DATA & computational Journalism

CONFERENCE

20TH/21ST JUNE 2018

CARDIFF UNIVERSITY
CARDIFF, WALES

CONFERENCE PROCEEDINGS

@DATAJCONF #DATAJCONF

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD School of Computer Science and Informatics / Ysgol Cyfrifiadureg a Gwybodeg
CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD School of Journalism, Media and Cultural Studies / Ysgol Newyddiaduraeth, y Cyfryngau ac Astudiaethau Diwylliannol
UCD DUBLIN School of Information and Communication Studies
UCD DUBLIN College of Social Sciences and Law
HACKS/HACKERS SWxSW
HACKS/HACKERS DUBLIN
Google News Initiative

Editors:
Bahareh Heravi, Martin Chorley, Glyn Mottershead

Authors:
Silvia Majo-Vazquez, Jason R. C. Nurse, Jun Zhao and Rasmus K Nielsen, Bahareh Heravi, Pablo Leon Villagra, Sarwar Islam, Megan Lucero, Brooks Paige and Tomas Petricek, Mariana Marasoiu

ISBN: 978-1-910963-23-4

Welcome to the 2nd European Data and Computational Journalism Conference!

The European Data and Computational Journalism Conference aims to bring together industry, practitioners and academics in the fields of journalism and news production and information, data, social and computer sciences, facilitating a multidisciplinary discussion on these topics in order to advance research and practice in the broad area of Data and Computational Journalism.

Held in Cardiff, Wales, UK the 2nd edition of the conference will present a mix of academic talks and keynotes from industry leaders. It will be followed by a day of workshops.

Submissions of academic research focused talks, industry talks, as well as hands-on workshops were invited for the conference, on the subjects of journalism, data journalism, computational journalism, and information, data, social and computer science.

Topics of interest include, but are not limited to:

- Application of data and computational journalism within newsrooms
- Data driven investigations
- Data storytelling
- Open data for journalism, storytelling, transparency and accountability
- Algorithms, transparency and accountability
- Automated, robot and chatbot journalism
- Newsroom software and tools
- 'Post-fact' journalism and the impact of data
- User experience and interactivity
- Data and Computational Journalism education
- Post-desktop news provision/interaction
- Data mining news sources
- Visualisation and presentation
- Bias, ethics, transparency and truth in Data Journalism
- Newsroom challenges with respect to data journalism, best practices, success and failure stories

Collected within these proceedings are the academic abstracts presented at the conference.

We would like to take this opportunity to thank the programme committee for their hard work reviewing submissions and helping us to come up with the fantastic line-up of talks for this year. Welcome to Cardiff, and welcome to DataJConf 2018!

Bahareh R Heravi, Martin J Chorley & Glyn Mottershead
DataJConf 2018 co-chairs

Fragmentation across media platforms: Mapping Audience behaviour on the web, Facebook and Twitter

Silvia Majo-Vazquez
University of Oxford
silvia.majo-vazquez@politics.ox.ac.uk

Jason R. C. Nurse
University of Oxford
jason.nurse@cs.ox.ac.uk

Jun Zhao
University of Oxford
junhao@cs.ox.ac.uk

Rasmus K. Nielsen
University of Oxford
rasmus.nielsen@politics.ox.ac.uk

Abstract: The study of audience fragmentation has recently attracted great interest. We contribute to this debate by analyzing the level of audience fragmentation on three different media platforms of the French online domain. To this end we map several audience networks: the networks of news users when navigating the web generally; the network emerging from news consumption on Facebook; and finally, the networks of paths emerging from Twitter news users. Using a community detection approach, we determine the level of fragmentation on each of these networks. Then, we provide evidence to identify the drivers that influence online exposure to political news. In particular, we analyze our data using inferential network analysis to account for the influence of news consumption patterns across platforms and control for self-organizing network forces. We contribute to the debate of audience fragmentation focusing in France as a case study and by providing evidence at the comparative level. We argue that it is essential to jointly study the community structure emerging from these different media landscapes, if we are to determine the extent that the online news domain is fragmented. Until now, scholars have made assumptions about the structure of digital news domain through the lens of one of them i.e. studying patterns of news consumption on the web or on one single social media platforms. We discuss the implications of these findings for theoretical and empirical accounts of how digital technologies contribute toward audience polarization and "echo-chambers".

Keywords: news audience, social network analysis, audience fragmentation, audience behaviour, news media

Introduction

Information is at the center of civic life and diverse news media diets are deemed essential for democracy (Katz, 1996; Delli Carpini, 2004). The extent to which people are exposed to cross-cutting i.e., diverse news information is a key determinant to explain levels of political polarization and identify mechanisms behind critically-informed public opinion. It is well stablished that the level of diversity of news consumption patterns can be approximated by measuring the level of structural fragmentation of the news distribution systems under study: the online domain (Webster and Ksiazek, 2012; Dvir-Gvirsman, 2016; Taneja and Webster, 2016), the online and offline domain (Fletcher and Nielsen, 2017) , the blogosphere (Adamic and Glance, 2005) and the social platforms (Bakshy, Messing and Adamic, 2015) . We built on this approach, which basically, measures the emergence of communities on those systems and the level of connections among them to measure the level of diversity of news consumption patterns. These studies so far, have not addressed the differences in the level of fragmentation across the difference platforms of the news domain neither have they studied whether those patterns influence each other. We want to contribute to the previous literature by filling this gap. To this end, we assess whether and to what extent the news-seeking patterns of users of social media platforms correlate with the structure of those who navigate the web. This leads to our main research question: Do news media platforms and the web significantly vary in terms of their structural fragmentation? Additionally, we also aim at providing evidence to identify the drivers that influence the emergence of those patterns and hence, the online exposure to news. To this end, we analyze our data using inferential network analysis (Exponential Random Graph Models, ERGM) to account for the driving forces behind the news consumption patterns across platforms and control for endogenous network dynamics.

Methods and Data

Our data strategy includes three novel datasets, which map the audience behavior across the web, Facebook and Twitter during the French electoral campaign in 2017. As shown on Figure 1, we analyze three different types of networks. First, we reproduce the paths emerging from news users on the web (data is provided on monthly basis and we average data from April and May, when the campaign took place). We use the web network as a baseline against which we compare the level of fragmentation in Facebook and Twitter. Nodes of the web network represent news media outlets and two nodes are connected if they share a certain amount of audience. In the Facebook and Twitter networks, nodes also represent news media outlets and ties represent news consumption by measuring the amount of audience that 1) commented on a pair of news sites on Facebook or 2) RT,

mention or reply a pair of news sites on Twitter. Therefore, ties here directly measure the amount of people that not only was exposed to news content but actually, those who read it.

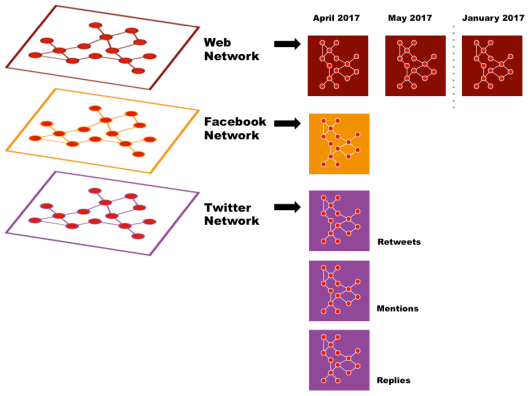

Figure 1 - *Figure 1. Model online media landscapes*

In order to operationalize the concept of media fragmentation, we use a community detection method. This technique reduces our three types of networks by classifying nodes into modules according to the density of connections: nodes in the same module have more connections to each other than to nodes in other modules. By measuring the number of modules, i.e. by obtaining the modularity score of each platform under study we bring a statistically robust measure of the level of audience fragmentation (Newman and Girvan, 2004; Borge-Holthoefer and Gonzalez-Bailon, 2015). Finally, we also apply an Exponential Random Graph Model (ERGM) to account for the influence of audience behavior on the web over the pattern of navigation on Facebook and Twitter. This type of inferential network analysis also allows to control for self-organizing network forces (Lusher, Koskinen and Robins, 2012).

Argument

We expect the patterns of news consumption on the web, Facebook and Twitter will show significant levels of variation. More precisely, we expect audience behavior in Facebook to be more fragmented than that of the web (H1) as a result the algorithmic mechanisms underpinning news consumption on this social platform and previous research in the field, suggesting high levels of audience overlapping on the news web (Webster and Ksiazek, 2012; Mukerjee, Majó-Vázquez and González-Bailón, 2018). Yet, there is also evidence to support the expectation that news consumption on Facebook will be significantly less fragmented than that on Twitter (H2). Therefore, we expect to find more communities emerging from our Twitter networks. The argument for this is twofold: Twitter has less prominent passive content personalization and therefore, what users see on their Twitter feeds is primarily shaped by their following network and hence, more affected by social segregation mechanism (Mutz and Martin, 2001); Furthermore, Twitter publics are more politically interested which we expect to lead to more fragmentation i.e., less audience overlapping, in their news consumption patterns,

References

Adamic, L. A. and Glance, N. (2005) 'The Political Blogosphere and the 2004 U . S . Election : Divided They Blog', pp. 36–43.

Bakshy, E., Messing, S. and Adamic, L. (2015) 'Exposure to ideologically diverse news and opinion on Facebook', Science. American Association for the Advancement of Science, 348(6239), pp. 1130–1132.

Borge-Holthoefer, J. and Gonzalez-Bailon, S. (2015) 'Scale, Time, and Activity Patterns: Advanced Methods for the Analysis of Online Networks', in Fielding, N., Lee, R., and Blank, G. (eds) Handbook of Online Research Methods. Second. Thousand Oaks: Sage Publications. Available at: http://ssrn.com/abstract=2686703.

Delli Carpini, M. X. (2004) 'Mediating Democratic Engagement: The Impact of Communications on Citizens' Involvement in Political and Civic Life', in Kee Kaid, L. (ed.) Handbook of Political Communication Research. Lawrence Erlbaum Publishers, pp. 395–434.

Dvir-Gvirsman, S. (2016) 'Media audience homophily: Partisan websites, audience identity and polarization processes', new media & society. SAGE Publications, p. 1461444815625945.

Fletcher, R. and Nielsen, R. K. (2017) 'Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication', Journal of Communication. Wiley Online Library, 67(4).

Katz, E. (1996) 'And Deliver Us from Segmentation', Annals of the American Academy of Political and Social Science. Sage Publications, Inc. in association with the American Academy of Political and Social Science, 546, pp. 22–33. Available at: http://www.jstor.org/stable/1048167.

Lusher, D., Koskinen, J. and Robins, G. (2012) Exponential random graph models for social networks: Theory, methods, and applications. Cambridge University Press.

Mukerjee, S., Majó-Vázquez, S. and González-Bailón, S. (2018) 'Networks of Audience Overlap in the Consumption of Digital News', Journal of Communication, Forthcommi. Available at: https://ssrn.com/abstract=3032609.

Mutz, D. C. and Martin, P. S. (2001) 'Facilitating communication across lines of political difference: The role of mass media', American Political Science Review, 95(1), pp. 97–114.

Newman, M. E. J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', Physical review E. APS, 69(2), p. 26113. doi: 10.1007/978-3-540-88192-6_64.

Taneja, H. and Webster, J. G. (2016) 'How do global audiences take shape? The role of institutions and culture in patterns of web use', Journal of Communication. Wiley Online Library, 66(1), pp. 161–182.

Webster, J. G. and Ksiazek, T. B. (2012) 'The Dynamics of Audience Fragmentation: Public Attention in an Age of Digital Media', Journal of Communication, 62(1), pp. 39–56. doi: 10.1111/j.1460-2466.2011.01616.x.

Data Journalism Education Globally

Bahareh R. Heravi
School of Information and Communication Studies
University College Dublin
Bahareh.Heravi@ucd.ie

Abstract: This paper explores data journalism education globally through examining a curated dataset of 219 unique modules or programmes on data journalism or related fields[1].

Keyword: data journalism, data-driven journalism, data journalism education, computational journalism, journalism education

Introduction

Data journalism, as we know it today, has been growing for the past 10 years. We are at a stage where many media organisations have data journalists in their newsrooms and organisations and/or are increasingly interested in hiring journalists with data skills. Similarly, academic and educational institutions are making concerted efforts to include data journalism in their programmes, which has led to a surge in the number of data journalism modules and programmes in the past three to five years. This paper studies the state of data journalism higher education globally by curating and analysing a list of 219 unique data journalism modules and programmes offered across the globe.

Method

To understand the state of data journalism education, I compiled a comprehensive global dataset of data journalism modules and programmes, or closely related modules. This dataset was compiled using data collected from a variety of existing data sources (e.g. including Nguyen's dataset), as well as a comprehensive additional search. The final dataset was composed of 219 unique entries, including data journalism programmes and data journalism modules, or closely related modules and is composed of the following fields/variables: 'id', 'title', 'theme', 'organisation', 'school/sub-org', module listing/code, type (full programme or module), level (UG/PG), credit, start year, latest offering, homepage, instructor 1, instructor 1 highest education level, instructor 2, instructor 2 highest education level.

Results

The results of the analysis of 219 data journalism related module show the US has the largest offerings in data journalism related modules and programmes, while in Europe only a scattered number of such modules and programmes exist: 148 of the modules are in the US, 8 are in Canada, which, excluding online courses, leaves only 63 modules and programmes outside of North America offering data journalism related topics altogether. Outside of North America, the UK, the Netherlands, Ireland and Australia are the countries with the highest number of data journalism related modules and programmes. The number of modules per country, where two or more modules/programmes are available in a country, are presented in Table 1.

USA	UK	Canada	Netherlands	Ireland	Australia	Italy
146	12	8	7	6	4	4

Switzerland	China	Germany	Hong Kong	Spain	Greece
4	3	2	2	2	2

Table 1: Number of data journalism related modules and programmes in each country, where more than two modules or programmes are present in a country, excluding the online courses (hence the difference between number of modules taught in text and in the Table). *N=219*.

[1] An extended version of this paper is published in Journalism Practice, doi: 10.1080/17512786.2018.1463167.

.

Overall there are 24 countries present in this dataset. Out of these 24 countries only the US, UK, Ireland, Germany, Canada, Spain and Hong Kong present a strong focus on data journalism as a programme of its own rights, with more than one module dedicated to data journalism, or having postgraduate programmes in data journalism. Out of all European countries listed in Table 1, only the UK (3 universities), Ireland (1 university) and Spain (1 university) present a strong focus on data journalism as a self-contained programme. The rest of the countries in Europe only provide one or two modules in this area.

Amongst the 219 modules and programmes in the dataset, there are only 25 programmes fully and specifically on data journalism. In other words only 25 universities around the world provide degrees or programmes dedicated to data journalism. Despite this, many universities consider data journalism an important topic, and there are 153 instances of stand-alone modules on data journalism in varying non-data journalism focused programmes. The rest are online, vocational or undefined.

In terms of topics taught data shows that Data Journalism and CAR (Computer Assisted Reporting) modules are the most frequently taught data related module in journalism school, forming 65% of all modules listed in the dataset. Such modules have their focus on a complete workflow of data journalism, from finding, collecting and cleaning data, to analysis, visualisation and communication, without diving deep into more complex topics or tools. They are sometimes also a prerequisite for more advanced modules. Modules focusing on data visualisation, coding and computational journalism are the more advanced modules in the offering (Fig. 1).

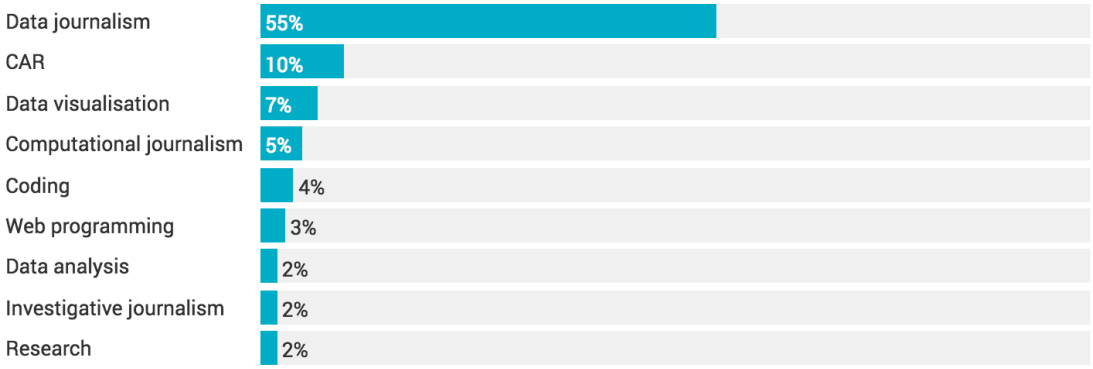

Figure 1. Themes with more than 5 modules in the dataset. *N=219*

In terms of the instructors of these programme, and despite these programmes being university programmes, only the teachers of 58 of these modules and programmes had PhD level education. A mere 49 has Masters level postgraduate degree and 38 were taught with individuals with Bachelors/undergraduate level education. This shows that while many academic/university modules and programmes are emerging in this area, not many instructors have the normal academic credential in this area, which is unlike many other academic disciplines such as social sciences or data and computer science.

Conclusion

While many journalism programmes still do not cover data skills, many have introductory offerings on the topic, and a few have more advanced offerings. Despite this growth, put next to results from the Global Data Journalism survey (Heravi, 2018; Heravi & Lorenz, 2018) - where journalists expressed lack of skills in data analysis skills, interest in learning those skills - data journalism related modules around the world overall fail to provide modules that target immediate needs of data journalists (as identify in Global Data Journalism survey) at a level above very basic data analysis. This finding calls for immediate attention in offering more advanced data and statistical analytics modules in journalism and data journalism programmes around the world.

This study reiterates that the time to include data journalism in journalism programmes has arrived. Data journalism is an interdisciplinary field, and while I expect we will see an increase of data

offerings in journalism programmes in the upcoming years, we require a broadened approach to data journalism training. This approach must be mature enough to facilitate entrants from the various disciplines that converge to create data journalism professionals.

References

Heravi, B. R. & Lorenz, M., 2018. The State of Data Journalism Globally: Practices, Skills, Challenges, Opportunities and Values., forthcoming.

Heravi, B. R., 2018. 'Data Journalism in 2017: A Summary of Results from the Global Data Journalism Survey'. In: Chowdhury, G.; McLeod, J; Gillet, V; Willett, P (eds). Transforming Digital Worlds: iConference 2018, Lecture Notes in Computer Science. Cham, Switzerland: Springer International. , pp.107-113.

You guessed it! Reflecting on preconceptions and exploring data without statistics

Pablo León-Villagrá
University of Edinburgh
pablo.leon@ed.ac.uk

Sarwar Islam
University of Leicester
si113@le.ac.uk

Megan Lucero
The Bureau of Investigative Journalism
meganlucero@tbij.com

Brooks Paige
Alan Turing Institute
bpaige@turing.ac.uk

Tomas Petricek
Alan Turing Institute
tomas@tomasp.net

Abstract: We live in times in which information is abundant, but trust in expert analysis is low. How can we make complex issues accessible for readers and overcome their preconceptions? We propose a novel way of presenting readers with data and raising awareness for individual bias and preconceptions. In our application, readers freely choose potentially relevant factors in societally relevant issues and reflect on their choices in an engaging, non-threatening way. We suggest that such an approach allows for more engaged and open-minded readers and as a result can facilitate democratic, data-centric debate.
Keywords: data-driven storytelling, open data, interactivity, transparency and trust, bias

Introduction

In times of partisan debates and filter bubbles shaping our news, trust in traditional journalism seems to be declining. At the same time, the amount of information and publicly available data is larger than ever before, in principle allowing anybody to gain insight and participate in an informed, democratic debate. Here we explore how journalism can build trust and make data more approachable by implementing data-driven environments for exploration and reflection.

Data-driven documents

Articles often try to simplify complex issues and in an attempt to make the argument approachable present only aggregate statistics to support their argument. As a result, these articles do not allow the readers to check the sources or assumptions made during analysis, which can lead to distrust. Acknowledging these shortcomings, there has been a surge of data-driven reporting tools (see for example Petricek, 2018). These tools can make articles more transparent and open, and encourage readers to scrutinize the analysis and explore further. The reader then can transform from a passive recipient into an active and informed participant in a debate. In this work, we follow the example of data-driven, transparent storytelling but focus more crucially on allowing the reader to engage and reflect on her own beliefs and bias.

Environments for engagement and reflection

One element of a good explanation is it's predictive quality. For example, say we are interested in regional differences in the quality of healthcare and potential sources and factors for better care. A thorough analysis of the matter will require us to examine measures of good care, for instance the time it takes to access care. In general, we will be interested in determining relevant factors that, say, successful regions share, as these are crucial for the evaluation of policy decisions. To find factors we might use domain knowledge to search data sources, potentially transforming or discarding inconclusive variables to derive our piece. In many situations the reader will not be aware of the steps preceding the analysis, nor the reasons for the transformations and as a result could dismiss or distrust our work. Additionally, readers rarely encounter an argument without personal views about what *the right* predictors should be or what the author intends with her piece.

Research has highlighted that correcting bias and engaging readers requires more than just a clear presentation of facts. In fact, corrections can sometimes increase misconceptions (Nyhan & Reifler, 2010). In general, corrections have a greater chance of being effective if they are communicated in non-challenging and appealing way (for an overview see Lewandowsky, Ecker & Cook, 2017). These results can not simply be explained on the basis of a lack of information literacy, as even numerically-savvy readers might choose to maintain their erroneous beliefs to protect their political self identity (Kahan et al., 2017). Given that corrections are difficult and mere presentation of facts can be unconvincing - how do we create an environment that encourages reflection and critical engagement? We propose that to encourage engagement and reflection it is crucial to allow the reader to freely voice and explore their preconceptions. Our application attempts to be a non-

challenging, open, exploration of the issue and the readers assumptions. In what follows we present a first prototype of such an environment.

Application

We chose NHS waiting times (for cancer patients) as our example topic and collected publicly available datasets[2]. Our focus was acquiring a breadth of variables that might be relevant for differences in waiting times instead of finding few, statistically relevant factors. The selection of variables has to allow readers to make their own choices without creating the impression that the document presupposes a particular interpretation. The application consists of three stages: selecting factors, repeatedly predicting a target variable and an overview of one's accuracy.

Selecting factors: After framing the issue, readers are presented with the dataset in form of (potentially predictive) variables. Readers are prompted to select variables that they deem important (see Figure 1). Since absolute numbers will not be informative for non-expert readers, all numerical quantities are expressed verbally (in terms of quantiles, e.g "slightly lower than average").

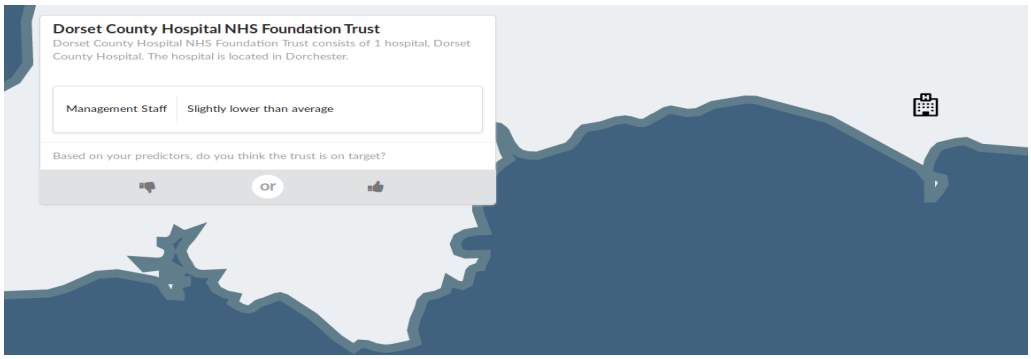

Figure 1. Readers try to predict if the NHS trusts fulfil the waiting time targets. They are given the trust location and general information, as well as the variables they selected. Here the reader chose Managerial Staff as a relevant predictor.

Predicting the target: Readers are shown the variables selected in the previous section for several instances of the target outcome and are asked to predict the target (in our example they had to judge if a particular NHS trust achieved the target for cancer waiting times). For an example, see Figure 2.

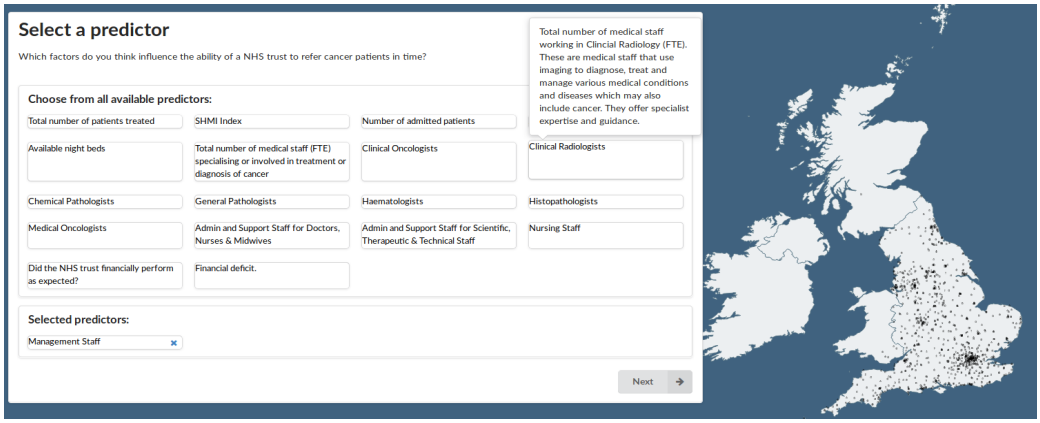

Figure 2. This article explores the ability of NHS trusts to fulfil waiting time targets for cancer patients. The variables the reader could select contained characteristics of NHS trusts (number of patients treated, staff numbers for different staff, day and night beds, financial indicators etc.).

Evaluation: After performing a number of predictions, readers are confronted with their predictive accuracy. This evaluation informs the reader about the predictive quality of their preconceptions. In

2 Waiting times data: https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/. Additional factors used: https://data.gov.uk/dataset/nhs-workforce-medical-staff and https://digital.nhs.uk/catalogue/PUB30092.

the future we plan to offer additional feedback at this stage - for example hints recommending features with stronger predictive accuracy or a comparison to other readers' performance. This evaluation and feedback can also be used as introduction to a traditional journalistic piece.

Conclusions

In this work we developed a prototype that allows readers express their preconceptions and reflect on their predictive accuracy. Our application is designed to be neutral - it does not attempt to instil an expert opinion, but instead tries to raise awareness for preconceptions and complex relationships. We suggest that reflection on preconceptions and awareness of complexity are fundamental for more engaged and open-minded readers and can facilitate more democratic debate. In the future we would like to test the efficiency of our application and explore further ways in which preconceptions and reader engagement can be incorporated into reporting.

References

Lewandowsky, S., Ecker, U.K. and Cook, J., 2017. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. Journal of Applied Research in Memory and Cognition, 6(4), pp.353-369.

Kahan, D.M., Peters, E., Dawson, E.C. and Slovic, P., 2017. Motivated numeracy and enlightened self-government. Behavioural Public Policy, 1(1), pp.54-86.

Nyhan, B. and Reifler, J., 2010. When corrections fail: The persistence of political misperceptions. Political Behavior, 32(2), pp.303-330.

Petricek, Tomas. "Tools for Open, Transparent and Engaging Storytelling." Proceedings of the International Conference on the Art, Science, and Engineering of Programming - Programming '17, 201?

Stories of storytelling about UK's EU funding

Mariana Marasoiu	Sarwar Islam	Luke Church	Megan Lucero	Brooks Paige	Tomas Petricek
University of Cambridge	University of Leicester	University of Cambridge	The Bureau of Investigative Journalism	Alan Turing Institute	Alan Turing Institute
mcm79@cam.ac.uk	si113@le.ac.uk	luke@church.name	meganlucero@tbij.com	bpaige@turing.ac.uk	tomas@tomasp.net

Abstract: In the context of open data, the analyst of that data is removed from the collection process. This can make data analysis extremely difficult, even impossible if the data needed to answer their questions has not been collected. We present a study aimed at exploring the difficulties of open data analysis using, as an example, the data available on EU funding to the UK. We report on some of the fundamental difficulties we observed whilst analysing this data and we suggest that by building a catalog of such difficulties we can explain the limitations of working with open data to the wider public and data publishers. Finally, we propose a methodological transition in how data analysis is viewed as part of a wider process.

Keywords: Open data, data journalism, autoethnography, data analysis

Introduction

Open data is increasingly becoming one of many tools that journalists use in their investigations. However, the process of analysing this data is challenging, from lack of appropriate and end-user accessible tools (Davies, 2010), to issues with the data itself, such as formatting, geo tagging or which data is being collected (Gurstein, 2011).

We describe several difficulties encountered whilst analysing the funding data on European Structural and Investment Funds (ESIF) to the UK. The dataset was identified by The Bureau Local3, a team of data journalists working with a large network of citizens and reporters across the UK. The Bureau was interested in tracing EU funding to community-level in order to support other local journalists wanting to report on the impact of Brexit in their area. Whilst a fairly specific example, it is a good illustration of the challenges of working with open data.

Methods

To investigate the difficulties of analysing the EU funding data, we conducted an autoethnographically-inspired study, recording each step of the analysis process. Autoethnography is a qualitative research method involving self-observation and self-reflection in which the author relates their thoughts, experience and behaviour to the wider social life, cultural belief system and practices of the ethnographic setting (Marechal, 2010). In the context of human-computer interaction (HCI), autoethnography has been used for requirements elicitation (Cunningham and Jones, 2005), for informing design (Neustaedter and Sengers, 2012) or for identifying design challenges and opportunities (Fernando et al., 2016). Since our goal was to understand the challenges of analysing a dataset, the results of our autoethnography are analytical rather than descriptive — we discuss these in the next section.

Our study also draws on more typical inspection and task analysis methods in HCI research, such as Cognitive Walkthrough (Polson et al., 1992) and task inspection. We kept two detailed diaries of our thoughts, experiences, actions and each low-level interaction with the tools used (in our case, Microsoft Excel and STATA) for analysing the EU subsidies data over several weeks. This exploration was directed towards specific goals, typical of what a local journalist may be interested in: i) analysing funding for apprenticeships in Middlesbrough and ii) analysing funding for skills before employment in Liverpool. Due to the limited time available for the study, we only covered downloading the data, formatting it, an initial analysis and an attempt to fill in some of the missing information.

[3] https://www.thebureauinvestigates.com/projects/the-bureau-local

Findings

The diary documenting the analysis on apprenticeships in Middlesbrough contained 98 slides with text and screen- shots (see Figure 1) and usage descriptions of 5 different tools and 16 websites. The second diary documenting the analysis on skills before employment in Liverpool extended over 11 A4 pages (8pt text and screenshots), primarily describing interaction with STATA, a statistics package, but also with PDFs and several websites.

We categorized the diary data through thematic coding (Gibbs, 2007), identifying difficulties across two dimensions: interface-related and data-related. The interface-related issues of the tools we used can be described by existing usability frameworks (e.g. Blackwell, in press; Green and Petre, 1996). The data-related issues were of two levels: concrete (e.g. missing data, file formats) and abstract. We focus here on such three, particularly common, abstract issues.

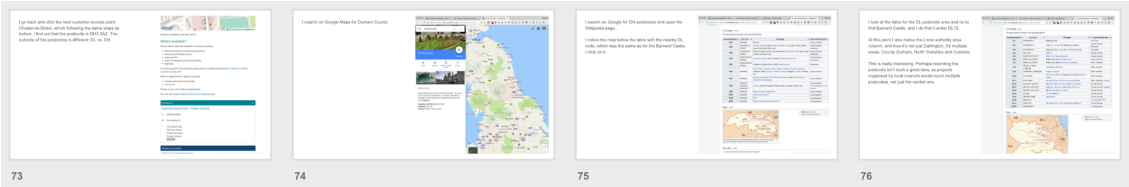

Figure 1. Four consecutive diary slides illustrating the granularity of the observation

Data/Question schema mismatch

Description: The current schema of the data is not suitable for answering the questions about the data. This may mean that the data will need to be restructured, or the level of detail needs to be changed: through aggregation if the granularity level is low, or collecting more fine-grained data if the granularity is too high.

Example: The funding data is organised by geographical regions and lists the name of each organisation that has received funds, the amount of funding and the contract period. Since there is no lower granularity geographical data, the analysis about local distribution of funds (e.g. county-level) cannot be done without collecting new data.

Entities live in multiple hierarchies

Description: What was assumed to fit within one hierarchy now needs to be split into two or more hierarchies.

Example: The address of the fund beneficiary can be different to the area affected by the funds, as organizations based in some part of the country can receive funding for doing work in another part of the country. What was initially a single category will need to be split into two categories “beneficiary address” and “benefiting area”.

Messy categories

Description: The categories are not clear-cut, e.g. the same type of information can be at different levels of detail.

Example: The size of the area that benefits from the funds varies widely, from individual addresses, to one or multiple counties, to entire regions. Recording this in a form that can be analysed is challenging.

Conclusions

Even though our analysis was restricted to data on EU funding to the UK, we believe that our findings can also be applied more broadly for explaining some of challenges of working with open data. For example, when the data analysts are removed from the process of data collection and publication (typical of open data), data/question schema mismatch is a relevant issue, resulting in the analysts needing to do new collection work themselves. Beyond the examples given, problems with multiple

hierarchies and messy categories can arise when merging multiple datasets, another typical task in data analysis. In the worst case, this results in manual labelling of all the data points. These observations suggest that there is a need for anticipating the kinds of questions and analyses the wider public would want to ask of open data. In some cases, the solution could be a more iterative, cyclical process of data collection, publication and analysis followed by refined collection etc., with feedback channels between the different actors. This is analogous to the transition from the waterfall process of software development to Agile methodologies that now dominate industry practice.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Blackwell, A.F. (in press). "A pattern language for the design of diagrams", in Richards, C. (Ed.), *Elements of Diagramming*.

Cunningham, S.J. and Jones, M. (2005), "Autoethnography: A Tool for Practice and Education", *Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Making CHI Natural*, ACM, New York, NY, USA, pp. 1–8.

Davies, T. (2010), *Open Data, Democracy and Public Sector Reform: A Look at Open Government Data Use from Data.gov.uk*, (unpublished Master's thesis), University of Oxford.

Fernando, P., Pandelakis, M. and Kuznetsov, S. (2016), "Practicing DIYBiology In An HCI Setting", *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, pp. 2064–2071.

Gibbs, G. (2007), "Thematic coding and categorizing", in Gibbs, G. (Ed.), *Qualitative Research Kit: Analyzing Qualitative Data*, SAGE Publications, Ltd, London, England, pp. 38–55.

Green, T.R.G. and Petre, M. (1996), "Usability analysis of visual programming environments: a 'cognitive dimensions' framework", *Journal of Visual Languages & Computing*, Elsevier, Vol. 7 No. 2, pp. 131–174.

Gurstein, M.B. (2011), "Open data: Empowering the empowered or effective data use for everyone?", *First Monday*, Vol. 16 No. 2.

Marechal, G. (2010), "Autoethnography", in Mills, A., Durepos, G. and Wiebe, E. (Eds.), *Encyclopedia of Case Study Research*, SAGE Publications, Vol. 2, pp. 43–45.

Neustaedter, C. and Sengers, P. (2012), "Autobiographical Design in HCI Research: Designing and Learning Through Use-it-yourself", *Proceedings of the Designing Interactive Systems Conference*, ACM, New York, NY, USA, pp. 514–523.

Polson, P.G., Lewis, C., Rieman, J. and Wharton, C. (1992), "Cognitive walkthroughs: a method for theory-based evaluation of user interfaces".

European Data and Computational Journalism Conference
Programme Committee 2018

- James Hamilton (Stanford University, U.S.)
- Nicholas Diakopoulos (University of Maryland, U.S.)
- Meredith Broussard (New York University, U.S.)
- Cheryl Phillips (Stanford University, U.S.)
- Eddy Borges Rey (University of Stirling, Scotland, U.K.)
- Marc Esteve del Valle (University of Groningen, The Netherlands)
- Miranda McLachlan (Goldsmiths, University of London, U.K.)
- Stefano Ceccon (The Times and The Sunday Times, U.K.)
- Jonathan Gray (University of Bath, U.K.)
- Paul Bradshaw (Birmingham City University, U.K.)
- Bahareh Heravi (University College Dublin, Ireland)
- Glyn Mottershead (Cardiff University, Wales, U.K.)
- Martin Chorley (Cardiff University, Wales, U.K.)

Thanks to all the committee for their timely and informative reviews!