

# Open Data Editor

## User research findings

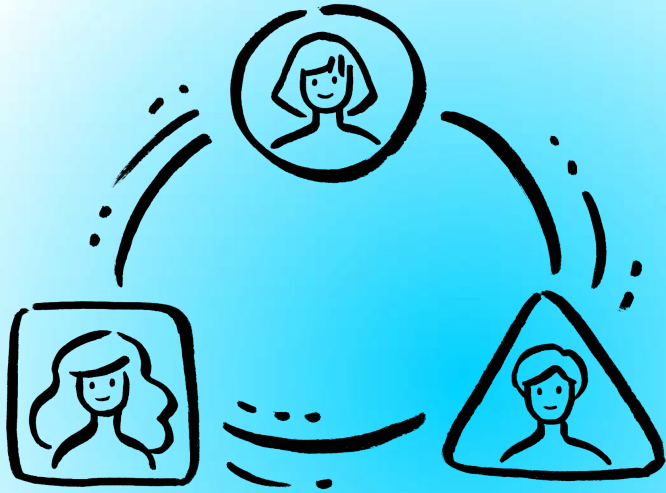


Open Knowledge

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

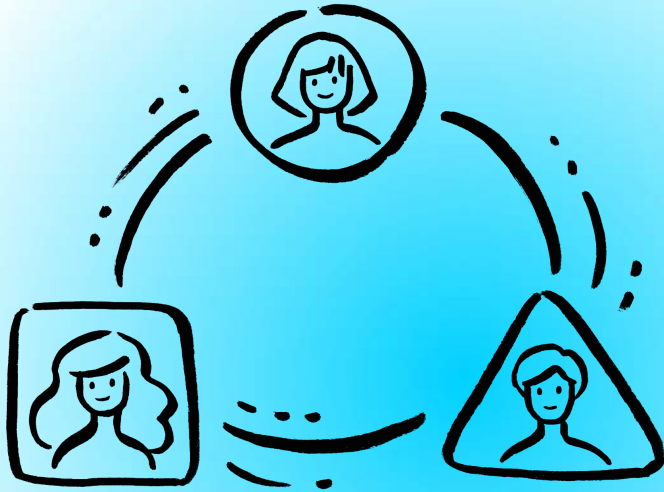
April 2024

# Main objectives



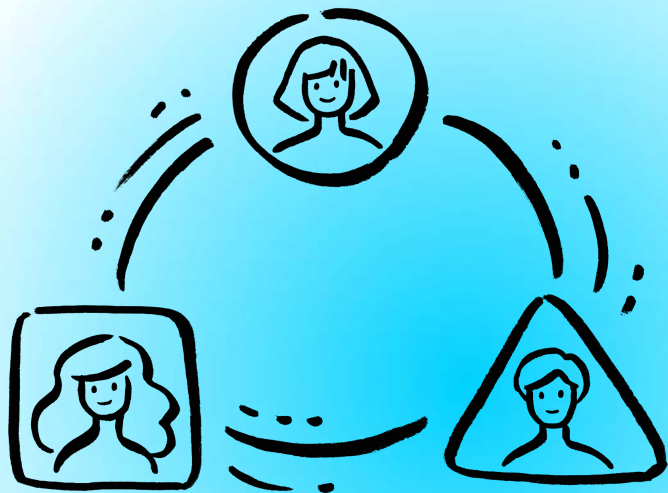
- To understand how people with competent data knowledge work with data.
- What tools they use.
- What challenges they face.

# Questions



- In which cases do you use data in your work?
- Which tools do you use?
- Why do you use those tools?
- Can you describe how you use data tools?/  
What do you do with them?
- What is your favourite tool for working with data? Why? Please, describe those features that make your work easier.

# Questions

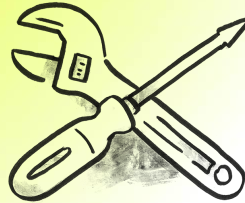


- What types of data formats do you work with in general?
- What challenges do you face when working with these formats?
- Tell me about the last time working with data was a challenge. What was the hardest part? Why was it hard? How did you solve the problem?
- Please describe, step by step, what you do when you get a data file to work with.

# Questions focused on



**Data exploration / Data  
quality assessment**



**Data cleaning**



**Data pains**

# Interviewees and profiles

**10 people in total (7 women and 3 men)**

- NGO Communications leader who works with data
- OSINT specialist / Journalist / University professor.
- University professor / Journalist/ Academic researcher.
- Data visualization designer and open data advocate.
- Learning and social impact specialist (data and AI user) / University professor.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Interviewees and profiles

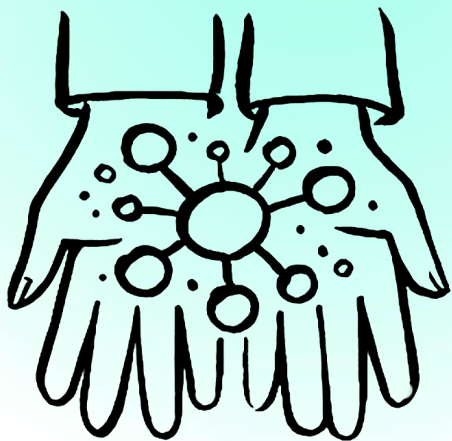
**10 people in total (7 women and 3 men)**

- Investigative journalist working on transnational investigations.
- Data Insight Analyst – Helps people develop projects to fight inequalities
- Data journalist with basic coding skills
- Journalist working with data – tables for infographics + storytelling / Private company.
- Research leader – NGO / Former public official (Ministry of Education)



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE



**We organized  
answers into 11  
categories to see  
patterns**





# Interviews' patterns and findings



Archivo Editor Ver Insertar Formatos Datos Herramientas Extensiones Ayuda



Compartir



100%

€

%

.0

.00

123

Predet...

-

10

+

B

I



A



C185

Research leader - NGO / Former public official (Ministry of Education)

	A	B	C	D	E
1	Int	Interview	Interviewee profile	Category	Description
2	1	2024-03-14	University professor / Journalis	Data use	For teaching, for journalistic investigations but also for academic purposes
3	1	2024-03-14	University professor / Journalis	Type of data / ...	Open data websites (governemnts), but also databases from NGOs.
4	1	2024-03-14	University professor / Journalis	Formats	csv and xls.
5	1	2024-03-14	University professor / Journalis	Tools	Excel and Google Sheets, for qualitative and quantitative tasks. To organize press art
6	1	2024-03-14	University professor / Journalis	Other relevant ...	Once used data with geographic coordinates. The dataset was delivered in Excel. SH
7	1	2024-03-14	University professor / Journalis	Data quality ev...	"I read the column headings (...), then the first thing I do is a pivot table (...)". Tool use
8	1	2024-03-14	University professor / Journalis	Most used func...	Pivot tables, some functions such as COUNTIF, count empty cells, etc. Also other fun
9	1	2024-03-14	University professor / Journalis	Pains	"When I try to explore a database, to see if I can find something (...) I get very frustrat
10	1	2024-03-14	University professor / Journalis	Pains	Different formats for dates, hours and differences in currencies (conversions).
11	1	2024-03-14	University professor / Journalis	Pains	Problems with formats: "I once received a [file with a] format that I didn't know what to
12	1	2024-03-14	University professor / Journalis	Pains	Manually copy data from a PDF "When there are few and when it is manually possible
13	1	2024-03-14	University professor / Journalis	Other relevant ...	Internet searches to find Excel functions.
14	1	2024-03-14	University professor / Journalis	Other relevant ...	"I'm still having a hard time incorporating ChatGPT for this [working with data ] (...) be
15	1	2024-03-14	University professor / Journalis	Data quality ev...	"The first thing I do is to check if data answers the question: "If I ask how much money
16	1	2024-03-14	University professor / Journalis	Metadata	"The times I have looked at the metadata they do not give me more information than v
17	2	2024-03-14	Learning and social impact spe	Data use	"To consolidate training data from the company I work for (...) and have a better under
18	2	2024-03-14	Learning and social impact spe	Tools	BigQuery, Google's cloud SQL engine: "It's very practical for me because you can inte
19	2	2024-03-14	Learning and social impact spe	Tasks	Tasks after downloading OurWorldInData dataset (artificial intelligence investments):



Data

Summary - Per category

Data validation

A1 🔍 List of tables

A B C D E F G H I J

1	<b>List of tables</b>	
2		
3	<a href="#">1. General summary</a>	<a href="#">7. Gains</a>
4	<a href="#">2. Data use</a>	<a href="#">8. Pains</a>
5	<a href="#">3. Formats</a>	<a href="#">9. Metadata</a>
6	<a href="#">4. Type of data (sources)</a>	<a href="#">10. Data quality</a>
7	<a href="#">5. Tasks</a>	<a href="#">11. Other relevant findings</a>
8	<a href="#">6. Most used functionalities</a>	<a href="#">12. Tools</a>
9		
10		

**General summary**

Category	COUNTA de Interviewee profile
Pains	51
Tasks	26
Other relevant findings	26
Gains	26
Tools	16
Data use	13
Most used functionalities	8
Formats	7
Type of data / Source	5
Data quality evaluation	5
Metadata	4

**Data use**

Category	Description
Data use	"I create infographics in Spanish (...) and v
Data use	"I use data to develop outreach content, tc
Data use	"To consolidate training data from the com
Data use	For data visualizations for news articles, b
Data use	For investigations. The interviewee uses s
Data use	For press clipping (measurement of press
Data use	For research purposes, to make decisions
Data use	For teaching, for journalistic investigations
Data use	The interviewee works with datasets to ex
Data use	To analyze data for internal insights and a
Data use	To reveal original stories
Data use	To translate research data into valuable in

**Formats**

Category	Description
Formats	csv and xls.
Formats	csv, xls and json
Formats	csv, xls and json
Formats	csv, xls, PPT, PC
Formats	csv, xls, SQL
Formats	xls, csv, json (so
Formats	xls, PDFs

**Type of data**

Category	Description
----------	-------------



# Data exploration actions

# Patterns

*"I upload [import/open] the csv to a spreadsheet, explore it, see what columns it has, what names those columns have, add filters to have a quick look at the unique rows of the dataset. After that, it depends on what I'm going to do, but I go and make pivot tables to get quick counts and do a first exploratory analysis (...) and then what I often find useful is...if I have a hypothesis, I take a piece of paper and make a conceptual map of what I need, taking into account the data I have...I start to make calculations or join concepts and then I go back to the table and start making pivot tables and calculations to start rejecting or validating that initial hypothesis".*



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Patterns

When exploring a dataset the interviewee uses filters to search for outliers, normalize data (names/countries written in a different way, check empty cells, data format for different columns (dates, for example)).

To explore datasets the interviewee creates a copy (first thing he does), creates pivot tables and histograms: *"Pivot tables allow me to immediately start cleaning the data"*.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE



**People working  
with data  
experience the  
same old problems**

# General findings

## *Pains*

- Pain 1: *"Date field in different format (...) there is no data validation...(...) things that are not standardized (...) / Categories that are not written in the same way, problems with formats in numbers, dates and units.*
- Pain 2: *"Making changes to a lot of data by hand because the original format of the database cannot be converted". // Date formatting is not always corrected when settings in the spreadsheet are updated.*



Open Knowledge

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# General findings

## *Pains*

- Pain 3: Data cleaning pains: normalize names, places, empty cells, numbers.
- Pain 4: Duplicated information: *"I remember a database of Covid symptoms where the symptoms were duplicated"*.
- Pain 5: *"One of the main challenges is data quality. It's basic and to be expected, but it's something that keeps happening. When I joined X [name of the company] I thought I was going to be in the panacea (...) my expectation when joining a technology company was to find better quality data and that didn't happen (...)"*



Open Knowledge

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE





# Most used tools



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Tools to explore and clean data

- Google Sheets.
- Excel / Open Office.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Tools to explore and clean data

- Open Refine.
- SQL.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Data visualization tools

- Flourish.
- Datawrapper.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# Other tools

- Add-ons
- Salesforce.
- Tabula



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE



# Other findings...



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

*"When I try to explore a database, to see if I can find something (...) I get very frustrated when I don't find [anything] because I don't know if I'm not searching properly or (...) I'm not applying the right techniques (...) I would like to know which steps I should follow when I have a dataset (...) to find something interesting"*

University professor / Journalist/  
Academic researcher.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

*"..the worst monster is when data disappears and you lose your work done"*

Investigative journalist working on transnational investigations.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE



*Google Sheets: just a few steps to convert csv into data "ready to use".*

Data journalist with basic coding skills



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

*"If you make a mistake in Google Sheets is relatively easy to find the error (...) you will see there is a problem in the data. In ChatGPT it is very difficult to see if you make a mistake ...."*

Data visualization designer and open data advocate.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

As a journalist the interviewee expresses concerns regarding the use of AI at work: *"If I am going to publish the whole database, I would upload it to ChatGPT, otherwise not (...). It is my data and the dataset that we are building"*.

Investigative journalist working on transnational investigations.



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

# What have we done so far?

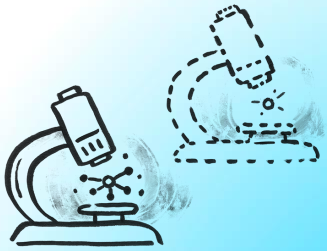


**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

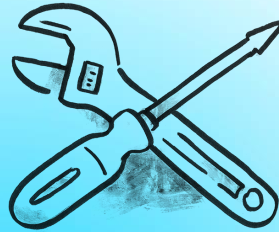
April 2024

# Five weeks recap!



## User research - Phase 1

Interview preparation + logistics.  
Ten people interviewed



## ODE repository

Reviewed of all existing issues  
to categorize them.  
Issue prioritization.  
Repository clean-up.



## Define first things to work on

Defined general issues to start  
working on.



## User personas

Created user personas for the  
ODE.

# Thanks!



**Open Knowledge**

FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

[okfn.org](https://okfn.org)  
[@okfn](https://twitter.com/okfn)