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.

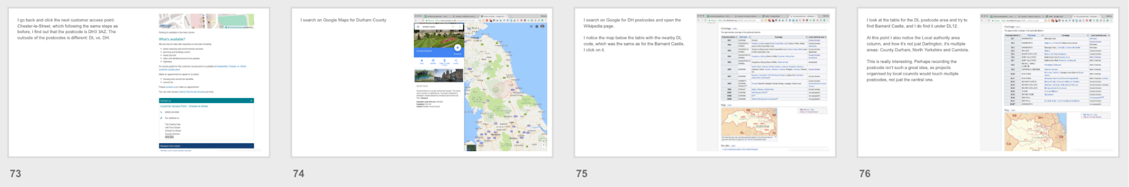
Our study also draws on more typical inspection and task analysis methods in HCI research, such as Cognitive Walkthrough (Polson et al., 1992) and task inspection. We kept two detailed diaries of our thoughts, experiences, actions and each low-level interaction with the tools used (in our case, Microsoft 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.


I go back and click the next customer access point: Chester-le-Street, which following the same steps as before, I find out that the postcode is DH3 3AZ. The outcode of the postcodes is different: DL vs. DH.

I search on Google Maps for Durham County

I search on Google for DH postcodes and open the Wikipedia page.

I notice the map below the table with the nearby DL code, which was the same as for the Barnard Castle. I click on it.

I look at the table for the DL postcode area and try to find Barnard Castle, and I do find it under DL12.

At this point I also notice the Local authority area column, and how it's not just Darlington, it's multiple areas: County Durham, North Yorkshire and Cumbria.

This is really interesting. Perhaps recording the postcode isn't such a great idea, as projects organised by local councils would touch multiple postcodes, not just the central one.

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”.