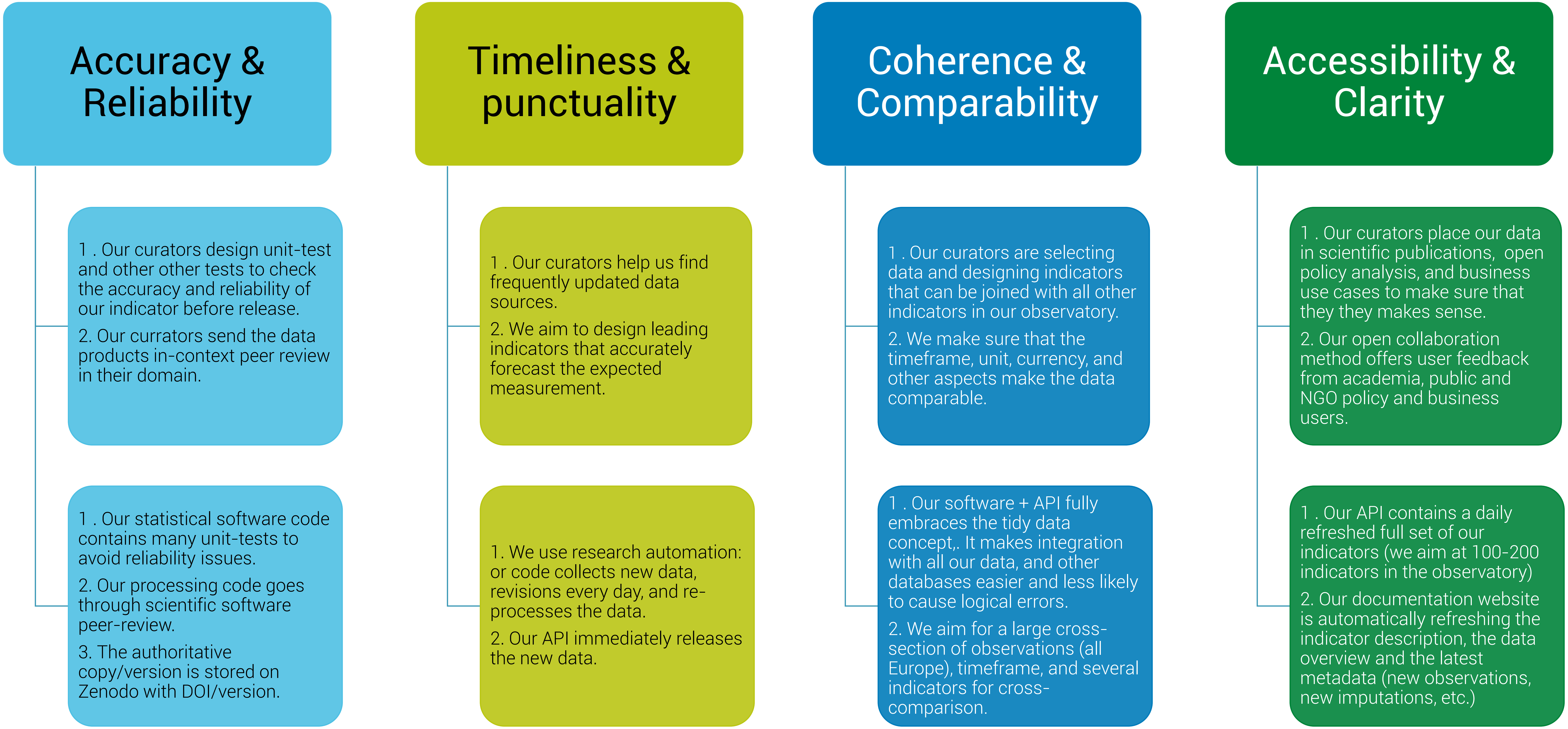
- Though data is usually valuable if it is not in isolation, open data is very difficult to join with other data. Almost never confirms the tidy data principles, which makes integration into databases or making composite indicators a very challenging data processing task.
- Haphazard use of measurement units (gram vs kilogram), currencies, metadata codes (regional boundaries change several thousand times just in the EU over a few years.)
- In short: open data requires investment into processing, unit testing, documentation to be usable. These are very costly operations, but we believe it can be done at scale and at a best value for money with open-source statistical code and research automation. This is what we do: create automated data observatories that reprocess, validate and automatically document open data to meet high statistical standards.

Why is Open Data Not Used?

- Open data is released after primary governmental, scientific or corporate use. It is not processed and organized to the new user's needs.
- The data is poorly documented, the primary user does not have an incentive to hire a data scientist or statistician to provide important metadata for the information: it needs to be reverse engineered to figure out important aspects of the data.
- The EU, OECD, an UN bodies are (co-)financing more than 60 permanent data collection points, so called '**observatories**' or '**data observatories**. Our market research found they almost never use any form of open data. We believe that it is a wasted opportunity to spend millions of euros on each data observatory's collection problem when billions worth of open data (at historical cost) is not even considered in them.



How can we build up the missing trust?



Accuracy & Reliability

1 . Our curators design unit-test and other other tests to check the accuracy and reliability of our indicator before release.

2. Our currators send the data products in-context peer review in their domain.

1 . Our statistical software code contains many unit-tests to avoid reliability issues.

2. Our processing code goes through scientific software peer-review.

3. The authoritative copy/version is stored on Zenodo with DOI/version.

Timeliness & punctuality

1 . Our curators help us find frequently updated data sources.

2. We aim to design leading indicators that accurately forecast the expected measurement.

1. We use research automation: or code collects new data, revisions every day, and re-processes the data.

2. Our API immediately releases the new data.

Coherence & Comparability

1 . Our curators are selecting data and designing indicators that can be joined with all other indicators in our observatory.

2. We make sure that the timeframe, unit, currency, and other aspects make the data comparable.

1 . Our software + API fully embraces the tidy data concept,. It makes integration with all our data, and other databases easier and less likely to cause logical errors.

2. We aim for a large cross-section of observations (all Europe), timeframe, and several indicators for cross-comparison.

Accessibility & Clarity

1 . Our curators place our data in scientific publications, open policy analysis, and business use cases to make sure that they they makes sense.

2. Our open collaboration method offers user feedback from academia, public and NGO policy and business users.

1 . Our API contains a daily refreshed full set of our indicators (we aim at 100-200 indicators in the observatory)

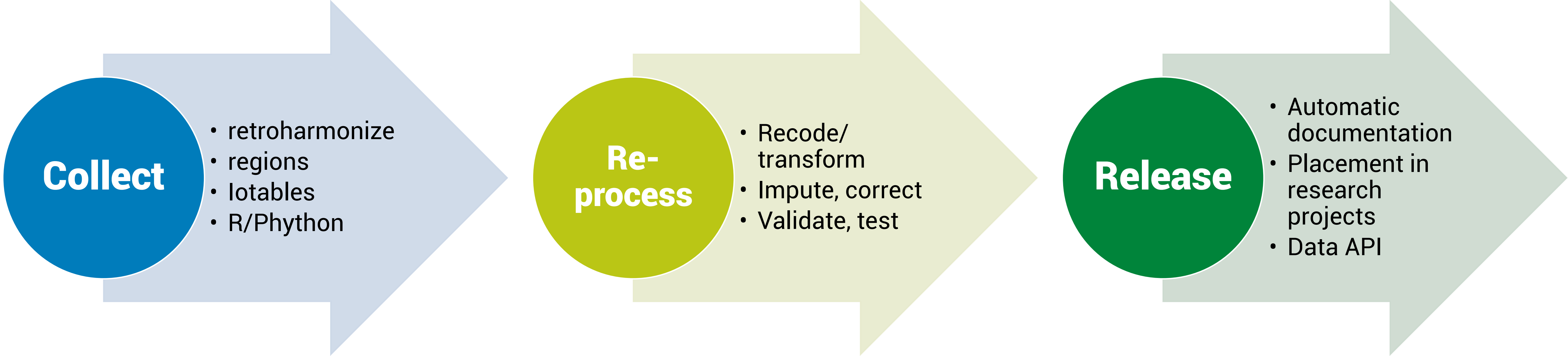
2. Our documentation website is automatically refreshing the indicator description, the data overview and the latest metadata (new observations, new imputations, etc.)

We aim to increase the quality of existing open governmental indicators, such as Eurostat products, and to design new indicators that are at least on the quality level of Eurostat's products. Our quality assurance follows the following methodology.

1. Towards a harmonised methodology for statistical indicators — Part 1: Indicator typologies and terminologies - 2014 edition (pdf)
2. Towards a harmonised methodology for statistical indicators — Part 2: Communicating through indicators (pdf)
3. Towards a harmonised methodology for statistical indicators — Part 3: Relevance for policy making (pdf)

What is our technology?

Technology – panning out gold from muddy open sources



Collect

- retroharmonize
- regions
- Iotables
- R/Phython

Re-process

- Recode/ transform
- Impute, correct
- Validate, test

Release

- Automatic documentation
- Placement in research projects
- Data API

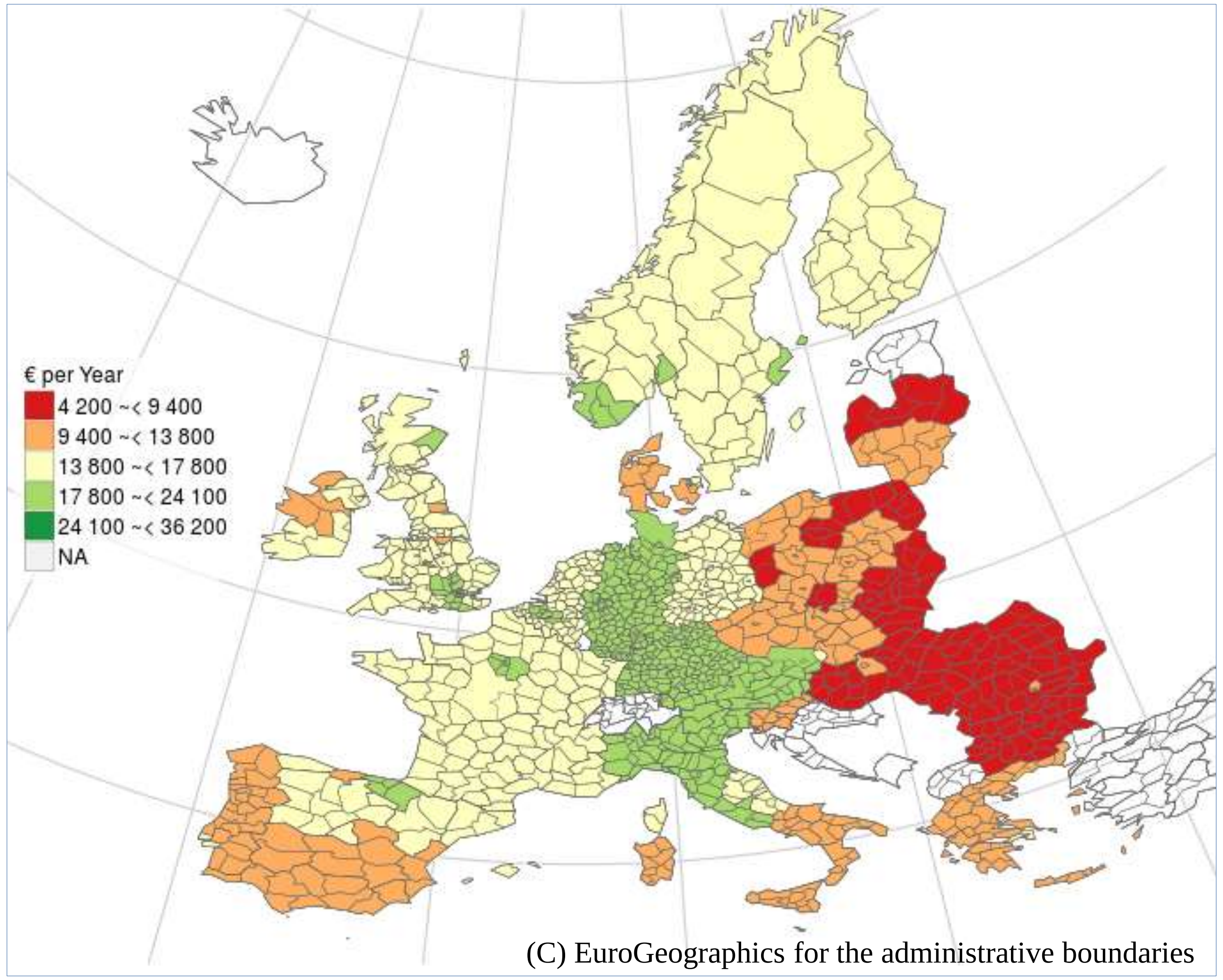
Our retroharmonize, regions and iotables software has each about 1000-2000 specialist users worldwide.
The users are potential collaborators to pan out more open data and potential clients to produce high-quality research products.

rOpenGov.org

Retrieval and Analysis of Eurostat Open Data with the eurostat Package

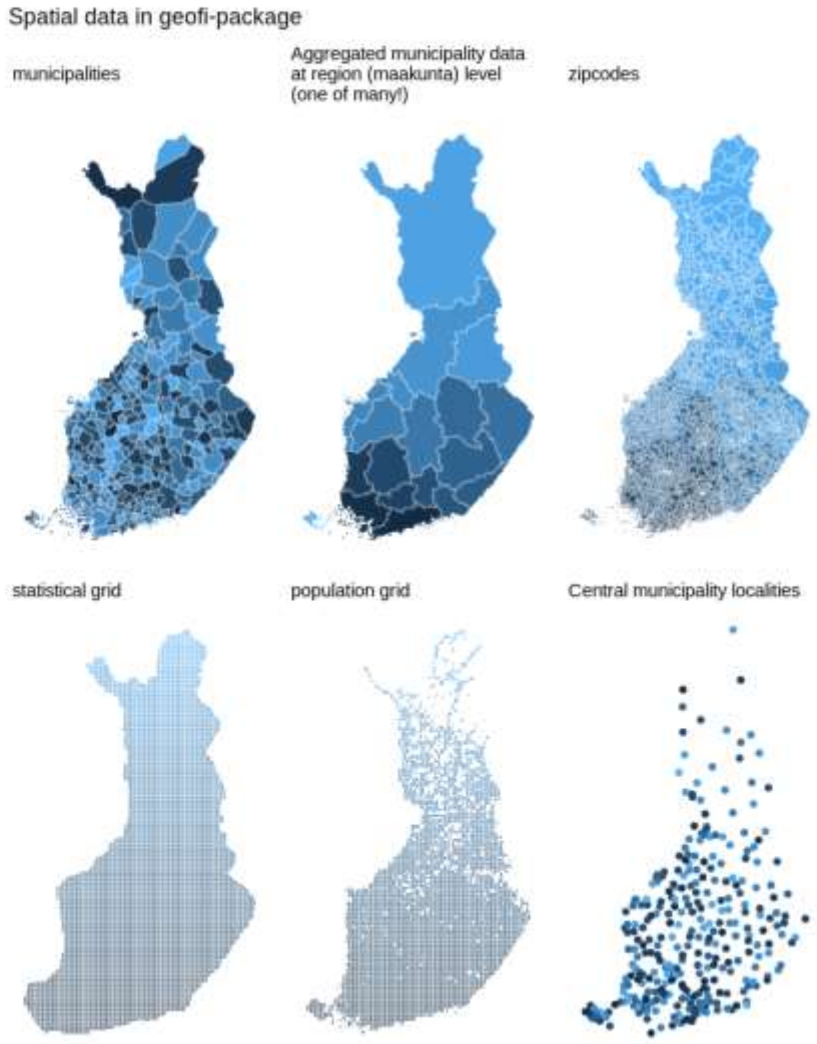
by Leo Lahti, Janne Huovari, Markus Kainu, and Przemysław Biecek

R Journal 9(1):385-392, 2017



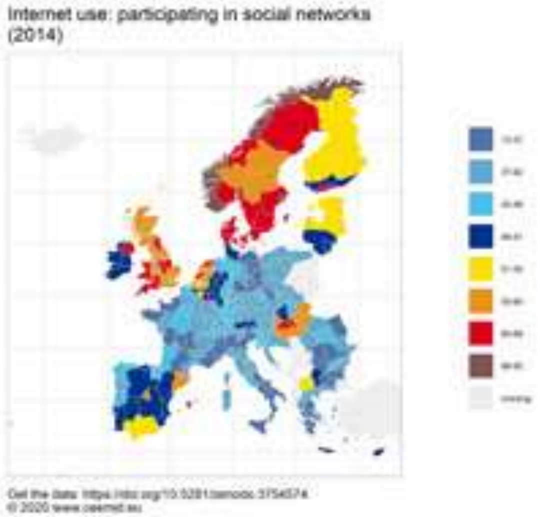
(C) EuroGeographics for the administrative boundaries

Eurostat open data: average household expenditure in 2011


Spatial data in geofi-package
municipalities
Aggregated municipality data at region (maakunta) level (one of many!)
zipcodes
statistical grid
population grid
Central municipality localities

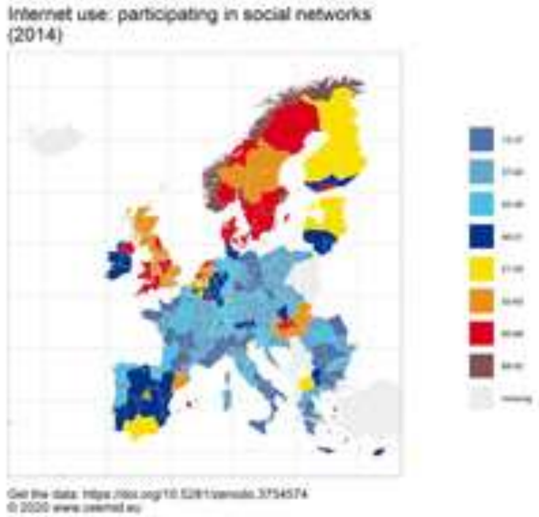
- Open developer network for open government data analytics in R
- 30 R packages in various stages of development; 10,000+ downloads/month
- Launched at the NIPS Machine Learning open-source software workshop 2013
- Active developers from 5 countries; coordinated by University of Turku, Finland
- This is a compilation of mature R packages that collectively provide tested tools to retrieve, refine, enrich, integrate, and analyse open government data from Eurostat, national statistical authorities, geospatial information, and other sources
- Seamless incorporation of open data streams with state-of-the-art statistical and probabilistic programming techniques and reproducible data science workflows

New Statistical Indicators



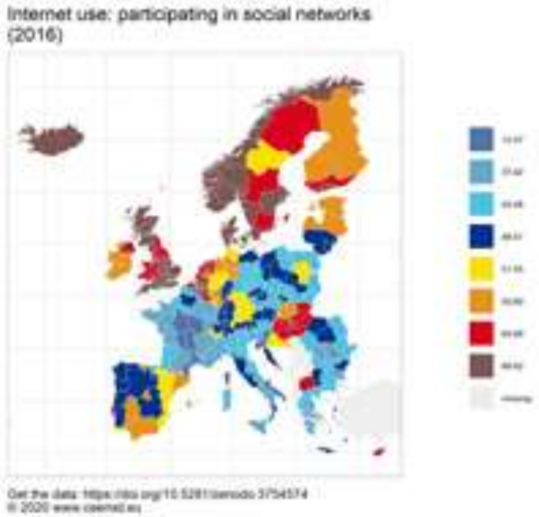
Internet use: participating in social networks (2014)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu



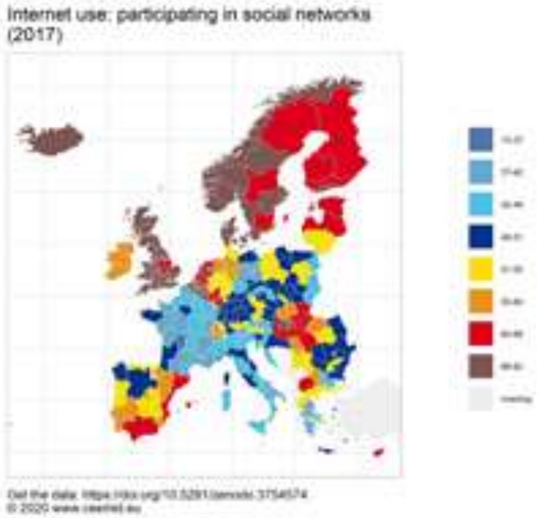
Internet use: participating in social networks (2014)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu



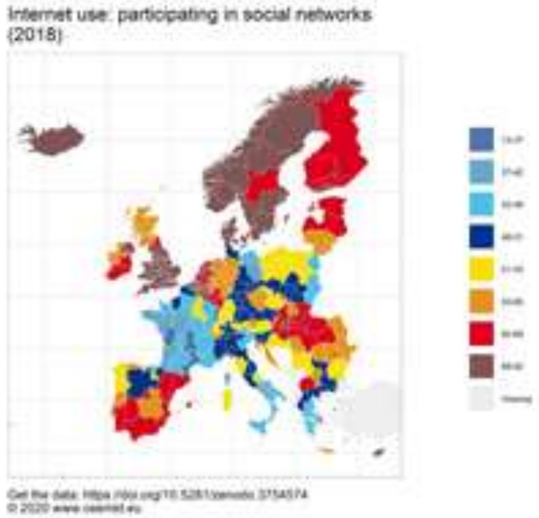
Internet use: participating in social networks (2016)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu



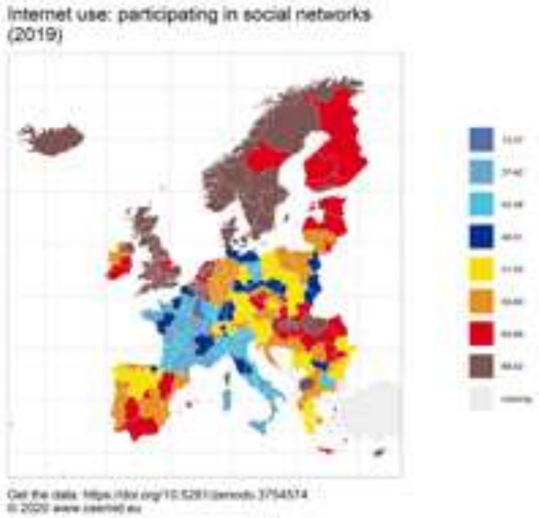
Internet use: participating in social networks (2017)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu



Internet use: participating in social networks (2018)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu

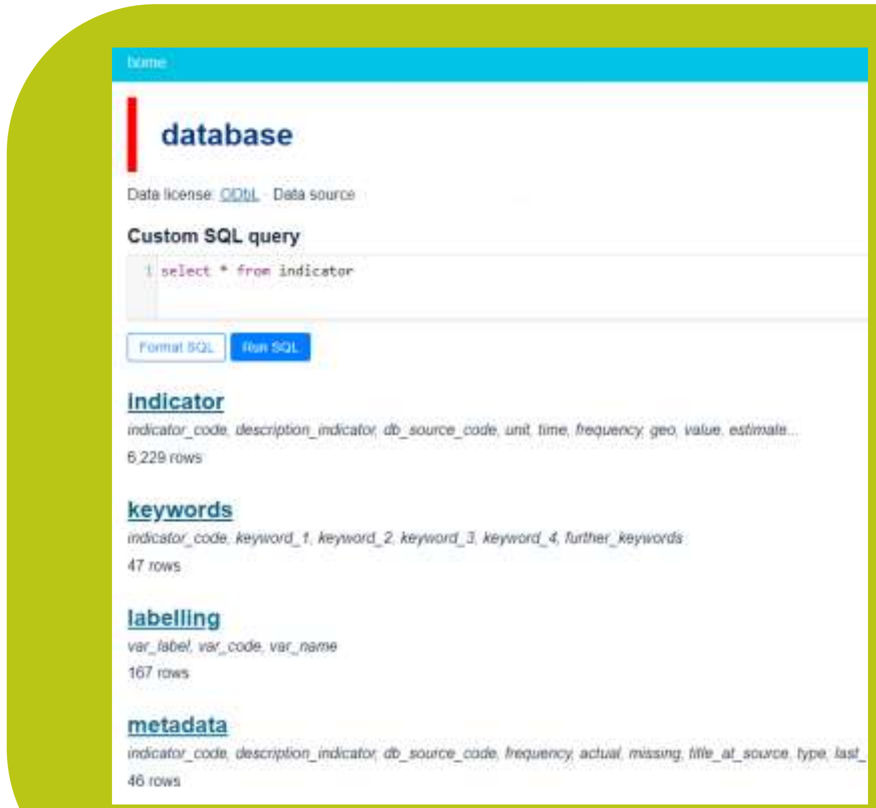


Internet use: participating in social networks (2019)

Get the data: https://doi.org/10.5281/zenodo.3754574
© 2020 www.ceemid.eu

- Statistical agencies and governments collect many-many times more data than what eventually is released as statistically aggregated data products (for example, the GDP or regional GDP indicator, literacy indicators, etc.)
- Via rOpenGov and our peer-reviewed statistical softwares, we have access to the raw data of Eurostat and other governmental and scientific agencies covered by the EU Open Data Directive, or similar legislation in other jurisdictions, and using the very same methodology of Eurostat, OECD, we can create similar statistical indicators ahead of the official publication date, or in details that are not published by the statistical agency.

What is our service?



database

Data license: ODbL · Data source

Custom SQL query

```
select * from indicator
```

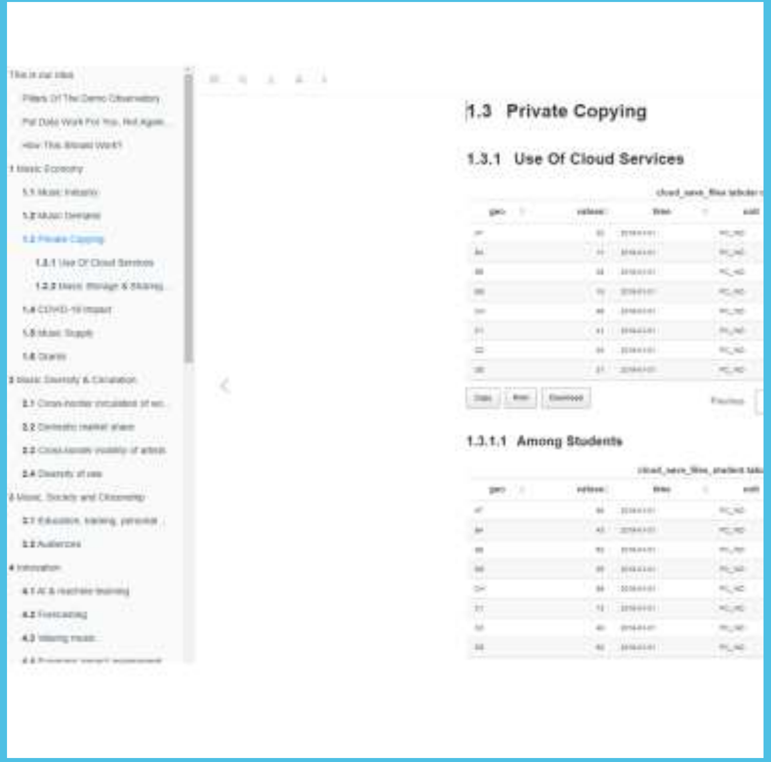
indicator
indicator_code, description_indicator, db_source_code, unit, time, frequency, geo, value, estimate...
6,229 rows

keywords
indicator_code, keyword_1, keyword_2, keyword_3, keyword_4, further_keywords
47 rows

labelling
var_label, var_code, var_name
167 rows

metadata
*indicator_code, description_indicator, db_source_code, frequency, actual, missing, title_at_source, type, last_*
46 rows

A modern, ODBL open API with daily refreshing indicator, processing and descriptive metadata. Download via SQL or in simple csv tables.

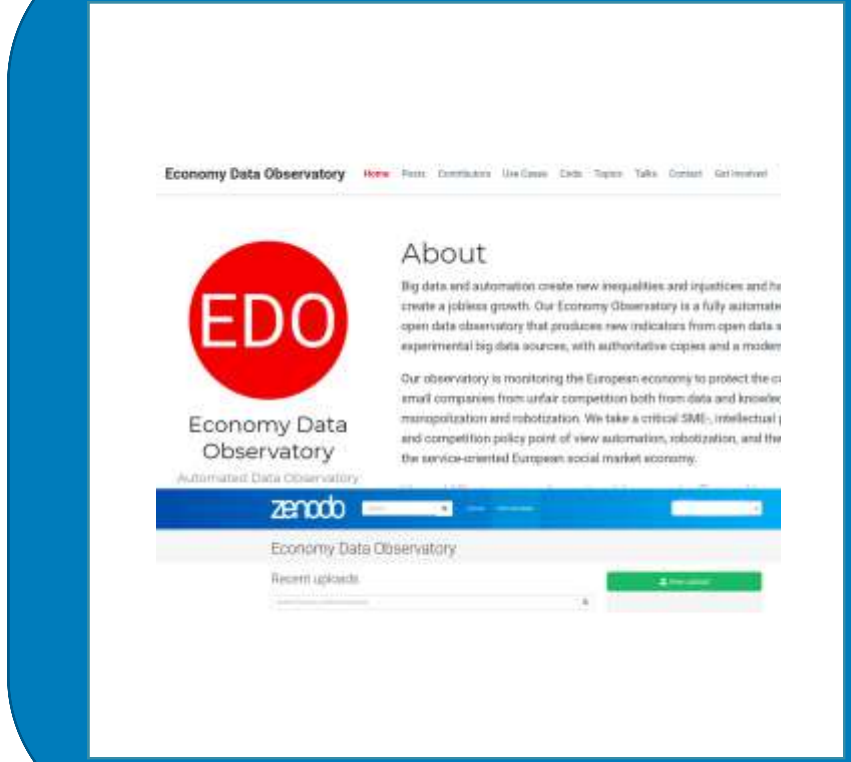


1.3 Private Copying

1.3.1 Use Of Cloud Services

1.3.1.1 Among Students

A daily refreshing long-form documentation with explanations, visualizations and human readable metadata.



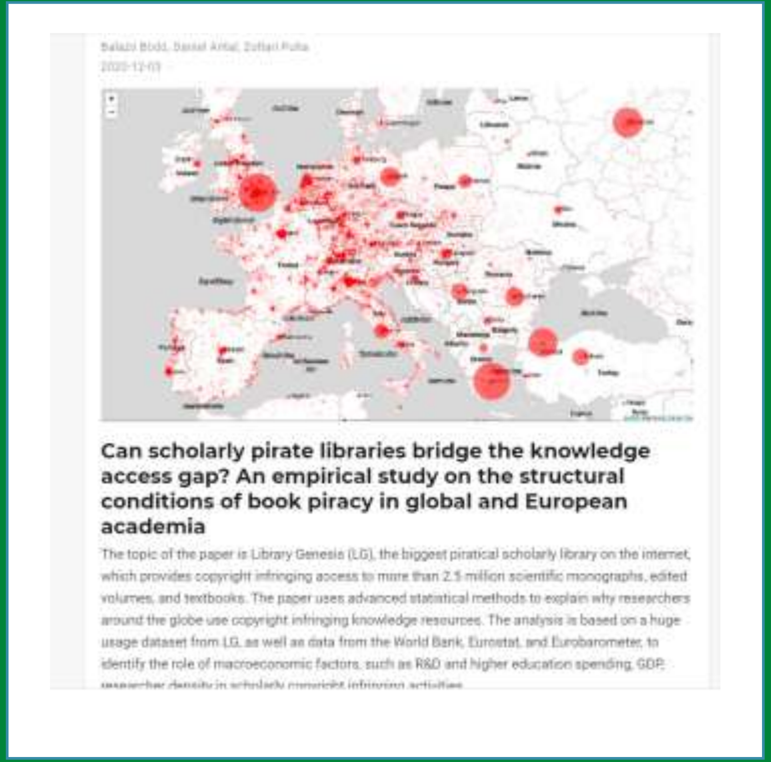
About

Big data and automation create new inequalities and injustices and ha create a jobless growth. Our Economy Observatory is a fully automate open data observatory that produces new indicators from open data s experimental big data sources, with authoritative copies and a moder

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

A dedicated community space on Zenodo for authoritative data copies with DOI.

Website with tutorials and use cases of the data from leading experts.


Balazs Bodo, Daniel Antal, Zoltan Puha
2020-12-03

Can scholarly pirate libraries bridge the knowledge access gap? An empirical study on the structural conditions of book piracy in global and European academia

The topic of the paper is Library Genesis (LG), the biggest piratical scholarly library on the internet, which provides copyright infringing access to more than 2.5 million scientific monographs, edited volumes, and textbooks. The paper uses advanced statistical methods to explain why researchers around the globe use copyright infringing knowledge resources. The analysis is based on a huge usage dataset from LG, as well as data from the World Bank, Eurostat, and Eurobarometer, to identify the role of macroeconomic factors, such as R&D and higher education spending, GDP, researcher density in scholarly copyright infringing activities.

We would like to integrate our data flow with your research workflow. Our service design team is looking forward to requests from public and NGO policy users, researchers and consultants.

Open governmental data:

We access re-usable public sector information covered by the Open Data Directive, and various freedom of information legislation.

Observatory specific text

Open scientifc data:

We access re-usable scientific data from the Zenodo repository

Observatory specific text

Data sharing:

We are encouraging our users and data curators to share their properietary data through us.

Data is getting exponentially more valuable in integration than in isolation. We incentives data sharing with 8 years of industry experience

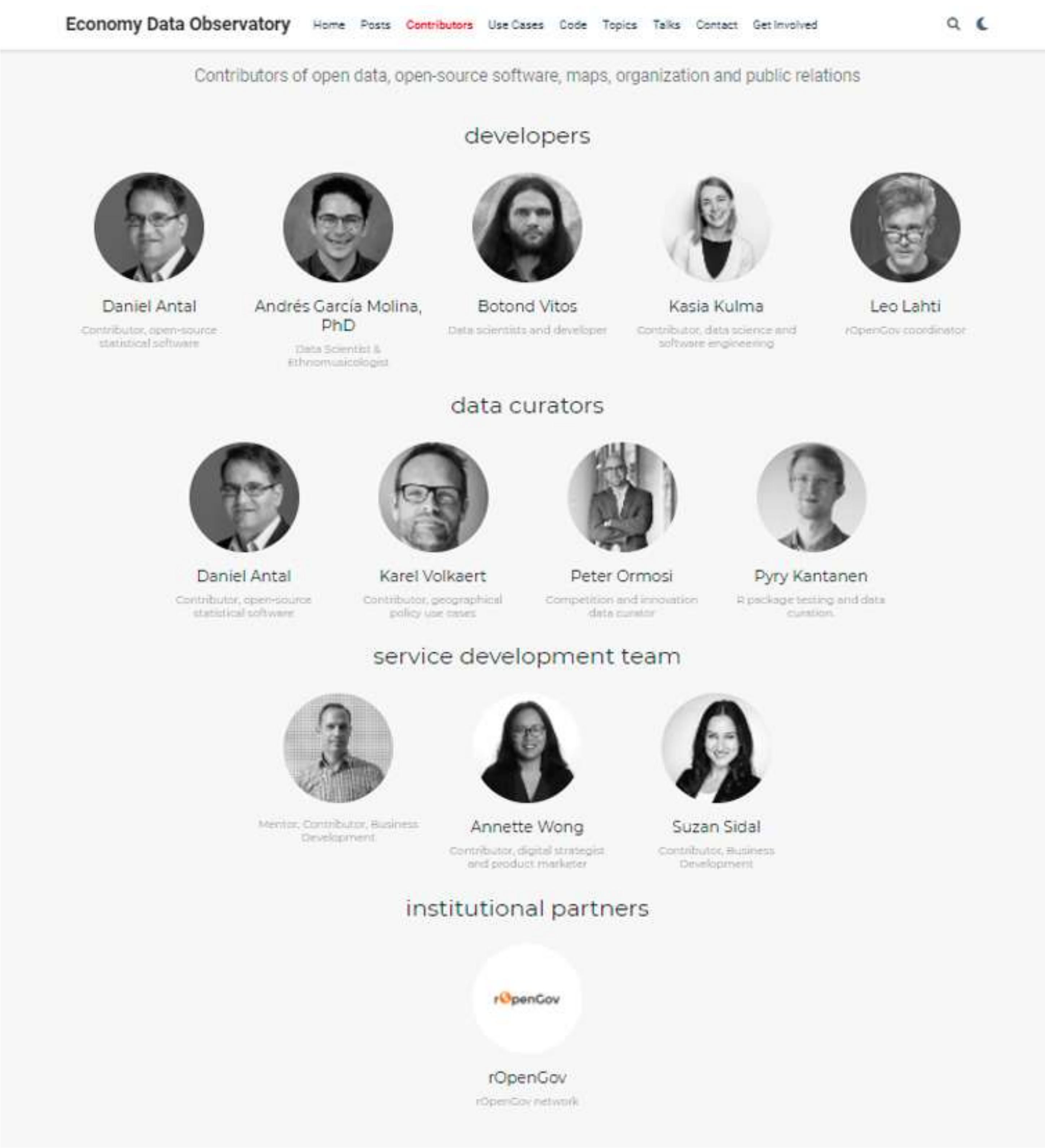
Big data sources, satellites:

We are monitoring various transitory but open APIs, satellite images and other continous "big data" sources and use novel statistical technology to capture them into permanent, reliable and timely statistical, business, policy or scientific indicators.

Data sources

It would be difficult to name all our data sources.

- We are working on connecting socio-economic data, survey data and sensory information on Europe's map.
- We are carrying out small area statistical estimates to map harmonized surveys about climate change attitudes more relevant on a regional and metropolitan area level.
- We are particularly interested in working the DG Competition Mergers database and the patent databases of the EU IPO. We believe that these data are open data under the Open Data Directive, but currently they are not served by an open API.



Contributors of open data, open-source software, maps, organization and public relations

developers

Daniel Antal
Contributor, open-source statistical software

Andrés García Molina, PhD
Data Scientist & Ethnomusicologist

Botond Vitos
Data scientists and developer

Kasia Kulma
Contributor, data science and software engineering

Leo Lahti
rOpenGov coordinator

data curators

Daniel Antal
Contributor, open-source statistical software

Karel Volkaert
Contributor, geographical policy use cases

Peter Ormosi
Competition and innovation data curator

Pyry Kantanen
R package testing and data curation.

service development team

Mentor, Contributor, Business Development

Annette Wong
Contributor, digital strategist and product marketer

Suzan Sidal
Contributor, Business Development

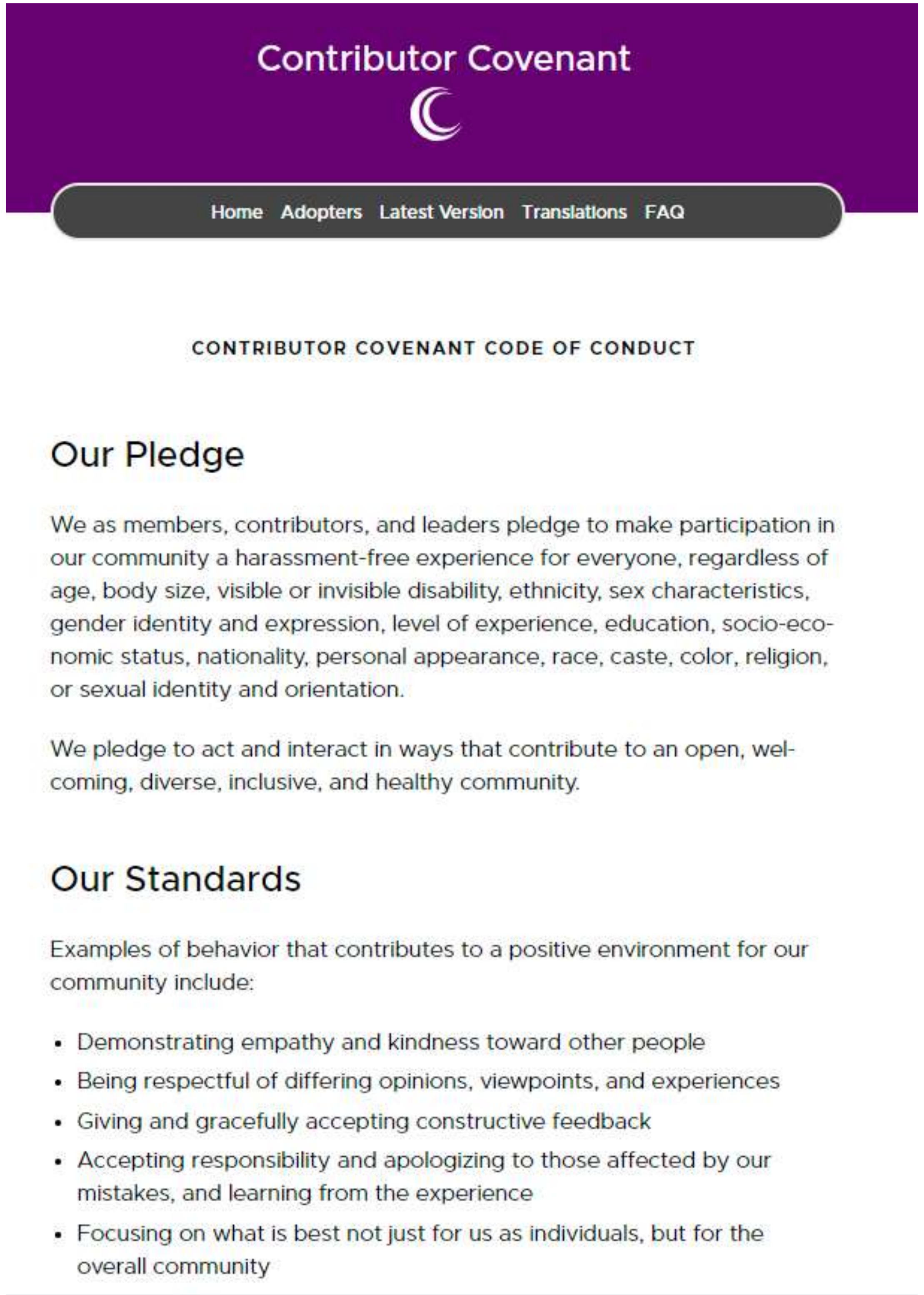
institutional partners

rOpenGov
rOpenGov network

Our **Economy Data Observatory** is being developed in an open collaboration with individuals, music industry stakeholders and research institutions.

The four team members in our EU Datathon 2021 submission form were selected in no particular order.

We are actively recruiting and added new contributors every day to our website.



CONTRIBUTOR COVENANT CODE OF CONDUCT

Our Pledge

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, caste, color, 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.

Our Standards

Examples of behavior that contributes to a positive environment for our community include:

- Demonstrating empathy and kindness toward other people
- Being respectful of differing opinions, viewpoints, and experiences
- Giving and gracefully accepting constructive feedback
- Accepting responsibility and apologizing to those affected by our mistakes, and learning from the experience
- Focusing on what is best not just for us as individuals, but for the overall community