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]*Data Science Institute, InfoLab21, Lancaster University, LA1 4YW, UK*
[2]*Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster University, LA1 4AP, UK*

**(l.thornton2);(v.neumamn)@lancaster.ac.uk*

Abstract

Research data centres (RDCs) in environmental science are currently facing challenges due to a number of factors. These include increased volume and heterogeneity of incoming data, transdisciplinary research, and a growing diversity of data consumers from academics through to private industry actors and governmental bodies. Many of these challenges relate to perceived trust in the data provided by the RDCs and in the data centres themselves. In this paper we explore these challenges and identify five distinct themes or 'mechanisms' (standardistation, supplementary information, interactivity, provenance and traceability, and the management of stakeholder interests). Using the lens of trust to situate these challenges in RDC practice, we discuss how these challenges and mechanisms relate to the emergence of new technologies such as blockchain. We report that there are many benefits that blockchain technology can have in RDC brokerage and data management, and in fostering trust in data centres by data producers and consumers. However we also note that this technology can also have unintended consequences, impacting upon the trust held by stakeholders. We conclude that trust is an appropriate construct for combating the challenges that RDCs face, but that in order to effectively design and implement these mechanisms, care should be taken with the underlying and often implicit intricacies. We recommend that these intricacies should be mapped out and planned before implementing technology, and that future work will upon this.

***Keywords*—** Trust; data brokerage; data centres; environmental science

1 Introduction

Environmental science encompasses a broad range of disciplines, from biodiversity (Roy et al., 2012) to climate change (Silvertown, 2009). In comparison to other domains, environmental data is highly heterogeneous and 'difficult' to work with (Blair, 2018). This is a consequence of a blend of disciplines that vary from a small number of 'big science' fields, that are large in data volume but homogeneous in instrumentation, format, content and structure and a larger number of 'small science' fields in which there are smaller amounts of data but with more heterogeneity and variety (Borgman, 2015). Combined, environmental science encompasses a blend of scientific practices and data. This data contributes to our collective knowledge of the natural world, and informs environmental policies and new research directions.

Environmental data is held in environmental research data centres (RDCs). RDCs archive, curate, and distribute data for data consumers (e.g. governmental bodies and academic researchers) and ensure that the interests of data producers are protected and data is available and accessible for a range of consumers (Welpton, 2017). Environmental RDCs curate a multitude of different varieties and sources of data for many scientific sub-disciplines and function as an intermediary between research and knowledge formation on the one hand and policy-making and legislation on the other.

This paper documents our research into understanding the challenges that environmental RDCs face (detailed in Section 2) in successfully acting as data brokers in the 21st century and their exploration into using distributed ledger technologies (commonly known as blockchain) to combat these challenges. Empirical research was undertaken during a collaboration between Lancaster University's Data Science Institute and Centre for Ecology & Hydrology as part of a Natural Environment Research Council (NERC) funded initiative as detailed in Section 3. Following this in Section 4, we present the findings of our research. We found that trust plays a fundamental role in successfully counter-

ing challenges faced by RDCs and develop five actionable mechanisms for fostering trust. Finally, in Section 5 we discuss the implications of our findings and consider the implementation of these trust mechanisms.

2 Background

There are significant changes in the practice of science and data that have contributed towards a series of contemporary challenges for scientists, academics, private sector and legislators alike (Borgman, 2015). As an intermediary between different parties, these challenges also impact upon RDCs. These relate to cultural shifts underlying scientific research and data practices in recent years. As discussed above, the nature of environmental data is changing. Environmental data is heterogeneous and is becoming increasingly more so with the proliferation of new data sources, such as remote sensing, citizen science and real-time data collection through the Internet of Things (IoT). Partly as a result of this, and due to the growth of computational processing power, there has been an 'information explosion' furthering the complexity of data management and distribution (Allcock et al., 2002; Korth, 1997). Environmental RDCs are therefore facing challenges in their ability to successfully curate a wide and diverse range of data for a number of scientific disciplines.

Furthermore, RDCs also face increased pressure to provide a greater amount of detailed, reliable data to a wider audience. This stems from a change in the level of scrutiny of scientific knowledge (Guimarães Pereira, 2006). Data and published findings are no longer taken at face value and there is a demand for deeper understanding. To gain this understanding and contribute to knowledge creation, there is a desire for additional information and meta data. Alongside this, there is also a push towards transdisciplinary research (Borgman, 2015) and a tradition of open access in the environmental sciences (Edwards, 2010). Combined this leads to specific challenges. Environmental RDCs must not only curate increasing amounts of heterogeneous data for a wide range of scientific domains and audiences, but have different requirements for these audiences dependent on their familiarity within the scientific discipline of the consumed data. This is also compounded by the increasing importance of impact, value promotion, and the re-utilisation (Birch, 2016). As a result of this, innovative technology such as blockchain and smart contracts are potential tools to address some of these challenges (Neisse, 2017). In order to gain further insight, we adopt the lens of trust as a means to deal with increasing complexities in environmental data centres.

Trust is a necessary part of any human interaction and co-operation (Knowles et al., 2018). Trust is traditionally thought of as person-to-person trust, which is based upon an assessment of another's reliability to act in a one's best interest (Clark, 2014; Gambetta, 1988; Riegelsberger, 2005). With regards to the contemporary challenges highlighted above, trust is necessary as it is a functional mechanism to reduce social complexity (Luhmann, 2000). Based on the belief that systems bring with them complexity, there are more possibilities to react to than possible thus extending the limits of human cognition. Trust is an effective mechanism to reduce complexity. However in digital and technological spaces trust performs differently, new forms of technology disrupt the levels of information at our disposal and the cues we use to assess and determine trustworthiness (Erickson, 2000; Knowles, 2018; O'Neill, 2002).

To overcome these difficulties, mechanisms are required to foster trust (Knowles, 2016; Lee, 2004). This means that any mechanisms to reduce complexity and to foster trust must consider how trust functions online, whilst also taking into account the lack of trust in a variety of (new) data sources, by an increasing number of data consumers who do not have a pre-existing level of trust in RDCs, and also by data producers whose interests must be protected. The motivations for mechanisms to foster trust are fourfold: to foster trust in data, in data producers, in data consumers, and in data centres.

3 Methodology

At the beginning of our empirical research, NERC expressed an interest in potential innovations to improve data management and RDC service provision. On the basis of our research, we considered specific mechanisms in the form of blockchain as a vehicle for delivering trust. New distributed ledger technologies have the potential to support the key role of data management, but also regain and build trust through attributes like immutability and the incorporation of smart contracts.

Semi-structured interviews with a range of internal and external stakeholders related to NERC's five data centres were undertaken. The interviewees held senior positions and were affiliated to several of NERC's RDCs. 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 on RDCs and blockchain with a larger audience of stakeholders was conducted, which included participants from different research councils, employees of research data centres, and members

of Lancaster University. At the workshop we presented our initial findings from the interviews and explored potential approaches to solving these challenges. We conducted ethnographic field work during the workshop, capturing the discussions, participants views, and outcomes of workshop tasks. This including taking field notes of discussions, Q&A sessions, presentations and short ad-hoc interviews with participants. In addition to this, we collected data arising from group tasks including post-it notes and flip chart papers. We examined the interview and workshop data and conducted a thematic analysis to identify patterns and overarching themes in our data. Combined, our empirical analysis contributes to the identification of key challenges faced by data centres as discussed in the next section. This is followed by a discussion on trust building mechanisms and the application of blockchain.

4 Findings

Our empirical research highlighted that many of the challenges environmental RDCs face are related to the evolving role of data brokerage as found in the literature, and the ways in which these challenges relate to the need for trust. As discussed, the increase in scrutiny of scientific knowledge, the transition to transdisciplinary research and open access data, and the changing nature of environmental data pose challenges for data centres. Following our empirical research, we analysed the results and have identified and categorised the opportunities to solve these challenges as mechanisms for fostering trust.

4.1 Standardisation

Environmental science data is often heterogeneous and there are few if any existing standards for data formats, labelling and aggregation. Our research found that this was the case for the environmental RDCs we studied. The feedback we received found a consensus for approaching standardisation first, before moving on to other challenges:

> "The main challenge is coming up with common vocabularies and nomenclatures that enable you to search across the data. [...] Then there's the issue of if you bring those data together in some aggregated dataset over the history of many decades, you will have brought together many datasets. There is then a standardisation issue about 'how did you bring that together?"

This is all the more important in the context of trust. Standards guarantee the objectivity of (positivist) scientific research insights and tangible results, as well as enabling reproducibility. Standards and continuity both reduce complexity and contribute towards trustworthiness (Clark, 2014; Knowles et al., 2015; Luhmann, 2000). Standardisation-could increase efficient data management for RDCs. The formalisation of standards also provides a foundation for a shared infrastructure (Borgman, 2015), therefore paving the way for the remaining challenges and mechanisms of trust. RDCs can play a role in the standardisation of methods and terminologies. However, this is time and labour intensive (Edwards, 2010). In this scenario, standardisation requires a conversation across scientific disciplines, including multiple stakeholders. In order for these standards to be implemented and accepted, data centres must be trusted by data producers and data consumers. Standard formats, terms, and methods for aggregation should foster trust by data consumers, in particular those from transdisciplinary backgrounds.

4.2 Supplementary Information

Supplementary information, in the form of meta-data, unique identifiers, or rich narrative, was identified as another mechanism to foster trust. Dependent on the origin and processing techniques used, amongst other things, data can have biases, ambiguities, and inaccuracies and can therefore carry inherent uncertainty (Lukoianova, 2014). For scientists and other data consumers, supplementary information would allow them to understand the complexities and narratives behind the data. Likewise, for those working in policy, gaining an understanding of the complexities and values of data would enable them to handle the uncertainty within the science-policy interface processes more effectively (Guimarães Pereira, 2006). Our findings indicated this:

> "Trust is an issue. I suppose if you step back from it, you would say it was a sensor, you bought it from a certain provider, it gets installed in a location and then it streams data to you, what is there to go wrong? You ought to trust it. In reality, they don't run perfectly. [...] So, the narrative that goes with that is going to be fairly critical."

In order to be able to trust the data, this supplementary information is necessary for certain audiences. Supplementary information will seek to foster trust in data by data consumers, enabling them to utilise data for their own purposes, be this academic research, commercial use, or for policy-making.

4.3 Interactivity

Our research found that interactivity is needed by some data consumers, reflecting the increase in the diversity of the audience. When we are unfamiliar with a situation and have no cues on which to base our trust, an exchange of words or questions asked allows us to begin to place trust (O'Neill, 2002). The development of interactive platforms would allow RDC service provision to evolve and accommodate differing user needs.

This notion of RDCs as service providers was discussed with participants who agreed that the ability to engage in a dialogue with data centres through an interface would be a welcome addition. For instance, a user could input their requirements and receive the relevant data held (or combination of) that may suit their needs. This would benefit users in terms of efficiency, reducing the time spent accessing several portals and manually searching for data. Given the lack of standardisation as discussed above, this may foster trust through ease of use and usefulness (Davis, 1989). Here trust would enable a reduction in complexity for consumers, but not a reduction in system complexity. Thus there is a tradeoff between increased workload in terms of infrastructure maintenance for RDCs on the on hand, and a mechanism fostering trust by data consumers, enabling a reduction in accessibility difficulties.

4.4 Provenance and Traceability

The fourth mechanism identified was provenance and traceability of data, which participants saw as essential to foster trust. Participants argued for better systems to question and gain insight into the journey of data (Bates, 2016). 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. As trust in records is built upon evidence of authenticity and reliability, enabling provenance is key (Sexton et al., 2017). This formalisation of provenance and traceability would also foster trust by data producers in RDCs 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. Evidence of the propagation and distribution 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. Further, in times of increased pressure to improve academic metrics, traceability would provide an additional feature to check how and by whom data was used.

4.5 Management of Stakeholders Interests

Trends towards open data and collaborative production of knowledge have changed relationships among stakeholders and often contribute towards tensions between them (Borgman, 2015). Moreover, the different mechanisms for establishing or maintaining trust are often in competition with one another (Knowles, 2016). The final mechanism for fostering trust is therefore to manage stakeholders' interests appropriately ensuring that the mechanisms, data, and other stakeholders are trusted by other parties.

Provenance is one such mechanism where the impacts of implementation must be considered. Data producers submit their data to data centres, which is then accessed by data consumers. Data producers need to trust in the data centres to preserve their data as supplied, and to ensure that their interests are protected, for instance that their data will not be taken and used inappropriately (Borgman, 2015). Similarly many data consumers need to have trust in the data and consequently trust in the data centres as providers of this data. A mechanism for provenance/traceability would enable data producers to trace where their data has gone to, and would enable data consumers to see where the data has come from. However, this may be undesirable for certain industries who prioritise the protection of business interests and intellectual property. In this instance the formalisation of provenance may result in a paradox whereby greater data quality assurance and trust increases for some data consumers but there is a decrease of engagement and trust for others. This could be problematic for RDCs looking to build their reputation and strengthen their engagement with non-academic consumers, or to increase efficiency by turning data into assets (Birch, 2016). The formalisation of provenance is also potentially problematic as provenance information may be difficult to convey because different audiences have have different needs. This is just one example, but it is an illustration of the fact that any attempt to formalise provenance as a mechanism to foster trust must consider the impacts on various stakeholders.

5 Discussion

As shown in the aforementioned challenges, the need for trustworthiness is often invisible and implicit, and it is only when we specifically draw out the challenges that trust, or the lack thereof, becomes visible. When considering mechanisms to foster trust, we found that the multiplicity of stakeholders equates to a multiplicity of trusts. For instance, provenance was discussed with regards to the implications that this may have on specific sub-sections of the RDC con-

sumer base. It is also the case that interactivity and supplementary information may not be required by all audiences. Whilst these mechanisms are important for those who are not familiar with environmental data, they may be relatively less useful for those who are experienced with it and may also add barriers to access, i.e. additional systems to navigate. Furthermore, with regards to supplementary information it may be difficult to encourage data producers to expend time on producing this additional information when they do not understand the benefits of it, i.e. if they are not data consumers and have not been in the situation of working with data from another field. Therefore, whilst these mechanisms may benefit certain types of data consumers they may not benefit all data consumer nor data producers and therefore this must be considered so as to not discourage the uploading of data or use of data centres.

With regards to trust, it is beneficial to consider the end-to-end nature of trust: to think of trust in and by RDCs as a (non-block)chain. Distributed ledger technologies are interesting as their primary purpose is the formalisation of a chain of immutable transactions (Bhaskaran et al., 2018). This technology could be used to record data entering and exiting RDCs, solving the issue of provenance as well as the supplementary information through automatic enforcement of meta-data formats. However, the issue of traceability is complex. There is a maximum of a first level trace in blockchain. The data can be traced within the chain of transactions, but if it is shared outside the blockchain, traceability is not resolved.

In addition to this, new technology, e.g. smart contracts may be beneficial to the brokerage role of RDCs. Indeed, creating, maintaining and, in some cases repairing trust is an essential element of data centres' daily work. But this may not be easily solved by using novel technologies. It should not be understated that technology does not specifically increase automation in all its forms. These technologies, e.g. blockchain in particular, remain labour and time intensive and could significantly burden RDC employees. Mechanisms and technologies to foster trust were seen to be beneficial by participants to the extent that they did not drastically reduce the ability of data centres to fulfill their role. If these mechanisms do not fit with existing systems or are too complex to navigate, it may be difficult to encourage their acceptance and adoption (Knowles et al., 2015).

Finally, it is all too easy to focus on one specific challenge or mechanism to foster trust. Our findings illustrated that trust is required in data, data centres and data producers by data consumers; trust is vital in data centres and data consumers by data producers; and trust is needed in data producers, data consumers, and other data centres by data centres. This is complex, and it is for this very reason that we suggest these relationships and their intricacies must be mapped out prior to implementing mechanisms to foster trust. The next stage for this work will consider these interactions to a deeper extent, and will seek to conduct further research into the actualities of implementing mechanisms, e.g. to design and apply mechanisms to foster trust and how certain technologies can support those trust mechanisms.

6 Conclusion

In this paper, we looked into the changes in the role of RDCs over time and through a number of factors. When we empirically explored the challenges that environmental RDCs face we were able to identify five mechanisms relating to the nature of trust and intertwined in the practice of data storage, access and re-use. Trust needs to be considered when contemplating the application of novel technological approaches such as blockchain. Whilst blockchain is in many ways a suitable mechanism for fostering trust, e.g. smart contracts and immutability, in many ways it does not solve all of the challenges facing RDCs. What's more, application of this technology may also heavily impact on other mechanisms or stakeholders. The next stages of research are to conduct further work into the interactions of trust and the implementation of trust building mechanismsin real life applications rather than purely theoretical considerations.

We conclude that the challenges faced by data centres need more empirical exploration particularly around the creation and mechanisms of trust. There is no one-fits-all solution, and any implementation must consider the complexities we have detailed. Consequently, we need to discuss and formulate clear prioritisations regarding the role of data centres, espcially as they are a valuable infrastructure within the research-policy interface.

Acknowledgements

This work was supported/funded by the EIDC "EnvChain" NERC Data Innovation funding award (February 2018) and the EPSRC NPIF DTP (Grant Number EP/R512564/1).

We would like to thank all interview and workshop participants for their time and valuable input.

References

Allcock, B. et al. (2002). "Data management and transfer in high-performance computational grid environments". In: *Parallel Computing* 28.(5), pp. 749–771.

Bates J., Lin Y. & Goodale-P. (2016). “Data Journeys: Capturing the Socio-material Constitution of Data Objects and Flows”. In: *Big Data & Society* 3.(2), pp. 1–12.

Bhaskaran, K. et al. (2018). “Double-Blind Consent-Driven Data Sharing on Blockchain”. In: *2018 IEEE International Conference on Cloud Engineering (IC2E)*, pp. 385–391.

Birch, K. (2016). “Rethinking Value in the Bio-economy: Finance, Assetization, and the Management of Value”. In: *Science, Technology, & Human Values* 42.(3), pp. 460–490.

Blair, G.S. (2018). “Complex Distributed Systems: The Need for Fresh Perspectives”. In: *Proceedings of the 38th IEEE International Conference on Distributed Computing Systems (ICDCS)*, pp. 1410–1421.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

Clark, D. (2014). “The Role of Trust in Cyberspace”. In: *Trust, Computing, and Society*. Ed. by R. Harper. Cambridge: Cambridge University Press. Chap. 2, pp. 17–37.

Davis, F.D. (1989). “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: *MIS Quarterly* 13.(3), pp. 319–340.

Edwards, P.N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Erickson T., & Kellogg W.A. (2000). “Social Translucence: An Approach to Designing Systems that Support Social Processes”. In: *ACM Transactions on Computer-Human Interaction* 7.(1), pp. 59–83.

Gambetta, G. (1988). *Trust: Making and Breaking Cooperative Relations*. Oxford: Basil Blackwell.

Guimarães Pereira Â., Guedes Vaz S.-& Tognetti S. (2006). *Interfaces between Science and Society*. Sheffield: Greenleaf Publishing.

Knowles B., & Hanson V.L. (2018). “Older Adults’ Deployment of ’Distrust’”. In: *ACM Transactions on Computer-Human Interaction* 25.(4), pp. 1–25.

Knowles, B. (2016). “Emerging Trust Implications of Data-Rich Systems”. In: *IEEE Pervasive Computing* 15.(4), pp. 76–84.

Knowles, B. et al. (2015). “Models and Patterns of Trust”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW ’15)*, pp. 328–338.

Knowles, B. et al. (2018). “Attending to the Problem of Uncertainty in Current and Future Health Wearables”. In: *Communications of the ACM* 61.(12), pp. 62–67.

Korth H.F. & Silberschatz, A. (1997). “Database Research Faces the Information Explosion”. In: *Communications of the ACM* 40.(2), pp. 139–142.

Lee J.D., & See K.A. (2004). “Trust in Automation: Designing for Appropriate Reliance”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.(1), pp. 50–80.

Luhmann, N. (2000). *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. Vol. 5. Lucius & Lucius.

Lukoianova T., & Rubin V.L. (2014). “Veracity Roadmap: Is Big Data Objective, Truthful and Credible?” In: *Advances in Classification Research Online* 24.(1), pp. 4–15.

Neisse R., Steri G. & Nai-Fovino I. (2017). “A Blockchain-based Approach for Data Accountability and Provenance Tracking”. In: *Proceedings of the 12th International Conference on Availability, Reliability and Security ARES ’17*, 14:1–14:10.

O’Neill, O. (2002). *A Question of Trust: The BBC Reith Lectures*. Cambridge: Cambridge University Press.

Riegelsberger J., Sasse M.A. & McCarthy-J.D. (2005). “The Mechanics of Trust: A Framework for Research and Design”. In: *International Journal of Human Computer Studies* 62.(3), pp. 381–422.

Roy, H. E. et al. (2012). “Understanding Citizen Science & Environmental Monitoring. Final Report on behalf of UK-EOF”. In: *NERC Centre for Ecology & Hydrology and Natural History Museum*.

Sexton, A. et al. (2017). “A Balance of Trust in the Use of Government Administrative Data”. In: *Archival Science* 17.(4), pp. 305–330.

Silvertown, J. (2009). “A New Dawn for Citizen Science”. In: *Trends in Ecology & Evolution* 24.(9), pp. 467–471.

Welpton, R. (2017). “Research Data Centres: The Role of Brokers for Negotiating Access to Data”. In: *Data for Policy 2017: Government by Algorithm?*