RESEARCH PROPOSAL: THE SOCIAL LIFE OF (OPEN) DATA - PILOT STUDY

TIM DAVIES (TGD1G11@SOTON.AC.UK)

1. Core questions

The Social Life of Data Pilot will develop and apply a draft methodology for a dataset centric study of open data in the context of a pilot project looking at the re-use of data published through the International Aid Transparency Initiative. The research is framed by two sets of questions reflecting the methodological and applied research concerns:

Methodological research: How far can provenance-focussed approaches to mapping the artefacts, agents and processes involved in the use of a specific set of open government data increase our understanding of the use of open government data, and support critical evaluation of open data policy and practice? What is involved in creating an effective map of open data provenance for this purpose?

Applied research: How are datasets published through the International Aid Transparency Initiative currently being used? How are different uses of the data related to each other, and to processes of local meaning-making? What factors are implicated in different patterns of data use? What interventions can be taken to support diverse uses of this open data resource?

2. Motivation and focus

Narratives about the impact of open data often draw on a number of implicit or explicit assumptions, including: downplaying the prior (non-open) distribution and uses of those datasets now available as open data; assuming civic or economic value is derived primarily from direct and frictionless use of open datasets; assuming that the large datasets available from government or industry contain the appropriate contents required to drive change; and assuming that such datasets present a neutral view of the world that will support more rational policy making. These assumptions rest upon the treatment of datasets in the abstract, isolated from social context and from the patterns of human and material agency that impact upon them. Policy arguments concerning open data also frequently operate on the basis of anecdotes and isolated case studies of particular use of an open dataset, assuming the release of similar datasets will de-facto drive similar outcomes, without considering the different factors involved in the successful and sustainable use of open data Kuk and Davies [2011].

Date: December 2011.

It is important, therefore, to have a richer understanding of open data use in practice. However, tracing the patterns of human and material agency exercised in the production, distribution and use of open datasets is made particularly challenging by the non-transactional open nature of their distribution, and the diverse range of artefacts that can result from open data use (e.g. individual facts taken into conversations and decision making processes, information and info-graphics, interactive interfaces and applications, innovations in existing services, or adapted and derived datasets - see Davies [2010]). The social life of data pilot aims to develop a clear method for researching and recording information on the use of particular open data, and to test the utility of the resulting record for exploring questions concerning how open data impacts upon collaboration, policy making, practitioner action, and upon processes of meaning-making with data. The method should also support the exploration of open data value-chains, and an analysis of approaches that can best support economic and social innovation based on open data.

2.1. **Pilot context: The International Aid Transparency Initiative (IATI).** The International Aid Transparency Initiative (IATI) is a political process[1] started by aid recipient governments to ensure donors publish timely and detailed information on aid resources. IATI has developed an XML standard[2] for sharing data, and during 2012 it is expected that 75% of Official Development Assistance (i.e. aid from government, bilateral and multilateral donors) will be published using this standard. IATI invites donors to publish their information in simple XML files on their own websites - usually generating these files from existing organisational management information systems. These files are catalogued in the IATI Registry[3], and users of the individual datasets variously aggregate them together to create a comprehensive dataset of available aid activity information, or select particular IATI data files to work with. Whilst the IATI political process centres on government donors, a number of NGOs are starting to publish details of their aid activities using the standard. Since the first IATI standard data files were published in early 2011 the data has seen an increasingly range of uses, although in many cases the development of applied uses of the data remain in their early stages.

Between May and December 2011 the researcher was heavily involved in supporting the development of tools, skills and resources for working with the IATI datasets, seeking to apply insights gained from Kuk and Davies [2011] to support the development of shared artefacts, shared knowledge, and communities of actors, to support work with the data[4] However, in the absence of a clear method to map out how these interventions have impacted upon highly distributed patterns of data use, evaluating interventions (including engaging in participative evaluation with data stakeholders[5]) is challenging.

[1] http://www.aidtransparency.net

[2] http://www.iatistandard.org

[3] http://www.iatiregistry.org

[4] .

[5] Where stakeholders is defined very broadly to include those involved in the production of data, those already engaged in the use of data, and those who stand to be affected by use of the data, either directly or indirectly.

3. Research design and method

The research design adopts a mixed-methods approach Brewer and Hunter [2006], Tashakkori and Teddlie [1998] in which data collection is split into two parts: generating a schematic map of open data use from documentary sources and fact-focussed e-mail interviews; and then discussing that map in qualitative key informant interviews or group discussions. The combination of methodological development and applied research in the design can be understood within the framework of strategic research, facing both academia and practice, seeking to balance a focus on problems likely to emerge in the medium-term future with concern for the state of scientific knowledge Daniel [1993].

The schematic mapping aspect of the project will draw upon the Provenance Data Model (PROV-DM) draft recently released by the WC3[6] to create a basic web application[7] for mapping the different artefacts, agents and processes associated with the production and use of a particular open dataset (or collection of datasets). The pilot will focus on data generated through the International Aid Transparency Initiative (IATI), and will draw upon documentary evidence published online concerning data use and upon short e-mail enquiries to identify additional uses of IATI open data. Through a process of iterative coding, a draft taxonomy for categorising artefacts and processes involved in open data production and use will be developed, and a detailed record of IATI data use manually entered into the web application creating a research database. The resulting database will be visualised in an interactive browser drawing on methods from social network analysis, showing how different artefacts around the open data under investigation are related, by people, processes and shared artefacts.

Visualisation of the map will be used as the basis for small semi-structured group discussion (likely to be conducted over Skype) and individual interviews with a number of key informants with a connection to the aid and development field and open aid data. They will be asked to consider how they would engage with different uses of IATI open data in their own work, and to add annotations to the visualisation of data use patterns highlighting important aspects of the context of the data which they feel are missing from the map.

3.1. Data collection and sampling strategy.

The map of open data will be based upon primary research and entering data into the custom built web application. The researcher is a participant observer in the general open data field, and specifically in the use of IATI data, having worked on contract for AidInfo, an executive member of the IATI Secretariat. It is anticipated that the initial mapping of uses of IATI data will be based on comprehensive and structured documentation of the researchers existing knowledge, backed by examination of available online material, open source code and blog posts. Following this, a snowball sample method of e-mail correspondence with key informants to identify as yet unknown uses of the data, or potential leads with respect to data uses, will be used. It is hoped that this process will be able to draw upon continued engagement with the IATI Secretariat, both as key informants, and

[6] http://www.w3.org/TR/2011/WD-prov-dm-20111018/

[7] Using the Django framework in Python

securing secretariat support to distribute a call for information on data uses more widely, potentially including a public call for information on the IATI Registry website. The goal of sampling for the map will be to capture as wide a range of data uses as possible, and to ensure all forms of data use[8] are represented, ideally with more than one example for each. Analysis will need to be sensitive to the fact that any map will have missing data, and time constraints during the pilot are likely to mean that the final map will not be comprehensive.

Group discussions and key informant interviews will take place once a draft version of the map has been produced. The sampling of key informants will be based on a purposive approach to select individuals with diverse perspectives on IATI data. It is anticipated no more than 7 key informants will be involved in the pilot phase of this Social Life of Data project.

4. Ethical considerations

Applied research, internet research, and open data research each give rise to particular research ethics questions and considerations. As an applied project in a context where I have had an existing role aligned to one of the institutions in the field, I need to be transparent about the objectives of the research being undertaken[9], and the relationship of the research to the institutions I have previously worked with and attendant to any power dynamics or biases that may occur during data collection by my association with IATI institutions. I need to be aware of where data already available to me is privileged data, and to take care to secure appropriate consent to use any such information and data in the research. This consideration taken into account, much of the information I will draw upon concerning use of IATI data is available in public and semi-public online spaces. Eysenbach and Till [2001] recommends assessing the potential for harm from any online data collection, and Association for Internet Research (AoIR) ethical guidelines note that the vulnerability of those involved in a study should be taken into account Ess and AoIR Ethics Working Committee [2002]. Mann and Stewart [2000] highlights that the extent to which an online space should be understood as public or not should depend not only on whether content is formally accessible across the Internet, but on how the users publishing that content understand the potential audience for it. In many cases, material on open data use is clearly placed into the public domain, with explicit license statements on source code, or with narratives about data use published on openly accessible blogs and message boards. However, even when information is in the public domain, combining it in new ways can have ethical implications[10] and as Dutton and Piper [2010] describe, doing business in the new network-enabled research environment may require researchers to more explicitly manage competing risks, rather

[8] With the categories of data use applied here to be developed through the process of iterative coding

[9] Transparency about the process of the research is also important to contribute to an increased awareness in the field of how publicly domain data may be used, and the contribute to an increase in the critically informed decision making of actors publishing content in the online public domain.

[10] For example, if an individual has worked on a project with data for their employer, but is also known to have worked on another data use project in their spare time without the knowledge of their employer it may not be appropriate to record that the same agent has been involved in both projects.

than the application of inflexible and potentially inappropriate best practices from non-digital research contexts. Neuhaus and Webmoor [2011] suggest this approach could be understood as an agile ethics for the online environment. A case-by-case evaluation of mailing list posts, posts on limited-access discussion boards, and source code without an explicit license will be required to assess where additional consent to use details from those postings should be obtained, or where some selective anonymisation should be applied. Particular attention should be given to any data that includes reference to individuals or organisations working in civil society contexts, particularly if situated in non-democratic states.

Where information is directly solicited from key informants by e-mail interview or online form respondents will be presented with a clear statement of how their replies will be used and asked to confirm their consent to data re-use[11] (See appendix).

Interviewees will be asked to give informed consent prior to the interviews, and will be offered the option of anonymity or identification in the written report.

Mann and Stewart [2000] highlights an additional potential ethical and legal challenge for this project. It is intended that the mapping dataset generated through the research will be published under an open license at the end of the project. However, UK data protection legislation sets out that data should only be used for the primary purpose for which it is collected, which, if the primary purpose is defined as producing project-specific research outcomes, would preclude opening up the data for others to use. Where consent is sought it becomes important then to ensure that consent permits secondary use of the data from the start Dutton and Piper [2010], and that the open publication of the data is done in such a way as to protect individuals concerned from any harms[12].

4.1. **Analysis.** Following from the mixed-model research design, the analysis of data will draw upon mixed methods, combining a range of qualitative visualisation techniques[13] with other qualitative techniques for analysis of interview data Miles and Huberman [1994]. Exploratory quantitative analysis of coded information in the schematic map will be used to suggest areas for qualitative attention; with scope for additional quantitative work to be offer confirmation for, or contradiction of, hypothesis generated in qualitative work.

[11] Where an respondent describes some proprietary use of data it may be possible to give a generalised description of this without revealing commercially sensitive information about that data use. PROV-DM offers the flexibility to record the connection between a use and a dataset without specifying all the details, and it may be possible to introduce semantics into the data model to mitigate agains the impacts of missing data where respondents do not consent to full data sharing.

[12] For example, if the researchers copy of the dataset includes e-mail addresses of agents identified in the data, it may be appropriate to publish a non-reversible hash of these, rather than the actual e-mail addresses in an open version of the dataset

[13] Neuhaus and Webmoor [2011] note that visualisation, the compression of large quantities of data through the visual register, is frequently foregrounded in web-eased research due to the nature and scale of the data

The analysis will focus on the production of two reports: the first will be prepared for a practitioner audience in partnership with the IKM Emergent programme and will highlight learning about processes of meaning-making around the dataset; the second report will critically evaluate the method adopted, and will explore future methodological developments to support the evaluation of open data initiatives. In addition, findings with practical relevance to IATI will be shared through short blog posts and a presentation to the IATI and AidInfo teams.

5. Pilot timescale and partnerships

- Development of the web application for recording data - Now - January 15th
- Data capture surveying data from IATI - 15th - 30th January
- Skype interviews with key informants - 30th January - 15th February
- Write up and evaluation - 15th - 28th February

A partnership with IKM Emergent for the production of a report centred on the meaning-making questions of the study is already arranged. The researcher will also explore possible partnerships with the International Aid Transparency Initiative and AidInfo programmes to support data collection and dissemination of findings.

References

John Brewer and Albert Hunter. *Foundations of multimethod research*. SAGE, 2006. ISBN 9780761988618. URL `http://books.google.com/books?id=l5rJOlwzBxcC`.

W. W. Daniel. Strategic research: Political convenience or practical reality? *Policy Studies*, 14(3):4–17, 1993. URL `http://www.informaworld.com/smpp/ftinterface content=a787296341 fulltext=713240930`

Tim Davies. *Open data, democracy and public sector reform: A look at open government data use from data. gov. uk*. Practical Participation, 2010. URL `http://scholar.google.co.uk/scholar?hl=en&q=open+data+and+democracy+davies&btnG=Sear`

William H. Dutton and Tina Piper. The Politics of Privacy, Confidentiality and Ethics: Opening Research Methods. In William H. Dutton and Paul W. Jeffreys, editors, *World Wide Research: Reshaping the Sciences and Humanities*. 2010.

Charles Ess and AoIR Ethics Working Committee. Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee. pages 1–33, 2002.

Gunther Eysenbach and James E Till. Ethical issues in qualitative research on internet communities. *BMJ*, 323(7321):1103–1105, November 2001. doi: 10.1136/bmj.323.7321.1103. URL `http://www.bmj.com http://www.bmj.com/cgi/content/short/323/7321/1103 http://www.bmj.com/cgi/reprint/323/7321/1103.pdf`.

George Kuk and Tim Davies. The Roles of Agency and Artifacts in Assembling Open Data Complementarities, 2011.

C. Mann and F. Stewart. *Internet communication and qualitative research: A handbook for researching online.* Sage Publications Ltd, 2000. URL `http://books.google.co.uk/books?hl=en&amp;lr=&amp;id=HkuLBLAGp6UC&amp;oi=fnd&amp;pg=PA1&am`

M. B Miles and A. M Huberman. *Qualitative data analysis: An expanded sourcebook.* SAGE publications, Inc, 1994.

Fabian Neuhaus and Timothy Webmoor. Agile Ethics for Massified Research and Visualization. *Information, Communication & Society*, (October 2011):1–23, October 2011. ISSN 1369-118X. doi: 10.1080/1369118X.2011.616519. URL `http://www.tandfonline.com/doi/abs/10.1080/1369118X.2011.616519`.

A. Tashakkori and C. Teddlie. *Mixed methodology: Combining qualitative and quantitative approaches.* Sage Publications, Inc, 1998.