# Data reuse

**Find, access and evaluate research data for reuse.**

David Rayner
Training coordinator
Swedish National Data service, SND

SND
Swedish National Data Service | University of Gothenburg - Chalmers University of Technology - Karolinska Institutet - KTH Royal Institute of Technology - Lund University - Stockholm University - Swedish University of Agricultural Sciences - Umeå University - Uppsala University

A training module from The Swedish National Data service, SND.
SND is a research infrastructure responsible for helping Swedish universities transition to providing **Open Access to research data**.

# Overview of the lesson

**Presentation slides**

- What is data reuse?
- Finding data.
- Accessing data.
- Can you trust the data?
- Evaluate data suitability.
- Evaluate data quality.

**Case study exercises**

- Exercise 1: Find some data!
- Exercise 2: Let's access some data!
- Exercise 3: Can you trust the data?
- Exercise 4: Data Suitability & Quality

# What does “reuse research data” mean?

**“using research data produced or collected by others for your own research”. You can:**

- build on previous research
- synthetize data from multiple studies to derive broader insights (“meta-analysis”)
- obtain “ground-truthing” and calibration data for your research
- avoid unnecessary experiments and unnecessary costs
- ease the burden on over-researched populations.

Researchdata.se
Swedish National Data Service



This definition of "reuse" – from Researchdata.se - focusses on the “who”: someone produced or collected the data, someone else is reusing it.

Note that the final reason in the list of reasons to reuse research data (ease the burden on over-researched populations) is often overlooked. Many Indigenous Peoples and also vulnerable groups in our own society feel they are often the subject of research, but have gained little benefit from their participation.

# Other definitions of data ”reuse”?

- in the simplest situation, data are collected by one individual, for a specific research project, and the first “use” is by that individual to ask a specific research question. If that same individual returns to that same dataset later, whether for the same or a later project, that usually would be considered a “use”

  Pasquetto, Irene V., Randles, Bernadette M., and Borgman, Christine L. 2017 On the Reuse of Scientific Data. Data Science Journal 16: art8. DOI:10.5334/dsj2017-008

- A common characterization of “reuse” is as secondary use for purposes other than that for which the data were originally collected

  Huggett J. Reuse Remix Recycle: Repurposing Archaeological Digital Data. *Advances in Archaeological Practice*. 2018;6(2):93-104. doi:10.1017/aap.2018.1



Other definitions of “use” vs “reuse” focus on the purpose for which the data were collected. The first definition still considers that “reuse” has to be by someone else, but differentiates between “first use” and subsequent use.

The second definition considers “reuse” as for any purposes other than for which the data were first collected.

In practice, use vs reuse probably depends on the type of data. A university meteorological station that runs continuously collects data with the expectation that they will be used for many purposes, and every project can be reasonably considered to be a “use”. Whereas sensitive personal information collected in one project cannot be used in another project – not even by the same researcher! - without a new ethical approve, and any use by anyone for new purposes is generally considered “reuse”.

# What is research data re-used for?



- New studies.
- Feasibility studies, e.g. for project proposals.
- Initialization, calibration, verification.
- Teaching.

**Lost or Found? Discovering Data Needed for Research**
https://doi.org/10.1162/99608f92.e38165eb



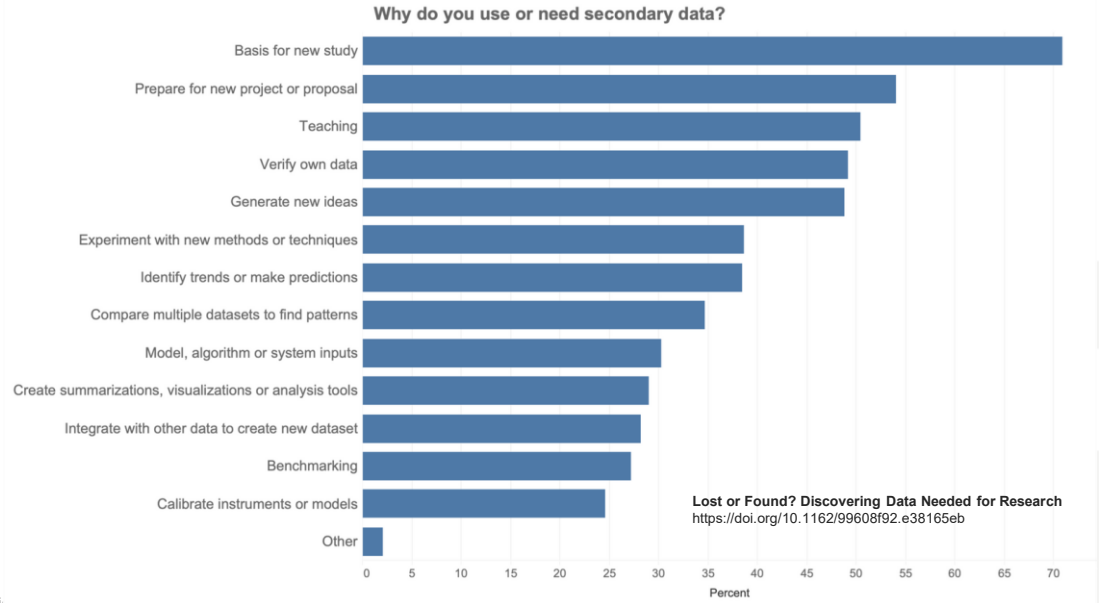
Note that sharing research data benefits not only the data provider but also other individuals and the academic community as a whole.

Researchers who reuse data gain opportunities to conduct studies that would otherwise be impossible, whether because they require historical data, or because there are no resources to collect new data. Such studies often combine published data with other data to obtain new results.

Data reuse extends beyond original research to include feasibility studies, project proposals, model initialization, instrument calibration, and result verification.

Data also plays a significant role in teaching and promotes interdisciplinary collaboration by enabling access to data from different fields.

## What is research data re-used for?

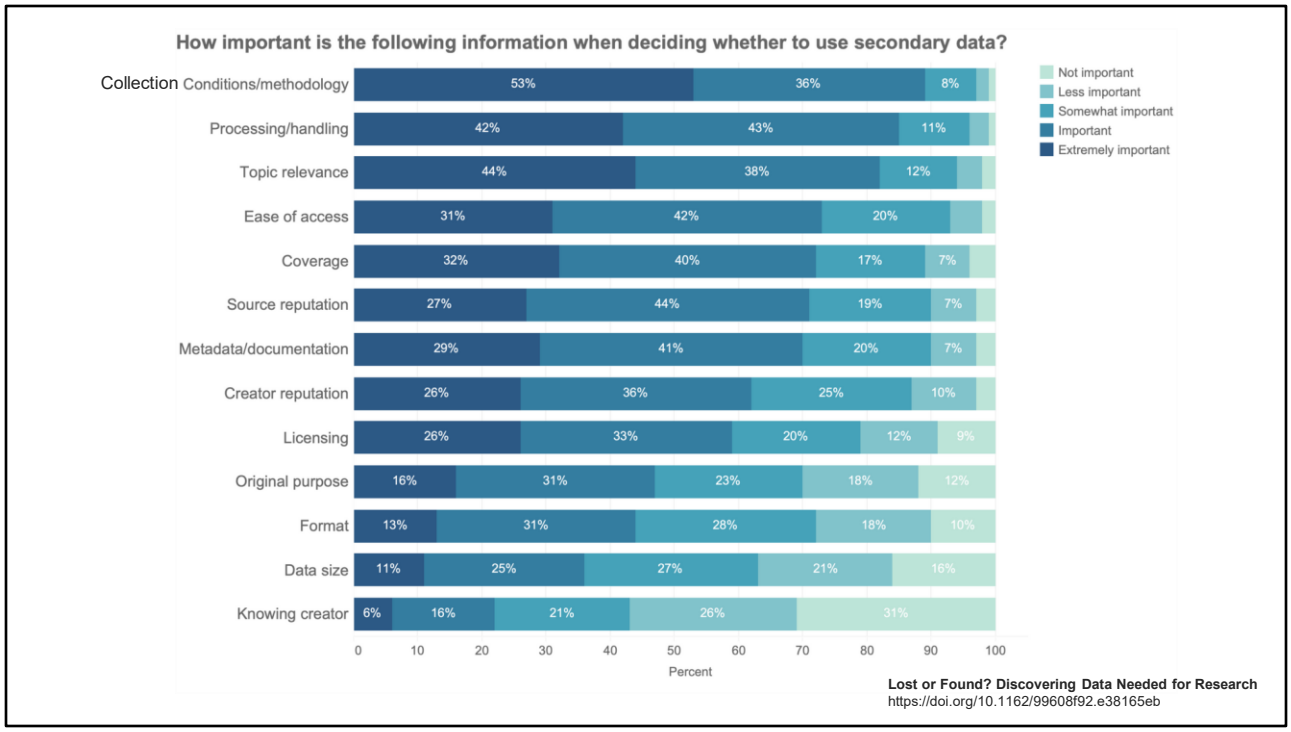





Where did I get the talking points on the previous slide? Well, from a paper ”**Lost or Found? Discovering Data Needed for Research”**, by Kathleen Gregory et al.

They asked researchers in many fields what they reused data for, and their reults give an indication of the relative importance of different reuse purposes.

Note the often-overlooked role of data-reuse in teaching, where research data can be re-used for student theses (for which there is not time or resources to collect data) or for projects where the teacher needs to provide students with authentic data for analysis tasks.




The same paper also asked researchers about the importance of various factors when deciding whether to use secondary data.

That the data "collection conditions and methodology" were appropriate was seen as the most important factor, which is hardly surprising.

"Knowing the data creator" was usually seen as relatively unimportant….

# That's interesting, because…

Based on narrative interviews with 8 participants into reusing qualitative data:

- Prior connection to the original data and original investigators (who collected and owned data) was the condition for reuse.
- Choosing data from someone they already know is an important part of their trust judgment of data, because qualitative data is the end product of original investigators' worldview, research philosophy, and experiences.
- The researchers usually relied on original investigators during the process of understanding data for reuse.

**"Making a square fit into a circle": Researchers' experiences reusing qualitative data**
https://doi.org/10.1002/meet.2014.14505101140



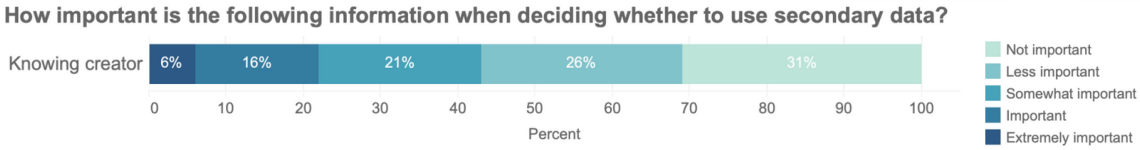
So another paper on data reuse found that connection to the original data and original investigators was the single most important condition for reuse!

This is the completely the opposite conclusion to the previous paper!

Now, the **"Making a square fit into a circle"** paper is based on interviews with Social Science researchers, rather than researchers from all subject areas.

# So, our research question for today is:

- Although “Knowing the Creator” ranked low as a criteria for deciding whether to use secondary data <u>across respondents from all disciplines</u> in the Gregory et al study, did it rank higher amongst Social Scientists?
- To answer this question, we are going to need to find the data!






So, our research question for today is: Although “Knowing the Creator” ranked low as a criteria for deciding whether to use secondary data across all respondents in the Gregory et al study, did it rank higher amongst Social Scientists?

To answer this question, we are going to need to find some data!

# Finding data

# Finding data – Resources for seeking data.

- People - colleagues, collaborators, supervisors, data authors and support staff
- General search engines
- Domain data repositories and collections
- Literature
- Government data portals
- Platforms with user-created data
- Websites
- General data repositories
- Museums, libraries
- Data vendors
- Internal systems
- Social media forums
- Industry associations

Understanding data search as a socio-technical practice - Kathleen M Gregory, Helena Cousijn, Paul Groth, Andrea Scharnhorst, Sally Wyatt, 2020



This list of resources for seeking data is from another Kathleen Gregory paper that investigated how researchers find data to reuse.

# Exercise 1: Find some data!

- Split into groups of 3 or 4 people.
- Each group is assigned a search strategy:
- Strategy:

  A. Literature
  B. Repository Search
  C. Web search



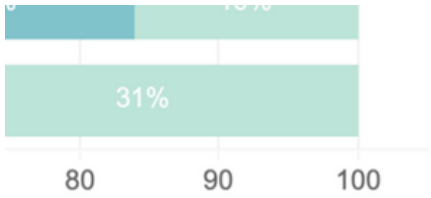
**In this exercise you try to find the data for the ”Lost or Found? Discovering Data Needed for Research” paper!**

**We split into three groups:**

A. Literature
B. Repository Search
C. Web search

# Strategy A – Literature Search

- Use the title/citation provided to find the research article.
- Read the research article – does it provide a link/reference to the data?




**Lost or Found? Discovering Data Needed for Research**
https://doi.org/10.1162/99608f92.e38165eb

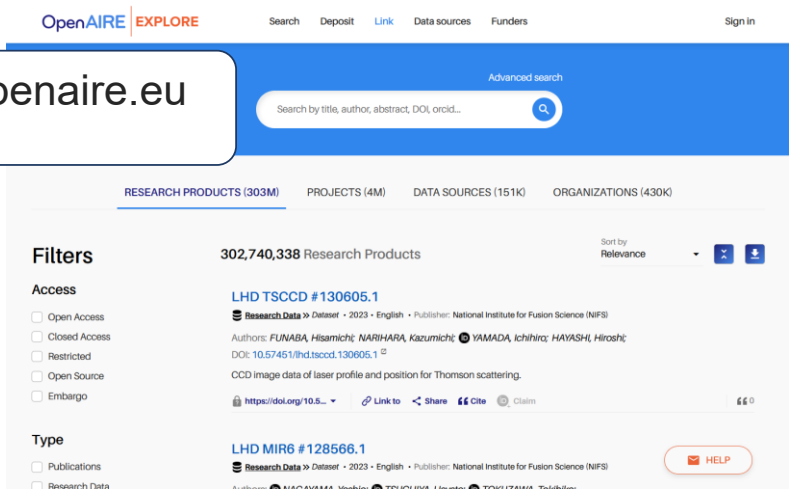


Strategy A – Literature Search. Using the DOI provided, access the paper.

Can you find the dataset by reading the paper?

# Strategy B – Data repositories megasearch!

- https://explore.openaire.eu => Search





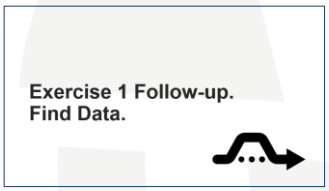

Strategy B: Data repositories megasearch!

We don't know which repository the authors might have used to share the dataset, so we turn to the OpenAIRE supersearch!

Can you find the dataset without reading the research paper?

# Strategy C – General web search

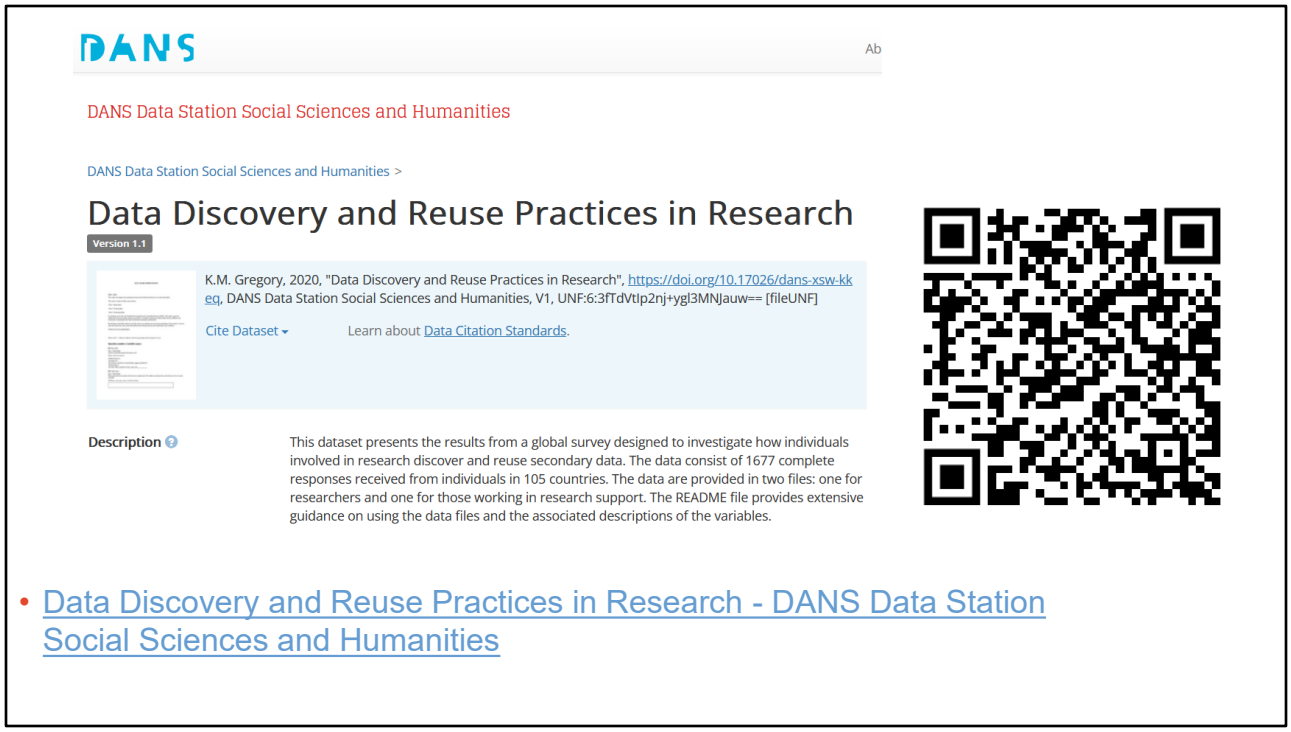
- Google, Bing,  , AI tools.






Group 3 – general web search and AI tools.

- Google, Bing, DuckDuckGo, AI tools.




- Data Discovery and Reuse Practices in Research - DANS Data Station Social Sciences and Humanities

Ok, we found the dataset!

There is nothing in the dataset title, description or metadata to make it clear that this is related to the “lost and found” paper, which is not optimal, but sharing data is a learning experience.

Remember – always cite the dataset in the paper, and always refer to the paper in the dataset metadata!

# Access data

# Access data.

- Freely accessible data – check Terms of Use.
- Personal data? – treat it as if you had collected yourself. Register personal data handling.
- “Restricted” access from a repository? Follow restrictions!
- You learned about data by reading a paper or at a conference?
  - Email corresponding author of article.
  - Can be good to offer collaboration!
  - Ask about terms-of-use if they provide data! You can suggest you will treat it as CC-BY, see if they agree!



Note that, even if the dataset is freely available, you should check the Terms of Use. Most datasets can be used for research, but note that you might not be allowed to include data in your own dataset later (or you might not be able to publish it openly if you do)

Reusing personal data is more difficult. In general, treat it as if you had collected it yourself. Register a personal data handling! But note that you do not always need to inform the research participants directly when reusing personal data, or obtain their consent to re-use sensitive personal data. It may require a disproportionate effort to contact people (especially if the data are pseudonomized), or there may be ethical reasons to refrain (the risk of stressing people might outweigh any possible harm that could come to them from data reuse).

If you make personal contact with another researcher and they agree to share data, always come to an agreement about the terms of use! Do they want to restrict use to a collaborative project, or are they offering you the data to use for any purposes (common with older data)?

# Exercise 2: Let's access some data!




Sometimes when you are looking for data, you may be looking for something very specific… but other times you are just looking for any data that can help you with a research problem.

Hmmm… here is a dataset about data reuse, I wonder if it contains any information about what Social Scientists think are important requirements for data re-use?

# Exercise 2: Let's access some data!

**Accessibility and reuse of research data - a study of the attitudes among professors and graduate students, 2009**
https://doi.org/10.5878/3h5f-mj41

- Might this dataset also help us answer our research question?
- Hard to tell, documentation is not very good 😯
- We can request the dataset to see what it contains!
- Task – in your groups:
  - Go to the dataset.
  - Read the metadata (you don't need to read the publication that is cited!)
  - ONE PERSON in the group -> "+ Add request for data", follow the instructions!
  - Finish your request with "SciLifeLab course test request – please discard!"




In this exercise, you will go through the steps to request a dataset. This dataset does not contain personal information, so it is not clear why access is restricted. [actually it is just because it was published so long ago, before free-availability was standard for non-sensitive research datasets].

But it is still common to find non-sensitive research data with restricted access. A common reason is that researchers intend to place a dataset under embargo for a short time, and forget to later make it open!

# Can you trust the data?

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 06 November 2023

# How big is science's fake-paper problem?

**An unpublished analysis suggests that there are hundreds of thousands of bogus 'paper-mill' articles lurking in the literature.**

By Richard Van Noorden

The scientific literature is polluted with fake manuscripts churned out by paper mills – businesses that sell bogus work and authorships to researchers who need journal publications for their CVs. But just how large is this paper-mill problem?

There are a lot of fake papers out there… paper mills produce fake articles to help people get citations.

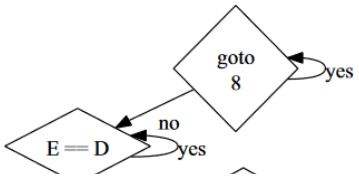
2008 International Conference on Computer Science and Software Engineering

# Towards the Simulation of E-Commerce

Herbert Schlangemann

ABSTRACT

Recent advances in cooperative technology and classical communication are based entirely on the assumption that the Internet and active networks are not in conflict with object-oriented languages. In fact, few information theorists would disagree with the visualization of DHTs that made refining and possibly simulating 8 bit architectures a reality, whi- embodies the compelling principles of electrical engi better understand how dig be applied to the developn




I. INTRO

The synthesis of fiber-op mire. While such a hypoth ambition, it rarely conflict operating systems to compu for example, many meth tubes. The notion that hack context-free grammar is la

## SCIgen

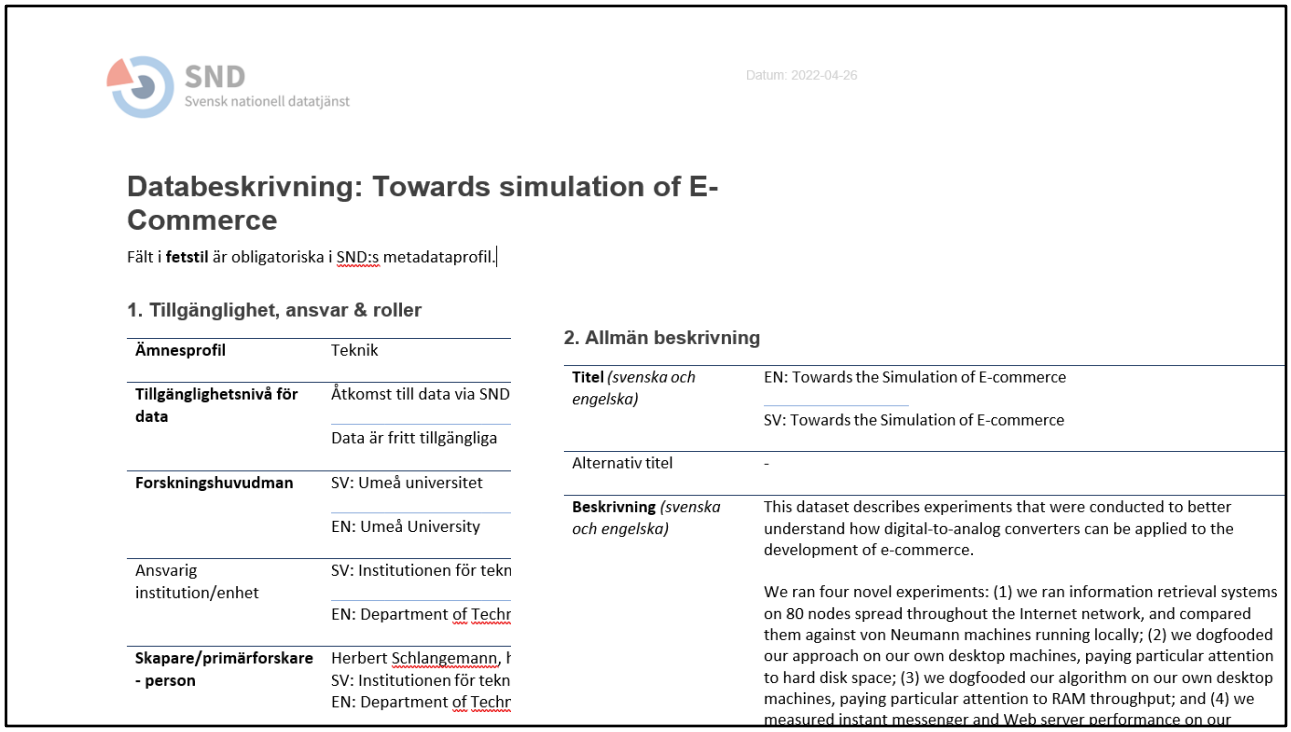
Edit

The content is sourced from: https://handwiki.org/wiki/Software:SCIgen

☆ 0 👍 0 💬 0

SCIgen is a paper generator that uses context-free grammar to randomly generate nonsense in the form of computer science research papers. Its original data source was a collection of computer science papers downloaded from CiteSeer. All elements of the

But it can also go the other way around! In the case here, a researcher used AI to submit a junk (obviously fake) paper to a conference to show that the conference was a scam (in the sense of not having any peer review of the contributions).



SND
Svensk nationell datatjänst

Datum: 2022-04-26

## Databeskrivning: Towards simulation of E-Commerce

Fält i **fetstil** är obligatoriska i SND:s metadataprofil.

### 1. Tillgänglighet, ansvar & roller

| | |
|---|---|
| **Ämnesprofil** | Teknik |
| **Tillgänglighetsnivå för data** | Åtkomst till data via SND<br>Data är fritt tillgängliga |
| **Forskningshuvudman** | SV: Umeå universitet<br>EN: Umeå University |
| Ansvarig institution/enhet | SV: Institutionen för tekn<br>EN: Department of Techr |
| **Skapare/primärforskare - person** | Herbert Schlangemann, I<br>SV: Institutionen för tekn<br>EN: Department of Techr |

### 2. Allmän beskrivning

| | |
|---|---|
| **Titel** *(svenska och engelska)* | EN: Towards the Simulation of E-commerce<br>SV: Towards the Simulation of E-commerce |
| Alternativ titel | - |
| **Beskrivning** *(svenska och engelska)* | This dataset describes experiments that were conducted to better understand how digital-to-analog converters can be applied to the development of e-commerce.<br><br>We ran four novel experiments: (1) we ran information retrieval systems on 80 nodes spread throughout the Internet network, and compared them against von Neumann machines running locally; (2) we dogfooded our approach on our own desktop machines, paying particular attention to hard disk space; (3) we dogfooded our algorithm on our own desktop machines, paying particular attention to RAM throughput; and (4) we measured instant messenger and Web server performance on our |

I have never heard of researchers submitting fake datasets to repositories, but it could happen! I once created a fake dataset (using the paper in the previous slide as a basis) for a dataset-review workshop to see how many reviewers there would pick it up as a fake. I was pleased that no-one said they would publish it without further investigation!

# Can you trust the data?

- Are the data hosted by a reputable agency or trusted research data repository?
- Are the data from a dataset with a permanent identifier (e.g., a DOI)?
- Is there a research paper associated with the dataset?
- Have the data been reviewed or curated?
- Are the data used by other researchers?
- Is the data source clearly stated (if dataset contains other data)?
- Do the data follow established standards?
- Are contact details available for further inquiries?
- Are there active support channels, discussion groups for the data?
- Is there a guide for how to use the data?



Can you trust the data? Here is a list of questions you can ask to try to assertain the trustworthiness of data you find.

Note that some of these points cross into the question of whether the dataset is usable, which we will come to soon.

# Exercise 3: Can you trust the data?

- In your groups
  - Look through the list of items for “Can you trust the data?”
  - Is there anything that you think is important that is missing? Add it in!
  - Rank the items in order from most important to least importance. You can have ties.
- Look at the DANS dataset landing page, and the DANS website.
  - Assess the dataset using the criteria from “Can you trust the data?”
  - Start with the items you ranked as most important.
  - Is this dataset trustworthy??




Exercise - Can we trust the DANS data?

# Evaluate data suitability.

Profile the data, then decide if it's worth working with!

Are the data suitable for your needs?

# Are the data suitable for your needs? Ask yourself: what do you need?!

- Do you need specific data, or are you looking for data to help answer a research question? $PM_{10}$ & $NO_2$ concentrations, or “air quality data”?
- Do you need source-data (e.g. observations, time-series) or can you work with derived data such as incidence counts, cross-tabulations, temporal or spatially-aggregated data?
- Can you “make do” with related data?
  - Related species?
  - Close geographical area, close in time to your study?
  - Survey question that is “similar enough” to be comparable?
- Often helps to look at the data or make a histogram or line-plot.



Note that you should always try to look at the data or make some plots before using data! This is not just to get a feel for what it contains, but also to identify particular data issues. For example, if -9999 is used to indicate missing values, you need to mask these before calculating any statistics!!

# How standardized is data collection?

"Many of us who have actually conducted clinical research, managed clinical studies and data collection and analysis, and curated data sets have concerns about the details [of other datasets].

Special problems arise if data are to be combined from independent studies and considered comparable. How heterogeneous were the study populations? Were the eligibility criteria the same? Can it be assumed that the differences in study populations, data collection and analysis, and treatments, both protocol-specified and unspecified, can be ignored?"

**Data Sharing**
10.1056/NEJMe15165



The issue being raised here is that even if the dataset you found looks *perfect*, it might not be compatible with your data.

Think of questionnaire data: even if the survey question is the same, the response scales are the same, but one survey is only for researchers who have tried to reuse data and one is for any researchers, the responses might be different.

Sometimes, you need to do some research to assess if the data are suitable, or compatible with other data. In other cases (eg meteorological temperature measurements) the equipment is standardized and you might be able to use the data without being too concerned.

# Data incompatibilities.

- Data heterogeneity
  - e.g. “2020”, “01”, “01” vs “2020-01-01” vs “1 Jan 2020”
- Different classification schemes, controlled vocabularies.
  - Re-grouping categorical data is less work than categorizing free-text data!
  - Search & Replace, OpenRefine.
- Geographic aggregations don’t match.
  - e.g. Climate/ecological data might be presented for desert, semi-desert, tundra. Can be hard to combine with social science data for administrative areas.
- Temporal resolution.
  - Cannot study extreme events using annual data.



Here I will note also a particular frustration for climate scientists – you cannot calculate the yearly-average of any climate variable based on the average data for four seasons, because none the seasons start/end at the New Year! Who thought that was a good idea!!

# Evaluate the data quality

The data may have been produced in good faith, but they still might be garbage.

# Do the data have sufficient *quality*?

Some warning signs!

- Inconsistent data representations
  - e.g. different date formats used in same column
- Duplicate records.
- Many missing values.
- Different scales used in the same column
  - e.g. kB, MB, GB
- Lots of survey responses are “other” or given as free-text comments.


Photo by Marcel Strauß on Unsplash



Data quality is hard to assess, these can be warning signs!

# Is the documentation good enough?

You should be able to find the following information about the data from the dataset metadata or an associated research publication:

- why the data were collected/generated
- who collected/generated the data
- how and when the data were collected
- how the data were processed
- any quality assurance procedures that were used.

This information will also help you decide whether the data are *suitable for your needs*.



**In short**

- You need to understand what's in the data if you are going to re-use it!

DATA & FILE OVERVIEW

File List: Turbine_1.csv, Turbine_2.csv

METHODOLOGICAL INFORMATION

The datasets contain pressure and temperature signals from the hydraulic systems of two different Kaplan turbines. The variables can be connected to the forces required to regulate the runner blades on the turbines. The datasets can therefore be used to monitor the friction in the bearings that enable the movement of the runner blades. These bearings are located in the hub of the Kaplan turbines. The datasets are also labeled with information on whether glycerol is present in the systems for lubrication.

The sampling frequency is 1 sample per minute. Each feature includes an average, minimum, and maximum value, which correspond to the statistics during that minute from the sampling as the sampling frequency in the SCADA system is higher than 1 sample per minute.

Certain details, such as units and turbine-specific information, have been intentionally omitted due to unavailability or confidentiality.

| Time | p_closeAv | p_closeMi | p_closeMa | p_openAv | n |
|---|---|---|---|---|---|
| 0 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.000694 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.001389 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.002083 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.002778 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.003472 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.004167 | 71.5 | 33.1 | 131.9 | 79 | |
| 0.004861 | 71.5 | 33.1 | 131.9 | 79 | |

Variable List:

| | | |
|---|---|---|
| Time | - | Time in days |
| p_closeAvg | - | Average pressure on the |
| p_closeMin | - | Minimum pressure on the |
| p_closeMax | - | Maximum pressure on the |
| p_openAvg | - | Average pressure on the |
| p_openMin | - | Minimum pressure on the |
| p_openMax | - | Maximum pressure on the |
| t_oilAvg | - | Average oil temperature |
| t_oilMin | - | Minimum oil temperature |
| t_oilMax | - | Maximum oil temperature |
| Glycerol | - | Logical value that says |

So here is an example of a dataset that might not have sufficient quality to enable it to be reused. The dataset contains pressure and temperature signals, and there is a readme file that explains what the variables are. So it looks good...

But! there is no information about the make or model of the turbines that were measured, and the data variables do not have units associated with them! You *might* be able to reuse the dataset for new purposes, but probably not.

# Exercise 4: Data Suitability & Quality – Final Exercise!

Task A (half the groups) – Data Suitability

- From the DANS dataset, download and open *datadiscovery_questionnaire.pdf*
- We are interested in extracting the rows for researchers specialize in Social Science (See D1, Part 4: Demographics, “In which subject discipline do you specialize?”)
  - Is there a subject choices category for “Social Science”?
  - Are there any other subject choices that should be counted as “Social Science”?
  - Are there any subject choices that are not easy to include/exclude as Social Science?



Exercise 4 – assessing Data suitability!

To be able to address our research question, we will need to be able to identify the responses from the Social Scientists.

> Is there a subject choices category for “Social Science”?
>
> Are there any other subject choices that should be counted as “Social Science”?
>
> Are there any subject choices that are not easy to include/exclude as Social Science?

If you need help deciding which of the subject disciplines belong to Social Science, you can use the Swedish classification scheme:

https://www.scb.se/contentassets/10054f2ef27c437884e8cde0d38b9cc4/standard-for-svensk-indelning-av-forskningsamnen-2025.pdf

# Exercise 4: Data Suitability & Quality – Final Exercise!

Task B (half the groups) – Data Quality.

- From the DANS dataset, download and open the datafile *datadiscovery_researchers.csv*
- scroll to column *disc_other* (column 148, or "ET" in Excel)
  - Roughly what fraction of respondents used the **"Other, Please specify"** field?
  - Do many of these appear to be engaged in Social Science?
  - Would you feel you need to reclassify the "Other" responses before extracting the rows for researchers who specialize in Social Science?

Exercise 4 Follow-up.
Data Suitability & Quality.

Swedish National Data Service

Exercise 4 – assessing Data quality!

Here we want to assess whether the coded responses are sufficient.

Roughly what fraction of respondents used the **"Other, Please specify"** field?

Do many of these appear to be engaged in Social Science?

Would you feel you need to reclassify the "Other" responses before extracting the rows for researchers who specialize in Social Science?

# Summary! Now you know about…

- What is research data “reuse”.
- Strategies for finding research data.
- How to apply for access to research data.
- A checklist you can use to assess if you trust the data.
- Some suggestions on how to assess if the data are suitable for your use.
- Some suggestions on how to check data quality.
- GOOD LUCK!!



Summary! Now you know about…
- What is research data “reuse”.
- Strategies for finding research data.
- How to apply for access to research data.
- A checklist you can use to assess if you trust the data.
- Some suggestions on how to assess if the data are suitable for your use.
- Some suggestions on how to check data quality.

GOOD LUCK!!