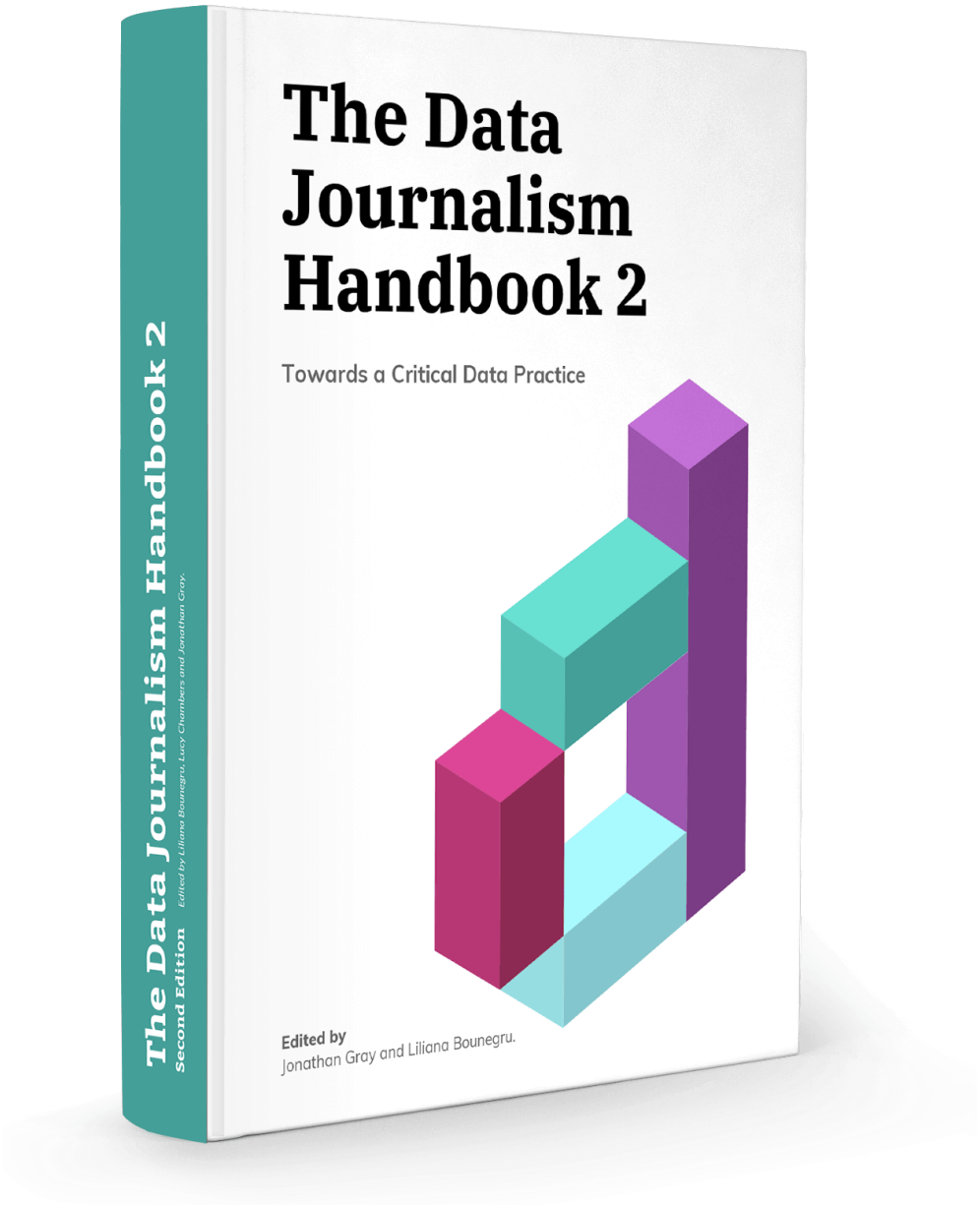
The Data Journalism Handbook 2



Chapters

60. Data Journalism with Impact
61. Beyond Clicks and Shares: How and Why to Measure the Impact of Data Journalism Projects
62. The Economics of Data Journalism
63. The Datafication of Journalism
64. Forms of Data Journalism
65. Data Journalism and its Publics
66. Data Journalism: In Whose Interests?
67. Indigenous Data Sovereignty
68. What is Data Journalism For? Cash, Clicks, and Cut and Trys
69. Statisticians and Journalists: Tales of Two Professions
70. Data Journalism and Digital Liberalism
71. Afterword: Data Journalism and Experiments in Reporting

Introduction

Written by: Jonathan Gray Liliana Bounegru

Data journalism in question

What is data journalism? What is it for? What might it do? What opportunities and limitations does it present? Who and what is involved in making and making sense of it? This book is a collaborative experiment responding to these and other questions. It follows on from another edited book, *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News* (O'Reilly Media, 2012).[1] Both books assemble a plurality of voices and perspectives to account for the evolving field of data journalism. The first edition started through a "book sprint" at MozFest in London in 2011, which brought together journalists, technologists, advocacy groups and others in order to write about how data journalism is done. As we wrote in the introduction, it aimed to "document the passion and enthusiasm, the vision and energy of a nascent movement", to provide "stories behind the stories" and to let "different voices and views shine through". The 2012 edition is now translated into over a dozen languages – including Arabic, Chinese, Czech, French, Georgian, Greek, Italian, Macedonian, Portuguese, Russian, Spanish and Ukrainian – and is used for teaching at many leading universities, as well as teaching and training centres around the world, as well as being a well-cited source for researchers studying the field.

While the 2012 book is still widely used (and this book is intended to complement rather than to replace it), a great deal has happened since 2012. On the one hand, data journalism has become more established. In 2011 data journalism as such was very much a field "in the making", with only a handful of people using the term. It has subsequently become socialised and institutionalised through dedicated organisations, training courses, job posts, professional teams, awards, anthologies, journal articles, reports, tools, online communities, hashtags, conferences, networks, meetups, mailing lists and more. There is also broader awareness of the term through events which are conspicuously data-related, such as the Panama Papers, which whistleblower Edward Snowden then characterised as the "biggest leak in the history of data journalism".

On the other hand, data journalism has become more contested. The 2013 Snowden leaks helped to establish a transnational surveillance apparatus of states and technology companies as a matter of fact rather than speculation. These leaks suggested how citizens were made knowable through big data practices, showing a darker side to familiar data-making devices, apps and platforms.[2] In the US the launch of Nate Silver's dedicated data journalism outlet FiveThirtyEight in 2014 was greeted by a backlash for its over-confidence in particular kinds of quantitative methods and its disdain for "opinion journalism".[3] While Silver was acclaimed as "lord and god of the algorithm" by *The Daily Show*'s Jon Stewart for successfully predicting the outcome of the 2012 elections, the statistical methods that he advocated were further critiqued and challenged after the election of Donald Trump in 2016. These elections along with the Brexit vote in the UK and the rise of populist right-wing leaders around the world, were said to correspond with a "post-truth" moment, characterised by a widespread loss of faith in public institutions, expert knowledge and the facts associated with them, and the mediation of public and political life by online platforms which left their users vulnerable to targeting, manipulation and misinformation.

Whether this "post-truth" moment is taken as evidence of failure or as a call to action, one thing is clear: data can no longer be taken for granted, and nor can data journalism. Data does not just provide neutral and straightforward representations of the world, but is rather entangled with politics and culture, money and power. Institutions and infrastructures underpinning the production of data – from surveys to statistics, climate science to social media platforms – have been called into question. Thus it might be asked: Which data, whose data and by

which means? Data about which issues and to what end? Which kinds of issues are data-rich and which are data-poor? Who has the capacities to benefit from it? What kinds of publics does data assemble, which kinds of capacities does it support, what kinds of politics does it enact and what kinds of participation does it engender?

Towards a critical data practice

Rather than bracketing such questions and concerns, this book aims to "stay with the trouble" as the prominent feminist scholar Donna Haraway puts it.[4] Instead of treating the relevance and importance of data journalism as an assertion, we treat this as a question which can be addressed in multiple ways. The collection of chapters gathered in the book aim to provide a richer story about what data journalism does, with and for whom. Through our editorial work we have encouraged both reflection and a kind of modesty in articulating what data journalism projects can do, and the conditions under which they can succeed. This entails the cultivation of a different kind of precision in accounting for data journalism practice: specifying the situations in which it develops and operates. Such precision requires broadening the scope of the book to include not just the ways in which data is analysed, created and used in the context of journalism but also more about the social, cultural, political and economic circumstances in which such practices are embedded.

The subtitle of this new book is "towards a critical data practice", and reflects both our aspiration as editors to bring critical reflection to bear on data journalism practices, as well as reflecting the increasingly critical stances of data journalism practitioners. The notion of "critical data practice" is a nod to Philip E. Agre's notion of "critical technical practice", which he describes in terms of having "one foot planted in the craft work of design and the other foot planted in the reflexive work of critique".[5] As we have written about elsewhere, our interest in this book is understanding how critical engagements with data might modify data practices, making space for public imagination and interventions around data politics.[6]

Alongside contributions from data journalists and practitioners writing about what they do, the book also includes chapters from researchers whose work may advance critical reflection on data journalism practices, from fields such as anthropology, science and technology studies, (new) media studies, internet studies, platform studies, the sociology of quantification, journalism studies, indigenous studies, feminist studies, digital methods and digital sociology. Rather than assume a more traditional division of labour such that researchers provide critical reflection and practitioners offer more instrumental tips and advice, we have sought to encourage researchers to consider the practical salience of their work, and to provide practitioners with space to reflect on what they do outside of their day-to-day deadlines. None of these different perspectives exhaust the field, and our objective is to encourage readers to attend to the different aspects of how data journalism is done. In other words, this book is intended to function as an multidisciplinary conversation starter, and – we hope – a catalyst for collaborations.

We do not assume that "data journalism" refers to a unified set of practices. Rather it is a prominent label which refers to a diverse set of practices which can be empirically studied, specified and experimented with. As one recent review puts it, we need to interrogate the "how of quantification as much as the mere fact of it", the effects of which "depend on intentions and implementation".[7] Our purpose is not to stabilise how data journalism is done, but rather to draw attention to its manifold aspects and open up space for doing it differently.

A collective experiment

It is worth briefly noting what this book is not. It is not just a textbook or handbook in the conventional sense: the chapters don't add up to an established body of knowledge, but are rather intended to indicate interesting directions for further inquiry and experimentation. The book is not just a practical guidebook of tutorials or "how tos": there are already countless readily available materials and courses on different aspects of data practice (e.g. data analysis and data visualisation). It is not just a book of "behind the scenes" case studies: there are plenty of articles and blog posts showing how projects were done, including interviews with their creators. It is not just a book of recent academic perspectives: there is an emerging body of literature on data journalism scattered across numerous books and journals.[8]

Rather the book has been designed as a *collective experiment* in accounting for data journalism practices and a *collective invitation* to explore how such practices may be modified. It is collective in that, as with the first edition, we have been able to assemble a comparatively large number of contributors (more than seventy) for a short book, and the editorial process has benefitted from recommendations from contributors. Through what could be considered a kind of curated "snowball editorial", we have sought to follow how data journalism is done by different actors, in different places, around different topics, through different means. Through the process we have trawled through many shortlists, longlists, outlets and datasets to curate different perspectives on data journalism practices. Though there were many, many more contributors we would have liked to include, we had to operate within the constraints of a printable book, as well as giving voice to a balance of genders, geographies and themes.

It is experimental in that the chapters provide different perspectives and provocations on data journalism, which we invite readers to further explore through actively configuring their own blends of tools, datasets, methods, texts, publics and issues. Rather than inheriting the ways of seeing and ways of knowing that have been "baked into" elements such as official datasets or social media data, we encourage readers to enrol them into the service of their own lines of inquiry. This follows the spirit of "critical analytics" and "inventive methods" which aim to modify the questions which are asked and the way problems are framed.[9] Data journalism can be viewed not just in terms of how things are *represented*, but in terms of how it organises *relations* – such that it is not just a matter of producing data stories (through collecting, analysing, visualising and narrating data), but also attending to who and what these stories bring together (including audiences, sources, methods, institutions and social media platforms). Thus we may ask, as Noortje Marres recently put it: "What are the methods, materials, techniques and arrangements that we curate in order to create spaces where problems can be addressed differently?". The chapters in this book show how data journalism can be an inventive, imaginative, collaborative craft, highlighting how data journalists interrogate official data sources, make and compile their own data, try new visual and interactive formats, reflect on the effects of their work and make their methods accountable and code re-usable.

The online beta of the book is intended to provide an opportunity to publicly preview a selection of chapters before the printed version of the book is published. We hope this process will elicit comments and encounters (and perhaps testing out in contexts of teaching and training) before the book takes its final shape. If the future of data journalism is uncertain, then we hope that readers of this book will join us in both critically taking stock of what journalism is and has been, as well as intervening to shape its future.

An overview of the book

To stay true to our editorial emphasis on specifying the setting, we note that the orientation of the book and its selection of chapters is coloured by our interests and those of our friends, colleagues and networks at this particular moment – including growing concerns about climate change, environmental destruction, air pollution, tax avoidance, (neo)colonialism, racism, sexism, inequality, extractivism, authoritarianism, algorithmic injustice and platform labour. The chapters explore how data journalism makes such issues intelligible and experienceable, as well as the kinds of responses it can mobilise. The selection of chapters also reflects our own oscillations between academic research, journalism and advocacy, as well as the different styles of writing and data practice associated with each of these. We remain convinced of the generative potential of encounters between colleagues in these different fields, and several of the chapters attest to successful cross-field collaborations.

After the introduction, the book starts with a "taster menu" on doing issues with data. This includes a variety of different formats for making sense of different themes in different places – including looking at the people and scenes behind the numbers for home demolitions in occupied East Jerusalem (Haddad), multiplying memories of trees in Bogota (Magaña), tracing connections between agricultural commodities, crime, corruption and colonialism across several countries (Sánchez and Villagrán), investigating extractive industries in Peru (Salazar), mobilising for road safety in the Philippines (Rey and Mendoza), putting carbon emissions into context (Clark), engaging

publics with data graphics on Instagram (Alaali), counting transgender lives (Talusan), and mapping segregation in the US (Williams). The chapters in this section illustrate a breadth of practices from visualisation techniques to building campaigns to engaging audiences around data on Instagram.

The third section focuses on how journalists assemble data, including projects on themes such as land conflicts (Shrivastava and Paliwal), air pollution (Naik and Salve) and knife crime (Barr). It also includes accounts of how to obtain and work with data in countries where it may be less easy to come by, such as in Cuba (Carmona et al) and China (Ma). Assembling data may also be a way of engaging with readers (Coelho) and assembling interested actors around an issue, which may in itself constitute an important outcome of a project. Gathering data may involve gradually and creatively piecing together fragments of information from disparate sources, including documents, interviews and investigative fieldwork (Boros). As well as using data, other types of stories may be surfaced by exploring how numbers are made (Verran).

The fourth section is concerned with different ways of working with data. This includes with graph databases (Haddou), algorithms (Stray), code (Simon) and varieties of digital and computational methods (Zhang; Rey). Contributors examine emerging issues and opportunities arising from working with sources such as text data (Maseda) and data from the web, social media and other online devices (Weltevrede). Others look at practices for making data journalistic work transparent, accountable and reproducible (Leon; Mazotte). Databases may also afford opportunities for collaborative work on large investigative projects (Díaz-Struck and Romera). Feminist thought and practice may also inspire different ways of working with data (D'Ignazio).

The fifth section is dedicated to examining different ways in which data can be experienced. Several pieces reflect on contemporary visualisation practices (Aisch and Rost; Stabe), as well as how readers respond to and participate in making sense with visualisations (Kennedy et al). Other pieces look at how data is mediated and presented to readers through databases (Rahman and Wehrmeyer), web based interactives (Bentley), TV and radio (de Jong) and comics (Amancio).

The sixth section is dedicated to emerging approaches for investigating data, platforms and algorithms. The digital is taken as a site of investigation, as highlighted by BuzzFeed News projects on viral content, misinformation and digital culture (Silverman). Chapters in this section examine different ways of reporting on algorithms (Diakopoulous), as well as how to conduct longer term collaborations in this area (Elmer). Several chapters look at how to work with social media data to explore how platforms participate in shaping debate, including storytelling approaches (Vo) and repurposing data to see how platforms and data industries see humans (Lavigne). A final chapter explores affinities between digital methods research and data journalism, including how data can be used to tell stories about web tracking infrastructures (Rogers).

The seventh section is on organising data journalism, and attends to different types of work in the field which is considered indispensable but not always prominently recognised. This includes the changing role of data journalism in newsrooms (Pilhofer; Klein); how data journalism has changed over the past decade (Rogers); how platforms and the gig economy shape cross-border investigative networks (Candea); entanglements between data journalism and movements for open data and civic tech (Baack); open source coding practices (Pitts); data journalism and gender (Vaca); audience measurement practices (Petre); archiving data journalism (Broussard); organising transnational collaborations (Ottaviani and Govindasamy); and the role of the #ddj hashtag in connecting data journalism communities on Twitter (Au and Smith).

The eighth section looks at training data journalists and the development of data journalism around the world. This includes chapters on teaching data journalism at universities in the US (Phillips); hackathons and bootcamps in Central Asia (Valeeva); and MOOCs and local training initiatives in Turkey (Dag). Others argue for the importance

of empowering marginalised communities to tell their stories (Constantaras), and caution against "digital universalism" and underestimating innovation in the "periphery" (Chan).

Data journalism does not happen in a vacuum and the ninth section surfaces its various social, political, cultural and economic settings. A chapter on the genealogies of data journalism in the United States serves to encourage reflection on the various historical practices and ideas which shape it (Anderson). Other chapters look at the economics and sustainability of data journalism (Steiger); data journalism as a response to broader societal processes of datafication (Lewis and Radcliffe); different forms and formats of data journalism (Cohen); the publics that data journalism assembles (Parasie); and how data journalism projects are valued through awards (Loosen). Two chapters reflect on different approaches to measuring the impact of data journalism projects (Bradshaw; Green-Barber). Others examine issues around data journalism and colonialism (Young) and indigenous data sovereignty (Kukutai and Walter).

The tenth and final section closes with reflections, challenges and possible future directions for the field. This includes chapters on opportunities and pitfalls of knowing society through data (Didier); data journalism and digital liberalism (Boyer); and whether data journalism can live up to its earlier aspirations to become a field of inspired experimentation, interactivity and play (Usher). An afterword from Noortje Marres reflects on data journalism as a form of reporting from the perspective of digital sociology.

Twelve challenges for critical data practice

Drawing on the time that we have spent exploring the field of data journalism through the development of this book, we would like to provide twelve challenges for "critical data practice". These consider data journalism in terms of its capacities to *shape relations* between different actors as well as to *produce representations* about the world.

1. How can data journalism projects account for the *collective character of digital data, platforms, algorithms and online devices*, including the interplay between digital technologies and digital cultures?
2. How can data journalism projects tell stories about big issues at scale (e.g. climate change, inequality, multinational taxation, migration) while also *affirming the provisionality* and *acknowledging the models, assumptions and uncertainty* involved in the production of numbers?
3. How can data journalism projects tell stories *both with and about data* including the various actors, processes, institutions, infrastructures and forms of knowledge through which data is made?
4. How can data journalism projects *cultivate their own ways of making things intelligible, meaningful and relatable through data*, without simply uncritically advancing the ways of knowing "baked into" data from dominant institutions, infrastructures and practices?
5. How can data journalism projects *acknowledge and experiment with the visual cultures and aesthetics that they draw on*, including through combinations of data visualisations and other visual materials?
6. How can data journalism projects make space for *public participation and intervention* in interrogating established data sources and re-imagining which issues are accounted for through data, and how?
7. How might data journalists cultivate and consciously affirm *their own styles of working with data*, which may draw on, yet remain distinct from fields such as statistics, data science and social media analytics?
8. How can the field of data journalism *develop memory practices to archive and preserve their work*, as well as situating it in relation to practices and cultures that they draw on?
9. How can data journalism projects collaborate around transnational issues in ways which *avoid the logic of the platform and the colony, and affirm innovations at the periphery*?
10. How can data journalism support marginalised communities to use data to *tell their own stories on their own terms*, rather than telling their stories for them?
11. How can data journalism projects develop their own *alternative and inventive ways of accounting for their value and impact in the world*, beyond social media metrics and impact methodologies established in other

fields?

12. How might data journalism *develop a style of objectivity which affirms, rather than minimises, its own role in intervening in the world* and in shaping relations between different actors in collective life?

Words of thanks

We are most grateful to Amsterdam University Press for being so supportive with this experimental project, including the publication of an online beta as well as their support for an open access digital version of the book. It is perhaps also an apt choice, given that several of the contributors convened at one of the first European conferences on data journalism which took place in Amsterdam in 2010. Open access funding is supported by a grant from the Netherlands Organisation for Scientific Research (NWO, 324-98-014).

The vision for the book was germinated through discussions with friends and colleagues associated with the Public Data Lab. We particularly benefited from conversations about the book with Andreas Birkbak, Erik Borra, Noortje Marres, Richard Rogers, Tommaso Venturini and Esther Weltevrede. We were also provided with space to develop the direction of this book through events and visits to Columbia University (in discussion with Bruno Latour); Utrecht University; the University of California Berkeley; Stanford University; the University of Amsterdam; the University of Miami; Aalborg University Copenhagen; Sciences Po, Paris; the University of Cambridge; London School of Economics; Cardiff University; Lancaster University and the International Journalism Festival in Perugia. Graduate students taking the MA course in data journalism at King's College London helped us to test the notion of "critical data practice" which lies at the heart of this book.

Our longstanding hope to do another edition was both nurtured and materialised thanks to Rina Tsubaki, who helped to gather support from the European Journalism Centre and the Google News Initiative. We are grateful to Adam Thomas, Bianca Lemmens, Biba Klomp, Letizia Gambini, Arne Grauls and Simon Rogers for providing us with both editorial independence and enduring support to scale up our efforts. The editorial assistance of Daniela Demarchi has been tremendously valuable in helping us to chart a clear course through sprawling currents of texts, footnotes, references emails and spreadsheets.

Most of all, we would like to thank all of the data journalism practitioners and researchers who were involved in the project (whether through writing, correspondence or discussion) for accompanying us, and for supporting this experiment with their contributions of time, energy, materials and ideas without which the project would not have been possible. This book is, and continues to be, a collective undertaking.

Works cited

Philip E. Agre, 'Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI', in *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*, ed. by Geoffrey Bowker, Susan Leigh Star, Bill Turner, and Les Gasser (Mahwah, NJ: Erlbaum, 1997), pp. 130–57.

Elizabeth Popp Berman and Daniel Hirschman, 'The Sociology of Quantification: Where Are We Now?':, *Contemporary Sociology*, 2018.

Dylan Byers. 'Knives out for Nate Silver'. *Politico*. 2014.

Jonathan Gray, Liliana Bounegru and Lucy Chambers, *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News*, O'Reilly Media, 2012.

Jonathan Gray, 'Three Aspects of Data Worlds', *Krisis: Journal for Contemporary Philosophy*, 2018: 1.

Jonathan Gray and Liliana Bounegru, ‘What a Difference a Dataset Makes? Data Journalism And/As Data Activism’. In *Data in Society: Challenging Statistics in an Age of Globalisation*, J. Evans, S. Ruane and H. Southall (eds). Bristol: The Policy Press, 2019.

Jonathan Gray, Carolin Gerlitz and Liliana Bounegru, ‘Data infrastructure literacy’, *Big Data & Society*, 5(2), 2018, pp. 1–13.

Donna J. Haraway. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press. 2016.

Celia Lury and Nina Wakeford, eds., *Inventive Methods: The Happening of the Social* (London: Routledge, 2012).

Richard Rogers, ‘Otherwise Engaged: Social Media from Vanity Metrics to Critical Analytics’, *International Journal of Communication*, 12 (2018), 450–72.

Behind the Numbers: Home Demolitions in Occupied East Jerusalem

When you look at the chart below (Figure 1), you will see a series of steady orange and black bars followed by a large spike in 2016. Once you take a closer look at the caption you'll understand that this chart shows the number of structures destroyed and people affected by Israel's policy of home demolitions.

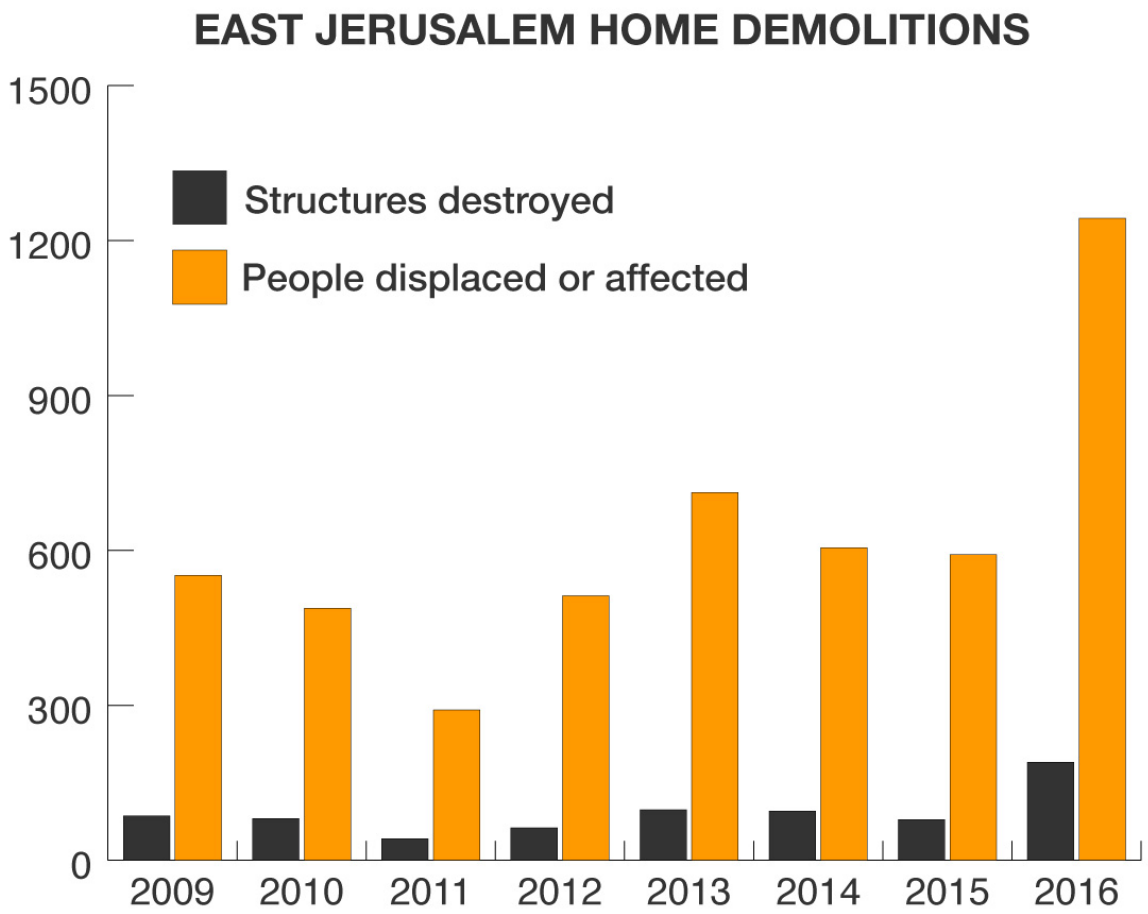


Figure 1: Al Jazeera graph showing East Jerusalem home demolitions, 2009-2016.

As Nathan Yau, author of Flowing Data, put it "data is an abstraction of real life". Each number represents a family, and each number tells a story.

Broken Homes is the most comprehensive project to date tracking home demolitions in East Jerusalem, a Palestinian neighbourhood that has been occupied by Israel for 50 years[1].

Working closely with the United Nations, Al Jazeera tracked every single home demolition in East Jerusalem in 2016. It turned out to be a record year, with 190 structures destroyed and more than 1,200 Palestinians displaced or affected.

We decided to tackle this project after witnessing an escalation in violence between Israelis and Palestinians in late 2015. The goal was twofold: to understand how Israel's home demolitions policy would be affected by the increased tensions, and to tell readers the human stories behind the data.

The project reveals the impact on Palestinian families through video testimony, 360-degree photos and an interactive map that highlights the location, frequency and impact of each demolition.

Our producer in Doha began coordinating with the UN in late 2015 to develop a framework for the project. The UN routinely gathers data on home demolitions, and while some of it is available online, other aspects - including GPS coordinates - are only recorded internally. We wanted to be able to show every demolition site on a map, so we began obtaining monthly data sets from the UN. For each incident, we included the demolition date, number of people and structures affected, a brief description of what happened, and a point on our East Jerusalem map showing the location. We cross-checked these with news reports and other local information about home demolitions. We then selected a case to highlight each month, as a way of showing different facets of the Israeli policy - from punitive to administrative demolitions, affecting everyone from young children to elderly residents.

Our reporter on the ground travelled throughout East Jerusalem over the course of the year to speak with many of the affected families, in order to explore their losses in greater depth and to photograph and record the demolition sites.

There was a broad range of responses from the affected families. The interviews had to take place in the physical location of the demolition, which could be a difficult experience for those affected, so sensitivity and patience was required at all stages, from setting up the meetings to recording the material.

On the whole, the families responded well to the project. They were very generous with their time and in sharing their experiences. In one instance, a man had written down a list of things he wanted to say to us. In another case, it took a few attempts to convince the family to take part. One family declined to meet with us and so we had to liaise with the UN and find another family willing to speak about their home demolition.

Image 2: Panoramic photograph of home demolished in May 2016.

Many news organisations, including Al Jazeera, have reported on individual home demolitions over the years. One of the main reasons for taking a data-driven approach this time was to clearly contextualise the scale of the story by counting each and every demolition. This context and fresh perspective is especially important when reporting on an ongoing topic to keep readers engaged.

A word of advice for aspiring data journalists: taking a data-driven approach to a story doesn't need to be technical or expensive. Sometimes just following and counting occurrences of an event over time is enough to tell you a lot about the scale of a problem. As long as your data gathering methodology remains consistent, there are many stories that you can tell using data that you might not otherwise report on. Also, be patient. We gathered data for an entire year to tell this story. The most important thing is to thoroughly storyboard exactly what data you need before sending any reporters out into the field. Most of the time you won't need any special equipment either. We used an iPhone to take all the 360 degree images and capture the specific GPS coordinates.

The project - released in January 2017 in English, Arabic and Bosnian - presents a grim warning about what lies ahead as Israel continues to deny building permits to 98 percent of Palestinian applicants, ramping up the pressure on a large and growing population.

Works Cited

Megan O'Toole et al., 'Broken Homes: A Record Year of Home Demolitions in Occupied East Jerusalem', *Al Jazeera*, 2017.

Multiplying Memories While Discovering Trees in Bogota

Written by: Maria Isabel Magaña

Bogotá holds almost 16% of the population of Colombia in just 1.775 km². You get the idea, it's crowded, it's furious. But it's also a green city, surrounded by mountains and many different trees planted. Most of the times, trees go unnoticed by its citizens in the midst of their daily life. Or at least that's what happened to the members of our data team except for one of our coders, who loves trees and can't walk down the street without noticing them. She knows all the species and the facts about them. Her love for nature in the midst of the chaos of the city is what got us thinking: has anybody, ever, talked about the trees that are planted all over town?

And that simple question was the catalyst for so many others: What do we know about them? Who is in charge of taking care of them? Are they really useful to clean the city's pollution? Do we need more trees in the city? Is it true that only the rich neighborhoods have tall trees? Are there any historical trees in town?

We began our investigation aiming to do two different things: connect the citizens with the green giants they see everyday and understand the reality of the city's arborization plan.[1]

To do so, we analyzed the urban census of tree planting in Bogotá that the Botanical Garden made in 2007, the only set of information available and that is updated every month. The Botanical Garden refused to give us the full data even after we submitted multiple freedom of information requests filled with legal arguments. Their position was a simple one: the data was already available in their DataViz portal. Our argument: you can only download 10,000 entries and the database is made up of 1.2 million entries. It's public data, just give it to us! Their answer: We won't give it to you but we will improve our app so you can download 50,000 entries.

Our solution? Reach out to other organizations that had helped the Botanical Garden collect the data. One of those entities was Ideca, which collects all the information related to the city's cadastre. They gave us the whole dataset in no time. We, obviously, decided to publish it so that everyone can access it (call it our little revenge against opacity). The Botanical Garden realized this and stopped any further conversation with us, and we decided not to continue a legal battle.

In addition, we included public data from the Mayor's Office of Bogotá and the National Census, to cross-reference information that we could analyze in relation to trees. Finally, we conducted interviews with environmental experts and forestry engineers that allowed us to understand the challenges the city faces. They had done so much work and so many investigations analyzing not only the reality of arborization, but also the history behind the trees in the city. And most of this work was largely unnoticed by authorities, journalists and many others.

The final product was an eight piece data project that showed the reality of the arborization plan of the city, mapped every single tree – with information about its height, species, and benefits for the city-, debunked many myths around tree planting, and told the stories of some historical trees in town. We used Leaflet and SoundCloud for the interactive and the design was implemented by our talented group of coders. We also used the StoryMapJS to allow the users to explore the historic trees of the city.

We decided how and which pieces were important for the story after researching many other similar projects and then partnered with a designer to create a nice UX experience. It was our first big data project and a lot of it involved trial and error as well as exploration.

More importantly, we involved citizens by inviting them to help us build a collaborative tree catalog and to share their own stories regarding the trees we had mapped. We did so through social media, inviting them to add information about tree species to a spreadsheet. Bogotá's residents continue to help us enrich the catalogue to this day. In addition, we shared a WhatsApp number where people could send voice notes with their stories about trees. We received almost 100 voice messages from people telling stories of trees where they had their first kiss, trees that taught them how to climb, that protected them from thieves or that were missed because they were cut down. We decided to include this audio as an extra filter in the visualization app, so users could also get to know the city's trees through people's stories.

The main article and visual was then republished by a national newspaper (both in print and online), and shared by local authorities and many citizens who wanted to tell their stories and transform the relationship people had with their environment. So far, people have used the map to investigate the city's nature and to support their own research on the city's trees.

For our organisation, this has been one of the most challenging projects we have ever developed. But it is also one of the most valuable ones because it shows how data journalism can be about more than just numbers: it can also play a role in creating, collecting and sharing culture and memories, helping people notice things about the places they live (beyond graphs and charts) and multiply and change the relations between people, plants and stories in urban spaces.

Works Cited

María Isabel Magaña, Ana Hernández, David Daza, Verónica Toro, Juan Pablo Marín, Lorena Cala. '[Áboles de Botogá]', Datasketch, [2017]

From Coffee to Colonialism: Data Investigations into How the Poor Feed the Rich

Written by: Raúl Sánchez Ximena Villagrán

At the beginning of 2016, a small group of journalists decided to investigate the journey of a chocolate bar, banana or cup of coffee from the original plantations to their desks. Our investigation was prompted by reports that all of these products were produced in poor countries and mostly consumed in rich countries.

Starting from that data we decided to ask some questions: How are labour conditions in these plantations? Is there a concentration of land ownership by a small group? What kinds of environmental damage do these products cause in these countries? So *El Diario* and *El Faro* (two digital and independent media outlets) joined forces to investigate the dark side of the agroindustry business model in developing countries.[1]

'Enslaved Land' project is a one year crossborder and data-driven investigation that comes with a subheading that gets straight to the point: "This is how poor countries are used to feed rich countries".[2] In fact, colonialism is the main issue of this project. As journalists, we didn't want to tell the story of the poor indigenous people without examining a more systemic picture. We wanted to explain how land property, corruption, organized crime, local conflicts and supply chains of certain products are still part of a system of colonialism.

In this project, we investigated five crops consumed widely in Europe and the US: sugar, coffee, cocoa, banana and palm oil in Guatemala, Colombia, Ivory Coast and Honduras. As a data driven investigation, we used the data to get from pattern to story. The choice of crops and countries was made based on a previous data analysis of 68 million records of United Nations World Trade Database (Fig. 1).

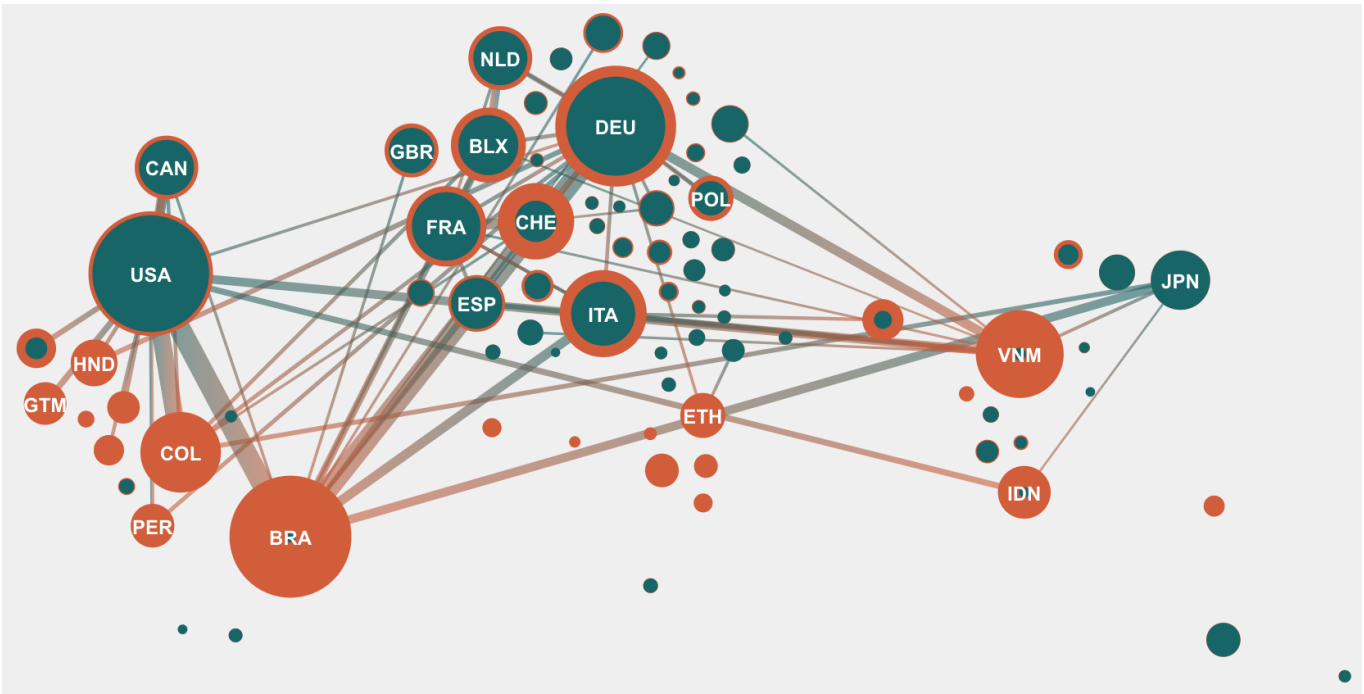


Figure 1: Network graph showing imports and exports of coffee in 2014.

This investigation shows how balance of power between rich and poor countries has changed from the 15th century to present and prove that these crops are produced thanks to exploitative, slave-like conditions for workers, illegal business practices and sustained environmental damage.

The focus of our stories were selected because of the data. In Honduras, the key was to use geographic information to tell the story. We compiled the land use atlas of the country and matched the surface of palm plantations with protected areas. We found that 7,000 palm oil hectares were illegally planted in protected areas of the country. As a result, our reporter could investigate the specific zones with palm plantations in protected areas. The story uses individual cases to highlight and narrate systemic abuse, such as the case of 'Monchito', a Honduran peasant that grows African palm in Jeannette Kawas National Park.

This project isn't only about land use. In Guatemala, we created a database of all sugar mills in the country. We dived into the local company registry to know the owners and the directors of the mills. Then we linked these people and entities with offshore companies using business public records of Panama, Virgin Islands and Bahamas. To find how they create and manage the offshore structure, El Faro had access to the Panama Papers database so we used that information to reconstruct how one of the biggest mills of the country worked with Mossack Fonseca law firm to avoid taxes.

A transnational investigation that aimed to discover corruption and businesses malpractices in third world countries is a challenge. We had to work in rural areas where there is no governmental presence, and in most cases the reporting had some risks. Also, we managed with countries where there is a considerable lack of transparency, non-open data and, in some cases, public administrations that didn't know what information they had.

Honduras and Guatemala were only a part of the investigation. More than 10 people worked together to produce this material. All this work was coordinated from the offices of eldiario.es in Spain and El Faro in El Salvador working alongside journalists in Colombia, Guatemala, Honduras and Ivory Coast.

This work was undertaken by not just journalists, but by editors, photographers, designers and developers who participated in the development and production process to make an integrated web product. This project would not have been possible without them.

We used an integrated scrolly-telling narrative for each of the investigations. For us, the way that users read and interact with the stories is as important as the investigation itself. We chose to combine satellite images, photos, data visualizations and narrative because we wanted the reader to understand the link between the products they consumed and the farmers, companies, and other actors involved in their production.

This structure allowed us to use a narrative where personal stories were as important as data analysis. For example, we told the story of John Pérez, a Colombian peasant whose land was stolen by paramilitary groups and banana corporations during the armed conflict, with a zoomable map that takes you from his plantation to the final destination of the Colombian banana production.

This project showed that data journalism can enrich traditional reporting techniques to connect stories about individuals to broader social, economic and political phenomena.

It was also published by Plaza Pública in Guatemala, Ciper in Chile and was included in the Guatemalan radio show ConCriterio. The latter led to a pronouncement from the Guatemalan Tax Agency asking for resources to fight against the tax fraud of sugar mills.

Works Cited

Sánchez et al, 'La tierra esclava' , April 2017.

Investigating Extractive Industries in Peru

This chapter is launching soon

Mobilising for Road Safety in the Philippines

This chapter is launching soon

Contextualising Carbon Emissions

This chapter is launching soon

Engaging Publics around Data Reporting on the Arab world with Instagram

This chapter is launching soon

Using Data Science and Visualization to Explore Segregation in the United States

This chapter is launching soon

Counting Transgender Lives

This chapter is launching soon

Documenting Land Conflicts Across India

Written by: Kumar Sambhav Shrivastava Ankur Paliwal

Land is a scarce resource in India. The country only has 2.4 percent of the world's land area but supports over 17% of the world's population. As one of the world's fastest growing economies, it requires large swathes of lands to fuel its ambitious agenda of industrial and infrastructure growth. At least 11 million hectares land is required for development projects in the next 15 years. But a huge section of India's population – mostly marginalised communities – depend on land for their sustenance and livelihood. Over 200 million people depend on forests while 118.9 million depend on farming land in India.

The competing demands cause conflicts. In many cases land is forcefully acquired or fraudulently grabbed by the state or private interests, dissenters are booked by the state agencies under false charges, compensation is paid partially, communities are displaced, houses are torched, and people get killed. Social disparities around the caste, class and gender also fuel land struggles. Climate change-induced calamities are making land-dependent communities further vulnerable to displacements. All this is reflected in the many battles taking place over land across India.

When we started writing about development issues in India, we came across many such conflicts. However, we realised it was not easy to sell those stories happening in remote corners of India to the editors in New Delhi. The mainstream media did not report on land conflicts except the ones that turned fatally violent or were fought in the national courts. The sporadic reporting by a few journalists had little impact. Voices of the people affected by such conflicts remained unheard. Their concerns remained unaddressed.

The reason, we thought, was that the reporters and the editors looked at the conflicts as isolated incidents. We knew land conflicts were one of the most important stories about India's political economy. But the question was how to sell it to editors and readers. We thought that if journalists could scale up their reporting on individual cases of conflicts to examine broader trends, their stories could not only have wider reach but might also show the intensity of various kind of conflicts and their impact on people, economy and the environment. The biggest challenge to achieving this was lack of a database which journalists could explore to see what trends are emerging around specific kind of conflicts such as conflicts about roads, townships, mining or the wildlife-protected areas. There was no such database of ongoing land conflicts in India. So we decided to build one.

In November, 2016, we started Land Conflict Watch, a research-based data journalism project, which aims to map and document all ongoing land conflicts in India. We developed a documentation methodology in consultation with academics working on land governance. We put together a network of researchers and journalists, who live across the country, to document the conflicts in their regions following this methodology.

For the purpose of this project, we defined land conflict as any situation that has conflicting demands or claims over the use or ownership of land, and where communities are one of the contesting parties. Ongoing conflicts where such demands or claims have already been recorded in a written or audio-visual format at any place, from the village level to the national level, are included. These records could be news reports, village assembly resolutions, records of public consultation for development projects, complaints submitted by people to government authorities, police records or court documents. Conflicts such as the property disputes between two private parties or between a private party and the government are excluded unless they directly affect broader publics.

The researchers and journalists track national and local media coverage about their regions and interact with local activists, community organisations and lawyers to find cases of conflicts. They then collect and verify information from publicly-available government documents, independent studies, and by talking to affected parties. Data such

as location of conflict, reasons behind the conflicts, number of affected people, affected area, land type – whether private, common or forest –names of the government and corporate agencies involved and a narrative summary of the conflict are documented.

Researchers file all the data into reporting-and-review software built into the Land Conflict Watch website. Data is examined and verified by dedicated reviewers. The software allows to-and-fro work flow between the researchers and the reviewers before the data is published. The dashboard, on the portal, not only presents the macro picture of the ongoing conflicts at national level but zooms in to give details of each conflicts, along with the supporting documents to back data, at the micro level. It also provides the approximate location of the conflict on the interactive map.

About 35 journalists and researchers are currently contributing. As of September 2018, the project had documented over 640 cases. These conflicts affect close to 7.3 million people and span over 2.4 million hectares of land. Investments worth $186 billion USD are attached to projects and schemes affected by these conflicts.

As a conflict is documented, it is profiled on the portal as well as on social media to give heads-up to national journalists and researchers. The project team then collaborates with journalists to create in-depth, investigative stories at the intersection of land rights, land conflicts, politics, economy, class, gender and the environment using this data. We also collaborate with national and international media to get these stories published. Many of these stories have been republished by other mainstream media outlets. We have also conducted trainings of journalists on the use of the database in finding and scaling up stories around land governance.

Land Conflict Watch is an ongoing project. Apart from designing stories, we also work with academics, researchers and students to initiate public debates. Land Conflict Watch's data has been cited by policy think-tanks in their reports. Land-governance experts have written op-eds in national newspapers using the data. We regularly get requests from research students at the Indian and foreign universities to use our data in their research. Non-profit organisations use land conflict data, documents and cases to strengthen their campaign to fight for the land rights of conflict affected communities. The stories inform people and help shape discourse around land rights and governance related issues in India.

Alternative Data Practices in China

Written by: Jinxin Ma

A couple of years ago, I delivered a presentation introducing data journalism in China at the Google News Summit, organized by Google News Lab. It was a beautiful winter day in the heart of the Silicon Valley, and the audience was a full room of a hundred or so senior media professionals mostly from western countries. I started by asking them to raise their hands if they think, firstly, if there is no good data in China, and secondly, if there is no real journalism in China. Both questions got quite some hands up, along with some laughters.

These are two common comments, if not bias, that I encounter often when I attend or speak at international journalism conferences. From my observation in the past six years, instead of having no data, there are huge amounts of data existing and being accumulated every day in China, and its quality is improving. Instead of having no real journalism, there are many journalists producing impressive stories every day, though not all of them ultimately get published.

Issue-driven Data Creation

Even before the term "data journalism" was introduced to China, data stories existed. While nowadays we normally use the term "data-driven stories" in China, there was a period when we see the contrary: instead of having data driving stories, we witnessed stories, or particular issues, driving the production of data. These are always issues that resonate with regular citizens, such as the air pollution.

Since 2010, the Ministry of Environment has published a real-time air pollution index, but one important figure was missing.[1] PM2.5, or pollutants that measure less than 2.5 micrometers in diameter, which would lead to irreversible harm to human bodies, was not published.

In the wake of the seriousness of air pollution and lack of official data of PM2.5, a nationwide campaign started in November 2011 called *I test the air for the motherland*, advocating for every citizen to contribute to monitoring air quality and publishing results to social media platforms.[2] The campaign was initiated by an environmental non-profit, with testing equipment crowd-funded from citizens, and they also provided training to interested volunteers. The movement was widely spread after a few online influencers joined forces, including Pan Shiyi, a well-known business leader, who then had more than 7 million followers on Sina Weibo, one of China's most widely used social media platforms.[3]

After two years of public campaigning, starting from January 2012, the data of PM2.5 was finally included in the government data release. It was a good start, but challenges remained. There was immediately observation on the discrepancies with the data released by the U.S. Embassy there, which brough doubts regarding the accuracy and accountability of the data.[4]

In terms of functionality, it was also not journalist-friendly. Despite hourly updates of the data from more than 100 cities, the information is only provided on a rolling basis on the webpage, but not downloadable as a dataset in any format. Though data has been centralized, historical data is not accessible for the public. In other words, without being able to write a script to scrape the data every hour and save it locally, it is impossible to do any analysis of trends over time or undertake comparisons between cities.

That was not the end of the story. Issue-driven data generation continues. When the data is not well structured or in a user-friendly format, and when data journalists struggle with limited technical skills, civil society or tech geeks can come in to provide support.

One early example back in 2011 was PM25.in, which scrapes air pollution data and releases it in a clean format. The site claims to have more than 1 billion search queries since they started operating.[5] Another example is Qing Yue, a non-governmental organization which collects and cleans environmental data from government websites at all levels, and then releases it to the public in user-friendly formats. Their processed data turns out to be widely used by not only data teams in established media outlets but also government agencies themselves for better policy making.

The generation of data and the rising awareness around certain issues have gone hand in hand. In 2015, a documentary investigating the serious air pollution took the country by storm. The self-funded film, entitled *Under the Dome*, exposed the environmental crisis of noxious smog across the country, and then traced after the roots of the problem and the various parties responsible.[6] The film has been compared with Al Gore's *An Inconvenient Truth* in both style and impact. In the storytelling, it presented lots of scientific data, charts to show analysis and explain the trends over the years, as well as a social network visualizations of corruption within environment and energy industries. As soon as it was released online, the film went viral and reached 200 million hits within 3 days, before it was censored and taken down within a week. But it had successfully raised public awareness and ignited a national debate on the issue, including around the accessibility and quality of air pollution data, and it has successfully made the country's leadership aware of the significance of the issue.

Two weeks after the release of the documentary, at the press conference the National People's Congress, addressing a question about air pollution which referred to the film, Premier Li Keqiang admitted that the government was failing to satisfy public demands to halt pollution, acknowledged some of the problems raised by the documentary, including lax enforcement of pollution restrictions, and emphasized that the government would impose heavier punishments to cut the toxic smog.[7] At the end of August 2015, the new Air Pollution Prevention and Control Law was released, and was implemented Jan 2016.[8]

Air pollution is only one example illustrating that even when data availability or accessibility is challenging, public concern with issues can lead to citizen contributions to data generation, as well as changing government attitudes and the availability of public sector data. In more established ecosystems, data can be more readily available and easy to use, and journalists' job can be more straightforward: to take data and use them as basis for stories. In China the process can be less linear, and the dynamics of citizen, government, civil society and media are much more interactive. Data, instead of just serving as the starting point for stories, can also be brought into the picture at a later stage and further enable new kinds of relations between journalists and the public.

Evolving Data Culture

The data environment in China has been changing rapidly in the past decade, partly driven by the dynamics described above, and partly due to other factors such as the global open data movement, rapidly growing internet companies, surprisingly high mobile penetration rate, etc. Data culture has been evolving around these trends as well.

Government legislation provides the policy backbone for data availability. To the surprise of many, China does have laws around Freedom of Information. *The State Council Regulations on the Disclosure of Government Information* was adopted in 2007 and came into force on May 1, 2008, which has a disclosure mandate and affirms a commitment on government transparency. Following the regulation, government agencies at all levels started dedicated web pages to disclose the information they have, including data. However, even though it gave journalists the right to request certain data or information from the authorities, in the first three years since the law enforcement, there is no publicly known cases of any media or journalists requesting data disclosure, according to a study in 2011 published by Caixin media.[9] The study revealed that, in 2010, the Southern Weekly, a leading newspaper, sent a testing request to 29 environmental bureau for certain information disclosure only got a 44%

response rate, and within media organizations there is normally no supporting system such as a legal team that could help the journalists push their demands further. One journalist who, in his personal capacity, actually took the government to the court for not disclosing information, ended up losing his job. The difficulties and risks that Chinese journalists encounter when leveraging legal tools can be much greater than their western peers.

In the wake of the global open data movement and increasing interest in big data, China was also reacting to these trends. In 2012, both Shanghai and Beijing launched their own open data portals, each with hundreds of datasets, around areas such as land usage, transportation, education, pollution monitoring, etc. In the following years, more than a dozen open data portals have been set up, not only in the biggest cities, but also in local districts and less developed provinces. The development was rather bottom-up, without standard template or structure for data release at the local level, which made the data collection at the user end not much easier. By 2015, the State Council has released the Big Data Development Action Plan, where open data was officially recognized as one of the ten key national projects, and a concrete timeline for opening government data was presented.[10] However, the official data is not always where journalists start, and also not always aligned with public interests and concerns.

On the other hand, the private sector, especially the technology giants such as Alibaba or Tencent, have over the years accumulated huge amount of data. According to its latest official results, Alibaba's annual active consumers have reached 601 million by September 30, 2018.[11] The e-commerce data from such a strong user base – equivalent to the entire Southeast Asian population – can reveal lots of trading trends, demographic shifts, urban migration directions, consumer habit changes, etc. There are also vertical review sites where more specific data is available, such as Dianping, the China equivalent of Yelp. Despite concerns around privacy and security, if used properly, those platforms provide rich resources for data journalists to mine.

One outstanding example in leveraging the big data is the Rising Lab, a team under the Shanghai Media Group, specializing on data stories about urban life.[12] The set-up of the Lab was an answer to the emerging trend of urbanization: China has more than 600 cities now, compared to 193 in 1978, with 56% of the population living in urban areas, according to a government report in 2016.[13] Shifting together with the rapid urbanization is rise of internet and mobile use, as well as lifestyle changes such as the rapid adoption of sharing economy models. These trends are having a big impact on data aggregation.

With partnership agreements and technical support from tech companies, the Lab collected data from frequently-used websites and apps by city dwellers, such as property price, number of coffee shops and bars, number of co-working spaces or easiness of public transportation, etc. reflecting various aspects of urban life. Coupled with its original methodology, the Lab has produced a series of city rankings on different aspects, such as commercial attractiveness, level of innovation, diversity of life, etc. The rankings and the stories are updated every year based on new data available but following the same methodology to ensure consistency. The concept and stories have been well received and even starting to influence urban planning policies and company's business decisions, according to SHEN Congle, Director of the Lab.

The Lab's success illustrated the new dynamics emerging between data providers, journalists, and citizens. It shows how softer topics also become a playground for data journalism, along side of the other pressing issues such as environmental crisis, corruption, judicial injustice, public health and money laundering. It also explores new potential business models for data journalism, as well as how data-based products can bring value to governments and businesses.

Readers' news consumption also has had an impact on the development of data journalism, with one being more visual and another being more mobile. Since 2011, infographics have become popular thanks to a few major news portal's effort to build dedicated vertices with infographics stories, mostly driven by data. In 2014, the story of the

downfall of the former security chief Zhou Yongkang, one of the nine most senior politicians in China, was the biggest news of year. Together with the news story, Caixin produced an interactive social network visualization to illustrate the complex network around Zhou, including 37 people and 105 companies or projects connected to him, and the relationship between these entities, all based on the 60,000-word investigative piece of its reporting team. The interactive received 4 million hits within one week, and another 20 million views on social media, according to Caixin.[14] The widespread of this project introduced the new ways of data storytelling to the public, and created the new appetite which didn't exist before.

Almost at the same time, the media industry was welcoming the mobile era. Like the Rising Lab, more and more data stories, like any other online content in China, is now disseminated mostly on mobile. According to the China Internet Network Information Center (CNNIC), more than 95% of internet users in the country have used a mobile device to access the internet in 2016.[15] WeChat, the domestic popular messaging app and social media platform, has reached 1 billion users in March, 2018.[16] The dominance of mobile platform means data stories in China are now not only mobile-first, but in many cases mobile-only. Such market demand led to a lot of lean, simple and sometime creative interactives that are mobile friendly.

In short, data culture in China has been evolving, driven by various factors from global movements to government legislation, from public demand to media requests, from new generations of data providers, to new generation of news consumers. The interdependent relationships between players have created very complex dynamics, where constraints and opportunities coexist. Data journalism has bloomed and advanced along its own path in China.

Practical Tips

This section is specifically for those who are looking to work on China-related stories, and wondering where to even get started. It won't be easy. You would have language barriers first, as most data sources are only available in Chinese. You would then have all the common issues with any data elsewhere: data accuracy, data completeness, data inconsistency, etc. Let's assume you have all the right skills to spot those issues and work on them.

First of all, who are the biggest players? Quite a number of the leading media outlets have established data teams, and it is good to follow their stories and talk to their reporters for tips. Here are a few ones you should know:

- Caixin Media, Data Visualization Lab;[17]
- The Paper, Beautiful Data Channel[18];
- Shanghai Media Group, The Rising Lab;[19]
- DT Finance.[20]

Secondly, where to find the data? A comprehensive list would be a separate handbook so here are just a few suggestions to get started:

1. Start with government websites, both central ministries and local agencies. You would need to know which department is the right one(s) for the data you are looking for, and you should check both the thematic areas of ministries (for example the Ministry of Environmental Protection) and the dedicated data website at the local level if it exists.
2. There will be data that you don't even expect - for example, would you expect that the Chinese government publicized millions of court judgements after 2014 in full text? Legal documents are relatively transparent in the U.S. but not in China. But the Supreme People's Court (SPC) started a database called China Judgments Online just doing that.
3. Once you find some data that could be useful online, make sure you download a local copy.

4. Sometimes the data is not available online. It is still common. Sometimes they are in the form of a government annual report published and you could order online, sometimes they are only available in paper archives behind certain offices. For example certain government agencies have the records of private companies but not all available online.
5. If the data is not at all released by government, check if any user-generated contents available. For example, the data of public health is very limited, but there are dedicated websites for hospital registration, or elderly centres, among others. If you could scrape and clean up the data, you would have valuable data to have a good overview of the topic.
6. Utilize databases in Hong Kong – from official ones like Hong Kong Companies Registry to independent ones such as Webb-site Reports. As mainland China and Hong Kong becoming politically and financially closer, more information is available there thanks to Hong Kong's transparent environment and legal enforcement, which may be valuable for tracing money.
7. There is data about China but not necessarily in China. There are international organizations or academic institutions that have rich China-related data. For example, The Paper used data from NASA and Harvard in one of its latest stories

Last but not least, while some of the challenges and experience are unique to China, lot of them could potentially provide some useful lessons for other countries, where the social, cultural and political arrangements have a different shape but similar constraints.

Works Cited

Nicole Jao, 'WeChat now has over 1 billion active monthly users worldwide', Technode, March 5, 2018

Edward Wong and Chris Buckley, 'Chinese Premier Vows Tougher Regulation on Air Pollution', New York Times, March 15, 2015

Zhao Lijian, Tonny Xie and Jenny Tang, 'How China's new air law aims to curb pollution', China Dialogue, December 30, 2015

Man-Chung Cheung, 'More than 95% of Internet Users in China Use Mobile Devices to Go Online', Marketer, February 2, 2017

'How does the Chinese media promote information disclosure' [中国媒体如何推进信息公开], ifeng, September, 2011

The State Council of the people's Republic of China, 'Govt report: China's urbanization level reached 56.1%', April 16, 2016

Ivan Anson et al, 'PM2.5', BestApp Studio, 2011.

Chai Jing, 'Chai Jing's review: Under the Dome – Investigating China's Smog', [柴静雾霾调查：穹顶之下], YouTube, March 1, 2015.

'Zhou Yongkang's People and Finances' [周永康的人与财], Caixin Data Visualization Lab, 2015

Reassembling Public Data in Cuba: How Journalists, Researchers and Students Collaborate When Information Is Missing, Outdated or Scarce

Written by: Yudivián Almeida Cruz Saimi Reyes Carmona

Postada.club is a small team. At the beginning, we were four journalists and a specialist in mathematics and computer science, who decided in 2014 to venture together into data journalism in Cuba. We also wanted to investigate the issues related to that practice. In Cuba, until that moment, there was no media outlet that was explicitly dedicated to data journalism. Postdata.club was the first.

Right now we are two journalists and a data scientist working in our free time in Postdata.club. In none of our jobs we directly perform data journalism, because Saimi Reyes is editor of a culture related website, Yudivián Almeida is a professor of the School of Math and Computer Science at the University of Havana and Ernesto Guerra is journalist in a magazine about popular science and technology. Our purpose is to be more than a media organization, an experimental space where we want to explore and learn about the nation we live in with and through data.

We set out to use open and public data and wanted to share both: our research and the way we do it. That's why we started using Github, the platform where Postdata.club lives. Depending on the requirements of each story we want to tell, we decide on the extension of our texts and the resources we will use, be they graphics, images, videos, audios. We focus on journalism with social impact, sometimes long form, sometimes short form. We are interested in all the subjects that we can approach with data, but, above all, those related to Cuba or its people.

For our investigations we work in two ways, depending on the data. Sometimes we have access to public and open databases. With these, we undertake data analysis to see if there may be a story to tell. Sometimes we have questions and go straight to the data to find the answers for a story. In other cases, we explore the data and in the process find elements that we believe may be interesting or questions arise whose answers may be relevant and which may be answered by that data source or by another.

If the information we get from these databases looks interesting, we complement it with other sources such as interviews and comparing with other information sources. Then, we think how to narrate the research with one or more written texts about the subject accompanied by visualizations to present insights from the data.

Other times – and on more than a few occasions – we have to create databases ourselves based on information that is public but not properly structured and use these as the basis for our analysis and inquiry. For example, to address the topic of Cuban elections, we had to build databases based on information from different sources. For this, we started with data published on the site of the Cuban Parliament, however, these were not complete, so we completed our databases with press reports and information published on sites related to the Communist Party of Cuba. Later, in order to approach the recently designated Council of Ministers, it was also necessary to build another database. In that case, information provided by the National Assembly was not complete and we used press reports, the Official Gazette and another informative sites to get a fuller picture. In both cases, we created databases in JSON format which were processed and used for most of the articles we conceived about the elections and the executive and legislative powers in Cuba.

In most cases we share such databases on our website with an explanation of our methods. However, our work in Cuba is sometimes complicated by the lack of some data that should be public and accessible. Much of the information we use is provided by government entities, but in our country many institutions are not properly represented on the Internet or do not publicly report all the information they should. In some cases we have gone directly to these institutions to request access to certain information, a procedure which is often cumbersome, but important.

For us, one of the biggest issues obtaining data in Cuba lies in its outdatedness. When we finally have access to information we are looking for, it is often not complete, or it is very outdated. Thus, the data may be available for consultation and download on a website, but the last date corresponds to five years ago. In some cases, we must complete the information by looking at different sites that are reliable. In others cases, we must go to printed documents, images or live sources that help us to work with recent information. This has made our way of working different depending on each investigation and the available data. These are the particularities of our environment and this is the starting point from which we set out to offer our readers good journalism that has a social impact. If the information we share is useful for at least one person, we feel it's worth it.

In addition to maintaining Postdata.club website, where we place the articles and stories that result from our research, we also want to extend this way of doing data journalism to other spaces. Thus, since 2017, we have taught a Data Journalism course to students of the journalism programme at the School of Communication of the University of Havana. This subject had barely been taught in our country and this therefore requires ongoing learning and preparation, while receiving feedback from students and other colleagues.

Through our exchanges with these future journalists and communication professionals we have learned many new ways of working and, surprisingly, we have found out new ways to access information. One of the things we do in these classes is to involve students in the construction of a database. There was no single source in Cuba to obtain the names of the people who have received national awards, based on their life's work in different areas and activities. With all of the students and teachers, we collected and structured a database of more than 27 awards since they began to be granted so far. This information allowed us to reveal that there was a gender gap in the awarding of prizes. Women received these prizes only 25% of the time. With this discovery we were able, together, to write a story that encouraged reflection about gender issues in relation to the national recognition of different kinds of work.

In 2017 also, we had another revealing experience that helped us to understand that, in many cases, we should dare not to settle for existing published databases and that we should not make too many assumptions about what is and isn't possible. As part of their final coursework, we asked students to form small teams to carry out their task. These were composed, in each case, by one of the four members of the Postdata.club team, two students of journalism and a student of computer science, who had integrated the course to achieve an interdisciplinary dynamic. One of the teams proposed to tackle new initiatives of self-employment in Cuba. Here, these people are called "cuentapropistas". What was a few years ago a very limited practice, is now rapidly growing due to the gradual acceptance of this form of employment in society.

We wanted to investigate the self-employment phenomenon in Cuba. Although the issue had been frequently addressed, there was almost nothing about the specificities of self-employment by province, the number of licenses granted per area activities, or trends over time. Together with the students, we discussed which questions to address and came to the conclusion that we lacked sources with usable data. In places where this information would have been posted publicly, there was no trace. Nor was there any information in the national press that contained a significant amount of data. Beyond some interviews and isolated figures, nothing was published widely.

We thought that the data would be difficult to obtain. Nevertheless, journalism students from our programme approached the Ministry of Labor and Social Security and asked for information about self-employment in Cuba. In the Ministry they were informed that they could give them the database and in a few days the students had it in their hands. Suddenly, we had all the information that interested many Cubans, and we could also share it, because, in fact, it was meant to be public. The Ministry did not have an up-to-date internet portal and we had wrongly assumed that the data was not accessible.

Students, along with the future computer scientist and journalist from Postdata.club, prepared their story about self-employment in Cuba. They described from the data, and in a detailed way, the situation of this kind of employment in the country. Coincidentally, the information came into our hands at a particularly active time on this subject. For those months, the Ministry of Labor and Social Security decided to limit the delivery of licenses for 28 activities of those authorized for non-state employment. We were thus able to quickly use the data we had to analyse how these new measures would affect the economy of the country and the lives of self-employed workers.

Most of our readers were surprised that we had obtained the data and that it was relatively easy to obtain. In the end it was possible to access this data because our students had asked the ministry and until today it's only in Postdata.club where this information is public, so everyone can consult and analyze it.

Doing data journalism in Cuba continues to be a challenge. Amongst other things, the dynamics of creating and accessing data and the political and institutional cultures are different from other countries where data can be more readily available. Therefore we must always be creative in looking for new ways of accessing information and, from it, to tell stories about issues that matter. It is only possible if we continue to try, and at Postdata.club we will always strive to be an example of how data journalism is possible even in regions where data can be harder to come by.

Narrating a Number and Staying with the Trouble of Value

Written by: Helen Verran

In 2009 the contribution to Australia's GDP from transactions in which the state purchased environmental interventions to enhance ecosystems value, from rural landholders in the Corangamite NRMR was calculated as AUD4.94 million.

The number that I narrate here emerged in a press statement issued by the government of the Australian State of Victoria in 2009. The media release announced the success of investment by the State Government in environmental conservation in one of Australia's fifty-seven Natural Resource Management Regions (NRMR). The environmental administrative region of grassy basalt plains that spreads east-west in south central Victoria is named Corangamite, an Aboriginal term that replaced a name bestowed by the first British pastoralists who in the mid nineteenth century invaded this country from Tasmania. They called the region 'Australia Felix' and set about cutting down all the trees. The squatters, who subsequently became landowners here, would in less than a century become a sort of colonial landed-gentry. In 2008, in operating the EcoTender Programme in the Corangamite NRMR, the Victorian government purchased ecosystems services value from the descendants of those squatters in pay-as-bid auctions. In 2009 the contribution to Australia's GDP from these transactions was calculated as AUD4.94 million. The announcement of this value was the occasion of the media release where I first met the number.

I doubt that any journalists picked up on the news promulgated in this brief including its numbered value; this number is hardly hot news. In the context of a press release the naming of a specific number value reassures. The national accounts are important and real, and if this regional government intervention features as a specified value contributing to the national economy, then clearly the government intervention is a good thing. The specification of value here claims a realness for the improvements that the government interventions are having. The implication is that this policy leads to good environmental governance. Of course, the actual value the number name (AUD 4.94 million) points to, what it implicitly claims to index, is not of much interest to anyone. That a number appears to correspond to something out-there that can be valued, is good enough for purposes of reassuring.

My narration of this number offers a mind-numbingly detailed account of the socio-technical means by which the number came to life. The story has the disturbing effect of revealing that this banal number in its workaday media release is a paper-thin cover-up. Profound troubles lurk. Before I begin to tell my story and articulate the nature of these profound troubles that seem to shadow any doing of valuation, even such a banal doing, let me pre-emptively respond to some questions that I imagine might be beginning to emerge for readers of the *Data Journalism Handbook*.

First, I acknowledge that telling a story of how a number has come to life rather than finding some means to promote visualization of what that number means in a particular context, is rather an unusual approach in contemporary data journalism. I can imagine a data journalist doubting that such story telling would work. Perhaps a first response is to remind you that it is not an either/or choice and that working by intertwining narrative and visualizing resources in decoding and interpreting is an effective way to get ideas across. In presenting such a an intertwining, journalists should always remember that there are two basic speaking positions in mixing narratives and visuals. One might proceed as if the visual is embedded within the narrative in which case you are speaking to the visual which seems to represent or illustrate something in the story. Or, you

can proceed as if the narrative is embedded in the visual in which case you are speaking *from within* diagram. This is a less common strategy in data journalism, yet I can imagine that the story I tell here could well be used in that way. Of course, switching between these speaking positions within a single piece is perhaps the most effective strategy.[1]

Second, you might see it as odd to tell a story of a very particular number when what clearly has agency when it comes to decision making and policy design, and what data journalists are interested in, is what can be made of datasets in mobilizing this algorithm or that. This worry might prompt you to ask about relations between numbers and datasets. The answer to such a query is fairly straight forward and not very interesting. There are many numbers in a dataset; the relation is a one-many relation albeit that numbers are assembled in very precise arrays. The more interesting question enquires about the relation between numbers and algorithms. My answer would be that while algorithms mobilise a protocol that elaborates how to work relations embedded in a database, numbers express a protocol that lays out how to work relations of collective being. Numbering is a form of algorithming and vice versa.[2] We could say that numbers are to algorithms as a seed is to the plant that might germinate from it; to mix metaphors, they have a chicken and egg relation. While there are certain interestingly different socio-technical characteristics of generating enumerated value by analogue means (mixing cognitive, linguistic, and graphic resources) of conventional enumeration as taught to primary school children, and contriving enumerated value by digital computation, it is the sameness that matters here: AUD4.94 million has been generated algorithmically and expresses a particular set of relations embedded in a particular data set, but it still presents as just a number.[3]

So now, to turn to my story. The intimate account of number making I tell here as a story would enable a journalist to recognize that the good-news-story that the government is slyly soliciting with its media release is not a straightforward matter. We see that perhaps a political exposé would be more appropriate. The details of how the number is made reveal that this public-private partnership environmental intervention program involves the state paying very rich landowners to do work that will increase the value of their own property. The question my story might precipitate is how could a journalist either celebrate or expose this number in good faith? When I finish the story, I will suggest that that is not the right question.

Narrating a Number

What is the series of socio-technical processes by which ecosystems services value come into existence in this PPP programme in order that this value might be traded between government as buyer and landowner as vendor? And exactly how does the economic value of the trade come to contribute to the total marginal gains achieved in the totality of Australian economic activity, Australia's gross domestic product (GDP)? I attend to this double-barrelled question with a step by step laying out of what is required for a landholder to create a product – 'ecosystems services value' – that can compete in a government organised auction for a contract to supply the government with 'ecosystem services value'. The messy work in which this product comes to life involves mucking around in the dirt, planting tree seedlings, fixing fences, and generally attempting to repair the damage done to the land perhaps by the landowner's grandparents, who heedlessly and greedily denuded the country of trees and seeded it with water hungry plants, in hopes of more grain or more wool and family fortune. Ecosystems services value is generated by intervening in environmental processes.

The value which is the product to be traded, begins in the work of public servants employed by a Victorian State Government department (at that time Department of Sustainability and Environment, DSE). Collectively these officials decide the areas of the State within which the administration will 'run' tenders. In doing this, *EnSym*, an environmental systems modelling platform is a crucial tool.This computing capacity is a marvel, it knows 'nature out there' as no scientist has ever known nature. Precise and focussed representations can be produced—probably overnight.

‘This software has been developed by the ecoMarkets team and incorporates science, standards, metrics and information developed within DSE, as well as many leading international and national scientific models.

EnSym contains three main tools – the ‘Site Assessment Tool’ for field work, the ‘Landscape Preference Tool’ for asset prioritization and metric building, and ‘BioSim’ for catchment planning’.[4]

Prioritizing and mapping the areas of the State where auctions will be established, specifying and quantifying the environmental benefits, the ecological values, that might be enhanced through on-ground conservation and revegetation works, are recorded in numerical form. They represent ecosystem properties in the out-there land. And the computer program can do more than that, it can also produce a script for intervention by humans. Just as the script of a play calls for production, so too does this script. And, as that script comes to life, nature out-there seems to draw closer. It ceases to be an entirely removed ‘nature out there’ and becomes nature as an infrastructure of human lives, an infrastructure that we might poke around in so as to fix the ‘plumbing’.

When the script for a choreographed production of collective human effort is ready, in the next step the government calls for expressions of interest from landholders in the project area. In response to submitted expressions of interest, a government officer visits all properties. We can imagine this officer as taking the general script generated by *EnSym* along to an actual place at a given time. He or she has a formidable translation task ahead.

The field officer assesses possible sites for works that might become a stage for the production of the script. The aim is to enhance the generation of the specified ecosystems services, so the officer needs to assess the likelihood that specified actions in a particular place will produce an increase in services provision from the ecosystem, thus increasing the value of that particular ecosystems service generated by that property, and through adding together the many such increases generated in this intervention program, by the state as a whole. Together the landowner and the government officer hatch a plan. In ongoing negotiation, a formalized management plan for specified plots is devised. The field officer develops this plan in contractable terms. Landholders specify in detail the actual work they will do to action the plan. Thus, a product that takes the form of a particular ‘ecosystems services value’ is designed and specified as a series of specified tasks to be completed in a specified time period: so many seedlings of this set of species, planted in this array, in this particular corner of this particular paddock, and fenced off to effect a conservation plot of such and such dimensions, using these materials.

Landholders calculate the cost of the works specified by the state, no doubt including a generous labour payment. They come up with a price the government must pay if it is to buy this product, a particular ‘ecosystems services value’. Here they are specifying the amount of money they are willing accept to undertake the specified works and hence deliver the ecosystems services value by the specified date. They submit relevant documents to the government in a sealed envelope.

So how does the subsequent auction work? Here *EnSym* becomes significant again in assessing the bids. Not only a knower of nature out there, and a writer of scripts for intervention in that ‘out there’ imagined as infrastructure, *EnSym* is also a removed judging observer that can evaluate the bids that have been made to produce that script, much like a Warner Brothers might evaluate competing bids to produce a movie. Bids are ranked according to a calculated ‘environmental benefits index’ and the price proposed by the landowner. We must suppose that the government buys the product which offers the highest ‘environmental benefits index’ per unit cost.

‘Bid assessment. All bids are assessed objectively on the basis of

- the estimated change in environmental outcomes
- the value of the change in environmental outcomes
- the value of the assets affected by these changes (significance)

- dollar cost (price determined by the landholder)

Funds are then allocated on the basis of best value for money'

When the results of the auction are announced, selected bidders sign a final agreement based on the management plan and submitted schedule of works as defined spatial and temporal organization. When all documents are signed, reporting arrangements are implemented and payment can begin.

'DSE forwards payment to signed-up landholders on receipt of an invoice. Payments occur subject to satisfactory progress against actions as specified in the Management Agreement'

This is a good thing, right?

What I have laid out is a precise description of how to buy and sell ecosystems services value. This takes me back to the press release. A quick reading of the media statement might leave a reader with the impression that AUD4.94 million is the value of the additional natural capital value that this government programme has generated. At first glance AUD4.94 million appears to be the marginal gain in Australia's natural capital value that was achieved in the program. But that is a mistake. AUD4.94 million is *not* the name of a natural capital value. I explain what this number name references below. At this point I want to stay with the product that has been bought and sold in this auction. This product is the trouble I want to stay with.

I want to ask about the value of the increase in "ecosystems services value" that this elaborate and rather costly government program has achieved. A careful reading of the details of the work by which this increase in value comes into being, reveals that nowhere and at no time in the process has that value ever been named or specified. The product that is so rigorously bought and sold is an *absence*. And worse there is literally no way that it could ever be otherwise. The program is a very elaborate accounting exercise for a means of giving away money. When this becomes clear to an outsider, it also becomes obvious that this actuality of what the exercise *is* has never been hidden. When it comes down to it, this program is a legitimate means for shifting money from the state coffers into the hands of private landowners.

Recognizing that this is a program of environmental governance in a liberal parliamentary democracy in which the social technology of the political party is crucial, let me as your narrator temporarily put on a party-political hat. Corangamite is an electorate that has a history of swinging between choosing a member of the left of centre party (Labor Party) or a member of the right of centre party (Liberal Party) to represent the people of the area in the Victorian Parliament. It is clearly in the interests of any government—left-leaning or right-leaning to appeal to the voters of the electorate. And there is no better way to do that than by finding ways to legitimately transfer resources from the state to the bank accounts of constituents. That there is no possibility of putting a number on the value of the product the state buys and the landowners sell here, is on this reading, of no concern.

So, let me sum up. Economically this program is justified as generating environmental services value. Described in this way this is a good news story. Taxpayer money used well to improve the environment and get trees planted to ameliorate Victoria's excessive carbon dioxide generation. Problematically the increase in the value of Victoria's natural capital cannot be named, articulated as a number, despite it being a product that is bought and sold. It seems that while there are still technical hitches, clearly, *this is a good thing*.

But equally, using a different economics this program can just as legitimately be described as funding the labour of tree planting to enhance property values of private landowners. It is a means of intervening to put right damage caused by previous government programs subsidising the misallocated labour of land clearing that in all likelihood

the landowners grandparents profited by, creating a benefit which the landowner continues to enjoy. On this reading the government policy effected in EcoTender is an expensive program to legitimately give away tax payer money. Clearly, *this is a bad thing*.

On not disrespecting numbers and algorithms: staying with the troubles of value

So, what is a journalist to do? Writing as a scholar and not as a journalist, I can respond to that obvious question only vaguely. In beginning I return to my claim that the number name used in the press release is a paper-thin cover-up to divert attention from lurking trouble. As I see it valuation always brings moral trouble that can never be contained for long. The right question to ask I think is, "How might a data journalist respond to that moral trouble?"

First, I clear up the matter of the AUD$4.94 million. What is this figure? Where does this neatly named monetary value come from? This is how it is described in an academic paper offering critical commentary on the EcoTender program

'Under this market-based model economic value from ecosystems services is created when the per-unit costs of complying with the conservation contract are less than the per-unit price awarded to the successful participants in the auction. While [for these sellers] some economic value is lost through the possibility of foregone production of marketed commodities, the participation constraint of rational landowners ensures that there will be a net increase in [economic] value created in the conduct of the auction'. [5]

Under the economic modelling of this policy, the assumption is that landowners will efficiently calculate the costs they will incur in producing the government's script for intervening in nature as infrastructure—in generating a more efficient performance of the workings of natural infrastructure. Of course, everyone assumes that a profit will be made by the landowner, although of course, it is always possible that instead of a profit the landowner will have miscalculated and made a loss, but that is of no interest to the government as the buyer of the value generated by the landowners' labour.

What is of interest to the government is the issue of how this economic transaction can be articulated in a seemly manner. Quite a problem when the product bought and sold has existence solely within the circuit of an auction. The solution to this problematic form of being of the product is the elaborate complex and complicated technology of the national accounts system. Establishing a market for ecosystems services value, the government wants to show itself as making a difference in nature. And the national accounts are the very convenient place where this can be shown in monetary terms. The 'environmental benefits index' the particular value on the basis of which the government has purchased a particular product—an environmental services value, is ephemeral. It exists solely as a flash, a moment in the auction.[6] Despite this difficulty in the form of its existence, by ingenious contrivance, both the means of buying and selling something that has a single ephemeral moment of existence is achieved, and evidence of the specific instance of economic activity can be incorporated into the national accounts, albeit that some economists have serious reservations about accuracy.[7]

AUD 4.94 million is remote from the action of the EcoTender program and from the nature it is designed to improve. But clearly if the government makes a statement that its programs have successfully improved a degraded and damaged nature it is best to find a way to indicate the extent of that improvement. It seems any number is better than none in this situation. And certainly, this is a happy, positive number. An unhappy, negative number that no doubt is available to the government accountants—the value of the cost of running the government program, would never do here. Why go on about this oddly out of place number name? Surely this is going a bit far? What is the harm of a little sleight of hand that is relatively easily picked up? My worry here is that this is misuse of a number that seems to be deliberate. It fails to respect numbers, and refuses to acknowledge the trouble that numbering, or in this case algorithming, always precipitates. It trashes a protocol.

My narrating of a number I found on a visit to a government website, has unambiguously revealed a government program that generates social goods and bads simultaneously. The sleight of hand number naming (using the precise value AUD 4.94 million in the media release) that I also found in my narration, points off to the side, at something that is always threatening to overwhelm us: valuation as a site of moral tension and trouble.

Is the big claim here that value is moral trouble that can never be contained for long? Value theory is a vast topic that has ancient roots in all philosophical traditions, and this is a rabbit warren of vast proportions that I decline to enter. I merely note that claims, often heard over that past thirty years, that the invisible hand of the market tames the moral trouble that tracks with value, is a dangerous exaggeration. Markets might find ways to momentarily and ephemerally tame value—as my story reveals. But the trouble with value always returns. Attending to *that* is the calling of the data journalist.

Here are a few suggestions on how a data journalist might respect numbers and algorithms—as protocols. When you are faced with an untroubled surface, where no hint of moral tension is to be found, but still something lurks, then 'prick up' your ears and eyes. Attune yourself to numbers and algorithms *in situ*; work out how to think with a number that catches at you. Find ways to dilate the peep-holes that number names cover. Cultivate respectful forms of address for numbers and algorithms in practicing curiosity in disciplined ways. Recognise that numbers have pre-established natures and special abilities that emerge in encounter; that the actualities of series of practices by which they come to be, matter. Be sure that when you can do these well enough, surprises lie in store. Interesting things happen inside numbers as they come to be.

Works cited

Helen Verran and Brit Ross Winthereik, 'Innovation with Words and Visuals. A Baroque Sensibility', in *Modes of Knowing*. eds. by John Law and Evelyn Ruppert, (Mattering Press: 2016).

Helen Watson, 'Investigating the Social Foundations of Mathematics: Natural Number in Culturally Diverse Forms of Life', *Social Studies of Science* 20, 1990, pp. 283-312.

Helen Verran, 'Two Consistent Logics of Numbering', Science and an African Logic, *Chicago University Press*, (2001).

Helen Verran, 'Enumerated Entities in Public Policy and Governance in Mathematics, Substance and Surmise', eds. by Ernest Davis and Philip Davis, (Springer International Publishing Switzerland, 2015, DOI 10.1007/978-3-319-21473-3_18.

Department of Sustainability and Environment, Victorian Government. EcoMarkets. EcoTender and BushTender. http://www.dse.vic.gov.au/ecom... (2008)

Jon Roffe, 'Abstract Market Theory', *Palgrave Macmillan*, 2015.

Gary Stoneham, Andrew O'Keefe, Mark Eigenraam, David Bain, 'Creating physical environmental asset accounts from markets for ecosystem conservation', *Ecological Economics* 82, (2012) pp. 114–122, p. 118.

Structured Thinking: The Case for Making Data

This chapter is launching soon

Making Data with Readers at La Nacion

This chapter is launching soon

Making Data for Investigations at Thomson Reuters, openDemocracy and Greenpeace

This chapter is launching soon

Mapping Pollution in Indian Cities

This chapter is launching soon

Accounting for Methods in Data Journalism: Spreadsheets, Scripts and Programming Notebooks

Written by: Sam Leon

With the rise of data journalism, ideas around what can be considered a journalistic source are changing. Sources come in many forms now: public datasets, leaked troves of emails, scanned documents, satellite imagery and sensor data. In tandem with this, new methods for finding stories in these sources are emerging. Machine learning, text analysis and some of the other techniques explored elsewhere in this book are increasingly being deployed in the service of the scoop.

But data, despite its aura of hard objective truth, can be distorted and mis-represented. There are many ways in which data journalists can introduce error into their interpretation of a dataset and publish a misleading story. There could be issues at the point of data collection which prevent general inferences being made to a broader population. This could, for instance, be a result of a self-selection bias in the way a sample was chosen, something that has become a common problem in the age of internet polls and surveys. Errors can also be introduced at the data processing stage. Data processing or cleaning, can involve geocoding, correcting misspelled names, harmonising categories or excluding certain data points altogether if, for instance, they are considered statistical outliers. A good example of this kind of error at work is the inaccurate geocoding of IP addresses in a widely reported study that purported to show a correlation between political persuasion and consumption of porn[1]. Then of course we have the meat of the data journalist's work, analysis. Any number of statistical fallacies may affect this portion of the work such as mistaking correlation with causation or choosing an inappropriate statistic to summarise the dataset in question.

Given the ways in which collection, treatment and analysis of data can change a narrative - how does the data journalist reassure the reader that the sources they have used are reliable and that the work done to derive their conclusions is sound?

In the case that the data journalist is simply reporting the data or research findings of a third-party, they need not deviate from traditional editorial standards adopted by many major news outlets. A reference to the institution that collected and analysed the data is generally sufficient. For example, a recent Financial Times chart on life expectancy in the UK is accompanied by a note which says: "Source: Club Vita calculations based on EuroStat data". In principle, the reader can then make an assessment of the credibility of the institution quoted. While a responsible journalist will only report studies they believe to be reliable, the third-party institution is largely responsible for accounting for the methods through which it arrived at its conclusions. In an academic context, this will likely include processes of peer review and in the case of scientific publishing it will invariably include some level of methodological transparency.

In the increasingly common case where the journalistic organisation produces the data-driven research, then they themselves are accountable to the reader for the reliability of the results they are reporting. Journalists have responded to the challenge of accounting for their methods in different ways. One common approach is to give a description of the general methodology used to arrive at the conclusions within a story. These descriptions should be framed as far as possible in plain, non-technical language so as to be comprehensible to the widest possible audience. A good example of this approach taken by the Guardian and Global Witness in explaining how they count deaths of environmental activists for their *Environmental Defenders* series.[2]

But – as with all ways of accounting for social life – written accounts have their limits. The most significant issue with them is that they generally do not specify the exact procedures used to produce the analysis or prepare the data. This makes it difficult, or in some cases impossible, to exactly reproduce steps taken by the reporters to reach their conclusions. In other words, a written account is generally not a reproducible one. In the example above, where the data acquisition, processing and analysis steps are relatively straightforward, there may be no additional value in going beyond a general written description. However, when more complicated techniques are employed there may be a strong case for employing reproducible approaches.

Reproducible data journalism

Reproducibility is widely regarded as a pillar of the modern scientific method. It aids in the process of corroborating results and to help identify and address problematic findings or questionable theories. In principle, the same mechanisms can help to weed out erroneous or misleading uses of data in the journalistic context.

A look at one of the most well-publicised methodological errors in recent academic history can be instructive. In a 2010 paper, Harvard's Carmen Reinhart and Kenneth Rogoff purposed to have shown that average real economic growth slows (a 0.1% decline) when a country's debt rises to more than 90% of gross domestic product (GDP).[3] This figure was then used as ammunition by politicians endorsing austerity measures.

As it turned out, the regression was based on an Excel error. Rather than taking the mean of a whole row of countries, Reinhart and Rogoff had made an error in their formula which meant only 15 out of the 20 countries they looked at were incorporated. Once the all the countries were considered the 0.1% "decline" became a 2.2% average increase in economic growth. The mistake was only picked up when PhD candidate Thomas Herndon and professors Michael Ash and Robert Pollin looked at the original spreadsheet that Reinhard and Rogoff had worked off. This demonstrates the importance of having not just the method written out in plain language - but also having the data and technology used for the analysis itself. But the Reinhart-Rogoff error perhaps points to something else as well - Microsoft Excel, and spreadsheet software in general, may not be the best technology for creating reproducible analysis.

Excel hides much of the process of working with data by design. Formulas - which do most of the analytical work in a spreadsheet - are only visible when clicking on a cell. This means that it is harder to review the actual steps taken to reaching a given conclusion. While we will never know for sure, one may imagine that had Reinhart and Rogoff's analytical work been done in a language in which the steps had to be declared explicitly (e.g. a programming language) the error could have been spotted prior to publication.

Excel based workflows generally encourage the removal of the steps taken to arrive at a conclusion. Values rather than formulas are often copied across to other sheets or columns leaving the "undo" key as the only route back to how a given number was actually generated. "Undo" histories of course are generally erased when an application is closed, and are therefore not a good place for storing important methodological information.

The rise of the literate programming environment: Jupyter notebooks in the newsroom

An emerging approach to methodological transparency is to use so-called "literate programming" environments. Organisations like Buzzfeed, The New York Times and Correctiv are using them to provide human readable documents that can also be executed by a machine in order to reproduce exactly the steps taken in a given analysis.[4]

First articulated by Donald Knuth in the 1980s, literate programming is an approach to writing computer code where the author intersperses code with ordinary human language explaining the steps taken.[5] The two main literate programming environments in use today are Jupyter Notebooks and R Markdown.[6] Both produce human

readable documents that mix plain English, visualisations and code in a single document that can usually be rendered in HTML and published on the web. Original data can be linked to explicitly and any other technical dependencies such as third-party libraries will be clearly identified.

Not only is there an emphasis on human readable explanation, the code is ordered so as to reflect human logic. Documents written in this paradigm can therefore read like a set of steps in an argument or a series of answers to a set of research questions.

> "The practitioner of literate programming can be regarded as an essayist, whose main concern is with exposition and excellence of style. Such an author, with thesaurus in hand, chooses the names of variables carefully and explains what each variable means. He or she strives for a program that is comprehensible because its concepts have been introduced in an order that is best for human understanding, using a mixture of formal and informal methods that reinforce each other."[7]

A good example of the form is found in Buzzfeed News' Jupyter Notebook detailing how they analysed trends in California's wildfires.[8] Whilst the notebook contains all the code and links to source data required to reproduce the analysis, the thrust of the document is a narrative or conversation with the source data. Explanations are set out under headings that follow a logical line of enquiry. Visualisations and charts are used to bring out key themes.

One aspect of the "literate" approach to programming is that the documents produced (as Jupyter Notebook or R Markdown files) may be capable of re-assuring even those readers who cannot read the code itself that the steps taken to produce the conclusions are sound. The idea is similar to Steven Shapin and Simon Schaffer's account of "virtual witnessing" as a means of establishing matters of fact in early modern science. Using Robert Boyle's experimental programme as an example Shapin and Schaffer set out the role that "virtual witnessing" had:

> "The technology of virtual witnessing involves the production in a reader's mind of such an image of an experimental scene as obviates the necessity for either direct witness or replication. Through virtual witnessing the multiplication of witnesses could be, in principle, unlimited. It was therefore the most powerful technology for constituting matters of fact. The validation of experiments, and the crediting of their outcomes as matters of fact, necessarily entailed their realization in the laboratory of the mind and the mind's eye. What was required was a technology of trust and assurance that the things had been done and done in the way claimed."[9].

Documents produced by literate programming environments such as as Jupyter Notebooks - when published alongside articles - may have a similar effect in that they enable the non-programming reader to visualise the steps taken to produce the findings in a particular story. While the non-programming reader may not be able to understand or run the code itself, comments and explanations in the document may be capable of re-assuring them that appropriate steps were taken to mitigate error.

Take for instance a recent Buzzfeed News story on children's home inspections in the UK.[10] The Jupyter Notebook has specific steps to check that data has been correctly filtered (Figure 1) providing a backstop against the types of simple but serious mistakes that caught Reinhart and Rogoff out. While the exact content of the code may not be comprehensible to the non-technical reader, the presence of these tests and backstops against error with appropriate plain English explanations may go some way to showing that the steps taken to produce the journalist's findings were sound.

```
In [11]: # Make sure that we've identified all private-sector owners
         assert (
             as_at_data_filtered
             .loc[lambda df: df["Sector"] == "Private"]
             ["Owner"].isnull()
             .sum()
         ) == 0
```

Figure 1: A cell from the Buzzfeed Jupyter notebook with a human readable explanation or comment explaining that its purpose is to check that the filtering of the raw data was performed correctly

More than just reproducibility

Using literate programming environments for data stories does not just help make them more reproducible.

Publishing code can aid collaboration between organisations. In 2016, Global Witness published a web scraper that extracted details on companies and their shareholders from the Papua New Guinea company register.[11] The initial piece of research aimed to identify the key beneficiaries of the corruption-prone trade in tropical timber which is having a devastating impact on local communities. While Global Witness had no immediate plans to re-use the scraper it developed, the underlying code was published on Github – the popular code sharing website.

Not long after, a community advocacy organisation, ACT NOW!, downloaded the code from the scraper, improved it and incorporated it into a their iPNG project that lets members of the public cross-check names of company shareholders and directors against other public interest sources.[12] The scraper is now part of the core data infrastructure of the site, retrieving data from the Papua New Guinea company registry twice a year.

Writing code within a literate programming environment can also help to streamline certain internal processes where others within an organisation need to understand and check an analysis prior to publication. At Global Witness, Jupyter Notebooks have been used to streamline the legal review process. As notebooks set out the steps taken to a get a certain finding in a logical order, lawyers can then make a more accurate assessment of the legal risks associated with a particular allegation.

In the context of investigative journalism, one area where this can be particularly important is where assumptions are made around the identity of specific individuals referenced in a dataset. As part of our recent work on the state of corporate transparency in the UK, we wanted to establish which individuals controlled a very large number of companies. This is indicative (although not proof) of them being a so-called “nominee” which in certain contexts - such as when the individual is listed as Person of Significant Control (PSC) - is illegal. When publishing the list of names of those individuals who controlled the most companies, the legal team wanted to know how we knew a specific individual, let’s say John Barry Smith, was the same as another individual named John B. Smith.[13] A Jupyter Notebook was able to clearly capture how we had performed this type of deduplication by presenting a table at the relevant step that set out the features that were used to assert the identity of individuals (see below).[14] These same processes have been used at Global Witness for fact checking purposes as well.

Potential nominee PSCs

Which PSCs currently control the most number of companies?

In [44]:
```
temp_df = active_psc_records.groupby(['name_elements.forename','name_elements.surname','month_year
_birth','address.postal_code'])[['company_number']]\
    .agg(unique_company_count).sort_values(by='company_number',ascending=False)
temp_df_for_viz = temp_df.copy()
temp_df_for_viz = temp_df_for_viz.reset_index()
temp_df_for_viz['Name'] = temp_df_for_viz['name_elements.forename'] + ' ' + temp_df_for_viz['name_
elements.surname']
temp_df_for_viz[['Name','company_number']].head(10).to_csv('data/viz/top_10_pscs.csv',index=False)
temp_df.head(10)
```

Out[44]:

				company_number
name_elements.forename	name_elements.surname	month_year_birth	address.postal_code	
Michael	Gleissner	1969-04-01	CT20 2RD	1193
Peter	Valaitis	1950-11-01	BS9 3BY	997
Waris	Khan	1979-02-01	W1G 9QR	639

Figure 2: A section of the Global Witness Jupyter notebook which constructs a table of individuals and accompanying counts based on them having the same first name, surname, month and year of birth and postcode.

Jupyter Notebooks have also proven particularly useful at Global Witness when there is need to monitor a specific dataset over time. For instance, in 2018 Global Witness wanted to establish how the corruption risk in the London property market had changed over a two year period.[15] They acquired a new snapshot of from the land registry of properties owned by foreign companies and re-used and published a notebook we had developed for the same purpose two years previously (Figure 2).[16] This yielded comparable results with minimal overhead. The notebook has an additional advantage in this context too: it allowed Global Witness to show its methodology in the absence of being able to re-publish the underlying source data which, at the time of analysis, had certain licensing restrictions. This is something very difficult to do in a spreadsheet-based workflow. Of course, the most effective way of accounting for your method will always be to publish the raw data used. However, journalists often use data that cannot be re-published for reasons of copyright, privacy or source protection.

While literate programming environments can clearly enhance the accountability and reproducibility of a journalist's data work, alongside other benefits, there are some important limitations.

One such limitation is that to re-produce (rather than just follow or "virtually witness") an approach set out in a Jupyter Notebook or R Markdown document you need to know how to write, or at least run, code. The relatively nascent state of data journalism means that there is still a fairly small group of journalists, let alone general consumers of journalism, who can code. This means that it is unlikely that the Github repositories of newspapers will receive the same level of scrutiny as say peer reviewed code referenced in an academic journal where larger portions of the community can actually interrogate the code itself. Data journalism may therefore be more prone to hidden errors in code itself when compared to research with a more technically literate audience. As Jeff Harris points out, it probably won't be long before we see programming corrections published by media outlets in much the same way as traditional that factual errors are published.[17] It is worth noting in this context that tools like Workbench (which is also mentioned in Jonathan Stray's chapter in this book) are starting to be developed for journalists, which promise to deliver some of the functionality of literate programming environments without the need to write or understand any code[18].

At this point it is also worth considering whether the new mechanisms for accountability in journalism may not just be new means through which a pre-existing "public" can scrutinise methods, but indeed play a role in the formation of new types of "publics". This is a point made by Andrew Barry in his essay, *Transparency as a political device*:

> "Transparency implies not just the publication of specific information; it also implies the formation of a society that is in a position to recognize and assess the value of – and if necessary to modify – the information that is made public. The operation of transparency is addressed to local witnesses, yet these witnesses are expected to be properly assembled, and their presence validated. There is thus a circular relation between the constitution of political assemblies and accounts of the oil economy – one brings the other into being. Transparency is not just intended to make information public, but to form a public which is interested in being informed"[19]

The methods elaborated on above for accounting for data journalistic working in themselves may play a role in the emergence of new groups of more technically aware publics that wish to scrutinise and hold reporters to account in ways not previously possible before the advent and use of technologies like literate programming environments in the journalistic context.

This idea speaks to some of Global Witness's work on data literacy in order to enhance the accountability of the extractives sector. Landmark legislation in the European Union that forces extractives companies to publish project-level payments to governments for oil, gas and mining projects, an area highly vulnerable to corruption, has opened the possibility for far greater scrutiny of where these revenues actually accumulate. However, Global Witness, and other advocacy groups within the Publish What You Pay coalition have long observed that there is no pre-existing "public" which could immediately play this role. As a result, Global Witness and others have developed resources and training programmes to assemble journalists and civil society groups in resource rich countries who can be supported in developing the skills to use this data to more readily hold companies to accounts. One component to this effort has been the development and publication of specific methodologies for red flagging suspicious payment reports that could be corrupt.[20]

Literate programming environments are currently a promising means through which data journalists are making their methodologies more transparent and accountable. While data will always remain open to multiple interpretations, technologies that make a reporter's assumptions explicit and their methods reproducible are valuable. They aid collaboration and open up an increasingly technical discipline to scrutiny from various publics. Given the current crisis of trust in journalism, a wider embrace of reproducible approaches may be one important way in which data teams can maintain their credibility.

Works Cited

Jacob Harris, 'Distrust Your Data', *Source*, 22 May 2014.

Ben Leather and Billy Kyte, 'Defenders: Methodology', *Global Witness*, 13 July 2017.

Donald Knuth, 'Literate Programming', Computer Science Department, Stanford University, Stanford, CA 94305, USA, 1984.

Andrew Barry, 'Transparency as a political device In: Débordements: Mélanges offerts à Michel Callon', Paris: Presses des Mines, 2010.

Carmen M. Reinhart and Kenneth S. Rogoff, 'Growth in a Time of Debt', *The National Bureau of Economic Research*, December 2011.

Donald E. Knuth, 'Literate Programming', Stanford, California: *Center for the Study of Language and Information*, 1992.

Steven Shapin and Simon Schaffer, 'Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life', Princeton University Press, 1985.

Richard Holmes and Jeremy Singer-Vine, 'Danger and Despair Inside Cambian Group, Britain's Largest Private Child Care Home Provider', BuzzFeed News, 26 July 2018.

Naomi Hirst and Sam Leon, 'Two Years On, We're Still in the Dark About the UK's 86,000 Anonymously Owned Homes', Global Witness, 7 December 2017.

Jacob Harris, 'The Times Regrets the Programmer Error', Source, 19 September 2013.

Global Witness, 'Finding the Missing Millions', 9 August 2018.

Ways of Doing Transparency in Data Journalism

This chapter is launching soon

Data Journalism: What's Feminism Got to Do With It?

This chapter is launching soon

Making Algorithms Work for Reporting

This chapter is launching soon

Coding in the Newsroom

This chapter is launching soon

Computational Reasoning at Full Fact and Urbs Media

This chapter is launching soon

Data Methods in Journalism

This chapter is launching soon

Exploring Relationships with Graph Databases

This chapter is launching soon

Text as Data: Finding Stories in Corpora

This chapter is launching soon

Online Devices and their Research Affordances for Data Investigations

This chapter is launching soon

How ICIJ Deals with Huge Data Dumps like the Panama and Paradise Papers

This chapter is launching soon

Data Visualisations: Newsroom Trends and Everyday Engagements

Written by: Helen Kennedy William Allen Rosemary Lucy Hill Martin Engebretsen Andy Kirk Wibke Weber

This chapter looks at both the production of data visualizations (henceforth "dataviz") in newsrooms and audiences' everyday engagements with dataviz, drawing on two separate research projects. The first is *Seeing Data*, which explored how people make sense of data visualizations, and the second is *INDVIL*, which explored dataviz as a semiotic, aesthetic and discursive resource in society.[1] The chapter starts by summarizing the main findings of an *INDVIL* sub-project focusing on dataviz in the news, in which we found that dataviz are perceived in diverse ways and deployed for diverse purposes. It then summarizes our main findings from *Seeing Data*[2], where we also found great diversity, this time in how audiences make sense of dataviz. This diversity is important for the future work of both dataviz researchers and practitioners.

Data Visualization in Newsrooms: Trends and Challenges

How is data visualization being embedded into newsroom practice? What trends are emerging, and what challenges are arising? To answer these questions, in 2016 and 2017 we undertook 60 interviews in 26 newsrooms across six European countries: Norway (NO), Sweden (SE), Denmark (DK), Germany (DE), Switzerland (CH), and the United Kingdom (UK). Interviewees in mainstream, online news media organisations included editorial leaders, leaders of specialist data visualization teams, data journalists, visual journalists, graphic/data visualization designers and developers (although some didn't have job titles, a sign in itself that this is a rapidly emerging field). We present some highlights from our research here.

Changing journalistic storytelling

The growing use of data visualization within journalism means that there is a shift from writing as the main semiotic mode to data and visualization as central elements in journalistic storytelling. Many interviewees stated that data visualization is the driving force of a story, even when it is a simple graphic or diagram.

> "The reader stats tell us that when we insert a simple data visualization in a story, readers stay on the page a little longer." (SE)

Dataviz are used with a broad range of communicative intentions, including: "to offer insight" (UK), "to explain more easily" (SE), "to communicate clearly, more clearly than words can" (UK), "to tell several facets in detail, which in text is only possible in an aggregated form" (DE), to make stories "more accessible" (DK), "to reveal deplorable state of affairs" (CH), "to help people understand the world" (UK). Data visualisation is used to emphasise a point, to add empirical evidence, to enable users to explore datasets, as aesthetic attraction to stimulate interest, and to offer entry into unseen stories.

These changes are accompanied by the emergence of multi-skilled specialist groups within newsrooms, with data and dataviz skills prioritized in new recruits. But there are no patterns in the organization of dataviz production within newsrooms – in some, it happens in data teams, in others, in visual teams (one of our dataviz designer interviewees was also working on a virtual reality project at the time of the interview) and elsewhere, in different teams still. And just as new structures are emerging to accommodate this newly proliferating visual form, so too newsroom staff need to adapting to learn new tools, in-house and commercial, develop new skills and understand how to communicate across teams and areas of expertise in order to produce effective data stories.

The 'mobile first' mantra and its consequences

Widespread recognition that audiences increasingly consume news on small, mobile screens has led to equally widespread adoption of a 'mobile first' mantra when it comes to producing dataviz in newsrooms. This means a turn away from the elaborate and interactive visualizations that characterized the early days of dataviz in the news, to greater simplicity and linearity, or simple visual forms with low levels of interactivity. This has led to a predominance of certain chart types, such as bar charts and line charts, and to the advent of *scrollytelling*, or stories that unfold as users scroll down the page, with the visualizations that are embedded in the article appearing at the appropriate time. Scrolling also triggers changes in visualizations themselves, such as zooming out.

> "Often in our stories we use the scrolling technique. It is not necessary to click but to scroll, if you scroll down, something will happen in the story." (DE)

Tools to automate dataviz production and make it possible for journalists who are not dataviz experts to produce them also result in the spread of simplified chart forms. Nonetheless, some interviewees are keen to educate readers by presenting less common chart types (a scatterplot, for example) accompanied with information about how to make sense of them. Some believe that pictures can also present data effectively – a Scandinavian national tabloid represented the size of a freight plane by filling it with 427,000 pizzas. Others recognize the value of animation, for example to show change over time, or of experimenting with zoomability in visualizations.

The social role of journalism

Linking a dataviz to a data source, providing access to the raw data and explaining methodologies are seen by some participants as ethical practices which create transparency and counterbalance the subjectivity of selection and interpretation which, for some, is an inevitable aspect of visualizing data. Yet for others, linking to data sources means giving audiences 'all of the data' and conflicts with the journalistic norm of identifying and then telling a story. For some, this conflict is addressed by complex processes of sharing different elements of data and process on different platforms (Twitter, Pinterest, Github).

This leads data journalists and visualization designers to reflect on how much data to share, their roles as fact providers and their social role more generally. Data journalist Paul Bradshaw sums this up on his blog:

> "How much responsibility do we have for the stories that people tell with our information? And how much responsibility do we have for delivering as much information as someone needs?"[3]

Former *Guardian* editor Alan Rusbridger (2009) raised a similar question about the social role of journalism when he points to the range of actors who do what journalism has historically done – that is, act as a gatekeeper of data and official information (eg FixMyStreet and TheyWorkForYou in the UK. He concluded 'I don't know if that is journalism or not. I don't know if that matters' (cf Baack 2018). Some of our interviewees work on large-scale projects similar to those discussed by Rusbridger – for example, one project collated all available data relating to schools in the UK and made this explorable by postcode to inform decision-making about school preference. So the question of what counts of journalism in the context of widespread data and dataviz is not easy to answer.

What's more, sharing datasets assumes that audiences will interact with them, yet studies indicate that online interactivity is as much a myth as a reality, with the idealized image of an active and motivated explorer of a visualized dataset contrasting with the more common quick and scrolling reader of news (eg Burmester et al 2010). Similarly, a study of data journalism projects submitted to the Nordic Data Journalism Awards concludes that interactive elements often offer merely an *illusion* of interactivity, as most choices already are made or predefined by the journalists (Appelgren 2017). This again calls into question the practice of sharing 'all of the data' and raises questions about the changing social role of journalism.

Trust, truth and visualisations 'in the wild'

Other elements of the process of visualizing data raise issues of trust and truth and also relate to how journalists think about the social role of journalism. One aspect of dataviz work that points to these issues is how journalists working with data visualization think about data and their visual representation. Some see it as a form of truth-telling, others as a process of selection and interpretation, and others still believe that shaping data visualizations through choices is a way of revealing a story and so is precisely what journalists should do. These reflections highlight the relationship between (dis)trust and presentation, and between perspective and (un)truthfulness.

In our current, so-called 'post-truth' context, in which audiences are said to have had enough of facts, data and experts and in which fake news circulates quickly and widely, our participants were alert to the potential ways in which audiences might respond to their data visualizations, which might include accepting naively, refuting skeptically, decontextualizing through social sharing, or even changing and falsifying. They felt that journalists increasingly need 'soft knowledge of internet culture', as one (UK) participant put it. This includes an understanding of how online content might be more open to interrogation than its offline equivalent, and of how data visualizations may be more likely to circulate online than text, floating free of their original contexts as combinations of numbers and pictures 'in the wild' (Espeland and Sauder 2007). This in turn requires understanding of strategies that might address these dangers, such as embedding explanatory text into a visualisation file so that the image cannot be circulated without the explanation. These issues, alongside concern about audiences' data and visualization literacy, inform and reshape journalists' thinking about their audiences.

How Do People Engage with Data Visualizations?

In this section, we look at dataviz in the news from the perspective of the audience. How do audiences engage with and make sense of the visualisations that they encounter in news media? Data journalists are often too busy to attend to this question. Data visualisation researchers don't have this excuse, but nevertheless rarely focus their attention on what end users think of the visualisations that they see.

Enter *Seeing Data*, a research project which explored how people engage with the data visualisations that they encounter in their everyday lives, often in the media. It explored the factors that affect engagement and what this means for how we think about what makes a visualisation effective. On *Seeing Data* we used focus groups and interviews to explore these questions, to enable us to get at the attitudes, feelings and beliefs that underlie people's engagements with dataviz. 46 people participated in the research, including a mix of participants who might be assumed to be interested in data, the visual, or migration (which was the subject of a number of the visualisations that we showed them) and so 'already engaged' in one of the issues at the heart of our project and participants about whom we could not make these assumptions.

In the focus groups, we asked participants to evaluate eight visualizations, which we chose (after much discussion) because they represented a diversity of subject matters, chart types, original media sources, formats and aimed either to explain or to invite exploration. Half of the visualisations were taken from journalism (BBC; *The New York Times*; *The Metro*, a freely distributed UK newspaper; and *Scientific American* magazine). Others came from organisations which visualise and share data as part of their work (the Migration Observatory at the University of Oxford; the Office for National Statistics (ONS) in the UK; and the Organisation for Economic Co-operation and Development (OECD).

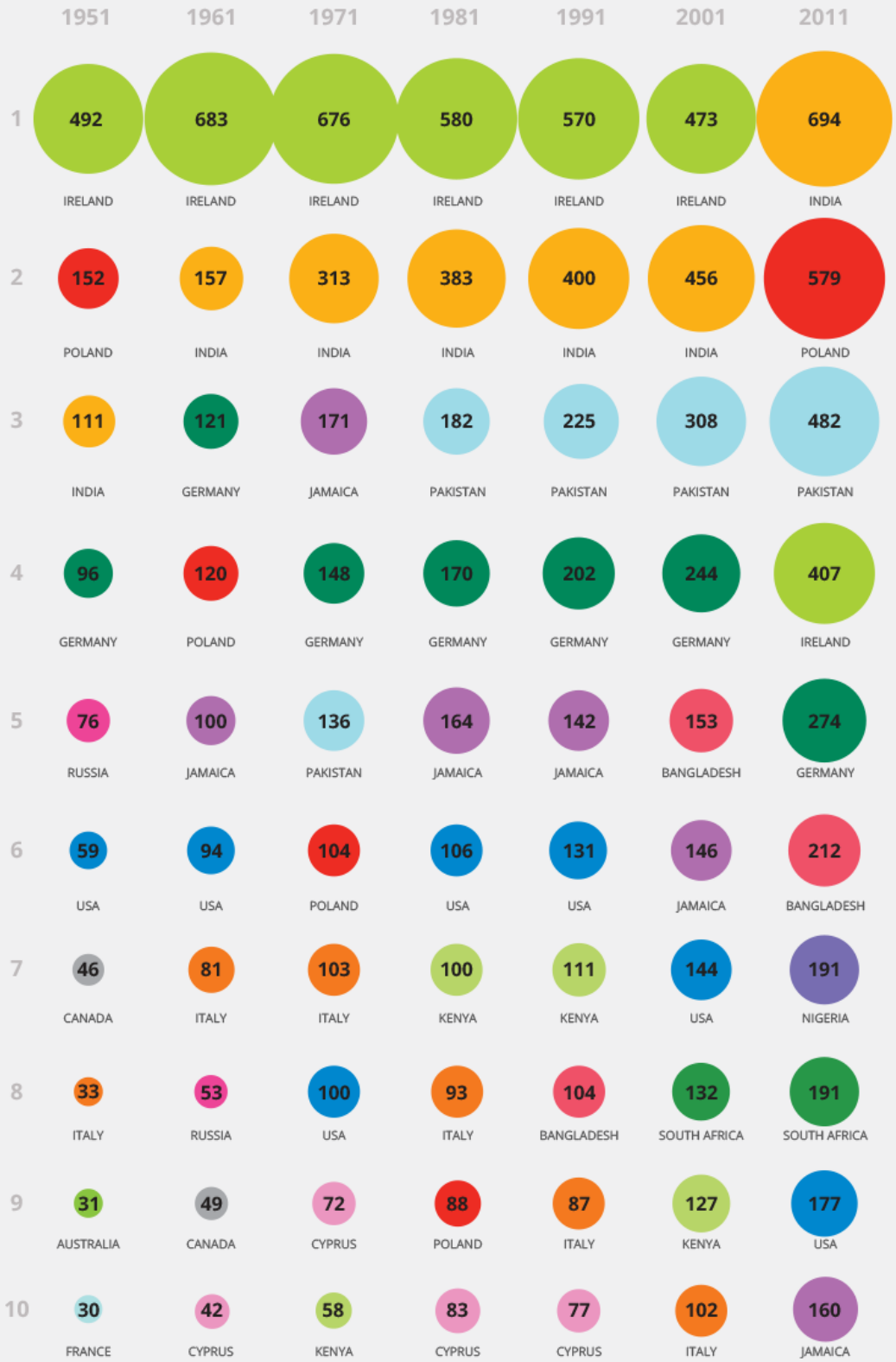
After the focus groups, seven participants kept diaries for a month, to provide us with further information about encounters with visualizations 'in the wild' and not chosen by us.

Non-UK born census populations 1951 - 2011

13% (7.5 MILLION) OF RESIDENTS IN ENGLAND AND WALES WERE BORN OUTSIDE THE UK, 2011

TOP TEN NON-UK COUNTRIES OF BIRTH

NUMBERS ARE IN THOUSANDS



ALL OTHER NON-UK BORN

1951 1961 1971 1981 1991 2001 2011

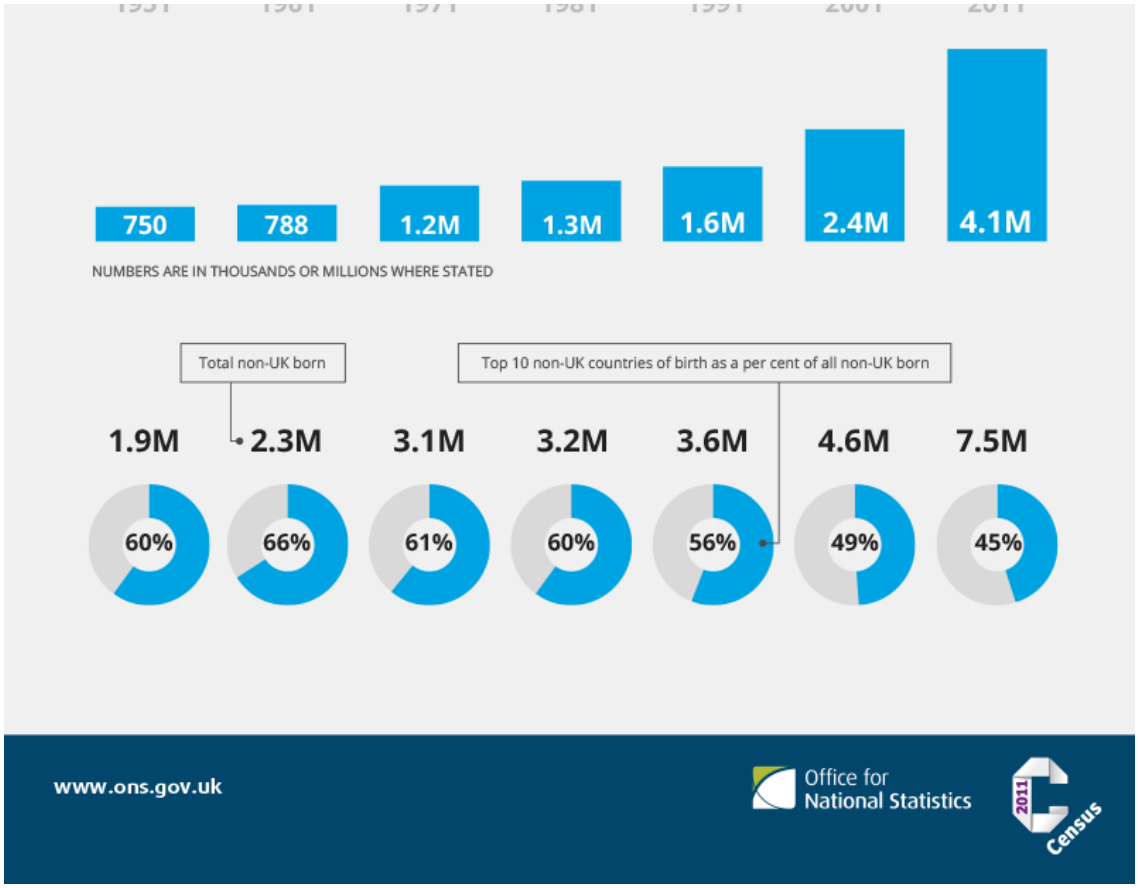


Figure 1: Non-UK Born Census Populations 1951-2011, Office for National Statistics)

Figure 2: Migration In The Census, produced for The Migration Observatory, University of Oxford

Factors which affect dataviz engagement

Subject matter

Visualizations don't exist in isolation from the subject matter that they represent. When subject matter spoke to participants' interest, they were engaged – for example with Civil Society professionals who were interested in issues relating to migration and therefore in migration visualizations. In contrast, one participant (who was male, 38, white British, an agricultural worker) was not interested in any of the visualizations we showed him in the focus group or confident to spend time looking. However, his lack of interest and confidence and his mistrust of the media (he said he felt they try 'to confuse you') did not stop him from looking at visualizations completely: he told us that when he came across visualizations in *The Farmer's Guide*, a publication he read regularly because it speaks to his interests, he would take the time to look at them.

Source or media location

The source of visualizations is important: it has implications for whether users trust them. Concerns about the media setting out to confuse were shared by many participants and led some to view visualizations encountered within certain media as suspect. In contrast, some participants trusted migration visualizations which carried the logo of the University of Oxford, because they felt that the 'brand' of this university invokes quality and authority. But during the diary keeping period, a different picture emerged. Participants tended to see visualizations in their favoured media, which they trusted, so they were likely to trust the visualizations they saw there too. One participant (male, 24, white British agricultural worker), who reads *The Daily Mail*, demonstrated this when he remarked in his interview that 'you see more things wrong or printed wrong in *The Sun* I think'. Given the ideological similarities between these two publications, this comment points to the importance of media location in dataviz engagement.

Beliefs and opinions

Participants trusted the newspapers they regularly read and therefore trusted the visualizations in these newspapers, because both the newspapers and the visualizations often fitted with their views of the world. This points to the importance of beliefs and opinions in influencing how and whether people take time to engage with particular visualizations. Some participants said they liked visualizations that confirmed their beliefs and opinions. But it is not just when visualizations confirm existing beliefs that beliefs matter. One participant (male, 34, white British, IT worker) was surprised by the migration data in an ONS visualization in Figure 1. He said that he had not realised how many people in the UK were born in Ireland. This data questioned what he believed and he enjoyed that experience. Some people like, or are interested in, data in visualizations that call into question existing beliefs, because they provoke and challenge horizons. So beliefs and opinions matter in this way too.

Time

Engaging with visualizations is seen as work by people for whom doing so does not come easily. Having time available is crucial in determining whether people are willing to do this 'work'. Most participants who said they lacked time to look at visualizations were women, and they put their lack of time down to work, family and home commitments. One working mother talked about how her combined paid and domestic labour were so tiring that when she finished her day, she didn't want to look at news, and that included looking at visualizations. Such activities felt like 'work' to her, and she was too tired to undertake them at the end of her busy day. An agricultural worker told us in an email that his working hours were very long and this impacted on his ability to keep his month-long diary of engagements with visualizations after the focus group research.

Confidence and skills

Audiences need to feel that they have the necessary skills to decode visualizations, and many participants indicated a lack of confidence in this regard. A part-time careers advisor said of one visualization: 'It was all these circles and colours and I thought, that looks like a bit of hard work; don't know if I understand'. Many of our

participants expressed concern about their lack of skills, or they demonstrated that they did not have the required skills, whether these were visual literacy skills, language skills, mathematical and statistical skills (like knowing how to read particular chart types), or critical thinking skills.

Emotions

Although last in our list, a major finding from our research was the important role that emotions play in people's engagements with data visualisations. A broad range of emotions emerged in relation to engagements with dataviz, including pleasure, anger, sadness, guilt, shame, relief, worry, love, empathy, excitement, offence. Participants reported emotional responses to: visualisations in general; represented data; visual style; the subject matter of data visualisations; the source or original location of visualisations; their own skill levels for making sense of visualisations.

For example, two civil society professionals used strong language to describe their feelings when they looked at the visualizations of migration in the UK shown in Figure 2. The data caused them to reflect on how it must feel to be a migrant who comes to the UK and encounters the anti-immigration headlines of the media. They described themselves as feeling 'guilty' and 'ashamed' to be British.

Other participants had strong emotional responses to the visual style of some visualizations. A visualization of film box office receipts by *The New York Times*[4] divided participants, with some drawn to its aesthetic and some put off by it:

It was a pleasure to look at this visual presentation because of the coordination between the image and the message it carries.

Frustrated. It was an ugly representation to start with, difficult to see clearly, no information, just a mess.

What this means for making effective visualisations

A broad range of understandings of what makes a visualisation effective emerged from our research. Visualizations in the media that are targeted at non-specialists might aim to persuade, for example. They all need to attract in order to draw people in, if they are to commit time to finding out about the data on which the visualization is based. Visualizations might stimulate particular emotions, which inspire people to look longer, deeper or further. They might provoke interest, or the opposite. An effective visualisation could:

- Provoke questions/desire to engage in discussions with others
- Create empathy for other humans in the data
- Generate enough curiosity to draw the user in
- Reinforce or back up existing knowledge
- Provoke surprise
- Persuade or change minds
- Present something new
- Lead to new confidence in making sense of dataviz
- Present data useful for one's own purposes
- Enable an informed or critical engagement with a topic
- Be a pleasurable experience
- Provoke a strong emotional response.

What makes a visualisation effective is fluid – no single definition applies across all dataviz. For example, being entertained by a visualization is relevant in some contexts, but not others. Visualizations have various objectives: to communicate new data; to inform a general audience; to influence decision-making; to enable exploration and analysis of data; to surprise and affect behaviour. The factors that affect engagement which we identified in our research should be seen as *dimensions* of effectiveness, which carry different weight in relation to different visualizations, contexts and purposes. Many of these factors lie outside of the control of data visualisers, as they relate to consuming, not producing, visualizations. In other words, whether a visualization is effective depends in large part on how, by whom, when and where it is accessed. Sadly, our research doesn't suggest a simple checklist which guarantees the production of universally effective visualizations. However, if we want accessible and effective data visualisations, it's important that journalists working with data visualisation engage with these findings.

Works Cited:

Stefan Baack, '*Knowing What Counts: how data activists and data journalists appropriate and advance datafication*', PhD thesis, *University of Groningen*, 238 p., 2018

Paul Bradshaw, 'No I'm not abandoning the term "storytelling", Alberto – just the opposite (and here's why)', *Online Journalism Blog*, 14 September 2017.

Wendy Nelson Espeland and Michael Sauder, 'Rankings and reactivity: how public measures recreate social worlds' *American Journal of Sociology*, 113(1), 1-40, 2007.

Alan Rusbridger, Why Journalism Matters', *Media Standards Trust Series*, August 2010.

Matthew Bloch, Shan Carter and Amanda Cox, 'The Edd and Flow of Movies: Box Office Receipts 1989 - 2008', *The New York Times*, 23 February 2008.

Kennedy, Hill, Allen and Kirk, 'Engaging with (big) data visualizations: Factors that affect engagement and resulting new definitions of effectiveness', *First Monday* 21:11, (2016).

Kennedy and Hill, 'The Feeling of Numbers: Emotions in Everyday Engagements with Data and Their Visualisation', *Sociology* 52:4, 2018, pp. 830-848.

Searchable Databases as a Journalistic Product

Written by: Zara Rahman

A still emerging journalistic format is the searchable online database – a web interface that gives access to a dataset, by newsrooms. This format is not new, but its appearance among data journalism projects is still relatively scarce.[1]

In this article, 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.[2] It is worth noting, too, that though 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 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 personalisation element is ProPublica's 'Dollars for Docs', which compiled data on payments to doctors and teaching hospitals that were made by pharmaceutical and medical device companies.[3] This topic and approach was mirrored by Correctiv and Spiegel Online to create Euros für Ärzte, who created 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, where the goal was to increase the accessibility of said data so that readers would be able to search it for themselves to, presumably, 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.[4]

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 2016 and 2015.[5] Their database used a variety of online reporting, as well as reader-input.

Another type of database involves taking an existing set of data and creating an interface where a reader can generate a report based on particular 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. The UK-based Bureau of Investigative Journalism compiles data from various sources gathered through their investigations, within a database called Drone Warfare.[6] The database created allows readers to select particular countries covered and the time frame, to create a report with visualisations summarising 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, OCCRP maintain and update OCCRP Data which allows viewers to search over 19 million public records.[8] In both these cases, the primary user of the tools is not envisioned to be the average reader, but instead journalists or researchers who would then carry out further research on whatever information is found using these tools.

Following are some of 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, they decided that member companies must publish payments to healthcare professionals and organisations in the countries they operate starting in July 2016.[9] Inspired by ProPublica's "Dollars for Docs" project, non-profit German investigative newsroom Correctiv decided to collect 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.[10] They named the investigation "Euros für Ärzte" ("euros for doctors").

In collaboration with German national news outlet Spiegel Online, documents and data were gathered from around 50 websites and converted from different formats to consistent tabular data. This data was then 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.[11] 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, via their events, organisations 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 more meta 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 would 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 Spiegel Online received dozens of complaints and legal threats from doctors that appeared in the database. As the data came from public, albeit difficult to find, sources, the legal team of Spiegel Online decided to defer most complaints to the pharma companies and only adjust the database in case of changes at the source.

Technical considerations of building databases

For a newsroom considering how to make a dataset available and accessible to readers, there are various criteria to consider, such as size and complexity of the dataset, 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 ~20,000 recipients and their financial connections to pharma companies required a custom software solution. They 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. The data was stored in a relational database separate from the content database to separate concerns. In their case, having a process and interface to adjust entries in the live database was crucial as dozens of upstream data corrections came in after publication.

However, smaller datasets 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 frontend 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 datasets accessible are JavaScript-based web applications with access to the dataset via API. This setup lends 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 frontend.

Affordances offered by databases

Databases within, or alongside, a story, provide a number of new affordances for both readers, and for 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 the existing infrastructure and methodologies.

Databases also potentially offer increased optimisation 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.

Optimising 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 data[12] 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.

Firstly: 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 included in such a database would presumably have both the expectation of such a development, and adequate resources to take action, whereas a healthcare professional likely is not 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 expands upon this, noting that in the US, schools are becoming more segregated despite (or perhaps because of) an increase in data available about 'school performance'.[13] 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.

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

For example, the absence of an organisation from a database of political organisations involved in organised crime may not mean that the organisation does not take part in organised crime itself; it simply means that there was no data available about their actions. Michael Golebiewski and danah boyd 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"[14]. This type of

absence of data in an otherwise data-saturated space also maps closely to what Brooklyn-based artist and researcher Mimi Onuoha refers to as a "missing data set" and highlights the societal choices that go into collecting and gathering data.[15]

Thirdly: 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 is, 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 agreed 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 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 remove consent of 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 these approaches which prioritise the rights of people reflected in the data all the way through the investigation, from data gathering to publication, are a core part of optimising (data) journalism for trust.[16]

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.

- 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.
- 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.
- 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.
- Allocate enough resources for maintenance over time: Keeping the data and database software current 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.
- Observe how readers are using the database: trends in searches or use might provide leads for future stories and investigations.
- 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

Adrian Holovaty, ' A Fundamental Way Newspaper Sites Need to Change', *Writing*, 6 September 2006.

Mike Tigas, Ryann Grochowski Jones, Charles Ornstein, and Lena Groeger, 'Dollars for Doctors', *ProPublica*, 28 June 2018.

Claire Fontaine, 'Driving School Choice', *Data & Society: Points*, 20 April 2017.

Michael Golebiewski and Danah Boyd, 'Data Voids: Where Missing Data Can Be Easily Exploited', *Data & Society*, May 2018.

Jay Rosen, 'Optimizing Journalism for Trust', *The Correspondent*, 14 April 2018.

The Web as a Medium for Data Visualisation

This chapter is launching soon

Developments in the Field of News Graphics

This chapter is launching soon

Understanding Conflicts with Data Comics

This chapter is launching soon

Data Journalism for TV and Radio

This chapter is launching soon

Telling Stories with the Social Web

Written by: Lam Thuy Vo

We have become the largest producers of data in history. Almost every click online, each swipe on our tablets and each tap on our smartphone produces a data point in a virtual repository. Facebook generates data on the lives of more than 2 billion people. Twitter records the activity of more than 330 million monthly users. One MIT study found that the average American office worker was producing 5GB of data each day[2]. That was in 2013 and we haven't slowed down. As more and more people conduct their lives online, and as smartphones are penetrating previously unconnected regions around the world, this trove of stories is only becoming larger.

A lot of researchers tend to treat each social media user like they would treat an individual subject — as anecdotes and single points of contact. But to do so with a handful of users and their individual posts is to ignore the potential of hundreds of millions of others and their interactions with one another. There are many stories that could be told from the vast amounts of data produced by social media users and platforms because researchers and journalists are still only starting to acquire the large-scale data-wrangling expertise and analytical techniques needed to tap them.

Recent events have also shown that it is becoming crucial for reporters to gain a better grasp of the social web. The Russian interference with the 2016 U.S. presidential elections and Brexit; the dangerous spread of anti-Muslim hate speech on Facebook in countries in Europe and in Myanmar; and the heavy-handed use of Twitter by global leaders — all these developments show that there's an ever-growing need to gain a competent level of literacy around the usefulness and pitfalls of social media data in aggregate.

How can journalists use social media data?

While there are many different ways in which social media can be helpful in reporting, it may be useful to examine the data we can harvest from social media platforms through two lenses.

First, social media can be used as a proxy to better understand individuals and their actions. Be it public proclamations or private exchanges between individuals — a lot of people's actions, as mediated and disseminated through technology nowadays, leave traces online that can be mined for insights. This is particularly helpful when looking at politicians and other important figures, whose public opinions could be indicative of their policies or have real-life consequences like the plummeting of stock prices or the firing of important people.

Secondly, the web can be seen as an ecosystem in its own right in which stories take place on social platforms (albeit still driven by human and automated actions). Misinformation campaigns, algorithmically skewed information universes, and trolling attacks are all phenomena that are unique to the social web.

How is social data used for journalistic stories

Instead of discussing these kinds of stories in the abstract, it may be more helpful to understand social media data in the context of how it can be used to tell particular stories. The following sections discuss a number of journalistic projects that made use of social media data.

Understanding public figures: social media data for accountability reporting

For public figures and everyday people alike, social media has become a way to address the public in a direct manner. Status updates, tweets and posts can serve as ways to bypass older projection mechanisms like interviews with the news media, press releases or press conferences.

For politicians, however, these public announcements — these projections of their selves — may become binding statements and in the case of powerful political figures may become harbingers for policies that need yet to be put in place.

Because a politician's job is partially to be public-facing, researching a politician's social media accounts can help us better understand their ideological mindset. For one story, my colleague Charlie Warzel and I collected and analyzed more than 20,000 of Donald Trump's tweets to answer the following question: what kind of information does he disseminate and how can this information serve as a proxy for the kind of information he may consume?

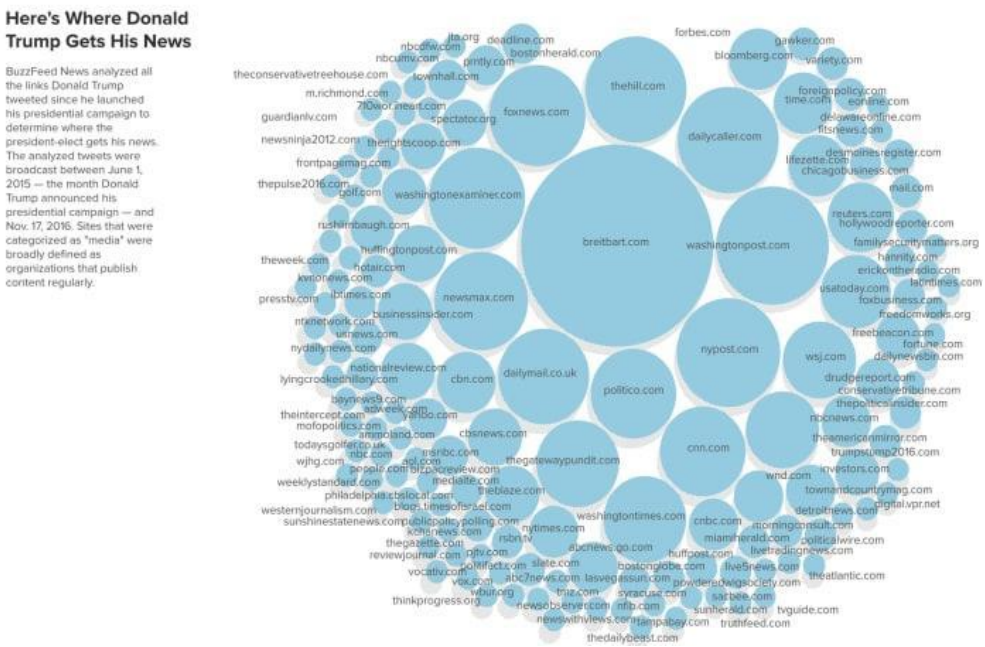


Figure 1: A snapshot of the media links that Trump tweeted during his presidential campaign

Social data points are not a full image of who we actually are, in part due to its performative nature and in part because these data sets are incomplete and so open to individual interpretation. But they can help as complements: President Trump's affiliation with Breitbart online, as shown above, was an early indicator for his strong ties to Steve Bannon in real life. His retweeting of smaller conservative blogs like The Conservative Tree House and News Ninja 2012 perhaps hinted at his distrust of "mainstream media."[3].

Tracing back human actions

While public and semi-public communications like tweets and open Facebook posts can give insights into how people portray themselves to others, there's also the kind of data that lives on social platforms behind closed walls like private messages, Google searches or geolocation data.

Christian Rudder, co-founder of OKCupid and author of the book Dataclysm had a rather apt description of this kind of data: these are statistics that are recorded of our behavior when we "think that no one is watching."

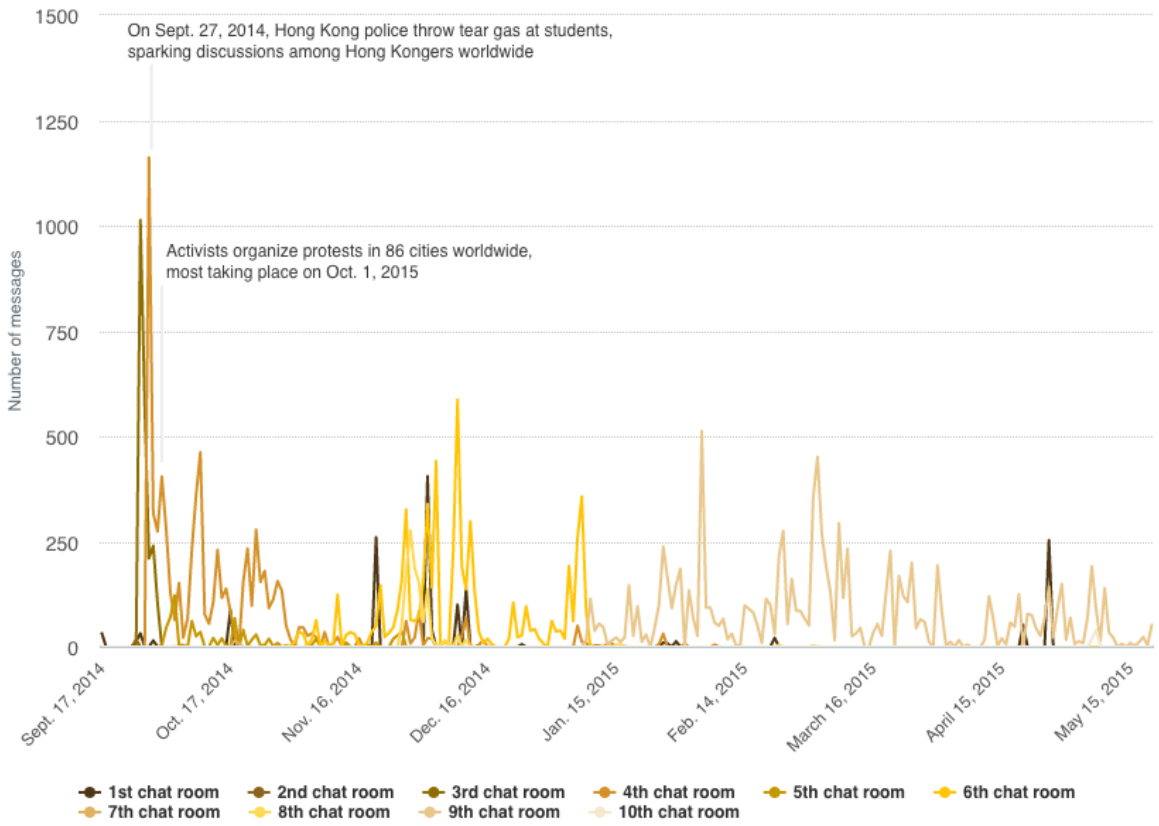
By virtue of using a social platform, a person ends up producing longitudinal data of their own behavior. And while it's hard to extrapolate much from these personal data troves beyond the scope of the person who produced them, this kind of data can be extremely powerful when trying to tell the story of one person. I often like to refer this kind of approach as a Quantified Selfie, a term Maureen O'Connor coined for me when she described some of my work.

Take the story of Jeffrey Ngo, for instance. When pro-democracy protests began in his hometown, Hong Kong, in early September of 2014, Ngo, a New York University student originally from Hong Kong, felt compelled to act. Ngo started to talk to other expatriate Hong Kongers in New York and in Washington, D.C. He ended up organizing protests in 86 cities across the globe and his story is emblematic of many movements that originate on global outrage about an issue.

For this Al Jazeera America story, Ngo allowed us to mine his personal Facebook history — an archive that each Facebook user can download from the platform[4]. We scraped the messages he exchanged with another core organizer in Hong Kong and found 10 different chat rooms in which the two and other organizers exchanged thoughts about their political activities.

The chart below (Figure 3) documents the ebbs and flows of their communications. First there's a spike of communications when a news event brought about public outrage — Hong Kong police throwing tear gas at peaceful demonstrators. Then there's the emergence of one chat room, the one in beige, which became the chat room in which the core organizers planned political activists well beyond the initial news events.

Jeffrey Ngo allowed Al Jazeera America to analyze his Facebook data to shed light on his political activities online. Below is a chart showing 10 chat rooms on Facebook that involved Leong, Ngo and other activists. Friends of friends would introduce one another on Facebook in these chat rooms, said Ngo. Soon chats became more formalized, with the ninth chat room being used by the most active leaders.



Source: Facebook data courtesy of Jeffrey Ngo / United for Democracy: Global Solidarity with Hong Kong Facebook group

Figure 3: United for Democracy: Global Solidarity with Hong Kong Facebook group. Source: Facebook data courtesy of Jeffrey Ngo.

Since most of their planning took place inside these chat rooms, we were also able to recount the moment when Ngo first met his co-organizer, Angel Yau. Ngo himself wasn't able to recall their first exchanges but thanks to the Facebook archive we were able to reconstruct the very first conversation Ngo had with Yau.

While it is clear that Ngo's evolution as a political organizer is that of an individual and by no means representative of every person who participated in his movement, it is, however, emblematic of the *kind* of path a political organizer may take in the digital age.

Phenomena specific to online ecosystems

Many of our interactions are moving exclusively to online platforms.

While much of our social behavior online and offline is often intermingled, our online environments are still quite particular because online human beings are assisted by powerful tools.

There's bullying for one. Bullying has arguably existed as long as humankind. But now bullies are assisted by thousands of other bullies who can be called upon within the blink of an eye. Bullies have access to search engines and digital traces of a person's life, sometimes going as far back as that person's online personas go. And they have the means of amplification — one bully shouting from across the hallway is not nearly as deafening as thousands of them coming at you all at the same time. Such is the nature of trolling.

Washington Post editor Doris Truong, for instance, found herself at the heart of a political controversy online. Over the course of a few days, trolls (and a good amount of people defending her) directed 24,731 Twitter mentions at her. Being pummeled with vitriol on the Internet can only be ignored for so long before it takes some kind of emotional toll.

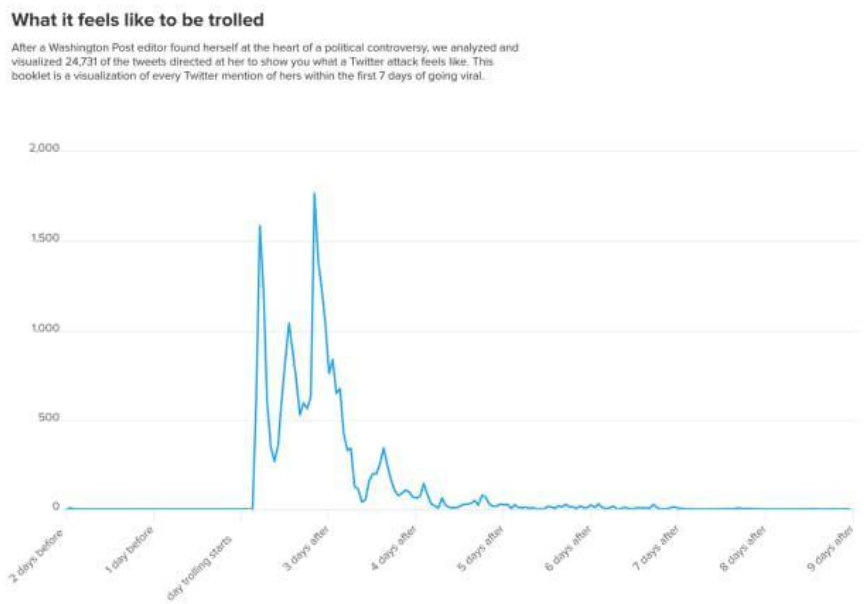


Figure 5: A chart of Doris Truong's Twitter mentions starting the day of the attack

Figure 5: *A chart of Doris Truong's Twitter mentions starting the day of the attack*[5]

Trolling, not unlike many other online attacks, have become problems that can afflict any person now - famous or not. From Yelp reviews of businesses that go viral — like the cake shop that refused to prepare a wedding cake for a gay couple — to the ways in which virality brought about the firing and public shaming of Justine Sacco, a PR

person who made an unfortunate joke about HIV and South Africans right before she took off on an intercontinental flight — many stories that affect our day to day take place online these days.

Information wars

The emergence and the ubiquitous use of social media has brought about a new phenomenon in our lives: virality.

Social sharing has made it possible for any kind of content to potentially be seen not just by a few hundred but by millions of people without expensive marketing campaigns or TV air time purchases.

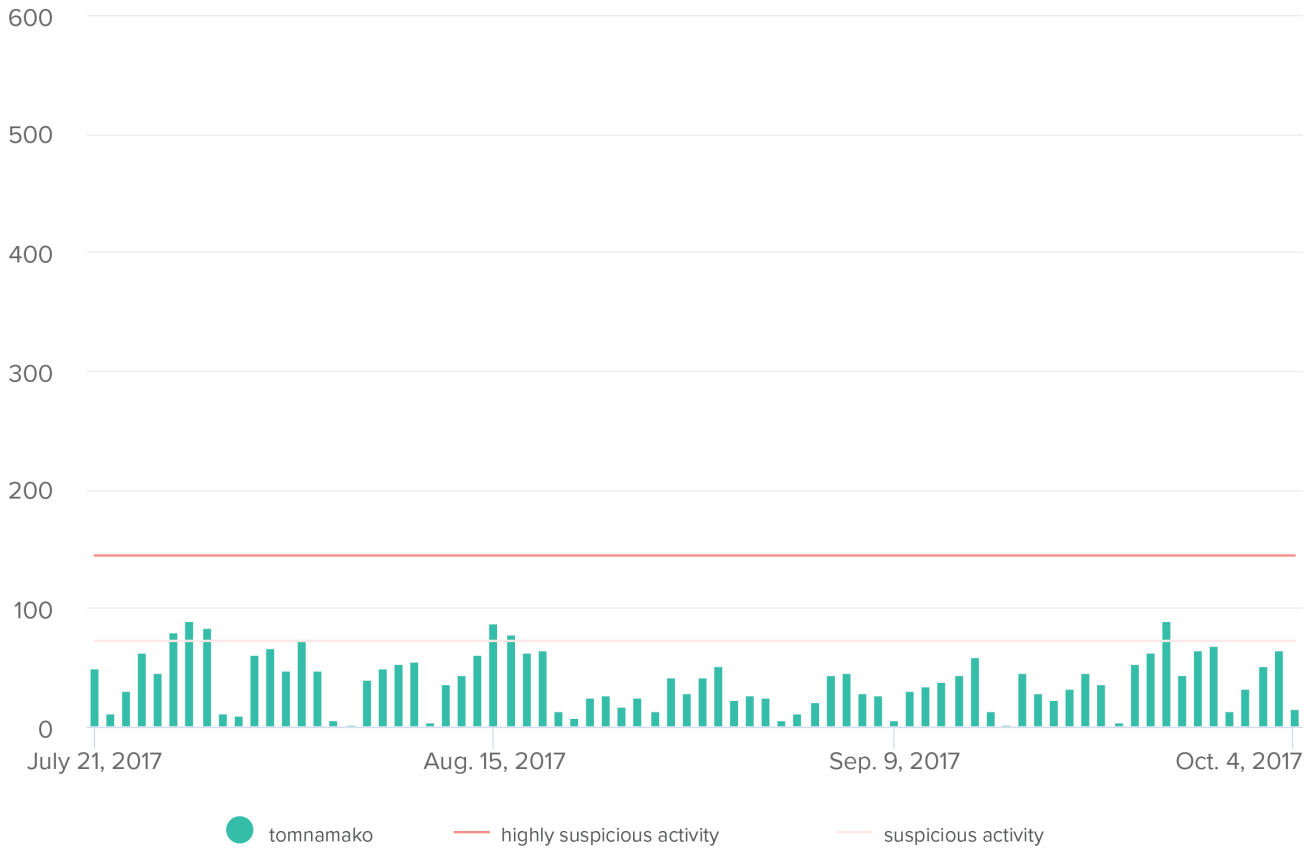
But what that means is that many people have also found ways to game algorithms with fake or purchased followers as well as (semi-)automated accounts like bots and cyborgs.[6]

Bots are not evil from the get-go: there are plenty of bots that may delight us with their whimsical haikus or self-care tips. But as Atlantic Council fellow Ben Nimmo, who has researched bot armies for years, told me for a BuzzFeed story: "[Bots] have the potential to seriously distort any debate [...] They can make a group of six people look like a group of 46,000 people."

The social media platforms themselves are at a pivotal point in their existence where they have to recognize their responsibility in defining and clamping down on what they may deem a "problematic bot." In the meantime, journalists should recognize the ever growing presence of non-humans and their power online.

For one explanatory piece about automated accounts we wanted to compare tweets from a human to those from a bot[7]. While there's no surefire way to really determine whether an account is operated through a coding script and thus is not a human, there are ways to look at different traits of a user to see whether their behavior may be suspicious. One of the characteristics we decided to look at is that of an account's activity.

For this we compared the activity of a real person with that of a bot. During its busiest hour on its busiest day the bot we examined tweeted more than 200 times. Its human counterpart only tweeted 21 times.



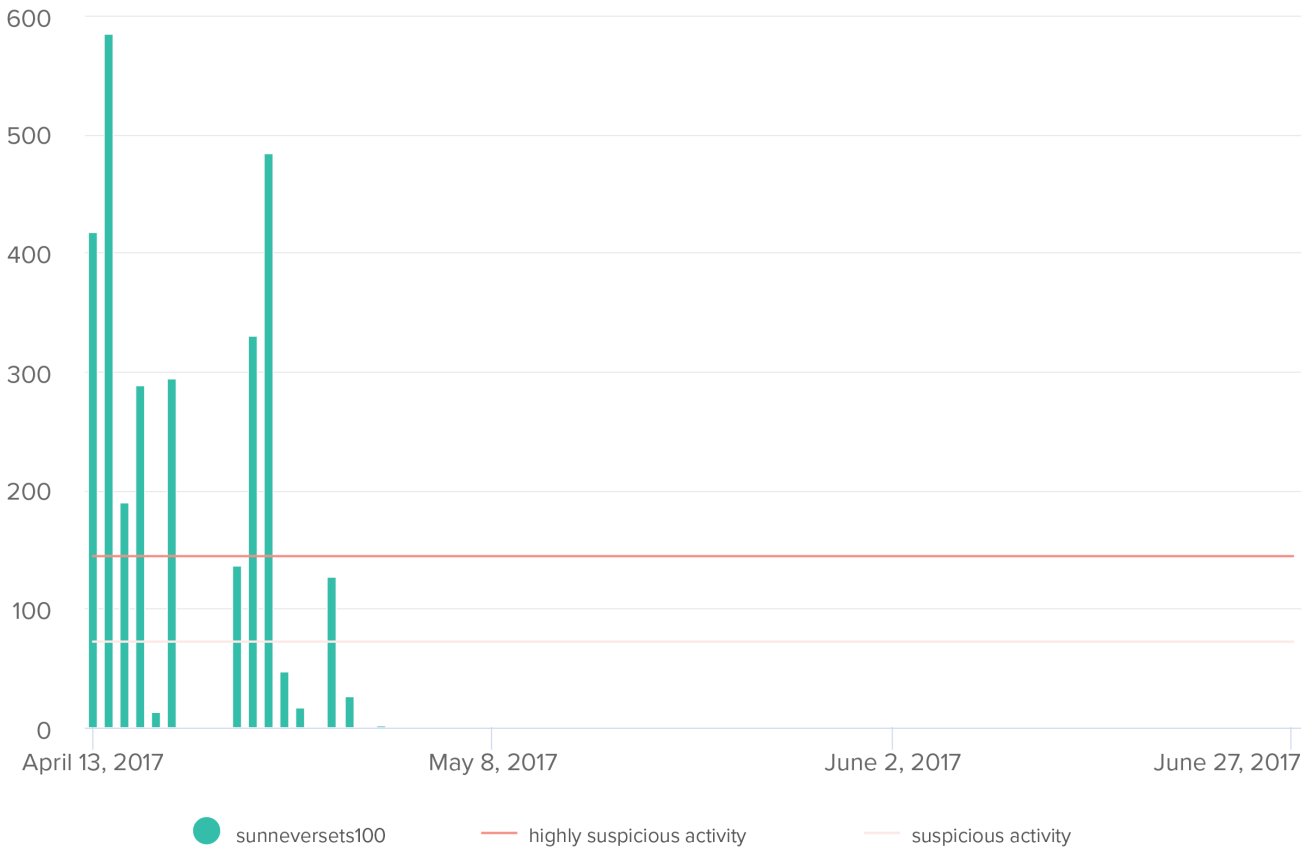


Figure 6: BuzzFeed News compared one of its own human editors' Twitter data, @tomnamako, and the data of several accounts that displayed bot-like activity to highlight their differences in personas and behavior. The first chart above shows that the BuzzFeed News editor's last 2,955 tweets are evenly distributed throughout several months. His daily tweet count barely ever surpassed the mark of 72 tweets per day, which the Digital Forensics

Research Lab designated as a suspicious level of activity. The second chart shows the bot's last 2,955 tweets. It was routinely blasting out a suspicious number of tweets, hitting 584 in one day. Then, it seems to have stopped abruptly.

How to harvest social data

There are broadly three different ways to harvest data from the social web: APIs, personal archives and scraping.

The kind of data that official channels like API data streams provide is very limited. Despite harboring warehouses of data on consumers' behavior, social media companies only provide a sliver of it through their APIs (for Facebook, researchers were once able to get data for public pages and groups but are no longer able to mine that kind of data after the company implemented restrictions on the availability of this data in response to the Cambridge Analytica. For Twitter, this access is often restricted to a set number of tweets from a user's timeline or to a set time frame for search).

Then there are limitations on the kind of data users can request of their own online persona and behavior. Some services like Facebook or Twitter will allow users to download a history of the data that constitutes their online selves—their posts, their messaging, or their profile photos—but that data archive won't always include everything each social media company has on them either.

For instance, users can only see what ads they've clicked on going three months back, making it really hard for them to see whether they may or may not have clicked on a Russia-sponsored post.

Last but not least, extracting social media data from the platforms through scraping is often against the terms of service. Scraping a social media platform can get users booted from a service and potentially even result in a lawsuit[8].

For social media platforms, suing scrapers may make financial sense. A lot of the information that social media platforms gather about their users is for sale—not directly, but companies and advertisers can profit from it through ads and marketing. Competitors could scrape information from Facebook to build a comparable platform, for instance. But lawsuits may inadvertently deter not just economically motivated data scrapers but also academics and journalists who want to gather information from social media platforms for research purposes.

This means that journalists may need to be more creative in how they report and tell these stories journalists may want to buy bots to better understand how they act online, or reporters may want to purchase Facebook ads to get a better understanding of how Facebook works[9].

Whatever the means, operating within and outside of the confines set by social media companies will be a major challenge for journalists as they are navigating this ever-changing cyber environment.

What social media data is *not* good for

It seems imperative to better understand the universe of social data also from a standpoint of its caveats.

Understanding who *is* and who *isn't* using social media

One of the biggest issues with social media data is that we cannot assume that the people we hear on Twitter or Facebook are representative samples of broader populations offline.

While there are a large number of people who have a Facebook or Twitter account, journalists should be wary of thinking that the opinions expressed online are those of the general population. As a Pew study from 2018 illustrates, usage of social media varies from platform to platform[10]. While more than two thirds of U.S. adults

online use YouTube and Facebook, less than a quarter use Twitter. This kind of data can be much more powerful for concrete and specific story, whether it is to examine the hate speech spread by specific politicians in Myanmar or to examine the type of coverage published by conspiracy publication Infowars over time.

Not every user represents one real human being

In addition to that, not every user necessarily represents a person. There are automated accounts (bots) and accounts that are semi-automated and semi-human controlled (cyborgs). And there are also users who operate multiple accounts.

Again, understanding that there's a multitude of actors out there manipulating the flow of information for economic or political gain is an important aspect to keep in mind when looking at social media data in bulk (though this subject in itself — media and information manipulation — has become a major story in its own right that journalists have been trying to tell in ever-more sophisticated ways).

The tyranny of the loudest

Last but not least it's important to recognize that not everything or everyone's behavior is measured. A vast amount of people often choose to remain silent. And as more moderate voices are recorded less, it is only the extreme reactions that are recorded and fed back into algorithms that disproportionately amplify the already existing prominence of the loudest.

What this means is that the content that Facebook, Twitter and other platforms algorithmically surface on our social feeds is often based on the likes, retweets and comments of those who chose to chime in. Those who did not speak up are disproportionately drowned out in this process. Therefore, we need to be as mindful of what is not measured as we are of what is measured and how information is ranked and surfaced as a result of these measured and unmeasured data points.

Works Cited

Patrick Tucker, 'Has Big Data Made Anonymity Impossible?', *MIT Technology Review*, 7 May 2013

Lam Thuy Vo, 'The Umbrella Network', *Al Jazeera America*, 3 June 2015.

Lam Thuy Vo, 'Twitter Bots Are Trying To Influence You', *BuzzFeed News*, 11 October 2017.

Julia Angwin, Madeleine Varner and Ariana Tobin, 'Facebook Enabled Advertiser to Reach "Jew Haters"', *ProPublica*, 14 September 2017.

Monica Anderson and Aaron Smith, 'Social Media Use in 2018', *Pew Research Centre*, 1 March 2018.

Lam Thuy Vo, 'Here's What It Feels Like To Be Trolled In Trump's America', BuzzFeed News, 2017

Lam Thuy Vo, 'Here's What We Learned From Staring At Social Media Data For A Year', BuzzFeed News, 2017

The Algorithms Beat: Angles and Methods for Investigation

The Machine Bias series from ProPublica began in May 2016 as an effort to investigate algorithms in society.[1] Perhaps most striking in the series was an investigation and analysis exposing the racial bias of recidivism risk assessment algorithms used in criminal justice decisions.[2] These algorithms score individuals based on whether they are a low or high risk of reoffending. States and other municipalities variously use the scores for managing pre-trial detention, probation, parole, and sometimes even sentencing. Reporters at ProPublica filed a public records request for the scores from Broward County in Florida and then matched those scores to actual criminal histories to see whether an individual had actually recidivated (i.e. reoffended) within two years. Analysis of the data showed that black defendants tended to be assigned higher risk scores than white defendants, and were more likely to be incorrectly labeled as high risk when in fact after two years they hadn't actually been rearrested.[3]

Scoring in the criminal justice system is of course just one domain where algorithms are being deployed in society. The Machine Bias series has since covered everything from Facebook's ad targeting system, to geographically discriminatory auto insurance rates, and unfair pricing practices on Amazon.com. Algorithmic decision making is increasingly pervasive throughout both the public and private sectors. We see it in domains like credit and insurance risk scoring, employment systems, welfare management, educational and teacher rankings, and online media curation, among many others.[4] Operating at scale and often impacting large swaths of people, algorithms can make consequential and sometimes contestable calculation, ranking, classification, association, and filtering decisions. Algorithms, animated by piles of data, are a potent new way of wielding power in society.

As ProPublica's Machine Bias series attests, a new strand of computational and data journalism is emerging to investigate and hold accountable how power is exerted through algorithms. I call this *algorithmic accountability reporting*, a re-orientation of the traditional watchdog function of journalism towards the power wielded through algorithms.[5] Despite their ostensible objectivity, algorithms can and do make mistakes and embed biases that warrant closer scrutiny. Slowly, a beat on algorithms is coalescing as journalistic skills come together with technical skills to provide the scrutiny that algorithms deserve.

There are, of course, a variety of forms of algorithmic accountability that may take place in diverse forums beyond journalism, such as in political, legal, academic, activist, or artistic contexts.[6] But my focus is this chapter is squarely on algorithmic accountability reporting as an independent journalistic endeavor that contributes to accountability by mobilizing public pressure. This can be seen as complementary to other avenues that may ultimately also contribute to accountability, such as by developing regulations and legal standards, creating audit institutions in civil society, elaborating effective transparency policies, exhibiting reflexive art shows, and publishing academic critiques.

In deciding what constitutes the beat in journalism, it's first helpful to define what's newsworthy about algorithms. Technically speaking, an algorithm is a sequence of steps followed in order to solve a particular problem or to accomplish a defined outcome. In terms of information processes the outcomes of algorithms are typically decisions. The crux of algorithmic power often boils down to computers' ability to make such decisions very quickly and at scale, potentially affecting large numbers of people. In practice, algorithmic accountability isn't just about the technical side of algorithms though—algorithms should be understood as composites of technology woven together with people such as designers, operators, owners, and maintainers in complex sociotechnical systems.[7] Algorithmic accountability is about understanding how those people exercise power within and through the system, and are ultimately responsible for the system's decisions. Oftentimes what makes an algorithm newsworthy is when it somehow makes a "bad" decision. This might involve an algorithm doing something it

wasn't supposed to do, or perhaps not doing something it was supposed to do. For journalism, the public significance and consequences of a bad decision are key factors. What's the potential harm for an individual, or for society? Bad decisions might impact individuals directly, or in aggregate may reinforce issues like structural bias. Bad decisions can also be costly. Let's look at how various bad decisions can lead to news stories.

Angles on Algorithms

In observing the algorithms beat develop over the last several years in journalism, as well as through my own investigations of algorithms, I've identified at least four driving forces that appear to underlie many algorithmic accountability stories: (1) discrimination and unfairness, (2) errors or mistakes in predictions or classifications, (3) legal or social norm violations, and (4) misuse of algorithms by people either intentionally or inadvertently. I provide illustrative examples of each of these in the following subsections.

Discrimination and Unfairness

Uncovering discrimination and unfairness is a common theme in algorithmic accountability reporting. The story from ProPublica that led this chapter is a striking example of how an algorithm can lead to systematic disparities in the treatment of different groups of people. Northpoint, the company that designed the risk assessment scores (since renamed to Equivant), argued the scores were equally accurate across races and were therefore fair. But their definition of fairness failed to take into account the disproportionate volume of mistakes that affected black people. Stories of discrimination and unfairness hinge on the definition of fairness applied, which may reflect different political suppositions.[8]

I have also worked on stories that uncover unfairness due to algorithmic systems—in particular looking at how Uber pricing dynamics may differentially affect neighborhoods in Washington, DC.[9] Based on initial observations of different waiting times and how those waiting times shifted based on Uber's surge pricing algorithm we hypothesized that different neighborhoods would have different levels of service quality (i.e. waiting time). By systematically sampling the waiting times in different census tracts over time we showed that census tracts with more people of color tend to have longer wait times for a car, even when controlling for other factors like income, poverty rate, and population density in the neighborhood. It's difficult to pin the unfair outcome directly to Uber's technical algorithm because other human factors also drive the system, such as the behavior and potential biases of Uber drivers. But the results do suggest that when considered as a whole, the system exhibits disparity associated with demographics.

Errors and Mistakes

Algorithms can also be newsworthy when they make specific errors or mistakes in their classification, prediction, or filtering decisions. Consider the case of platforms like Facebook and Google which use algorithmic filters to reduce exposure to harmful content like hate speech, violence, and pornography. This can be important for the protection of specific vulnerable populations, like children, especially in products like Google's YouTube Kids which are explicitly marketed as safe for children. Errors in the filtering algorithm for the app are newsworthy because they mean that sometimes children encounter inappropriate or violent content.[10] Classically, algorithms make two types of mistakes: false positives and false negatives. In the YouTube Kids scenario, a false positive would be a video mistakenly classified as inappropriate when actually it's totally fine for kids. A false negative is a video classified as appropriate when it's really not something you want kids watching.

Classification decisions impact individuals when they either increase or decrease the positive or negative treatment an individual receives. When an algorithm mistakenly selects an individual to receive free ice cream (increased positive treatment), you won't hear that individual complain (although when others find out, they might say it's unfair). Errors are generally newsworthy when they lead to increased negative treatment for a person,

such as by exposing a child to an inappropriate video. Errors are also newsworthy when they lead to a decrease in positive treatment for an individual, such as when a person misses an opportunity. Just imagine a qualified buyer who never gets a special offer because an algorithm mistakenly excludes them. Finally, errors can be newsworthy when they cause a decrease in warranted negative attention. Consider a criminal risk assessment algorithm mistakenly labeling a high-risk individual as low-risk—a false negative. While that's great for the individual, this creates a greater risk to public safety by letting free an individual who goes on to commit a crime again.

Legal and Social Norm Violations

Predictive algorithms can sometimes test the boundaries of established legal or social norms, leading to other opportunities and angles for coverage. Consider for a moment the possibility of algorithmic defamation.[11] Defamation is defined as "a false statement of fact that exposes a person to hatred, ridicule or contempt, lowers him in the esteem of his peers, causes him to be shunned, or injures him in his business or trade."[12] Over the last several years there have been numerous stories, and legal battles, over individuals who feel they've been defamed by Google's autocomplete algorithm. An autocompletion can link an individual's or company's name to everything from crime and fraud to bankruptcy or sexual conduct, which can then have consequences for reputation. Algorithms can also be newsworthy when they encroach on social norms like privacy. For instance, Gizmodo has extensively covered the "People You May Know" (PYMK) algorithm on Facebook, which suggests potential "friends" on the platform that are sometimes inappropriate or undesired.[13] In one story, reporters identified a case where PYMK outed the real identity of a sex worker to her clients.[14] This is problematic not only because of the potential stigma attached to sex work, but also out of fear of clients who could become stalkers.

Defamation and privacy violations are only two possible story angles here. Journalists should be on the lookout for a range of other legal or social norm violations that algorithms may create in various social contexts. Since algorithms necessarily rely on a quantified version of reality that only incorporates what is measurable as data they can miss a lot of the social and legal context that would otherwise be essential in rendering an accurate decision. By understanding what a particular algorithm actually quantifies about the world—how it "sees" things – it can inform critique by illuminating the missing bits that would support a decision in the richness of its full context.

Human Misuse

Algorithmic decisions are often embedded in larger decision-making processes that involve a constellation of people and algorithms woven together in a sociotechnical system. Despite the inaccessibility of some of their sensitive technical components, the sociotechnical nature of algorithms opens up new opportunities for investigating the relationships that users, designers, owners, and other stakeholders may have to the overall system.[15] If algorithms are misused by the people in the sociotechnical ensemble this may also be newsworthy. The designers of algorithms can sometimes anticipate and articulate guidelines for a reasonable set of use contexts for a system, and so if people ignore these in practice it can lead to a story of negligence or misuse. The risk assessment story from ProPublica provides a salient example. Northpointe had in fact created two versions and calibrations of the tool, one for men and one for women. Statistical models need to be trained on data reflective of the population where they will be used and gender is an important factor in recidivism prediction. But Broward County was misusing the risk score designed and calibrated for men by using it for women as well.[16]

How to Investigate an Algorithm

There are various routes to the investigation of algorithmic power: no single approach will always be appropriate. But there is a growing stable of methods to choose from, including everything from highly technical reverse engineering and code inspection techniques, to auditing using automated or crowdsourced data collection, or even low-tech approaches to prod and critique based on algorithmic reactions.[17] Each story may require a different

approach depending on the angle and the specific context, including what degree of access to the algorithm, its data, and code is available. For instance, an exposé on systematic discrimination may lean heavily on an audit method using data collected online, whereas a code review may be necessary to verify the correct implementation of an intended policy.[18] Traditional journalistic sourcing to talk to company insiders such as designers, developers, and data scientists, as well as to file public records requests and find impacted individuals are as important as ever. I can't go into depth on all of these methods in this short chapter, but here I want to at least elaborate a bit more on how journalists can investigate algorithms using auditing.

Auditing techniques have been used for decades to study social bias in systems like housing markets, and have recently been adapted for studying algorithms.[19] The basic idea is that if the inputs to algorithms are varied in enough different ways, and the outputs are monitored, then inputs and outputs can be correlated to build a theory for how the algorithm may be functioning.[20] If we have some expected outcome that the algorithm violates for a given input this can help tabulate errors and see if errors are biased in systematic ways. When algorithms can be accessed via APIs or online webpages output data can be collected automatically.[21] For personalized algorithms, auditing techniques have also been married to crowdsourcing in order to gather data from a range of people who may each have a unique "view" of the algorithm. AlgorithmWatch in Germany has used this technique effectively to study the personalization of Google Search results, collecting almost 6 million search results from more than 4,000 users who shared data via a browser plugin (as discussed further by Christina Elmer in her chapter in this book).[22] Gizmodo has used a variant of this technique to help investigate Facebook's PYMK. Users download a piece of software to their computer that periodically tracks PYMK results locally to the user's computer, maintaining their privacy. Reporters can then solicit tips from users who think their results are worrisome or surprising.[23]

Auditing algorithms is not for the faint of heart. Information deficits limit an auditor's ability to sometimes even know where to start, what to ask for, how to interpret results, and how to explain the patterns they're seeing in an algorithm's behavior. There is also the challenge of knowing and defining what's expected of an algorithm, and how those expectations may vary across contexts and according to different global moral, social, cultural, and legal standards and norms. For instance, different expectations for fairness may come into play for a criminal risk assessment algorithm in comparison to an algorithm that charges people different prices for an airline seat. In order to identify a newsworthy mistake or bias you must first define what normal or unbiased should look like. Sometimes that definition comes from a data-driven baseline, such as in our audits of news sources in Google search results during the 2016 U.S. elections.[24] The issue of legal access to information about algorithms also crops up and is of course heavily contingent on the jurisdiction.[25] In the U.S., Freedom of Information (FOI) laws govern the public's access to documents in government, but the response from different agencies for documents relating to algorithms is uneven at best.[26] Legal reforms may be in order so that public access to information about algorithms is more easily facilitated. And if information deficits, difficult to articulate expectations, and uncertain legal access aren't challenging enough, just remember that algorithms can also be quite capricious. Today's version of the algorithm may already be different than yesterday's: as one example, Google typically changes its search algorithm 500-600 times a year. Depending on the stakes of the potential changes, algorithms may need to be monitored over time in order to understand how they are changing and evolving.

Recommendations Moving Forward

To get started and make the most of algorithmic accountability reporting I would recommend three things. Firstly, we've developed a resource called Algorithm Tips, which curates relevant methods, examples, and educational resources, and hosts a database of algorithms for potential investigation (first covering algorithms in the U.S. Federal government and then expanded to cover more jurisdictions globally)[27]. If you're looking for resources to learn more and help get a project off the ground, that could be one starting point.[28] Secondly, focus on the outcomes and impacts of algorithms rather than trying to explain the exact mechanism for their decision making.

Identifying algorithmic discrimination (i.e., an output) oftentimes has more value to society as an initial step than explaining exactly how that discrimination came about. By focusing on outcomes, journalists can provide a first-order diagnostic and signal an alarm which other stakeholders can then dig into in other accountability forums. Finally, much of the published algorithmic accountability reporting I've cited here is done in teams, and with good reason. Effective algorithmic accountability reporting demands all of the traditional skills journalists need in reporting and interviewing, domain knowledge of a beat, public records requests and analysis of the returned documents, and writing results clearly and compellingly, while often also relying on a host of new capabilities like scraping and cleaning data, designing audit studies, and using advanced statistical techniques. Expertise in these different areas can be distributed among a team, or with external collaborators, as long as there is clear communication, awareness, and leadership. In this way, methods specialists can partner with different domain experts to understand algorithmic power across a larger variety of social domains.

Works Cited

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Bias', *ProPublica*, May

2016.

Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm', *ProPublica*, May 2016.

Cathy O'Neil, 'Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy', *Broadway Books*, 2016.

Frank Pasquale, 'The Black Box Society: The Secret Algorithms That Control Money and Information', *Harvard University Press*, 2015.

Virginia Eubanks, 'Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor', *St. Martin's Press*, 2018.

Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures', *Digital Journalism* 3 (3), (2015).

Nicholas Diakopoulos, 'Sex, Violence, and Autocomplete Algorithms', *Slate*, August, 2013.

Nicholas Diakopoulos, 'Rage Against the Algorithms', *The Atlantic*, October, 2013.

Tega Brain and Surya Mattu, 'Algorithmic Disobedience' n.d.

Taina Bucher, 'If... Then: Algorithmic Power and Politics', *Oxford University Press*, 2018.

Nick Seaver, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems', *Big Data & Society*, 4(2), (2017).

Mike Ananny, 'Toward an Ethics of Algorithms', *Science, Technology & Human Values* 41 (1), (2015).

Bruno Lepri, et al., 'Fair, Transparent, and Accountable Algorithmic Decision-making Processes', *Philosophy & Technology*, 84(3), (2017).

Jennifer Stark and Nicholas Diakopoulos, 'Uber seems to offer better service in areas with more white people. That raises some tough questions', *Washington Post*, March 2016.

Sapna Maheshwari, 'On YouTube Kids, Startling Videos Slip Past Filters', *New York Times*, November 2017.

Nicholas Diakopoulos, 'Algorithmic Defamation: The Case of the Shameless Autocomplete', *Tow Center*, August 2013.

Seth C. Lewis, Kristin Sanders, Casey Carmody, 'Libel by Algorithm? Automated Journalism and the

Threat of Legal Liability', *Journalism & Mass Communication Quarterly* 80(1), (2018).

Kashmir Hill, 'How Facebook Figures Out Everyone You've Ever Met', *Gizmodo*, November 2017.

Kashmir Hill, 'How Facebook Outs Sex Workers', *Gizmodo*, October 2017.

Daniel Trielli and Nicholas Diakopoulos, 'How To Report on Algorithms Even If You're Not a Data Whiz', *Columbia Journalism Review*, May 2017.

Jeff Larson, 'Machine Bias with Jeff Larson', *Data Stories Podcast*, October, 2016.

Nicholas Diakopoulos, 'Automating the News: How Algorithms are Rewriting the Media', *Harvard University Press*, (2019).

Nicholas Diakopoulos, 'Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens', in *Towards glass-box data mining for Big and Small Data*, ed by Tania Cerquitelli, Daniele Quercia and Frank Pasquale, (Springer, June 2017), pp 25-44 .

Colin Lecher, 'What Happens When An Algorithm Cuts Your Health Care', *The Verge*,

March 2018.

Steven Michael Gaddis, 'An Introduction to Audit Studies in the Social Sciences', in *Audit Studies Behind the Scenes with Theory, Method, and Nuance*, ed. by Michael Gaddis, (Springer, 2017), pp 3-44.

Christian Sandvig, et al, 'Auditing algorithms: Research methods for detecting discrimination on Internet platforms', presented at *International Communication Association preconference on Data and Discrimination Converting Critical Concerns into Productive Inquiry*, (2014).

Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures', *Digital Journalism* 3 (3), (2015).

Valentino-DeVries, Jennifer, Jeremy Singer-Vine, and Ashkan Soltani, 'Websites Vary Prices, Deals Based on Users' Information', *Wall Street Journal*, 24 December 2012.

Kashmir Hill and Surya Mattu, 'Keep Track Of Who Facebook Thinks You Know With This Nifty Tool', *Gizmodo*, January 2018.

Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark and Sean Mussenden, 'I Vote For – How Search Informs Our Choice of Candidate', in *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, eds by M. Moore and D. Tambini, June 2018.

Esha Bhandari and Rachel Goodman, 'Data Journalism and the Computer Fraud and Abuse Act: Tips for Moving Forward in an Uncertain Landscape', *Computation + Journalism Symposium*, (2017).

Nicholas Diakopoulos, 'We need to know the algorithms the government uses to make important decisions about us', *The Conversation*, May 2016.

Katherine Fink, 'Opening the government's black boxes: freedom of information and algorithmic

accountability', *Digital Journalism* 17(1), (2017).

Robert Brauneis and Ellen Goodman, 'Algorithmic Transparency for the Smart City', *20 Yale Journal of Law & Technology 103*, 2018.

Daniel Trielli, Jennifer Stark and Nicholas Diakopoulos, 'Algorithm Tips: A Resource for Algorithmic Accountability in Government', *Computation + Journalism Symposium*, October 2017.

Algorithms in the Spotlight: Collaborative Investigations at Spiegel Online

Written by: Christina Elmer

The demand for transparency around algorithms is not new in Germany. Already in 2012, SPIEGEL ONLINE columnist Sascha Lobo called for the mechanics of the Google search algorithm to be disclosed[1], even if this would harm the company. The reason: Google can shape how we view the world, for example through the autocomplete function, as a prominent case in Germany illustrated. In this case, the wife of the former Federal President had taken legal action against Google because problematic terms were suggested in the autocomplete function when her name was searched for. Two years later, the German Minister of Justice repeated this appeal, which was extended again by the Federal Chancellor in 2016: algorithms should be more transparent, Angela Merkel demanded.[2]

In the past few years, the topic of algorithmic accountability has been under constant discussion at SPIEGEL ONLINE – but initially only as an occasion for reporting, not in the form of our own research or analysis project. There may be two primary reasons why the German media began experimenting in this area later than their colleagues in the United States: On the one hand, journalists in Germany do not have such strong freedom of information rights and instruments at their disposal; on the other hand, data journalism does not have such a long tradition compared to the United States. SPIEGEL ONLINE has only had its own data journalism department since 2016 and is slowly but steadily expanding this area. It is of course also possible for newsrooms with smaller resources to be active in this field - for example through cooperations with organisations or freelancers. In our case, too, all previous projects in the area of algorithmic accountability reporting have come about in this way. This chapter will therefore concentrate on collaborations and illustrate which lessons we have learned from them.

Google, Facebook, Schufa – three projects at a glance

Our editorial team primarily relies on cooperation when it comes to the investigation of algorithms. In the run-up to the 2017 federal elections, we joined forces with the NGO AlgorithmWatch to gain insights into the personalization of Google search results[3]. Users were asked to install a plugin that regularly performed predefined searches on their computer. A total of around 4,400 participants donated almost six million search results and thus provided the data for an analysis that would challenge the filter bubble thesis – at least regarding Google and the investigated area.

For this project, our collaborators from AlgorithmWatch approached SPIEGEL ONLINE, as they were looking for a media partner with a large reach for crowdsourcing the required data. While the content of the reporting was entirely the responsibility of our department covering internet and technology related topics, the data journalism department supported the planning and methodological evaluation of the operation. Furthermore, the backup of our legal department was essential in order to implement the project in a way which was legally bulletproof. For example, data protection issues had to be clarified within the reporting and had to be fully comprehensible for all participants involved in the project.

Almost at the same time, SPIEGEL ONLINE cooperated with ProPublica to deploy their AdCollector in Germany during the months before the elections.[4] The project aimed to make the Facebook ads targeting of the German parties transparent. Therefore, a plugin collected the political ads that a user sees in her stream and revealed those ads that are not displayed to her. For this project, SPIEGEL ONLINE joined forces with other German media such as Süddeutsche Zeitung and Tagesschau – an unusual constellation of actors who usually are in competition

with each other, but one that seemed necessary in the public interest in order to reach as many people as possible. The results could also be published in journalistic stories, but the clear focus was on transparency. After two weeks, already around 600 political advertisements had been collected and made available to the public.

ProPublica's Julia Angwin and Jeff Larson brought the idea of a collaboration to the annual conference of the not-for-profit association netzwerk recherche in Hamburg, where they held a session on algorithm accountability reporting. From the very beginning, the idea was developed both with technical and methodological experts from different departments in the newsroom of SPIEGEL ONLINE. The exchange with our previous cooperation partners of the NGO AlgorithmWatch was also very valuable for us in order to shed light on the legal background and to include it in our research. After the conference, we expanded the idea further in regular telephone conferences. Later on, our partners from other media outlets were also involved.

In 2018, SPIEGEL ONLINE is supporting a major project aimed at investigating an extremely powerful algorithm in Germany – the Schufa credit report, which is used to assess the creditworthiness of private individuals. The report should show how high the probability is that someone can pay his bills, pay the rent or service a loan. It can therefore have far-reaching implications for a person's private life and a negative effect on society as a whole. For example, it is conceivable that the score increases social discrimination or treats individuals unequally, depending on whether more or less data on them is available. Also, incorrect data from integrated sources or mix-ups could be fatal for individuals.

However, the underlying scoring is not transparent; which data is taken into account in which weighting is not known. And those affected do not always notice anything of the process. This makes Schufa a controversial institution in Germany – and projects like OpenSCHUFA absolutely vital for public debate on algorithmic accountability, in our opinion.[5]

The project is mainly driven by the NGOs Open Knowledge Foundation (OKFN) and AlgorithmWatch, SPIEGEL ONLINE is one of two associated cooperation partners together with Bayerischer Rundfunk (Bavarian Broadcasting). The idea for this project came up more or less simultaneously with several parties involved. After some successful projects with the NGOs AlgorithmWatch and OKFN as well as with the data journalism team of Bayerischer Rundfunk, SPIEGEL ONLINE was included in the initial discussions.

The constellation posed special challenges. For the two media teams, it was important to work separately from the NGOs in order to ensure their independence from the crowdfunding process in particular. Therefore, although there were of course discussions between the actors involved, neither an official partnership nor a joint data evaluation is possible. This example emphasizes how important it is for journalists to reflect on their autonomy, especially in such high-publicity topics.

Making OpenSCHUFA known was one of the central success factors of this project. The first step was to use crowdfunding to create the necessary infrastructure to collect the data, which will be collected later in 2018 via crowdsourcing. The results are to be jointly evaluated by the partners in the course of the year in anonymized form. The central question behind it: Does the Schufa algorithm discriminate against certain population groups, and does it increase inequality in society?

As of March 2018, the campaign was largely successful. The financing of the software could be secured within the crowdfunding framework.[6] In addition, more than 16,000 people had already requested information to Schufa in order to obtain their personal data. These reports will later be the basis for the analysis of the algorithm and its effects.

Resonance and success indicators

Concerning their results, both the Facebook and the Google project were rather unspectacular and did not show the assumed effects. Political parties apparently hardly used Facebook's targeting options and the much-cited Google filter bubble proved to be unmeasurable within the crowdsourcing in Germany. In any case, to us it was more relevant to increase literacy with algorithms amongst our readers and to illustrate their functionalities and risks.

The assumption that we have succeeded in making the topic more widely known can be supported by the reach of exemplarily articles. The introductory article at the start of the Schufa project reached a large audience of around 335,000 readers, the majority of whom, however, came to the article via internal channels like our homepage. This was different in the field report with our author's personal story, which was read by around 220,000 people. A fifth of them reached the article via social media channels, which is well above the average. So apparently, it has been possible to reach new target groups with this topic. The reading time was also clearly above normal – with an average of almost three minutes. In addition, the topic was widely discussed in the public and in many media, as well as at several conferences.

What about the impact on the everyday reality? As a first step, it was important for us to anchor the topic in the public consciousness. So far, we have not seen any fundamentally different way political actors deal with publicly effective algorithms. We hope, however, that such projects will ultimately increase the pressure on legislation and standards for transparency in this area.

In any case, more effort would be needed in this area. With the discussed projects we were able to work on specific aspects of relevant algorithms, but of course it would be advisable to focus much more resources on this topic. It's great news that the pioneering work of Julia Angwin and Jeff Larson will be developed through a new media organisation focusing on the social impact of technology, which can devote more attention to this topic. Further experimentation is very much needed, partly because there is still some scope for action in the regulation of algorithms. The field of algorithmic accountability reporting has only developed in recent years. And it will have to grow rapidly to meet the challenges of an increasingly digitized world.

Organising collaborative investigations

Working together in diverse constellations not only makes it easier to share competencies and resources, it also allows a clear definition of roles. As a media partner, SPIEGEL ONLINE can work as a more neutral commentator without being too deeply involved in the project itself. The editors remain independent and thus justify the trust of their readers. Of course, they also apply their quality criteria to the reports within such a project – for example, by always giving any subject of their reporting the opportunity to comment on accusations. Compared to the NGOs involved, these mechanisms may slow media partners down more than they are comfortable with, but at the same time they ensure that readers are fully informed by their reports – and that these will enrich public debate in the long term.

Addressing these roles in advance has proven to be an important success criterion for collaborations in the field of algorithmic accountability. A common timeline should also be developed at an early stage and language rules for the presentation of the project on different channels should be defined. Because after all, a clear division of roles can only work if it is communicated consistently. This includes, for example, a clear terminology on the roles of the different partners in the project and the coordination of disclaimers in the event of conflicts of interest.

Behind the scenes, project management methods should be used prudently, project goals should be set clearly and available resources have to be discussed. Coordinators should help with the overall communication and thus give the participating editors the space they need for their investigations. To keep everyone up to date, information channels should be kept as simple as possible, especially around the launch of major project stages.

Regarding the editorial planning, the three subject areas were challenging. Although the relevance and news value were never questioned in general, special stories were needed to reach a broad readership. Often, these stories focused on the personal effects of the algorithms examined. For example, incorrectly assigned Schufa data made it difficult for a colleague from the SPIEGEL ONLINE editorial team to conclude an Internet contract. His experience report impressively showed what effects the Schufa algorithm can have on a personal level and thus connected with the reality of our audience's lives[7].

Thus, we tailored the scope of our reporting to the interests of our audience as far as possible. Of course, the data journalists involved are also very interested in the functioning of the algorithms under investigation – an interest that is extremely useful for research purposes. However, only if these details have a relevant influence on the results of the algorithms can they become the subject of reporting – and only if they are narrated in a way that is accessible for our readers.

Internally in the editorial office, support for all three projects was very high. Nevertheless, it was not easy to free up resources for day-to-day reporting in the daily routine of a news-driven editorial team - especially when the results of our investigations were not always spectacular.

Nevertheless, the topic of algorithmic accountability reporting is very important to us. Because in Europe we now still have the opportunity to discuss the issue in society and to shape how we want to deal with it. It is part of our function as journalists to provide the necessary knowledge so that citizens can understand and shape this scope. And as far as possible, we also take on the role of a watchdog by trying to make algorithms and their effects transparent, to identify risks and to confront those responsible. To achieve this, we have to establish what might otherwise be considered unusual collaborations with competitors and actors from other sectors.

What we have learned from these projects

1. **Collaborate where possible**. Only in diverse teams can we design a good setup for investigating such topics and also join forces – an important argument given both the scarcity of resources and legal restrictions that most journalists have to cope with. But since these projects bring together actors from different systems, it is crucial to discuss the underlying relevance criteria, requirements and capabilities beforehand.
2. **Define your goals in a comprehensive way**. Raising awareness for the operating principles of algorithms can be a first strong goal in such projects. Of course, projects should also try to achieve as much transparency as possible. At best we can check whether algorithms have a discriminatory effect – but project partners should keep in mind that this is surely a more advanced goal that requires the availability of extensive datasets.
3. **Implement such projects with caution**. Depending on the workload and the day-to-day pressure of the journalists involved, you might even need a project manager. Be aware that the project timeline may conflict with the requirements of the current reporting from time to time. Take this into account in communicating with other partners and, if possible, prepare alternatives for such cases.
4. **Dedicate yourself to research design**. To set up a meaningful design that produces useful data, you might need specialized partners. Close alliances with scientists from computer science, mathematics and other thematically related disciplines are particularly helpful for investigating some of the more technical aspects of algorithms. Furthermore, it may also be useful to cooperate with social and cultural researchers to gain a deeper understanding of classifications and norms that are implemented in them.
5. **Protect the data of your users very carefully**. If algorithms are to be investigated, you may use data donations from usersin order to consider as many different cases as possible. Especially in such crowdsourcing projects, legal support is indispensable in order to ensure data protection and to take into account the requirements of the national laws and regulations. If your company has a data protection officer, involve them in the project early on.

Works Cited

Sascha Lobo, ‘Was Bettina Wulff mit Mettigeln verbindet’, *SPIEGEL ONLINE*, September, 2012.

Sorge vor Kartell, ‘Maas hätte gerne, dass Google geheime Suchformel offenlegt’, September, 2014.

Fabian Reinbold, ‘Warum Merkel an die Algorithmen will’, *SPIEGEL ONLINE*, October, 2016.

AlgorithmWatch: Datenspende BTW17, September 2017

Julia Angwin and Jeff Larson, ‘ Help Us Monitor Political Ads Online’, *ProPublica*, September, 2017.

OpenSCHUFA project website www.openschufa.de

OpenSCHUFA crowdfunding framework https://www.startnext.com/openschufa

Von Philipp Seibt, ‘Wie ich bei der Schufa zum "deutlich erhöhten Risiko" wurde’, *SPIEGEL ONLINE*, March, 2018.

How Do Platforms See Humans?

This chapter is launching soon

Investigations into the Digital: Reporting on Misinformation, Platforms and Digital Culture at BuzzFeed News

This chapter is launching soon

Telling Data: Digital Methods for Analysing Web Trackers and Other Natively Digital Objects

This chapter is launching soon

Archiving Data Journalism

Written by: Meredith Broussard

In the first edition of the *Data Journalism Handbook*, published in 2012, data journalism pioneer Steve Doig wrote that one of his favorite data stories was the Murder Mysteries project by Tom Hargrove[1]. In the project, which was published by Scripps Howard News Service, Hargrove looked at demographically detailed data about 185,000 unsolved murders and built an algorithm to suggest which murders might be linked. Linked murders could indicate a serial killer at work. "This project has it all," Doig wrote. "Hard work, a database better than the government's own, clever analysis using social science techniques, and interactive presentation of the data online so readers can explore it themselves."

By the time of the second edition of the *Data Journalism Handbook*, six years later, the URL to the project was broken . The project was gone from the web because its publisher, Scripps Howard was gone. Scripps Howard News Service had gone through multiple mergers and restructurings, eventually merging with Gannett, publisher of the USA Today local news network.

We know that people change jobs and media companies come and go. However, this has had disastrous consequences for data journalism projects.[2] Data projects are more fragile than "plain" text-and-images stories that are published in the print edition of a newspaper or magazine.

Ordinarily, link rot is not a big deal for archivists; it is easy to use Lexis-Nexis or ProQuest or another database provider to find a copy of everything published by, say, the New York Times print edition on any day in the twenty-first century. But for data stories, link rot indicates a deeper problem. Data journalism stories are not being preserved in traditional archives. As such, they are disappearing from the web. Unless news organizations and libraries take action, future historians will not be able to read everything published by the Boston Globe on any given day in 2017. This has serious implications for scholars and for the collective memory of the field. Journalism is often referred to as the "first draft of history." If that first draft is incomplete, how will future scholars understand the present day? Or, if stories disappear from the web, how will individual journalists maintain personal portfolios of work?

This is a human problem, not just a computational problem. To understand why data journalism isn't being archived for posterity, it helps to start with how "regular" news is archived. All news organizations use software called a content management system (CMS), which allows the organization to schedule and manage the hundreds of pieces of content it creates every day, and also imposes a consistent visual look and feel on each piece of content published. Historically, legacy news organizations have used a different CMS for the print edition and for the web edition. The web CMS allows the news organization to embed ads on each page, which is one of the ways that the news organization makes money. The print CMS allows print page designers to manage different versions of the print layout, and then send the pages to the printer for printing and binding. Usually, video is in a different CMS. Social media posts may or may not be managed by a different application like SocialFlow or Hootsuite. Archival feeds to Lexis-Nexis and the other big providers tend to be hooked up to the print CMS. Unless someone at the news organization remembers to hook up the web CMS too, digital-first news isn't included in the digital feeds that libraries and archives get. This is a reminder that archiving is not neutral, but it depends on deliberate human choices about what matters (and what doesn't) for the future.

Most people ask at this point, "What about the Internet Archive?" The Internet Archive is a treasure, and the group does an admirable job of capturing snapshots of news sites. Their technology is among the most advanced digital archiving software. However, their approach doesn't capture everything. The Internet Archive only collects publicly available web pages. News organizations that require logins, or which include paywalls as part of their

financial strategy, cannot be automatically preserved in the Internet Archive. Web pages that are static content, or plain HTML, are the easiest to preserve. These pages are easily captured in the Internet Archive. Dynamic content, such as Javascript or a data visualization or anything that was once referred to as “Web 2.0,” is much harder to preserve, and is not often stored in the Internet Archive. “There are many different kinds of dynamic pages, some of which are easily stored in an archive and some of which fall apart completely,” reads an Internet Archive FAQ. “When a dynamic page renders standard html, the archive works beautifully. When a dynamic page contains forms, JavaScript, or other elements that require interaction with the originating host, the archive will not contain the original site's functionality.”

Dynamic data visualizations and news apps, currently the most cutting-edge kinds of data journalism stories, can’t be captured by existing web archiving technology. Also, for a variety of institutional reasons, these types of stories tend to be built outside of a CMS. So, even if it were possible to archive data visualizations and news apps, (which it generally isn’t using this approach), any automated feed wouldn’t capture them because they are not inside the CMS.

It’s a complicated problem. There aren’t any easy answers. I work with a team of data journalists, librarians, and computer scientists who are trying to develop tech to solve this thorny problem. We’re borrowing methods from reproducible scientific research to make sure people can read today’s news on tomorrow’s computers. We’re adapting a tool called ReproZip that collects the code, data, and server environment used in computational science experiments. We think that ReproZip can be integrated with a tool such as Webrecorder.io in order to collect and preserve news apps, which are both stories and software. Because web and mobile based data journalism projects depend on and exist in relation to a wide range of other media environments, libraries, browser features, and web entities (which may also continually change), we expect that we will be able to use ReproZip to collect and preserve the remote libraries and code that allow complex data journalism objects to function on the web. It will take another year or two to prove our hypothesis.

In the meantime, there are a few concrete things that every data team can do to make sure their data journalism is preserved for the future.

1. **Take a video**. This strategy is borrowed from video game preservation. Even when a video game console is no more, a video play-through can show the game in its original environment. The same is true of data journalism stories. Store the video in a central location with plain text metadata that describes what the video shows. Whenever a new video format emerges (as when VHS gave way to DVD, or DVD was replaced by streaming video), upgrade all of the videos to this new format.
2. **Make a scaled-down version for posterity**. Libraries like Django-bakery allow dynamic pages to be rendered as static pages. This is sometimes called “baking out.” Even in a database with thousands of records, each dynamic record could be baked out as a static page that requires very little maintenance. Theoretically, all of these static pages could be imported into the organization’s content management system. Baking out doesn’t have to happen at launch. A data project can be launched as a dynamic site, then it can be transformed into a static site after traffic dies down a few months later. The general idea is: adapt your work for archiving systems by making the simplest possible version, then make sure that simple version is in the same digital location as all of the other stories published around the same time.
3. **Think about the future**. Journalists tend to plan to publish and move on to the next thing. Instead, try planning for the sunset of your data stories at the same time that you plan to launch them. Matt Waite’s story “Kill All Your Darlings” on Source, the Open News blog, is a great guide to how to think about the life cycle of a data journalism story. Eventually, you will be promoted or will move on to a new organization. You want your data journalism to survive your departure.

4. **Work with libraries, memory institutions, and commercial archives**. As an individual journalist, you should absolutely keep copies of your work. However, nobody is going to look in a box in your closet or on your hard drive, or even on your personal website, when they look for journalism in the future. They are going to look in Lexis-Nexis, ProQuest, or other large commercial repositories. To learn more about commercial preservation and digital archiving, Kathleen Hansen and Nora Paul's book *Future-proofing the News: Preserving the First Draft of History* is the canonical guide for understanding the news archiving landscape as well as the technological, legal, and organizational challenges to preserving the news.

Works Cited

Katherine Boss and Meredith Broussard, 'Challenges of archiving and preserving born-digital news applications', *IFLA Journal* 42:3, (2017), pp. 150-157.

Meredith Broussard, 'Preserving news apps present huge challenges', *Newspaper Research Journal* 36:3, (2015), pp. 299-313.

ProPublica, 'A Conceptual Model for Interactive Database Projects in News' (2016),

Meredith Broussard, 'The Irony of Writing Online About Digital Preservation',*The Atlantic*, 20 November 2015.

Meredith Broussard, 'Future-Proofing News Apps', *Media Shift*, 23 April 2014.

Data Journalism's Entanglements with Civic Tech

Written by: Stefan Baack

While computer-assisted reporting was considered a practice exclusive to (investigative) journalists, data journalism is characterized by its entanglements with the technology sector and other forms of data work and data culture. Compared to computer-assisted reporting, the emergence of data journalism in the US and in Europe intersected with several developments both within and outside newsrooms: the growing availability of data online, not least due to open data initiatives and leaks; newsrooms hiring developers and integrating them within the editorial team to better cope with data and provide interactive web applications; and the emergence of various 'tech for good' movements that are attracted to journalism as a way to use their technological skills for a 'public good'. This has contributed to an influx of technologists into newsrooms ever since data journalism emerged and became popular in the 2000s in the West and elsewhere. However, the resulting entanglements between data journalists and other forms of data work are distinct in different regions. Moreover, data journalism is connected to new, entrepreneurial forms of journalism that have emerged in response to the continued struggle of media organizations to develop sustainable business models. These new types of media organizations, e.g. nonprofit newsrooms like ProPublica or venture-backed news startups like BuzzFeed, tend to question traditional boundaries of journalism in their aspiration to 'revive' or 'improve' journalism, and technology and data often play a key role in these efforts.[1]

The entanglements between data journalism and other forms of data work and data cultures create new dependencies, but also new synergies that enable new forms of collaboration across sectors. Here I want to use the close relationship between data journalism and civic tech as an example because in many places both phenomena emerged around the same time and mutually shaped each other from an early stage. Civic tech is about the development of tools that aim to empower citizens by making it easier for them to engage with their governments or to hold them accountable. Examples of civic tech projects are OpenParliament, a parliamentary monitoring website that, among other things, makes parliamentary speeches more accessible; WhatDoTheyKnow, a freedom of information websites that helps users to submit and find freedom of information requests; and FixMyStreet, which simplifies the reporting of problems to local authorities.[2]

Civic technologists and data journalists share some important characteristics. First, many practitioners in both groups are committed to the principles of open source culture and promote sharing, the use of open source tools and data standards. Second, data journalists and civic technologists heavily rely on data, be it from official institutions, via crowdsourcing or other sources. Third, while differing in their means, both groups aspire to provide a public service that empowers citizens and holds authorities accountable. Because of this overlapping set of data skills, complementary ambitions and joint commitment to sharing, civic technologists and data journalists easily perceive each other as complementary. In addition, support from media organizations, foundations like the Knight Foundation, and grassroots initiatives like Hacks/Hackers have created a continuous exchange and collaborations between data journalists and civic technologists.

The tension between expanding and reinforcing the journalistic 'core'

Based on a case study in Germany and the UK that examined how data journalists and civic technologists complement each other, we can describe their entanglements as revolving around two core practices: facilitating and gatekeeping.[3] Facilitating means enabling others to take actions themselves, while gatekeeping refers to the traditional journalistic role model of being a gatekeeper for publicly relevant information. To illustrate the difference,

parliamentary monitoring websites developed by civic technologists are intended to enable their users to *inform themselves*, e.g. by searching through parliamentary speeches (facilitating), but not to *pro-actively push* information to them that is deemed relevant by professionals (gatekeeping). Facilitating is about individual empowerment, while gatekeeping is about directing public debate and having impact.

What characterizes the entanglements between data journalists and civic technologists is that practices of facilitating and gatekeeping are complementary and can mutually reinforce each other. For example, civic tech applications not only facilitate ordinary citizens; data journalists can use them for their own investigations. Investigations by journalists, on the other hand, can draw attention to particular issues and encourage people to make use of facilitating services. Moreover, information rights are essential for both facilitating and gatekeeping practices, which creates additional synergies. For example, data journalists can use their exclusive rights to get data that they then share with civic technologists; while journalists can profit from civic tech's advocacy for stronger freedom of information rights and open data policies.

New entrepreneurial forms journalism play a particular role in the relationship between data journalism and civic tech, as they are more open towards expanding traditional gatekeeping with civic tech's notion of facilitating. For example, ProPublica has developed several large, searchable databases intended to facilitate not the engagement of ordinary citizens with their governments, but journalistic investigations by local newsrooms who do not have the resources and expertise to collect, clean and analyze data themselves. Another nonprofit newsroom from Germany, Correctiv, has taken a similar approach and even integrated the freedom of information website of the Open Knowledge Foundation Germany into some of its applications to enable users to directly request further information that is then automatically added back to Correctiv's database.[4]

While these examples illustrate that there is a growing number of organizations that expand traditional notions of journalism by incorporating practices and values from other data cultures, there is also the opposite: data journalists that react to the similarities in practices and aspirations with other fields of data work by embracing their professional identity as gatekeepers and storytellers. Those journalists do not necessarily reject civic tech, but their response is a greater specialization of journalism, closer to notions of traditional, investigative journalism.

The opportunities of blurry boundaries

In sum, data journalism's entanglements with other fields of data work and data culture contribute to a greater diversification of how 'journalism' is understood and practiced, be it towards an expansion or a reinforcement of traditional values and identities. Both journalists themselves, and researchers can consider data journalism as a phenomenon embedded in broader technological, cultural and economic transformations. I have focused on the entanglements between data journalists and civic technologists in this article, but I would like to point out two key lessons for data journalists that are relevant beyond this particular case:

1. Benefitting from blurry boundaries: Journalists tend to describe a lack of professional boundaries towards other fields as problematic, but the synergies between civic technologists and data journalists demonstrate that blurry boundaries can also be an advantage. Rather than perceiving it primarily as problematic, data journalists also need to ask whether there are synergies with other fields of data work, and how to best benefit from them. Importantly, this does not mean that journalists necessarily have to adopt practices of facilitating themselves. While there are examples of that, journalists who reject this idea can still try to find ways to benefit without sacrificing their professional identity.
2. Embracing diversity in professional journalism: The findings of my study reflect how 'journalism' is increasingly delivered by a variety of different, more specialized actors. This diversification is raising concerns for some of the journalists I interviewed. For them, media organizations that adopt practices of facilitating might weaken

their notion of 'hard', investigative journalism. However, journalists need to acknowledge that it is unlikely that there will be one definite form of journalism in the future.

In sum, a stronger awareness of both the historical and contemporary ties to other forms of data work and data culture can help journalists to reflect about their own role, and to be better aware of not just new dependencies, but also potential synergies that can be used to support and potentially expand their mission.

Works Cited

Andrea Wagemans, Tamara Witschge, and Frank Harbers, 'Impact as driving force of journalistic and social change', SAGE journals, 2018.

Stefan Baack, 'Practically Engaged. The entanglements between data journalism and civic tech'. *Digital Journalism*, 6(6), 673–692, 2018.

Nikki Usher, 'Venture-backed News Startups and the Field of Journalism'. *Digital Journalism*, 5(9), 1116–1133, 2017.

Data Feudalism: How Platform Journalism and the Gig Economy Shape Cross-Border Investigative Networks

This chapter is launching soon

Data Journalism in the Newsroom

This chapter is launching soon

Data Journalism Culture

This chapter is launching soon

Organising Cross-Border Data Journalism Initiatives: Case Studies in Africa

This chapter is launching soon

A Decade of Data Journalism: 2009-2019

This chapter is launching soon

Open Source Coding Practices in Data Journalism

This chapter is launching soon

Data Journalism and Gender

This chapter is launching soon

The #ddj Hashtag – Eunice Au (GIJN) and Marc Smith (Connected Action)

This chapter is launching soon

Data-Driven Editorial? Considerations for Working with Audience Metrics

This chapter is launching soon

Data Journalism By, About and For Marginalised Communities

Written by: Eva Constantaras

I do data journalism in countries where things are widely considered to be going badly – as in not just a rough patch, not just a political hiccup, but entire political and economic systems failing. In such places, one reads that corruption has paralyzed the government, citizens are despondent and civil society under siege. Things are going terribly. Producing data journalism in some of the most impoverished, uneducated and unsafe parts of the world has brought me to an important conclusion about data journalism. Injustice, inequality and discrimination are ubiquitous, insidious and overlooked in most countries. Journalists I work with have unflinchingly embraced new tools to, for the first time, measure just how bad things are, who is suffering as a result, whose fault it is and how to make things better. In these contexts, journalists have embraced data as a means to influence policy, mobilize citizens and combat propaganda. Despite the constraints on free press, data journalism is seen as a means to empowerment.

What I bring and would like to explore in this piece is a commitment to data journalism by, about and for marginalized communities. By attending to different aspects of injustice, inequality and discrimination, and their broader consequences on the lives of marginalised communities, we render them visible, measurable and maybe even solvable. These stories engage journalists deeply rooted in marginalized communities. They tap into issues that groups which face institutional discrimination care about to foster citizen engagement. They are disseminated through local mass media to reach the most people and pressure governments into making better decisions for the whole country. Here are five kinds of data journalism stories that attend to the interests and concerns of marginalised communities in Afghanistan, Pakistan, Kenya, Kyrgyzstan and the Balkans.

1. Why are people going hungry if our country has enough resources to feed everyone?

In Kenya, donors were funding exactly the wrong food programs. A 12-minute, television story by NTV's Mercy Juma about Turkana, an isolated, impoverished region of northern Kenya, revealed that malnutrition in children is a growing problem as drought and famine becomes more intense and frequent. Money goes to emergency food aid, not long-term drought mitigation. The same money spent on one year of emergency food aid could fund a food sustainability programme for the entire county and its nearly million residents, according to draft policies in parliament. She threatened to pull her story when editors wanted to edit out the data: her story depended on engaging donors, enraging citizens and embarrassing the government mostly through television, but also in print and summarized online[1].

She convinced donors with the strength of her data. She sourced climate, agricultural and health data from government ministries, public health surveys, donor agencies and the Kenyan Red Cross. The USAID Kenya mission saw the data visualization demonstrating that one year of USAID emergency food aid could fund the entire Kenya Red Cross food sustainability strategy for Turkana. She demonstrated the health impact on children of delays and the stark contrast to countries growing food in deserts. She was invited to present her findings at the USAID office in Nairobi and in 2015, USAID's agriculture and food security strategy shifted from humanitarian aid to sustainable agriculture[2].

She won over public opinion with the intimate documentation of families starving in Turkana. She spent three days with the families featured in the piece along with a Turkana translator and videographer. The station phone was ringing off the hook before the story finished airing with Kenyans seeking to donate money to the families featured in the story. Due to the massive reaction to the story from individuals and organizations, within hours the station

established a relief fund for Turkana County. This and follow up stories on the desperate famine situation in Northern Kenya prompted daily attention in the Kenyan media, which has historically shown a lack of interest in the plight of the isolated and impoverished regions of northern Kenya. Her main audience connected to a strong, human story and not the data that would suggest donations could be more wisely invested in development.

The government succumbed to public and donor pressure. The Drought Monitoring Committee asked Juma to share data from her story because they claimed they were not aware that the situation had become so desperate, though the same department had tried to charge her for access to the data when she began her investigation. Based on Juma's water shortage data, the Ministry of Water plans to travel to Turkana to dig more boreholes. The government, through the Ministry of Planning and Devolution, released Sh2.3 billion ($27 million) to go towards relief distribution in Turkana County, a development that Juma followed closely. Due to the massive reaction to the story from individuals and organizations, food sustainability legislation that redirected aid was finally introduced into the Senate in May that year[3]. Juma has continued to produce data-driven features on the disconnect between public perception, donor programs and policy, including in Teen Mums of Kwale, an investigation on the impact of contraceptive use on teen pregnancy rates in a conservative part of the country.[4]

2. How do we ensure our justice system is protecting the marginalized?

In Afghanistan, Pajhwok Afghan News data team used data to probe the impact of two policies lauded as key for progress towards justice in the country: the 2009 Afghanistan's Law on the Elimination of Violence Against Women and the Afghanistan National Drug Control Strategy (2012-2016) and found two unexpected casualties of these policies: abused women and rural labourers. Though Afghanistan does not have an access to information law, many agencies that receive donor funding, including the women's affairs and counter-narcotics ministries, are contractually obligated to make that data available.

Five years after the domestic violence law took effect, Pajjhwok Afghan wanted to track the fate of abusers and the abused. The team obtained the data on the 21,000 abuse cases from the Ministry of Women Affairs and several UN agencies tasked with tracking cases from registration to final verdicts and mediation. They found that in the worst country in the world to be a woman, the widely lauded law has channeled women through a local mediation process entrenched in traditional chauvinism that usually lands her right back with her abuser[5]. Two years later, Human Rights Watched published a study confirming PAN's findings; the law and mediation have failed Afghan women[6]. Even if more women had access to the court system, which boasts a high rate of conviction for abusers, there remains the thorny issue of what to do with divorced women in a society were women do not work.

Similar practical challenges arise in the enforcement of Afghanistan's drug strategy. The United Nations Office of Drugs and Crime was granted rare access to prisoners convicted of drug charges and handed over the raw survey data to the Pajhwok team. Analysis of survey findings revealed that the policy has landed mostly poor illiterate drivers and farmers in prison while most drug kingpins walk free[7]. Most also reported that they planned to go right back to labouring in the drug trade once they are released as it is the only way to support their families in isolated rural areas.

These stories served a threefold purpose for the Pajhwok data team: reality check policies developed from a Western legal lens, highlight the consequences of economic marginalization by both gender and location and provide data-driven public interest content in Dari, Pashtu and English for a diverse audience.

3. How do we ensure a quality education for everyone?

Access to education, often regarded as a great equalizer, has allowed marginalized communities to quantify a government's failure to provide basic public services and push local leaders towards reform. In a series of stories, developer-cum-journalist Abdul Salam Afridi built a beat around education access among the disadvantaged, which landed him on the shortlist for the Data Journalism Awards for his portfolio. In his first story, he used official government education statistics and nationwide education survey data to show that parents in the remote tribal region of Khyber Pass, who out of desperation were sending growing numbers of children to private schools were making a bad investment. His story showed that most graduating students in both public and private schools fail basic standardized tests[8]. Further stories on public education in the Federally Administrative Tribal Areas, where Salam himself is from, and KP probe the reasons behind failing schools[9].

Another story based on student rosters for the national vocational training program and national job listings revealed a huge gap between skills and market demand. The investigation revealed that the country is training IT specialists and beauticians when it needs drivers and steel workers, leaving over half of their alumni unemployed, largely because of who was behind the project. Funded by the German government development fund, GiZ, the Pakistan government did its own analysis, came to the same conclusion and quickly overhauled to program with new course offerings aligned with more needed jobs skills[10].

An inherent advantage to data driven beat reporting among marginalized communities is that the journalist can stay on the story after the initial scandal is forgotten. What these stories also have in common is that they use data not just to report the problem, but also what can be done about it. These journalists gathered data to measure the problem, the impact, the causes and the solution. Globally, there is a push for accessible data journalism by, about and for marginalized communities to win their trust and engage them in civic life.

Data journalism under constraints

Much of the division in academia about the long-term viability of data journalism stems from a split over whether its aim is to produce high profile interactive product or fact-based public interest reporting. Journalists in developing countries use data to answer basic questions about institutionalized gender discrimination, prejudicial justice systems and willful neglect of the hungry and deliver that information to as many people as they can. They do this knowing that these problems are complicated and policies are still very unlikely to change as a result. Data journalists in the West, with access to better resources, data and free media, and a more responsive government are often not seizing the opportunity to ensure that in such tumultuous times, we are addressing the information needs of marginalized citizens and holding government accountable.

Most of these problems were invisible before and will become invisible again if journalists stop counting. Data journalism at its best is by, about and for those who society has decided do not count. Luckily civil society, activists, academics, governments and others are working together to do a better job of counting those who have been left out. Journalists have a vital role in ensuring that these are problems people are talking about and working to fix. Everything was terrible, is terrible and will be terrible unless we keep counting and talking. Year after year, we need to count the hungry, the abused, the imprisoned, the uneducated, the unheard because everywhere on earth, things are terrible for someone.

Works Cited

Mercy Juma, 'When will Kenya have enough to feed all its citizens?', Daily Nation, 28 January 2014.

NTV Kenya, '#TeenMumsOfKwale', YouTube, 2 October 2016.

Abdul Qadir Munsef and Zarghona Salehai, 'Cases of Violence Against Women,' Pajhwok Afghan News, 11 May 2016.

United Nations Assistance Mission in Afghanistan, ‘Injustice and Impunity: Mediation of Criminal Offences of Violence Against Women’, May 2018.

Navid Ahmad Barakzai and Ahsanullah Wardak, ‘Most Jailed Drug Offenders Are Poor, Illiterate’, Pajhwok Afghan News, 28 September 2016.

Abdul Salam Afridi, ‘In KP, Parents Still Prefer Private Over Public Schools’, News Lens, 18 February 2017.

Abdul Salam Afridi, ‘Half of FATA Schools Functioning in Dire Straits’, News Lens, 16 June 2017.

Abdul Salam Afridi, ‘Despite Huge Investment The Outlook of Education in KP Remains Questionable’, News Lens, 2 March 2018.

Abdul Salam, ‘TVET Reform Programmes Targeting Wrong Skills’, Data Journalism Pakistan, 16 September 2017.

Teaching Data Journalism at Universities in the United States

This chapter is launching soon

Data Journalism MOOCs in Turkey

This chapter is launching soon

Hackathons and Bootcamps in Kyrgyzstan: Reflections on Training Data Journalists in Central Asia

This chapter is launching soon

Data Journalism, Digital Universalism and Innovation in the Periphery

This chapter is launching soon

Genealogies of Data Journalism

Written by: C.W. Anderson

Introduction

Why should anyone care about the history of data journalism? Not only is "history" a rather academic and abstract topic for most people, it might seem particularly remote for working data journalists with a job to do. Journalists, working under tight deadlines and with a goal of conveying complicated information quickly and understandably to as many readers as possible, can be understandably averse to wasting too much time on self-reflection. More often than not, this reluctance to "navel-gaze" is an admirable quality; when it comes to the practices and concepts of data journalism and computational reporting, however, a hostility towards historical thinking can be a detriment that hampers the production of quality journalism itself.

Data journalism may be the most powerful form of collective journalistic sense making in the world today. At the very least, it may be the most positive and positivistic form of journalism. This *power* (the capacity of data journalism to create high-quality journalism, along with the rhetorical force of the data journalism model), *positivity* (most data journalists have high hopes for the future of their particular subfield, convinced it is on the rise) and *positivism* (data reporters are strong believers in the ability of method-guided research to capture real and provable facts about the world) create what I would call an empirically self-assured profession. One consequence of this self-assurance, I would argue, is that it can also create a whiggish assumption that data journalism is always improving and improving the world. Such an attitude can lead to arrogance and a lack of critical self-reflexivity, and make journalism more like the institutions it spends its time calling to account.

In this chapter I want to argue that a better attention to history can actually improve the day-to-day workings of data journalism. By understanding that their processes and practices have a history, data journalists can open their minds to the fact that things in the present could be done differently because they might have once been otherwise. In particular, data journalists might think harder about how to creatively represent uncertainty in their empirical work. They might consider techniques through which to draw in readers of different political sensibilities and persuasion that go beyond simple stating factual evidence. They might, in short, open themselves up to what Science and Technology Studies scholars and historians Catherine D'Ignazio and Lauren Klein have called a form of "feminist data visualization," one that rethinks binaries, embraces pluralism, examines power and considers context (D'Ignazio and Klein 2018; see also D'Ignazio's chapter in this book). To accomplish these changes, data journalism more than most forms of journalistic practice) should indeed inculcate this strong historical sensibility due to the very nature of its own power and self-assurance. No form of history is better equipped to lead to self-reflexivity, I would argue, than the genealogical approach to conceptual development pioneered by Michel Foucault, and embraced by some historians of science and scholars in science and technology studies.

"Genealogy," as defined by Foucault and who himself draws on the earlier work of Nietzsche, is a unique approach to studying the evolution of institutions and concepts over time and one that might be distinguished from history as such. Genealogical analysis does not look for a single, unbroken origin of practices or ideas in the past, nor does it try to understand how concepts developed in an unbroken and evolutionary line from yesterday to today. Rather, it focuses more on *discontinuity* and *unexpected changes* than it does on the presence of the past in the present. As Nietzsche noted, in a passage from the *Genealogy of Morals* quoted by Michel Foucault:

the "development" of a thing, a practice, or an organ has nothing to do with its progress towards a single goal, even less is it the logical and shortest progress reached with the least expenditure of power and resources. Rather, it is the sequence of more or less profound, more or less mutually independent processes of overpowering

that take place on that thing, together with the resistance that arises against that overpowering each time, the changes of form which have been attempted for the purpose of defense and reaction, as well as the results of successful counter-measures. Form is fluid; the "meaning," however, is even more so.[1]

A "genealogy of data journalism," then, would uncover the ways that data journalism evolved in ways that its creators and practitioners never anticipated, or in ways that may have even been contrary to their desires. It would look at the ways that history surprises us and sometimes leads us in unexpected directions. This approach, as I argued earlier, would be particularly useful for working data journalists of today. It would help them understand, I think that they are *not* working in a pre-defined tradition with a venerable past; rather, they are mostly making it up as they go along in ways that are radically contingent. And it would prompt a useful form of critical self-reflexivity, one that might help mitigate the (understandable and often well-deserved) self-confidence of working data journalists and reporters.

I have attempted to write such a genealogical account in my book, *Apostles of Certainty: Data Journalism and the Politics of Doubt*. In the pages that follow, I want to summarize some of the main findings of the book, and discuss ways that its lessons might be helpful for the present day. I want to conclude by arguing that journalism, particularly of the datafied kind, could and should do a better job demonstrating what it does not know, and that these gestures towards uncertainty would honor data journalism's origins in the critique of illegitimate power rather than the reification if it.

Data Journalism Through Time: 1910s, 1960s and 2000s

Can journalists use data – along with other forms of quantified information such as paper documents of figures, data visualizations, and charts and graphs –in order to produce better journalism? And how might that journalism assist the public in making better political choices? These were the main questions guiding *Apostles of Certainty: Data Journalism and the Politics of Doubt*, which tried to take a longer view of the history of news. With stops in the 1910s, the 1960s, and the present, the book traces the genealogy of data journalism and its material and technological underpinnings, and argues that the use of data in news reporting is inevitably intertwined with national politics, the evolution of computable databases, and the history of professional scientific fields. It is impossible to understand journalistic uses of data, I argue in the book, without understanding the oft-contentious relationships between social science and journalism. It is also impossible to disentangle empirical forms of public truth telling without first understanding the remarkably persistent Progressive belief that the publication of empirically verifiable information will lead to a more just and prosperous world. *Apostles of Certainty* concluded that this intersection of technology and professionalism has led to a better journalism but not necessarily to a better politics. To fully meet the demands of the digital age, journalism must be more comfortable expressing empirical doubt as well as certitude. Ironically, this "embrace of doubt" could lead journalism to become more like science, not less.

The Challenge of Social Science

The narrative of *Apostles of Certainty* grounds itself in three distinct U.S. time periods which provide three different perspectives on the development of data journalism. The first is the so-called "Progressive Era" which was a period of liberal political ascendancy accompanied by the belief that the both state *and* ordinary citizens, informed by the best statistics available, could make the world a more just and humane place. The second moment is the 1950s and 1960s, when a few journalism reformers began to look to quantitative social science, particularly political science and sociology, as a possible source of new ideas and methods for making journalism more empirical and objective. They would be aided in this quest by a new set of increasingly accessible databases and powerful computers. The third moment is the early 2010s, when the cutting edge of data journalism has been

supplemented by “computational” or “structured” journalism. In the current moment of big data and “deep machine learning,” these journalists claim that journalistic objectivity depends less on external referents but rather emerges from within the structure of the database itself.

In each of these periods, data-oriented journalism both *responded to* but also defined itself *in partial opposition to* larger currents operating within social science more generally, and this relationship to larger political and social currents helped inform the choice of cases I focused on in this chapter. In other words, I looked for inflection points in journalism history that could help shed light on larger social and political structures, in addition to journalism. In the Progressive Era[2], traditional news reporting largely rejected sociology’s emerging focus on social structures and de-personalized contextual information, preferring to retain their individualistic focus on powerful personalities and important events. As journalism and sociology professionalized, both became increasingly comfortable with making structural claims, but it was not until the 1960s that Philip Meyer and reformers clustered around the philosophy of Precision Journalism began to hold up quantitative sociology and political science as models for the level of exactitude and context to which journalism ought to aspire. By the turn of the 21st century, a largely normalized model of data journalism began to grapple with doubts about replicability and causality that were increasingly plaguing social science; like social science, it began to experiment to see if “big data” and non-causal forms of correlational behavioralism could provide insights into social activity.

Apostles of Certainty thus argues implicitly that forms of journalistic expertise and authority are never constructed in isolation or entirely internally to the journalistic field itself. Data journalism did not become data journalism for entirely professional journalistic reasons, nor can this process be analyzed solely through an analysis of journalistic discourse or “self-talk.” Rather, the type of expertise that in the 1960s began to be called data journalism can only be understood *relationally*, by examining the manner in which data journalists responded to and interacted with their (more authoritative and powerful) social scientific brethren. What’s more, this process cannot be understood solely in terms of the actions and struggles of humans, either in isolation or in groups. Expertise, according to the model I put forward in *Apostles of Certainty*, is a networked phenomenon in which professional groupings struggle to establish jurisdiction over a wide variety of discursive *and* material artifacts. Data journalism, to put it simply, would have been impossible without the existence of the database, but the database as mediated through a particular professional understanding of what a database was and how it could be deployed in ways that were properly journalistic (for a more general attempt at this argument about the networked nature of expertise, see Anderson 2013). It is impossible to understand journalistic authority without also understanding the authority of social science (and the same thing might be said about computer science, anthropology, or long-form narrative non-fiction). Journalistic professionalism and knowledge can never be understood solely by looking at the field of journalism itself.

The Persistence of Politics

Data journalism must be understood genealogically and in relation to adjacent expert fields like sociology and political science. All of these fields, in turn, must be analyzed through their larger conceptions of politics and how they come to terms with the fact that the “facts” they uncover are “political” whether they like it or not. Indeed, even the desire for factual knowledge is itself a political act. Throughout the history of data journalism, I argue in *Apostles of Certainty*, we have witnessed a distinct attempt to lean on the neutrality of social science in order to enact what can only be described as progressive political goals. The larger context in which this connection is forged, however, has shifted dramatically over time. These larger shifts should temper any enthusiasm that what we are witnessing in journalism is a teleological unfolding of journalistic certainty as enabled by increasingly sophisticated digital devices.

In the Progressive Era, proto-data journalists saw the gathering and piling up of quantitative facts as a process of social and political enlightenment, a process that was nonetheless free of any larger political commitments. By collecting granular facts about city sanitation levels, the distribution of poverty across urban spaces, statistics about church attendance and religious practice, labor conditions, and a variety of other bits of factual knowledge-- and by transmitting these facts to the public through the medium of the press-- social surveyors believed that the social organism would gain a more robust understanding of its own conditions of being. By gaining a better understanding of itself, society would improve, both of its own accord and by spurring politicians toward enact reformist measures. In this case, factual knowledge about the world spoke for itself; it simply needed to be gathered, visualized, and publicized and enlightenment would follow. We might call this a "naïve and transparent" notion of what facts are – they require no interpretation in and of themselves, and their accumulation will lead to positive social change. Data journalism, at this moment, could be political without explicitly stating its politics.

By the time of Philip Meyer and the 1960s, this easy congruence between transparent facts and politics had been shattered. Journalism was flawed, Meyer and his partisans argued throughout the 1950s and 1960s, because it mistook objectivity for simply collecting a record of what all sides of a political issue might think the truth might be and allowing the reader to make their own decisions about what was true. In an age of social upheaval and political turmoil, journalistic objectivity needed to find a more robust grounding, and it could find its footing on the terrain of objective social science. The starting point for journalistic reporting on an issue should not be the discursive claims of self-interested politicians but rather the cold, hard truth gleaned from an analysis of relevant data with the application of an appropriate method. Such an analysis would be *professional but not political*; by acting as a highly professionalized cadre of truth-tellers, journalists could cut through the political spin and help plant the public on the terrain of objective truth. The directions this truth might lead, on the other hand, were of no concern. Unlike the earlier generation of blissfully and naively progressive data journalists, the enlightened consequences of data were not a foregone conclusion.

Today I would argue that a new generation of computational journalists has unwittingly reabsorbed some of the political and epistemological beliefs of their Progressive Era forbearers. Epistemologically, there is an increasing belief amongst computational journalists that digital facts in some way "speak for themselves," or at least these facts will do so when they have been properly collected, sorted, and cleaned. At scale, and when linked to larger and internally consistent semantic databases, facts generate a kind *correlational excess* in which troubles with meaning or causality are washed away through a flood of computational data. Professionally, data journalists increasingly understand objectivity as emerging from within the structure of the database itself rather than as part of any larger occupational interpretive process. Politically, finally, I would argue that there has been the return of a kind of "crypto-progressivism" amongst many of the most studiously neutral data journalists, with a deep-seated political hope that more and more data, beautifully visualized and conveyed through a powerful press, can act as a break on the more irrational or pathological political tendencies increasingly manifest within western democracies. Such, at least, was the hope before 2016 and the twin shocks of Brexit and Donald Trump.

Certainty and Doubt

The development of data journalism in the United States across the large arc of the 20th century should be seen as one in which increasingly exact claims to journalistic professional certitude coexisted uneasily with a dawning awareness that all facts, no matter what their origins, were tainted with the grime of politics. These often-contradictory beliefs are evident across a variety of data-oriented fields, of course, not simply just in journalism. In a 2017 article for *The Atlantic*, for instance, science columnist Ed Yong grappled with how the movement toward "open science" and the growing replicability crisis could be used by an anti-scientific Congress to demean and defund scientific research. Yong quoted Christie Aschwanden, a science reporter at FiveThirtyEight: "it feels like there are two opposite things that the public thinks about science," she tells Yong. "[Either] it's a magic wand that

turns everything it touches to truth, or that it's all bullshit because what we used to think has changed … The truth is in between. Science is a process of uncertainty reduction. If you don't show that uncertainty is part of the process, you allow doubt-makers to take genuine uncertainty and use it to undermine things".[3] These thoughts align with the work of STS scholar Helga Nowotny, who argues in *The Cunning of Uncertainty* that "the interplay between overcoming uncertainty and striving for certainty underpins the wish to know" (Nowotny 2016). The essence of modern science—at least in its ideal form—is not the achievement of certainty but rather the fact that it so openly states the provisionality of its knowledge. Nothing in science is set in stone. It admits to often know little. It is through this, the most modern of paradoxes, that its claims to knowledge become worthy of public trust.

One of the insights provided by this genealogical overview of the development and deployment of data journalism, I would argue, is that data-oriented journalists have become obsessed with increasing exactitude and certainty at the expense of a more humble understanding of provisionality and doubt. As I have tried to demonstrate, since the middle of the 20th century journalists have engaged in an increasingly successful effort to render their knowledge claims more certain, contextual, and explanatory. In large part, they have done this by utilizing different forms of evidence, particularly evidence of the quantitative sort. Nevertheless, it should be clear that this heightened professionalism —and the increasing confidence of journalists that they are capable of making contextualized truth claims— has not always had the democratic outcomes that journalists expect. Modern American political discourse has tried to come to grips with the uncertainty of modernity by engaging a series of increasingly strident claims to certitude. Professional journalism has not solved this dilemma; rather it has exacerbated it. To better grapple with the complexity of the modern world, I would conclude, journalism ought to rethink the means and mechanisms by which it conveys its own provisionality and uncertainty. If done correctly, this could make journalism more like modern science, rather than less.

Works Cited

C. W. Anderson, 'Towards a Sociology of Computational and Algorithmic Journalism', *New Media & Society* 15:7, pp. 1005–1021, 2013.

C. W. Anderson, 'Apostles of Certainty: Data Journalism and the Politics of Doubt', New York, NY, *Oxford University Press*, 2018.

Michel Foucault, 'Power/Knowledge: Selected Interviews and Other Writings', New York, NY, *Vintage*,1980, 1972-1977.

Helga Nowotny, 'The Cunning of Uncertainty', London, UK: *Polity Press*, 2016.

Ed Yong, 'How the GOP Could Use Science's Reform Movement Against It', *The Atlantic*, 5 April 2017.

Data-Driven Gold-Standards: What the Field Values as Award-Worthy Data Journalism and How Journalism Co-Evolves with the Datafication of Society

Written by: Wiebke Loosen

Introduction: Journalism's response to the datafication of society

Perhaps better than in the early days of data journalism, we can understand the emergence of this new reporting style today as one journalistic response to the datafication of society.[1] Datafication refers to the ever-growing availability of data that has its roots in the digitalization of our (media) environment and the digital traces and big data that accrue with living in such an environment[2]. This process turns many aspects of our social life into computerized data — data that is to various ends aggregated and processed algorithmically. Datafication leads to a variety of consequences and manifests itself in different ways in politics, for instance, than it does in the financial world or in the realm of education. However, what all social domains have in common is that we can assume that they will increasingly rely on an ever more diverse range and greater amount of data in their (self-) sense making processes.

Situating the datafication of journalism in relation to the datafication of wider society helps us also to look beyond data journalism, to recognize it as "only" one occurrence of, and to better understand, journalism's transformation towards a more and more data-based, algorithmicized, metrics-driven, or even automated practice[3]. In particular, this includes the objects and topics that journalism is supposed to cover, or, put differently, journalism's function as an observer of society: The more the fields and social domains that journalism is supposed to cover are themselves 'datafied', the more journalism itself needs to be able to make sense of and produce data to fulfil its societal role. It is this relationship that is reflected in contemporary data journalism which relies on precisely this increased availability of data to expand the repertoire of sources for journalistic research and for identifying and telling stories.

Awards: A means to study what is defined and valued as data journalism

One way of tracing the evolution of data journalism as a reporting style is to look at its output. While the first studies in journalism research tended to focus more on the actors involved in its production and were mainly based on interviews, more and more studies have recently been using content analysis to better understand data journalism on the basis of its products[4]. Journalism awards are a good empirical access point for this purpose for several reasons: Firstly, award submissions have already proved to be useful objects for the analysis of genres and aspects of storytelling (e.g. Wahl-Jorgensen 2013).[5] Secondly, data journalism is a diffuse object of study that makes it not only difficult, but, rather, preconditional, to identify respective pieces for a content analysis. The sampling of award nominees, in turn, avoids starting with either a too narrow or too broad definition – this strategy is essentially a means of observing self-observation in journalism as such pieces represent what the field itself regards as data journalism and believes that they are significant examples of this reporting style. Thirdly, nominations for internationally oriented awards are likely to influence the development of the field as a whole as they are highly recognized, are considered to be a kind of gold-standard and as such also have a cross-border impact. In addition, looking at international awards allows us to investigate a sample that covers a broad geographical and temporal range.

However, it is also important to keep in mind that studying (journalism) awards brings with it different biases. The study we are drawing from here is based on an analysis of 225 nominated pieces (including 39 award-winning pieces) for the Data Journalism Awards (DJA) – a prize annually awarded by the Global Editors Network[6] – in the years 2013 to 2016[7]. This means that our sample is subject to a double selection bias: at first it is self-selective, since journalists have to submit their contributions themselves in order to be nominated at all. In the second step, a more or less annually changing jury of experts will decide which entries will actually be nominated. In addition, prizes and awards represent a particular form of "cultural capital" which is why award-winning projects can have a certain signal effect for the field as a whole and serve as a model for subsequent projects[8]. This also means that awards not only represent the field (according to certain standards), but also constitute it. That is, in our case, by labelling content as data journalism, the awards play a role in gathering together different practices, actors, conventions, and values. They may be considered, then, to have not just an *award-making function* but also a *field-making function*. This means that award-worthy pieces are always the result of a kind of "co-construction" by applicants and jurors and their mutually shaped expectations. Such effects are likely to be particularly influential in the case of data journalism as it is still a relatively new reporting style with which all actors in the field are more or less experimenting.

Evolving but not revolutionizing: Some trends in (award-worthy) data journalism

Studies that analyze data-driven pieces generally demonstrate that the evolution of data journalism is by no means a revolution in news work. As a result, they challenge the widespread belief that data-driven journalism is revolutionizing journalism by replacing traditional methods of news discovery and reporting. Our own study broadly concurs with what other empirical analyses of "daily" data journalism samples have found[9]. These only represent fairly limited data collections, but they do at least allow us to trace some developments and perhaps, above all, some degree of consistency in data journalism output.

In terms of who is producing data-driven journalism on an award-worthy level, results show that the 'gold-standard' for data journalism, that is, worthy of peer recognition, is dominated by newspapers and their online departments. Over the four years we analyzed, they represent by far the largest group among all nominees as well as among award-winners (total: 43.1%; DJA-awarded: 37.8%). The only other prominent grouping comprises organizations involved in investigative journalism such as *ProPublica* or the *The International Consortium of Investigative Journalists* (ICIJ), who were awarded significantly more often than not. This might reflect awards' inherent bias towards established, high-profile actors, echoing findings from other research that data journalism above a certain level appears to be an undertaking for larger organizations that have the resources and editorial commitment to invest in cross-disciplinary teams made up of writers, programmers and graphic designers[10]. This is also reflected in the team sizes: Of the 192 projects in our sample that had a byline, they named on average just over five individuals as authors or contributors and about a third of projects were completed in collaboration with external partners who either contributed to the analysis or designed visualizations. This seems particularly true for award-winning projects that our analysis found were produced by larger teams than those only nominated ($M = 6.31$, $SD = 4.7$ vs $M = 4.75$, $SD = 3.8$).

With regards to the geographies of data journalism that receives recognition in this competition, we can see that the United States dominates: nearly half of the nominees come from the U.S. (47.6%), followed at a distance by Great Britain (12.9%) and Germany (6.2%). However, data journalism appears to be an increasingly global phenomenon as the number of countries represented by the nominees grew with each year, amounting to 33 countries from all five continents in 2016.

Data journalism's reliance on certain sources influences the topics it may or may not cover. As a result, data journalism can neglect those social domains for which data is not regularly produced or accessible. In terms of topics covered, DJA-nominees are characterized by an invariable focus on political, societal, and economic issues

with almost half the analyzed pieces (48.2 percent) covering a political topic. The small share of stories on education, culture, and sports – in line with other studies – might be unrepresentative of data journalism in general and instead result from a bias towards 'serious' topics inherent in industry awards. However, this may also reflect the availability or unavailability of data sources for different domains and topics or, in the case of our sample, the applicants' self-selection biases informed by what they consider worthy of submission and what they expect jurors to appreciate. In order to gain more reliable knowledge on this point of crucial importance, an international comparative study that relates data availability and accessibility to topics covered by data reporting in different countries would be required. Such a study is still absent from the literature but could shed light on which social domains and topics are covered by which analytical methods and based on which data sources. Such an approach would also provide valuable insight to the other side of this coin: the blind spots in data-driven coverage due to a lack of (available) data sources.

One recurring finding in content-related research on data journalism is that it exhibits a 'dependency on pre-processed public data' from statistical offices and other governmental institutions[11]. This is also true of data-driven pieces at an award-worthy level: we observed a dependence on data from official institutions (almost 70% of data sources) or other non-commercial organizations such as research institutes, NGOs and so on as well as data that are publicly available, at least, on request (almost 45%). This illustrates, on the one hand, that data journalism is making sense of the increased availability of data sources, but on the other, that it also relies heavily on this availability: the share of self-collected, scraped, leaked, and requested data is substantially smaller. Nonetheless, data journalism has been continually linked to investigative reporting, which has 'led to something of a perception that data journalism is all about massive data sets, acquired through acts of journalistic bravery and derring-do'[12]. Recent cases such as the 'Panama Papers' have contributed to that perception[13]. However, what this case also shows is that some complex issues of global importance are embedded in data that require transnational cooperation between different media organizations. Furthermore, it is likely that we will see more of these cases as soon as routines can be further developed to continuously monitor international data flows, for example in finance, not merely as a service, but also as deeper and investigative background stories. That could stimulate a new kind of *investigative data-based real-time journalism*, which constantly monitors certain finance data streams, for example, and searches for anomalies.

Interactivity counts as quality criterion in data journalism, but interactivity is usually implemented with a relatively clear set of features – here our results are also in harmony with other studies and what is often described as a "lack of sophistication" in data-related interactivity[14]. Zoomable maps and filter functions are most common, perhaps because of a tendency to apply easy-to-use and/or freely available software solutions which results in less sophisticated visualizations and interactive features. However, award-winning projects are more likely to provide at least one interactive feature and integrate a higher number of different visualizations. The trend towards rather limited interactive options might also reflect journalists' experiences with low audience interest in sophisticated interactivity (such as gamified interactivity opportunities or personalisation tools that make it possible to tailor a piece with customised data). At the same time, however, interactive functions as well as visualizations should at best support the storytelling and the explanatory function of an article - and this requires solutions adapted to each data-driven piece.

A summary of the developmental trends over the years shows a somewhat mixed pattern as the shares and average numbers of the categories under study were mostly stable over time or, if they did change, they did not increase or decrease in a linear fashion. Rather, we found erratic peaks and lows in individual years, suggesting the trial-and-error evolution one would expect in a still emerging field such as data journalism. As such, we found few consistent developments over the years: a significantly growing share of business pieces, a consistently and significantly increasing average number of different kinds of visualisations and a (not statistically significant, but) constantly growing portion of pieces that included criticism (e.g. on the police's wrongful confiscation methods) or

even calls for public intervention (e.g. with respect to carbon emissions). This share grew consistently over the four years (2013: 46.4% vs 2016: 63.0%) and was considerably higher among award winners (62.2% vs 50.0%). We can interpret this as an indication of the high appreciation of the investigative and watchdog potential of (data) journalism and, perhaps, as a way of legitimizing this emerging field.

From data journalism to datafied journalism - and its role in the data society

Data journalism represents the emergence of a new journalistic sub-field that is co-evolving in parallel with the datafication of society — a logical step in journalism's adaptation to the increasing availability of data. However, data journalism is no longer a burgeoning phenomenon, it has, in fact, firmly positioned itself within mainstream practice. A noteworthy indicator of this can again be found when looking at the Data Journalism Awards: the 2018 competition introduced a new category called "innovation in data journalism"; it appears that data journalism is no longer regarded as an innovative field in and of itself, but is already looking for innovative approaches in contemporary practice[15].

We can expect data journalism's relevance and proliferation to co-evolve alongside the increasing datafication of society as a whole – a society in which sense making, decisions, and all kinds of social actions increasingly rely on data. Against this background, it is not too difficult to see that the term "data journalism" will become superfluous in the not too distant future because journalism as a whole, as well as the environment of which it is part, is becoming increasingly datafied. Whether this prognosis is confirmed or not: the term "data journalism", just as the term "data society", still sensitizes us to fundamental transformation processes in journalism and beyond. This includes how and by what means journalism observes and covers (the datafied) society, how it self-monitors its performance, how it controls its reach and audience participation, and how it (automatically) produces and distributes content. In other words, contemporary journalism is characterized by its transformation towards a more data-based, algorithmicized, metric-driven, or even automated practice.

However, data is not a "raw material"; it does not allow direct, objective or otherwise privileged access to the social world[16]. This circumstance is all the more important for a responsible data journalism as the process of society's datafication advances. Advancing datafication and data-driven journalism's growing relevance may also set incentives for other social domains to produce or make more data available (to journalists) and we are likely to see the co-evolution of a 'data PR', that is, *data-driven public relations* produced and released to influence public communications for its own purposes. This means that routines for checking the quality, origin and significance of data are becoming increasingly important for (data) journalism and raises the question of why there may be no data available on certain facts or developments.

In summary, I can organize our findings according to seven 'Cs' - seven challenges and underutilized capacities of data journalism that may also be useful for suggesting modified or alternative practices in the field:

1. **Collection**: Investigative and critical data journalism must overcome its dependency on publicly accessible data. More effort needs to be made in gaining access to data and collecting them independently.
2. **Collaboration**: Even if the 'everyday' data-driven piece is becoming increasingly easier to produce; more demanding projects are resource and personnel intensive and it is to be expected that the number of globally relevant topics will increase. These will require data-based investigations across borders and media organisations, and, in some cases, collaboration with other fields such as science or data activism.
3. **Crowdsourcing**: The real interactive potential of data journalism lies not in increasingly sophisticated interactive features but in crowdsourcing approaches that sincerely involve users or citizens as collectors, categorizers, and co-investigators of data.[17]
4. **Co-Creation**: Co-creation approaches, as they are already common in the field of software development, can serve as a model for long-term data-driven projects. In such cases, users are integrated into the entire

process, from finding a topic to developing one and maintaining it over a longer period.

5. **Competencies**: Quality data journalism requires teams with broad skill sets. The role of the journalist remains important, but they increasingly need a more sophisticated understanding of data, data structures, and analytical methods. Media organizations, in turn, need resources to recruit data analysts who are increasingly desirable in many other industries.
6. **Combination**: Increasingly complex data requires increasingly sophisticated analysis. Methods that combine data sources and look at these data from a variety of positions could help paint more substantial pictures of social phenomena and strengthen data journalism's analytical capacity.
7. **Complexity**: Complexity includes not only the data itself, but its increasing importance for various social areas, political decision-making, etc., as well; in the course of these developments, data journalism will increasingly be confronted with data PR and 'fake data'.

What does this mean? Taking into account what we already know about (award-winning) data journalism in terms of what kinds of data journalism are valued, receiving wide public attention (such as the Panama Papers), and contributing to a general appreciation of journalism, what kinds of data journalism do we really want? In this regard, I would argue that data journalism is particularly relevant in its unique role as a responsible kind of journalism as part of the data society; that is data journalism:

- That investigates socially relevant issues and makes the data society understandable and criticizable by its own means;
- Aware of its own blind spots while asking why there are data deficiencies in certain areas and whether this is a good or a bad sign;
- Actively tries to uncover data manipulation and data abuse;
- Keeps in mind, explains, and emphasizes the character of data as "human artifacts" that are by no means self-evident collections of facts, but are often collected in relation to very particular conditions and objectives[18].

At the same time, however, this means that data journalism's peculiarity, its dependency on data, is also its weakness. This limitation concerns the availability of data, its reliability, its quality, and its manipulability. A responsible data journalism should be reflexive about its dependency on data - and it should be a core subject in the discussion on ethics in data journalism. These conditions indicate that data journalism is not only a new style of reporting, but also a means of intervention that challenges and questions the data society, a society loaded with core epistemological questions that confront (not only) journalism's assumptions about *what* we (can) know and *how* we know (through data).

These questions become more urgent as more and increasingly diverse data is incorporated at various points in the "circuit of news": as a means of journalistic observation and investigation, as part of production and distribution routines, and as a means of monitoring the consumption activities of audiences. It is in these ways that datafied journalism is affecting: (1) *journalism's way of observing the world* and constructing the news from data; (2) the very core of journalism's performance in *facilitating the automation of content production*; (3) the *distribution and circulation of journalism's output* within an environment that is shaped by algorithms and their underlying logic to process data; (4) what is *understood as newsworthy* to increasingly granularly measured audience segments.

These developments present (data) journalism with three essential responsibilities: to critically observe our development towards a datafied society, to make it understandable through its own means, and to make visible the limits of what can and should be recounted and seen through the lens of data.

Works Cited

Julian Ausserhofer, Robert Gutounig, Michael Oppermann, Sarah Matiasek, Eva Goldgruber, 'The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork', *Journalism*, (2017).

Eddy Borges-Rey, 'Towards an epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity', *Journalism*, (2017).

Christine L. Borgman, 'Big data, little data, no data: Scholarship in the networked world', *MIT Press, Cambridge*, 2015.

James F English, 'Winning the Culture Game: prizes, awards, and the rules of art', *New Literary History* 33(1), (2002), pp. 109-135.

Megan Knight, 'Data journalism in the UK: A preliminary analysis of form and content', *Journal of Media Practice* 16(1), (2015), pp. 55–72.

Klaus Krippendorf, 'Data', in *The International Encyclopedia of Communication Theory and Philosophy*, ed. By Klaus Bruhn Jensen and Robert T. Craig, Volume 1 A-D, (Wiley Blackwell, 2016), pp. 484-489.

Wiebke Loosen, 'Four forms of datafied journalism. Journalism's response to the datafication of society', *Communicative Figurations*, (Working Paper No. 18), (2018).

Wiebke Loosen, Julius Reimer and Fenja De Silva-Schmidt, 'Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the data journalism awards 2013-2016', *Journalism*, (2017).

Sylvain Parasie, 'Data-driven revelation? Epistemological tensions in investigative journalism in the age of 'big data'', *Digital Journalism* 3(3), (2015), pp. 364-380.

Cindy Royal and Dale Blasingame, 'Data journalism: An explication', *#ISOJ* 5(1), (2015), pp. 24-46.

Constance Tabary, Anne-Marie Provost, Alexandre Trottier, 'Data journalism's actors, practices and skills: A case study from Quebec', *Journalism: Theory, Practice, and Criticism* 17(1), (2016), pp. 66-84.

Jose van Dijck, 'Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology', *Surveillance & Society* 12(2), (2014), pp. 197–208.

Karin Wahl-Jorgensen, 'The strategic ritual of emotionality: a case study of Pulitzer Prize-winning articles', *Journalism: Theory, Practice, and Criticism* 14(1), (2013), pp. 129–145.

Mary Lynn Young, Alfred Hermida and Johanna Fulda, 'What makes for great data journalism? A content analysis of data journalism awards finalists 2012–2015', *Journalism Practice*, (2017), pp. 115-135.

Data Journalism with Impact

Written by: Paul Bradshaw

If you've not seen *Spotlight*, the film about the Boston Globe's investigation into institutional silence over child abuse, then you should watch it right now. More to point – you should watch right through to the title cards at the end.[1]

A list scrolls down the screen. It details the dozens and dozens of places where abuse scandals have been uncovered since the events of the film, from Akute, Nigeria, to Wollongong, Australia. But the title cards also cause us to pause in our celebrations: one of the key figures involved in the scandal, it says, was reassigned to "one of the highest ranking Roman Catholic churches in the world."

This is the challenge of impact in data journalism: is raising awareness of a problem "impact"? Does the story have to result in penalty or reward? Visible policy change? How important is impact? And to whom?

These last two questions are worth tackling first. Traditionally impact has been important for two main reasons: commercial, and cultural. Commercially, measures of impact such as brand awareness and high audience figures can contribute directly to a publication's profit margin through advertising (increasing both price and volume) and subscription/copy sales[2]. Culturally, however, stories with impact have also given news organisations and individual journalists 'bragging rights' among their peers. Both, as we shall see, have become more complicated.

Measurements of impact in journalism have, historically, been limited: aggregate sales and audience figures, a limited pool of industry prizes and the occasional audience survey were all that publishers could draw on.

Now, of course, the challenge lies not only in a proliferation of metrics, but in a proliferation of business models, too, with the expansion of non-profit news provision in particular leading to an increasing emphasis on impact and discussion about how that might be measured[3].

Furthermore, the ability to measure impact on a story-by-story basis has meant it is no longer editors that are held responsible for audience impact, but journalists too.

Measuring impact by the numbers

Perhaps the easiest measure of impact is sheer *reach*: data-driven interactives like the BBC's '*7 billion people and you: What's your number?*' engaged millions of readers in a topical story; while at one point in 2012 Nate Silver's data journalism was reaching one in five visitors to the New York Times.[4]

Some will sneer at such crude measures — but they are important. If journalists were once criticised for trying to impress their peers at the expense of their audience, modern journalism is at least expected to prove that it can connect with that audience. In most cases this proof is needed for advertisers, but even publicly-funded universal news providers like the BBC need it too, to demonstrate that they are meeting requirements for funding.

Engagement is reach's more sophisticated relation, and here data journalism does well too: at one editors' conference for newspaper publisher Reach, for example, it was revealed that simply adding a piece of data visualisation to a page can increase dwell time (the amount of time a person spends on a page) by a third. Data-driven interactivity can transform the dullest of subjects: in 2015 the same company's David Higgerson noted that more than 200,000 people put their postcodes into an interactive widget by their data team based on deprivation statistics — a far higher number, he pointed out, "than I would imagine [for] a straight-forward 'data tells us x' story"[5].

Engagement is particularly important to organisations who rely on advertising (rates can be increased where engagement is high). but also those for whom subscriptions, donations and events are important: these tend to be connected with engagement too.

The expansion of non-profit funding and grants often comes with an explicit requirement to monitor or demonstrate impact which is about more than just reach. *Change* and *action* in particular - political or legal – are often referenced. The International Consortium of Investigative Journalists (ICIJ), for example, highlight the impact of their Panama Papers investigation in the fact that it resulted in "at least 150 inquiries, audits or investigations … in 79 countries", alongside the more traditional metric of almost 20 awards, including the Pulitzer Prize.[6] In the UK a special place is reserved in data journalism history for the MPs' expenses scandal. This not only saw The Telegraph newspaper leading the news agenda for weeks, but also led to the formation of a new body: the Independent Parliamentary Standards Authority (IPSA). The body now publishes open data on politicians' expense claims, allowing them to be better held to account and leading to further data journalism.

But policy can be much broader than politics. The lending policies of banks affect millions of people, and were famously held to account in the late-1980s in the US by Bill Dedman in his Pulitzer-winning "Color of Money" series of articles. In identifying racially divided loan practices ("redlining") the data-driven investigation also led to political, financial and legal change, with probes, new financing, lawsuits and the passing of new laws among the follow-ups.[7]

Fast-forward 30 years and you can see a very modern version of this approach: ProPublica's Machine Bias series shines a light on algorithmic accountability, while the Bureau Local tapped into its network to crowdsource information on algorithmically targeted 'dark ads' on social media.[8] Both have helped contribute to change in a number of Facebook's policies, while ProPublica's methods were adopted by a fair housing group in establishing the basis for a lawsuit against the social network.[9] As the policies of algorithms become increasingly powerful in our lives — from influencing the allocation of police to Uber pricing in non-white areas — holding these to account is becoming as important as holding more traditional political forms of power to account, too.[10]

What is notable about some of these examples is that their impact relies upon — and is partly demonstrated by — collaboration with others. When the Bureau Local talk about impact, for example, they refer to the numbers of stories produced by members of its grassroots network, inspiring others to action, while the ICIJ lists the growing scale of its networks: "LuxLeaks (2014) involved more than 80 reporters in 26 countries.[11] Swiss Leaks (2015) more than 140 reporters in 45 countries". The figure rises to more than 370 reporters in nearly 80 countries for the Panama Papers investigation: 100 media organisations publishing 4,700 articles.[12]

What's more, the data gathered and published as a result of investigations can become a source of impact itself: The Offshore Leaks database, the ICIJ points out, "is used regularly by academics, NGOs and tax agencies".

There is something notable about this shift from the pride of publishing to winning plaudits for acting as facilitators and organisers and database managers. As a result, collaboration has become a skill in itself: many non-profit organisations have community or project management roles dedicated to building and maintaining relationships with contributors and partners, and journalism training increasingly reflects this shift too.

Some of this can be traced back to the influence of early data journalism culture: writing about the practice in Canada in 2016, Alfred Hermida and Mary Lynn Young noted "an evolving division of labour that prioritizes inter-organisational networked journalism relationships".[13] And the influence was recognised further in 2018 when the Reuters Institute published a book on the rise of collaborative journalism, noting that "collaboration can become a story in itself, further increasing the impact of the journalism".[14]

Changing what we count, how we count it, and whether we get it right

Advanced technical skills are not necessarily required to create a story with impact. One of the longest-running data journalism projects, the Bureau of Investigative Journalism's Drone Warfare project has been tracking US drone strikes for over 5 years.[15] Its core methodology boils down to one word: persistence.[16] On a weekly basis Bureau reporters have turned 'free text' reports into a structured dataset that can be analysed, searched, and queried. That data — complemented by interviews with sources — has been used by NGOs and the Bureau has submitted written evidence to the UK's Defence Committee[17].

Counting the uncounted is a particularly important way that data journalism can make an impact — indeed, it is probably fair to say that it is data journalism's equivalent of 'giving a voice to the voiceless'. The Migrants Files, a project involving journalists from over 15 countries, was started after data journalists noted that there was "no usable database of people who died in their attempt to reach or stay in Europe".[18] Its impact has been to force other agencies into action: the International Organization for Migration and others now collect their own data.

Even when a government appears to be counting something, it can be worth investigating. While working with the BBC England Data Unit on an investigation into the scale of library cuts, for example, I experienced a moment of panic when I saw that a question was being asked in Parliament for data about the issue.[19] Would the response scoop the months of work we had been doing? In fact, it didn't — instead, it established that the government itself knew less than we did about the true scale of those cuts, because they hadn't undertaken the depth of investigation that we had.

And sometimes the impact lies not in the mere existence of data, but in its representation: one project by Mexican newspaper El Universal, Ausencias Ignoradas (Ignored Absences), puts a face to over 4,500 women who have gone missing in the country in a decade.[20] The data was there, but it hadn't been broken down to a 'human' level. Libération's Meurtres conjugaux, des vies derrière les chiffres does the same thing for domestic murders of women, and Ceyda Ulukaya's Kadin Cinayetleri project has mapped femicides in Turkey.[21]

When data is bad: impacting data quality

Some of my favourite projects as a data journalist have been those which highlighted, or led to the identification of, flawed or missing data. In 2016 the BBC England Data Unit looked at how many academy schools were following rules on transparency: we picked a random sample of 100 academies and checked to see if they published a register of all their governors' interests, as required by official rules. One in five academies failed to do so — and as a result the regulator Ofcom took action against those we'd identified.[22] But were they serious about ensuring this would continue? Returning to the story in later years would be important in establishing whether the impact was merely short-term, or more systemic.

Sometimes the impact of a data journalism project is a byproduct — only identified when the story is ready and responses are being sought. When the Bureau Local appeared to find that 18 councils in England had nothing held over in their reserves to protect against financial uncertainty, and sought a response, it turned out the data was wrong.[23] No-one noticed the incorrect data, they reported. "Not the councils that compiled the figures, nor the Ministry of Housing, Communities and Local Government, which vetted and then released [them]". Their investigation has added to a growing campaign for local bodies to publish data more consistently, more openly, and more accurately.

Impact beyond innovation

As data journalism has become more routine, and more integrated into ever-complex business models, its impact has shifted from the sphere of innovation to that of delivery. As data editor David Ottewell wrote of the distinction in 2018:

> "Innovation is getting data journalism on a front-page. Delivery is getting it on the front page day after day. Innovation is building a snazzy interactive that allows readers to explore and understand an important issue. Delivery is doing that, and getting large numbers of people to actually use it; then building another one the next day, and another the day after that."[24]

Delivery is also, of course, about impact beyond our peers, beyond the 'wow' factor of a striking datavis or interactive map — on the real world. It may be immediate, obvious and measurable, or it may be slow-burning, under the radar and diffuse. Sometimes we can feel like we didn't make a difference — as in the case of the Boston Globe's Catholic priest — but change can take time: reporting can sow the seeds of change, with results coming years or decades later. The Bureau Local and BBC do not know if council or schools data will be more reliable in future — but they do know that the spotlight is on both to improve.

Sometimes shining a spotlight and accepting that it is the responsibility of others to take action is all that journalism can do; sometimes it takes action itself, and campaigns for greater openness. To this data journalism adds the ability to force greater openness, or create the tools that make it possible for others to take action.

Ultimately, data journalism with impact can set the agenda. It reaches audiences that other journalism does not reach, and engages them in ways that other journalism does not. It gives a voice to the voiceless, and shines a light on information which would otherwise remain obscure. It holds data to account, and speaks truth to its power.

Some of this impact is quantifiable, and some has been harder to measure — and any attempt to monitor impact should bear this in mind. But that doesn't mean we shouldn't try...

Works Cited

Alan Rusbridger, 'Alan Rusbridger: Who Broke the News?', *The Guardian*, 31 August 2018.

Christoph Schlemmer, 'Speed is Not Everything: How News Agencies Use Audience Metrics', *Reuters Institute for the Study of Journalism*, (2016).

Elia Powers, 'Selecting Metrics, Reflecting Norms', *Digital Journalism* 6:4, (2018), pp. 454-471.

Dylan Byers, '20% of NYT Visitors Read 538', *Politico*, 11 June 2012.

David Higgerson, 'How Audience Metrics Dispel The Myth That Readers Don't Want To Get Involved With Serious Stories' 14 October 2015.

Will Fitzgibbon and Emilia Diaz-Struck, 'Panama Papers Have Had Historic Global Effect – And the Impacts Keep Coming', *ICIJ*, 1 December 2016.

Maeve McClenagan, 'Campaigners Target Voters With Brexit "Dark Ads"', *The Bureau of Investigative Journalism*, 18 May 2017.

Mary Clare Jalonick, 'Facebook Vows More Transparency Over Political Ads', *The Seattle Times*, 27 October 2017.

Maurice Chammah, 'Policing The Future', *The Marshall Project*, 2 March 2016.

Jennifer Stark, 'Investigating Uber Surge Pricing: A Data Journalism Case Study', Global Investigative Journalism Network, 2 May 2016.

Mar Cabra, 'How ICIJ went from having no data team to being a tech-driven media organisation', ICIJ, 29 November 2017.

Uri Blau, 'How Some 370 Journalists in 80 Countries Made the Panama Papers Happen', *Nieman Reports*, 6 April 2016.

Alfred Herminda and Mary Lynn Young, 'Finding The Data Unicorn', *Digital Journalism* 5:2, (2016), pp. 159-176.

Richard Sambrook, 'Global Teamwork: The Rise of Collaboration in Investigative Journalism', eds. By Richard Sambrook, *Reuters Institute for the Study of Journalism*, 2018.

BBC, 'Libraries Lose a Quarter of Staff As Hundreds Close', *BBC News*, 29 March 2016.

Maria Crosas Batista, 'How One Mexican Data Team Uncovered the Story of 4,000 Missing Women', *Online Journalism Blog*, June 2016.

Data Driven Journalism, 'Kadincinayetleri.org: Putting femicide on the map', *Data Driven Journalism*, 10 February 2016.

BBC, 'Academy Schools Breach Transparency Rules', BBC News, 18 November 2016.

Gareth Davies, 'Inaccurate and Unchecked: Problems with Local Council Spending Data', *The Bureau of Investigative Journalism*, 2 May 2018.

David Ottewell, 'The Evolution of Data Journalism', Towards Data Science, 28 March 2018.

Beyond Clicks and Shares: How and Why to Measure the Impact of Data Journalism Projects

Written by: Lindsay Green-Barber

Journalism and impact

While many journalists balk at the idea of journalistic impact, in fact, contemporary journalism, as a profession, is built on a foundation of impact: to inform the public so we can be civically engaged and hold the powerful to account. And while journalists worry that thinking about, talking about, strategizing for, and measuring the positive (and negative) impact of their work will get too close to crossing the red line from journalism into advocacy, practitioners and commentators alike have spent many column inches and pixels hand wringing about the negative effects of "fake news," misinformation, and partisan reporting on individuals, our society, and democracy. In other words, while journalists want to avoid talking about the impact of their work, they recognize the serious social, political, and cultural impacts of "fake news."

What's more, prior to the the professionalization of journalism in the late 1800s and early 1900s, journalism was a practice in influence, supported by political parties and produced with the express goal of supporting the party and ensuring its candidates were elected.[1] Thus, in an historical perspective, journalism's professionalization and embrace of (the myth of) neutrality are actually quite new.[2] And journalism's striving for "neutrality" was not a normative decision, but rather a function of changing economic models and a need to appeal to the largest possible audience in order to generate revenue.[3]

Given the concurrent and intimately related crises of the news industry business model and lack of public trust in media in the United States and Western Europe, one might argue that journalism's turn away from acknowledging its impact has been an abdication of responsibility, at best, and a failure, at worst.

But there are signs of hope. In recent years, some media organizations have begun to embrace the fact that they are influential in society. The proliferation of nonprofit media, often supported by mission driven philanthropic foundations and individuals, has created a Petri dish for impact experimentation. Many commercial media have also come around to the idea that communicating the positive impact of their work with audiences is a strategy for building trust and loyalty, which will hopefully translate into increases in revenue. For example, in 2017, the Washington Post added "Democracy dies in darkness" to its masthead, embracing (and advertising) its role in our political system. And CNN created an "Impact Your World" section on its website, connecting world events, its reporting, stories of "impact," and pathways for audience members to take action, from hashtag campaigns to donations.[4]

Media organizations have also begun to try new strategies to maximize the positive impact of their work, as well as to use non-advertising metrics and research methods to understand the effectiveness of these strategies. While, in some cases, digital metrics can be useful proxies for impact measurement, advertising metrics like unique page views or even more advanced analytics like time spent on a page are meant to measure the reach of content without consideration of the effects of this content on an individual.

I would like to propose a framework for media impact, that is a change in the status quo as a result of an intervention, that includes four types of impact: on individuals; on networks; on institutions; and on public discourse. These types of impact are interrelated. For example, as journalism often assumes, reporting can increase individuals' level of knowledge about an issue, resulting in them voting in a particular way and ultimately affecting

institutions. Or, a report may have immediate effects on institutions, such as a firing or a restructuring, which then trickles down to impact individuals. However, impact that is catalyzed by journalism often takes time and involves complex social processes.

Different types of journalism are better equipped for different types of impact. For example, James T. Hamilton shows that investigative reporting can save institutions' money by uncovering malfeasance, corruption, or wrongdoing and spurring change. And documentary film has proven to be particularly effective in generating new and/or strengthened advocacy networks to promote change.[5]

The remainder of this chapter explores the relationship between data journalism and impact, demonstrating how data journalism can contribute to various types of social change. It then suggests methods for how data journalism's effectiveness might be measured, and what journalists and news organizations can do with this information.

Why data journalism

While journalists employ data journalism for many reasons, there are two that come to the fore: first, to provide credible evidence to support claims made in storytelling; and second, to present information to audiences as data, rather than text-based narrative. The practice of data journalism is built on a foundational value judgement that data are credible, and by extension, a journalistic product that includes data reporting is credible - and potentially more so than it would be without.

Data reporting that is used to communicate information as static numbers, data, charts, graphs, or other visuals is similar to other journalistic formats (i.e., text, video, audio) in that it is essentially a linear form of communicating selected information to an audience. Data reporting that is made available to audiences through a news interactive is a unique form of storytelling in that it assumes an audience person will interact with the data, ask their own questions, and search for answers in the the data at hand. Thus, the "story" depends upon the user as much as it does on the journalism.

Even this rough hewn version of data journalism implicates all four types of impact.

Individuals

Data journalism tends to focus on individual audience members as the potential unit for change, providing audiences with credible information so that they may become more knowledgeable and, by extension, make more informed decisions. And while data journalism as a scaffolding for traditional, linear storytelling increases audience trust in the content, news or data interactives provide the greatest potential for data journalism to have an impact at the level of individuals.

With a data interactive, that is a "big interactive database that tells a news story", a user can generate their own question and query the data to look for answers.[6] Media companies often assume that data interactives will allow audience to do deep dives and explore data, find relevant information, and tell stories. In an analysis of data interactives by one news organization, the author of this chapter found that the most successful data apps, meaning those that were highly trafficked and deeply explored, were part of a full editorial package that included other content, have the ability to look up geographically local or relevant data, have a high degree of interactivity, are aesthetically pleasing and well-designed, and that load quickly.[7]

ProPublica's *Dollars for Docs* is a classic example of data journalism in that it accesses significant amounts of data, in this case about pharmaceutical and medical device companies' payments to doctors, structures the data, and presents it to audiences as an interactive database with the goal to inspire individuals to conduct their own

research and possibly take action[8]. The project instructs audience to “use this tool” to search for payments to their doctors, and, in a sidebar, says, “Patients, Take Action. We want to know how you've used or might use this information in your day to day lives. Have you talked to your doctor? Do you plan to? Tell us”.[9]

Networks

Data journalism provides credible information that can be used by networks (formal and/or informal) to strengthen their positions and work. For example, advocacy organizations often use data reporting to bolster their claims in public appeals or in legal proceedings, especially in cases where the data are not publicly available. Journalism’s practice of requesting access to data that are not available in the public realm, analyzing these data, and publishing the findings, absorbs costs that would otherwise be insurmountable for individuals or networks.[10]

Institutions

Data journalism can generate reporting that institutions work hard to keep hidden, as they are evidence of corruption, malfeasance, wrongdoing, and/or incompetence. When this information comes to light, there is pressure on institutions to reform - resulting from the threats associated with elections on politicians or market forces on publicly held companies.

For example, the International Consortium of Investigative Journalism’s *Panama Papers* collaborative investigation analyzed more than 11.5 million to uncover “politicians from more than 50 countries connected to offshore companies in 21 tax havens.”.[11] This investigation led to the resignation of politicians, such as Iceland’s Prime Minister Sigmundur David Gunnlaugsson, investigations of others, like Pakistan’s former Prime Minister Former Pakistan Prime Minister Nawaz Sharif (who was sentenced to ten years in jail in 2018), and countless other institutional responses.

Public discourse

Because data journalism can often be broken down into smaller parts, whether geographically, demographically, or by other factors, the data can be used to tell different stories by different media. In this way, data journalism can be localized to generate a shift in public conversation about issues across geographic locations, demographic groups, or other social boundaries.

The Center for Investigative Reporting has published national interactive datasets about the Department of Veterans Affairs, one with average wait times for veterans trying to access medical care at VA hospitals, and a second with the number of opiates being prescribed to veterans by VA systems. In both cases, local journalism organizations used the datasets as the baseline to do local reporting about the issues.

So, how can data journalists strategize for impact?

You’ve done the hard work: you got access to data, you crunched the numbers, structured the data, and you have an important story to tell. Now what?

A high-impact strategy for data journalism might follow the following five steps:

1. **Set goals**

What might happen as a result of your project? Who or what has the power an/or incentive to address any wrongdoing? Who should have access to the information you’re bringing to light? Ask yourself these questions to decide what type or types of impact are reasonable for your project.

Once you have goals for your project, identify the important target audiences for the work. What source of news and information do these audiences trust? How might they best access the information? Do they need an interactive, or will a linear story more effective?

How will you and your news organization engage with audiences, and how will audiences engage with your work? For example, if you've identified a news organization other than your own as a trusted source of information for a target audience, collaborate. If your data interactive has important information for an NGO community, hold a webinar explaining how to use it.

Depending upon your goals and content and engagement plans, select the appropriate research methods and/or indicators in order to track progress and understand what's working and what's not working. While media often refer to "measuring" the impact of their work, I prefer the term "strategic research," as both qualitative and quantitative research methods should be considered. The sooner you can identify research methods and indicators, the better your information will be. (The following section discusses measurement options in greater depth.)

You've invested time and resources in your data journalism reporting, content, engagement, and measurement. What worked? What will you change next time? What questions are still outstanding? Share these learnings with your team and the field to push the next project further ahead.

How do we "measure" the impact of our work?

As alluded to earlier, media impact research has been dominated by advertising metrics. However, ad metrics, like page views, time on page, and bounce rate are potential proxies for some impact. They are meant to measure the total exposure of content to individuals without concern for their opinions about the issues, whether or not they have learned new information, or their intent to take action based upon the content. When considering the impact of content on individuals, networks, institutions, and public discourse, there are other innovative qualitative and quantitative methods that can be used to better understand the impact of reporting on individuals, networks, institutions, and public discourse. This section explores a handful of promising research methods for understanding the impact of data journalism.

Analytics

Media metrics can be used as proxies to for desired outcomes like increased awareness or increased knowledge. However, media companies should be intentional and cautious when attributing change to analytics. For example, if a data journalism project has as its goal to spur institutional change, unique page views are not an appropriate metric of success; mentions of the data by public officials in documents would be a better indicator.

Experimental research

Experimental research creates constant conditions under which the effects of an intervention can be tested. University of Texas Austin's Center for Media Engagement has conducted fascinating experimental research about the effects of news homepage layout on audience recall and affect, and of solutions-oriented reporting on audience affect for news organizations. Technology companies are constantly testing the effects of different interactive elements on users. Journalism organizations can do the same to better understand the effects of data interactives on users, whether in partnership with universities or by working directly with researchers in-house from areas like marketing, business development, and audience engagement.

Surveys

Surveys, while not the most leading edge research method, are a proven way to gather information from individuals about changes in interest, knowledge, opinion, and action. Organizations can be creative with survey design, making use of technology that allows for things like return visit triggered pop-ups or tracking newsletter click through to generate a survey pool of potential respondents.

Content analysis

Content analysis is a research method used to determine changes in discourse, over time. This method can be employed to any text-based corpus, making it extremely flexible. For example, when an organization produces content with the goal of influencing national public discourse, it could conduct a post-project content analysis on the top ten national newspapers to determine the influence of its stories. If the goal is to influence a state legislature, an organization can use post-project content analysis on publicly available legislative agendas.[12] Or, if the goal is to make data available to advocacy networks, post-project content analysis could be used to analyze an organization's newsletters.

Content analysis can be conducted in at least three ways. At the most basic level, a news organization can search for a project's citations in order to document where and when it has been cited. For example, many reporters create Google news alerts using a keyword from their reporting, together with their surname, in order to determine in what other outlets a project is picked up. This is not methodologically sound, but it provides interesting information and can be used to do a gut check about impact. This process may also generate additional questions about a project's impact that are worth a deeper dive. Many organizations use news clipping services like Google News Alerts or Meltwater for this purpose.

Rigorous content analysis would identify key words, data, and /or phrases in a project, then analyze their prevalence pre- and post-publication in a finite corpus of text to document change. Computational text analysis goes a step further and infers shifts in discourse by advanced counting and analysis techniques. These more rigorous content analysis methods likely require a news organization to partner with trained researchers.

Looking ahead: Why journalists should care about the impact of data journalism

To stay relevant, journalism must not only accept that it has an impact on society, but embrace that fact. By working to understand the ecosystem of change in which journalism functions, and its specific role within this system, the industry can work to maximize its positive impact and demonstrate its value to audiences.

Data journalists, with their understanding for the value and importance of both quantitative and qualitative data, are well positioned for this endeavor. By articulating the goals of data journalism projects, developing creative audience engagement and distribution strategies, and building sophisticated methods for measuring success into these projects, reporters can lead this movement from within.

Works Cited

Lindsay Green-Barber, 'Changing the conversation: The VA backlog,' *The Center for*

Investigative Reporting, 2015.

Lindsay Green-Barber, 'What

makes a news interactive successful? Preliminary lessons from The Center for

Investigative Reporting,' *The Center for*

Investigative Reporting, 2015.

Lindsay Green-Barber, 'Waves

of Change: The Case of Rape in the Fields,' *The Center for Investigative Reporting*, 2014.

Lindsay Green-Barber and Pitt Fergus, 'The Case for

Media Impact: A Case Study of ICIJ's Radical Collaboration Strategy,' *Tow Center for Digital Journalism*, 2017. https://academiccommons.columbia.edu/catalog/ac:jdfn2z34w2

Tim Groseclose and Jeffrey Milyo, "A Measure of Media Bias," *The Quarterly Journal of Economics* (4), (2005), pp. 1191-1237.

James Hamilton, 'All the News That's Fit to Sell: How the Market Transforms into News', *Princeton University Press*, 2004.

James Hamilton, 'Democracy's Detectives: The Economics of Investigative Journalism', *Cambridge: Harvard University Press*, 2016.

Scott Klein, 'The Data Journalism Handbook', *O'Reilly Media*. 2012.

The Economics of Data Journalism

This chapter is launching soon

The Datafication of Journalism

This chapter is launching soon

Forms of Data Journalism

This chapter is launching soon

Data Journalism and its Publics

This chapter is launching soon

Data Journalism: In Whose Interests?

This chapter is launching soon

Indigenous Data Sovereignty

This chapter is launching soon

What is Data Journalism For? Cash, Clicks, and Cut and Trys

The daily refreshing of Five Thirty Eight's interactive 2016 election map forecasts was all but ritual among my fellow Washingtonians, from politicians to journalists to students to government workers and beyond. Some of this ilk favored The New York Times' Upshot poll aggregator; the more odds-minded of them, Real Clear Politics, and those with more exotic tastes turned to The Guardian's US election coverage. For these serial refreshers, all was and would be right with the world so long as the odds were ever in Hillary Clinton's favor in the US presidential election's version of Hunger Games, the bigger the spread, the better.

We know how this story ends; Nate Silver's map, even going into election day, had Hillary Clinton likely to win by 71.4%. Perhaps it's due time to get over the 2016 US election, and after all, obsession with election maps is perhaps a particularly American pastime, due to the regular cycle of national elections – though that's not to say that a world-wide audience isn't also paying attention.[1] But until link rot destroys the map, it's there, still haunting journalists and Clinton supporters alike, providing fodder for Republicans to remind their foes that the "lamestream media" is "fake news". Politics aside, the US 2016 presidential election should not be forgotten by data journalists: even if the quantification was correct to anyone's best knowledge, the failures in mapping and visualization have become one more tool through which to dismantle journalists' claim to epistemic authority (or more simply, their claim to be "authorized knowers").

Yes, it is unfair to conflate data journalism as electoral prediction – it certainly is far more than that, particularly from a global vantage point, but this sometimes seems that this what data journalism's ultimate contribution looks like: endless maps, clickable charts, and calculators prone to user error, over-simplification, and marginalization regardless of the rigor of the computation and statistical prowess that produced them. With the second edition of this handbook is now in your hands, we can declare that data journalism has reached a point of maturation and self-reflection, and as such, it is important to ask "What is Data Journalism For?"

Data journalism, as it stands today, still only hints at the potential it has offered to reshape and reignite journalism. The first edition of this handbook began as a collaborative project, in a large group setting in 2011 at a Mozilla Festival, an effort I observed but quickly doubted as ever actually materializing into a tangible result (I was wrong); this second edition is now being published by the University of Amsterdam Press and distributed in the US by the University of Chicago Press with solicited contributors, suggesting the freewheeling nature of data journalism has been exchanged somewhat in return for professionalism, order, and legitimacy. And indeed, this is the case: data journalism is mainstream, taught in journalism schools, and normalized into the newsroom[2]. Data journalism has also standardized and as such, has changed little over the past five to seven years; reviews of cross national data journalism contests reveal limited innovation in form and topic (most often: politics), with maps and charts still the go-to[3]. Interactivity is limited to what is considered "entry level techniques" by those in information visualization (Young, Hermida and Fulda,, 2017); moreover, data journalism has not gone far enough to visualize "dynamic, directed, and weighted graphs".[4] Data journalists are still dealing with pre-processed data rather than original "big data" - and this data is "biggish," at best – government data rather than multi-level data in depth and size of the sort an ISP might collect.

This critique I offer flows largely from a Western-centered perspective, if not-US centered perch, but that does not undermine the essential call to action I put forward: data journalists are still sitting on a potentially revolutionary toolbox for journalism that has yet to be unleashed. The revolution, however, if executed poorly, only stands to further undermine both the user-experience and knowledge-seeking efforts of news consumers, and at worst, further seed distrust in news. If data journalism just continues to look like it has looked for the past five to ten years,

then data journalism does little to advance the cause of journalism in the digital and platform era. Thus, to start asking this existential question about "What is data journalism for?" I propose, that data journalists, along with less-data focused but web-immersed journalists who work in video, audio, and code, as well as the scholars that poke and prod them, need to rethink data journalism's origin story, its present rationale, and its future.

Data Journalism in the US: The Origin Story

The origin story is the story we tell ourselves about how we and why we came to be, and is more often than not filled with rose-tinted glasses and braggadocio than it is reality. The origin story of data journalism in the US goes something like this: In the primordial pre-data journalism world, data journalism existed in an earlier form, as computer-assisted reporting, or was called that in the US, which offered an opportunity to bring social science rigor to journalism.

In the mythos of data journalism's introduction to the web, data journalists would become souped-up investigative journalists empowered with superior computational prowess of the 21st century who set the data (or documents) free in order to help tell stories that would otherwise not be told. But beyond just investigating stories, data journalists also were to somehow save journalism with their new web skills, bringing a level of transparency, personalization, and interactivity to news that news consumers would appreciate, learn from, and of course, click on. Stories of yesteryear's web, as it were, would never be the same. Data journalism would right wrongs and provide the much needed objective foundation that journalism's qualitative assessments lacked, doing it at a scale and with a prowess unimaginable prior to our present real-time interactive digital environment replete with powerful cloud-based servers that offload the computational pressure from any one news organization. Early signs of success would chart the way forward, and even turn ordinary readers into investigative collaborators or citizen scientists, such as with The Guardian's MP scandal coverage or WNYC's Cicada project, which got a small army of New York-area residents to build soil thermometers to help chart the arrival of the dreaded summer instincts. And this inspired orchestration of journalism, computation, crowds, data, and technology would continue, pushing truth to justice.

The Present: The 'Hacker Journalist' as Just Another (Boring) Newsroom Employee

The present has not moved far past the origin story that today's data journalists have told themselves, neither in vision nor in reality. What has emerged has become two distinct types of data journalism: the "investigative" data journalism that carries the noble mantle of journalism's efforts forward, and daily data journalism, which can be optimized for the latest viral click interest, which might mean anything from an effort at ASAP journalistic cartography to turning public opinion polling or a research study into an easily shareable meme with the veneer of journalism attached. Data journalism, at best, has gotten boring and overly professional, and at worst, has become another strategy to generate digital revenue.

It is not hyperbole to say that data journalism could have transformed journalism as we know it – but hitherto it has not. At the 2011 MozFest, a headliner hack of the festival was a plugin of sorts that would allow anyone's face to become the lead image of a mock-up Boston Globe home page. That was fun and games, but The Boston Globe was certainly not going to just allow user-generated content, without any kind of pre-filtering, to actually be used on its home page. Similarly, during the birth of this first Data Journalism Handbook, the data journalist was the "hacker journalist," imagined as coming from from technology into journalism or at least using the spirit of open source and hacking to inspire projects that bucked at the conventional processes of institutional journalism and provided room for experimentation, imperfection, and play – tinkering for the sake of leading to something that might not be great in form or content, but might well hack journalism nonetheless[5]. In 2011, the story was of outsiders moving into journalism, in 2018, the story is of insiders professionalizing programming in journalism, the spirit of innovation, invention, has become decidedly corporate, decidedly white-collar, and decidedly less fun.[6]

Boring is ok, and serves a role. Some of the professionalization of data journalism has been justified with the "data journalist as hero" self-perception – data journalists as those who, thanks to a different set of values (e.g. collaboration, transparency) and skills (visualization, assorted computational skills) could bring truth to power in new ways. The Panama and Paradise Papers are perhaps one of the best expressions of this vision. But, investigative data journalism requires time, effort, and expertise that goes far beyond just data crunching, and includes many other sources of more traditional data, primarily, interviews, on-location reporting, and documents. Regularly occurring, groundbreaking investigative journalism is an oxymoron, though not for lack of effort – the European Data Journalism Network, the US' Institute for NonProfit News, and the Global Investigative Journalism network – showcase the vast network of would-be investigative efforts. The truth is that a game-changer investigation is not easy to come by, which is why we can generally name these high-level successes on about ten fingers and the crowd-sourced investigative success of The Guardian MP example from 2010 has yet to be replaced by anything newer.

What's past is prologue when it comes to data journalism. Snow Fall, The New York Times' revolutionary immersive storytelling project that won a Pulitzer in 2012, emerged in December 2017 as "Deliverance from 27,000 Feet" or "Everest". Five years later, The New York Times featured yet another longform story about a disaster on a snowy mountain, just a different one (but by same author, John Branch). In those five years, "Snowfall" or "Snowfalled" became shorthand within The New York Times and outside it for adding interactive pizzaz to a story; after 2012, a debate raged not just at The Times but in other US and UK newsrooms as to whether data journalists should be spending their time building pre-built tools that could auto-Snowfall any story, or work on innovative one-off projects.[7] Meanwhile, Snow Fall, minimally interactive at best in 2012, remained minimally interactive at best in its year-end 2017 form.

> "But wait," the erstwhile data journalist might proclaim "Snow Fall isn't *data journalism* – maybe a fancy trick of some news app developers, but there's no data in Snow Fall!" Herein lies the issue: maybe data journalists don't think Snow Fall is data journalism, but why not? *What is data journalism for if it is not to tell stories in new ways with new skills that take advantage of the best of the web?*

Data journalism also cannot just be for maps or charts, either, nor does mapping or charting data give data journalism intellectual superiority over immersive digital journalism efforts. What can be mapped is mapped. Election mapping in the US aside, the ethical consequences of quantifying and visualizing the latest available data into clickable coherence needs critique. At its most routine, data journalism becomes the vegetables of visualization. This is particularly true given the move toward daily and evenly demand for data journalism projects. Perhaps it's a new labor statistic, city cycling data, recycling rates, the results of an academic study, visualization because it can be visualized (and maybe, will get clicked on more). At worst, data journalism can oversimplify to the point of dehumanizing the subject of the data that their work is supposed to illuminate. Maps of migrants and their flows across Europe take on the form of interactive arrows or genderless person icons, as human geographer Paul Adams argues, digital news cartography has rendered the refugee crisis into a disembodied series of clickable actions, the very opposite of what it could as journalism to make unknown "refugees" empathetic and more than a number.[8] Before mapping yet another social problem or academic study, data journalists need to ask: to what end are we mapping and charting (or charticle-ing for that matter)?

And somewhere between Snow Fall and migration maps lies the problem: *What is data journalism for?* The present provides mainly evidence of professionalization and isomorphism, with an edge of corporate incentive that data journalism is not just to aid news consumers with their understanding of the world but also to pad the bottom lines of news organizations. Surely that is not all data journalism can be.

The Future: How Data Journalism Can Reclaim its Worth (and Be Fun, too)

What is data journalism for? Data journalism needs to go back to its roots of change and revolution, of inspired hacking and experimentation, of a self-determined vision of renegades running through a tired and uninspired industry to force journalists to confront their presumed authority over knowledge, narrative, and distribution. Data journalists need to own up to their hacker inspiration and hack the newsroom as they once promised to do; they need to move past a focus on profit and professionalism within their newsrooms. Reclaiming outsider status will bring us closer to the essential offering that data journalism promised: a way to think about journalism differently, a way to present journalism differently, and a way to bring new kinds of thinkers and doers into the newsroom, and beyond that, a way to reinvigorate journalism.

In the future, I imagine data journalism as unshackled from the term "data" and instead focused on the word "journalism." Data journalists presumably have skills that the rest of the newsroom or other journalists do not: the ability to understand complicated data or guide a computer to do this for them, the ability to visualize this data in a presumably meaningful way, and the ability to code. Data journalism, however, must become what I have called interactive journalism – data journalism needs to shed its vegetable impulse of map and chart cranking as well as its scorn of technologies and skills that are not data-intensive, such as 360 video, augmented reality, and animation. In my vision of the future, there will be a lot more of BBC's "Secret Life of the Cat" interactives and New York Times' Dialect Quizzes; there will be more projects that combine 360 video or VR with data, like Dataverse's effort funded by the Journalism 360 immersive news initiative. There will be a lot less election mapping and cartography that illustrates the news of the day, reducing far-away casualties to clickable lines and flows. Hopefully, we will see the end of the new trend toward interactives showing live-time polling results, a new fetish of top news outlets in the US). Rather, there will be a lot more originality, fun, and inspired breaking of what journalism is supposed to look like and what it is supposed to do. Data journalism is for accountability, but it is also for fun and for the imagination; it gains its power not just because an MP might resign or a trendline becomes more clear, but also because ordinary people see the value of returning to news organizations and to journalists because journalists fill a variety of human information needs – for orientation, for entertainment, for community, and beyond.

And to really claim superior knowledge about data, data journalists intent on rendering data knowable and understandable need to collect this data on their own – data journalism is not just for churning out new visualizations of data gathered by someone else. At best, churning out someone else's data makes the data-providers' assumptions visible, at worst, data journalism becomes as stenographic as a press release for the data provider. Yet many data journalists do not have much interest in collecting their own data and find it outside the boundaries of their roles; as Washington Post data editor Steven Rich explained, in a tweet, the Post "and others should not have to collect and maintain databases that are no-brainers for the government to collect. This should not be our fucking job".[9] At the same time, however, the gun violence statistics Rich was frustrated by having to maintain are more empowering than he realized: embedded in government data are assumptions and decisions about what to collect that need sufficient inquiry and consideration. The data is not inert, but filled with presumptions about what facts matter. Journalists seeking to take control over the domain of facticity need to be able to explain why the facts are what they are, and in fact, the systematic production of fact is how journalists have claimed their epistemic authority for most of modern journalism.

What data journalists is for, then, is for so much more than it is now – it can be for fun, play, and experimentation. It can be for changing how stories get told and invite new ways of thinking about. But it also stands to play a vital role in re-establishing the case for journalism as truth-teller and fact-provider; in creating and knowing data, and being able to explain the process of observation and data collection that led to a fact, data journalism might well become a key line of defense about how professional journalists can and do gather facts better than any other occupation, institution, or ordinary person ever could.

Works Cited

Paul C. Adams, 'Migration Maps with the News: Guidelines for ethical visualization of mobile populations', *Journalism Studies* 1:21, (2017), pp. 527-547.

Seth C. Lewis and Nikki Usher, 'Trading zones, boundary objects, and the pursuit of news innovation: A case study of journalists and programmers', *Convergence* 22:5, (2016), pp. 543-560.

Norman P. Lewis, and Stephenson Waters, 'Data Journalism and the Challenge of Shoe-Leather Epistemologies', *Digital Journalism* 1:18, (2017). pp. 719-736.

Seth C. Lewis and Nikki Usher, 'Open source and journalism: Toward new frameworks for imagining news innovation', *Media, Culture & Society* 35:5, (2013), pp. 602-619.

Wiebke Loosen, Julius Reimer, and Fenja De Silva-Schmidt, 'Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016', *Journalism*, (2017).

Christina Niederer, Wolfgang Aigner, and Alexander Rind, 'Survey on visualizing dynamic, weighted, and directed graphs in the context of data-driven journalism' *Proceedings of the International Summer School on Visual Computing*, (2015), pp. 49-58.

Nikki Usher, 'Interactive journalism: Hackers, data, and code', *Urbana-Champaign: University of Illinois Press*, 2016.

Statisticians and Journalists: Tales of Two Professions

This chapter is launching soon

Data Journalism and Digital Liberalism

This chapter is launching soon

Afterword: Data Journalism and Experiments in Reporting

This chapter is launching soon