UK Data Service

Proceedings from the workshop on

Supporting human rights organisations to deliver insights from data

29-30 October 2015

University of Essex

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Contents

Supporting human rights organisations to deliver insights from data: workshop report and outcomes UK Data Service, 29-30 October 2015

Introduction

At the end of October the **UK Data Service** (the Service) organised a workshop on "Supporting human rights organisations to deliver insights from data" at the University of Essex, with support from the Economic and Social Research Council (ESRC). This was in the spirit of the ESRC's ongoing work in engaging with civil society.

For this workshop Human rights organisations were targeted for two reasons:

- First, that Essex is well-recognised in this field, with its well-established Centre for Human Rights and the current University Chancellor, Shami Chakrabarti, who is also director of Liberty, the British civil liberties advocacy organisation.
- Second, data about the people represented by these charities faces particular challenges around privacy and data protection, and hence, wider sharing; challenges that the Service is uniquely positioned to address.

Human Rights organisations attending the workshop included those involved in supporting victims of human trafficking, torture, unfair trials, conflict and war and other vulnerable situations. Speakers from other civil society and philanthropic bodies attended as well as academics and journalists. From the Service's own interest in increasing the amount of data and evidence that can be shared for transparency and analytic purposes, looking more closely into helping provide solutions for safely sharing and utilising data would benefit the broader civil society sector.

The workshop provided a forum for participants to engage with one another to discuss strategies, tools and skills required for civil society organisations to become better knowledge managers. In this context 'knowledge' includes the organisation's own datasets, external contextual data sources and measures of evidence; while 'manager' covers the capacity to handle, report and campaign using this data-based evidence. A wide range of scenarios and data challenges and opportunities were considered, helpfully brought along by the participating organisations. These scenarios covered both legal and ethical challenges of data and their management. Other challenges covered skills and the capacity to maximise the opportunity presented by the data.

Opening session

Louise Corti, Associate Director at the **UK Data Archive** and organiser of this event introduced how the meeting would be organised and set out the interests of the UK Data Service, ESRC's flagship national data investment, in engaging with the data holdings and needs of civil society organisations.

Christina Rowley from the **ESRC** followed by introducing the perspective and investments from the ESRC. She highlighted the data infrastructure they support through the UK Data Service and other projects, and outlined the What Works Centres; independent institutions, part-funded by UK government which aimed to provide robust research evidence for policy and practice in the area of: crime reduction, local economic growth, tackling poverty, early intervention and wellbeing.

Christina went on to describe the ESRC's current Civil Society investments though partnership funding, Centres, such as the Wales Institute for Social and Economic Research, Data and Methods, the new large grant in human rights, big data and technology at Essex, and funding for the NCVO Civil Society Almanac. She mentioned that all ESRC funding opportunities include funding to support collaborative working and knowledge exchange (KE), for example through the standard grants scheme. The KE aspect was important, as it aimed to influence the development of policy and practice, shape legislation, alter behaviour, change concepts and discourses and provide capacity building. Christina finished by highlighting the annual **Celebrating Impact Prize** which was a great opportunity for showcasing exemplary policy and practice-related work coming out of ESRC funding.

Neil Serougi, trustee at the charity, **Freedom from Torture**, and co-organiser of this event, provided the keynote speech. He addressed the challenges facing civil society organisations, from changing perceptions of the role that charities should play in society to the current reduced public support for charities, and the challenges they face in navigating a more politicised landscape. He discussed how, by using data more efficiently and effectively, human rights organisations can change people's perceptions of the work they do, and further help the public empathise with the individuals and groups with which charities work. Neil continued by highlighting the problem with the sector's largely 'traditional' use of data to illustrate key performance indicators (KPIs) and producing the kind of reports that are no longer an effective method of demonstrating the impact of their work. There is now a pressing need to use data in a meaningful and intelligent way to both present the outcomes of a charity's work and to help supporters and funders interact with their work.

To conclude and to set the scene for the following two days of debate, Neil left participants with a warning that "the equation of more information with more impact is not necessarily true by default. [So] data is important in the quest to make an impact but, more important in our line of work is being able to make it mean more than the sum of its parts; in other words to deliver more than an array of numbers that usually depict what most of us intuitively would expect to see anyway".

Session 1: In house data collection: what do you have, what do you need and what skills do you have to analyse the data?

Civil society organisations have been collecting increasing amounts of operational, research and evaluation data but with tight budgets and limited capacities; few fully exploit or share the data that they collect. This session, led by **Louise Corti,** aimed to focus on the types of information that organisations collect themselves, identifying barriers to sharing and understanding those data. The session examined what capacity and skills civil society organisations need to undertake good data practices.

Attendees had been asked in advance questions about the data they hold, scope and format, and whether any were shared. Data held included monitoring and evaluation data, qualitative and quantitative, held in a variety of formats; within bespoke customer database systems e.g. SalesForce, spreadsheets, word processing documents, email attachments, SPSS and paper. When asked about data handling and analysis capacity, replies ranged from: in-house software specific skills to input and run queries; in-house analysts in larger organisations; collaboration - drawing on skills elsewhere; and many felt that that they needed more training on issues of storage security, and data protection. When asked about data sharing many did this when: submitting summary data to higher level bodies; reporting to comply with legal requirements; and for analysis purposes where data were transferred to a third party, but typically did not share data for the benefit of others gaining insight. Some saw no need to share data; others might if asked. Issues raised were that: it was not always their data to share; confidentiality concerns; and international data sharing restrictions. In summary, while some would like to share data, a fair bit of preparatory work would need to be done to make this possible.

All agreed that keeping accurate records in an optimised way, that can be reported upon, offers a solid foundation for demonstrating transparency; complying with both the Information Commissioner with respect to data, and with Charity Commission procedures by providing a record of due diligence to donors.

The session opened with **Tracey Gyateng** from **New Philanthropy Capital** (NPC), a think tank which helps charities make more use of their data to understand the needs of their beneficiaries, measure impact, plan scenarios and improve operational effectiveness. NPC helps civil society organisations to lever government datasets in order to better understand their effectiveness of their interventions. For example, NPC has been working with the Ministry of Justice to set up the Justice Data Lab, which provides the social enterprise *Blue Sky* with non-disclosive data on the re-offending rates of people who used their services compared to a control group. The data are matched confidentially at the level of the individual by the Justice Data Lab, analysed, and the aggregated results returned to the Blue Sky team. NPC are also working with the Department of Work and Pensions on a similar scheme focused on employment outcomes. By providing quantified data on whether people who have used a charity's services are more likely to avoid committing further crimes or return to employment, this mechanism gives charities an evidenced measure of the 'hard outcomes' of their activities.

Tracey outlined a number of barriers charities face sharing their data or using external information sources. Consent can be an issue; vulnerable, hard-to-reach groups may be reluctant to give consent for their data to be shared with government departments. Charities may also be unaware data are available, unable to spare the time or money to invest in data, lacking the skills to analyse data or understand the results, worried about what the data may tell them, or concerned about privacy and ownership. The importance the future role of the national services like the Administrative Data Service in providing safer and easier access for civil society organisations to government data was stressed.

Nigel Fielding of the **Department of Sociology** at the University of Surrey used the prison reform charity the Howard League, with which he has worked for many years, to illustrate the rich and deep information resource that a charity could offer. The League's information base goes back to the 18th century and spans original pamphlets and tracts as well as surveys documenting parliamentary processes and policies relating to prison reform. As an operational campaigning charity, this information base continues to grow and has great potential for academic researchers. However, the League has few resources to systematically organise the materials it holds in a way that would make them easier to discover and share. Although only a fraction of the League's materials held are in digitised form, Nigel emphasised the power of computer assisted software qualitative software (CAQDAS) packages in managing textual and visual information in order to help civil society organisations to leverage information resources in order to advance their agendas. The CAQDAS Networking Group that he set up in the 1990s can offer support for organisations in choosing software and training in how to exploit their features. Nigel highlighted the useful role of involving students in the work of the League through placements, who are incentivised by the direct experience they gain from criminal justice work by undertaking office work but also getting directly involved in campaigns. This work experience includes going into prisons on fact-finding missions to talk to prisoners about the League's work, and meeting with prison governors and staff to discuss policy and practice. Students also prepare briefings and help draft speeches for the League's senior staff.

In her talk, *Driving a Ferrari into the desert and leaving it there*, **Roisin Read** from the ESRC's Making Peacekeeping Data Work project at the **University of Manchester** spoke of how humanitarian organisations manage security data, in particular relating the conflict and peacekeeping intervention in Darfur. Her research indicates that in conflict situations, data collection guidelines are often not followed in practice, meaning that information is not always gathered in ways that ensure the safety of informants. In Darfur, for example, the consequent fear of reprisals has deterred people from reporting human rights violations or security incidents to UNAMID [the UN-African Union peacekeeping mission in Darfur]. "People stopped reporting incidents … there was a widespread perception that if you shared information with UNAMID, you would be targeted by the security services."

The talk title derived from a staff member in the information management division of a large NGO who observed that the elaborate and expensive systems for storing, analysing and managing humanitarian field data are simply too far ahead of any desire or capacity to use them. One of the project's interviewees also stressed that in the humanitarian sphere, operational data must be timely. "I need information to be quick and dirty. Good information too late is useless."

The final speaker in this session, **Ingvill Mochmann** of the **GESIS-Leibniz Institute for the Social Sciences** in Cologne, spoke about data collection and sharing within an international network for research on children born as a result of war (usually a child whose father has been a member of an enemy, allied or peacekeeping force and mother a local citizen). Although there have been children born though conflict across time and nations, an understanding of their experiences, needs and rights is fragmented and limited. The *Children Born of War* project is a global virtual network that brings together research findings and data collections on the topic. Sources include survey data, administrative data, interviews, letters, photos, medical records, church records, and biographies, in short "everything you can access". The field is very new and a participatory research approach is used to engage older, adult children born of war in the development of questionnaires for younger groups in order include topics and experiences relevant to themselves.

Even as adults, children born of war are often vulnerable, hidden populations. They may be stigmatised and subject to discrimination throughout their lives. This presents a number of practical and ethical challenges including the re-traumatisation of participants and the different ethical regulations of data access, storage and sharing across countries.

Session 2: Making an Impact: Using Data beyond Key Performance Indicators

David Walker, contributing editor at **The Guardian** introduced the session by recommending that a pragmatic approach should be adopted when using data and posed the question, "does a huge amount of data help to promote those in conflicts across the globe?" David continued by suggesting that academics often think "to know, is to do"; and stressed that in order to move forward we need to combine both knowledge and the enterprise of the 'doers' based in civil society, which should forge greater data utility and awareness.

Outlining the work of **Medical Aid for Palestinians** (MAP), who partner with the University of Beirut to harness academic analysis, **Bob Jones** demonstrated the value of data in an example where outputs were used by MAP to both raise questions of government, and to provide valuable outputs to government to help inform policy change. Impact is a crucial part of fundraising, with research indicating that donors are more likely to donate around the times during which an event such as a massacre has occurred, when public awareness is heightened. As MAP is largely funded by individual donations, using data to create infographics that demonstrate statistical evidence is key to informing public awareness. Anticipating public interpretation and understanding of the information has to be carefully managed to avoid any potentially damaging misinterpretations.

Emma Prest told us about how her own organisation, **DataKind UK**, works. DataKind UK, a chapter in an international network, is a charity that has as its core vision the use of data in the ‘service of humanity’, creating and nurturing a 'data-for-good' community made up of enthusiastic data scientists. Their work applies data science tools and techniques to social situations and involves putting together teams of volunteer data scientists to contribute to 'Data Dive' events which aim to collect, analyse and visualise data to support better decision making for charities. These hands-on knowledge exchange sessions focus on providing answers to specific questions; bridging the gap between communities and making data owners aware of the rich resources they have available to them.

To emphasise the value of this methodology, Emma outlined a research project where DataKind had worked with an NGO which gives money directly to the poorest villages in Kenya and Uganda. The NGO realised that one of the indicators of poverty was whether there was a thatched roof or a metal roof on the house; a metal roof indicated a richer household and a thatched roof, indicated a poorer household. By using satellite imagery and learning algorithms, DataKind were able to create a model to classify the indicators – metal and thatched roofs – so that those on the ground were able to work more efficiently to identify the poorest communities, enabling the NGO to better manage resources and deliver to villages/communities most in need.

Megan Lucero, Data Journalism Editor at **The Times and Sunday Times**, discussed the reinvention of investigative journalism and its on-going development to demonstrate how computing can advance accountability and public interest in reporting. Using recent scandals on doping within the field of athletics and the corruption within FIFA, Megan outlined the work her team has conducted to ensure the validity of data resources, legal compliance and external evidence to support these ‘breaking news’ claims and stories.

While collaboration enables us to move forward, combining expertise and data sources to provide support for fundraising and evidence for research and policy change, it is not without its challenges. Data literacy and user engagement with stories has to be carefully managed by the authors/publishers to avoid damaging misunderstanding and misuse of information. Use of incomplete data to publish quickly can also cause harm through exposing unsound findings. Megan demonstrated some engaging interactive visual outputs from her stories that were simple yet effective, allowing the reader to interrogate available data to build up a greater contextual picture of the news item.

Megan spoke of a final caution with ‘data-driven journalism’: it can lead to the pursuit of topics for which there is available data rather than focusing on topics that need to be investigated! The session highlighted that the effective visualisation of synthesised results and simplified outputs built on solid and robust data can help us tell a story powerfully.

Session 3: Ethical frameworks and governance

The final session of the first day addressed legal and ethical challenges in using and sharing data. **Libby Bishop** from the Service opened the session by summarising the themes that had emerged from case studies, prepared in advance by participants, of the key ethical issues their own organisation were confronting. Key points arising were:

- the dilemma that sharing data may help clients, but also may, in some cases, create risks either to participants or to trust relations with CSOs, even if data protection regulations are fully met;
- the challenges of anonymisation, including understanding legal requirements and finding practical tools to anonymise data;
- how gaining informed consent presented difficulties with some groups, such as children, those lacking capacity, or those who might be retraumatised by relating their experiences of violence.

Libby's spoke on how the Service protects data using its **5 SAFES** framework of **SAFE DATA - SAFE PROECTS - SAFE PEOPLE - SAFE SETTNGS - SAFE OUTPUTS**, and also shared some practical tools, such as consent form templates that can be used as a basis for designing consent forms.

Jim Vine of **the Housing Associations' Charitable Trust** (HACT) followed by presenting an example of successful data sharing by a CSO to help its clients through using a 'trusted intermediary'. He described a project called *Community Insight* which levers the power of information across organisations, providing high-level visualisations to detailed reports on local neighbourhoods. Jim gave an example of where data on low-wage jobs were mapped with local bus routes to reveal a mismatch: jobs were being created in places that were not well provisioned with public transport. As a result, the local community is attempting to provide alternative modes of community transport. Jim went on to present a concise summary of the Data Protection Act and its implications for CSOs. One important point he emphasised was a reminder that not all uses (processing) of personal data must be consented; for example if a CSO can demonstrate a 'legitimate interest' for processing the data and, crucially, there is no 'prejudicial effect' on the individuals' providing data.

To conclude the session, participants worked in groups on ethical issues arising within their own organisations, taking turns to present their situations and getting feedback and actionable suggestions from others. One group had the challenge of gaining written consent from children lacking capacity, with one participant who suggested that audio-recorded might be used to capture consent, but was not sure how to implement this. Others in this group were familiar with the situation and advised that the combination of recorded consent, transcribing the recordings to create a written record, and working with guardians works well. Other problems arising in groups were: how to ensure that a client understands the implications when they agree to having their representation used for campaigning; and how to apply a general risk mitigation strategy when deciding on the level of anonymisation needed for data sharing.

Session 4: Exploring opportunities for using third party data sources to provide context

The first session of the second day examined the potential for using third party data sources to provide broader contextual knowledge for organisations working in the field of human rights. Various forms of data are collected and, often these data can be made available to researchers. Organisations require awareness about to how to locate and access these sources and to understand how to evaluate and analyse them. This session sought to answer questions relating to how the strategic goals of CSOs could be achieved by using innovative methods and bringing together different forms of formal and informal knowledge. With the increasing importance of administrative records or 'big data' in this landscape, this session also addressed the theoretical and technical challenges researchers face in trying to analyse these data.

The session began with a presentation from **Hersh Mann** who manages the User Support and User Training area for the Service. Hersh began by providing an overview of the Service and explained that data held and made available comes in different forms - quantitative and qualitative. Illustrating his talk with examples drawn from topics like housing, poverty, human development, and minority rights, he demonstrated how the Service can facilitate access to data that cover a very eclectic mix of topics. Hersh also explained how the Service assists users in tracking down and helping provide access to data that are not already available. The Service's expertise in how to share data and their trusted relationships with key bodies like government departments and other organisations that create data, can often provide solutions that users may not have thought possible.

Sian Oram from **King's College London,** was the second speaker in this session. Delivering her talk, entitled *Mental health responses to human trafficking: qualitative data tools*, she explained how a current project, PROTECT – Provider Responses, Treatment, and Care for Trafficked People - she was involved in, aimed to understand how people are identified as trafficked within mental health services, and the challenges professionals experience in responding to trafficked people's mental health needs. She explained that while mental disorder is prevalent among survivors of human trafficking, little is known about health professionals' experiences of identifying and providing care for them. Using the South London and Maudsley (SLAM) NHS Foundation Trust Case Register Interactive Search (CRIS) database, comprehensive clinical electronic health records were used to identify trafficked patients. Content analysis was used to establish how people were identified as trafficked, and thematic analysis was used to explore the challenges experienced in responding to mental health needs.

The final presentation of the morning session was by **Matt Williams and Luke Sloan**, both from **Cardiff University**. Their talk, *Gaining Insights from Social Media Data: Collection, Analysis and Interpretation*, began by showing how they have been using data from social media to understand social reactions to major news stories and how demographic information can be derived from the use of Twitter. They highlighted the enormous growth in the creation of data that has resulted from the use of social media and the potential that these data have in helping researchers interpret social problems.

Introducing the Cardiff Online Social Media ObServatory (COSMOS), Matt and Luke described a technical architecture for the computational analysis of social media data, and explained the challenges involved in storing and interrogating such large data collections. By using methods such as word frequency counts, network analysis, and geospatial clustering, they showed how they analysed tweets and mentioned how these new forms of information could be used to complement existing sources of data. In the second part of the talk they presented case studies on, for example, cyberhate and the Ebola outbreak. These they used to show how the COSMOS platform allowed them to gain insights into the way information spreads or is not propagated online.

In the final session over lunch participants were invited to discuss issues of their choice with the range of experts present.

Summary

Summing up after such an intensive and positive event like this is easier to say, but harder to enact; we should not let the momentum lapse. Ideas and debates emerged about the way that civil society organisations can work in a data-centric society; what might this mean for the evolution of organisations? The meeting explored very different forms of data, from summary statistics to qualitative classification and social media. In thinking about producing evidence beyond the KPI indicator, there is a need to consider novel ways in which this varied information can be marshalled to influence policy, develop impact strategies or persuade more people to donate.

Even gathering information from participating organisations for the workshop sessions, it became clear that there was an anxiety among participating organisations about the task; and in some cases, the feeling that they might not be able to operate at the level 'expected' from the sessions. What we in the Service very helpfully took away from this meeting was that the formal language of social research can be a barrier to engagement; terms like 'data' and 'analysis' might better be replaced with 'information; and 'insight'.

In the CSO sector there is less investment, and certainly very few bespoke roles and responsibilities purely dedicated to making information and data work in terms of evidence. It is a truism that many of the participants went back to their jobs the following Monday morning hit by all the mounting work they were unable to do on Thursday and Friday; and it is unrealistic to expect them to become bona-fide researchers with all the skills and the time, the abilities and capabilities.

With less money to upskill staff to undertake research activities, coproduction (with experts on a pro bono basis) is likely to be an important model. More and more organisations need to prove the case for investment in sound and ethical information resources. Good outcomes can help demonstrate this in a feedback loop.

It would be useful for the community to explore the potential to develop a network of organisations that could be better informed and enabled regarding partnership opportunities, such as with ESRC research funding. Another idea for future direction of travel is how the academic community might work up a brokerage system akin to the DataKind model; where university academics can help civic society organisations move to that next level, both in terms of capacity and capability. The national co-ordinating centre for public engagement (NCCPE) might be the kind of organisation that could help with this. There are platform that already exist within universities that highlight collaboration interests of staff, such as Piirus at Warwick University, designed to match academics with particular interests.

The Service is keen to play its part in helping engage with CSO data issues and this event is our first dedicated initiative in this area. We already have a number of great resources that we can offer, such as running tailored data management workshops and a vast catalogue of data that can be used to provide rich context. There are a number of concrete follow up actions that we can take, within the next 6-12 months:

- run a similar engagement event again next year;
- targeted promotion of our data;
- a stall at the next national Charity Fair or at NCVO's annual 'Evolve' event in June 2016;
- a webinar on introducing useful data sources for the sector;
- a webinar on how to (easily) create data visualisations using free tools such as MS Excel, Google table tools and using the COSMOS social media platform;
- bespoke training and guidance on how to safely manage and share data for organisations collecting data about beneficiaries and leading to the possibility of data deposits that we can curate and provide access to;
- guidance on how to separate legal (i.e., intellectual property, Data Protection) and ethical issues, showing consideration of any possible harm of releasing documents;
- take practical steps to introduce these ideas to the international arena in our work with international data organisations such as the Research Data Alliance.

Presentations and bios from the event are available at: ukdataservice/news-and-events/eventsitem/?id=4128. As the talks were so valuable, edited proceedings will also be available before the end of the year also from this site.

Keep in touch with us at: https://www.ukdataservice.ac.uk/about-us/contact.
Contact our help desk at: https://www.ukdataservice.ac.uk/help/get-in-touch.

Introduction

Christina Rowley, ESRC: ESRC's Civil society engagement and agenda

Presentation slides at: http://www.ukdataservice.ac.uk/media/604185/hrdw_rowley29102015.pdf

Some of you many know that, in January 2015, the ESRC published its latest Strategic Plan. In that plan, we outline our mission, and highlight our priorities in terms of the types of strategic activities in which we engage. Our mission focuses on promoting and supporting high quality research and training; developing and supporting data infrastructure; meeting the needs of users and beneficiaries; and communicating and promoting social sciences to the general public. What we did not do in that document, in order to be able to be more agile, was to lay out specific thematic priorities for forthcoming commissioning. The ESRC has three current priorities. These are:

(1) The UK in a Changing Europe initiative, led by Professor Anand Menon at Kings College London;

(2) A strand of work funded under the auspices of Research Councils UK (RCUK), working with the Engineering and Physical Sciences Research Council and the Arts and Humanities Research Council, around conflict crime and security, which used to be known as Global Uncertainties and is now called the RCUK Partnership for Conflict, Crime and Security (PaCCS).

(3) A key priority for us over the last year or so has been what we have called Urban Transformations but, in terms of the RCUK umbrella term, is now Urban Living – the RCUK Urban Living Partnership.

With the links in the presentation slides you can find out more about these priorities, from our website and from the RCUK site.

We are developing new areas of thematic priority and will be publishing more information about these shortly. At the moment I can tell you there are five areas under consideration by the Council and the committees. These are housing, macro-economy (or new approaches towards macro-economic modelling), mental health, new ways of living and working in a digital age and productivity. Not all of these will be taken forward but it will be some combination of these. Some of these may, for example, be delivered through steers in our Centres and Large Grants (CLG) Competition, which will open in spring 2016. Others may be delivered through strategic commissioning, where we focus in on an area and launch a call to commission something specifically on that theme. Some of you in the room may also be aware that in January 2016 we will be holding a consultation workshop on housing, because that priority is further along the planning stages and we have been working with some stakeholders to map out what an investment in housing might look like.

In terms of data and the ESRC's data priorities, I am not going to go through all of the bullet points in the slides but you can see there what our concerns are, and what we wish to prioritise over the next few years. One of the things we are keen to do, and why this event is so important, is to ensure that academic and non-academic partners, in today's case civil society organisations (but for us this would also include central local, and devolved government partners, as well as business partners – partners from all sectors, essentially) can work together around data and related research to answer questions that are really useful to those non-academic partners (as well as being of interest to academic researchers). On slide 6, you can see just a few examples of data investments that ESRC currently supports. We have also made significant investments, alongside government departments, in various What Works Centres and I know that there are one or two of you in the room who will be familiar with the What Works network. There are six centres at the moment but there is also a looser network of centres, which are engaged in similar activities, broadly speaking – namely, translating what evidence there is into usable information for policy-makers and practitioners.

In terms of recent ESRC research investments that are relevant, I wanted to flag just a few here. In particular, a new large grant, funded in the 2014/15 CLG Competition investigating the ethical relationships between Human Rights, Big Data and Technology has just started at the University of Essex (the dedicated website is not up and running yet but more information can be found at https://www.essex.ac.uk/hrc/research/bigdata.aspx).It is hoped that the research that that project undertakes will feed into and connect up with the data investments at Essex and elsewhere, the concerns might be around questions of privacy and freedom of information in a 'big data' era (an era which is, in part, characterised by the repurposing of data). We made a strategic investment in the NVCO Civil Society Almanac earlier in 2015, for three years' worth of funding. Some of you may use the Almanac in your own work. We also have four Civil Society Data Partnerships and I do not think I need to say too much about those because some of them will be presenting later on. The WISERD Centre – that is the Wales Institute for Social and Economic Research, Data and Methods – was funded in the 2013/14 CLG Competition. They are working particularly with a number of Welsh civil society organisations (although their findings and research will have applicability beyond the Welsh context). That is based at Cardiff, as well as other Welsh universities, if you are interested in finding out more. Previously we have funded the Third Sector Research Centre at the University of Birmingham and the Centre for Charitable Giving and Philanthropy at CASS Business School.

In a slightly different vein, although also working very closely with civil society partners, we co-fund some large grants through the Connected Communities programme (which is led by the Arts and Humanities Research Council). The Connected Communities programme is very much focused on the co-production of research, and all projects funded under the auspices of Connected Communities have community partners, either with Project Partner or Co-Investigator status, so they really are working on research projects collaboratively. Two large grants that ESRC leads on are (1) the Imagine programme, at the University of Sheffield, and (2) Productive Margins, at the University of Bristol. I would also like to flag the National Centre for Research Methods, which is another ESRC investment that has a specific remit to offer training courses and events, not just for academics but to include non-academic, such as civil society actors, so there might be some training courses and events on their website that are of interest to you.

Then we get to how the ESRC is seeing engagement with civil society more generally, which is that strategic programmes, programmes that focus specifically on civil society, are valuable but actually we want to embed engagement with civil society across all that we do, so we have established some larger investments – Impact Acceleration Accounts – at a number of institutions, in 24 universities and research organisations, (see slide 11). That funding is to ensure that those institutions can give their academics the ability to engage (with civil society organisations, charities, foundations and so on, as well as with non-academic partners from other sectors) when that engagement is needed, when projects arise, rather than having to go through a very long-winded process of applying to us for very small pots of money. That funding has been devolved, and the decision-making has been devolved to the universities themselves. It may be that a university near you has one of those, and it would be worth looking into the opportunities, if you want to make contact and find out how you as a charity can work with researchers, or build a network, get a project up and running, or do some work that derives from existing research. Of course, you do not have to stick to the university closest to you. It could be that there is a particularly relevant scholar at the University of Bangor or

Edinburgh or Exeter, and you are based in London – that is no problem, you can use the universities' websites to find people with whom you might be interested in working.

What I have put in the last line on slide 9 is 'Pathways to Impact'. We refer to our standard grants scheme as responsive-mode, because it is responsive to the proposals that are submitted, offering funding for curiosity-driven projects. There is a completely open remit to that scheme, as long as it is at least fifty per cent economic and/or social science. What we would like to see from academics is more organisations – charities, local government, businesses, devolved and central government, international actors – being included as project partners, and as co-investigators, so that research projects are genuinely co-produced. But in cases where it is perhaps not appropriate or feasible to have non-academic co-investigators, we still want all academics to think very carefully about who might find the project interesting, and to whom the findings might actually be valuable – for academics to talk to them, and to find out information that may have a bearing on the research design, not just towards the end of the project, when the more concrete information and findings are coming out of that project. We want academics to think "who should I be informing of these, who should I be listening to and how can I do that?" and put those plans, and those activities, in the proposal front and centre, and to cost them in, so that we can provide adequate funding for those activities to take place. And we ask them to do that in the section of the application form where they are expected to tell us about their planned "Pathways to Impact".

It might be as straight forward as the project holding some breakfast briefings for policy-makers in London and the cost of that will be people travelling to the events, a bit of catering, venue hire and so on. Or it might be saying, "we are working with the NSPCC and we want to ensure we have got children's views on board, we want to set up a special children's advisory group and ensure that they are commenting on the research, and there will be ethical implications that come with extra costs, such as…" – those activities that are planned to maximise the potential for impact can then all be included in the award funding. We are looking to academics to tell us honestly about their research and engagement plans, to hear their carefully considered assessment of who the most important people and organisations are with which they need to engage and how best they feel they can do that, whether that's "these are the right organisations and that's why we have been in contact with them and we know they are ready to work with us" or "we know it is going to be difficult to work with these organisations but we are going to try, this is a bit risky but here's why we should be attempting this".

So, really, for us – rather than segregating or separating out civil society as just one strand of what we do – embedding is what we are trying to do now. All of our funded projects have to consider their pathways to impact and it really is about including engagement, and support for that engagement, across all that we fund, and all that we do. In terms of what we mean by "Pathways to Impact", there are a variety of different ways in which projects can have impacts on the world. Some of it is conceptual: changing the ways in which we talk about concepts or ideas, and having an impact on media discourses, for example. Some of it is capacity building, and I think that today is a really good example of that – ensuring that people feel confident with research and handling data, for example. And then there is obviously instrumental impact – such as policy-relevant work – but I really want to stress that "policy relevance" is not the only way in which we (ESRC), understand or value impact. We do have a very broad understanding of what impact can look like.

Finally, we are just at the stage of re-commissioning our Doctoral Training Centres (DTCs – in the process, changing the name slightly, to Doctoral Training Partnerships, or DTPs). One of the particular opportunities that Doctoral Training Centres/Doctoral Training Partnerships are encouraged to offer is student placements. Those placements are three months for a doctoral student to go and be based at a different organisation and to assist with a research or data project where their expertise can help an organisation. So, for those of you who might be interested in having a doctoral student placed with you for three months, you can make contact and see how they run that scheme at their institution. Again it does not have to be the geographically closest DTC/DTP, although it might be – but we do not want all London-based organisations to offer all the opportunities only to the London DTPs! I should say that those are not free to the organisations that want a placement but I do know that some institutions make special provision for working with charities and the non-profit sector. So it is worth making contact even if you think you are a cash-strapped organisation.

Key Note

Neil Serougi: Standing out from the crowd. The value of 'data as evidence' in Civic Society

Presentation slides at: https://www.ukdataservice.ac.uk/media/604181/hrdw_serougi29102015.pdf

It is good to see you all and I am sure the two days will be as productive as it will be enjoyable.

The focus of the conference has been as you know on civic society, and it's long overdue. Culturally economically and politically the role of organisations such as we represent today has in the wider schemata changed to the point where it encompasses a wider set of expectations.

Civic society is a now nuanced term beyond its seemingly simple description of those organisation that operate within the helping sector, doing good works with a large dose of voluntarism, patching up the holes in the safety net.

Now civic society also constitutes a set of expectations that are integral to *provision and change* agendas that are at times calibrated to chime with policy re-design at the highest level. This has had several implications of which one particularly stands out; the pressure to meet expectations that are often driven by political exigencies rather than what necessarily works in the best interests of the Charity and its stakeholders.

What this has engendered is a duality in the identity of charities. On the one hand, the ethos of charitable unconditional giving and service remains a strong driver in the activities of civic organisations, but on the other, there is an emerging operational context in which eligibility has crept into the discourse about disadvantaged and marginalised groups and communities.

The importance of this should not be underestimated. Civic society organisations have always occupied a crucial role in supplying the information - often antithetical to popular opinion – to those in power across a range of institutions.

This was brought home to me at a recent United Nations Association (UNA) conference where the importance of civic society organisations in sensitising the public and political domain to landscapes which are not normally the priority – especially in those areas where access to information is at times weak and contested – it was fully recognised as an invaluable source of reliable knowledge. It is true on a domestic level too. A vast amount of information is introduced through organisations whose business is understanding the experiences of those who suffer abuses or disadvantage, the effects of which are hidden and at times misunderstood. Now though, with a new ideological imprint on the charitable sector, information and how we use it to shape perceptions about ourselves and the missions we undertake has become pivotal to not only resisting the temptation to fall in with the dominant eligibility discourse but also proving our value around delivering 'what works'.

It is an information challenge that requires a major upskilling and for organisations that are not one of the so called "big players", it is imperative that we embrace this challenge in a very competitive market. Society has become more data centric so we must follow, and in doing so, increase the potential for both sustainability and increased impact.

By the end I would like to think that we will have made the first steps towards creating a better understanding of what is required to do this and effectively 'stand out from the crowd'.

But what does standing out from the crowd mean? Well, in its simplest form, it means being able to be more effective than others and in ways which that we know could give us the extra leverage whether it be in policy domain or funding campaigns. It will not be easy, 'Tesco-isation' has meant that the environments we work in are to a large extent dominated by the big players.

Recent figures regarding development charities for example, estimated that whilst the degree of concentration in the market share of four of our largest development charities had declined from 70% to 50%, the growth rate between development charities reflected considerable differences. It was still very difficult for other players to grow into that market.

So 'standing out' for smaller and medium charities is already difficult but even for the "bigger fish", life is getting tougher in the face an array of external and internal factors that are exacting real pressures on our ability to make an impact.

There are several aspects to this which I want to touch on and which I believe have implications for how we might use and utilise data as a counter-response.

One of these is political in nature and might be thought of as the emergence of a civic authoritarianism, where perceptions about civic duties are increasingly reflecting sceptical attitudes and decidedly less accommodating attitudes to those who need help.

Now it would be naive to assume that we have moved from a society where everything was rosy, but the environment in which we now work and routinely make judgements has been layered with a set of new social assumptions; the most obvious being the prominence of a culpability agenda dressed up as eligibility, across the lives of those we hope to help and whose needs we seek to address.

It is perhaps inevitable that as our social and public policy has become replete with adversarial themes – and by this I mean the reluctance to accept at face values what people may say about their own needs and circumstances – the bar feels much higher now in terms of getting public acceptance for our own organisations. The scepticism is greater and it appears that the more desperate the group, the harder it is to believe the legitimacy of their stories.

There is real evidence of this. Analysis of 49,000 responses to the National Survey of Third Sector Organisations showed that organisations dealing with socially excluded groups are especially vulnerable, as these are areas where relatively limited philanthropic money has been available.

It follows then that this is doubly difficult for Human Rights Organisations who work with disempowered groups and also within a field that in its own right is under assault.

Human Rights and International support for its ethics and its values is facing a particularly hard time. In a world of fast moving events and conflicts new pressures have assaulted our sense of what is legitimate and proportionate. We have a relentless undercurrent of scepticism at best, and hostility at worst about its validity, and an invitation to see human rights as a corollary of a soft touch narrative. Whether people deserve human rights has surreptitiously infiltrated the way we engage the misery and oppression altering assumptions about the nature of the problem and of the people who rely on us to support them.

A common perception of this in the eyes of the public is that Human Rights issues are now the preserve of vested interests and in some cases a 'legal industry' that uses its provisions to protect the guilty or culpable. At its crudest it plays into a sentiment of unfair systems that mean others are getting advantages "not open to the rest."

We now face a situation where perhaps for the first time Human rights does not automatically elicit the support that it used to.

This is a reality that gets to the heart of our task today….how can we begin to build the capacity and capability required to start using data as evidence that can refute or at least redress the imbalance in some of the myths that are hard to shift.

If Civic Authoritarianism is one variable that has made life more difficult for public acceptance of what we do, then another perhaps more obvious challenge is the declining economic and material standards of households. The fact that the country has experienced increasing hardships since the recession of 2008 has been instrumental in reversing decades of growth in funding. This has been explored at length by the Third Sector Research Centre which concluded that we will almost certainly face a situation where the viability of future plans will be under more pressure than ever before.

It is a view borne out by the Charity Commission which revealed that the recent trend in our income is distinctively down compared to previous years. For example, data from a group of around 50,000 charities that have reported financial information every year since 1998, revealed that a higher proportion have experienced year-on-year decline in real income since 2008 and that a high proportion of these have witnessed more than a 25% decrease. There is no evidence that we will be returning to pre-2008 levels in the near or medium term.

We need to be careful here though. It is not an automatic relationship as it is in contrast to the UK recession in the early 1990s, which did not have a serious effect on charitable giving.

So what does this mean? It seems to say that there is a definite shift in attitudes about the genuineness of charities that in economic downturns justifies opting out.

Cathy Pharoah (co-director of the ESRC Research Centre for Charitable Giving) found in a public survey that the public wanted there to be less charities. Now this leads me to the second external factor: the declining status of our type of organisations in the eyes of the public.

So why has this happened? Civic Authoritarianism may explain trends where certain groups can be easily conflated with social problems and ideas that they might contrast with their own life chances, but the scepticism about civic organisations has other inflexions that suggest scepticism originating from different dynamics.

For example, the rise of profitable philanthropy. I think the emergence of social enterprise models where traditional ideas about charitable giving have been skewed by the notion of the Third Sector, where for profit organisations predominate, has introduced a value for money or a wasted money backstory to how people see giving. More than ever there is a need to use data to deliver evidence about value in ways that challenge the wasted money assumptions.

Finally there is the wavering Public Trust in Charities. This is a serious development and is in no small part due to the recent incidents that, as the Information Commissioner has said, has meant that there is a real danger that charity is becoming a “dirty word”. Kids Company is the most recent, but before that we had stories in the media about charities that shared data and ultimately found out that the details of donors had been sold onto unrelated agencies that were then cold calling households. Inevitably the idea that Charities no longer have the same aims and ethos that previously defined them and they are less trustworthy, has been a negative consequence we could all have done without. These issues around surveillance rights and commercial penetration of private space has brought into sharp focus the ethics of our practice

There are 4 Government sponsored investigations underway as a result of all of the above breaches in trust:

1. Sir Stuart Hetherington, CEO of NCVO;
2. The Chief Information Officer;
3. The Fundraising Standards Board; and,
4. The Charity Commission

The privacy debate is a complex one and there is a need to avoid the polarities that can come to characterise the arena i.e. the nothing versus everything scenario. It can become esoteric and extraordinarily dense and overly defined by information governance 'good practice'. Maybe the best way to look at it though is through a data jurisprudence approach where the issues are separated from the information governance industry and shaped within organisations by the promulgation of a rights and ethics narrative.

So if we accept that we face an environment where there are more charities, less funding, growing scepticism and less trust, we clearly have a major project on our hands to change perceptions.

In fact this brings into sharp relief a tension which I believe is going to define the informatics landscape in the next decade. How to make our data mean something beyond its own intrinsic qualities, or put another way – how we add value.

It is not one that can be dodged. As the policy environs become more data centric, the influencers adopt new forms of communications built around sophisticated data processes, and public services increasingly operate within a data sphere that is multi-faceted. CSOs like ours have to respond by becoming more intuitively aware, skilled and adept at using new methods to deliver evidence beyond the format of the key performance indicator (KPI).

What might this mean?

Well for example, what if we were to derive data from multiple mediums and use the integrated data to predict better? What if we were to think about using new and novel forms of data collection to provide profiles that align different experiences with structured data analysis? If as is quite possible, the 'civic consciousness' through which we sift relevance and most certainly trigger 'agreement' emotions about worthiness is now partly being redesigned by the multi-media and digital world in which we interact as well as transact, then we may need to rethink how we achieve use data to achieve impact.

So a big challenge is to initiate civic society organisations like ours into a new way of thinking about how we view and exploit data and in particular, understand the rules of engagement.

We need to move it from a set of descriptors to a set of constructions that can shape a narrative across different mediums and indeed emotions.

The advent of the internet and its myriad of personal and social technologies means we now operate within a new reality where identities, attachments and language is shaped by technologies that use data almost ideologically - in a way that structures our thinking about relevance as well as impact; where the message will resonate effectively only if it makes sense across a wide range of information sources.

I know in my own organisation, Freedom from Torture, that this has been discussed at length and we recognise that the need to move to the next level requires more than the ability to deliver a powerful message: rather we need to prove that addressing the needs of our survivors is correlated closely with other representations that resonate with the identities of those we wish to engage.

This is very different from eliciting unconditional sympathy….it is more to do with constructing a plausible commonality and is predicated on new ways of data management and expression.

So all well and good, but there is one last challenge which sits firmly within our own orbit and it is Organisational Culture and resistance to – or at least fear of – the informatics agenda. In my experience this is more a state of mind than founded on real premises but nevertheless it is real because it feels real.

Organisational Culture

So one of the main challenges is to do with a state of mind in an organisation. A state of mind that can infiltrate a culture into an organisation's mindset, that lessens its capability to take advantage or exploit its data potential. What I mean by this is that there is a default position on the part of many that when issues surrounding data come up, it is approached from a narrow perspective that either pigeon holes it as the province of a data cruncher (who occupies a particular niche most likely tied up with technology), or reduces it to the lowest common denominator i.e. a very rudimentary self-expression of the message and therefore a communication exercise. In this context the richness of the eclectic forms of data that we could use and express in different novel ways is diminished. Often data as a concept is reduced to a description of routine numerical trends which prevails over the potential to use other forms of integrated analysis that is the hallmark of what is called the big data agenda.

In this respect we have to accept that we may be out of step with certainly the younger generations of our target audiences who multi-task with ease when it comes to technology, and accordingly, will assimilate information in a multi-dimensional and eclectic fashion as opposed to the indicator driven templates that have prevailed in previous bureaucratic structures typical of the 20th century. And yet, this is not an easy transition. All of us will have heard the term "Big Data" and many recoiled at its complexity, not least regarding what it exactly means.

To some extent we have seen this in the challenges that we have had in putting this conference together. Numerous organisations, have I think, failed to take up the opportunity to attend simply because its themes felt threatening and on unfamiliar territory. It illustrates some very real issues about where we need to focus efforts in the future.

So the challenge is for the sector not only to get better at what it does with data and to 'up skill, but to see the potential to meet our challenges across a range of themes by intelligently marshalling our data and information assets in its broadest sense - and in such a way that we can demonstrate not only what works but for whom and why.

The development of counter narratives built on good intelligence that can articulates a new reality will be central to reshaping public ideas about both the work we do, and crucially our relevance and value to them.

Conclusion

Now the conference today is focussed around lots of themes that hopefully will build on some of what I have said but I want to issue a word of warning up front, and it is this; the equation of more information with more impact is not necessarily true by default.

In the NHS we pursued information relentlessly and ultimately could not see the wood through the trees. So data is important in the quest to make an impact, but more important in our line of work is being able to make it mean more than the sum of its parts - in other words deliver more than an array of numbers that usually depict what most of us intuitively would expect to see anyway.

So standing out is not easy and is certainly not helped by the temptation to use data at times as a signpost for authoritative analysis and at the expense of our other innate discriminating faculties.

Session 1: In house data collection: what do you have, what do you need and what skills do you have to analyse the data

Louise Corti: Introduction

We are going to move onto the next session which is all about data. We have four speakers who are going to talk about various experiences of handling data and the challenges of doing things with data. But first of all I asked you to complete a short task. This was about your own experience of what data are collected in your organisation, how you felt it was being handled and capacity or skills to interpret these data, and finally about whether you've shared data. Thank you for contributing.

I undertook a short qualitative analysis of your responses using ten responses. I thought I would just summarise what we found (also set out in Table 1). Generally people are using evaluation data, which we would expect, personnel data is being held, there's a range of quantitative measures, performance measures, KPIs and qualitative information too. So there's a whole range of information being held.

When we asked about the range of system for storing data there was some use of bespoke databases, commercial customer tracking systems like SalesForce to keep track of beneficiaries. There was also other kinds of databases which I guess people are purchasing. That is an interesting point because as it's a proprietary system data can get 'stuck' in there. Sometimes it's quite hard to export data and exploit it outside of the confines of the system

Quite a lot of paper is being held, word processed documents are used as expected, as well as Excel spreadsheets, information and attachments held in inboxes, SPSS files and web blogs of various activities. All in all, quite a range of things that are being collected; and, in the modern digital world, it's not unexpected really.

In terms of handling and analysis, there was a mixture of answers, depending on how large the organisation is, and how many people are actually running these kinds of activities. So first of all, if you use a bespoke database in-house then at least one person will know how to use that; how to handle basic queries and to get things out. Whether or not there is capacity within those softwares to allow more complex queries on data, I don't know, but it would be interesting to hear your views about what you can do with a bespoke database and what kind of queries are routinely can run. Some organisations that do have analysts who are preparing all this stuff. Some are drawing on external skills or organisations to help them co produce and co-develop some of their outputs.

There seems to be a general view that although the databases are handling basic information needs, that more needed information would be useful on looking at storage and security issues when handling data; such as data within email inboxes. We also find this to be the case with academic researchers and we do a lot of training on research data management. Many academics really don't have a good understanding of transferring data - what you can and can't transfer using Dropbox, such as for confidential data. These kinds of practices are happening every day even in academia, the place where you might expect more attention to be paid to that, but actually it's to do with the digital world - keeping up with digital data storage and security, formats and back up, often felt to be a dry or dull area.

There are some useful practical lessons to know about in this area and once you know about them it's all quite straight forward; but if you don't know about things like how to encrypt files or devices, it seems like its quite scary. So there is something to think about there - how can we support our users and data depositors on those critical issues.

In terms of data sharing, your responses were really quite mixed. That's what we are very interested in. some of you are actually working with academics already, sharing bits of data with them to undertake bespoke analyses. Others are handing data to another organisation to do the data analysis, handling and reporting from them. These last two activities are not data sharing per se, but transferring data from one organisation to another and back. Others reports that they have to report data due to legal requirements while others are actually submitting summary data up to larger or international organisations. It will be interesting to hear what kind of things these are, but I imagine they are numeric measures.

Typically then, there is not a lot of data sharing for all the reasons that we are going to be discussing today, mostly concerning ethics and data protection. Some of you didn't see the need need because the data might not seem that useful and it might be quite summary. Others expressed a desire to share data if they were asked, but would need to consider all the legal and ethical challenges. Some claimed that that it was not *theirs* to share, as it didn't belong to them. We'll talk about ethics and confidentiality and consent this afternoon, but there might also be international restrictions across data boundaries. Other stated that they would like to share data, but in terms of scaling that up, they'd need to do some internal work to get their house in order first. I think that is something that experts in this room can probably help with.

So these are the kind of things that have been raised by you, and I think it's a great way to enter into some of our talks in this session.

Our first talks are on UK focussed activities, while the second speaker will address international data matters. I am going to invite Tracy to come and talk first followed by Nigel.

Table 1

Summary of data questions

Types of data

- Monitoring and evaluation data
- Qualitative and quantitative data

Range of formats

- Bespoke database systems e.g. SalesForce, Raisers Edge, Prime, Sharepoint
- Paper
- Excel sheets
- Word documents
- Attachments in Inboxes
- SPSS files
- Weblogs

Data handling and analysis

- Often in-house software specific skills to input and run queries
- Some larger organisation have in-house analysts
- Some collaborate to draw on gain skills elsewhere
- May need more training on issues of storage security, and data protection

Data Sharing

Some do:

- For academics

- To another organisation solely for analysis purposes
- For reporting due to legal requirements
- Submitting summary data to higher level bodies

But typically not:

- See no need
- Might if asked
- Not our data to share
- Confidentiality concerns
- International data sharing restrictions
- Would like to, but preparatory work would need to be done

Session 1: In house data collection: what do you have, what do you need and what skills do you have to analyse the data?

Nigel Fielding: Prisoner of the Past or Hidden Resource: Documentary Records

Presentation slides at: https://www.ukdataservice.ac.uk/media/604176/hrdw_fielding29102015.pdf

PPT 1: Title

I am a criminologist working mostly on policing and the courts. I started my UK career lecturing at Hendon Police College, then got a job at the University of Surrey in 1978, where I have been ever since. My appointment included running the Home Office sponsored programme for training future probation officers. That brought me into contact with the Howard League for Penal Reform, which is the oldest criminal justice charity in the world.

PPT 2: Logo of Howard League

The 18th century prison reformer John Howard was a sickly child who inherited a fortune but whose Calvinist faith led him away from the customary life of the wealthy. He was taken captive on a mercy mission to Portugal, an experience that started his lifelong interest in prison reform. As high sheriff of Bedfordshire he inspected the local prison and, disgusted with what he found, began a national tour inspecting prisons, publishing The State of the Prisons in 1777. Thereafter he clocked up 42,000 miles visiting prisons not only in Britain but across Europe.

PPT 3: John Howard statue

Howard was somewhat of an eccentric – he believed that infectious disease could be avoided by sleeping wrapped in blankets soaked in cold water and despite pleas not to go he set out age 63 to visit prisons in Russia, where he died of typhus contracted on a prison visit. He was the first civilian to have a statue erected to him in St Paul's Cathedral.

PPT 4: Objects of the Howard League

Almost 80 years after his death the Howard Association was founded in London, with the aim to promote 'the most efficient means of penal treatment and crime prevention'. In its first annual report, in 1867, it declared its priorities were to promote reformative and properly remunerated prison labour, and to abolish capital punishment. The Association merged in 1921 with the Penal Reform League, becoming the Howard League as it is today. There are several other bodies named after John Howard, in New Zealand, Australia and elsewhere. In Canada the John Howard Society runs much of the country's prisoner rehabilitation programme.

PPT 5: John Howard Society of Canada

In England, the League's recent campaigns have been highly successful. Under its present director, the interestingly named Frances Crook, the League has developed a special focus on the imprisonment of children, especially girls. England has the lowest age of criminal responsibility in Europe, and imprisons more children than any other western European country. The League has led campaigns to restrict children's imprisonment and recently secured cancellation of the government's plan to build the largest youth prison in Europe. Perhaps best known is its victory in the campaign to revoke the former justice minister Chris Grayling's policy of banning books being sent into prison. A partial victory was secured under Grayling and the policy was then not only rescinded by his successor Michael Gove but the old limit of a dozen on how many books you can have in your cell was removed.

PPT 6: Howard Journal of Criminal Justice

As well as its reform activities, the League publishes the longest-established criminology journal in the world, the Howard Journal of Criminal Justice. It was established in 1921. I served as its editor for 13 years, during which the League was prey to London's office rental market, which imposed constant pressure not only on how it was staffed and operated but also on the maintenance of its archives, equipment and material resources. The League moved to permanent offices, bought outright by donations, in 2000. The offices were opened by Betty Boothroyd, former Speaker of the House of Commons. The League has many prominent supporters, including Michael Palin, Cherie Booth, and Monty Don, and a former President was the playwright John Mortimer.

A last major area of activity is the League's legal team. This was founded in 2002 and now has a contract with the Legal Services Commission to provide legal services to people under age 21 held in prisons in England and Wales. Some of the team work on a pro bono basis, others are funded under the contract.

The League has a staff of 32, which seems quite large, but few of these are salaried posts and even fewer are permanent. It has long had a student association, but it got serious about progressing its agenda by making use of students about ten years ago.

PPT 7: The Howard

Its student section produces a tabloid-type newspaper with a nominal cover price that brings in income and is used to raise awareness on campuses. It runs annual prize schemes for best undergrad and best postgrad dissertations on a topic relating to criminal justice. The cash prizes of £1000 each, and the popularity of criminology degrees, ensures a high profile. A small number of students are employed on a waged basis, but most of the students working for the League do so as placement students. They generally have their expenses covered but little else. The incentive is the direct experience they get in criminal justice work, so it appeals to people with a career interest in the field. As this implies, while students do a lot of the office work, they are also directly involved in campaigns. This includes going into prisons on fact-finding missions, to talk to prisoners about the League's work, and to meet with prison governors and staff to discuss policy and practice. Students also prepare briefings and help draft speeches by the League's senior staff.

I have gone into this history in a bit of detail because it gives us an idea of the information resource that a charity like the League represents. Like any charity, its best resource is its people, but in the League's case the people are able to draw on a rich and deep resource of prison reform materials. These span not only Howard's original pamphlets and tracts, and surveys of the state of the prisons, but documents of all kinds about the parliamentary process relating to the prisons and their reform, Home Office policy documents, and similar materials about other countries. The League's information base goes back to the 18th century.

PPT 8: Dartmoor Prison

So when a prison is inspected and faults revealed in, say, its production of clean and wholesome food, the League can not only document that the problem has been found in previous inspections dating back, in the case of Dartmoor, to Napoleonic times, but to say what previous inspectors recommended and what the prison service committed to. Alongside this information base is a collection of prison studies. Plus, of course, the editions of the Journal, each issue including a Penal Policy File reporting on the policy documents the League received that quarter.

The League's present interest is, naturally, in its campaigns. Like every charity, it wants to make a difference and instinctively this takes the form of lobbying, politicking and awareness-raising. But that also means it is sitting on a resource that keeps growing steadily as its collections accumulate. Periodically one can see the evidence of the information base the League holds in its campaign materials, publications based on research by people who have consulted its materials, and student dissertations. Since the student operation has developed this resource has been more systematically organized and it is recognised that it cannot only promote research for criminal justice reform but improve the League's revenue, for instance, as a partner in externally-funded research. Such research usually includes funding for researchers who can contribute to the work of the League.

The position is thus one of some successes and much promise. But the League is not alone in the field of criminal justice reform. There are scores of such charities in the UK.

PPT 9: Criminal Justice Alliance

Just one amongst several umbrella organizations, the Criminal Justice Alliance, founded in 2011, has over 80 member organizations. The Alliance is a kind of packaging and networking service and has to take care not to encroach on its members' brief. No doubt there is value in coordinating campaigns via such a body but these activities could be more valuable if information was more systematically tended by the charities, and they were more inclined to share what they hold.

When information resources are exploited it is most often in aid of a particular kind of case – ones involving vulnerable defendants like young women, for instance – or a particular policy, like Grayling's book ban. This gives an organization's exploitation of its information resources a bitty, disconnected character. Moreover, materials may be collated for such purposes and when the spotlight moves on the materials remain in that collation, making it hard to re-assemble them for a new purpose. Often only a fraction of materials are held in digitized form. That is not to say that digitization solves everything. The police researcher Peter Manning gives the case of crime analysts in a large US police department. They wanted to analyse trends in local crime and asked where the records were. They were taken to the basement of the force data centre.

PPT 10: Stack of Old Computers

There, piled on top of each other, were layers of obsolete computer hardware, oldest on the bottom. The records were held in unsupported formats on dead machines that were not interoperable.

I mentioned the League's use of student volunteers. They are an asset because they are, in management speak, 'agile'. They are open to new ideas, keen to try things out that they have recently learned, and enthusiastic. But they are there only a short time, at most, 6-7 months if it is a nominal one year placement. Turnover works against organizational memory. Much of the material artefacts produced in their work for the League will go with them. To an extent this applies to other volunteers, but students are likely to have some fundamental research skills and knowledge of methods and often use the placement to do primary research for a project or dissertation, so can make a large contribution to the organization's research capacity. So it is a good idea to make the closing stage of their time there explicitly to do with documenting what they have done and the materials they have used, organising the materials in a way that will be useful to future workers.

If an organization wants to exploit the materials it has, and perhaps to scale up its value by sharing with other organizations, human resources are not the only kind it needs. If we take the standard distinction between quantitative (numerical) and qualitative (textual) data it is likely that the organisation will have a considerable amount of qualitative data. A great deal of the donkey work in exploiting such material involves clerical type data management. Materials that were born digital or can be digitized can be organized and analysed using qualitative software or 'CAQDAS'.

PPT 11: CAQDAS

As well as text, CAQDAS provides tools to work with visual material – still or moving images. Most CAQDAS packages can perform basic forms of quantitative analysis as well, and export information to statistics packages. Some also provide tools to work with geo-referenced data. Moreover, there are several freeware packages that do these things and can be freely downloaded. Others are available on a 'freemium' basis, with basic features free and more advanced ones requiring a paid license, or the developer will make the full package available for a limited time. A one day workshop is enough to pick up the essentials of these packages, and online self-learning resources are freely available from the CAQDAS project at Surrey.

PPT 12: Nexus Network

Recently the ESRC funded the 'Nexus Network', concerned with issues relating to food, energy, water and the environment. It held a workshop at Sussex this summer to discuss how NGOs could be mediators between researchers and society, acting as a 'boundary organization'. An example is the York Environmental Sustainability Institute. Although its 120 members are based at the University of York, it operates as a broker between the academics and NGOs. Its recent work includes a project in India at the junction of food, water and climate change issues. As the monsoon becomes less reliable due to climate change, new types of rice are needed to maintain yields. But the search for new cultivars did not begin with botany but with social science.

PPT 13: Rice farmers

The essential requirement expressed by the women who work harvesting rice was that it should not take any longer to pick than existing varieties. Because the Institute and local NGOs regarded them as an information resource, the scientists working on the genomics of the new cultivar were able to start from the farm workers' request. The project partner is a farming cooperative with 40,000 members, so the project will have a widespread impact. This is an example of the wider point that it is often possible to get bigger and quicker change as an NGO by working with the business sector rather than with government. While firms will eventually compete to supply the seeds and rootstock, at this initial stage they are cooperating with each other and the rice farmers. So at a certain stage of an intervention an NGO can benefit from the pre-competitive cooperative stage where businesses are working with each other and project partners.

PPT 14: Summing Up

In this talk I hope to have given an idea of the ways that an NGO can leverage its information resources to advance its agenda without incurring high costs. I've particularly focused on student volunteers and on IT, but another part of the academic landscape is also relevant. UK academics have their research evaluated every five years and this determines the amount of one of the major government funding streams to universities. Other countries have similar systems. In the UK, 20% of the points in the 2014 evaluation exercise were awarded on the basis of whether the research had real-world impact. That means that UK academics are now strongly motivated to demonstrate impact, because their career advancement partly depends on it. We already know that the 2020 exercise will increase the percentage dependent on impact, at the very least to 25% but, if the Treasury prevails, to much more than that. Academics have already been told that they should develop relationships with those external bodies who have an interest in their research and that, alongside business, NGOs are prime impact partners. So it is likely that if an NGO is sitting on an information resource, it will be able to cultivate mutually-beneficial relationships with high calibre researchers who can provide expertise in exploiting such resources.

Session 1: Driving a Ferrari into the desert and leaving it there? The challenges of information management UN peacekeeping and humanitarian NGOs

Roisin Read

Presentation slides at: https:/www.ukdataservice.ac.uk/media/604180/hrdw_read29102015.pdf

I'm Roisin Read, I work at the Humanitarian and Conflict Response Institute at the University of Manchester and I coordinate the making peace-keeping data work for the international community project which is funded by the ESRC. It is quite a mouthful. It looks at the use and production of UN peacekeeping data. We have a historical dataset from the UNAMID mission in Darfur but we also look at the politics of data - production, questions, about the wider humanitarian information landscape and as part of the investigations into this we started talking to the wider humanitarian community about the information they collect, the dataset we have is primarily security incident data and this is the kind of data that we were initially primarily interested in obviously contains a lot of human rights information as well. One of the things that we found was there was a really interesting thing going on in our background conversations where the UN were telling us that one of their main sources of information about security incidents was NGOS then NGOs was telling us that there main sources of security incident information was the UN. There was no clear input into this system so this got me interested and I actually bid for an impact acceleration account money to look into this humanitarian information landscape particularly as it relates to security data.

In the run up to the World Humanitarian summit there's lots on monitoring and evaluation data, lots of needs data but very little attention being paid to the security data that is collected. One of the problems I found with this is that there is no clear idea of what security data is either by the UN or the NGOs its very much blurred with wider situational awareness data its often told on a person to person basis the margins of meetings was a phrase that I got told by one of my interviews but most of this kind of security information data particularly in context like Sudan where there is quite an antagonistic relationship with the government was happening informally and then as I was putting together this bid for the IAA, I spoke to people within NGOs one of whom told me his colleague works for one of the big humanitarian organisations in information management, described their current approach to information management like this: its like buying a state of the art car, driving it into the desert and leaving it there. He said that all the attention had been going onto these new systems, what these new systems could offer and a lot of money was being spent on the new kind of data management systems but people weren't using them, they weren't fit for purpose, they weren't solving peoples' needs so they were just carrying on using the existing methods that they had used before and not engaging with these new systems. So there was the investments for storing, analysing and managing data was simply so far ahead of the desire or the capacity to actually use them. So I took this as my starting point and continued speaking to humanitarian professionals, attended several workshops and run some workshops to looking at how knowledge evidence is collected, information is managed within the humanitarian sector. So then my interest is in primarily security data I found that there is really no way to distinguish this from the broader information environment.

So overwhelmingly the biggest criticism I've heard from the interviews that I've done with humanitarian professionals is that there is simply a clarity or understanding around why data is being collected be it security data, needs assessment data, it's just done because it's always been done. This has particularly serious implications not only for generating value from that data if we had this culture of well we've collected this data before so we'll collect it again, but also in terms of the expectations of what will be done with that data. In the case of Sudan, there are extremely serious implications on the part of those who are sharing data with international organisations so there is an expectation that something will happen as a result of their sharing data particularly when it comes to sensitive security data. One of my colleagues on a peace-keeping project I should mention that Celia Russell is also one of our project members, did interviews with former front line data gatherers from the UNAMID mission as well as with Darfurian refugees in Chad about their experiences of collecting or sharing information with the UN. One of our findings has been that people had stopped reporting incidents to UNAMID because they had initially this expectation that doing so would improve their position, their protection. People would be more secure because they had shared this information with UNAMID. Actually what we found in the case of our interviews is that people were targeted by the security services for speaking to UNAMID and there was a widespread perception that if you shared information with the UNAMID mission that you would then be targeted by the security services. So people just stopped sharing information.

This was related to another problem that was identified which was a problem with the basic front line ability of staff to collecting that information. They don't know what the data is for they also often haven't been trained in data gathering and the issues which hampers their ability to collect useful information and also to make claims about what that data will be used for relating it back to the first point. So again from the UNAMID staff very few of them had formal training those that had where primarily from two PowerPoint presentations in one of the offices in the UNAMID mission just one of the field offices covering investigating and reporting human rights violations and documenting what they should do with a special focus on interviewing in practice. So I've included one of the slides from this training that they received. But again this was not wide spread this was one office that had done this and that actually a lot of our interviews said that the problem was that these just weren't being followed in practice.

As I've mentioned the refugees in Chad said there is an extremely high degree of suspicion about sharing information and they felt that entire transcripts of testimony that they had given were being, that the security services had access to it. There is also a further risk that the interviews who worked for UNAMID have it themselves because of their position working there and their lack of training in data protection had put them at risk as well as the conduits of that information. There is several that suggested that the lack of training of staff members saying that interviews were conducted in public places, there was security services staff around and that people just weren't following the guidelines in practice, consent was particularly problematic, people said it was sought but given verbally but there was serious doubt as to whether it was truly informed consent as the potential consequences of sharing information were never explained and also what was being done with that information wasn't really being explained.

So the overall consequence of this is as part of this data gathering, people were actually being put at risk and the information collected by the UNAMID mission itself because of its precarious position within Sudan, wasn't actually able to add any value to protection from having this information. They were just collecting it because it was politically impossible for them not to collect human rights information but actually it decreased their ability to access areas in need. This is just one of the other slides from that. And then again overwhelmingly this problem of data sharing comes up again and again. Sharing between agencies, there is a real lack of understanding of what actually the rules are, what is the international law on data sharing and what were the codes of conduct which should be in place and at the workshops I've been to as part of the world humanitarian summit again and again this is one of the things that people say needs to be clarified. What are the rules, what should be the rules, what information can be shared, what can't, which country takes precedence, which countries data protection rules take precedence, is it the country you are in, country the organisation is based in, who should be in charge of deciding that, what kinds of information can you share under what circumstances and this is the problem again with building the relationship with academics. Can we share this information? As a project we've had a huge problem with the dataset that we initially began with because it was very unclear who owns the data, who the rights holder of the data is, is it the mission, is the African Union, is it the UN? So we have come up against some of these problems ourselves.

And then finally the difficulty of finding inputs into this system so one of the humanitarian interviews I spoke to said that these reports seem to throw around intuitive feelings that there is no clear input into the system and in his spare time this particular interviewee was trying to go through a stack of reports on the situation in Sudan and trying to find the source for particular incidents just because he was so frustrated because it was really unclear where any of this information was coming from and it was just this rumour Chinese whispers somebody said something to someone at this meeting that this had happened and then it escalated and it became a fact got into all the reports but wasn't clear where this information was coming from. The other problem is that the information seems to be flowing upwards so this is based on my colleague who worked in the UNAMID mission, this is his understanding of where information goes, so upwards is actually down on my slide, but so it started with the incident itself and then you might have witnesses or people who have got the information but then it would just go further and further up and be pushed up but it was very unclear what was being done with it at each stage so there is a process of editing of summarising but it wasn't being shared horizontally it was just flowing upwards to the security council level ultimately and the DPKO but seemingly not very much was done with that information and I just wanted to point out this not only seriously restricts the ground level understanding of that but can have other implications and this is the example from my partner on the IAA project which is MSF and they have pointed out that with regard to security information and how they manage security information, initially what they did was take it on a case by case basis from one of those staff experienced in security they would look at the circumstance and try and understand what happened and actually at that level it was a lot of personal dynamics, maybe someone had rubbed someone up the wrong way or that they had been behaving in a particular manner that aggravated something in the local community or somebody just simply didn't like them but actually once you take that up to the headquarters level and aggregate that as a statistical problem, you start then getting solutions which come with very universal application that we must respond to security in this way and they traced a direct link between looking at security incidents against their staff at the aggregate level and the way that aid has retreated behind fortified aid compounds and that actually this has really serious implications on the relationships on the ground between aid workers and their recipients and they draw a line to that then because they don't have that relationship on the ground with the participants there security deteriorates and that leads to targeting of aids staff in itself so this breakdown and that you can have if you look at the wrong level of information, or look at information at the wrong level to analyse it. Then, just to end, there is a general sense and this has been mentioned again and again that it is not the amount of information that is being collected that's the problem, it’s that we don't really know what it’s for and we're not really sure what we are doing with it but actually there are different needs and we need to take these different data requirements within organisations but also between organisations and manage them differently that we need to take them at the right level, some organisations said that they wanted lots of information and then to be able to use it like intelligence analysts others said they needed more detailed long term information and that wasn't enough and just to finish, I thought this was a particularly nice quote from one of my participants.

Louise - that's brilliant. What's really interesting there I think is this idea of provenancing claims that have come out and certainly in academia there is an interest in transparency and actually supporting a claim back to the original data and I think it’s almost the same kind of issue how do you do that safely and effectively but so it doesn't take too much time so it’s really interesting.

Session 1: In house data collection: what do you have, what do you need and what skills do you have to analyse the data?

Ingvill C. Mochmann: Children Born of War: expanding the evidence base on hidden populations.

Presentation slides at: https://www.ukdataservice.ac.uk/media/604179/hrdw_mochmann29102015.pdf

What I am going to talk about is the on data collection and data sharing we are doing in the international research on children born of war (CBOW).

I am a political scientist from Bergen in Norway and I have been working in Germany in the Data Archive, working closely with the UK Data Archive (UKDA) for 20 years. At the same time I am professor of international politics and vice president for research at Cologne Business School and an affiliated expert at Harvard Humanitarian Initiative. So I work in disciplinary international methodology and I think this is important when it comes to experiences that you have from the field when you work on these kinds of topics.

I will focus on population - we talked about the 'vulnerable population' a couple of times - and the group I am going to talk about is particularly vulnerable. I do not know if you can see it clearly but here you see a woman going through the streets with a shaved head and holding a little baby in her arms. Most people from Europe know about this particular population of children that were born from relationships being either loving or abusive situations of soldiers from armed forces in Europe, during the war and in post war periods. And actually this is a topic most people consider to be at the part of a particular time but what I am going to show you, and hopefully you will agree with me, is that this is a topic without time and space. It has always been and it is likely to always be, as long as you have any kinds of conflict.

The data and the sources on CBOW that we have from previous conflicts, I think are essential to know about when we look into present day conflicts and that is where Non-governmental Organisations (NGOs) and charities come in because, what we realised very quickly, is that actually even getting data and information on this group requires acceptance of the concept of a group called CBOW . Having a definition of who we are talking about is a part of the process of getting it on the agenda because it was a non-topic - a total taboo. So we depended on lay researchers and their organisations and the intersection between academics, NGOs, Intergovernmental Organisations (IGOs), governments and the military.

Just very briefly: who are the CBOW? We have decided that these are "a child that has a parent that was a part of an army peace keeping force and the other parent a local citizen". We do not, at this point, distinguish between consensual and abusive relationships because the evidence we have so far is that this is really difficult to distinguish clearly in times of war. The United Nations (UN) defined at one point, all kinds of relationships between UN peacekeeping forces and local women as exploitation and relationships are forbidden. With the consequence that nobody is looking for what actually is going on, as the problem is not supposed to exist. But we know that is not the case. So a hidden population is a population which is difficult to access and very often then also vulnerable - you do not know the size, you do not know how to ask them, you do not know how to involve them in research.

Some pictures of different categories of CBOW to show you there are real people behind it. These are actual children born of war from different types of conflict throughout different wars. When it comes to sources, we use several different ones, basically everything accessible from personal documents, letters, photographs to medical records, administrative records etc. Very often this information was found when the mothers died. This information was used as the first step to actually go and look for facts, real facts. It was often a first indication that there may be something more in their life than you actually thought there would be. Also films very often provide first information on relationships and possible CBOWs, and different kinds of media, now in today's conflicts of course there is the internet and blogs, but also court records. I've worked with people from the international criminal court and every time we hear about court cases in present conflict zones where sexual violence has been used, we also ask is the pregnancy level higher, or are there any children because nobody seems to ask about them. These children are particularly vulnerable as we know from many conflicts that the children are not registered, are abused, discriminated against and stigmatised in society. This is something we see from Uganda, from the Democratic Republic of Congo (DRC) and even from Bosnia, where even the community tried to embed the mothers and their children.

In addition to these sources we have different surveys, quantitative and qualitative both from Norway, Denmark, Netherlands and more recently from Austria and Germany. Colleagues are working on CBOW in DRC and Bosnia, where, partly the same questionnaires are applied. Based on the information we have so far, I have developed a framework of factors important to the life development of CBOW, and it gives us an indication of what is relevant to this group, and helps us to know where we can we look for data and what kinds of data may be relevant to collect in present day conflicts. We have come far in this research field in the past decade. A colleague of mine was actually talking about the children born of war at the UN general assembly in June. In recent studies, we involved lay researchers from various national associations and asked them to help us to develop the questions, help us to check whether we are asking the real questions that are of relevance to them and their peers.

We will talk more about ethical issues later, but just very briefly, one of the practical challenges with the data collection was actually that when you involve lay researchers, you have certain difficulties in agreeing on different kinds of practical issues regarding questionnaire design and methods. Also the participants told us that it was partly very exhausting for them and the confrontation with past experiences, had been emotionally difficult. So now we have psychologists on board and we involve them. The validity of memories is another issue and I think it very important that when you involve and you work with lay researchers, to be aware that all are different; the lay researchers, CSO, charities, researchers etc. Furthermore many studies are carried out repeatedly because it becomes fashionable and there is funding available. Increasingly, people tell me that those participating in focus groups, for example in studies on sexual violence in the DRC, actually ask what is in it for them.

Louise: A couple of things that you raised for me is how we could ask front line organisations to gather data that is useful for others, because actually the best point of contact to get information for, if it was structured information that was useful, but how one can do that is difficult. I think the second issue there is over researching, as you say there are so many groups that are collecting testimonies and things about conciliation and other things but actually, is there evidence that that research is having an impact?

Session 2: Making an Impact: Using Data Beyond Key Performance Indicators

David Walker: Introduction

I am a journalist by background but have also spent a bit of time over the years involved in social research governance. Many congratulations to Louise and Neil for making today happen, and to the ESRC.

Recognising the huge opportunities for the media, for the producers of knowledge, civil society organisations, in exploiting data, I need to register a few caveats. They include pointing out that the purpose of our organisations is rendering service to people. We want to maximise our effectiveness and that the utilisation and collection of data are important. But we must not forget our ends; we should be pragmatic about means.

Human Rights organisations, governments and general public are increasingly turning to images, videos and text to investigate and understand the human impact of conflicts, disasters and political violence. It remains unclear however whether this data actually improves communities' ability to protect the rights of vulnerable individuals around the world, particularly those who lack reliable and secure access to the internet or those whose rights are violated in private – we have to acknowledge that 'private' can mean 'not visible to our western/northern eyes'.

Another caveat. There is a slight tendency in universities to believe that 'to know is to do'. Generating then curating data are processes; the object is putting the data to use to provide services to real people and to influence policy. Knowledge – enriched by data – is not an end in itself. Those of you who work for Non-government Organisations (NGOs) and civil society bodies are often eclectic about where the data comes from. Your test is utility – though we must make strong efforts to validate the data's provenance. Academic researchers sometimes seem to believe that knowledge generated within a university (and subject to peer review) should have primacy. We, however, have to weigh its usefulness and practicability; we have to be pragmatic. Perhaps there is a slight element of tension there between the two approaches.

Exploiting big data puts a premium on quantitative skills (on the part of NGOs and researchers) and numerical understanding (by the public). The Royal Statistical Society has tried to assess how far the public grasps probability, the calculus of risk and so on. Its findings are cautionary. Despite the best efforts of those presenting data in new and exciting ways – visualisation – the public has only limited grasp of quantities. That means we have to be careful that the data we put before our audiences is comprehensible.

Ideally, the data (or conclusions based on it) are presented as part of a 'story' – the data has to tell us something that we did not know. Usually, to provoke media interest, the story has to feature people. It has to play on the public's capacity to sympathise and (or judge). Big or small, data has to tell a tale.

We have, also, to accept that the data, even encased in a good story, will not necessarily change attitudes. Visualisation and graphical presentation of data are important but might not, in themselves, help accomplish your mission, which might be to move debate and secure new directions in public policy. Erroneous public views can persist, regardless of the evidence, for example on the amount of social security fraud or migration volumes. NGOs working on the Middle East also contend with deep bias – from media proprietors, for instance, which skews coverage in the press.

NGOs with a mission to campaign and lobby will recognise the challenge. They are having to deal with recent changes in legislation and the regulatory regime for charities, which is less favourable to campaigning. Utilising data will not insulate charities from the charge of doing 'politics', nor NGO staff's conviction that they are alerting the public to a moral scandal. The wrath of William Shawcross (chair of the Charity Commission) will not be stayed by rigorous exploitation of the database.

Sources of data are expanding. Tools for exploiting it are becoming more rigorous and sophisticated. But for charities, old questions about mission and effectiveness have not gone away.

Session 2: Making an Impact: Using Data Beyond Key Performance Indicators

Bob Jones: The challenges of using data to assess health needs and impact in conflict zones

Presentation slides at: http://www.ukdataservice.ac.uk/media/604177/hrdw_jones29102015.pdf

I don't know where to start really. The work of Medical Aid for Palestinians, and the way we use data, relates to a lot of the areas that have been talked about today. I will avoid talking about the lobbying act and the political situation in the UK and will instead focus on the health impacts of conflict and data gathering and the challenges of this context.

David mentioned long lasting, slow to move issues. Palestine is definitely one of those. Palestinians are one of the longest suffering refugee populations in the world and hold a unique place in UK public opinion. Support for Palestinians in the UK is broad and, partly for this reason, MAP is 70% funded by individual donations. Additionally any time there is a protest on Palestine you might get 100,000 people out like last year, or large numbers anyway, so in that sense we have a lot of public backing but as mentioned by others with regard to the lobbying act and other legislation, we are also operating in a shrinking area in terms of civil society space.

MAP has been around for 31 years. Founded in 1984 after the Sabra and Shatilla massacres in Lebanon. Our founders were orthopaedic surgeons who were working in the camps at the time. They came back to the UK after the massacres and founded MAP. Since then 90% or more of our work has been focused on programming - health programmes, working on capacity building within the ministry of health, employing midwives in refugee camps, mobile health clinics in the Jordan Valley, things like that. But predominantly working with local partners.

In February this year we expanded what was our communications team to be an advocacy team which is now 3 of us. Since then we have done a lot more UN engagement, EU, UK parliament and other advocacy. The drive behind this was that we have worked for many years on the symptoms of occupation, the health needs, essentially. I hate to use the pun, but applying a bandage to the occupation. We decided that in addition to our health programmes we also need to be addressing the wider socio-political determinants of health. For that we do rely on data, both qualitative and quantitative, our own data and that of our partners. In that sense this workshop is central to how we are trying to evolve our work from being an organisation focused on a small team in the UK doing mostly fundraising - alongside work with local partners through our offices in Palestine and Lebanon, to being a medium-sized organisation. We received a lot of donations with the Gaza attacks last year and we are now trying to build the organisation in ways that make use of our own data from health projects in order to try and improve the context of those health issues.

I also have a monitoring and evaluation colleague here with me, Jane, who can attest that there are many other ways we could be using our data at the moment, and actually its monitored in different ways from different offices so one of the things we are keen to develop is how we standardise how we monitor and evaluate our projects to get that data and how we can use that to inform research, reports and our advocacy in the end.

It was mentioned before by Neil about policy input and how the policy debate is being more and more informed by engagement with civil society and obviously with academics as well. I've got one example of that here. We did a project with DFID following the Gaza attacks last year and we had a request from a parliamentarian a couple of months ago. She had attended one of our public events where a doctor who had been in Gaza last year mentioned about multiple amputees and the situation of people with multiple amputations in Gaza. Following this event she wrote to us saying how can I put questions on this to parliament, what kinds of issues should I raise. So we helped frame the question - and then DFID came to us and said how do we answer this question. In one sense this is quite unique for Gaza, in that few people have access, but it does show quite a reliance on the information coming from civil society.

As I say, Gaza is unique in many ways. No delegation of MPs from the UK has been to Gaza since 2009. Ministers can go sometimes but this is front-benchers so they usually aren't very public with what happens afterwards. One of our campaigns is to advocate for MPs to get out to Gaza, for them to experience it. Someone mentioned before about aid-accountability, this comes into it here. UK money is being spent on projects but MPs can't go in and see these. Part of our work involves taking delegations out to Palestine and trying to campaign for those delegations to have access to Gaza and other areas.

This trend is actually worryingly also repeated at the international level as well. Since February when we started engaging more with UN bodies we have been involved with three consultations: the UN Special Rapporteur for Palestine; the Office of the High Commissioner for Human Rights; and a Commission of Inquiry, which was set up for Gaza. These bodies have been getting in touch with NGOs to say we are looking for information into what has happened since the attacks can you help us. Who has been affected, what has happened since, how has it recovered because again the UN commission isn't allowed into Gaza either. So the fact that we have access puts us in this position of privilege in a sense of providing some of the information, but also obviously bring a responsibility in making sure that information is representative.

That access itself is central to how we are limited in our campaigning. A lot of humanitarian organisations we find are very cagey about their advocacy because of humanitarian access - we are no different in that sense. When we went to the human rights council in March and June, we were the only humanitarian organisation that was giving evidence because we also work on human rights issues alongside that. It does put us in quite a precarious position and there are quite a lot of civil society organisations that have been set up specifically to target human rights and humanitarian organisations working on Palestine, which in a sense is good for us in someways - it makes us a lot more aware of how we need to be careful with our messaging. Every report we do gets taken apart and every message will be analysed to see if we are being biased in certain ways. We must focus on a story that has been heavily documented so we have video evidence, photo evidence and it has been documented by a range of people, it means that our messages are very solid.

One of the reports that we did included this infographic page. We worked with our local partner called Al Mezan Centre for Human Rights and a UK based organisation called Lawyers for Palestinian human rights (LPHR). Al Mezan provided the data, they documented every incident of attack during the Gaza attacks. the impact that had, the stories and the wider data, quantitative and qualitative, so we worked with them to put this together from what was originally a private submission. It took a long time to make it publicly consumable and we used info graphics to try and draw up some of the main messages. Some of these are a little bit more obvious - we've got numbers of attacks from 2008/9 - 2014 demonstrating the difference in escalation, attacks on ambulances, hospitals and this 511 stat was one of the main stats we were using when we were talking to media organisations. This was a new stat, no one else had access to this dat.
We were saying that 511 of the 2217 people who were killed never received medical assistance. It sounds like a powerful stat, and it is, but it is also very problematic because we don't know whether they died because medical assistance was prohibited from accessing them. We had to make this very clear when we were talking to the media, clarifying what this stat actually meant but this is an example of how using data in a very challenging environment becomes very problematic because we have to very carefully manage what that data ends up saying. We can't be seen to sign up to a message from data which we cant actually say or we don't know that from the data ourselves. So in that sense our indicators need to be very specific when we are using data.

All data is seen as political in this area. We get different data from the UN, NGOs, both local and international, Governments - Israeli and Palestinian. Which data you use is very political because of that. The data sets will have different methodologies, different assumptions built into them, which is why we decided to partner with Al Mezan as a local organisation with the stories all written into their data because it was seen as a bit more of a legitimate way of narrowing that field down.

Interestingly, a week before we released this report, Israel released its own report which was about 5 times the length and significantly more driven by data, but their own data with different definitions of who counts as a civilian, for example, which again is a challenge to this environment of being so competitive with data narratives. The week before we released this one they released one and the day after we released ours the UN released there's. That part was planned. When you are talking to media they have access to all these repots and you are trying to prove that the narrative you are putting forward through your own data is the legitimate narrative in this competing environment.

That's the situation in particular to the Gaza attacks last year but when we are taking these issues globally we require broader coalitions. We are also part of the safeguarding health in conflict coalition, which is an international coalition of organisations trying to highlight attacks on ambulances and medical practitioners globally - and yet again they use a lot of data mostly qualitative at the moment but again trying to be more quantitative. I think generally there is a sense that because data gathering is so difficult in conflict situations and there are so many competing narratives of what goes on, its very hard to get agreement within a large coalition internationally - so they tend to be very top level reports and not particularly useful for advocacy in some ways.

Data is also highly contested on specific medical issues. There was a lot of controversy recently around a UN report in Gaza which was talking about neo-natal deaths. A UN report came out demonstrating that neo-natal deaths in Gaza had gone up. The ministry of health immediately came out, and this is from the Gaza ministry of health, this is wrong and the rates had gone down. The way this was represented was of course embarrassing for the doctors and those who were working in the ministry of health. It came out later on that the sample size was small and the way the questions were asked was problematic. Women would turn up and they would ask them - Is your previous child still alive? - The way they interpreted that data was very problematic. But I will not go into too much detail on that here. The clear message in this case was that when you attempt to expand a complex narrative to more than a micro scale, everything becomes more contentious and there are so many more different narratives.

One of the things we are looking into at MAP is partnerships with Universities actually. We are looking at a partnership at the moment with the American University in Beirut to do an assessment of maternal and child health services in Lebanon. Essentially bringing a more academic analysis to the project work we already do and try and draw up the main themes and use that, to back up the advocacy messages in our reporting. I think in a sense because we have this unique position, particularly within Gaza, there is a greater possibility for the findings of this work to contribute to academic research - but also advocacy messages to inform policy.

I have one example here of good use of data - not actually done by us though, but this is aspirational. One of the things we want to try and do is map data from our work to display trends in a simpler format. This work (The Gaza Platform) was done by Al Mezan, Amnesty International, Forensic architecture and the Palestinian Centre for human rights. They pulled together all of the stories and case studies and mapped every attack that happened last year. The method of firing used, the casualty type and the dates. You can also see all the stories here as well. Just one of the things I've filtered to demonstrate this tool is on artillery. 305 people were killed by artillery, 290 of those were civilians and 98 of those were children. That already tells a story about the use of artillery in heavily populated areas and when you combine this with the narratives that come from military organisations like *Breaking the Silence* - ex-military human rights organisations, you get more of an idea why this has happened. Policies put in place by the military, how they use artillery, how they distinguish between targets. We wouldn't use this particular data tool because its not specifically medical (although they have mapped attacks on medical facilities), but we are looking at doing something similar with medical referrals, and a wide range of other issues.

If you have any questions I'd be happy to answer them now or later.

Session 2: **Making an Impact: Using Data Beyond Key Performance Indicators**

Emma Prest: **Enabling the application of pro bono data science to humanitarian problems**

Presentation slides at: https:/www.ukdataservice.ac.uk/media/604184/hrdw_prest29102015.pdf

I'm Emma, I am the General Manager of DataKind UK and I am going to tell you a bit about DataKind, who we are and how we help non-profit organisations. Plus I have two case studies for you. I think it is fair to say my talk is a little bit different from the one that has just gone. This session is about going beyond key performance indicators (KPIs). We do a lot of work with non-profits to use their data internally so it is a bit more inward looking, about how data are used for strategies as opposed to outward thinking and campaigning.

DataKind was founded four years ago in New York. Our fearless leader Jake Porway was working for the New York Times in research and development. In his spare time he would go to hackathons, where people would spend weekends not sleeping or showering, sitting next to others with brilliant machine learning skills or superb coders and getting really excited about their skills and their ability to change the world. In the end they would produce an app that would help you, for example, park your car. Jake got so frustrated and felt so unfulfilled with what was achieved at these events. So he thought, what if we take these tools and techniques that private companies are using, specifically data science tools and techniques, that private companies use to increase their profits and apply them to non-profits working on really tough social issues. So that was the impetus for DataKind. We now have six global chapters around the world: UK, Dublin, Bangalore, Singapore, San Francisco, DC and New York and we are all adopting the DataKind model and developing it in slightly different ways depending on our contexts.

We work with a lot of data science volunteers. They generally have full time day jobs in the private sector and maybe a PhD in maths or stats. or physics. They have a burning desire to volunteer in their spare time. We bring these data scientists together and we help scope and manage data scientist projects for non-profits. We have actually tapped into an amazing appetite amongst the data science community so we have lots and lots of data scientists, hundreds of data science volunteers signed up to volunteer with us in the UK.

What is data science? Data scientists love predicting things and classifying things. They use machine learning techniques. I will give you an example of when we use machine learning. We worked with a non-governmental organization (NGO),called Give Directly - they are all about giving money directly to some of the poorest villages in Kenya and Uganda. They realised that one of the indicators of poverty was whether you had a thatched roof or a metal roof on your house. If you had a metal roof you were richer; thatched roof you were poorer. Using satellite imagery, and learning algorithms we were able to create a machine learning model so they could identify which villages had which roofs. Previously they had teams on the ground going to villages and noting down roofs types. We built a machine learning model so the computer classified which houses had thatched or metal roofs. We were able to help Give Directly much more detail, so they could efficiently work out where the poorest communities were, using fewer resources.

Statistics normally has a model that you can put your data into, whereas with data science you often have no model and it is a bit more exploratory. You can analyse unstructured data e.g. text and photos and the plethora of data coming off the internet. I have to say we do a lot of work that is not pure data science because a lot of charities who come to us have a slightly lower level of data needs, so we do a whole range of things. We bring together teams of volunteer data scientists, we help to translate because the data scientists who have their own jargon and the NGOs who have different jargon. It can take quite a while before everyone is speaking the same language. We demystify data science and explain how it can help a charity, we deal with a lot of very sensitive data and help NGOs to share that with volunteers and we ultimately try and create solutions that can help the NGOs do their jobs better.

We have various programs. We have free services such as meet ups where we bring data people together with charities. These are evening events, getting them in the same room and an opportunity for charities to pick some data science brains. We do data weekends which is like a hackathon, but much more structured and successful. We will get 80-100 data scientists turning up for the weekend and we will have three or four charity projects, usually exploratory projects. For example charities bring along their database of data that they have collected for a decade. They want to know what is in it.

We also do longer term projects where we will spend six months with an organisation working on a project. At the moment we are working with Age UK Islington, to help them predict who of their clients are most vulnerable.

Now, on to case studies. Not really human-rights or humanitarian but nevertheless they are solid case studies, so I am still going to use them. One of them is a group called Buttle UK. They are a small organisation of about 30 staff based in the UK. They give crisis grants to people in need - often families and young people-that are in need of money to buy something like a washing machine. A lot of them do not have money to buy white goods, they have been made homeless, they have moved somewhere with basic furniture or, it can be clothes or supplies or short term grants of a couple of hundred pounds. Buttle came to us and said, they had their head around the quantitative side, but every time someone applied to them for one of these small grants, the vulnerable family or the social worker filled in a form with loads of text data which no one had ever looked at. Their idea was to see what is in this text data, what are these stories that people are telling us that we are missing. What are the relationships between all these individual grants that are coming to us.

They came to us with a great amount of data. They had all this narrative text data. They also had general data on family finances, reasons why they needed the grant, the amount of the grant, what it was for, location, dates... really rich data. During a DataDive weekend, 80 data scientists got stuck into their data and it was really fascinating. A couple of really immediate insights popped up. Bed wetting turned out to be surprisingly common and they realised that it was an indicator of high levels of stress in a household which was often linked to high levels of debt. They started to see some specific terms in these grants, showing how bad the families' situations were. Could they better understand families' needs and be better at pointing them to other services? They started to see the repetition of people coming back for things, e.g. a family would come with domestic violence issues and they would come back six months later with child development issues. Understanding these cycles and patterns in the data was really important.

Buttle took these findings to their next board meeting and they discussed whether they should be doing something more than just providing short term crisis grants. If we are seeing these patterns should we have a longer term programme? Should there be other kinds of support we should be signposting them to? They are now looking at a second stream of longer term support for the most vulnerable families.

But even more, they started to do some internal text analysis themselves. It was their monitoring and evaluation expert who came to the DataDive. She got really excited watching the data scientists and at some point suddenly went "hold on I think I can automate a huge part of my job". She went away and learnt Python. Previously, for a whole day every month, she had to pluck out key figures to give back to management and she realised that if she applied a script she could do it all automatically. She went away and learnt some code and at the click of a button, now does that job in a few seconds, rather than what used to be a few days' work. It is a really great case study showing an organisation what is in their own data, but what is much more interesting here, is that they were able to adapt and react based on the findings. I think that is because they are small and more agile compared to the larger charities. They are able to realise the importance and take it to their board to get the whole organisation thinking about it, which is huge and does not always happen at the end of the DataDive.

What we try to do at DataKind is be a catalyst, inspire and show people what to do with their data, but it is up to them to go on and change based on it. We think of it as organisational behaviour change. We want non-profits to change the way they make decisions. We want them to be more data driven and we want them to have a greater impact based on their data.

The second case study, is from a few months ago when we worked with Centrepoint who are a youth homelessness charity in London. They work with 8,000 young people a year. They work in Yorkshire, London and the North East. They have recently established the youth homelessness data bank which has a team of three people who are trying to pull together data on the state of the youth homelessness sector. They came to a DataDive with a very simple question. How many young people are facing homelessness in the UK?

This is quite a key question to the homelessness sector. It is hard to work out what services are needed and where if you do not know the scale of the problem. It is also hard to advocate on an issue if you don't know the scale of the problem. Centrepoint had put in freedom of information requests to every local authority in England, Scotland and Wales. Half of them replied and they replied in the most horrible, ugly formats you can imagine. Some were PDF's, some were loose data in an email. Centrepoint scraped all this data together and they came to the DataDive and they had this idea that if we can tell what we do know - we have data for some local authorities - can we then estimate what it would be for other local authorities with similar characteristics?

Using a lot of open data they pulled together the general characteristics of different local authorities. They had a massive group of 25 data science volunteers on the project for the weekend. They broke into sub teams so one was responsible for cleaning data, one for trying to work out what the model would be using dummy data, the other was pulling in all the contextual open data. They got to work. And the data was so much dirtier than expected. It took ten hours for the data to be in a usable format, but amazingly it all came together. By the end of the weekend they had a clean dataset which, to be fair, for Centrepoint was amazing. This is Jessie who is the only analyst at Centrepoint and it would have taken two months of her job just pulling this data together. So Jessie was over the moon, she had a clean dataset to work with.

But more importantly they were able to start mapping it. Here is the data from the local authorities that we knew on the left, and on the right are the estimates. What this really meant was that officially there are 15,000 young people who turn up to local authorities at

risk of homelessness. At the DataDive we found that the findings were more like 95,000. A team of six volunteers then carried on working with Centrepoint for three weeks after the DataDive and they found that the figures were closer to150,000. So this is big! Big not just for Centrepoint but also other youth homelessness charities .In September, when Centrepoint produced a report, they were able to say this is how bad things are and this is the kind of data we need. This is the structure and format we need local authorities to be producing data in. They immediately were able to use it for some advocacy work but they have also taken that all away and they are continuing work on the youth homelessness databank as a long term project. They are sharing that data with other youth homelessness organisations so they can try and co-ordinate and work out where they should be providing services. That is a really nice particular project where the right data skills at the right time can transform it.
We have some tips that we have learnt while trying to do this kind of thing with charities in the past year and a half:

- Buy in is really important - it sounds easy, it really is not. You need the Chief Executive to be on board to give staff time to work on this kind of thing, but you also need the person who is doing all the data collection and the data cleaning to make sure they appreciate why sound data collection is important and that entering clean data can help a project;
- You need to know what questions you can ask, and this is the fundamental point where we start with charities. What data do you have? What do you want to know? Understanding what kinds of questions you can ask and what other data should you collect.
- Using it for more than impact assessment. I would say most charities at the moment have a fundraising team that might have some data analysis skills. They may do some impact assessment using data wrangling skills. But data can be used for so much more. It can be used for understanding need, trying to look at strategy longer term, thinking about continuous improvement.
- Talk about your data. When you are having meetings and are making decisions ask, what does the data say? Know where to bring in data expertise. At DataKind we do not think every charity needs a data scientist. Most likely they probably just need to know when to bring someone in who does have the skills. They need to know what type of questions they can ask of that person.
- Giving staff and stakeholders access to the data is really important. We did a big project with Citizens Advice Bureau (CAB) and part of that project tried to democratise data and tried to open the data up to the staff. The staff have really smart questions and hunches and thought "oh this policy has changed here, we want to know how it is effecting people who are asking for help in our bureau around the country". We created a prototype dashboard so a lot of the staff could play around and explore the data which made their life a lot easier.
- Let your data person dream. If you have a data person, do not hide them in the corner and not include them in anything!

Questions: Louise - can I just ask about financing, resourcing something like a DataDive. How does that work because you need to hire space and get people, what are the logistics of it?

A: A lot of companies like to give us their offices for free so that helps! But we get sponsors, so sometimes it is the data companies that sponsor us. Other times DataDives are funded by for example, Which, the consumer group and the Young Foundation funded one. Some of the bigger charities want to investigate issues or fund a DataDive because it is a really interesting way to get lots of data science people, who you probably wouldn’t otherwise be able to afford looking at your questions and your problems.

Q: Louise - so do you do quite a lot of work in reaching out to those communities or those possible sponsors or, do they come to you? Are you like a magnet now?

A: Yes, I do not do anything; they seem to come to me!

Q: Louise - and just one note also in terms of the networking and how you get your data scientists, how you keep them together, getting them on board, is it social media that is keeping them together?

A: Again they come to me. We do not have any shortage of interest. That is partly because DataKind is a really strong brand. Jake Porway did his own TV show on the National Geographic channel and he is well known. A lot of data scientists go to conferences and come across him and they realise that there is a local London group and then come to me. There is some turnover, the same as with other volunteers, so sometimes they drop out of a project and you have to replace them and a lot of volunteer management goes on.

Q: You must be inundated with requests from small charities to help them. How do you select which ones?

A: We are not. Partly because we have no publicity because we are really small and we do not have capacity. We are slightly terrified that the charity sector are going to discover us at some point! We have about one request a week, half of those are not for data science, half of those are much lower level requests so a lot of it is just a bit of data therapy. I get on the phone and talk to them about what they are trying to do, at what point are they ready to come back to us, at what point we should put them in touch with some analyst help. So of those that are ready, then we look at whether they have the organisational buy in because that is the key thing for us. Does that charity have the time to engage with us? They are not always at the right place to be taking on a project with us. The longer term projects require some funding, so we jointly fundraise together. Once you have gone through all that it actually ends up with us not being overwhelmed. We have the right number at the moment.

Q: You said that one of the findings from one of the projects was to give access to data inside, what about outside, is there anybody coming that you help that can share data somehow and do you advise or recommend on those issues?

A: Yes. The project with CAB we had them sharing their data with St Mungo's Broadway, a big homelessness charity. They linked their data on an individual basis, but that is rare. That happened because we had some funding and we were able to spend time talking to them and their legal teams. It is not common that people come to us asking that. With CAB we did have a surreptitious aim of working with them which was to open up their data. We encourage charities to open data, but most of them do not even know what is in their own data. I feel that they are a step below being ready to open data up

Q: But you do address it? You do enter in the discussion that maybe they think about giving it to a wider audience?

A: As part of DataDives we take them through the process. The data has to be anonymised, non-disclosure agreements (NDAs) have to be signed, they open it up to a roomful of 80 volunteers, we see it as baby steps really. If they have done that once, then maybe they would be more open to doing it in a different context, but most charities are just terrified of the Data Protection Act and getting fined by the Charity Information Commissioner so will not do it.

Q: A lot of charities are not prone to storing their data electronically and I wonder if there is anything you can do to work with charities that are primarily paper based?

A: We like to work with charities that already have digital data because it is probably not the best use of our data scientist’s time to be using their maths PhDs to deal with data on paper. We do advise on how to get hold of data e.g. do a survey, or scrape a website. But we try not to focus on that because we think there are others better placed to help at that slightly different end of the spectrum.

Q: More of an observation rather than a question; in top down ownership/top down buy in, it is a problem that nearly all organisations, public, private and third sector struggle with but I wonder if there is some merit in seeing everything we are talking about today and wrestling with as part of knowledge management, if we actually re-articulate it as a knowledge management strategy. People who are turned off or at least glaze over when you talk about data and IT might be more open to the idea of putting in the resources and the times and include it in the strategies of the organisation as a whole?

Session 3: Ethical frameworks and governance

Libby Bishop: Introduction

Presentation slides at: https://www.ukdataservice.ac.uk/media/604168/hrdw_bishop29102015.pdf

Firstly my thanks for your dedication and listening. I am going to do a bit of an overview of what we are doing around data sharing at the UK Data Service and point you to more resources and places to go for help. We will then hear from Jim about the challenges of the Data Protection Act (DPA) and other issues, and then we are going to discuss some of the ethics cases that many of you sent in.

So let me get started here on ethics issues in human rights data. I titled the PowerPoint quite carefully. It does not say ethics answers in using human rights data and sadly I am afraid we may have more questions than answers. This is new territory for us to move into and we do not want to presume things that we do not really know. But the interaction that Tracey and I had that she alluded to earlier, I thought was really quite revealing about what she thinks she gained from talking to the UK Data Service, and certainly what we got from working with her.

After we spoke, she ended up organising a meeting around consent and anonymisation issues with representatives from the Information Commissioner's Office (ICO). One of the things that was beneficial for her was finding out that the problem [consent for sharing data with third parties] was familiar and lots of other people were having that problem too. This is something that we find often with ethics issues, that people are struggling, and feeling like they are the only ones grappling with this issue, and it can be informative simply to find out that these are known challenges being studied. Also the interaction helped her to formulate some sharper questions to pose to the ICO about precisely what she was trying to do, and what she wanted to be able to do, where there were barriers and obstacles and so forth.

Probably one of the reasons we have gone forward with this workshop is that we found many people, like Tracey, had heard of the UK Data Service but had not thought of us as a resource for this kind of information, and I think part of this may be a rebranding exercise for the UK Data Service in that we would like to be known as a resource for these kinds of issues and questions. One of the things that we do well is to act as a negotiator, go between, broker, matchmaker even, between people, for example, who deposit data and people who want to use data, and we bring them together. We can get access to data sources that people are having difficulty getting hold of and that is something I think we might be able to build on and develop in terms of going forward with the Civil Society sector.

Let me start out with the feedback from the ethics cases you sent to me. There are indeed a lot of challenges and ethics dilemmas: people see huge potential benefits from data sharing, using data in richer and more innovative ways and, yet inevitably there is risk. Sometimes those risks are small and the problem is that they might be over exaggerated in some cases. Other times the risks are severe, even violence, torture, or death and those clearly need mitigation and extreme measures for protection. There are lots of ways the data can help you be more effective. They can help internally with the efficiency, they can help with your campaigning, they can help with reporting requirements but there are real risks, loss of trust, breaches of confidentially, and potential for harm or violence.

You described two specific areas of tension. One is that there are tensions between "objective" data, and the culture of scientific research that calls for objectivity, and with data for advocacy. These uses may not be incompatible, but they do need to be negotiated carefully. Similarly a number of you raised issues with representation - representation in data, representation of the individual people, your clients, and the people you are trying to help. Representations can be very valuable for making connections with the public and with funders, but obviously these representations can be inaccurate or they can be misused by others. Those are some of the big dilemmas that you are grappling with.

In addition, a number of you raised specific challenges of anonymisation: does the Data Protection Act (DPA) even apply, does it matter to me if I anonymise data, do I need to anonymise, how do I go about doing that, and which tools to use. Some of you are using in-house tools; some of you are building your own. Similarly for informed consent, there is a lot of uncertainty around this. Is it even required? If it is not required legally, may it still be required ethically and how to get it if indeed it is required? And you described specific challenges around consent with children, images, facial images, those lacking capacity, such as people with dementia, people with learning disabilities and so forth. There is a lot of commonality with people raising similar sorts of issues but a fair bit of differentiation as well.

Next I will do an overview of where the UKDS plays in this space, around giving people advice on data management, on consent, anonymisation and so forth. We work with a couple of different frameworks; the main one is called 'Five Safes': Safe People, Safe Projects, Safe Settings, Safe Outputs and Safe Data. So there is a detailed brochure coming out shortly. But the idea and the main point to get across is there is no one of those "Safes" that is sufficient in terms of protecting the kinds of data that we hold, and in many cases we need combinations of training people, vetting projects, checking the statistical outputs of projects before they can be used in publication, and controlling access to data. The point is that it is a broader framework that needs to be in place in order to accomplish our mission which is to 'enable safe, ethical and legal sharing of data'. Even personal and sensitive data which would be deemed that way under the DPA can be shared.

The other area I will point you to is the 'Manage Data' area of the website, this is more or less the topic list for data therapy, that preparation you might need before you are ready for data scientists. This is exactly the kind of work that we do with researchers now so the whole point of our data management is to work with researchers as they are creating data to make that data better, in terms of its own reuse capability for the individual researcher, but also to allow that data potentially to be reused by others, to be archived and shared down the line. That list of topics is everything from writing data management plans to considering the legal and ethical issues of sharing, to working with copyright issues and documentation, metadata, formats, storage, how to do collaborative research, and things like tools for encryption. We run these courses in different formats, from a couple of hour's session to two-day workshops and we customise it to specific audiences. We go through topics like informed consent, how do you write a combination of an information sheet and a consent form that takes into account use of the data after the research project itself is completed. How do you word that? What kind of wording do you need to avoid? How do you write consent form that meets the requirements of the data protection laws and Jim is going to be speaking in more detail about data protection.

There are key principles of the DPA: being clear about the purpose of research, what is involved in participation, benefits and risks, ways of withdrawing, how the data are going to be used initially and for secondary uses and strategies to ensure confidentiality and so forth. Most researchers are fairly comfortable with writing these kinds of consent forms except for that last part about paying attention to onward sharing of data and that is where they need some help in understanding what that actually means and what can they write in the forms. We try to coach people about being clear about the risks and dangers, at the same time not putting excessive warnings and being absurdly risk adverse to a point where you deter anyone from participating in your research project, which of course is a real risk. We have actual sample forms of consent and they are here: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/consent-forms

Similarly, for anonymising data again these are just a few of the points that we provide for anonymising qualitative data. We have oodles of information on these techniques on our website and in training materials, and it is actually all in the book: Managing and Sharing Research Data; A Guide to Good Practice, Sage Publications and you should pick up your copy. Key points are things like planning ahead for anonymisation and how to avoid over anonymisation. I think, increasingly the community is moving away from over-reliance on anonymisation towards the idea that excessive wiping out of content from textual data (in particular), reduces the value and quality of the data. If you reduce quality to the point where you are making it unusable for the immediate project or for future projects, then you should look at other strategies in terms of restricting access to that data rather than anonymising it to the point where it loses value. We have examples of projects with what we see as difficult data, and show how it was actually anonymised, why perhaps less anonymisation was needed than you might have guessed and detailed explanations for that.

Where you are trying to avoid over anonymisation, the other 'third leg in the stool' that we talk about is the ability to control who can see the data, that is, regulating access to it in various forms and this is an area where we have gotten quite sophisticated at the Service. We have got three different levels of access: Open, Safeguarded, and Controlled, and they are regulated according to who can see it, the different types of agreements that people sign up to. I recognise that not everyone has got the UK Data Service sitting behind you, but in fact through our ReShare system, which is now open to anyone, there are three similar levels of access: Open, Safeguarded (meaning available to registered users only) and the final category is Permission only, where only people to whom you give permission can see the data. There is a growing sophistication of the infrastructure for data sharing and that is something that I think is important for people to know about.

People come back and say "but this data is too sensitive it cannot be shared it is too compromising, it's too threatening", but there are many archives and repositories around the world that hold *even* this kind of data. In Northern Ireland there is an archive of interviews of people involved in the 'Troubles', and yes, they actually tell stories of murders, plots and so forth. Some of these datum may have special circumstances, people who have put it in may be long dead or there are conditions under which this material is now available, but for almost any conflict situation you can think of: truth and reconciliation commissions, the Witness to Guantanamo project, archives of interviews with the Balkan reconciliation stories, and more, so there is a growing number of examples of archives out there handling this sort of material. None of it is easy, all of it usually requires special conditions of some kind, but in my area of qualitative data, that is the story of our life. If you work in qualitative data its always unstructured, it is always messy, it always needs special conditions of some kind and this is an extension into the realms of human rights data as well.

Let us now turn to what is on the forefront in terms of consent? Everyone knows that the consent model, if you will, is challenged. Everyone is well aware that consent conditions in things like the websites that we visit is utterly pointless, no one is reading the terms and conditions for this kind of data use. If you write terms and conditions that are complicated enough to explain what you are really doing with data, they are 40 pages long and no one can read them and so there is a bigger issue about consent. People are moving forward with trying to think of different ways of doing this. One model is dynamic consent, where you do not explain things all at once, you do not even attempt to put it all in a one-time conversation, you think of consent in a dynamic continuous model with the need to go back again and again to the people you are gathering data from and seek reconsent for continuing to use the material or for substantially different uses of the material. A similar and related kind of idea is happening with the themes of broader and enduring consent. This is what is happening in a lot of the genomics and biobanks and you can imagine how on earth do you promise what future uses might be? We have absolutely no idea if you had asked five years ago even what sort of uses would gene data have, you would have been wrong. In some ways it is unethical to promise people that you know what secondary uses of data might be especially in some very fast moving areas of science, so instead they are saying: give us your gene data, we have no idea how it may be used, and are you OK with that? Now we can debate whether that is really informed and do we want people giving consent in those sorts of conditions, and that is a reasonable debate, but this is the direction that some things are moving with some of the biobanks and genetic area.

As people have already pointed out, significant and important national and international differences exist, so the situation in Germany is much more strict than is the UK and for those of you who have followed data protection privacy issues, the 'Safe Harbour' ruling and disputes between the EU and the

US are very volatile as well. In addition, I think that a lot of people are not aware of the fact, that at least for administrative data (that is data collected by governments for routine operation) informed consent is actually not required. The DPA has what is called "alternative basis" for organisations that need to use personal data for their ongoing operation that do not require formal consent, which is why you do not give formal consent for things like your tax forms.

In other cases where consent is not possible but review is deemed a good idea, such as areas in medical research and in some of the large surveys like Understanding Society, they use access review committees, the idea being that there is some kind of representative panel that decides who can have access to the data. Now is that a perfect model for consent? Well it may not be the actual people whose data are included in the dataset but it is some kind of attempt to represent to the group and some attempt at getting stakeholder participation on the issue of data sharing. That's something that's going on with a number of the large surveys and the Administrative Data Research Network (ADRN) here at the UKDS has been quite active in seeking out a great deal of participation, starting with a long consultation project before that service was set up and later including citizen panels as a way to get direct participation in decisions about projects and research.

So the challenges are tough, there is not always enough information out there and in some cases there is too much and it is hard to find and it is hard to know what the good stuff is. So we will pull out these things on our resource list, there are things like guidance for how do you do a Privacy Impact Assessment (PIA). A PIA is a bit like an ethics review procedure, it is a checklist, a series of questions on what to think about around privacy implications. Is it a bit tedious? Yes. Will it force you to think through some issues which might not have crossed your plate? Yes. It will do that too and it is something that the ICO recommends. The ICO has a data sharing code of practice and they have an area that's specific for charities and third sector. There is some guidance on anonymisation coming out of the UK Anonymisation Network.

Now this is a little bit of a quick aside. I gave a different talk on this and my theme was ethics is complicated and messy and there is no app for that. We cannot actually solve this with technology or a little app. Well guess what. There is an app for that! An app for ethical decision making. I was entirely sceptical and thought oh my god the worst just happened, what have they done. Well this is actually an incredibly well respected applied ethics centre at Santa Clara University, called the Markkula Center for Applied Ethics. I am not saying the app is flawless but they go through a quick introduction to the main themes in ethics around utilitarian, virtue , and rights approach, how to reconcile them when your ethics framework leads you to contradictory conclusions. It is a useful little tool so I would never say there is the answer but there is an app for that!

This is, however, I think the most valuable thing I have read about ethics and data in the last year or two. This came out of the Nuffield Council on Bioethics and these are the concluding recommendations on collecting data in bio medical and health areas. These are the guidelines for what are the 'morally responsible expectations' about governance and use of data, that is, the four principals that we should try and follow:

1. Respect for persons;
2. Respect for established human rights;
3. Participation of those with morally relevant interests and
4. Accounting for decisions.

These principles bring out the ideas of transparency and accountability and we could call these "mom and apple pie" but nonetheless, when you look at what they are trying to do, I think they are very potent and powerful guidelines in terms of how to move forward with data sharing. So we have got lots of data, no doubt about it. We love it, we think it is 'cool stuff' and it can potentially answer a lot of questions, but we also need to keep those principals in mind, and we also cannot forget that there are people behind every single one of those data points. If we can do that, I think it will help us to make more ethical decisions. So I shall hand over to Jim at this point who will give you an example of how he is moving forward in his area.

Session 3: Ethical frameworks and governance

Jim Vine: Using a trusted intermediary to create insights from shared datasets in a collaborative sector

Presentation slides at: https://www.ukdataservice.ac.uk/media/604182/hrdw_vine29102015.pdf

I'm Jim Vine from HACT, (the Housing Associations' Charitable Trust). HACT works for housing associations and the wider social housing sector in delivering innovative projects; our activity is quite broadly drawn but includes a focus on data projects and a wider research portfolio in support of housing associations.

HACT is a charitable organisation and are also largely a social enterprise, by which I mean that most of our income is traded income with the housing associations on a project-by-project basis. We go out to the sector and say 'is this something that you are interested in and if so, will you fund it? Will it deliver something of value to you?'

Our principal sector is housing associations but we also engage with the wider social housing sector, including ALMOs (arm's-length management organisations) and council housing departments. Together they house 17-18% of households in the country and they are mostly sub-market rental homes, but they have a range of other housing types around that.

Community Insight (www.communityinsight.org)

I am mostly going to talk about some of the legal issues that we've encountered and our approaches to addressing those in our big data and data science projects. Before that, I thought I would start by putting this in the context of a different system that we've worked on, Community Insight, in order to draw upon another example of where we have done something to create this benefit for organisations of working with the trusted intermediary. The big advantage of Community Insight has been that as a sector you can develop something once and then deploy it many times. The development costs that went into Community Insight to get it up and running were well into the tens of thousands of pounds and over the 3 or so years it has been running it must be well into the hundreds of thousands of pounds to add on new features, new data, and to keep on top of everything.

What Community Insight does for the housing sector (and other public and third sector organisations) is give easier access to data. You can define an area by clicking around it on a map and quickly generate reports that access a lot of open data, giving you a detailed neighbourhood profile. That's delivered as a relatively low cost subscription service to organisations. The advantage of the approach is that no one organisation would have wanted to build a system like Community Insight for itself; it would have been prohibitively expensive. But splitting the development across 80, 90, 100 housing associations it makes a lot of sense.

These data science projects are where we get into the ethical and legal issues. With Community Insight we are dealing with open data that the government puts out, so I am comfortable working on the assumption that the government has done the job of anonymisation well enough in the first place. Certainly if it is data that has been released on an aggregated area basis that will almost certainly be the case. I suppose if you are using an open dataset and you discover that you think that you see cases in the data that seem to be a little too identifiable, the ethical course of action would probably be to contact the data provider and say 'hang on I've spotted something here that means I can tell this data relates to Mrs B

I used to call our work in this area 'big data' but now tend to call it 'data science'. By this I mean using relatively advanced analytic techniques to try and generate novel insight for the housing providers that we are working with. But generally speaking it's where you are dealing with your own data and doing

So what I am going to talk about, while the slides cover the technical details from the ICO, is really a description of the process that HACT has gone through. Although we discussed new things that ethics becomes more of an issue.

I should begin by stressing that I am not a lawyer and this not a formal statement of law. As far as possible I have pulled out quotes on a lot of the slides that are from the website of the Information Commissioner's Office (ICO) and linked to various documents there, so you will be able to refer to those for a more official expression of the regulator's view of the law. The rest of the material you should treat as the opinion of someone who has been through the process of setting up projects in this area, rather than anything definitive. The slides contain the quotes and links to where you can find them on the ICO website for ease of reference.

it with the ICO I should say that we found that it is not sort of regulator that was able to give definitive confirmation that our proposed approach was acceptable. We did find the ICO very approachable and very willing to give input and make suggestions, but at the end of our discussions the response to asking whether the approach was OK was more like an absence of a 'no' rather than a definitive 'yes'. We shared our plan with them and the response was effectively 'I don't see anything wrong with that'. We felt that we had done all we could and we feel we have a responsible approach in place and the ICO had not said it was a bad approach and we were happy to proceed on that basis. Different organisations will take different views on that and it is worth remembering that the ICO does have some pretty serious enforcement powers. I believe it has taken enforcement action against housing associations in the past and is not averse to using its powers in respect of third sector organisations if it identifies a data breach. Do take care, and do make contact with the ICO. But this is the approach that we have taken that we think is OK.

Although this is a session on ethics, I am mostly going to be talking about legal compliance and the framework we have adopted to meet out legal requirements. My justification for doing so is that I think that being legally compliant in this area actually takes you quite a long way to having an ethical approach. The main legislation that we are talking about is the Data Protection Act (DPA). Compliance with legislation is obviously a minimum requirement for running a project, but because the law is quite well developed in this area to protect data subjects' interests, a legal project will naturally avoid many of the worst ethical risks. If you are thinking through the things the Data Protection Act requires of you it will be raising many of the most important ethical questions for data handling. In some areas you may wish to simply meet your legal obligations, or in others you might feel that, once you start thinking about it, you want to go a bit further.

Part of my rationale for saying that legal compliance supports ethical practice is because the law is quite widely drawn. It covers any information that's held on a computer or is intended to be held on a computer, so it is pretty broadly drawn in terms of the data that falls under the remit of the DPA. Also what constitutes personal data under the DPA is widely drawn: any data that relates to an identifiable living human being (the slides cover the detailed description of that breadth of coverage). Again the ICO has further guidance on this if you are in doubt but my summary would be that it is broadly drawn; if you have data on individual people you should probably work on the assumption that it is likely to fall within the DPA.

I would also raise a note of caution here around anonymisation. Whether the information constitutes being an identifiable person or not is not just whether there are names in the data. It is also about whether you could 'triangulate' back and find out who that person is. Doing proper anonymisation that ensures that people cannot be re-identified from the anonymised dataset is a job in itself. It is not one we've had to deal with so I am not an expert on how you would do it, but if you need to do that then I would recommend you seek expert support on that.

There are extra conditions apply where you are talking about sensitive personal data (see the slide for details). There is a set categories that constitute sensitive personal data and there are even more safeguards in the law around that. Our view has been that there is quite a lot of analysis we can do without needing to use sensitive personal data so we are building experience of doing analysis with the housing sector's data and just not touching these more complex areas for the time being. If your analysis requires the use of sensitive personal data then you will need to be aware of the greater protections in place and ensure that you comply with those additional requirements.

The definition of 'processing' is another area that is broadly drawn. It includes holding information and carrying out any operation on that information. Pretty much any analysis that could be performed will fall somewhere within one of the list of things considered to be processing for the purposes of the DPA.

The DPA specifies different roles of data subjects, data controllers and data processors. See the slide for definitions and links to the guidance on this matter. Briefly, the data controller is the organisation responsible for the data and the data processor is some other organisation that is doing processing on behalf of the data controller. The ICO website has a guide on this, examining the differences in detail, which runs to about 20 pages; if you think this is a grey area in your case then you will want to refer to that.

Collaborative sector

I have described in this context HACT working in a collaborative sector. I have chosen that phrase to capture a few things. The first thing is that I am avoiding saying 'non-profit' or something like that because lot of housing associations that we work with view themselves as 'profit-for-purpose' organisations or similar. Also, social housing is a sector that crosses both the non-state (mostly charitable) sector and the public sector (i.e., including both housing associations and local authorities).

Most importantly, the description of this being work in a collaborative sector speaks to why this type of project can work: whilst there are certainly areas where housing associations are in competition with each other for grant or access to some resources, it is a collaborative sector in the sense that for a lot of their services there is so much excess need that no organisation could meet on its own. They are not generally in competition for tenants, and anything they can do to achieve better outcomes for their tenants and communities is for the good of all so they can collaborate and achieve more together. I suspect that applies to other charitable sectors. There will be pieces where you feel you are in competition (and that's fine) but in other parts of your activity where you can bring your data together to have impact then you can collaborate.

Creating insights

I mentioned that I'm talking here about data science. Briefly, what this mostly comprises for us is building models to identify relationships including predictive models. The strength in this context in bringing multiple organisations' data together is that it can help to detect relationships that one organisation on its own would not be able to detect from its smaller dataset.

When we start considering creating insights, one of the things we have to do to ensure the analysis is legal is establish the conditions for processing. This is one of the conditions that has to be met under the DPA for processing to be legal. The main grounds for processing that are most likely to apply are probably the subject having given consent to the processing or the one that we have tended to use, which is the 'legitimate interests' condition.

If you rely on the legitimate interests condition to do your analysis you have to demonstrate that the processing is needed for the legitimate interest of the data controller, i.e. the organisation responsible for the data. This has to be balanced against the interests of the data subjects, i.e. the people you hold data on, but that is not an absolute prohibition on doing things that are not in data subjects' interests. If there is a serious mismatch between competing interests, the data subjects' interests will trump the interests of the data controller, but, say, something is in the interests of the data controller and neutral for data subjects, you do not need to prove that it will be beneficial for the data subject.

You definitely have to weigh that up the interests on each side, but we have found that for a lot of our intended analyses we ask is this in the interests of the housing association if we do this analysis and generate this insight? Would the housing association be able to do something useful with this? If the answer to that is 'yes', we ask what it means for the tenants and when we have gone through what housing associations might do with the insight we find that it is likely to be beneficial to tenants too.

For example, if we are looking to produce a model on rent arrears and identify who is at an elevated risk of falling into rent arrears in order to offer some support to them before they fall into arrears.
The implications for the tenants seem likely to either be neutral if we do not come up with the model, or neutral if they are not one of the people who gets identified as being at risk, or positive if they get proactive support, since that might mean that they do not go on to fall into rent arrears and avoid having that stress and heartache and debt hanging around their shoulders. So we often find that those interests are well enough aligned or neutral to the data subjects, but you do have to run that check and establish that there is not that serious mismatch, making some analysis detrimental to the data subjects.

Besides consent and legitimate interests there are some more conditions for processing that are detailed in the slides. There are partial exemptions (with limitations) for processing that is conducted for research, statistical or historical purposes. Conversely, as I mentioned earlier, with sensitive personal information you have to do more. Processing those data needs extra conditions on top, which you can find described in detail on the ICO website.

Shared datasets

The processing we are doing is bringing together data from multiple organisations. What we are doing is not 'shared' in the sense of 'data sharing agreements' that you might have come across. A data sharing agreement is required where you have two data controllers and they are passing data between each other for some purpose.

In our case, HACT is acting as a data processor and we are taking data separately from each of those housing associations. We get sight of all of it but they do not get sight of each other's data. That is the model that we applied. You could presumably work out a process for having multiple data sharing agreements between different organisations but that is not one that we have done.

Trusted intermediary

We have acted as a trusted intermediary and the trust is very important in this process. For the data controller it is particularly important to underpin that trust with checking out that the processes and approaches are in place and assure yourself that the organisation has the procedures you need.

We act as a data processor. The big thing for the data controllers in this relationship is that data processors do not legally speaking have their necks on the line for the ICO. If something goes wrong, it is the data controller that the ICO has regulatory power over. However, the reputational responsibility will be more widely distributed. By which I mean that one of the things that we have said, and that has been quite compelling to housing associations, is that we have a reputation in the housing sector that we are keen to protect. If we were to mistreat their data in some way we may not be legally liable, but our reputation would be harmed. We would expect to get a lot of negative coverage in the trade press, and potentially in the mainstream media as well. Obviously the confidence they gain from our reputational interests is only a backup to their being able to check our technical processes and legal framework, and having their techies and lawyers sign off that those processes are all appropriate. But those in governance positions do seem to find it useful to see that we have a strong interest of our own in keeping the project on the right side of the law, even if any legal enforcement action would not be taken against us.

The benefit of this type of arrangement is that the data processor is acting in place of the data controller. The data controller needs to establish the grounds for processing, and once that is established the data processor can go away and conduct the analysis.

Investigation documents

This is a structure that we developed as part of that process to support that governance function and to support the housing associations in terms of giving us the authority to do work. We produce with them an investigation document that spells out what form the analysis will take and they will sign it off. Formally speaking, data controllers are the ones saying ‘please go and do this analysis for us’ but of course we support that process, drawing on our understanding of both the operational issues that the insight might address and the technical possibilities for analysis.

Analysis plans

On our newer project we are moving towards developing analysis plans. These are intended to support the robustness of the analysis we undertake. I mention it here because I think robustness is an ethical issue. Doing analysis that is as reliable as you can make it if you are going to be using it to inform decisions then you want those decisions to be based on the most robust insight you can generate. You do not want decisions to be based on fishing trips or data dredging. So we are increasingly drawing up - in advance of touching the data - a plan that says what we are going to do, and how we are going to do the analysis. In the future we are planning to lodge these analysis plans in a publicly accessible repository so that people can see, once we have conducted the analysis at the end, that we have done what we planned. If we do exactly what we plan and publish in advance then that will allow us to demonstrably show that we are not going on fishing trips and that our conclusions are more robust as a result.

The main reason why you should not go on data fishing trips is because correlation is not causation. The slide shows a chart with a correlation between crude oil imports from Norway and drivers killed in collision with a railway train from a website that has lots of these spurious correlations. Spending a short while with these helps to remind you why not to go fishing for correlations, because if you have enough data you are always bound to find a relationship somewhere in there, but if you have not gone about looking for it in a structured fashion, the relationships that you find probably mean nothing.

Session 4: Exploring opportunities for using third party data sources to provide context

Louise Corti: Introduction

So this last session before we sum up and have our surgeries is really about what external data sources you can use to complement your insights and findings to supplement your own information. We have a number of speakers who are talking about the things that they are working on, from social media to some of the sources we have at the Archive; and some tools that are also being used in qualitative research. The idea is that we really want to hear from you about what kind of data sources you think you can use, or you would like to use; and to see if we can help in any way in collating them all or trying to make them more available. So we are going to start off with Hersh from the UK Data Service who is going to tell you how to get hold of data we have and how you can request us to go and seek data through our brokering role. We do not always get data on request, but we do try!

Hersh Mann: Sourcing 'society' data from the UK Data Service and beyond

Presentation slides a: https://www.ukdataservice.ac.uk/media/604178/hrdw_mann29102015.pdf

Thanks Louise. I will just tell you a little bit about who I am and what I do. My name is Hersh and I work at the UK Data Archive (UKDA) for the UK Data Service where I focus on User Support and Training. This means that we help anyone who comes to us to access data and to answer their questions. For example, if they have any queries about data they are already using, or if they need advice on what kinds of data sources are available, what they can find, what kind of variables there are, why documentation might be missing for a particular study or anything like that. Those questions come to us, and my team will help you resolve these difficulties so that you can use your data more effectively. Before I get into this, some people in the room will already be registered users of the UK Data Service. Put your hands up if you are already using data. OK, there are a few of you, that is good. A lot of hands did not go up, so is that because we do not have anything that is useful for you or because you have never heard of us before? Who has never heard of us before? If nothing else you will leave this session informed about the Service and what we can do for you because there are thousands of data collections that we have available, and contrary to popular belief we are not only here for UK Universities. We have users from many different sectors from all over the world and the data are free to use for academic purposes. So this is a very very quick overview of what I am going to talk about. I will give you an introduction to what the UK Data Service is and explain what kinds of data we have and how you can go about accessing these materials, how you can go about finding things and what kinds of things you can do to help us facilitate access to data from other sources as well.

The UK Data Service is funded by the ESRC, it is your 'one stop shop' for a wide range of different data sources, that I will go through in a little while. As well as facilitating access to data we also provide support in other kinds of ways if you are using data. We provide training and my team will go to universities to speak to academics, to students, to librarians, explaining what kinds of resources we have and how we can help you perform your research. We also provide video guides, tutorials, all kinds of downloadable booklets and related materials which span all of our different data collections, whether you are using UK Data, international macro data, census materials or qualitative data. So who is it for? It is for absolutely anyone. The vast majority of data collections are available to anyone who registers with us and whilst there might be some licensing restrictions for some data collections, by registering with us, you could be using our End User Licence data within five minutes of identifying a source of interest. That is our new home page. Just launched this week, so do go and have a look and explore that. At the top you have got some menu items and if you use those you will be able to navigate around the site very easily. We categorise our data collections under broad headings depending on, for instance, the survey design and their coverage.

For example, our survey microdata collections include both cross-sectional and longitudinal forms, but we also hold aggregate statistics as well. Our international sources of data would be of interest to you if you are looking for aggregate data. We also have aggregate and micro data collections relating to the UK Census from 1971 onwards and, finally, we hold a range of qualitative and mixed methods data. If you are on the web pages and you want to find your way around then I recommend you use our 'key data' pages.

At the top of all of our web pages there is menu item called 'use data' and under that is a link for 'key data' and this is an effective way of getting into the collection, because it points you to the most popular data collections and gives you a good overview of what is available. But they are not the only data collections that are there, I am going to talk about some of the smaller scale studies in our collection as well. These data come from a wide variety of sources and whilst the big large scale surveys come from central government, the Office for National Statistics (ONS), Department for Work and Pensions, HMRC, and places like that, we also receive data from the UN, OECD, and World Bank, to name just a few. Individual research institutions will also deposit data with us and if you have funding from the ESRC then you are obliged to offer any data that you have collected as part of that grant back to the UKDA and that requires that you have a very good research management plan. So if any of you ever do receive funding from the ESRC, as part of your application proposal you will need to create a data management plan and do whatever is necessary to make sure you have done the best you can to make sure these data are available to researchers at the end of the project. You will you need to have good consent agreements, make sure that your participants are 'fully informed' as to what kind of research project they are involved and ensure that they know their data will handled securely.

Those of you who are using survey and micro data will know that you will need to be analysing these data using software packages like SPSS and Stata but I do know that we have some R users in the room which is becoming increasingly popular, and the beauty of using survey micro data is that you can construct your own tables and do your own analyses. You are not limited to the kinds of tables that may already have been produced by someone else, and here are some of our major key UK surveys. I mention these just to give you a flavour of the kinds of data that come to us from the ONS and other agencies as well. There should be some surveys there that are familiar to you. The Labour Force Survey is conducted every quarter and this is where our unemployment statistics come from. We hold all of the health surveys for England, Wales and Scotland and the crime surveys from England and Wales are interesting ones because not all crime is reported so some researchers would argue that the crime survey gives you a much better picture as to what the crime situation is in the UK. I also know that some people in the room are probably already using the housing surveys and citizenship surveys so there is no need for me to dwell on those.

There are good reasons to be using those particular data collections not only because they are widely available and have long time series, but because you know that you are going to get very good quality data if you are using any one of those surveys. They have good designs, the methodology is sound, they have large samples and they are well-documented so if you do get access to any of these data collections you are going to have the information to actually use it in an informed way. They cover lots of topics which will be of interest to people in the room. I have listed a small selection here: health, work, crime, social attitudes, family expenditure, environmental behaviours, leisure, wealth and assets and lots more.

In addition to the UK microdata survey collections that we have we also have these longitudinal panel surveys, and again there should be some there that familiar to you. The top three that I have listed, the National Child Development Survey (NCDS), the 1970 British Cohort Study and the Millennium Cohort have involved children from birth, and researchers from the Centre for Longitudinal Studies go back to those children every few years to collect new data. The NCDS started in 1958 with 170,000 children born in a particular week and the researchers go back to these same people, who are now approaching their 60s. The 1970 Cohort and Millennium Cohort are continuations of that particular survey model.

The other two surveys that I am going to highlight are the British Household Panel Survey (BHPS) and Understanding Society or UK Household Longitudinal Study (UKHLS), both of which are conducted by our neighbours at Institute for Social and Economic Research (ISER) here at the University of Essex. The BHPS started in 1991 and it followed 5,500 households and they collected data over 18 years to cover a really wide range of topics including: income, labour markets, household composition, household expenditure, housing conditions, health, education and that survey continues in a slightly different form. The original BHPS has now merged into Understanding Society which is an even more ambitious project. The researchers are expanding that one up to 40,000 households with 100,000 individuals and this is now the biggest household panel survey in the world. They have also expanded this survey to cover more topics including benefit payments, and most importantly, they have sought consent in this particularly survey to linking to health and education admin data. If you are interested in survey methodology the Understanding Society survey also incorporates an innovation panel. This is because it is such a big survey that they can afford to do experiments during their data collection, for example they have been investigating how different incentives affect response rates. What happens if the household members are told that they will only get their £15 for participating if every other person in the household also participates? How does that affect your response rate? Imagine being the one person in that household who did not participate and preventing everyone else from getting their £15! You might not be very popular.

As well as these big surveys we also have other types of data collections. I undertook a very quick search using some topics which came up yesterday that I know are going to be of interest to people in the room. I will not go through all of them. Obviously there are surveys of charities and organisations in the third sector but we also have some other collections from smaller scale projects. The items listed here at the bottom of the slide are all qualitative data collections and would be useful to some of you here.

In addition to the microdata that we have, we also have international macro data collections. These are regularly updated and then deposited with us by the organisations like the OECD, UN, the International Monetary Fund (IMF), the International Energy Agency, and the World Bank. The OECD, IMF and World Bank data are completely open access and there are more international macro data available to everyone than there ever has been. These data cover a very wide range of themes, such as, demography, governance, human development, social expenditure, education, land use, and science and technology. The attraction of using our international macro data is that you do not need any special software and you do not need any special statistical skills to generate these tables. Everything is done within your web browser using the UKDS.Stat platform so you can just chose the data collection that you are interested in and then start building your tables by selecting the variables that are of interest to you.

The other types of data that we hold are qualitative data collations, because when we talk about ‘data’ we are not only referring to numbers. Qualitative data and textual information typically derive from sources like interviews or field notes and you can search for those easily using our search tool ‘DISCOVER’. As an illustration I have listed some of our most popular qualitative sociology data collections on this slide.

If you would like to start searching for data you would use DISCOVER and you can do this via the search box on the home page. If you type in anything there you will come to a list of results with filters on the left-hand side. To illustrate, I did a search on ‘human rights’ and it produced 484 results. That is quite a lot to ‘sift through’ so on the left hand side you have various facets that you can use to target your search more precisely by data type, by subject area, by geographical variables, or by time coverage. I am not going to go through all of these, but using that human rights search results list, I came up with a few examples which might be of interest to some of you and this is just to give you a flavour of the diversity of the data collections that are available. Some of them have ‘human rights’ and ‘collective rights’ in the title but then you have got topics on ‘minorities’, ‘people are risk’, ‘workers’ rights’, and ‘refugees’. Just looking at this list, is there anything here that looks interesting and useful for your research? Yes? Great! Some of the nodding heads are people who did not know about us at the beginning, so I think we are making some good progress here. That is really good.

You have heard about all of the rich resources that are available. How do you get access? If you go online and look at any of our catalogue records you should always read those in full because it gives you the background as to what that data collection is, who collected it, how it was collected, what it is all about and the documentation will be freely available as well. If you identify a data collection of interest you will, in most cases, be able to download it after you register with us. If you are in UK Higher Education you will be able to log in and register with us using the username and password you have from your University. If you are not from a UK University you would first apply for some UKDA credentials. It is done very easily. You just complete a form online, we send your username and password, and then you can use that to log in and within 5 minutes of receiving your username and password you can be logged in, you can be registered and you can be downloading data of interest to you at your own desktop wherever you happen to be around the world. It is not difficult and people are amazed that they can do it so quickly. The only thing that we ask you to do is provide a little description as to what your research project is about. That is not because we are ‘checking up’ on you but that information is valuable to our funders and it also gives us an idea of what sectors people are coming from so that we know what kinds of training resources we might need to produce and what kinds of themes we might need to cover on our supporting pages.

I am now going to look at other sources of data and you do not need me to tell you what other sources of data are useful to you, because you know what you are interested in and you are experts in your field. I came up with the list of sources shown here because my background happens to be in political science so these are things that I knew of anyway. These are widely known, but there are many more and the key thing to do, if you know of a data collection somewhere, is to ask for it and come to us as well. We will try to negotiate access on your behalf if data are not already shared for research purposes. As you have been hearing over the course of yesterday and again today, data sharing is a good thing. What we need to do is to educate on good practice to facilitate data preservation and sharing, and promote re-use and citation. Everyone wants to make an impact and we can help demonstrate the impact of your work as well because we are interested to know what people are doing with data. If you have obtained data from the UK Data Service you should let us know what kinds of work you have been doing. If you have published something let us know about it and we can feature you in a case study. We have a list of over 160 case studies on our web pages now and more in the pipeline. If you provide a case study for us we could promote it through our communications channels and this would enhance the profile of your work. This slide shows a couple of examples I picked out. The first describes work performed by a policy institute in Washington DC and then we have a paper on housing done by the Strategic Society Centre. I highlight these because we would like to get more examples from outside the traditional academic sector.

The final thing I will talk about before closing this presentation is to mention the different ways of getting in touch with us. If you are new to our service then we have all kinds of online resources that will enable you to understand what data collections are available. We do hear sometimes from people that they are just overwhelmed when they come to our web pages, they look at what is available and do not know where to begin. To help with that we have a suite of web pages specifically designed for people who are new and there is guidance here on how you can conduct a search, how you can find relevant data and, most importantly, how you can get in touch with the Help Desk Team.

Session 4: Exploring opportunities for using third party data sources to provide context

Sian Oram & Mike Emberson: Mental health responses to human trafficking: qualitative data tools.

Published paper, Domoney et al. l(2015) 'Mental health service responses to human trafficking: a qualitative study of professionals' experiences of providing care', BMC Psychiatry, Vol 15, available at: http://www.biomedcentral.com/1471-244X/15/289

Session 4: Exploring opportunities for using third party data sources to provide context

Matt Williams & Luke Sloan: Gaining insights from social media data: Collection, analysis and interpretation

Presentation slides at: http://www.ukdataservice.ac.uk/media/604183/hrdw_williams29112015.pdf

Matt Williams

I am Matt Williams and I am here today with Luke Sloan. We are from the Social Data Science Lab. We work with two other colleagues, Dr Pete Burnap and Professor Omer Rana, from the School of Computer Science & Informatics at Cardiff University.

What we would like to do is give you a bit of context around using social media as data in the social sciences, then go through the particular issues by going through what we call the six V's: Volume, Velocity, Variety, Veracity, Virtue and Value. We will then move on to two case study examples: i) hate speech and social media, and ii) the discussion of Ebola on social media. In both we will show you how the COSMOS software that we have developed enables researchers to access and visualise the data in different ways.

In their pioneering article in *Science*, Lazer and colleagues argue that corporate giants such as Facebook, Google and Twitter have been using social data with advanced computing to mine and interpret it for half a decade. Until recently, Savage and Burrows argued academic social scientists have been left in an 'empirical crisis', lacking the access, infrastructure and skills to marshal these data. Ruppert et al. note how the disciplines of the social sciences face the challenge of how increasingly ubiquitous digital devices and the data they produce are reassembling their research methods apparatus. The exponential growth of social media uptake and the availability of vast amounts of information from these networks has created a fundamental methodological and technical challenge for social science. These challenges (and affordances) can be summarised as the 6 Vs: volume, variety, velocity, veracity, virtue and value.

Volume refers to the vast amount of socially relevant information uploaded on computer networks globally every second. Ninety percent of the world's data was created in the two years prior to 2013 (BIS 2013). This is partly due to the global adoption of social media over the past half a decade. Within the UK alone there are 15 million registered Twitter users posting on average 30 million tweets per day. Of these online social interactions, a sizable portion are relevant to social science research. A comparison with conventional curated and administrative data on crime shows how these new datasets far exceed the size of those that researchers have come to rely on to gain their scientific insight. The whole UK Data Archive currently holds between 2.2 and 15 terabytes of data. These sizes are dwarfed by the volume of social media data being produced daily that are relevant to social science. However, an important point to make about the size of these data is that we are rarely talking about big data *analytics*. We are talking about big data *management*. This is the pressing size issue for social scientists and public sector and third sector organisations. It is about sifting through vast amounts of social media data to whittle down a data set that is usable to address research questions. Once we've managed and tamed the big data, we very often end up with relatively small datasets for analysis. In our research an average size dataset for analysis is around 200,000 cases. But we start off with 200 million cases. Therefore it is about data management and refining the data so that the analysis stage can be done on a desktop. The sifting and storing and searching has to be done on more sophisticated hardware and software.

Velocity refers to the speed at which these new forms of data are generated and propagated by users. Recent social unrest illustrates how social media information can spread over large distances in very short periods of time. **Variety** relates to the heterogeneous nature of these data, with users able to upload text, images, audio and video. Peat (2010) describes how the average citizen is now a walking eye on the world, a citizen journalist, with the ability to take photographs, add captions and upload to the Internet for millions to see. This multi-modal mixed dataset can be harnessed by researchers. However, unlike qualitative and quantitative data that are often labelled, coded and structured within matrices and ordered transcripts, big 'social' data are messy, noisy and unstructured. **Veracity** relates to the quality, authenticity and accuracy of these messy data. The difficulty related to posing questions for social media analysis can result in findings that are superficial and lacking deep insight. To mitigate this shortcoming, we advocate triangulating social media communications with more conventional sources, such as curated and administrative data. Instead of big social data acting as a surrogate for established sources, they should instead augment them, adding a hitherto unrealised longitudinal extensive dimension to existing research strategies and designs. For the first time, this allows social scientists to study social processes as they unfold in real time at the level of populations, while drawing upon gold standard static qualitative and quantitative metrics to inform interpretations. Furthermore, we have shown that the near ubiquitous adoption of smartphone technology and social media amongst groups that are underrepresented in official survey collection exercises means these new data sources may provide better coverage of such populations.

Virtue relates to the ethics of using this new form of data in social research. The ESRC Framework for Research Ethics highlights the two key principles of informed consent and harm to participants. But it is not practically possible to seek informed consent from Twitter users in big 'social' data research, and Twitter's Terms of Service require users to consent for Twitter to share any content posted with third parties. A recent survey by our Lab found that 74 per cent of social media users knew that when accepting Terms of Service they were giving permission for their information to be accessed by third parties. Eighty-two per cent of respondents were 'not at all concerned' or only 'slightly concerned' about university researchers using their social media information (however this dropped to 56 per cent for police access). We may argue therefore that researchers in this field must accept that consent has been provided, as long as researchers adhere to basic principles of social science ethics, while ensuing results are presented at an aggregate level. Additional individual level consent should be sought if researchers wish to directly quote online communications.

Finally, **value** links the preceding five Vs – only when the volume, velocity and variety of these data can be computationally handled, and the veracity and virtue established, can social scientists begin to marshal them and extract meaningful information. However, to date, few academic social research studies have collected and analysed social media data. This is primarily due to the lack of existing computational infrastructure to support us in gaining access to and analysing these data, and the lack of interdisciplinary working practices. In order to make sense of this rich material, we advocate the establishment of interdisciplinary teams of computer and social scientists using parallel computing infrastructure.

At this point I'd like to talk about the Lab's COSMOS Web Observatory software and the collection and analysis of social media data, primarily Twitter data. The Social Data Science Lab was established essentially to democratise these data amongst academics and public sector organisations; to provide an access point so that individuals can download our software and start collecting and analysing social media free of charge. You can request to download from our web page and view a set of instruction videos (socialdatalab.net/software). Essentially there are different forms of analyses you can select - frequency of tweets over time, networks of tweeters (mentions and retweets), the sentiment of expression, the gender, age and location of tweeters and so on. In a moment Luke will also talk about the ability of the software to link to other forms of data, including curated and administrative forms (we currently have a link to the ONS API which allows us to import census data which we can then overlay with social media data to give us a more interesting picture of what might be going on in a certain context). This is what the platform looks like. This is an example of data we collected around the Boston marathon. As you can see here before the explosion, we have tweeters talking about the marathon roughly in line with the marathon route itself. Here we can identify a spike in communications where the traffic goes up markedly at a point in time. If we move the mouse over to that bar showing the anomaly, we find the content of the tweets change completely. It's clear the explosion has happened and the tweets about the event divert from the path of the marathon that we saw earlier on, spreading information of the attack across the city.

In terms of data collection, we have been collecting 1% of all Twitter data for the past 3 or 4 years for our database of around 5 billion tweets, which we have been using to run some of our experiments on identifying the gender and age of twitter users. But we cannot share that data unfortunately. The main reason for Twitter precluding the sharing of their data is quite simple. Twitter wish to honour the privacy of their users, and sharing historic datasets may jeopardise this. For example, if someone deletes a tweet that may not be reflected in our dataset that was collected in real time from the Twitter API. People can delete tweets months after they have made a tweet which really makes it difficult to maintain the privacy of individuals when sharing data. To go back to the Twitter API and query it for a deletion is a very arduous task and we don't have the resources to make this happen at this point in time. However, there is some hope. Twitter do allow you to share the tweet IDs. Each tweet you collect has an ID and that ID can then be used to query Twitter again and they can return data that has not been deleted allowing you to perform a secondary analysis. But this process does require a level of technical proficiency. If you do not collect in real-time, you have the option of purchasing it from Twitter's subsidiary GNIP. Unfortunately Twitter data is very expensive. On average it costs a minimum of $500 a day just to search on their systems and we have paid no less than £2000 on a simple search with a resulting dataset of around one million cases.

Now I'd like to talk about the first case study that used the COSMOS software and data from Twitter to study the production and propagation of cyber hate following the Woolwich terror attack. To give you some context I'd like to go over a few studies that have examined hate crime offline and its link to events. On a European scale, Legewie established a significant association between anti-immigrant sentiment and the Bali terrorist bombing using Eurobarometer data. King and Sutton found an association between terrorist acts and a rise in hate crime incidents in the US. Convincingly, they show that following the 9/11 terrorist attack law enforcement agencies recorded 481 hate crimes with a specific anti-Islamic motive, with 58 percent of these occurring within two weeks of the attack (4 percent of the at risk period of 12 months). In the UK Hanes and Machin found significant increases in hate crimes reported to the police in London following 9/11 and 7/7. These latter two studies were reliant upon police recorded hate crime data, and both sets of authors acknowledge the significant problems of non-reporting and lack of temporal granularity. The first two studies found that a sharp de-escalation was evident following the spike in hate crimes following the trigger event, indicating that event-specific motivated hate has a 'half-life'. These authors conclude hate crimes cluster in time and tend to increase, sometimes dramatically, in the aftermath of antecedent 'trigger' or galvanizing events, such as terrorist acts. They postulate that hate crimes are communicative acts, often provoked by events that incite a desire for retribution in the targeted group, towards the group that share similar characteristics to the perpetrators.

A focus on the temporal dimension of hate crimes allows for a study of their escalation, duration, diffusion, and de-escalation following trigger events. However, there are limitations in offline data: i) low temporal granularity; ii) issues with under reporting in official police data (particularly in the case of hate crimes); and iii) the retrospective nature of reporting and problems with witness and victim recall. Forms of naturally occurring online data, such as social media communications, while noisy and unstructured, lend themselves to temporal analysis. This is primarily due to the time-stamps that accompany all items of online social media communication. Further, researchers have argued that users of social media act like a distributed sensor network, often identifying events before the authorities and traditional media. Furthermore, users of social media are more likely to express emotional content due to phenomena such as deindividuation. Therefore, we argue that following trigger events, such as terrorists acts, social media users are often first publish a reaction, and given there are now over 2.5 billion users of social media (Smith 2014) these online communications provide rapid (to the second) insight into social reaction on an unprecedented scale.

What is particularly unique about social media communications, when compared to more traditional interviews or survey methods, is that user posts can be endorsed and spread by other users (in the case of Twitter 're-tweeted'), creating an information flow or propagation network that can be studied. For example, researchers can use such networks to identify what information or sentiment is being endorsed and propagated by users, and which users have the most or least influence in the spread of such messages. This locomotive, extensive and linked dimension of social media data allows us to study the fine grained (i.e. seconds instead of days, months or years) escalation, duration, diffusion, and de-escalation of social reaction following events, often far in advance of research using conventional curated or administrative data. In this study we examined the emergence and propagation of cyberhate following the Woolwich terror attack using part of Stan Cohen's framework for studying disasters: impact, inventory and early reaction. We focused upon Twitter and analysed approximately half a million tweets two weeks following the event.

The data collection period spanned a month following the terrorist event in Woolwich. Data were derived from the Twitter social media network. This network differs from others such as Facebook, in that it is public and the data are freely accessible by researchers. Twitter also has an open friendship network (non- reciprocal linking between users means that the followed are not required to follow their followers) resulting in a digital 'public agora' that promotes the free exchange of opinions and ideas. As a result, Twitter has become the primary space for online citizens to publicly express their reaction to events of national significance. A hashtag convention has emerged amongst Twitter users that allows tweets to be tagged to a topic which is searchable. The term 'trending' is used to describe hashtags that become popular within the tweet-stream, indicating a peak or pulse in discussion usually surrounding an event. Data were collected via the Twitter streaming Application Programming Interface (API) based on a manual inspection of the highest trending keyword following the event (i.e. ''Woolwich''), the most common strategy in the field of information diffusion online. This strategy produces robust samples due to the interactive nature of keywords and hashtags, where followers of events on Twitter actively seek out the most popular, or trending topics/hashtags in order to identify relevant information and subsequently add to the flow by replicating the keyword or hashtag in their posts. This selection procedure generates a census of tweets containing the most common keyword, and hence a large sample of all tweets about the event in question. An examination of Web search trends using the ''Woolwich'' keyword to query the Google Trends service indicated that an issue attention cycle around this event (the duration within which public attention to this event rises and falls away) spanned 15 days. This time window also maps onto the combined durations of Cohen's impact and early reaction phases in our dataset. This became the analysis sampling time frame for our study, during which we collected N=427,330 tweets.

These slides show snapshots of the COSMOS Dashboard showing geographic and temporal distributions of Twitter communications in the 15 day analysis window. As the content of tweets is not directly quotable in academic research[1], the COSMOS platform allows for the visualisation of tweet

[1] Twitter Terms of Service forbid the anonymisation of tweet content (screen-name must always accompany tweet content), meaning that ethically, informed consent should be sought from each tweeter to quote their post in research outputs. However, this is impractical given the number of posts generated and the difficulty in establishing contact (a direct private message can only be sent on Twitter if both parties follow each other). Therefore it is not ethical to directly quote tweets that identify *individuals* without prior consent.

content in the form of a WordCloud that presents an aggregate overview of the thousands of posts, while maintaining the anonymity and confidentiality of users[2]. In the first hour of data collection we can identify a relatively sparse geographic distribution of Twitter traffic across the UK (far left) and London (centre bottom). This is not unusual given that approximately 1 per cent of Twitter users enable their geo-location. Nevertheless, the relative distribution over time allows us to monitor the spread of social reaction. The WordCloud (centre top) is based on the full Twitter sample in the first hour of analysis (10,080 original tweets), not just geo-located data. Therefore this summary of content provides a window on the thousands of original commutations being sent during the initial reaction on social media, where size of word represents frequency of use in this period. The content in this early stage reflects the act ("attack", "killing", "murder", "london"), speculation as to the perpetrators' nationalities ("nigerian") religious backgrounds and possible motivation ("islam", "muslim", "religion"), and is devoid of any details on the victim apart from possibly gender ("man"). The focus on the perpetrator was likely fuelled by the YouTube video of the attacker that was uploaded within minutes of the event. Of particular salience to this paper is the presence of the terms "edl" and "hate". A closer inspection shows that the former term was being used to discuss the English Defence League's (EDL) various activities, mostly in a negative tone (e.g. criticism of a speech made by Tommy Robinson about Woolwich, and the rejection of an EDL donation to the charity Help for Heroes), while the latter term was being used in counter-hate speech tweets (e.g. shame on the EDL and BNP for spreading hate), as opposed to hateful tweets. Given the relatively low number of hateful tweets compared to the overall volume of communications no racist or religious slurs appear in the WordCloud. This initial hour of the study window, characterised by a lack of firm details and a degree of speculation, is akin to Cohen's impact stage in which citizens display an "unorganised response to the death, injury and destruction".

The next slide is a snapshot of the first four days in the study window. What is immediately apparent is the geo-tagged communications are more voluminous and widespread across the UK, with concentration in line with population density. London emerges as a hotspot for communications around the event, in particular the Woolwich region. The WordCloud (133,275 original tweets) shows the content of communications has shifted from disorganised speculation to specific details on the identity of the victim ("lee", "rigby", "drummer"), a perpetrator ("michael", "adebolajo") and the progress of the case ("arrested"). This more organised form of communication resonates with Cohen's inventory stage, where the mass media begin to play a central role, communicating to citizens the "preliminary picture" of the event. However, in our case it is worthy of note that the identity of both perpetrators was broadcast on Twitter hours ahead of conventional media. The next slide is a snapshot of the final three days of the study window. The geo-tagged communications plots represent the full 15 days and the cumulative volume of data now allows us to plot a more complete picture of the national social response, where we can observe dense clusters around London, the Midlands and Manchester (Lee Rigby's family home). The WordCloud (11,399 original tweets) shows Twitter communications have shifted from the specific details (victim, perpetrator, case status) onto boarder issues linked to the event ("british", "muslims", "islam", "religion", "terrorism", "media", "edl", "cameron"). This shift was possibly influenced by widespread media converge of the speeches made by David Cameron and Tony Blair in the latter stages of the study window. Cameron's mention of the EDL in his speech is partly responsible for their presence in the WordCloud, while the presence of "terrorism" for the first time is likely due to the use of the term in conjunction with radicalisation and Islam in the speech by Tony Blair. This shift towards the broader issues related to the event is akin to Cohen's reaction stage, where the images emerging in the inventory are "crystallized into more organised opinions and attitudes".

[2] Tweets from public organisations, such as government departments and police services, are deemed quotable as no individual can be identified.

Of the 210,807 original tweets posted about the terrorist event in the 15 days following, 1,878 tweets (1 per cent) were identified by the validated supervised machine classifier as containing BME or religious hate related terms at the moderate (e.g. 'send them home', 'deport them' etc.) or extreme (e.g. 'muslim scum') level. This shows that *event specific* cyberhate targeted towards the perpetrators and BME and religious groups associated with them, was present in social media communications following the Woolwich terrorist attack. The targeted nature of the cyberhate (e.g. containing the term Woolwich) demonstrates that the attack acted as an antecedent trigger event, galvanising tensions and sentiments against groups that shared similar characteristics to the suspected perpetrators. This is the first evidence to suggest spikes in hate crimes and incidents following such events are not confided to offline settings. The statistical model shows that only Far Right Political Agents emerged as significantly associated with the production of cyberhate. Political organisations (e.g. BNP, EDL) and party members made up the minority of these agents, with more general far right identifying groups and individuals making up the majority. Proportionally, compared to Other Agents, the odds of producing cyberhate on Twitter were three times larger for these agents following the terrorist event. A closer inspection of the tweets produced by these agents reveals that it was the more general far right identifying groups and individuals who produced cyberhate and mostly at a moderate level (e.g. 'send them home'). The odds for including hashtags were higher for cyberhate tweets, while they were lower for containing URLs. This may suggest those wishing to promote hateful content online and sensitise others to minority groups in society via contagion use hashtags to enhance the discoverability of their content. Conversely, URLs are possibly less common in hateful tweets given linked content (most often a popular media source) is unlikely to corroborate racist opinion and biased speculative rumours. The positive association between a rise in news headlines about the event and cyberhate tweets evidences the link (albeit relatively weak) between old and new media. The odds of the production of extreme cyberhate increased by a magnitude of 1.3 for every 100 additional news headlines produced. During early stages following the event (e.g. impact and inventory) tweeters may be fuelled by coverage in the press who have a role in 'setting the agenda' and 'transmitting the images', especially those who wish to spread hate, biased rumours and speculation.

Next we modelled the *size* of information flows following the event. The most novel and salient finding was that tweets containing cyberhate were negatively associated with the size of information flows emanating from the Woolwich terrorist event. None of these tweets were statistically likely to form *large* information flows following the event. Tweets containing hate terms were 45 per cent less likely to be retweeted as compared to tweets not containing such content. Original and retweeted cyberhate peaked during the early stages following the event (impact stage) and sharply declined over the four days following (inventory stage). Given that terrorist events have been shown to increase levels of anti-immigrant sentiment and hate crimes and incidents offline, it is surprising to find a lack of cyberhate *propagation* in terms of size following the Woolwich attack. In line with this finding, we found tweets containing positive words and phrases (e.g. 'warm wishes to the family of Lee Rigby', 'brave family', 'respect for armed forces' etc.) as opposed to negative words or phrases, were 38 per cent more likely to form large information flows. Finally, type of agent emerged as significant. Compared to the reference category Other Agent, tweets emanating from News Agents were more likely to be retweeted by a factor of 4.3, providing evidence to support the notion that traditional media messages maintain their role in '*setting the agenda'* and '*transmitting the images'* in the age of social media. Police Agents were also more likely to be retweeted by a factor of 5.7. The first finding is novel and shows for the first time that Twitter users propagate police tweets following terrorist events in the UK. It is evident that users are sharing police requests for information (e.g. metpoliceuk: "We are appealing for anyone who may have witnessed the incident in #Woolwich to contact us via the Anti-Terrorist Hotline"), case updates (e.g. metpoliceuk: "Two men aged 22 and 28 arrested on suspicion of murder remain in hospital in a stable condition #woolwich") and general commentary (e.g. PoliceFedICC: "EDL marches on Newcastle as attacks on Muslims increase tenfold in the wake of

Woolwich machete attack")[3]. Evidence from the US suggests this pattern of Twitter user behaviour was also evident following the Boston Marathon terrorist attack. The finding that Far Right Political Agents were the least likely to have content retweeted is consistent with the previous finding in relation to cyberhate.

Next we modelled the survival of tweets over the 15 day analysis window. A positive estimate in the Cox regression model is interpreted as increasing hazards to survival, and therefore reduces the duration of the information flow. Cyberhate is negatively associated with long lasting information flows, emerging as having the highest positive hazard ratio (1.19) of the variables of interest. News Agents emerged as significantly negatively associated with hazards to survival, indicating tweets from such agents were likely to last longer in the study period. Counter intuitively, Far Right Political Agents emerged as having the second highest negative hazard ratio, after Police Agents, indicating that information flows emanating from these types of agents were likely to outlast those emanating from other agents at some point in the 15 day analysis window. To better aid interpretation we used Kaplan–Meier (KM) survival estimation to plot the survival functions of cyberhate and Agent Type. Flows containing extreme BME or religious hate terms die out rapidly, between 20-24 hours following the event. Tweets containing moderate BME or religious hate terms last a little longer, between 36-42 hours before dying out. Tweets containing no cyberhate show a longer survival curve. This evidence confirms that extreme cyberhate was propagated in social media networks following this event in the immediate impact stage, which was then replaced with the propagation of moderate cyberhate in the inventory stage, and finally little or no propagation of cyberhate in the early reaction phase. This sharp de-escalation resonates with the work of Legewie and King and Sutton who postulate that the increase in offline anti-immigration sentiment and hate crimes and incidents following terrorist events has a half-life. It seems likely that this offline pattern is replicated online in social media networks. It is evident that information flows emanating from Far Right Political Agents outlast all other agent types up to 36-42 hours after the event, at which point they lose ground to News Agents, whose information flows last the longest (excusing Other Agent). The survivability of tweets emanating from Police Agents in the first 24-hour window, and their subsequent demise at around 36 hours is a novel and policy relevant finding.

To conclude this case study, the dominance of traditional media and police information flows during the impact and early inventory stage was accompanied by small (in terms of size as determined by retweeting volume) but sustained information flows emanating from far right political groups and individuals. The small but sustained nature of these flows indicates that there is limited endorsement of these twitter narratives, but where there is support it emanates from core group who seek out each other's messages over time. Therefore, contagion of cyberhate information flows is contained and unlikely to spread widely beyond such groups. Furthermore, preliminary analysis not covered here evidences the presence of counter-cyberhate speech following the terrorist event. These narratives from Twitter users either directly or indirectly challenge cyberhate. It remains to be seen if such self-regulation in social media networks represents a form of responsibilisation that can lighten the burden of policing the 'cyber-streets'.

Question - what is your definition of a long period of time? Answer: We use something called 'the issue attention cycle' which is a theoretical position that suggests that the attention of the public wanes around an event. We used Google searches for the term Woolwich and interest seemed to tail off after about two weeks following the event.

Luke Sloan (continues)

[3] Usernames and text reproduced here as the tweet accounts belong to public organisations i.e. the Metropolitan Police and the Police Federation.

Hi everybody my name is Luke Sloan and I work with Matt in the Social Data Science Lab. I am a political scientist by training and I am very much interested in demographics. One of the reasons why social media will never entirely supersede survey methodology, despite people fearing that it will, is that we do not know who uses it. We have no standard data so we do not know the demographic characteristics. For some platforms, such a Facebook, it is more obvious who the user is, as people will put their sex or gender on there as well as their date of birth, occupation and educational attainment. Although of course, there is an assumption that they are telling the truth. But Facebook data is hard to get hold of as its hidden whereas twitter data is easy to get hold of, but demographic life. So a lot of the work I have been doing is about trying to find proxies for gender, for age and for occupation.

I am also going to do a bit more on geographical location as well. The bread and butter of social science, and the principle interest of most people in this room, will be to look at how different groups respond in different ways to social phenomena – and by different groups we might mean ethnicity, nationality, first language, whether someone has a disability or not. So I am going to give a brief talk about how I have identified some demographic characteristics, discussing how these techniques are proxies and are not perfect, and afterwards we can have discussions about them. These techniques are the first foot holds in the mountain that we need to climb as we try and encourage people to think about other ways to refine the methods that we use. After going through this, I want to actually apply them to some data I have collected on Ebola to show that we must be capturing something with our demographic identifications, because the data show that groups differ e.g. men were talking about different things and they were interacting in different networks to women.

So one way of identifying gender, and the way we will do it on COSMOS, is to automatically detect the gender of a user by looking at their first name. We use a database of 40,000 names which specifies whether they are male, female, unisex or unknown. This assumes that people are using their genuine first names and there are probably about 40% we cannot classify, because people do not put their first names in their profile. We use cleaning algorithms so that we can draw out first names from usernames and we try and identify things and break things up. Names such as 'BBCbreaking' would be unclassified, and one of the important things to always bear in mind is that, unlike in a survey, every Twitter account is not necessarily a person or a respondent. It could be an organisation, it could be a news outlet, it could be a group.

However, for those we can classify as male or female and those we can identify as being from the UK, 48.8% are male and 52.2% are female – it is exactly the same as the 2011 Census data. I nearly fell off my chair when I saw that so it could be just a massive coincidence or it could honestly be that there is no gender bias on Twitter that we can detect. It could be that actually there are a lot more men on Twitter but they do not use their first name, they use some kind of *pseudonym* and I am open to that possibility, but it does seem to be the case that there is no gender bias in presence. That does not mean that there is not a gender bias in behaviour networks, number of tweets, number of followers and so on and so forth. Actual *behaviour* might be affected by gender.

Now one of the pieces of analysis we did to test our ability to pick up differences between men and women, was to look at the London Olympics. We collected tweets containing the hash tag '#teamGB' and we collected on Super Saturday - so what you can see here is sentiment score down the side. The negative sentiment score is below and then we have positive above. I am afraid we have used blue and pink - not very good for a social scientist but just because I have to rush through this quite quickly - so pink down the bottom here is negative female sentiment, pink at the top here is positive female sentiment.

And, as one of the things Matt picked up, we have great geographic granularity, we also have wonderful temporal granularity (by the second) so on super Saturday when something happens in the stadium with ‘#teamGB’ we get an instant response on Twitter so when you see an instant peak this is really handy for identifying reactions to, for example, speeches on the run up to the election or any event you can pick up. In this data you find we have these 3 peaks here in positive sentiment for women and at all of these 3 points we can identify as being ‘Female on Twitter’, (or people choosing to present themselves as female on Twitter), are using much more positive language. What we find is that this is when Mo Farrah starts his race, when Mo Farrah moves up to 3rd and when Jessica Ennis wins a medal.

Now regardless of what you think of the way we are identifying males or females there is clearly something going on - these groups mean something because it appears that female tweeters are much more positive in their use of language, in sentiment around this event. But it might just be an Olympic thing, and one of the important things that Matt mentioned is that the event is very important, the context in which something is occurring is very important and when we look at Ebola we see some interesting differences to.

Location is going to be incredibly important for those of you with international remits, even UK remits as well, and there are three ways of identifying geography all at different levels. So I am going to start with the bottom one -geotagged tweets. So, about just under 1% of all the twitter content produced globally is geotagged and what that means is that when someone presses the button to actually post their message, it records their exact co-ordinates to the metre. You can follow someone through the day based on what they are tweeting and track them – it is a bit scary actually.

If you have it switched on you might want to switch it off. But that is incredibly useful for us because if we have got someone tweeting about their fear of crime we can locate them in a geographical area for which we know criminal activity for which we know census data on socio economic deprivation. So for geotagged tweets we can place people within a context, within the social context and that is incredibly useful and that is why a lot of the stuff about mapping and laying that over tweets is very useful. You can use content of tweets too to identify location, so if people are talking about places or mundane geographies or if they are saying “just down the road” and if you happen to know where they live (which you may not) or if they say “just gone down to the Dog and Duck” you can try and use that to approximate where people are talking about.

People talk about specific places such as the ‘Millennium Stadium’, ‘Wembley’, so on and so forth – that is quite easy to locate. But the amount of effort and compilation time to actually turn that into something useful is tremendous, although there is some interesting integration work going on between Twitter and Foursquare and what Foursquare have is an outstanding location database. So for people who use Foursquare when you check in its taking your lat long but then its saying “oh you are in the Millennium Stadium” so it checks you in at the Millennium Stadium rather than just giving you coordinates. So there could be some interesting movements there. The other source of geography is user profile information. So any of you who have Twitter profiles might have the city you come from… or is it the city you identify with or the city you did live in two years ago and you have not updated it? So, not reliable data. If someone says Manchester, first of all you have to work out, is it Manchester UK, Manchester USA and actually it does not mean that they identify with that place. Plus, Manchester as a geographical unit is not very useful. However, if you are simply trying to locate people within countries it might be feasible to assume that people spend most of the time in the country in which they are writing their profile. So that might be a way of locating people in very large geographic areas.

Here is some data we collected over a week and we mapped it all. You can see what interests me is certainly Europe and North America, whilst Africa is not actually very good for Twitter data. I mean there might be people using Twitter there but perhaps the infrastructure does not exist to support geotagging - you need a mobile phone for start and you need to have access to 3G GPS network to be able to do that. Australia population centres are obvious but not much in China as they tend to use other platforms such as Weibo. Certainly not much in Russia and the interesting work I have recently done (which I have not reported on here) shows that the language of the interface someone uses determines whether they turn geotagging on or not. There does not really seem to be any gender difference - men are not more likely that women or vice versa, but certainly if people are interacting with the Russian interface of Twitter which I am using as a proxy for Russian being their first language, they are much less likely to have geotagging switched on than say someone in the UK. There are some interesting differences there and it is not uniformed. It is culturally constructed behaviour in that sense.

Q: What are all the dots in the sea? A: So those are people either on boats or using their phone on a plane when they should not. When I first saw this I thought "my god we have got it wrong what are all these dots?" but it is really interesting and actually if you were to put major plane routes and shipping routes on this it might start to make more sense.

But you know all these people down here are probably doing scientific research in Antarctica. So there is a really interesting issue around anonymity there. If you are in the UK we ca not see you but if you are in Antarctica we know what you are doing. Now the other really interesting thing is age. Age is quite difficult to track and the way we do it is by having a load of simple rules which check people's profile information for any double integer which ends with "years" or starts with "I am". So we will try it by proxy and do some human validation to check and its actually really accurate and what we find is that we can get age data from around 0.35% of Twitter users, which is a very small percentage although actually still a lot of users in the scheme of things. Filter that by those you can locate within the UK and that is an even smaller sub section of those users with geotagging switched on. You can see an interesting age profile which probably would not surprise you but no one has proved this before.

The majority of users on Twitter are between 13 and 25 years old and then it very much tails off toward the end. This is the age distribution by population proportion according to the 2011 census. There is an age bias on Twitter and that is no surprise. However, again, these are small proportions but you are talking over 60,000 over the age of 40 in the UK who are in theory in this age group using Twitter, so there are still a lot of people there and the question then becomes about how you chose to control for this, how you to work with your sample and whether or not you can adjust and calibrate. We have had some interesting discussions with the big data team at the Office for National Statistics (ONS) and post hoc calibration and all the problems with it. Any statisticians in the room or demographers are going to be spitting out their coffee at that idea but we like to try these things and push the boundaries.

The other thing we can do is 'occupation' so we can look for signatures of jobs and occupational terms so you might well have outreach officer, web designer, lawyer or whatever and we will pick that up and match it to the database of occupations provided by the ONS and then map that into NS-SEC groups. It is a bit messy because a term like 'manager' could mean anything in the UK… but if someone says they are a lawyer it is pretty obvious where they should be grouped. When we look at this we find that there is a lot of confusion around the creative industry so people saying they are a writer, they are a photographer, web designer which might be a hobby rather than occupation so there is some confusion there. However human validation suggests that, particularly for the higher managerial professional jobs that are very obvious such as lawyer, doctor, and so on and so forth, the automatic classification is quite effective. What you find is that lower managerial is over represented on Twitter compared to the population at large and that is largely because people say the are a photographer or they identify with their hobbies and likes rather than there careers in meat packing at Aberystwyth for example. So that is probably why it is under representative of lower supervisory and technical.

So to apply some of this, particularly the stuff around gender and location, to Ebola, here is some data I collected in a moment of inspiration. I set up a data collection for any tweet containing 'Ebola', a really simple query and if you use COSMOS you can do this as well and you set it off and leave it running for a few days and it continues to collect. So between Tuesday 22nd and 24th I collected over 180,000 tweets containing the word 'Ebola' and this was not even the peak of the crisis. Because I was not collecting during the peak of the crisis I did not exceed the 1% of all Twitter traffic, so that meant I had got everything with the term 'Ebola' in. The Twitter API gives you 1% of everything for free so you can choose to basically say "yes I will have the random 1% that Twitter provide me" or "yes I will have anything containing this key term" and as long as the search parameters do not exceed 1% of global traffic you will get 100% of things containing that. So if you are collecting data at the moment on some humanitarian crisis in Africa then you would probably get everything. If you were collecting data on the World Cup Twitter traffic would probably exceed 1% so you would start to be limited.

So all those 0.94% had geo tagging enabled at the time. What is really interesting is looking at if any of these demographic differences manifest online, so we have got a public health scare outbreak and I was really interested to see if men/women react differently to it online (because in the real word we suspect that they would, so do it manifest into virtual?). This slide shows the split by male and female and this just shows that the frequencies, the patterns of frequencies are not that different – it is not as though women are talking much more about it than men that we can tell, so there are no surprises there, however the network analysis shows us something different. Male and female tweeters (and I will say it once more - those we can identify as male or female with issues around how they choose to present themselves) operate in different networks, so what we have here is this person that is the name of the user but in the male network that person is incredibly central so we have lots of points around there that are not connected to everything but this person is. In the female network we have got different users in the centre of the network, and this is an example of how demographic difference seems to be manifest online with people interacting with different networks and that might be behaviours, retweets, production of original content, it might be followers so there is something interesting going on there.

Now if you were dealing with a culture for instance or an outbreak or any situation where you knew you had to get hold of male tweeters for any reason, perhaps because they are enactors or carriers of something, knowing that the network is different is incredibly important because this person is central so you could include this person in your tweet and your communication strategy. So the work that we are doing around the horse meat scandal at the moment with the Food Standards Authority is looking at who the central actors are, so when another scare comes out they know how to maximise the exposure of their tweets by mentioning key people, key actors, whether a URL or picture increases the life of a certain tweet being retweeted and so on and so forth. However it may be very context specific.

This is a Word Cloud, the word is bigger the more frequently the word is mentioned. They are quite similar except for the real odd thing that '#fightEbola' only appears in female tweeters cloud, whereas it did not feature frequently enough in male clouds – I have got no idea why this is, as it strikes me as an obvious hash tag that would be universal. It would be really interesting to unpick this. All these slides will be online and you can have another look.

Now these are for the tweets that I could map – 1,000 or so geotagged tweets and this is where they were located. And what you find is that most of the talk about Ebola is in the western world which is not affected by Ebola as much as Africa. Very little in Russia or China, they are not concerned at all even allowing for the low number of geo taggers in that area it is very small, very little said about it (although they may not be using Twitter!). So what I want to do is show you how you can use COSMOS to identify differences in discourse based on geographical location. I have highlighted North America here and we have got Africa here and you can use satellite view as well if you want. A Word Cloud illustrates that what you find in Africa is the passing on of news, passing on information whereas in America you have got lots of things about blame, you have got references to Obama, you have got swear words, you have got a different discourse, perhaps a discourse of blame and fear compared to a discourse of just trying to get the job done and sort things out. So there is a very big difference based on geography there and if you did not divide it by geography what you would end up with is this heterogeneous mess and you would not be able to find anything sensible in it. This is what Matt was saying about noise. There is a lot of data and you need to find ways to filter it to split it to try and make substantive sense of what is going on.

One other thing I just want to talk about is when you put gender, geography and sentiment together. Again you can do this in COSMOS, so what I have got here, sorry pink and blue again just for simplicity, are pink dots that represent geotagged tweets which were produced, we think, by a female user and blue dots by males. What we can also do then is change the size of the dots according to sentiment so what I have done here is specify that the more positive the sentiment, the bigger the dot. And you get something like this. What you can see is a really interesting dominance of a male positive sentiment particularly in Europe and Scandinavia with a focus on female positive sentiment in the UK which is not characteristic of the rest of Europe - people are looking to all manner of different patterns I am sure. But the interoperability of the tools means that you can bring multiple things together and all of a sudden you start telling a much richer story about what is going on. Down in Australia and Asia, these dots are all small which probably means that they were sentiment neutral or they were not very positive or perhaps they were passing on information so there was nothing positive to say. Where you have lots of bigger dots suggests that perhaps people are giving opinions and sentiment is typically inflated positive or negative when people are giving their views rather than passing on information. That is the other thing with Twitter of course, you do not know why people are using it, you do not know if a retweet is actually an endorsement and how people are using the platform.

So in summary, for demographic differences it looks like the demographic tools we have are picking up real difference that manifest online and I would be happy to talk to anyone afterwards as to how they think they can apply this to their third sector areas and/or organisations because this software is free for you to use and download, you just need to ask us.

Summing Up: Neil Serougi

When we started off on this particular project over a year ago I have to say I had no idea what this would look like at the end of the journey. It was something that emerged from an idea - and all credit to the ESRC especially in this respect - that was born out of a good intention to explore a new terrain rather than us having a fixed notion of what civil society and civic society organisations would need in terms of data skills, information and knowledge. As David Walker said yesterday, it is the first of its kind.

Consequently, we were not able to actually go down any ‘well-trodden paths’ or understand and know how others had actually addressed these needs in the past. I think I began to understand the scale of the task when I started to try to elicit exactly whether there was sufficient interest within the sector, and found lots of organisations were unresponsive, partly because of fear of the unknown. Data and the whole idea of being able to use it in newer novel ways was something I think they felt might be a step too far, and whilst they listened to what I had to say, they looked at the agenda and thought this was great stuff, but could not see themselves being able to operate at that level.

So even being here today is a sign of success and I want to especially thank Louise and Libby who have actually mentored me over the last year in terms of being able to get to this point where you folks have been able to enjoy, I hope the word enjoy is the right one, a day and a half of expert advice and new debates that even I had not thought of. Some of the material that Hersh Mann from the UK Data Service did earlier today was particularly impressive and enabling. I thought I knew the UK Data Service and what it it did but this was absolutely great stuff – innovative and exciting.

So just reflecting on what we've thought about over the last one and a half days: I've written down four or five things. It’s by no means exhaustive but this is what I have taken from it.

I certainly enjoyed the idea and subsequent debates that came out of us thinking about civic society and what it means and, particularly what it means for our evolution as organisations working in a very data-centric society. By exploring the very different forms of data that this entails we explored some really substantive themes beyond conventional measures. Certainly, this morning regarding the social media and the stuff that Sian did around qualitative methods, we heard a welcome addition to conferences like this, which can often be top heavy with quantitative analysis. I really want to reiterate that it is really is important for us in our types of organisations when we think about producing evidence to go beyond the role of data as the key performance indicator (KPI) and think about other novel ways in which we can marshal information into the kind of policy impact strategies that we need to persuade others.

We also looked at the current state of play. I think I was always anxious when we were asking you to do pieces of self-assessment work on top of the day job, but this actually increased your participation and gave you an investment in the proceedings. Consequently I think that some of the information that came out of knowing what you use, what your challenges are, the types of software and hardware, all contributed to valuable exchanges throughout the conference regarding resource requirements and skills deficits. It brought home just how investment in civic society organisations is very much constrained by the variability of funding; unsurprisingly there is not as much money to go around and there is certainly not as much money to up skill researchers and that is a really good point.

When I was trying to get people to come to this event there was little evidence of bespoke roles and responsibilities that were dedicated to making information and data work in terms of research driven evidence. Rather it was fragmented across different roles and responsibilities within organisations. So when I was looking at the information that had come back around your data needs, it was a salutary lesson that there is still a need to prove the case for investment in information resources within civic society organisations.

Then we looked at the ethical challenges. I just think this is going to grow and grow in importance. Every day, or what feels like every day, there is some news item about charities that cannot help but erode public trust. I think this is where we can use information in novel ways, and we talked about the need to start using data in ways which reasserted the value of what we do in our own context. We cannot win an argument about 'our value' if people in communities themselves experiencing austerity start from a position that money is wasted. We have to start to show that what we do has real value for them as well as for our own clients. The debate that took place around that whole ethical context in which we promote our cause and how, was really quite instructive.

And I think that leads me onto the last point now and it is that we need to move on, we need to think of next steps and some things we can do in the short term and Louise is going to show us what that might mean in practice. A lot can be done, but in particular, I do feel that it is important that we do not allow the momentum to lapse in terms of putting in place a solution around brokerage. A lot of the people who are around these tables will go back and then be hit on Monday by all the work you were unable to do whilst at this conference. It's unrealistic that you are going to become bona-fide researchers with all the requisite skills, time and abilities so there needs to be some future direction of travel towards a brokerage system where universities and academics can help civic society organisations move to that next level both in terms of capacity and capability. Society as a whole is inexorably moving towards this new type of landscape and we cannot afford to be left behind.

So, all that remains is for me to say a big thank you to you folks. I think you have made it a really successful event by your contributions and participation and hopefully it won't be the last time that I see you. You are always welcome to actually email me, as you know I am Vice-Chair of Freedom from Torture, so a lot of the issues that you actually deal with I experience as well and so I will always be happy to receive an email from you if you want any help and advice.

Louise Corti

What we did earlier is to outline some of the things that form part of the core remit of the UK Data Service - to service people who want to use data, or who want to give us data. That is out remit across sectors engaged in research, and whilst we cannot promise to support everyone doing everything, there are a number of things we can do without needing to invest new resources.

One thing we can easily do is to run another one of these next year. I think this has been a really good discussion space; word gets round that it seems to be useful. We would really like your feedback on whether the format might be changed at all, for example, whether there needs to be more discussion time factored in. We would like to hear about your own experience of these two days.

We already run a programme of data management workshops. There is a core team of us who have written a book, the Sage Best Practice Handbook for Researchers on Managing and Sharing Data, written for researchers by researchers. We do these quite a lot of training with a range of audiences and do get participants from civil society organisations. We could easily run a tailored workshop – over one or two days -specifically for this sector looking at key issues surrounding how you are storing data, backing it up, data protection, confidentiality and consent; routine areas we cover in our training. We are not expecting you to have deep research skills, but that we can focus down on the core challenges of looking after data and thinking about sharing useful information assets you hold - pitched at a suitable level for those attending.

One thing we think we will definitely do is a webinar on how you can visualise data easily using free tools, without needing to invest in bespoke analytic tools. One of our colleagues, Rob Dymond-Green, was going to be coming today to talk on how you can use free Google tools to visualise data. He would be a good person to bring in to deliver this webinar. Megan has agreed that she will do something on using Excel tools to visualise data, so I think between them and maybe Matt, if your team would like to do something on social media data using COSMOS, we could have a great webinar. This would enable learning through examples and because these are free tools, they present a much lower barrier.

Again, following from Hersh's talk, we can easily run seminars or webinars on what are the useful data sources for this sector. We already run these events as part of our outreach programme. We think that having a stall at the annual Charities Fair meeting might be something we could do, covering both sides of data sharing and of data use.

We have been collating some useful resources that speakers and participants have spoken about over the last day and a half, and I think we will mount on a web page. However, I am all too aware of the danger of creating an up-to-date information resource, like so many information brokering sites do. These links support supporting data matters, but we cannot promise that we will check all the links all the time because resources go out of date so fast!

Then of course we do run an archiving service for anyone who wishes to share data via us. That is core part of our remit, so we can talk with you or your organisation on a one-on-one basis if you like.

So these are the kind of things that we can do. We would like to hear from you whether there are other things that are in demand, and where we could contribute. Talking to a European colleague here as part of the bigger enterprise called the Research Data Alliance - where 'data experts' in the world work together on bigger solutions, such as cloud computing solutions for developing countries. There is a lot of good stuff going on there, but it does tend to be but quite high-level. However, there may be some very practical things we can do to introduce the kinds of needs we have encountered at this forum into the international arena.

Is there anything else that anyone can think of that would be useful for them? If you cannot think of anything specific now, just let us know. Our doors are always open for conversation!

Q: Chris Carleton, ESRC: Can I just highlight that if you ever do get, or you feel like you have a research project in a department of academics, you can apply to the ESRC. There are certain costs that you can claim relating to that as well. We have a scheme that is open all year round to any aspect of social science details of which are on our website.

Louise: An idea we talked about was in identifying the right academics to partner with, and how you know who they are and what their skills are. We really like the DataKind model which utilises a brokering system. There are lots of academics who would like to do pro-bono work, and some of them already do, but maybe there is some kind of ‘dating agency’ which could promote, say, the availability and skills of retired academics who would love to do something with their time. I actually envision the concept of a Silver Matrix’ where we can harness all the brain power of retired academics, and get them to do useful data-related things. The academic sector could certainly think about some kind of dating agency like that, maybe gleaning from Emma ways to do the DataKind approach for academics instead of commercial sector data scientists.

Supporting human rights organisations to deliver insights from data

Date: 29-30 October 2015

Location: University of Essex

Day 1

Thursday 29 October 2015

10.00	Coffee and registration
10.30	Welcome *Louise Corti, UK Data Service*
10.40	Introduction: ESRC's Civil Society engagement and agenda *Christina Rowley, ESRC*
11.00	Key Note, *Neil Serougi, Trustee, Freedom from Torture:* **Standing out from the crowd. The value of 'data as evidence' in Civic Society**
11.30-13.00	**Session 1: In house data collection: what do you have, what do you need and what skills do you have to analyse the data?**

Session leader: Louise Corti, UK Data Service

Content: This session will focus on the types of data that organisations collect themselves, identifying pain points in collecting, cleaning, managing or collating data. It will look at what capacity and skills civil society organisations need (in house or in partnership) to undertake good data practices. We will hear from organisations on how they manage data, and will use facilitated small group discussions to illuminate gaps and needs.

Speakers:

- *Tracy Gyateng, New Philanthropy Capital*: **The growth in data production collection. Opportunities and Challenges**
- *Nigel Fielding, University of Surrey:* **Prisoner of the Past or Hidden Resource: Documentary Records**

Gathering data for international human rights work:

- *Roisin Read, Humanitarian and Conflict Response Institute, University of Manchester:* **Driving a Ferrari into the desert and leaving it there? The challenges of information management UN peacekeeping and humanitarian NGOs.**
- *Ingvill C. Mochmann, GESIS-Leibniz Institute for the Social Sciences, Cologne:* **Children Born of War: expanding the evidence base on hidden**

populations.

- *Discussion*

Prerequisite: Organisations will submit a brief overview of data types held and an overview of in-house data skills and capacity.

Potential workshop outcomes:

- Overview of needs analysis within civil society organisations;
- Guide on how to collect usable and sharable data (building on existing guidance)

13.00–13.45 Lunch break

13.45–15.30 **Session 2: Making an Impact: Using Data beyond Key Performance Indicators**

Session leader: David Walker, The Guardian

Content: In this session we will look at how data gets translated into knowledge, be it at strategic or operational levels; for engagement, campaigning, community education or media reporting. A key aspect of this session will be to explore and understand how we might maximise the potential of data and intelligence with opinion formers/changers. We will examine the value of co-production, identifying opportunities where partnership models might improve the means to deliver 'knowledge benefits'. We will also look at the role of data in human rights reporting; understanding how human rights are received in the media and how evidence is articulated. A key outcome will be the opportunity to understand how data journalism may have changed reporting and what gets seen as relevant as opposed to just interesting.

Speakers:

- *Bob Jones, Medical Aid for Palestine (MAP):* **The Challenges of Using Data to Assess Health Needs and Impact in Conflict Zones**
- *Emma Prest, Manger DataKind UK:* **Enabling the application of pro-bono data science to humanitarian problems**
- *Megan Lucero, Data Journalism Editor, The Times and Sunday Times*: **Computational investigative journalism and how computing can advance accountability and public interest reporting.**
- *Discussion*

Potential workshop outcomes:

- Brief overview of how reporting and campaigning is carried out in participant’s own organisations
- Collated case studies from successful organisations

15.30–16.00 Tea

16.00–17.30 **Session 3: Ethical frameworks and governance**

Session leader: Libby Bishop, UK Data Archive

Content: In this session we will hear from a small panel of participants who will present short case studies of ethical issues encountered and/or successfully dealt with in the use of data from their own organisations. We will discuss under what circumstances data can be shared amongst civil society organisations using data sharing agreements to help foster collaboration. In identifying 'risk' and 'harm' from a data sharing perspective, we will examine disclosure risk vs. social benefits examining it in the Human Rights context. Who are the morally relevant participants?

Prerequisite: Organisations will submit a short case of an ethical issue or challenge they have experienced

Speakers:

- *Jim Vine, Director of Evidence, Data and Insight, Housing Association Charitable Trust*: **Using a trusted intermediary to create insights from shared datasets in a collaborative sector**
- *Civil Society Organisation case studies*

Potential workshop outcomes:

- Summary of ethical statements in organisations and solutions that have been/ could be used to mitigate these
- Versions of the cases suitable for public sharing with comments from organisation members about how the solutions were implemented (based on cases in the Association for Research Ethics' monthly newsletter)
- A short guide on making data shareable and optimal pathways to access.

17.30 Round up of todays' issues and questions

18.00–19.00 Welcome drinks and networking

19.30 – Workshop group meal, Bistro, Wivenhoe House

Day 2

Friday 30 October 2015

09.00–11.00 **Session 4: Exploring opportunities for using third party data sources to provide context**

Session leader: Louise Corti, UK Data Service

In this morning session we look at the potential for using third party data sources to provide broader context knowledge for organisations. What skills are needed to analyse these sources and how much trust can we place in them? How could meeting strategic goals of civil society organisations benefit from using multiple methods and bringing together different forms of formal and informal knowledge? What is the role of administrative records or 'big data' in this landscape? What kinds of tools are available to analyse these sources of data?

Speakers:

- *Hersh Mann, UK Data Service:* **Sourcing 'society' data from the UK Data Service and beyond**
- *Sian Oram, King's College London & Mark Emberson, Medaille:* **Mental health responses to human trafficking: qualitative data tools**
- *Matt Williams & Luke Sloan, Social Data Science Lab, Caridff University:* **Gaining Insights from Social Media Data: Collection, Analysis and Interpretation**
- *Discussion*

Prerequisite: Organisations will consider potential data sources participants want to use or think could be useful.

Potential workshop outcomes:

- Database of suggested useful third party data sources

11.00–11.30 Coffee break

11.30–12.00 **Session 5: Summing up: what has been learned and what are next step challenges?**

Session leader: Neil Serougi

In this session before lunch we will review what we have heard and discussed over the past two days and think about, 'now what'? We ask participants to identify one concreate activity that they may now be better positioned to pursue.

12.00–13.00 **Session 6: Surgeries with data experts and lunch**

Session facilitators: UK Data Service staff and speakers

This session allows participants to discuss in more depth any issues that their organisation has with respect to the collection, management and use of data/information with experts from the workshop. This can either be joining a table to discuss a specific topic led by experts, or a one-to-one discussion about any topic of their choice.

13.30-14.30 Prebooked taxis to Colchester Station