Trusted Brokers?: Identifying the Challenges Facing Data Centres

Lauren Thornton[1], Victoria Neumann[1], Gordon Blair[1], Nigel Davies[1], and John Watkins[2]

[1]{l.thornton2; v.neumann; g.blair; n.a.davies}@lancaster.ac.uk
[2]{jww}@ceh.ac.uk

Extended Abstract

Environmental research data centres operate at the intersection between research, industry, and government as they operate as 'data brokers', who archive, curate, and distribute data [7]. They are a valuable infrastructure to the provision of data for policy, directly accessed by government researchers and indirectly used to inform policy based upon academic research. However, in recent years there are new challenges impacting their role as data brokers. These arise from trends pushing towards more transdisciplinary research and the increasing volume and diversity of data, impacting data management. This paper highlights these contemporary challenges, with a particular focus on the necessity of trust for data management and data brokerage. We show that these challenges are sometimes competing and need to be addressed both individually and as a whole. In this way, data centres will not only be able to meet the new demands placed upon them but more importantly, maintain this essential function between data and policy.

Our empirical research is based on a collaboration between the Natural Environment Research Council (NERC) and the Data Science Institute of Lancaster University. Semi-structured interviews with a range of internal and external stakeholders related to NERC's five data centres were undertaken. These interviews were used to gain insight into the challenges facing data centres presently and when looking towards the future. Following this, a one-day workshop with a larger audience of stakeholders was conducted. At the workshop we presented our findings and explored potential approaches to solving these challenges. We conducted ethnographic field work during the workshop, analysing the discussions held and capturing participants views. Combined, our empirical analysis contributes to the identification of key challenges faced by data centres.

The challenges we identified are related to the transition of data centres' role as brokers and the trust in them to carry out this role. Traditionally, data centres have maintained data archives within each scientific domain. However, in recent years the volume, velocity, variety and veracity of environmental science data has increased thus requiring complex data management and distribution [1, 4]. We found that standardisation techniques for labelling and aggregating data were not consistent. This is compounded by the increase in the heterogeneity of environmental data [4]. Standardisation would enable efficient data management and coherence, thus eliciting trust in data centres.

Alongside this, a transition towards open access and transdisciplinary research has meant that there now exists a more diverse set of users to cater to. In response to this, data centres are now looking to develop interactive platforms. This would allow their service provision to accommodate differing user needs. One participant noted that, in order for an academic from a different domain or a non-academic to trust data, they would not only need to trust the data centre as the source, but also require contextual supplementary information in order to interrogate the data thoroughly. Moreover, by gaining an understanding of the complexities and values of data we can handle the uncertainty within the science-policy interface processes more effectively [5].

Another important issue was provenance, which participants saw as also essential to foster trust. To gain a greater understanding of data, such as the implicit assumptions and uncertainties, participants argued for better systems to question and gain insight into the journey of data [2]. A formal chain of data would enable users to question where the data has come from and identify any underlying factors that may affect any results derived. Thus, the traceability of data is essential for researchers and policy makers alike to foster trust. This formalisation of provenance would also foster trust by data producers in data centres as brokers. We found that data producers can often be reticent when it comes to uploading data, for fear of this data being taken to produce potentially erroneous results by data users. Thus, evidence of the propagation of data may foster trust as they

can assess where data has gone to and be aware of its re-use allowing them to counter any unfolding issues. This is particularly important as combined, this would allow us to look at policy decisions and the evidence used to inform them, and vice versa, enable data producers to see how their data has been used to inform decision-making.

During the discussion it became clear that trust in data and data centres is paramount. This trust ranges from data producers to data users as academics, public and private sector workers. However, whilst trust is entangled with each challenge, the different mechanisms establishing or maintaining trust are competing with each other [6]. For example, whilst increased data traceability is important for policy makers and scientific communities, it is undesirable for certain industries as they are often protecting business models. This might result in a decrease of engagement by these third parties owing to a lack of trust in sheltering their business interests. This creates a paradox in which greater data quality assurance increases but trust decreases. This could be problematic for data centres looking to increase efficiency by turning data into assets [3] and strengthening engagement with industry. This illustrates how multi-faceted trust is. Future work needs to look deeper into the mechanisms for trust and how new technical approaches might support the work of data centres in building and maintaining trust. This could lead to an expansion what environmental data for policy means beyond the classical governmental enforced regulations to sector-specific polices (e.g. the development of industry self-regulating policies).

We conclude that the challenges faced by data centres need more empirical exploration particular around the creation and mechanisms of trust. Trust is especially important regarding the impact of data and data centres for informed policy making. Combating challenges is not an easy task as there is no one-fits-all solution. Consequently, we need to discuss and formulate clear prioritisations regarding the role of data centres as a valuable infrastructure ought to develop and function within our society to become a data broker of the 21st century.

Acknowledgements

This work was supported/funded by the EIDC "EnvChain" NERC Data Innovation funding award (February 2018). We would like to thank all interview and workshop participants for their time and valuable input.

References

[1] Allcock, B., Bester, J., Bresnahan, J., Chervenak, A.L., Foster, I., Kesselman, C., Meder, S., Nefedova, V., Quesnel, D., and Tuecke, S., "Data management and transfer in high-performance computational grid environments," *Parallel Computing*, vol. 28, no. 5, pp. 749–771, 2002.

[2] Bates, J., Lin, Y., and Goodale, P., "Data Journeys: Capturing the Socio-material Constitution of Data Objects and Flows," *Big Data & Society*, vol. 3, no. 2, pp. 1–12, 2016.

[3] Birch, K., "Rethinking Value in the Bio-economy: Finance, Assetization, and the Management of Value," *Science, Technology, & Human Values*, vol. 42, no.-3, pp. 460–490, 2016.

[4] Blair, G.S. "Complex Distributed Systems: The Need for Fresh Perspectives," *Proceedings of the 38th IEEE International Conference on Distributed Computing Systems*, 2018.

[5] Guimarães Pereira, Â., Guedes Vaz, S., and Tognetti, S. *Interfaces between Science and Society*, Sheffield: Greenleaf Publishing, 2006.

[6] Knowles, B., "Emerging Trust Implications of Data-Rich Systems," *IEEE Pervasive Computing*, vol. 15, no. 4, pp. 76–84, 2016.

[7] Welpton, R. "Research Data Centres: The Role of Brokers for Negotiating Access to Data," *Data for Policy 2017: Government by Algorithm?*, 2017.