The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

Understanding the 'Intensive' in 'Data Intensive Research': Data Flows in Next Generation Sequencing and Environmental Networked Sensors

Ruth McNally and Adrian Mackenzie,
ESRC Cesagen, Lancaster University

Allison Hui,
David C. Lam Institute for East-West Studies, Hong Kong Baptist University

Jennifer Tomomitsu,
University of the Arts, London

Abstract

Genomic and environmental sciences represent two poles of scientific data. In the first, highly parallel sequencing facilities generate large quantities of sequence data. In the latter, loosely networked remote and field sensors produce intermittent streams of different data types. Yet both genomic and environmental sciences are said to be moving to data intensive research. This paper explores and contrasts data flow in these two domains in order to better understand how data intensive research is being done. Our case studies are next generation sequencing for genomics and environmental networked sensors.

Our objective was to enrich understanding of the 'intensive' processes and properties of data intensive research through a 'sociology' of data using methods that capture the relational properties of data flows. Our key methodological innovation was the staging of events for practitioners with different kinds of expertise in data intensive research to participate in the collective annotation of visual forms. Through such events we built a substantial digital data archive of our own that we then analysed in terms of three traits of data flow: durability, replicability and metrology.

Our findings are that analysing data flow with respect to these three traits provides better insight into how doing data intensive research involves people, infrastructures, practices, things, knowledge and institutions. Collectively, these elements shape the topography of data and condition how it flows. We argue that although much attention is given to phenomena such as the scale, volume and speed of data in data intensive research, these are measures of what we call 'extensive' properties rather than intensive ones. Our thesis is that extensive changes, that is to say those that result in non-linear changes in metrics, can be seen to result from intensive changes that bring multiple, disparate flows into confluence.

If extensive shifts in the modalities of data flow do indeed come from the alignment of disparate things, as we suggest, then we advocate the staging of workshops and other events with the purpose of developing the 'missing' metrics of data flow.

International Journal of Digital Curation (2012), *7*(1), 81–94. http://dx.doi.org/10.2218/ijdc.v7i1.216

 doi:10.2218/ijdc.v7i1.216

Data Flows and their Intensive Properties

This paper reports preliminary findings from research on 'data flow' in the contemporary life sciences. We explore the problem of how to characterise the movement of data in the cases of next generation sequencing (NGS) and environmental networked sensors (ENS). In both settings, there is said to be a 'data deluge' (a term that itself has roots in the early 1990s in the Human Genome Project). Both can also be said to epitomise data intensive research (DIR) (Atkinson & De Roure, 2010; Atkinson et al., 2010b; Hey, Tansley, & Tolle, 2009) in the life sciences. There are undoubtedly very interesting changes going on around bulk movements of data, whether this is referred to as the data deluge, data science, democratising data, open data, data sharing, or DIR. These changes have major implications for science, government, industry and popular culture at every scale, from individuals to global civil society and global climate. Therefore, scientists involved in DIR are calling for better sociologies and geographies of data.[1]

There are a number of ways of thinking about the bulk movement of data, which, it is claimed, will bring about fundamental changes in the nature of science (Anderson, 2008). We could broadly contextualise data flow in the context of the knowledge economy, and for NGS and ENS in particular, within the bioeconomy (OECD, 2009). At the other end of the spectrum, we could track data flow changes in particular settings, for instance, with a particular sequencing platform or a particular kind of environmental sensor (for example, Borgman et al., 2007). However, the key point is that whatever is happening to data flow, the change is not a single change that just happens at one point in time. Rather, changes in movements of data have duration, they have uneven dynamics, and work on many different scales. In this paper, we treat the changes in data mobility associated with NGS and ENS as phenomena to be mapped and understood, but not in terms of a fundamental change in the nature of science, due to an overwhelming quantity of data. We are interested in ways of sensing and making sense of the qualities and relations of people, instruments, infrastructures, conventions and institutions that impel altered modalities of data movement. The notion of data flow draws on the burgeoning sociological field of mobilities studies (Urry, 2000). Much of sociology and geography today is concerned with flows of people and things, and how to make sense of them. Mobilities studies take particular interest in how systematic movements of people, things and information reproduce the social world (Sheller & Urry, 2006). Studying data flows from the perspective of mobilities means thinking about how such flows are relational and performed.

Undoubtedly, there are many extensive changes associated with data deluges or DIR. Indeed, these are the changes most often reported and described: changes in size, volume or amounts of data, databases, servers, processors, or bandwidth; or the number of bioinformaticians, statisticians or data scientists needed to analyse the data. Such changes usually stand in as the main way of comprehending data flow, for example, see the recent IDC report on the zettabyte age (Gantz & Reinsel, 2011). In our study settings, we explore some ways of sensing data flows that are derived more

[1] For example, Alex Szalay, addressing the Data-Intensive Research Workshop at the e-Science Institute in Edinburgh, 15-19 March 2011.

from their *intensive* properties than from their extensive quantities. Intensive properties – a metaphor borrowed from physics – refer to properties of a system that are independent of scale. Such properties are said to be 'scale-invariant properties'; they do not depend on measures of size. In physical systems, extensive changes can be seen as derived from, or even driven by, intensive processes. Indeed changes in intensive properties can account for changes in regime or 'phase shifts'. Hence, intensive properties are deeply implicated in any account of change. If we treat data flow in terms of intensive processes, i.e., processes associated with phase shifts or changes in flow regimes, the question becomes: what is analogous to the role of temperature, pressure and density in data flows? What are the intensive variables or intensive properties in data flows? While we don't have a simple answer to this, we developed research methods that allow relational properties of data flows to be studied, and we analyse and discuss *replicability* (how data flows repeat and propagate), *durability* (the timing, temporalities and coordination of data flows) and *metrology* (how durability and replicability become measurable) as intensive properties of data flows.

Data Flow in Data Intensive Research: Two Scenarios

We selected our cases because they offer contrasting examples of DIR, as illustrated by the following simplified scenarios derived from their respective literatures. In the NGS scenario, data are generated in laboratories and relate to one particular class of biomolecule: the nucleic acids. With the commercialisation of next generation sequencers, genome sequencing has undergone a stepwise increase in speed and volume and a stepwise reduction in cost. In June 2011, 1622 NGS instruments were recorded globally, including 712 in USA, 199 in China and 132 in the UK.[2] The rise in sequencing capacity is 'democratising' sequencing as individual laboratories, and not just large multinational consortia, commission data to address biological questions in projects that they initiate independently (BBSRC, 2011). The availability of NGS data is catapulting sequence data to the forefront of biological experimentation, where it is used to address questions about gene function and regulation, explore genome diversity, and study gene-environment interaction. As a result, biological, biomedical and environmental research are converging on genome sequence data as the main data type (see Hawkins, Hon, & Ren, 2010; Licatalosi & Darnell, 2010; Mardis, 2011; Metzker, 2009; Snyder et al., 2009).

Like NGS, ENS is also named after its data producing instruments, only in this case the instruments are sensors embedded and remotely operating in the wild. In recent times the use of sensors has proliferated as they have become smarter, cheaper and more efficient (with lower energy consumption, and higher data storage and transmission). ENS uses many different types of sensors that directly or indirectly measure a range of environmental variables, gathering meteorological, oceanographic and seismic data, as well as data on river flow, dissolved oxygen concentration, salinity, light levels, temperature, humidity and nutrient flux. Environmental sensors do not operate alone: they are linked together in networks on many scales. At one end of the spectrum are large-scale global networks, such as the Global Seismographic Network; at the other are localised networks with multifunction nodes that monitor a

[2] Next Generation Genomics - World Map of High Throughput Sequencers: http://pathogenomics.bham.ac.uk/hts/

small habitat in great detail. ENS gathers and works with data in a diversity of data formats: digital and analogue, spatial and temporal, alphanumeric and image, fixed and moving (see Collins et al., 2006; Hart & Martinez, 2006; Hamilton et al., 2007; Porter et al., 2009).

NGS and ENS can be said to represent extreme ends of the spectrum in DIR. They designate different sources of data (sequencers, sensors), employ different experimental and analytical approaches, and enjoy different modes and levels of investment. In NGS, a single instrument produces data for many different experiments, whereas in ENS, a single study may deploy many different instruments (sensors) for its sole use. They thus epitomise very different data flow 'topographies', albeit with increasing connections.

Methods for Re-Enacting Data Flows

Our investigation of data flows in NGS and ENS consisted of a mixture of document analysis, observation and exercises using visualisations as provocations. We organised two workshops at the e-Science Institute in Edinburgh with domain and technical experts from the UK and USA. We followed these with a focus group comprised of biologists and environmental scientists from the Lancaster Environment Centre (LEC), none of whom were very involved in high-throughput or large-scale DIR. Our objective was to stage events for the re-mapping, re-measuring and re-visualisation of data flow in NGS and ENS. Our emphasis on practice stems from scholarship in science and technology studies, which subscribes to the notion that methods, objects of analysis and ideas are not separate, but rather entangled and produced together (Barad, 2007; Law, 2004; Mol, 2005; Haraway, 1999). This performative take encourages a more fluid approach to data gathering, with the understanding that methods (ours and those of DIR) *enact* realities at the same time that they attempt to describe them.

Drawing from previous experience of scientists' keen interest and investment in diagrams and data graphics, we sought to harness their expertise in reading such figures. Our key methodological innovation was the *collective annotation* of the visual forms prevalent in the literature in these two fields, such as graphics of data metrics, data volume, data flow, workflow, data integration and fusion. Key questions explored during these annotation exercises related to what is flowing, how it flows, what is rendered invisible, and alternative ways of representing data flow. We also facilitated group discussions based on the results of this shared work on visual forms. Through our research practices we built a substantial electronic archive of presentations, scientific papers, workshop notes, coded transcripts, photographs, annotated visuals and videos of the annotation process. These materials were analyzed in terms of our concerns with durability, replicability and metrology, and provided a narrative of how scientists use metrics to orient themselves to data flows.

Results of the Investigation

Trait 1: Durability

While it may seem banal to recognize that data flows exist in time, examining durability as an intensive property highlights the shifting and competing temporalities of DIR. Data flows are performed through stuttering temporalities, rather than being continuous or ever-present phenomena, as descriptors such as 'the data deluge' might suggest.

The durability of data flows is an implicit concern in DIR wherever collecting, storing, curating, distributing, sharing and archiving data for use and re-use occur. Durability addresses *when* data come into being and the timing of this in relation to other events and temporalities. For data to flow in ENS, the networks have to be ready at the right time for the environment, which yields data in accordance with its own temporalities: the rhythms of seasons, migratory patterns, climate changes and cycles of reproduction. As participants discussed, understanding how the temporality of data flows and environmental events interact is crucial for distinguishing between data generated by real events and those that are artefacts. In NGS, the timing of data collection is more likely to be governed by the time taken for sample preparation, and 'time can be wasted' taking advice on correcting experimental redesign (Focus Group). When biological and environmental studies involve time series data, they are vexed by the timing of sampling. In ENS, for data from different instruments and networks to become integrated and flow forward together, data collection has to be synchronised in time: incorrect time stamps can render data unusable.

> "Time stamps were a big issue – notoriously bad. Sad stories about non-synch'ed datasets." (participant in an ENS workshop)

The durability of data flows are also marked by disciplinary tensions. Although collaborating on the same DIR project, technical and domain experts occupied different 'time zones'. In one ENS project there was only a limited period of time during its 10 years when the network yielded data of use to the environmental scientists. The flow of publications from the technical research preceded the biological ones:

> "The initial period was all about battery life, sensors, networks. They realized in the middle that it was important to keep the human in the loop – that coincided with about two years of useful data [for application scientists]. At the end of that, the technology was mature enough for the application scientists to take it with them and use it. The technology people got bored at this point and moved on to doing mobile applications – kicked environmental scientists out of the loop." (participant in an ENS workshop)

Durability also addresses how data flows change over time: how they endure, not by remaining the same, but by being flexible and adaptive. Architectures of data flow endure amidst constant change (for example, in methods, funding and commercial environments, global collaboration and competition). The disruptions and changes in

research projects are opportunities for data flows to adapt. If not, durability becomes about ephemerality or transience when flow ceases because data are deleted, abandoned or become inaccessible. Versions of this were found in both NGS and ENS.

> "Projects can change from being one type of project into another … People who got grants to do exome capture are now going to complete genomics to get analysis." (participant in a NGS workshop)

> "When you are trying to do some research you have to use the latest technology because it's not deemed sexy otherwise" (Focus Group)

> "It is the Achilles heel of every semantic integration technology that it is not robust with changes. They use the most robust one (in practice). At the moment, in terms of reliable technology, it is not that scalable. The problem is mainly that modifications cause you to have a propagation effect on the mappings." (participant in a ENS workshop)

> "Changes in representation are easier to track, but standards can only really control syntax, they can't control more than that. And what goes on in people's mind is different. Changes in conceptualisation are harder to track because they reside inside humans." (participant in an ENS workshop)

In summary, initiating and sustaining data flow in NGS and ENS is contingent upon the synchronisation of instruments with the temporalities of environments, the synchronisation of data collection across instruments and experiments, and the synchronisation of professional 'time zones'. Moreover, the type of data that flows within a single experiment is liable to change with available technology, and project modifications can disrupt existing data flow infrastructures. Durability in NGS and ENS requires coordination, synchronisation and adaptation of people, things and ideas throughout the flow of data.

Trait 2: Replicability

If durability is about one particular data flow, replicability is about propagating productive solutions to other settings. The growth and expansion of data flows is tied to how their practices and architectures repeat, multiply and increase in number. By investigating replicability, we can bring into focus the conditions of possibility for scalability. What has to be fixed, stabilised or remain the same for something to propagate and scale-up? These facets are not reducible to standard measures of experimental replication.

As with durability, we found domain specific differences between NGS and ENS. The temporal and spatial specificity of environmental data pose severe limits on the replicability of ENS data infrastructures and data flows. In one ENS case, the chronic risk of missing unique data events and the irreplaceability of lost data led to the creation of a fault detection group to monitor data flow in real time. In NGS, by

contrast, replicability is almost too easy, and can undermine the value of existing data, thus interfering with the infrastructures of data durability:

> "Short read sequencing is so cheap, it's a disposable item. It's cheaper to make and analyse your own data than to download someone else's." (participant in a NGS workshop)

Replicating successful data flows is not just about the propagation of high throughput instruments and networks; it is also about infrastructures that propagate practices. If practices are not replicable and standardised, if they remain bespoke and embodied, how will they scale? In the NGS workshop, this aspect of replicability was discussed and debated in relation to the so-called bioinformatics 'bottleneck' (see BBSRC, 2011):

> "Bioinformaticians are doing the same things over and over again. Everyone has to continue reinventing the wheel. Rinse and repeat all over the world." (participant in a NGS workshop)

The use of workflow systems and data analysis pipelines to capture and transfer good laboratory and analytical practices was promoted. However, discussion pointed out the tension between this and the inherently innovative nature of research:

> "Most of these things [workflows] are moving targets – in our experience for mapping and assembly, how often do we change a version of it? Hourly seems to be the response." (participant in a NGS workshop)

> "I don't think we will ever get to fixed workflows. You will never get around to having to write new code for projects. The driver of that is the science. Science has to be novel and therefore cannot reuse whole systems. That novelty is what makes you have to write new bits of code." (participant in a NGS workshop)

There was concern about how to monitor the quality of the particular workflows and pipelines that were being propagated when standardisation was the agreed goal. Moreover, in practice propagation of what was agreed to be a good standard to adopt was found to lead to a proliferation of variants:

> "Well-oiled cogs meshing perfectly would be nice. However, when you look at the proliferation of minimal information checklists, they are domain specific. The result is a kind of Tower of Babel effect at the moment." (participant in a NGS workshop)

The graphs and charts prevalent in NGS associate step-wise increases in data flow with the diffusion of instrument innovation. However, this fails to acknowledge that DIR is intensively collaborative, even on relatively small projects:

> "Can't do this on your own – have to have a massive team – computer scientists, engineers, domain scientists, people to keep spirits up." (participant in an ENS workshop)

Thus, distinct shifts in data flow are not just about adding more instruments, or more efficient instruments, but about enhanced collaboration, and achieving this is challenged by the difficulty of synchronising different disciplines and funding cycles.

> "It generally takes time to demonstrate the efficacy of new methods. No matter how exciting or how personally accepting, you have to clearly demonstrate it works as well as previous methods, or better, and then wait for acceptance from the discipline before you go too far." (participant in an ENS workshop)

Bringing these disparate things together may require systemic changes in order for the collective effort to mesh. An example of what this entails comes from a project dependent upon the participation of amateur ornithologists as human 'sensors':

> "One of our projects – called eBird – is a global project. The concept is to get volunteers to go out and, using fairly standard protocols, collect their observations of birds […] When the project first started, we couldn't get anybody to do that. The notion was that eBird wasn't useful to the volunteers. So eBird needed to change how the volunteers thought about citizen science data. This changed in 2005 with the launch of eBird 2.0. Last Tuesday they collected more data than they did in 2004." (participant in an ENS workshop)

In summary, the meaning and value of replicability in both NGS and ENS is not self-evident. What constitutes too much replicability and too little, and what should and should not be standardised, are questions that have to be negotiated and re-negotiated. Moreover, the relationship between replicability and enhanced data flow is not straightforward. The dramatic increase in data production in eBird 2.0 was the result of a radical redesign of the system and a radical reconfiguration of the (human) sensors as enthusiastic hobbyists rather than worthy citizens. Stepwise increases may require qualitative, systemic change, for example, in the reconfiguration of the network, the forms of collaboration, and epistemic cultures. Finally, the durability of a data flow and its replicability interact in complex ways, sometimes reinforcing and sometimes undermining one another.

Trait 3: Metrology

Metrology is about how data flow and related things are measured, and how these metrics affect what people do. The very notion of DIR is elaborated by references to measures of size, speed, and cost. Diverse data metrics are a constitutive condition of DIR in practice. The size of a dataset, the speed of a network connection, the error rate of a remote field sensor or a sequencing machine are key considerations in making data flow. By describing flows in standard terms (summary numbers, graphs of volume, speed or cost) so that they can be evaluated and taken into account, metrics act as instruments that allow people to see data flow, a flow that otherwise would remain somewhat amorphous and difficult to grasp. Measures of flow are how differences of scale, cost, time and various forms of scientific and practical value are brought together. In a certain sense, metrology makes data flow.

In both NGS and ENS, the explicit use of metrics abounds. Both fields exhibit a 'data-metrics deluge', with metrics attached to the numbers of machines and observations, cost and size of storage and bandwidth, estimates of uncertainty, energy costs, work-time and processing time, and growth rates for all of these things. Novel metrics were also devised to convey the accuracy of sensors, the popularity of data standards, and the benefits of data deposition:

> "Recently an ecologist determined you could more accurately determine the onset of spring through public webcams using green divided by blue than by using remote sensing data." (participant in an ENS workshop)

> "Is there any benefit to having standards? Look at ProteoRED MIAPE satisfaction survey. 95% of people like MIAPE. Papers with data in ArrayExpress get cited more than equivalent papers that don't have data in ArrayExpress." (participant in a NGS workshop)

Analysis of metric talk pointed to how metrics play a role in making data flow. Domain and technical scientists in both fields were aware of growth curves (of costs, time, work, storage and bandwidth) and often acted in relation to them, for example, by attempting to 'keep within the curves' by delaying data collection to wait for the cost curve to shift, or by shifting data management strategies to keep the volume of data beneath available storage space.

At the same time, many discussions, interventions and presentations at the workshops and focus group demonstrated a complicated awareness of metrics that were missing. While presentations often provided common metrics of size and cost, the group annotation of visual images highlighted the metrics that were not represented (see Figures 1-2). For example, in NGS, a common response to the graphs and tables illustrating the falling price of sequencing was to point out the missing costs of bioinformatics. Particularly in cases where participants were unfamiliar with the papers from which images had been drawn, it became clear that the relevance of metrics is context specific. While specific concerns organize visualisations of metrics, these concerns often remain invisible, and this made it difficult for some participants to annotate the images themselves (see Figure 2). Moreover, "visualisation and justification are tied together" (participant in an ENS workshop). This raises questions about how anticipated audiences and common metrics shape some data flows to the exclusion of others.

 doi:10.2218/ijdc.v7i1.216

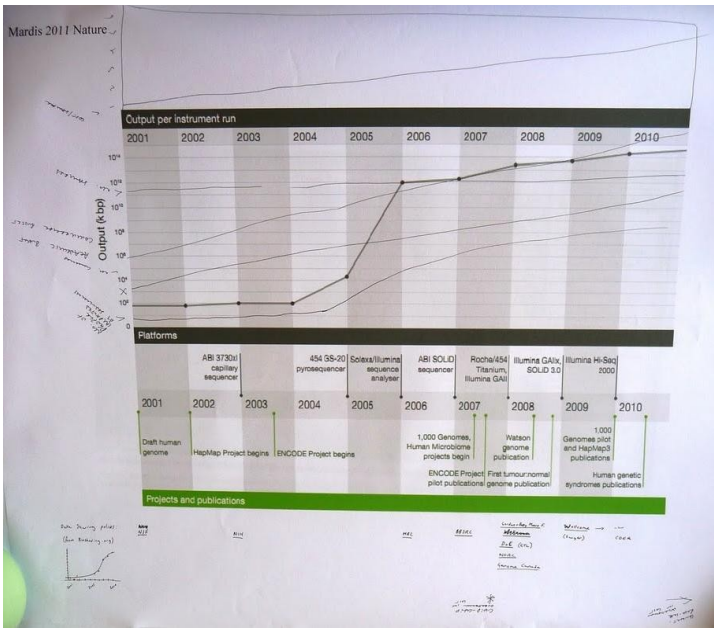

Figure 1. Collective annotation and adding missing metrics (NGS workshop)

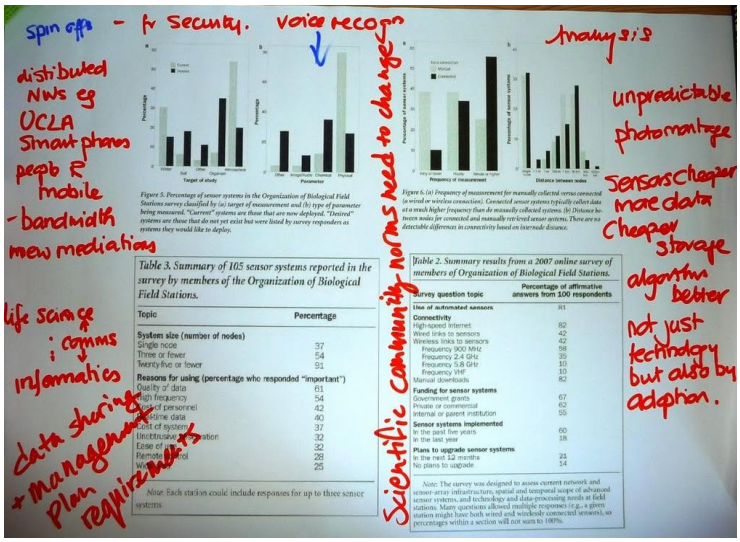

Figure 2. Collective annotation and adding missing metrics (ENS workshop)

In summary, the pervasiveness of metrics in NGS and ENS, and an acute awareness of missing metrics are two sides of the same coin. They both point to how the metrics of data flow are taken into account when making decisions. Thus, data metrics are not only ways of describing data flows, but are often invoked as guides or instruments for change.

Summary and Implications

The changes identified as most desirable for realising the power of DIR are often premised on the replicability and durability of data flows. It is argued that the sometimes stunning success of DIR in special cases needs to be made more durable

rather than transient, and more replicable rather than unique (see Atkinson et al., 2010a). We think this is where the analysis of data flow can play an important role. Our analysis of data flows in NGS and ENS highlights the need to map the features and forms that underpin durability and replicability because these features shape the topographies that condition how data flow. Features of the data topographies we have described that make a difference to flow include the different distributions of instruments, rates of accumulation of data, patterns of coordination and collaboration, and relative openness and closure of various scientific settings to economic, civic and political forces. In a given data topography, the entanglement of different sciences with economic processes and other political and social contexts gives rise to intensive differences.

Successful DIR stages a transition or a 'phase change' of some kind. Such changes result in non-linear changes in metrics, as they enrol new groupings and associations of people and things. A good example is the e-Bird project discussed earlier, and the engagement with human 'sensors' as bird-watching hobbyists rather than as citizen scientists. Extensive changes, that is to say, altered modalities of flow described as scaling up and speeding up, can thus be seen as a result of intensive changes, changes that bring multiple disparate flows into confluence.

Our research with NGS and ENS practitioners explored how they relate to the available metrics, and illustrated how they read metrics in ways that allow them to navigate, steer and coordinate relations between things and people. Metrology provides metrics for sensing and making sense of the relations between these disparate elements. In an important sense, metrics and metrology are the instruments which allow confluences or intensive changes to be brought into view and acted upon. Thus, data metrics not only measure changes in data flow; they are also agents of change that impel altered modalities of data movement. Making and seeing metrics allows one to see what kinds of transformations and changes are involved in marshalling and federating disparate things.

Our findings are based on a small scale study and involving just two examples of DIR. Moreover, NGS and ENS are 'small' science compared to the Australian Square Kilometre Array Pathfinder, CERN's Large Hadron Collider, and astronomy's Pan-STARR's array of celestial telescopes. These experiments are typically characterised in terms of their data 'firehose' instruments and a plethora of metrics. Yet, as with the cases we have studied, these are metrics of the extensive properties of data flow. Moreover, these 'big' science DIR experiments involve hundreds of scientists working in many countries and speaking numerous languages. Future studies with scientists from 'big' science DIRs have the potential to identify many more factors and relationships that condition the durability and replicability of data flow, and bring to light additional 'missing' metrics that are used to make sense of the relations between these disparate elements.

Our findings run counter to the suggestion that DIR can move out steadily and uniformly into new fields through the uptake and adoption of standards and infrastructures, practices and technologies. The implication of our research is that practices, technologies and infrastructures of DIR move unpredictably because of the uneven terrain presented by data flow topographies, and because different parties work with different metrics and ways of bringing them into communication have not been developed. The available surveys, measures and maps, even for a relatively

 doi:10.2218/ijdc.v7i1.216

narrowly defined and highly scrutinised case such as NGS, are rather impoverished and sparse in detail. Data flow topographies that allow people to locate what they are doing and what others are doing are still poorly developed, and there are insufficient data flow metrics that express relations between things for planning and making comparisons.

Our findings on the durability of data flow have implications for the practices of collecting, storing, curating, distributing, sharing and archiving of data. Specifically, they point to the need to foreground and take into account the importance of relationality to the durability of data flow, to flow as enacted. There is a need to attend to the coordination, synchronisation and adaptation of interdependent people, things and ideas that initiate and maintain data flows. This has implications for how data flow is described, and for the kinds of information that are recorded in data provenance efforts. Similarly, a focus on the conditions under which data flows occur might open the possibility of widening the scope of data transparency beyond the provision of open access. The 'virtual witnessing' of data flows, their evaluation and propagation from one experimental setting to another, may require a new 'literary technology' (Shapin & Schaffer, 1985) for DIR, a new way of describing the migration of data that captures the relationalities and interdependencies between the heterogeneous entities that condition its flow. In this way, discussions of data transparency and provenance may develop beyond a concern for origin to incorporate other dynamics of data flow.

Our research has highlighted some topographic features that condition the flow of data, and identified the importance of data metrics that relate these features and aid navigation. However, such metrics themselves only come about through innovations and interventions that develop new ways to express relations between things. Our experience with the process of designing and conducting workshops and other encounters with domain and technical experts suggests to us the potential of this kind of intervention as a way of purposely working on the metrics of data flow across different groups and settings.

Author's Note

McNally and Mackenzie made equal contributions to the design, conduct and analysis of the research and the writing of this manuscript. They were assisted by Hui and Tomomitsu in data gathering and the preparation of the manuscript for publication.

Acknowledgments

This research was undertaken with support from the e-Science Institute, Edinburgh.[3] The support of the Economic and Social Research Council (ESRC) is also gratefully acknowledged. This work is part of the Research Programme of the ESRC Genomics Network at Cesagen.[4] We also acknowledge constructive feedback from three anonymous reviewers.

[3] e-Science Institute: http://www.esi.ac.uk/research-themes/20

[4] Centre for Economic and Social Aspects of Genomics: http://www.genomicsnetwork.ac.uk/cesagen/

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired, 16*(7). Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Atkinson, M., & De Roure, D. (2010). *Realising the power of data-intensive research.* Technical report of the National e-Science Centre: Edinburgh, UK.

Atkinson, M., Kersten, M., Szalay, A., & van Hemert, J. (2010a). *Data-intensive research theme.* Report of the National e-Science Centre: Edinburgh, UK. Retrieved from http://www.esi.ac.uk/files/esi/Theme15-proposal.pdf

Atkinson, M., De Roure, D., van Hemert, J., Jha, S., McNally, R., Mann, B., Viglas, S., & Williams, C. (Eds.) (2010b). *Data-intensive research workshop report.* Report of the National e-Science Centre: Edinburgh, UK. Retrieved from http://dl.dropbox.com/u/3073925/DIRWS.pdf

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning.* London, UK: Duke University Press.

BBSRC. (2011). *BBSRC review of Next Generation Sequencing.* Retrieved from http://www.bbsrc.ac.uk/web/FILES/Reviews/1102-next-generation-sequencing.pdf

Borgman, C.L., Wallis, J.C., Mayernik, M.S., & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. Paper presented at the ACM/IEEE Joint Conference on Digital Libraries 2007, Vancouver, BC.

Collins, S.L., Bettencourt, L.M.A., Hagberg, A., Brown, R.F., Moore, D.I., Bonito, G., Delin, K.A., Jackson, S.P., Johnson, D.W., Burleigh, S.C., Woodrow, R.R., & McAuley, J.M. (2006). New opportunities in ecological sensing using wireless sensor networks. *Frontiers in Ecology and the Environment 4*(8).

Gantz, J.F. & Reinsel, D. (2011). *Extracting value from chaos: IDC digital universe survey.* Retrieved from http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

Hamilton, M.P., Graham, E.A., Rundel, P.W., Allen, M.F., Kaiser, W., Hansen, M.H., & Estrin, D.L. (2007). New approaches in embedded networked sensing for terrestrial ecological observatories. *Environmental Engineering Science 24*(2).

Haraway, D. (1999). Situated knowledges: The science question in feminism and the privilege of partial perspective. In M. Biagioli (Ed.), *The Science Studies Reader* (pp. 172-188). New York: Routledge.

Hart, J. & Martinez, K. (2006). Environmental Sensor Networks: A revolution in the earth system science? *Earth-Science Reviews 78*(3-4).

Hawkins, R.D., Hon, G.C., & Ren, B. (2010). Next-generation genomics: An integrative approach. *Nature Reviews Genetics 11*(7).

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery.* Microsoft Research.

Law, J. (2004). *After method: Mess in social science research.* London: Routledge.

Licatalosi, D.D. & Darnell, R.B. (2010). Applications of Next-Generation Sequencing RNA processing and its regulation: Global insights into biological networks. *Nature Reviews Genetics 11*(1).

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature 470*(7333).

Metzker, M.L. (2009). Sequencing technologies - the next generation. *Nature Reviews Genetics 11*(1).

Mol, A. (2005). *The body multiple: Ontology in medical practice.* (2nd Ed.) London: Duke University Press.

OECD. (2009). *The bioeconomy to 2030: Designing a policy agenda.* Retrieved from http://www.oecd.org/documenten_2649_36831301_42864368_1_1_1_1,00.html#Chapters_abstracts

Porter, J.H., Nagy, E., Kratz, T.K., Hanson, P., Collins, S.L., & Arzberger, P. (2009). New eyes on the world: Advanced sensors for ecology. *BioScience 59*(5).

Shapin, S., Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life.* Princeton: Princeton University Press.

Sheller, M. & Urry, J. (2006). The new mobilities paradigm. *Environment and Planning A 38.*

Snyder, L.A.S., Loman, N., Pallen, M.J., & Penn, C.W. (2009). Next-Generation Sequencing: The promise and perils of charting the great microbial unknown. *Microbial Ecology 57*(1).

Urry, J. (2000). *Sociology beyond societies: Mobilities for the twenty-first century.* London: Routledge.