28. Searchable Databases as a Journalistic Product

Zara Rahman and Stefan Wehrmeyer

Abstract

Exploring the responsible data challenges and transparency opportunities of using public-facing searchable databases within a data journalism investigation.

Keywords: databases, responsible data, crowdsourcing, engagement, data journalism, transparency

A still emerging journalistic format is the searchable online database—a web interface that gives access to a data set, by newsrooms. This format is not new, but its use in data journalism projects is still relatively scarce (Holovaty, 2006).

In this chapter, we review a range of types of databases, from ones which cover topics which directly affect a reader's life, to interfaces which are created in service of further investigative work. Our work is informed by one of the co-author's work on Correctiv's Euros für Ärzte (Euros for Doctors) investigation, outlined below as an illustrative case study.¹ It is worth noting, too, that although it has become good practice to make raw data available after a data-driven investigation, the step of building a searchable interface for that data is considerably less common.

We consider the particular affordances of creating databases in journalism, but also note that they open up a number of privacy-related and ethical issues on how data is used, accessed, modified and understood. We then examine what responsible data considerations arise as a consequence of using data in this way, considering the power dynamics inherent within, as well as the

¹ <https://correctiv.org/recherchen/euros-fuer-aerzte/> (German language)

consequences of putting this kind of information online. We conclude by offering a set of best practices, which will likely evolve in the future.

Examples of Journalistic Databases

Databases can form part of the public-facing aspect of investigative journalism in a number of different ways.

One type of database which has a strong personalization element is *ProPublica's* Dollars for Docs. It compiled data on payments to doctors and teaching hospitals that were made by pharmaceutical and medical device companies.² This topic and approach was mirrored by *Correctiv* and *Der Spiegel* to create Euros für Ärzte, a searchable database of recipients of payments from pharmaceutical companies, as explained in further detail below. Both of these approaches involved compiling data from already-available sources. The goal was to increase the accessibility of said data so that readers would be able to search it for themselves to, for instance, see if their own doctor had been the recipient of payments. Both were accompanied by reporting and ongoing investigations.

Along similar lines, the *Berliner Morgenpost* built the Schul Finder to assist parents in finding schools in their area. In this case, the database interface itself is the main product.³

In contrast to the type of database where the data is gathered and prepared by the newsroom, another style is where the readers can contribute to the data, sometimes known as “citizen-generated” data, or simply crowdsourcing. This is particularly effective when the data required is not gathered through official sources, such as *The Guardian's* crowdsourced database The Counted, which gathered information on people killed by police in the United States, in 2015–2016.⁴ Their database used online reporting as well as reader input.

Another type of database involves taking an existing set of data and creating an interface that allows the reader to generate a report based on criteria they set. For example, the Nauru Files allows readers to view a summary of incident reports that were written by staff in Australia's detention centre on Nauru between 2013 and 2015.⁵ The UK-based Bureau

2 <https://projects.propublica.org/docdollars/>

3 <https://interaktiv.morgenpost.de/schul-finder-berlin/#/>

4 <https://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database>

5 <https://www.theguardian.com/australia-news/ng-interactive/2016/aug/10/the-nauru-files-the-lives-of-asylum-seekers-in-detention-detailed-in-a-unique-database-interactive>

of Investigative Journalism compiles data from various sources gathered through their investigations, within a database called Drone Warfare.⁶ The database allows readers to select particular countries covered and the time frame, in order to create a report with visualizations summarizing the data.

Finally, databases can also be created in service of further journalism, as a tool to assist research. The International Consortium of Investigative Journalists created and maintain the Offshore Leaks Database, which pulls in data from the Panama Papers, the Paradise Papers, and other investigations.⁷ Similarly, Organized Crime and Corruption Reporting Project (OCCRP) maintains and updates OCCRP Data, which allows viewers to search over 19 million public records.⁸ In both cases, the primary user of the tools is not envisioned to be the average reader, but instead journalists and researchers envisioned to carry out further research on whatever information is found using these tools.

The list below summarizes the different considerations in making databases as a news product:

- **Audience:** aimed at readers directly, or as a research database for other journalists
- **Timeliness:** updated on an ongoing basis, or as a one-off publication
- **Context:** forming part of an investigation or story, or the database itself as the main product
- **Interactivity:** readers encouraged to give active input to improve the database, or readers considered primarily as viewers of the data
- **Sources:** using already-public data, or making new information public via the database

Case Study: Euros für Ärzte (Euros for Doctors)

The European Federation of Pharmaceutical Industries and Associations (EFPIA) is a trade association which counts 33 national associations and 40 pharmaceutical companies among its members. In 2013, the federation decided that, starting in July 2016, member companies must publish payments to healthcare professionals and organizations in the countries they operate (EFPIA, 2013). Inspired by *ProPublica*'s Dollars for Docs project, the non-profit German investigative newsroom Correctiv decided to collect

6 <https://www.thebureauinvestigates.com/projects/drone-war/>

7 <https://offshoreleaks.icij.org/>

8 <https://data.occrp.org/>

these publications from the websites of German pharmaceutical companies and create a central, searchable database of recipients of payments from pharmaceutical companies for public viewing. They named the investigation Euros für Ärzte (Euros for Doctors).

In collaboration with the German national news outlet *Der Spiegel*, documents and data were gathered from around 50 websites and converted from different formats to consistent tabular data. This data was further cleaned and recipients of payments from multiple companies were matched. The total time for data cleaning was around ten days and involved up to five people. A custom database search interface with individual URLs per recipient was designed and published by Correctiv.⁹ The database was updated in 2017 with a similar process. Correctiv also used the same methodology and web interface to publish data from Austria, in cooperation with derStandard.at and ORF, and data from Switzerland with Beobachter.ch.

The journalistic objective was to highlight the systemic influence of the pharmaceutical industry on healthcare professionals through events and organizations, and the associated conflicts of interest. The searchable database was intended to encourage readers to start a conversation with their doctor about the topic, and to draw attention to the very fact that this was happening.

On a different level, the initiative also highlighted the inadequacy of voluntary disclosure rules. Because the publication requirement was an industry initiative rather than a legal requirement, the database was incomplete—and it's unlikely that this will change without legally mandated disclosure.

As described above, the database was incomplete, meaning that a number of people who had received payments from pharmaceutical companies were missing from the database. Consequently, when users search for their doctor, an empty result can either mean the doctor received no payment or that they denied publication—two vastly different conclusions. Critics have noted that this puts the spotlight on the cooperative and transparent individuals, leaving possibly more egregious money flows in the dark. To counter that, Correctiv provided an opt-in feature for doctors who had not received payments to also appear in the database, which provides important context to the narrative, but still leaves uncertainty in the search result.

After publication, both Correctiv and *Der Spiegel* received dozens of complaints and legal threats from doctors who appeared in the database. As the data came from public, albeit difficult to find, sources, the legal team of *Der Spiegel* decided to defer most complaints to the pharma companies and only adjust the database in case of changes at the source.

9 <https://correctiv.org/thema/aktuelles/euros-fuer-aerzte/>

Technical Considerations of Building Databases

For a newsroom considering how to make a data set available and accessible to readers, there are various criteria to consider, such as size and complexity of the data set, internal technical capacity of the newsroom, and how readers should be able to interact with the data.

When a newsroom decides that a database could be an appropriate product of an investigation, building one requires bespoke development and deployment—a not insignificant amount of resources. Making that data accessible via a third-party service is usually simpler and requires fewer resources.

For example, in the case of Correctiv, the need to search and list around 20,000 recipients and their financial connections to pharma companies required a custom software solution. Correctiv developed the software for the database in a separate repository from its main website but in a way it could be hooked into the content management system. This decision was made to allow visual and conceptual integration into the main website and investigation section. To separate concerns, the data was stored in a relational database separate from the content database. In this case, having a process and interface for adjusting entries in the live database was crucial as dozens of upstream data corrections came in after publication.

However, smaller data sets with simple structures can be made accessible without expensive software development projects. Some third-party spreadsheet tools (e.g., Google Sheets) allow tables to be embedded. There are also numerous front-end JavaScript libraries to enhance HTML tables with searching, filtering and sorting functionalities which can often be enough to make a few hundred rows accessible to readers.

An attractive middle ground for making larger data sets accessible are JavaScript-based web applications with access to the data set via API. This setup lends itself well to running iframe-embeddable search interfaces without committing to a full-fledged web application. The API can then be run via third party services while still having full control over the styling of the front end.

Affordances Offered by Databases

Databases within, or alongside, a story, provide a number of affordances for both readers and newsrooms.

On the reader side, providing an online database allows readers to search for their own city, politician or doctor and connects the story to their own life. It provides a different channel for engagement with a story on a more personal level. Provided there are analytics running on these search queries, this also gives the newsroom more data on what their readers are interested in—potentially providing more leads for future work.

On the side of the newsroom, if the database is considered as a long-term investigative investment, it can be used to automatically cross-reference entities with other databases or sets of documents for lead generation. Similarly, if or when other newsrooms decide to make similar databases available, collaboration and increased coverage becomes much easier while reusing existing infrastructure and methodologies.

Databases also potentially offer increased optimization for search engines, thus driving more traffic to the news outlet website. When the database provides individual URLs for entities within, search engines will pick up these pages and rank them highly in their results for infrequent keyword searches related to these numerous entities—the so-called “long tail” of web searches, thus driving more traffic to the publisher’s site.

Optimizing for search engines can be seen as an unsavoury practice within journalism; however, providing readers with journalistic information while they are searching for particular issues can also be viewed as a part of successful audience engagement. While the goal of the public database should not be to compete on search keywords, it will likely be a welcome benefit that drives organic traffic, and can in turn attract new readership.

Responsible Data Considerations

Drawing upon the approach of the responsible members of the data community, who work on developing best practices which take into account the ethical and privacy-related challenges faced by using data in new and different ways, we can consider the potential risks in a number of ways.¹⁰

First is the question of the way in which power is distributed in this situation, where a newsroom decides to publish a database containing data about people. Usually, those people have no agency or ability to veto or correct that data prior to publication. The power held by these people depends very much upon who they are—for example, a politically exposed person (PEP) included in such a database would presumably have both the

10 <https://responsibledata.io/what-is-responsible-data/>

expectation of such a development and adequate resources to take action, whereas a healthcare professional would probably not be expecting to be involved in an investigation. Once a database is published, visibility of the people within that database might change rapidly—for example, doctors in the *Euros für Ärzte* database gave feedback that one of the top web search results for their name was now their page in this database.

Power dynamics on the side of the reader or viewer are also worth considering. For whom could the database be most useful? Do they have the tools and capacity required to be able to make use of the database, or will this information be used by the already-powerful to further their interests? This might mean widening the scope of user testing prior to publication to ensure that enough context is given to properly explain the database to the desired audience, or including certain features that would make the database interface more accessible to that group.

The assumption that more data leads to decisions that are better for society has been questioned on multiple levels in recent years. Education scholar Clare Fontaine (2017) expands upon this, noting that in the United States, schools are becoming more segregated despite (or perhaps because of) an increase in data available about “school performance.” She notes that “a causal relationship between school choice and rampant segregation hasn’t yet been established,” but she and others are working more to understand that relationship, interrogating the perhaps overly simplified relationship that more information leads to better decisions, and questioning what “better” might mean (Fontaine, 2017).

Second is the question of the database itself. A database on its own contains many human decisions; what was collected and what was left out, and how it was categorized, sorted or analyzed, for example. No piece of data is objective, although literacy and understanding of the limitations of data are relatively low, meaning that readers could well misunderstand the conclusions that are being drawn.

For example, the absence of an organization from a database of political organizations involved in organized crime may not mean that the organization does not take part in organized crime itself; it simply means that there was no data available about their actions. Michael Golebiewski and Danah Boyd (2018) refer to this absence of data as a “data void,” noting that in some cases a data void may “passively reflect bias or prejudice in society.” This type of absence of data in an otherwise data-saturated space also maps closely to what Brooklyn-based artist and researcher Mimi Onuoha (2016) refers to as a “missing data set,” and highlights the societal choices that go into collecting and gathering data.

Third is the direction of attention. Databases can change the focus of public interest from a broader systemic issue to the actions of individuals, and vice versa. Financial flows between pharmaceutical companies and healthcare professionals are, clearly, an issue of public interest—but, on an individual level, doctors might not think of themselves as a person of public interest. The fact remains, though, that in order to demonstrate an issue as broader and systemic (as a pattern, rather than a one-off), data from multiple individuals is necessary. Some databases, such as the Euros für Ärzte case study mentioned above, also change boundaries of what, or who, is in the public interest.

Even when individuals agree to the publication of their data, journalists have to decide how long this data is of public interest and if and when it should be taken down. The General Data Protection Regulation (GDPR) will likely affect the way in which journalists should manage this kind of personal data, and what kinds of mechanisms are available for individuals to rescind consent to their data being included.

With all of these challenges, our approach is to consider how people's rights are affected by both the process and the end result of the investigation or product. At the heart is understanding that responsible data practices are ongoing approaches rather than checklists to be considered at specific points. We suggest that approaches which prioritize the rights of people reflected in the data throughout the entire investigation, from data gathering to publication, are a core part of optimizing (data) journalism for trust (Rosen, 2018).

Best Practices

For journalists thinking of building a database to share their investigation with the public, here are some best practices and recommendations. We envision these will evolve with time, and we welcome suggestions.

First, ahead of publication, develop a process for how to fix mistakes in the database. Good data provenance practices can help to find sources of errors. Second, build in a feedback channel. Particularly when individuals are unexpectedly mentioned in an investigation, there is likely to be feedback (or complaints). Providing a good user experience for them to make that complaint might help the experience. Third, either keep the database up to date, or clearly mark that it is no longer maintained. Within the journalistic context, publishing a database demands a higher level of maintenance than publishing an article. The level of interactivity that a database affords

means that there is a different expectation of how up to date it is compared to an article. Fourth, allocate enough resources for maintenance over time. Keeping the data and database software up to date involves significant resources. For example, adding data from the following year to a database requires merging newer data with older data, and adding an extra time dimension to the user interface. Fifth, observe how readers are using the database. Trends in searches or use might provide leads for future stories and investigations. Finally, be transparent: It's rare that a database will be 100% "complete," and every database will have certain choices built into it. Rather than glossing over these choices, make them visible so that readers know what they're looking at.

Works Cited

- EFPIA. (2013). About the EFPIA Disclosure Code. European Federation of Pharmaceutical Industries and Associations. https://efpia.eu/media/25046/efpia_about_disclosure_code_march-2016.pdf
- Fontaine, C. (2017, April 20). Driving school choice. *Medium*. <https://points.datasociety.net/driving-school-choice-16f014d8d4df>
- Golebiewski, M., & Boyd, D. (2018, May). Data voids: Where missing data can be easily exploited. *Data & Society*. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf
- Holovaty, A. (2006, September 6). A fundamental way newspaper sites need to change. Adrian Holovaty. <http://www.holovaty.com/writing/fundamental-change/>
- Onuoha, M. (2016, February 3). On missing data sets. <https://github.com/MimiOnuoha/missing-datasets>
- Rosen, J. (2018, May 14). Optimizing journalism for trust. *Medium*. <https://medium.com/de-correspondent/optimizing-journalism-for-trust-1c67e81c123>

About the Authors

Zara Rahman is deputy director at the Engine Room and a fellow at the Digital Civil Society Lab at Stanford University's Centre for Philanthropy and Civil Society.

Stefan Wehrmeyer is the founder of FragDenStaat.de, Germany's Freedom of Information Portal, and works as a data journalist and developer.