Day 6: Data management and ethics

Guiding Principles for Data Management II

Sebastian Ramirez-Ruiz
Hertie School

What people think working with data looks like...

How it really is...

You have to wear many hats...

Research data management (RDM) lifecycle

Table of contents

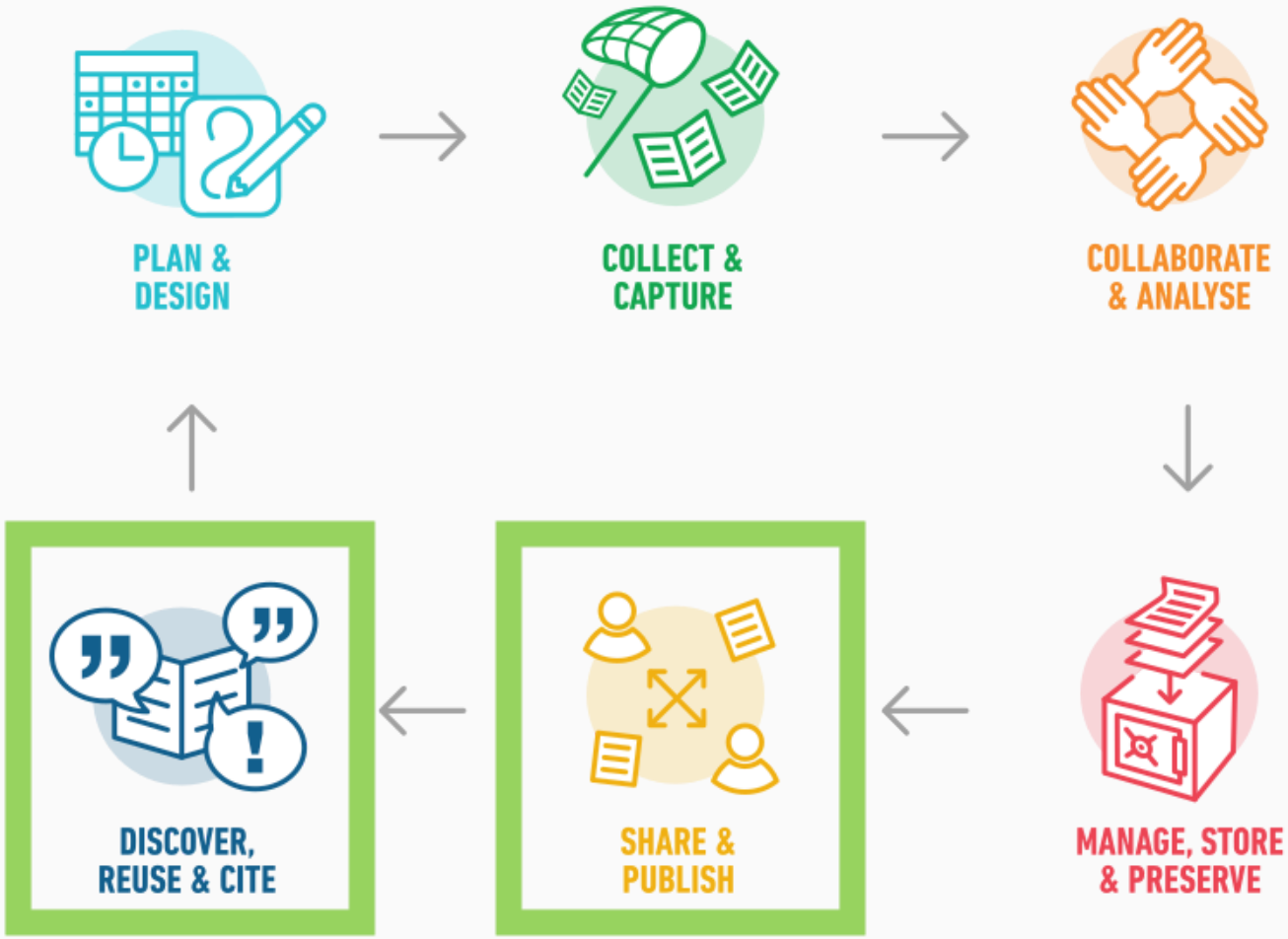
1. Archiving and publishing

2. Open government data

3. Discovering

Research data management (RDM) lifecycle

Hertie School



A show of hands!

Raise your hand if you:

- Can state the difference between data archiving and data publishing
- Can list some of the benefits of data publishing
- Can differentiate between different data publication services (data journals, self-archiving, a data repository)
- Know which data repositories fit the needs of different data sources
- Know of ways to promote your government data publication

Archiving and publishing

Promises of high-quality data

High-quality data have the potential to be reused in many ways. *Archiving* and *publishing* your data properly is at the core of making your data FAIR and will enable both your future self as well as others to get the most out of your data.

Data archiving

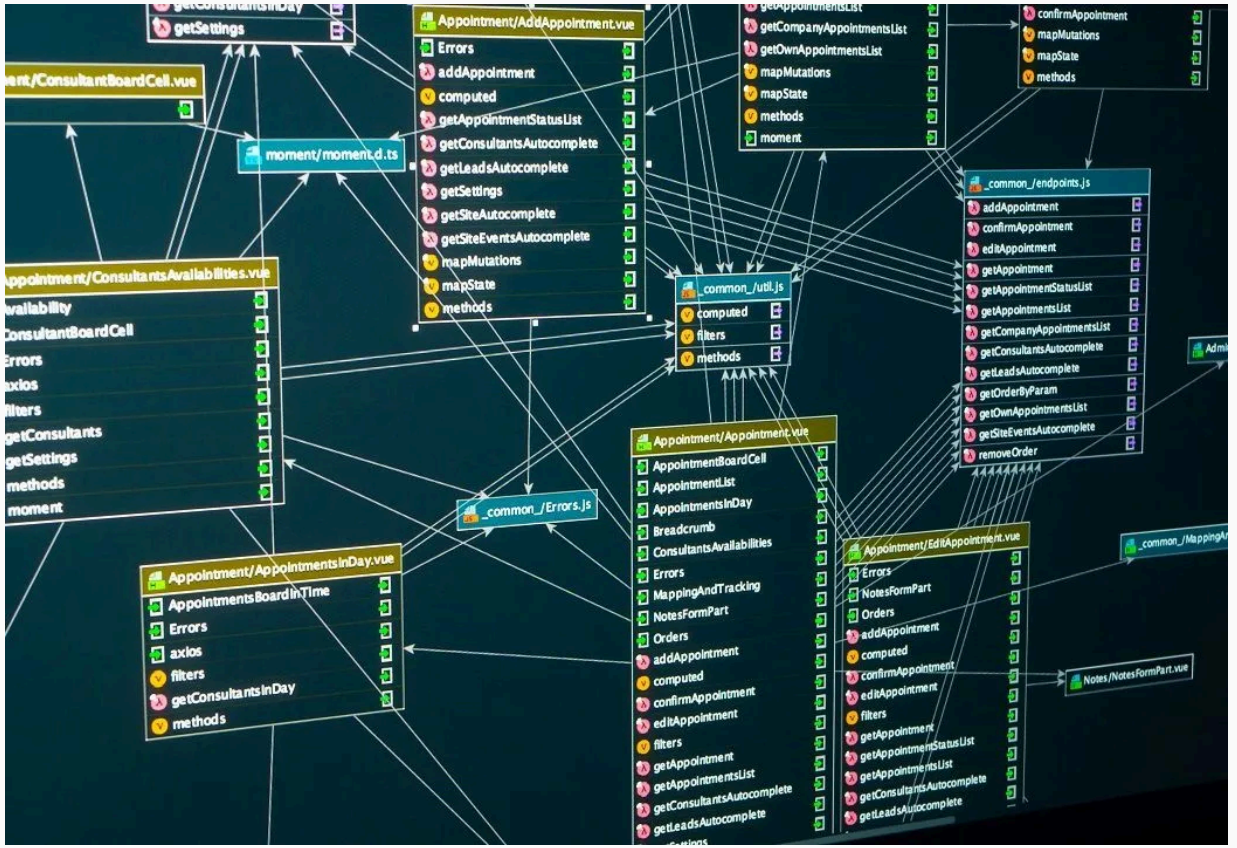
Archiving data for future reference

Data archiving is about storing and preserving data for the long term. When you archive your data, you make sure you can read and access the data later on. You can then also allow access to others for verification purposes when such a request arrives. In all cases, you should store your data safely, in a suitable file format, with adequate documentation.

Data publishing

Publishing data for reuse

To make your data reusable for purposes beyond the one for which you collected them, you can **publish your data**. Publishing your data is the act of publicly disclosing the research data you have collected, making them findable, accessible, and reusable.



Data publication in academia

Scientific progress

Data archiving and publication has direct benefits for the research itself (more robust), for the discipline and for science in general by enabling new collaborations, new data uses and establishing links to the next generation of researchers.

Norms of the field

Depending on field may be more prone. General practice in *computational social science*.

Drivers

- Funders
- Publishers

Selecting data for publication

Should you publish your data or shouldn't you?

Selecting data for publication

Should you publish your data or shouldn't you? And if so, which part of it?

Selecting data for publication

Should you publish your data or shouldn't you? And if so, which part of it?

- Sometimes this question is **straightforward to answer**:
  - the funder of the research demands it
  - strong researcher inclination
- Still, *not all data are created equal* and data publishing does involve an investment of researcher resources. **Some datasets have a more obvious reuse potential than others.**

Selecting data for publication (cont.)

Does your dataset have reuse potential?

- Does your data have potential value in terms of reuse,
  - national/international standing and quality,
  - historical importance,
  - uniqueness,
  - originality,
  - size,
  - scale,
  - costs of data production or innovative nature of the research?
- Could you foresee that secondary analyses on your data would benefit science? Or policy analysis?

If your answer to any of these is yes, your dataset has serious reuse potential.

Selecting data for publication (cont.)

Is your dataset reusable?

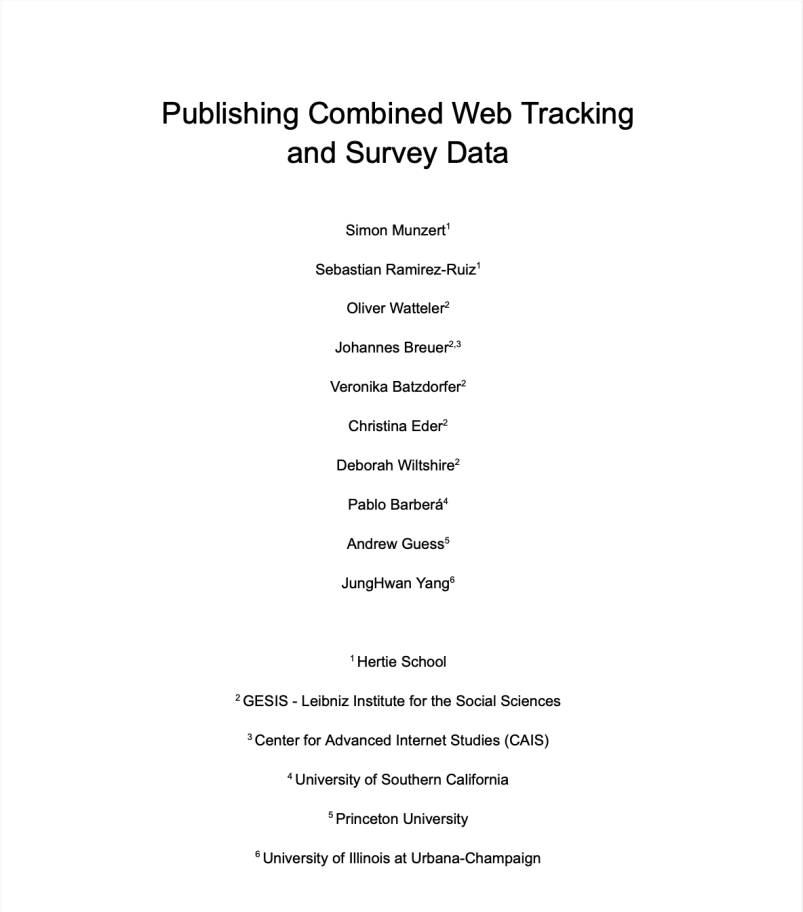
To be suitable for reuse, your dataset must be **functionally usable**.

- Can the data be accessed and utilized? (e.g., *readable format*)
- Is there sufficient metadata available to allow future users to comprehend your data? (e.g., *documentation*)
- Are there any legal restrictions that prevent the data from being published?

If you have addressed these practical considerations and your data hold potential value for reuse, you are 'good to go'.

Example from our own experience

Hertie School



Publishing Combined Web Tracking and Survey Data

Simon Munzert[1]

Sebastian Ramirez-Ruiz[1]

Oliver Watteler[2]

Johannes Breuer[2,3]

Veronika Batzdorfer[2]

Christina Eder[2]

Deborah Wiltshire[2]

Pablo Barberá[4]

Andrew Guess[5]

JungHwan Yang[6]

[1] Hertie School

[2] GESIS - Leibniz Institute for the Social Sciences

[3] Center for Advanced Internet Studies (CAIS)

[4] University of Southern California

[5] Princeton University

[6] University of Illinois at Urbana-Champaign

`Source`: Munzert, Simon, et al. "Publishing Combined Web Tracking and Survey Data." (2023). https://doi.org/10.31219/osf.io/y4v8z

How do I protect the privacy of my research subjects?

Hertie School

POLITICS & Media

MEOF
MEDIA EXPOSURE AND OPINION FORMATION

Table of Contents

1. Demographics	**8**
Wave number (wave)	8
Respondent ID (person_id)	8
Survey start time (starttime)	8
Survey end time (endtime)	9
Weight (weight)	9
Gender (gender)	9
Year of birth (birthyr)	10
Race (race)	10
Education (educ)	10
Marital status (marstat)	11
Children under age 18 in household (child18)	11
Employment (employ)	12
Household income (faminc_new)	12
State (inputstate)	13
Religion (religpew)	15
Born Again or Evangelical Christian (pew_bornagain)	15
Religious service attendance (pew_churatd)	16
Religion importance (pew_religimp)	16
Prayer frequency (pew_prayer)	17

`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

How do I protect the privacy of my research subjects?

Hertie School

```
R> dat$url[1:20]
```

```
##  [1] "mail.google.com/mail/u/0#inbox"
##  [2] "bing.com/search?q=fu%C3%9Fnagel+w%C3%A4chst+nach+oben+und+verhornt&form=EDGHPT&qs=PF&cvid=a97dc78299
##  [3] "preisvergleich-pilot.de/sport-und-freizeit.html"
##  [4] "http://tile-service.weather.microsoft.com:443"
##  [5] "http://3c.1und1.de:443"
##  [6] "https://www.trendfrage.de/"
##  [7] "questler.de/intern/paidmail"
##  [8] "http://www.ebesucher.de/surfbar/swhtmw.106_tobias_wutz"
##  [9] "http://www.google.de:443"
## [10] "http://www.ebesucher.de/surfbar/dermaerker"
## [11] "tracking.dpd.de/parcelstatus?query=01505160077643&locale=en_DE"
## [12] "ma-shops.de/shops/search.php?searchstr=kreuzz%C3%BCge&catid=904&submitBtn=Finden&lang=de&PHPSESSID="
## [13] "studien.usuma.de/cc3/survey/intweb.dll/project/wsusuma/CAWI170607_OG1/usrname=GMI&passwd=WW2017&PIN=2
## [14] "amazon.de/Samsung-J5550-Fernseher-Triple-Tuner/dp/B00U57X8OI/ref=sr_1_4?s=home-theater&ie=UTF8&qid=15
## [15] "keypanel.de/?ID=13"
## [16] "http://outlook.office.com:443"
## [17] "ci-marketing.de/_p.php?userid=3673&mailid=229933&mc=6495367958"
## [18] "http://www.farmerama.com:443"
## [19] "https://www.amazon.de/b/ref=gbph_ftr_m-8_ee12_page_263?node=7361369031&pf_rd_p=7bde80db-c7fc-47f3-a66
## [20] "http://www.webmie.com:443"
```

`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

How do I protect the privacy of my research subjects?


```
R> # parse Google requests
R> google_maps_df ← filter(dat, str_detect(url, "/maps"))
R>
R> # how much of the URLs are Google Maps URLs?
R> nrow(google_maps_df)
```

```
## [1] 12206
```

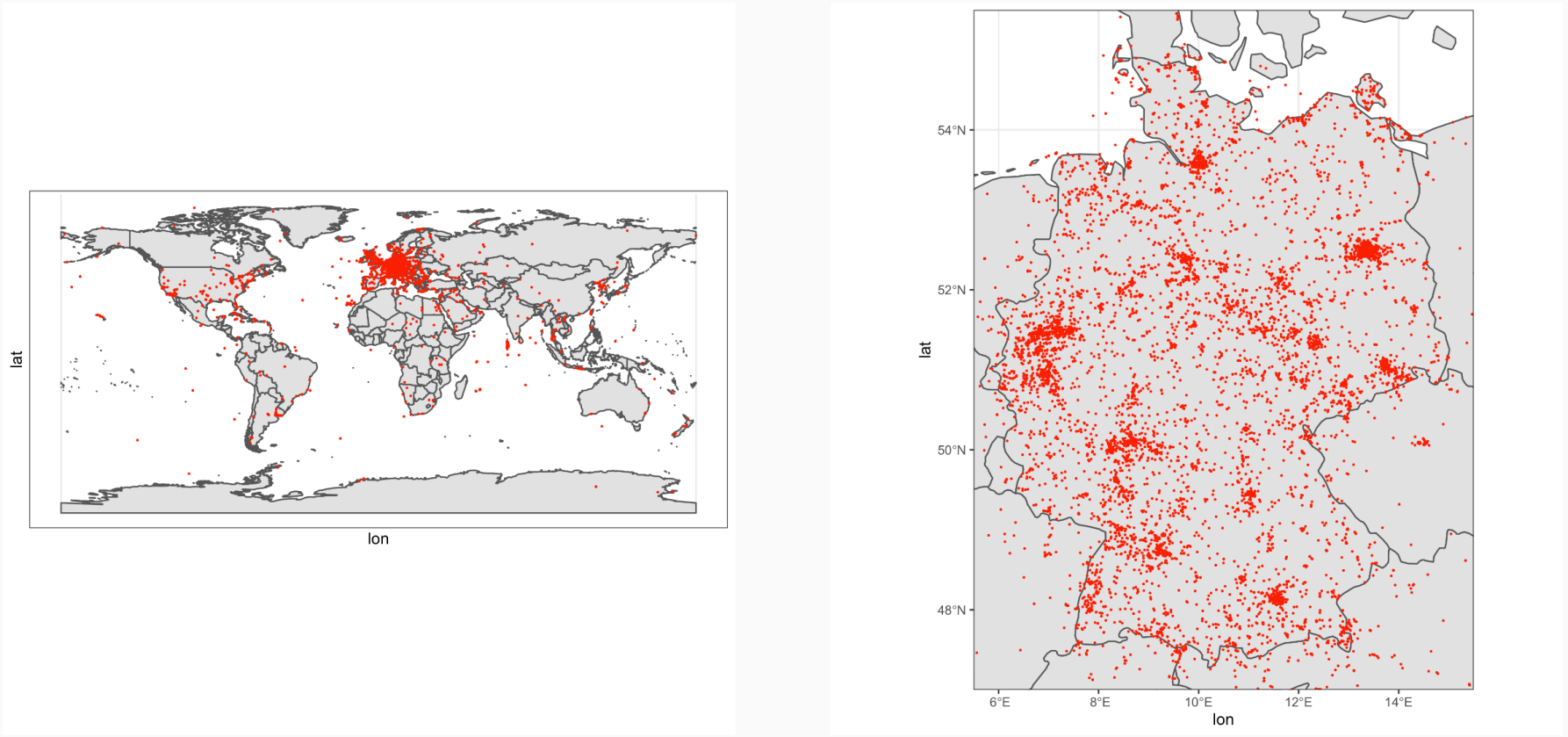
```
R> nrow(google_maps_df)/nrow(dat) # about 0.6%
```

```
## [1] 0.006104038
```

```
R> # extract coordinates
R> google_maps_df$coord_raw ← str_extract(google_maps_df$url, "@[-[:digit:]]{1,4}\\.[:digit:]+,[-[:digit:]]{1,4
R> google_maps_df$lat ←  str_extract(google_maps_df$coord_raw, "[-[:digit:]]{1,4}\\.[:digit:]+") %>% as.numeric
R> google_maps_df$lon ←  str_extract(google_maps_df$coord_raw, "[-[:digit:]]{1,4}\\.[:digit:]+$") %>% as.numeri
```

`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

How do I protect the privacy of my research subjects?



`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

How do I protect the privacy of my research subjects?


```
R> # parse ebay sites
R> ebay_df ← filter(dat, str_detect(url, "ebay"))
R>
R> # plz
R> ebay_plz ← ebay_df %>% filter(domain == "ebay-kleinanzeigen.de", str_detect(path, "^s\\-[:digit:]{5}")) %>%
R> length(ebay_plz)
```

```
## [1] 319
```

```
R> ebay_plz[1:20]
```

```
##  [1] "86720" "92660" "26655" "97789" "52457" "21149" "24983" "78234" "66346"
## [10] "08523" "14715" "01936" "21149" "12529" "14715" "72415" "78244" "21149"
## [19] "24306" "86720"
```

`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

How do I protect the privacy of my research subjects?


```
R> # parse ebay sites
R> ebay_df ← filter(dat, str_detect(url, "ebay"))
R>
R> # plz
R> ebay_plz ← ebay_df %>% filter(domain == "ebay-kleinanzeigen.de", str_detect(path, "^s\\-[:digit:]{5}")) %>%
R> length(ebay_plz)
```

```
## [1] 319
```

```
R> ebay_plz[1:20]
```

```
##  [1] "86720" "92660" "26655" "97789" "52457" "21149" "24983" "78234" "66346"
## [10] "08523" "14715" "01936" "21149" "12529" "14715" "72415" "78244" "21149"
## [19] "24306" "86720"
```

`Source` Munzert, Simon, Barberá, Pablo, Guess, Andrew M., & Yang, JungHwan (2022). Media Exposure and Opinion Formation in an Age of Information Overload (MEOF). GESIS, Cologne. https://doi.org/10.4232/1.13981.

Let's take one minute to discuss in which ways **publishing these data can be challenging.**

Licensing your data

If you publish your data in a data repository, you may need to apply a licence agreement to your data

Licensing your data

If you publish your data in a data repository, you may need to apply a licence agreement to your data

- A licence agreement is a legal arrangement between the depositor of the data set and the repository
  - It stipulates what users are allowed to do with the data

Licensing your data

If you publish your data in a data repository, you may need to apply a licence agreement to your data

- A licence agreement is a legal arrangement between the depositor of the data set and the repository
  - It stipulates what users are allowed to do with the data

To make re-use as likely as possible you may want to choose a licence which:

- Makes data available to the **widest audience** possible
- Makes the **widest range of uses** possible

Overview of Creative Commons licences

Licence	Can I copy & redistribute the work?	Is it required to attribute the author?	Can I use the work commercially?	Am I allowed to adapt the work?	Can I change the licence when redistributing?
CC0	Y	N	Y	Y	Y
CC BY	Y	Y	Y	Y	Y
CC BY-SA	Y	Y	Y	Y	N
CC BY-ND	Y	Y	Y	N	Y
CC BY-NC	Y	Y	N	Y	Y
CC BY-NC-SA	Y	Y	N	Y	N
CC BY-NC-ND	Y	Y	N	N	Y

This table is inspired by the CESSDA Guide

Access categories

Publishing data in a data repository does not automatically make them openly accessible.

Access categories

Publishing data in a data repository does not automatically make them openly accessible. **(Sensitive) personal data can still be protected by limiting access to the data.**

Access categories

Publishing data in a data repository does not automatically make them openly accessible. (Sensitive) personal data can still be protected by limiting access to the data. **Access controls can permit control down to an individual file level, meaning that mixed levels of access control can be applied to a data collection.**

Access categories

Publishing data in a data repository does not automatically make them openly accessible. (Sensitive) personal data can still be protected by limiting access to the data. Access controls can permit control down to an individual file level, meaning that mixed levels of access control can be applied to a data collection.

Many data repositories operate a three-tiered approach to data access:

- **Open access**
  - Data that can be accessed by any user whether they are registered or not. Data in this category should not contain personal information unless consent is given (see 'Informed consent').
- **Access for registered users** (safeguarded)
  - Data that is accessible only to users who have registered with the archive. This data contains no direct identifiers but there may be a risk of disclosure through the linking of indirect identifiers.
- **Restricted access**
  - Access is limited and can only be granted upon request. This access category is for the most sensitive data that may contain disclosive information.Restricted access requires the long-term commitment of the researcher or person responsible for the data to handle the upcoming permission requests.
- **Embargo**
  - Besides offering the opportunity for restricted access 'for eternity' most data repositories allow you to place a temporary embargo on your data. During the embargo period, only the description of the dataset is published. The data themselves will become available in open access after a certain period of time.

Other access

Special access

- Access through a *safe connection*
- Access in a *safe environment*

Open metadata for (sensitive) personal data

This was one of the solutions for our survey and web-tracking data archiving and publishing effort.

Other access

Special access

- Access through a *safe connection*
- Access in a *safe environment*

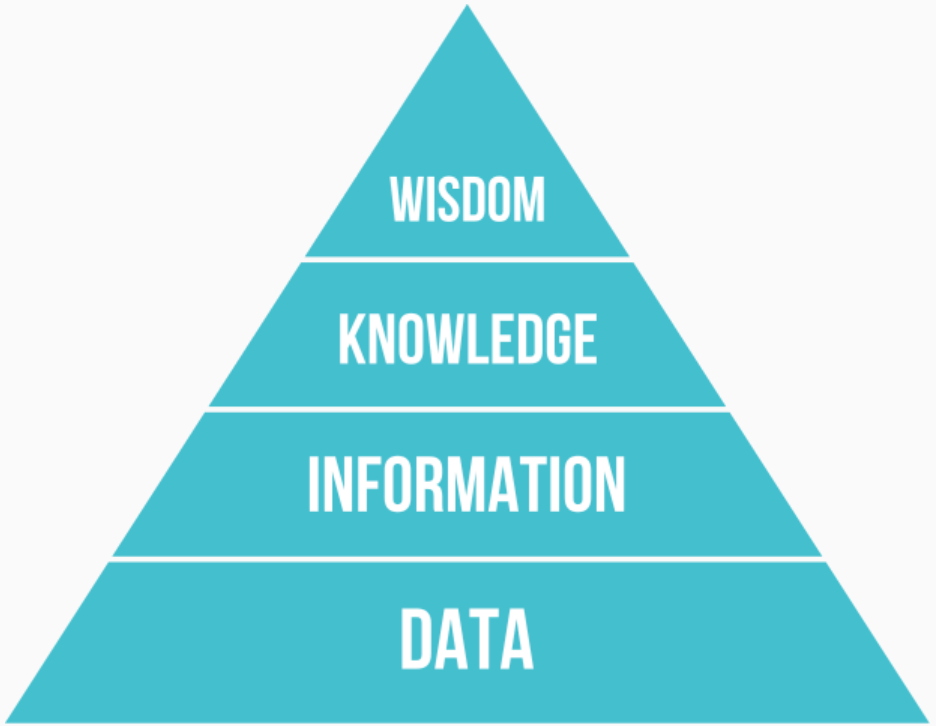
Open metadata for (sensitive) personal data

This was one of the solutions for our survey and web-tracking data archiving and publishing effort.

Open Government Data

What are data?

- Data are disembodied **facts**, **signs**, and **symbols**.
- We often define them by their **source** (*e.g. administrative, historical, medical, etc.*) and their **formats** (*e.g., numerical, textual, still image, geospatial, audio, video, and software.*)
- It is thought about as the basis of the **knowledge hierarchy** under the *DIKW pyramid*.



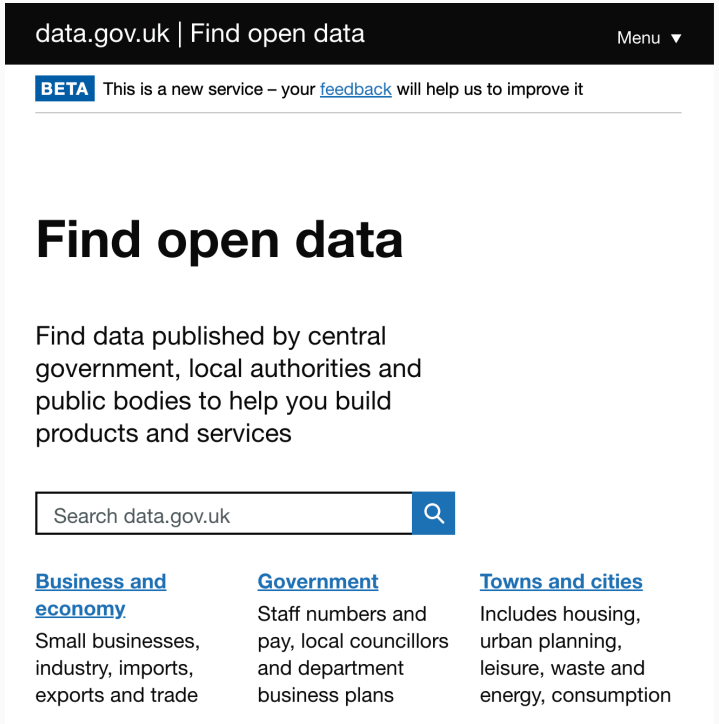
What is open data?

Open data

- Data that can be freely *used*, *modified*, and *shared* by anyone

What characteristics does it have?

- **Non-proprietary**: Available in formats that are open and standard without any restrictions
- **Machine-Readable**: Provided in formats that can be easily processed by computers
- **No Restrictions**: Free from any legal, financial, or technical barriers.
- **Comprehensive**: Covers all necessary aspects to provide a complete understanding



Why Open Government Data (OGD)?

- Public organizations **produce** and **collect** a broad range of different types of data to perform their tasks.
    - Demographic information
    - Socioeconomic data
    - Epidemiological data
    - Healthcare use
    - Industry data
    - Crime data
    - Performance
    - ...
- **These data tend to be extensive and central**

Government data can be a significant resource for increased transparency

The actors behind Open Government Data

There are two key civil society actors advocating for increased opennes of information, documents and datasets held by public bodies

- **Right to information movement**: promotes a public right of access to information from a **human rights perspective**

- **Open government data movement**: uses predominantly **social and economic arguments** to encourage the opening up of government data, such as:
  - Information hitting the public domain can create conditions for inclusive service delivery and more participation
  - Can stimulate the economy by allowing third parties to innovate using public data

Many argue that these data should be public domain, since **governments hold predominantly data emanating from citizens.**

The Open Government Data Principles

These principles were presented in 2007 during an **Open Government Working Group Meeting**

Government data shall be considered open if it is made public in a way that **complies with the principles** below:

- **Complete**: **All public data are made available**. Public data are data that is not subject to valid privacy, security or privilege limitations.
- **Primary**: Data are **as collected at the source**, with the highest possible level of granularity, not in aggregate or modified forms.
- **Timely**: Data are made available as **quickly** as necessary to **preserve the value** of the data.
- **Accessible**: Data are **available to the widest range of users** for the widest range of purposes.
- **Machine processable**: Data are **reasonably structured** to allow automated processing.
- **Non-discriminatory**: Data are **available to anyone**, with no requirement of registration.
- **Non-proprietary**: Data are available in a **format over which no entity has exclusive control**.
- **License-free**: Data are **not subject to any copyright, patent, trademark or trade secret regulation**. Reasonable privacy, security and privilege restrictions may be allowed.

Principles for Improving Federal Transparency in the US

Hertie School

10 principles proposed by Former US Federal Chief Information Officer **Vivek Kundra** before the US House of Representatives in 2011

- **Build end-to-end digital processes**: **Automate transfer of data** between systems to increase productivity, protect data integrity, and speed data dissemination. Capitalize on game-changing technologies to increase transparency.
- **Build once, use often**: Architect **systems for re-use and share platforms to reduce costs**, streamline systems and processes, reduce errors, and foster collaboration.
- **Tap into golden sources of data**: **Pull data directly from authoritative sources** to improve data quality, shorten processes and protect data integrity.
- **Release machine-readable data and encourage third party applications**: Make data **machine-readable** to allow the public to easily analyse, visualise and use government information.
- **Use common data standards**: Develop and use **uniform, unique identifiers and data standards** to ease the flow of data and reduce system complexity.

Principles for Improving Federal Transparency in the US

10 principles proposed by Former US Federal Chief Information Officer **Vivek Kundra** before the US House of Representatives

- **Validate data up front**: **Correct errors during collection** and at the point of entry to block bad data from ever entering the system.
- **Release data in real time and preserve for future use**: **Release data as quickly as feasible** to enhance its relevance and utility while maintaining future accessibility.
- **Reduce burden**: **Collect data once and use it repeatedly**. Pull from existing data sets to reduce costs and burden and to increase productivity and uniformity.
- **Protect privacy and security**: **Safeguard the release of information to increase public trust**, participation, preserve privacy, and protect national security. Open Government doesn't mean vulnerable government.
- **Provide equal access and incorporate user feedback**: Provide a common view of data to all **stakeholders to foster collaboration**. Incorporate user feedback to help identify high-value, meaningful data sets, set priorities, to continuously drive and improve future planning and processes.

Business Models Archetypes for value creation

Some actors make a business argument for Open Government Data. They identify some archetypes of business models that can create revenue from this practice:

- **Suppliers** that publish data, including the public sector;
- **Aggregators** that pool publicly available data and combine it to produce useful insights to be used by the various users;
- **Apps developers** that enable users to make more informed decisions (e.g. apps building on crime data, transport data);
- **Enrichers** that are large and established businesses producing significant quantities of open data and combine it with their own proprietary sources to provide services (e.g. insurers, retailers); and
- **Enablers** that are organisations that don't make money out of open data but provide platforms and technologies that others can use (e.g. websites that enable data sources of all types to make subsets of their data available to seek solutions from the public.

Potential beneficiaries from OGD practices

Government

Benefits

- **Efficiency**: Improves government operations and resource allocation.
- **Fraud Reduction**: Helps reduce fraud and errors.
- **Innovative Services**: Enables smarter, personalized public services.
- **Transparency**: Strengthens accountability and legitimacy.
- **Quality Interactions**: Enhances interactions between government and users.

Examples

- **Tax Gap Reduction**: More accurate tax collection.
- **Public Service Delivery**: Innovative apps and services

Potential beneficiaries from OGD practices (cont.)

Citizens

Benefits

- **Participation**: Encourages public participation in designing responses to public needs.
- **Informed Choices**: Provides information for making informed personal decisions.
- **Quality of Life**: Potentially improves overall quality of life through better services.

Examples

- **Fix My Street (UK)**: Citizens report local issues.
- **Chicago 311**: Government portal for public service requests.
- **Personal Data Access**: Crime rates, emissions, education stats.

Potential beneficiaries from OGD practices (cont.)

Civil society

Benefits

- **Informed advocacy**: Facilitates the identification of public needs.
- **Watchdogs**: Open government data allows civil society organizations (CSOs) to track government activities, expenditures, and decisions.
- **Facilitating partnerships**: Can encourage collaboration between CSOs, government agencies, and other stakeholders

Examples

- **Transparency International**: Corruption tracking.
- **Code for America**: Government data to build digital tools that improve public services.
- **Global Witness**: Investigate and expose illegal activities related to natural resources (land use, mining permits, deforestation)

The uses of OGD

1. Data to fact

Individuals seek specific facts in newly open datasets.

Applications:

- Civic or bureaucratic engagement
- Business planning
- Personal choices

Methods:

- Online interfaces
- Downloading spreadsheets

The uses of OGD (cont.)

2. Data to information

Creation of static representations and interpretations of data.

Manifestations:

- Visualizations
- Blog posts
- Infographics
- Written reports

The uses of OGD (cont.)

3. Data to Interface

Interactive access and exploration of datasets.

Examples:

- Searchable mapping mash-ups
- Tools for browsing large datasets and crowdsourcing feedback

Components:

- Static data interpretations
- Summary statistics
- Algorithmically derived assessments

The uses of OGD (cont.)

4. Data to Data

Sharing derived data in new formats or augmented forms.

Methods:

- Reformatting original datasets
- Combining with other data
- Creating APIs
- Easy downloads of dataset subsets

Advantages:

- Real-time updates
- Avoids re-downloading entire files

The uses of OGD (cont.)

5. Data to Service

Services where OGD supports operations behind the scenes.

Examples:

- Routing messages reporting issues like potholes to the appropriate authority
- Use of boundary data for service efficiency

The five star open data scheme

The five-star data openness scheme was developed and presented in 2010 by **Tim Berners-Lee a co-creator of the world wide web.**

⭐ data resources that have been made available on the network under the terms of an open license (in any format);

⭐⭐ resources made available in the form of structured data (for instance, an spreadsheet instead of a scan)

⭐⭐⭐ data in the document saved in an open format such as CSV;

⭐⭐⭐⭐ URI-tagged resources that are totally searchable;

⭐⭐⭐⭐⭐ data combined with different data that provides context to it.

Challenges to OGD

- **Policy challenges**
  - Disclosure policies, limits in data transparency, copyright issues
  - Lack of procedures and standards
- **Technical Challenges**
  - Variability in data sets, formats, and standards
  - Complexity of data access and source identification
  - Need for IT infrastructure and privacy enhancement
- **Economic and Financial Challenges**
  - Financial burden on governments
  - Training, technology, and infrastructure costs
  - Proprietary software and sensitive data issues

Challenges to OGD

- **Organizational Challenges**
  - Institutional structures and leadership
  - Independent oversight and data transparency
  - Supporting data release workflows and cultural orientation
- **Cultural Challenges**
  - Public interest and awareness
  - Stakeholder buy-in
- **Legal Challenges**
  - Variability in legal frameworks across countries
  - National security, privacy, and commercialization issues
  - Ambiguities in intellectual property rights

Responses to PSI Directive

OGD infrastructure providers

ckan

Home Features Showcase Solutions Support Blog Events FAQ Docs View on Github Login

CKAN for Government

CKAN is used by national and regional government organisations throughout the European Union, the Americas, Asia and Oceania to power a variety of official and community data portals.

Schedule a chat to speak to a CKAN steward who specialises in open data portals for government.

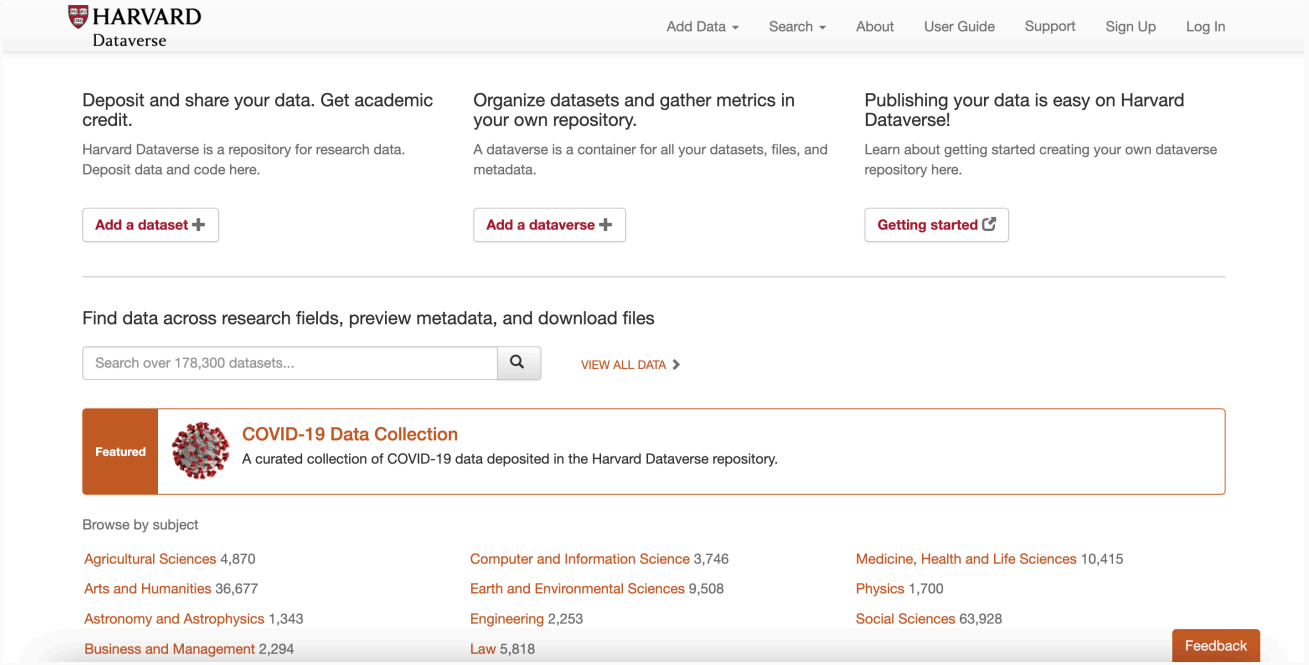
Schedule a chat

Discovering

Why data discover?

Hertie School

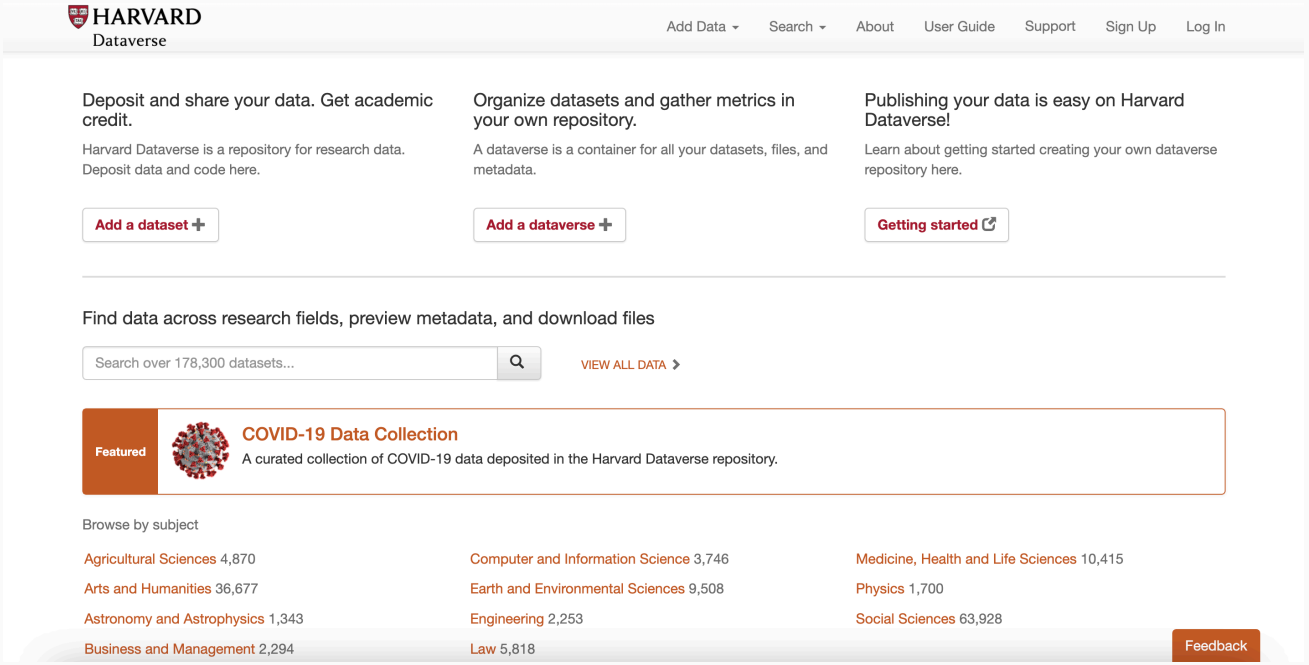
- **Save costs and time**



Why data discover?

Hertie School

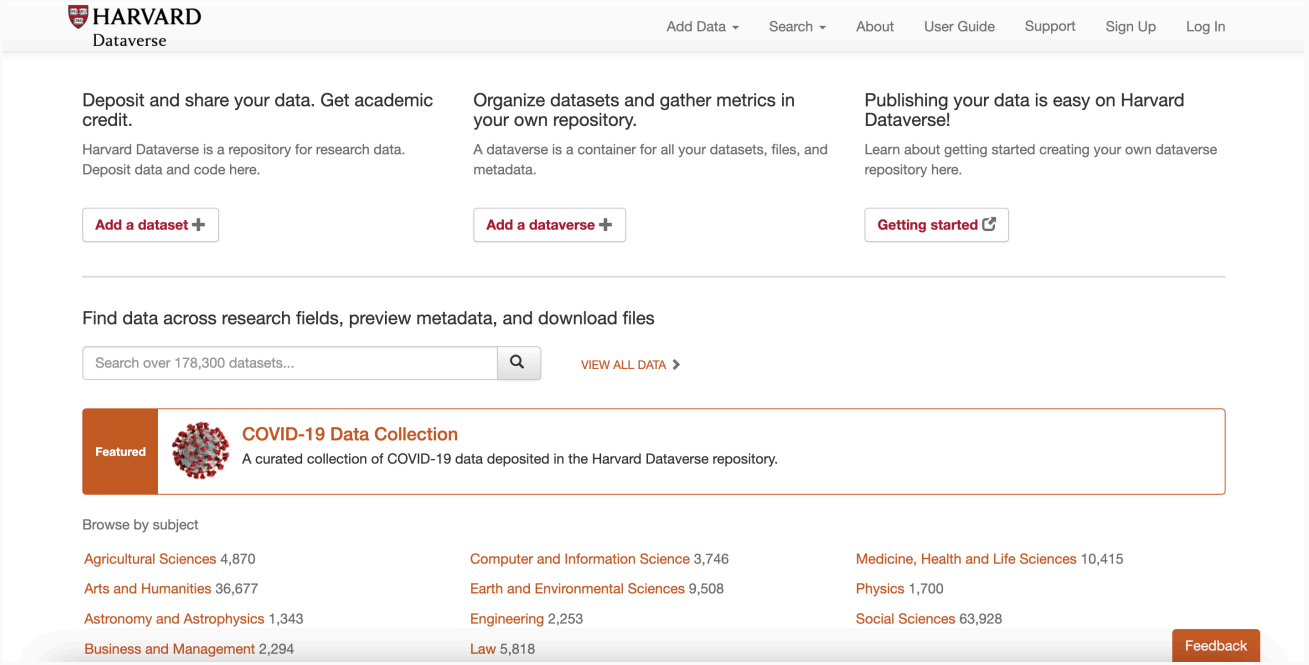
- Save costs and time
- **Compare results and replicate studies**



Why data discover?

Hertie School

- Save costs and time
- Compare results and replicate studies
- **Integrate into a larger environment**



How to find data?

What do you need?

- Identify the domain
- Measures and constructs
- What data types

Locate potential data hosts

- Ask (you've identified some people working on it)
- Data repositories
- Search engines and data aggregators (e.g., Google Dataset search)
- Data catalogues (GESIS) and data journals (Scientific Data)

Set up a search query

Ask for help

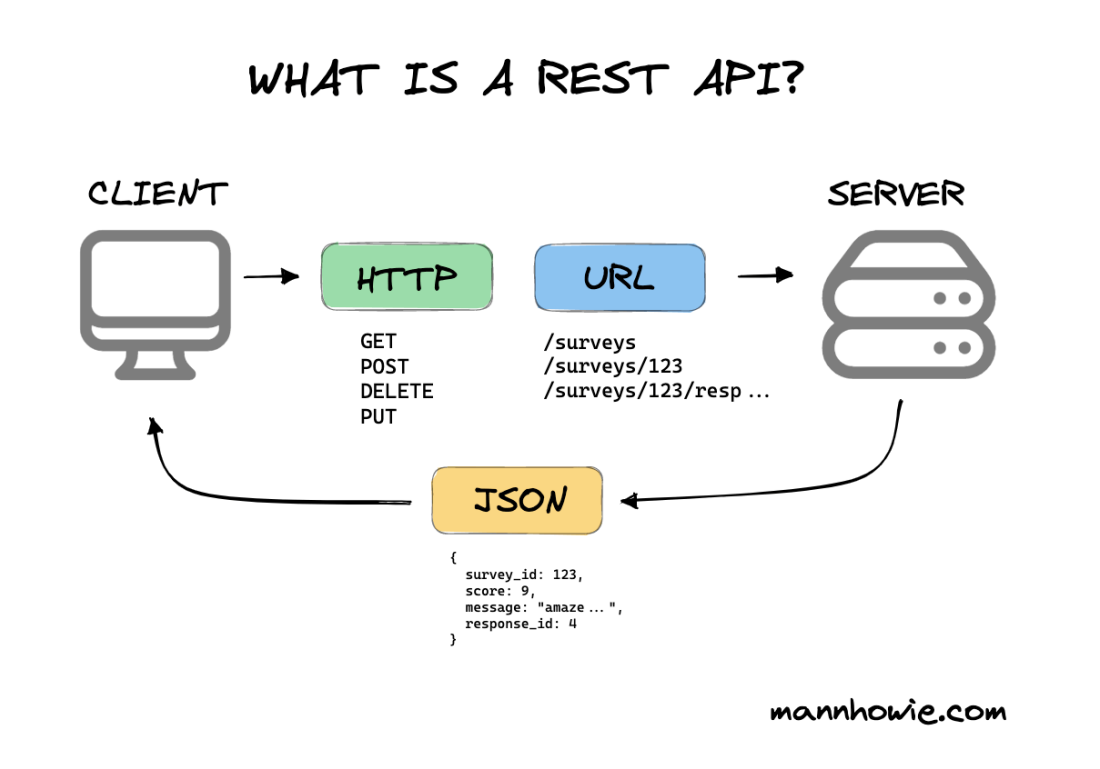
Once you have found data

Evaluate data quality

- **What** information was collected?
- **Who** collected the data? **When**? And, **where**?
- **Why** was the data created? E.g., different purposes for data collection are research, social policy, marketing etc.
- **How** was the data collected? You need detailed information about the methodology.
- How was the data **processed**?
  - Were there any changes in data? Who adjusted data in what way after it was collected?
  - To which manipulations was the data exposed?
- Were **consistency** and **logic checks** employed?
  - Is the data "clean", i.e. were nonlogical and erroneous values deleted?
- What quality assurance procedures were used? Did researchers use verified measurement tools?
  - Documentation

APIs (Application Programming Interface)

A set of **rules and protocols** for building and interacting with software applications. Act as intermediaries that allow different software systems to communicate with each other.



Meta

FORT

Approach Researcher Products Key Research Areas

Researcher API

RESEARCHER API

One of the first Meta APIs designed for academics

TikTok for developers

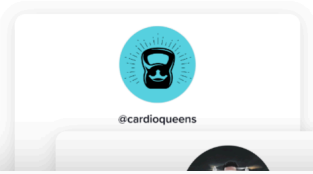
Products Docs Support Blog Log in

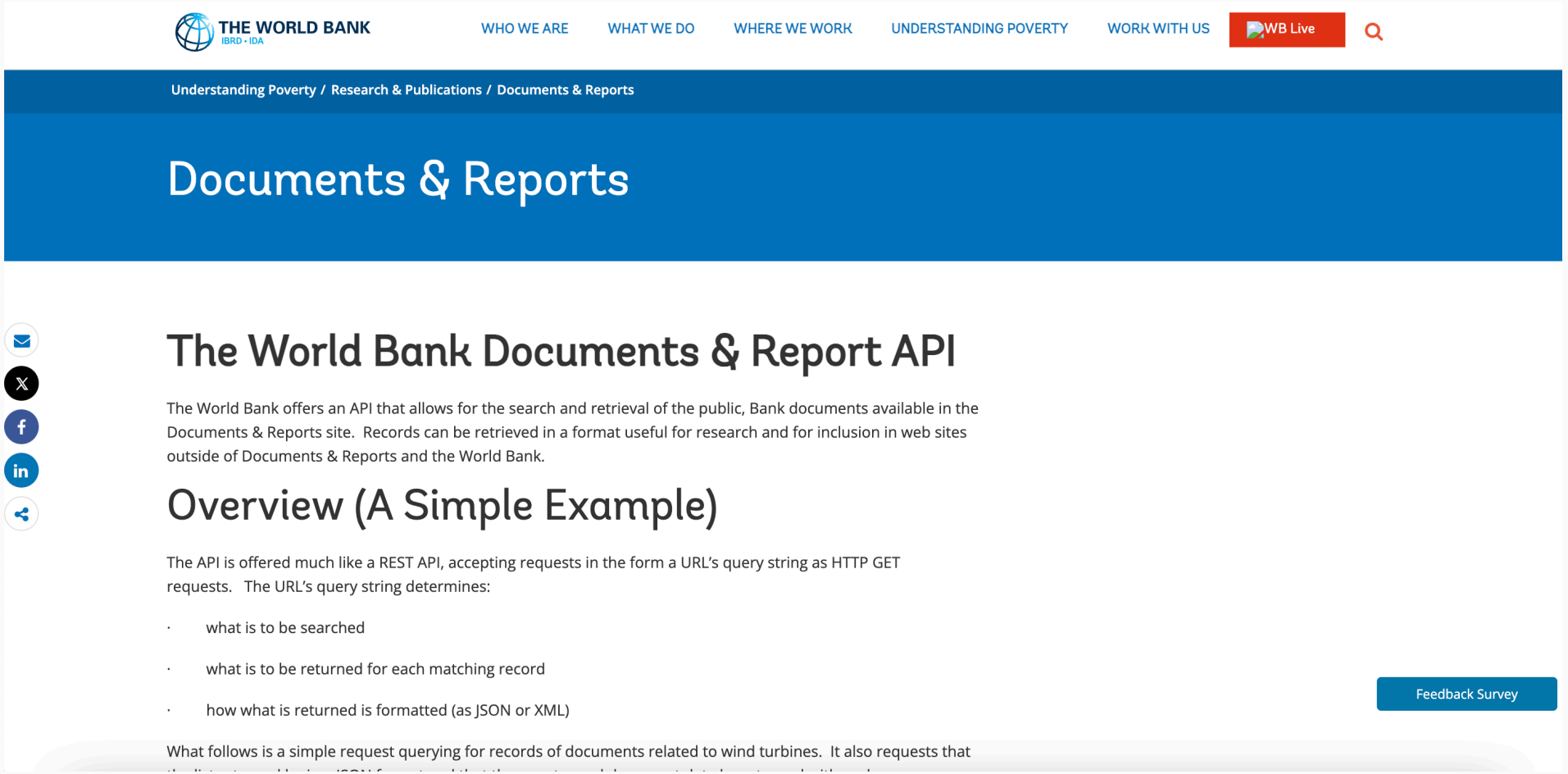
Research Tools

TikTok supports independent research about our platform. Through our Research Tools, qualifying researchers in the U.S. and Europe can apply to study public data about TikTok content and accounts. We're working to provide increased access to Research Tools in the future.

What data is available?

With the research tools, researchers can access public data on:




THE WORLD BANK
IBRD • IDA
WHO WE ARE
WHAT WE DO
WHERE WE WORK
UNDERSTANDING POVERTY
WORK WITH US
WB Live
Understanding Poverty / Research & Publications / Documents & Reports
Documents & Reports
The World Bank Documents & Report API
The World Bank offers an API that allows for the search and retrieval of the public, Bank documents available in the Documents & Reports site. Records can be retrieved in a format useful for research and for inclusion in web sites outside of Documents & Reports and the World Bank.
Overview (A Simple Example)
The API is offered much like a REST API, accepting requests in the form a URL's query string as HTTP GET requests. The URL's query string determines:
· what is to be searched
· what is to be returned for each matching record
· how what is returned is formatted (as JSON or XML)
What follows is a simple request querying for records of documents related to wind turbines. It also requests that
Feedback Survey

Questions?