Data Curation in the Era of Research Infrastructures

Nana Kwame Amagyei[1] and Elena Parmiggiani[2]

[1] Norwegian University of Science and Technology, Trondheim, Norway
nkamagye@ntnu.no
[2] Norwegian Universy of Science and Technology, Trondheim, Norway
parmiggi@ntnu.no

Abstract. Governments and funding bodies enact policies to set up Research Infrastructures (RIs) to integrate data from different environmental monitoring research sites. The goal of such policies is to publicly open access to research data to support research and innovation and to develop policies for sustaining RIs and the environment at large. However, little is known about the data curation practices of environmental monitoring research scientists at the origins of data. Informed by practice theories on the constitutive entanglement of the social and the material in everyday organisational work, the study zooms-in on the data curation practices utilised by environmental monitoring research scientists to create data that is of good quality and reliable for sharing. Early findings are based on observations and semi-structured interviews of participants in environmental research sites, who collaborate on taking samples of animal and plant species in marine and terrestrial environments. We find that data curation at the origins of data are rife with challenges and opportunities in: data competence, data management and data quality practices. This poster submission concludes with implications of such practices for data governance of large-scale RIs.

Keywords: Data Curation · Research Infrastructures · Environmental Monitoring.

1 Introduction

1.1 Problem and motivation

Two-thirds of the European Union's (EU) economic growth is a result of research and innovation [1]. This accounts for fifteen percent of all productivity gains in Europe [12]. Research Infrastructures (RIs) are a key component of this research and innovation development and they play a key role in the advancement of knowledge, technologies and the development of policies for a sustainable environment. Breakthroughs in research usually comprise combining reliable data in newer and more innovative ways often across disciplines [1]. The relationship between research and innovation - and thus RIs - is usually diverse, complex, and largely dependent on quality data. A critical success factor for producing quality

data, which is usually missing in discussions of long-term and sustainable use of RIs are: data curation practices of humans at the origins of data [2, 3]. In that, by focusing on RIs and their characteristic to make data accessible, emphasis is shifted away from the data curation practices employed by environmental monitoring research scientists at the *"origins"* or at the very early stages *before* data becomes digital. Equally important data curation issues which have received relatively little attention include: the choice of data collection technologies with their material characteristics, the technology's interaction with the research object of interest, the researcher's situated decisions during the process of collecting data including efforts to discover, clean and transform environmental monitoring data into digital data: all of which play a significant role in producing quality data.

Researchers have discovered that several challenges related to interactions between technologies, research objects and methods for producing and managing data unfold throughout the ecological monitoring research process as scientists attempt to understand the environment and create digital data [3, 4, 7]. This work therefore examines one such challenge – *data curation practices at the origins of data* – including how environmental monitoring research scientists investigate and resolve the issues that they face as they collect, clean and transform ecological data into digital data. Such understanding is important for two reasons, first, because data curation practices at the origins of data play a crucial role in producing quality data about the environment [2, 7]and as well, data curation practices at the origins of data involve a great deal of thinking, planning, and preparing, which are necessary for preparing ecological data to become digital data [3, 4, 7]. Consequently, by examining data curation practices at the origins of ecological monitoring research data we intend to contribute to theoretical discussions on data curation as an important concept in RIs that is rife with political, ethical, social and technical issues which are yet to be unearthed in current discussions. Our use of 'origins' attempts to account for the challenge in specifying a distinct moment at which ecological monitoring research phenomena become digital data [2]. This PhD study therefore attempts to answer the questions:

1. *What data curation practices are employed by ecological monitoring research scientists to collect, clean and transform ecological data into digital data?*
2. *How do such data curation practices at the origins of data influence a sustainable RI over the long-term?*

2 Literature Review

2.1 RIs are instantiated in relation to practice

RIs mean different things to different users in different contexts and emerge when situated practices are afforded by large-scale solutions [9]. RIs seek to establish solutions that support situated practices of people, processes, procedures, tools, and technology [5]. They are characterised by "openness to number and types of

users, interconnections of numerous systems, dynamically evolving portfolios of systems and shaped by an installed base of existing systems and practices." [6]. Thus suggesting that focus areas such as existing tools, methods and systems for curating environmental monitoring data, as well as interdisciplinary collaborations and policies on RIs offer benefits, but also bring about new issues to the long-term success of RIs. For example, in environmental monitoring, research and data scientists may be required as part of their job description to continually manage heterogeneous datasets from different sensors and to ensure that digital sensing devices that collect data are appropriately calibrated and ready for algorithmic uses. However, funding bodies may propose requirements for sharing these data with the public or other research institutes. This will further bring about new forms of data curation practices such as newer methods and technologies to add information about the provenance of data to make them more meaningful for sharing. This suggests that the open access to research data agenda present opportunities for understanding the central challenges of data curation in today's era of RIs.

3 Research Case

The European Long-Term Environmental Research (eLTER) in Norway is a network for Norwegian environmental monitoring research sites engaged in long-term, site-based ecological research [11]. eLTER Norway is cordinated by Norwegian Institute for Nature Research (NINA). Its mother organisation, eLTER, has several national sites spread across Europe. Each eLTER site produces and uses heterogeneous forms of data: including data on water bodies, air, temperature, animals and so on, that are collected both manually through scientist observations and recording, and through automated sensing devices or Internet-of-things (IoT) and satellites. Data from all national networks are to be integrated unto the Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research Infrastructure (eLTER RI) with the goal to "provide researchers with access to over 500 sites across Europe and biogeographical regions, to establish and offer harmonised and standardised data, services and training useful to citizens and experts in their joint efforts to find sustainable solutions to the Grand Societal Challenges" [11]. *Forskningsrådet*. The primary focus of this PhD is thus to examine data curation practices in RIs within eLTER Norway. The objective is to establish an empirically based perspective of the social and technical factors that concur in the work of environmental research scientists.

4 Research Design

According to Walsham (2006) interpretive methods of research begin from the assumption that our knowledge of the domain of human action, is a social construction by human actors [9]. Thus, the philosophical approach to be employed in this work is an interpretive paradigm. This is because we are concerned with

understanding the socio-technical context of data curation practices, including: the sensing technologies, algorithms and social processes by which environmental monitoring research scientists work within RIs, and through which they influence and are influenced by RI policies. The participants directly involved in this study (unit of analysis) are environmental monitoring research and data scientists carrying out their day-to-day tasks to make sense of and produce data useful for sharing within RIs.

To address the questions of this study, we will identify informants based on existing connections in eLTER Norway for which the second author has leads and access to. Case studies are helpful to understand the complex social phenomena and to investigate important characteristics of work processes [9]. As a result, the first author has interviewed and observed about twenty-two informants involved with eLTER Norway sites. This will continue over an extended period of time with visits to environmental research stations planned for upcoming months. We utilise observations, semi-structured interviews and ethnographic practices to collect and analyse qualitative data on the day-to-day practices, work processes, and information systems and infrastructures adopted by our informants to prepare, integrate, and make sense of heterogeneous datasets.

5 Early Findings, Implications and Conclusion

Early findings show that environmental monitoring data are first generated through the process of sampling, that is, the practice of taking samples of research objects in their natural environment of habitat. This sampling practice should be understood as having three levels of detail: as a data competence practice, as a data management practice and as a data quality practice. Data competence practice refers to the expertise needed to understand, analyse, communicate and transform environmental monitoring data into digital data. Data management practice refers to efforts to create effective processes and policies that comply with organisational, national, and international regulations. Data quality practice refers to approaches for harnessing the competency and creativity of ecological research scientists in order to balance their use of data management plans and data competences.

These findings have implications for our understanding of data governance for sustaining RIs over the long-term. For example, which particular forms of data governance, including approaches to aggregate data from different environmental monitoring research sites can support data quality practices in RIs? In what ways can the data competencies of environmental research scientists be harnessed to meet open access goals? How do different data management practices utilised by different research sites shape RIs over the long-term?

There is significant wealth of heterogeneous data generated in epistemic fields of study. It is therefore important to recognise and configure data curation as a resource within RIs to facilitate timely responses to environmental issues and develop effective policies for more sustainable societies.

References

1. ESFRI Whitepaper, https://www.esfri.eu/esfri-white-paper. Last accessed 28 Oct 2021.
2. Halfmann, G. (2020). Material Origins of a Data Journey in Ocean Science: How Sampling and Scaffolding Shape Data Practices. In Data Journeys in the Sciences (pp. 27-44). Springer, Cham.
3. Leonelli, S. (2019). Philosophy of biology: the challenges of big data biology. Elife, 8, e47381.
4. Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. Big data society, 1(1), 2053951714534395.
5. Ciborra, C. U., Hanseth, O. (1998). From tool to Gestell: Agendas for managing the information infrastructure. Information Technology People.
6. Monteiro, E., Pollock, N., Hanseth, O., Williams, R. (2013). From artefacts to infrastructures. Computer supported cooperative work (CSCW), 22(4-6), 575-607.
7. Leonelli, S., Tempini, N. (2020). Data journeys in the sciences (p. 412). Springer Nature.
8. Yakel, E. (2007). Digital curation. OCLC Systems Services: International digital library perspectives.
9. Parmiggiani, E., Karasti, H. (2018, August). Surfacing the arctic: politics of participation in infrastructuring. In Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2 (pp. 1-5).
10. Johari, J. (2009). Interpretivism in Information System (IS) Research. Integration Dissemination, 4.
11. eLTER-RI, 2021, https://elter-ri.eu/. Last accessed 28 Oct 2021.
12. The Economic Rationale for Public RI Finding and its Impact, Brussels 2017. https://ri-links2ua.eu/object/document/326/attach/KI0117050ENN$_0$02.*pdf*.