Which obstacles does Data-Driven Journalism have to overcome?

Johanna Fulda and Stefanie Neubert

LMU Munich
{Johanna.Fulda,Stefanie.Neubert}@campus.lmu.de

Abstract. Data-driven journalism is a topic which could become very important for journalism in the future. However, the usage of data-driven journalism spreads more or less slowly. Some publishing companies already make use of data-driven journalism in few projects, but this type of journalism is still in an early stage of development. In this article we present an overview over past popular projects involving data-driven journalism, it's current usage and it's usage potential in the future. We analyse the problems which might cause the slow spreading of this type of journalism and make suggestions to overcome these problems.

Keywords: data, journalism, data-driven journalism, computational journalism

1 Introduction

Data-driven journalism is a different type of journalistic activity. Usually, journalists have a topic to process and thus do research for data. Data-driven journalism assumes a huge amount of raw, unsorted data and tries to find interesting facts to create stories out of the data. Thus, the processing of data-driven journalism is the reverse to the processing of the usual, well-known journalism.
Data-driven journalism can be categorized into four consecutive steps [2, p. 1]:

- finding data,
- interrogating data,
- visualizing data and
- evaluating data.

Finding data can include technical skills or good contacts. If no data is published, there is no possibility to start a data-driven journalism project. To interrogate data, topic specific difficulties like the used jargon and the topic's context have to be understood. In addition, data has to be made machine-readable to process it easily in the further steps. Visualizing and evaluating the data is required to provide a clear representation of the data which is then published for the readers.

2 Data-driven journalism since WikiLeaks

The first step of data-driven journalism can be realized by WikiLeaks. WikiLeaks is a popular platform covering huge amounts of materials. It provides the data needed for any further steps, e.g., the data can be inspected by journalists and newspapers looking for new stories, without the requirement of technical skills to search the web for data.

2.1 Processing of WikiLeaks

As mentioned above, WikiLeaks covers the data for further processing. However, it only provides the raw material in terms of huge amounts of unsorted documents, which contain jargon and are not necessarily machine-readable. There is no possibility to realize the further steps of data-driven journalism, i.e., to generate visualizations or stories out of the data. These steps have to be managed with other tools. Additionally, WikiLeaks does not filter the data which is published. This means that data containing sensitive information is published despite of legal regulations. Thus, a journalist judges whether to publish sensitive data or not.

2.2 Journalism dealing with large amounts of data

Exemplifying the *Guardian*, projects concerning large amounts of data are processed as follows. Since editors and readers do not have the ability to access such large amounts of data in a reasonable way, programmers gain access to the data by providing a user interface. This user interface allows editors and readers to filter and search data as well as to analyse it. This processing leads to the disclosure of interesting facts [13, p. 118]. The approach to use a person gaining access to the data and one or more persons analysing it is frequently used for average projects. Some publishing companies make use of external agencies in extensive projects [7, p. 1].
Another aspect to mention is crowdsourcing, which means to include the readers into the processing. The published data of the British government is an example for this aspect. A web application was provided by the *Guardian* which allowed the readers to search over 500 000 documents for interesting data [11, p. 7].
Note that in the processing an extra person is required to gain access to the data. This poses an additional cost factor which does not exist in the usual journalism and thus might be a reason for publishing companies not to process many data-driven journalism projects.

2.3 Meaningful Examples

Afghanistan war logs. There are several examples showing a successfull deployment of data-driven journalism. The most popular example is the release of Afghanistan war logs [10]. A huge amount of war logs of the Afghanistan war

were published by WikiLeaks and analysed by readers of the *Guardian*. The geographic data of the war logs was mapped. The result indicated the spread of the insurgery, the hotspots of the war and the relocation of the war from 2004 to 2009. The publication of the geographic maps were subject of many media reports. Critisism was aroused since the war logs contained detailed information about the actions of the war and the numbers of civilian casualties and were not intended for publishing [10, p. 3–5].
In this example, WikiLeaks was used to realize the first step of data-driven journalism, *finding data*. A potential problem might be the access to data for WikiLeaks itself. Persons have to exist leaking the data to WikiLeaks. If the data is not allowed to be published, these persons act illegal.

Crime maps. Crime maps can be considered as another meaningful example [12]. Crime maps depict approximate geographic locations of criminal activity. In contrast to the publication of data tables, the publication of the data in crime maps allows people to identify critical locations in a more intuitive way [12, p. 2]. The publication of crime maps on the Internet arises conflicts, since victims could be indentified if there are taken no actions to protect their privacy. Further, economic disadvantages can arise in areas of high amounts of crime, such as low house prices or high insurance premiums [12, p. 10]. As an example, fig. 1 shows a crime map of London.
A problem arising for the data-driven journalism project might be the requirement to filter the sensitive data by eliminating names of victims and offenders as well as the exact location and date.

3 Data-driven journalism and newspapers

3.1 Usage today

Open Data Movement Like many organisations these days, the open data movement wants more transparency. In this special case they claim transparency and accessibility in data. Their request is that all the data that is available on the Internet should be in a uniform machine-readable format [6]. That would make it easy to access and to evaluate the data. Also they want data concerning matters of public interest or generally of public authorities to be open, free and easily accessible. In Germany the *OpenData Network* came up to push forward this progress. They say that the trust in the government and the understanding of many issues would raise and the political apathy of the poulation would be combated [9], [8]. Even if there are some constitutions that already publish their data on their websites (e.g., the federal ministry for finances in Germany), there are many points that still lack until they fulfil the requirements of the OpenData Network. As an example, they publish the finance plans of the years in huge PDF files, where every expenses are listed in detail, but to make these files accessible through an API they have to be reformatted by hand which is a quite fault-prone and tedious process. So the OpenData Network wants these establishments to

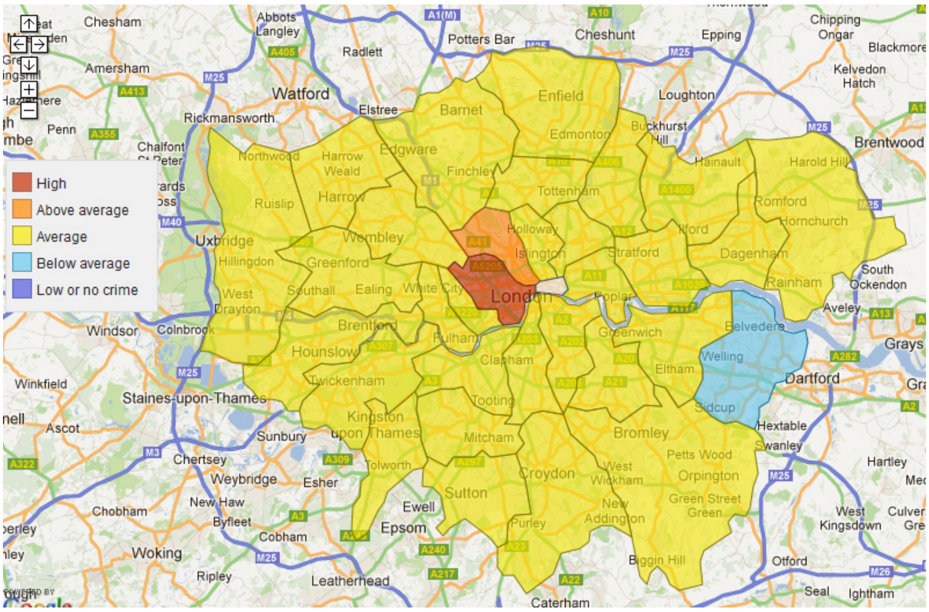

Fig. 1. A crime map of London, showing the relative crime rate of each borough. Source: http://maps.met.police.uk/

use an XML or JSON format or at least put their data into a table (e.g. in Excel – even if that would be a proprietary format – but more accessible than just a poorly structured text file) [6], [9].

Recent remarkable examples Like mentioned above, the federal ministry for finances shows its financial plans online. That means the data is available, but, because they put the data in a PDF format, it is not accessible. The nonprofit organization *Open Knowledge Foundation Germany e.V.* took all these information out of the proprietary format, translated it into a machine-readable format and created a website called *OffenerHaushalt.de*. There they used the translated data for interactive visualisations (The evaluation is dedicated to the visitor). But, as they say it was a long and unpleasant process and because it had to be done "by hand" it is not guaranteed, that there aren't some mistakes in it [4]. All these work could be avoided if the establishments just gave away the raw data.

One good example is the open data of *The World Bank* which attends to the economic development of less developed countries. For over 30 years now, they gather international projections of the population. It is all about economy, environmental protection and social projects. Through this collected data, they can assess the impact of their help. Since 2010 the data is available for everybody at their website *data.worldbank.org*. There they invite the visitors to use them for visualizations and evaluations, to illustrate global issues. The data is machine-

readable and divided into many different categories, so it is very comfortable and easy to access the desired data. There are already countless interactive and clear visualisations (and it is becoming more and more) [5].

Publishing companies The mentioned examples (and there are of course many more) are projects of very motivated people or of huge organisations. These people want to show what is possible through appropriate access to data and have spared no expenses. But what is it like for companies which want to use it, but don't have that much energy and time for it? E.g. publishing companies, especially newspapers or magazines, respectively their websites. The Guardian has some kind of a pioneering role for data journalism, but what about an average daily paper in Germany? We asked the head of the online department of the Süddeutsche Zeitung, Stefan Plöchinger. He is a big enthusiast in data-driven journalism and already published different quite big projects based on data. E.g. the *Zugmonitor*, see fig. 2, which was published in March 2012. There they worked together with *OpenDataCity* (a company specialized in data journalism) and visualized each train connection since October 2011. It shows the delays of every single train and thus reveals the weak points of the rail network. Through an API the data can be viewed at a specific time in the past and also in a live mode, in this way one can check which train is delayed at this very moment.
Such a project needs many steps and different types of expertise. In the example of the "Zugmonitor" finding the data already took much time, because the data was not freely available and the Deutsche Bahn also didn't publish it. Then there was the programming and implementing of an API, the designing with regards to usability (operability and comprehensibility) and finally the editorial evaluation. [7]
Even if this was one of the projects with the most amount of work, Plöchinger says that they try to realise data driven projects every one or two months. Especially for events like the European Championships, the Olympic Games or political elections, bigger projects can be planned detailed. Even if this format is better suited for interactive use (online media), the results of the evaluation of big data can also be presented in the print version of the newspaper (e.g., you can only show the relevant datasets) [7]. So the question arises, how much work is reasonable and affordable for such a project.

4 Detected obstacles

There are various obstacles existing, which can complicate or even obstruct data-driven journalism projects. Coming back to the four steps of the processing of data-driven journalism, most of the problems arise in the first two steps. A great problem is the availability of data referring to the first step, *finding data*. Most of the data is not published, even tough it is retained somewhere. People giving access to this data act illegal. As an example, the Afghanistan war logs were published illegally. Another obstacle is the format of the published data, referring to the second step, *interrogating data*. Governments publish their data, but in a

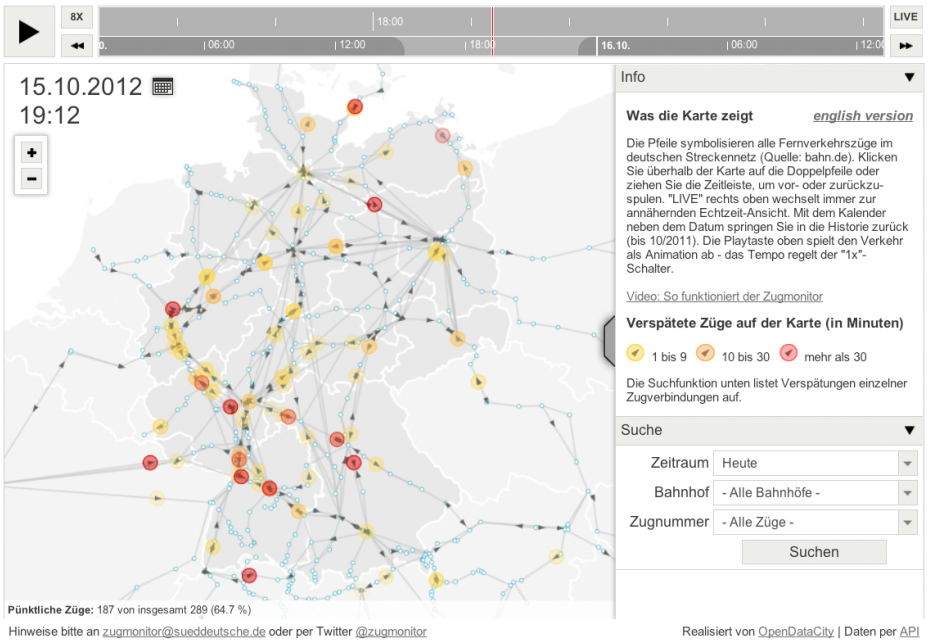

Fig. 2. The live view of the "Zugmonitor", published on sueddeutsche.de in March 2012. Source: http://zugmonitor.sueddeutsche.de/

form which is not machine-readable and thus cannot be interrogated easily. The data might contain jargon as well. This data has to be made machine-readable by copying the contents of the paper manually which poses a great waste of time. The open data movement demands the publication of data in a machine-readable form. Another problem arising in the second step is the existence of sensitive data, which has to be eliminated before it is published, e.g., the crime data has to be filtered before crime maps can be published. Another problem arising is the fact that the data might not contain any interesting facts and thus the project ends without any results.
Especially the obstacles referring to the *interrogating data* step cause a huge amount of work, since it has to be done manually. This in turn causes an imense cost factor, which is not present in usual journalism projects.

Profitability for publishing companies It is generally hard to calculate profitability for online media, because mostly it is for free and the financing is achieved over advertisement. The advertising expense depends on the average number of visitors. Therefore it is not possible to calculate work intensive projects like effort equals reward, but it is more important to attract people, so that they will come back. Plöchinger says it is more about marketing. To establish successfully on the market of online news you have to be eager to experiment. If the visitor finds new and interesting things on a website, he or she

will probably consider to use this site to catch up on news in the future and with that the ranking and the interest of companies to advertise there raises. So it is profitable indirectly and "indeed economically to do those things" says Plöchinger. [7]
The main reason why there are only few companies which use this form of journalism is that there are still too less people who can handle big amounts of data. A data jounalist needs to have an informatical background or at least be quite affine with data management, such as Excel charts or databases.

4.1 Usage potential

Capability of necessary techniques is growing The mentioned problem of the missing ability of many journalists to work with data will be less important in the future, because the upcoming generation of digital natives grew up with computers and the necessary tools and is also taught at school how to work with computer programs and tools. Also the basic knowledge of databases is getting part of the curriculum in school. [11, p. 11], [1, p. 16]

More data is available In the future more data will be available. First of all because of the possibilities that come with the digitalization of our environment – all collected data is stored digitally so it is almost no effort to make it public – and secondly because of the demand for more transparency. Even if it will not be as transparent as many political parties or freedom fighters wish, public establishments can no longer justify why they keep their collected data as a secret. If something is paid with taxes of the citizens, they need to have the right to see what happens with it. Also if something affects the general public, it should be free and available for everybody [2].

More tools available Also the development of helpful tools to visualize data is growing. Even if there are already many quite helpful and easy-to-use tools, only few people use them, maybe because they find it still too complicated or confusing. But because there are more and more cases of application, there will be more and more examples to copy from, or to let them inspire the journalists.[11, p. 9]

More budget available And last but maybe most important, the budget for data-driven projects will raise. Publishing companies spread their money on fixed issues, such as news, articles, reportages, photographs, infograhics etc. So far there is no budget provided for time-consuming data-driven journalism. Publishing companies have to become aware of this new form of journalism and have to include it into their financial plan.[3, p. 8]

5 Conclusion and Future Prospects

Data-driven journalism will be more present in the future and is getting a new approch to storytelling. The preparation for visualizations of huge amounts of data is getting easier and more accessible. Especially the generation of the digital natives is familiar with the required tools and techniques. Of course data-driven journalism is not appropriate for each topic. It only makes sense if the data lets you discover new insights. If the data reveals facts that were not obvious before and maybe even unexpected or enlightening, publishing companies will definitely profit from this way of journalism. The question of profitability cannot be answered explicitly, because the turnover of online journalism is not so obvious, but even if it will take some more time to get it over its teething problems, as a tool of marketing it definitely makes sense and has the chance to be used more often in the future.

References

1. Tanja Aitamurto, Esa Sirkkunen, and Pauliina Lehtonen. Trends in data journalism. pages 1–27, 2011.
2. Stefan Baack. A new style of news reporting: Wikileaks and data-driven journalism, July 2011.
3. Anna Daniel, Terry Flew, and Christina Spurgeon. The promise of computational journalism. In K. McCallum, editor, *Media, Demcoracy and Change: Refereed Proceedings of the Australian and New Zealand Communications Association Annual Conference*, pages 1–19, Canberra, ACT, 2010. Australia and New Zealand Communication Association.
4. Open Knowledge Foundation Deutschland. OffenerHaushalt - Hinweis zu Datenqualität, 2010.
5. Anke Domscheit-Berg. Auch Die Weltbank Legt Ihre Daten Offen - Armut Bekämpfen mit Mashups, April 2010.
6. Open Knowledge Foundation. The open data handbook, 2010-2012.
7. Johanna Fulda. Interview with Stefan Plöchinger about Datenjournalismus. December 2012.
8. Sean Maguire. Can data deliver better government? *The Political Quarterly*, 82(4):522–525, October–December 2011.
9. Lorenz Matzat. Definitionen: OpenData, OpenGovernment, Gov2.0 und Co., August 2010.
10. John O'Loughlin, Frank D. W. Witmer, Andrew M. Linke, and Nancy Thorwardson. Peering into the fog of war: The geography of the WikiLeaks Afghanistan War Logs, 2004–2009. *Eurasian Geography and Economics*, 51(4):472–495, 2010.
11. Sylvain Parasie. 'Hacker' journalism – a new utopia for the press?
12. Jerry H. Ratcliffe. Damned if you don't, damned if you do: Crime mapping and its implications in the real world. *Policing and Society*, 12(3):211–225, 2002.
13. Simon Rogers. Wikileaks und der investigative Datenjournalismus. Wie wir beim Guardian mit den Wikileaks-Dateien umgehen. In Heinrich Geiselberger, editor, *WikiLeaks und die Folgen*, chapter 3, pages 118–127. Suhrkamp, 2011.