A data paper without actual data?

Diego Antolinos Basso

RSE Lightning Talks - 20260226 @ EPFL

Contact

- work: diego.antolinos@unine.ch
- personal: email@diegodiego.net
- social: diegantobass@fosstodon.org

Profile

- I'm a research software engineer in social sciences and humanities
- I'm a computational linguist, python developer and open source advocate

Today

- Can I publish a method/data paper without actually sharing any data?
- 10 minutes presentation with slides
- Interdisciplinary discussion on data (papers)

Draft abstract

This article describes the constitution of data corpora in the project Pandemic Data [insert ref]. It is a step by step depiction of the collection, verification and enrichment of the data, with the goal of identifying obstacles met and tactics of circumvention implemented during the project. We aim to contribute to the digital journalism and social sciences litterature on methods. Specificaly on the importance of transparency in data collection processes, at a time of data-ification of social sciences...

Project

Pandemic Data: Production, diffusion et compréhension des données en temps de pandémie

- FNS NPR80 "Covid-19 in Society" 3 years project
- Digital Journalism / Marketing / Systèmes d'information
- "How did we produce and narrate data during the covid pandemic ?"

Nota bene: I came onboard midway

Corpora

- 35k "covid" articles
- ↳ ~800 data visualizations
- 1500 official press releases
- 40 interviews

Data management plan

- Not allowed to share the data
- Must be destroyed within 5 years
- Unclear about secondary data
- Legally binding via FNS
- Negociating a new agreement?

But...

It feels like the story of the quirks and idiosyncrasies of the datasets are an important part of the project!

Data collection "flow" chart

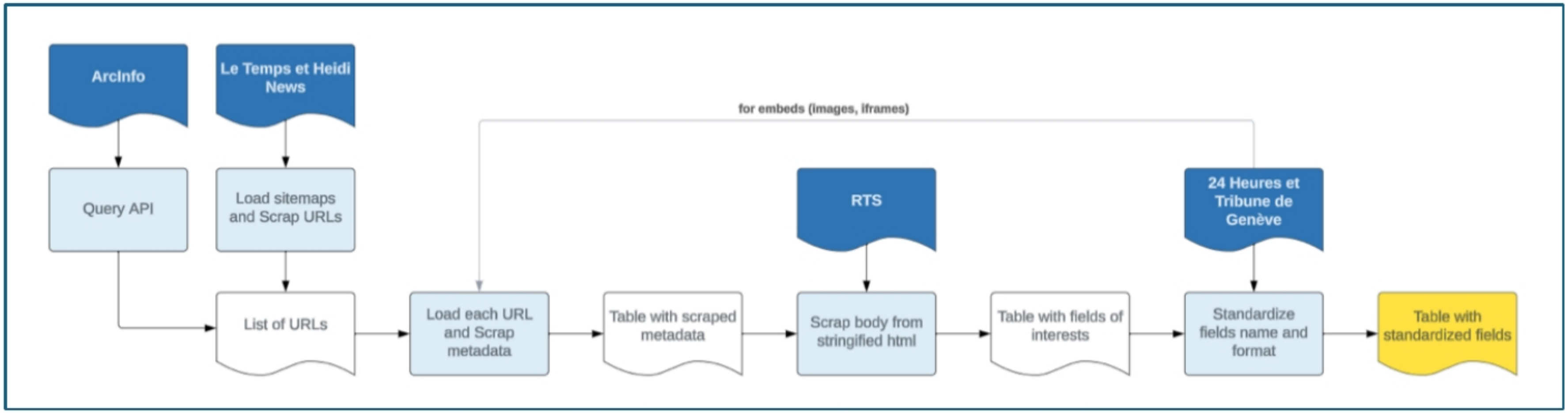
ArcInfo
Le Temps et Heidi News
for embeds (images, iframes)
Query API
Load sitemaps and Scrap URLs
RTS
24 Heures et Tribune de Genève
List of URLs
Load each URL and Scrap metadata
Table with scraped metadata
Scrap body from stringified html
Table with fields of interests
Standardize fields name and format
Table with standardized fields

Figure 1: Flowchart of data collection and cleaning

Credit : Céline Dupuis, Panda 2024

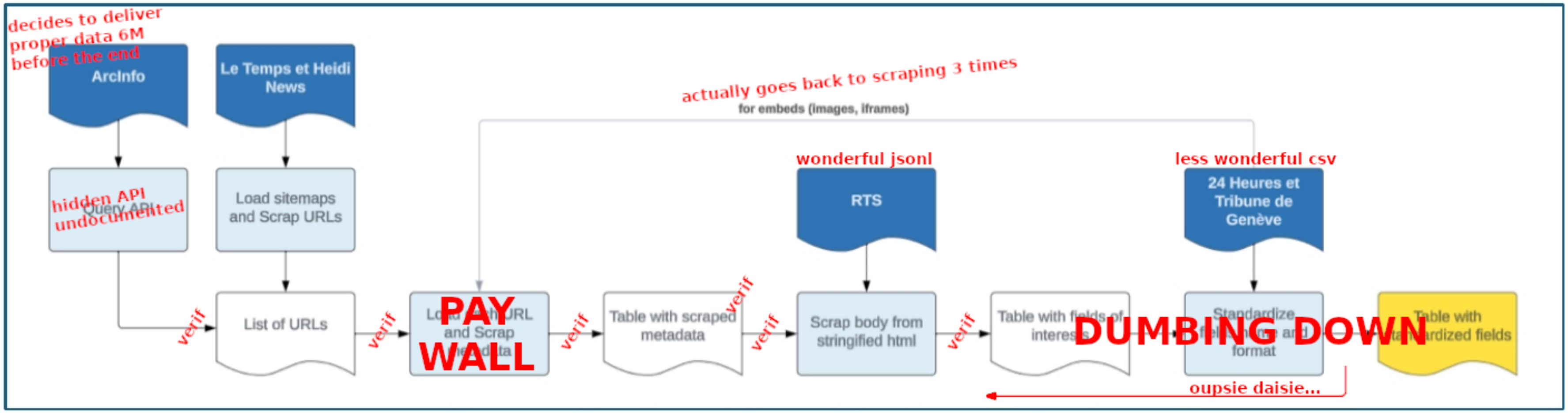
decides to deliver
proper data 6M
before the end
ArcInfo
Le Temps et Heidi News
actually goes back to scraping 3 times
for embeds (images, iframes)
wonderful jsonl
less wonderful csv
hidden API
undocumented
Query API
Load sitemaps and Scrap URLs
RTS
24 Heures et Tribune de Genève
verif
List of URLs
verif
Load each URL and Scrap Metadata
PAY
WALL
verif
Table with scraped metadata
verif
verif
Scrap body from stringified html
verif
Table with fields of interests
Standardize fields naming and format
DUMBING DOWN
Table with standardized fields
oupsie daisie...

Figure 1: Flowchart of data collection and cleaning

Credit : Diego Antolinos, Panda 2026

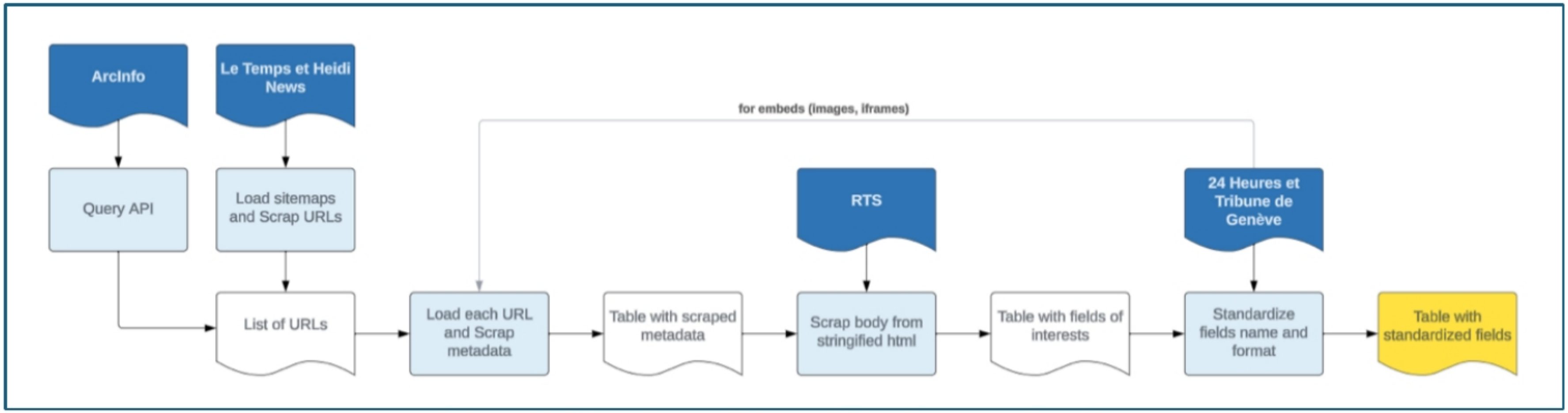
ArcInfo
Le Temps et Heidi News
Query API
Load sitemaps and Scrap URLs
for embeds (images, iframes)
RTS
24 Heures et Tribune de Genève
List of URLs
Load each URL and Scrap metadata
Table with scraped metadata
Scrap body from stringified html
Table with fields of interests
Standardize fields name and format
Table with standardized fields

Figure 1: Flowchart of data collection and cleaning

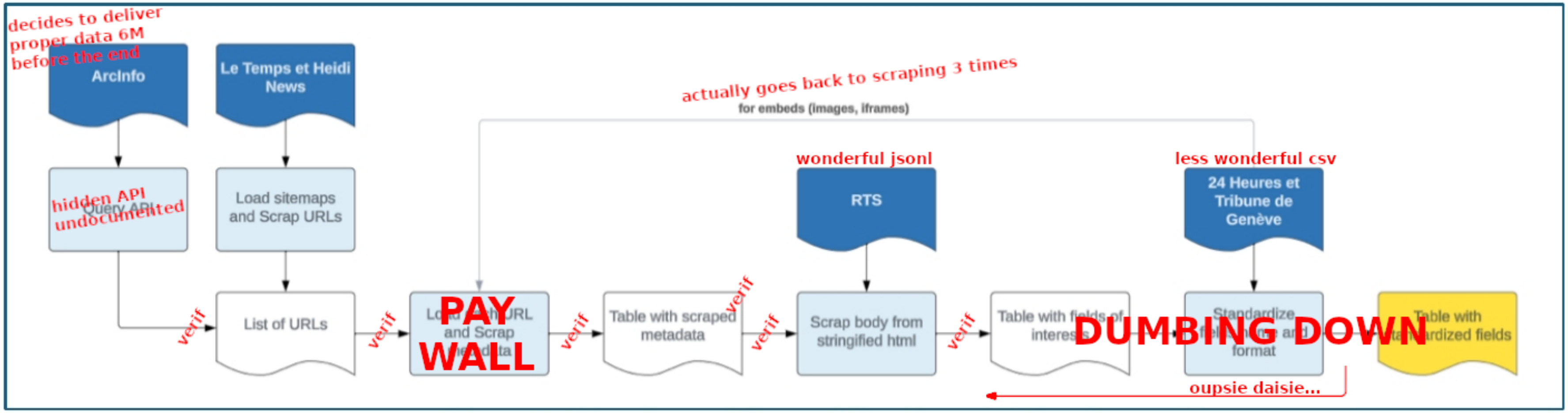
decides to deliver proper data 6M before the end
ArcInfo
Le Temps et Heidi News
actually goes back to scraping 3 times
for embeds (images, iframes)
wonderful jsonl
RTS
less wonderful csv
24 Heures et Tribune de Genève
hidden API undocumented
Query API
Load sitemaps and Scrap URLs
verif
List of URLs
verif
PAY WALL
Load each URL and Scrap metadata
verif
Table with scraped metadata
verif
verif
Scrap body from stringified html
verif
Table with fields of interests
DUMBING DOWN
Standardize fields name and format
Table with standardized fields
oupsie daisie...

Figure 1: Flowchart of data collection and cleaning

Example argument

"Dumbing down in data collection"

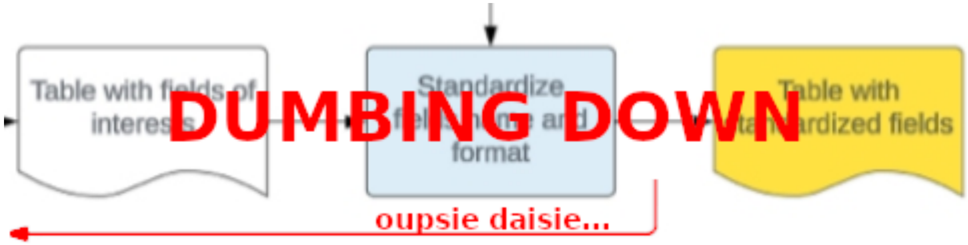
Table with fields of interests
Standardize field name and format
Table with standardized fields
DUMBING DOWN
oupsie daisie...

5 partners, 5 data sources, 5 times the headache

When collecting from multiple data sources, each with their practices/formats/tools, you can only keep what (metadata) you have in common between all sources

Dumbing down?

[showing the actual data]

Summary:

- 1 project partner can't deliver data
- Hidden API found from which data acquired
- Partner delivers data different from API
- What do?

Consequences

- Data practices influence hypotheses and results
- Dependency on data providers / partners
- Leveling of the richness of information collected
- Choices made early on have later consequences
- Very boring data curation problems tend to be made invisible

Other obstacles-arguments

- Filtering and creating secondary data
- Manual annotation plan of datapoints
- Intrication of primary and secondary data
- Scraping and its many epistemic problems

Open questions

- What is considered a proper data paper in your discipline?
- What's a method paper?
- Are the conditions in which a dataset has been collected explicited in your field? Is it something important and why?
- Does it sound obvious and uninteresting, yet crucial?

Notes

Making visible the dirty work of data wrangling strengthen the results. It increases reproducibility, shift back value on groundwork engineering and emphasize a scientific transparent approach to data science.