E-Science Institute: Mini-Theme Final Report

Title of Mini-Theme:	Data Flows in Genomic and Environmental Science: Durability, Replicability and Metrology
Theme Leaders:	Dr Ruth McNally & Dr Adrian Mackenzie
Research Team:	Jennifer Tomomitsu and Allison Hui
Start Date:	January 2011
Duration:	6 months
Report Date:	July 2011
Target Audience:	Those who run next generation sequencing facilities and scientists who use NGS data, genomics tools, platforms and infrastructures in both public and private settings; scientists developing environmental sensor networks and using these data; statisticians, mathematicians, informaticians and computer scientists involved in developing ways of managing, analysing, interpreting, storing and visualising data from data intensive research; digital data curators; Research Councils and other funding bodies.
Acronyms used:	NGS = Next Generation Sequencing ENS = Environmental Networked Sensors DIR = Data Intensive Research CESAGEN = Centre for Economic and Social Aspects of Genomics ESRC = Economic and Social Research Council NCBI = National Centre for Biotechnology Information (USA) SRA = Sequence Read Archive EMBL = European Molecular Biology Laboraboty EBI = European Bioinformatics Institute BBSRC = Biotechnology and Biological Sciences Research Council (UK) CENS = Centre for Embedded Network Sensing (USA) UCSC = University of California Santa Cruz

Executive Summary

e-Science Institute mini-theme

January – July 2011

Data Flows in Genomic and Environmental Science:

Durability, Replicability and Metrology

Ruth McNally and Adrian Mackenzie

ESRC Centre for Economic and Social Aspects of Genomics, Lancaster University

Addressing the Data-Intensive Research (DIR) Workshop at the e-Science Institute (e-SI) in Edinburgh in March 2010, Alex Szalay called for research towards a 'sociology of data'. A year later, we took up this challenge when, from January until July we ran an e-SI mini-theme on data flows in genomic and environmental science. Our focus on the 'flow' of data picks up on the prevalence of watery metaphors (deluge, tsunami, drowning, sea) attached to the movement of Big Data. Our chosen case studies epitomise two very different data flow topographies: in Next Generation Sequencing (NGS) for genomics, highly parallel sequencing facilities generate large quantities of sequence data; in Environmental Networked Sensors (ENS) loosely networked remote and field sensors produce streams of different data types.

Drawing on the sociological field of mobilities studies, we conceptualise data flow as being relational and performed. Our starting premise was that whatever is happening in data flow, it is not a single change happening at just one point in time, but that changes in the movement of data have duration, they have uneven dynamics and they work on many scales. We devised methods for sensing and make sense of the qualities and relations of people, things, places and ideas that impel altered modalities of data flow. Our key methodological innovation was the staging of events for practitioners with different levels and kinds of expertise in data intensive research to participate in the 'collective annotation' of visual forms. We built a substantial digital archive of data based on scientific papers, workshop and focus group notes and presentations, coded transcripts, photographs and videos. We analysed our data with respect to three related traits, or intensive qualities, of data flow: durability (the timing, temporalities and coordination of data flows); replicability (how data flows propagate), and metrology (how durability and replicability and other traits of data flow become measurable).

The findings from a preliminary analysis of our data are that studying data flow with respect to these three traits provides better insight into how doing DIR involves the coordination of many different people, things, places, knowledge and institutions. These disparate elements are the features and forms that shape the topographies of data flow and condition when and whether non-linear changes take place. We argue that whilst much attention is given to phenomena such as the scale, volume and speed of data in DIR, these are metrics of what we call 'extensive' changes in data flow rather than its intensive ones. Our conclusions are 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. The key role that metrology plays is the provision of metrics for sensing and making sense of the relations between these disparate elements. Thus, data metrics are not only a measure of change in data flow, but are also instruments that impel altered modalities of data movement. The making of metrics is a change-making process. However, available maps and metrics of data flow lack sufficient detail. 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 enriching the topographic maps and developing the missing metrics of data flow across different groups and settings as a way to lead to changes in e-Science.

Objectives

The key objectives of this theme were:

- to develop an awareness of the problems, obstacles, friction points or gaps that hinder transformations or reshaping of data flows to do better e-science;
- to identify practices and devices in the conduct of e-science that sustain collaborative development;
- to develop an awareness of some alternative ways of thinking about data flows in genomics and environmental sciences;
- to develop alternative socio-technical models that open up new avenues for interdisciplinary collaboration on devices and practices for research with high-throughput data flows.

Chronology of Events [1]

i) Workshop: Data Flows in Next Generation Sequencing

Venue: E-Science Institute, Edinburgh
Date: March 16, 2011
Event URL: http://www.nesc.ac.uk/esi/events/1198/#registration
Event wiki: http://www.nesc.ac.uk/esi/events/1185/
Presentations: http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=1185
Number attended: 31
Number of speakers: 8
Description: As a multidisciplinary enterprise, NGS requires collaboration between domain experts and technical experts and also entails experiments and experimental forms of actions. The workshop explored the arrangements that bring specific application and technical areas together for this e-science. Participants were invited from centres including the Wellcome Trust Sanger Institute, the BBSRC Genome Analysis Centre (TGAC), EMBL's European Bioinformatics Institute (EBI), Eagle Genomics Ltd., the University of Edinburgh and the James Hutton Institute (formerly Scottish Crop Research Institute). Eight speakers identified the challenges posed by data generation and circulation in NGS.

The agenda interspersed expert presentations with linked group activities. Preparation for the workshop began with analysis of documents for the recent literature on NGS. This analysis was used to inform the selection of speakers and participants, the organisation of the agenda and the focus and design of the group activities.

The challenges of assembling, moving, analysing and storing NGS data were a central focus in a number of presentations. This led to further discussions concerning data integration and reduction, and accounting for error and loss as data are moved around. Moreover, the durability of data in regards to storage (i.e. what to keep and why) was flagged as a concern, raising further questions about the durability of the data archives themselves, especially in the light of the recent announcement by the NCBI to phase out funding for its Sequence Read Archive (SRA).[2] A number of speakers presented workflow platforms, architectures and computer management systems for data management and analysis. This brought into focus points of collaborative tension between the developers of generic software systems on the one hand, and frontline bioinformaticians on the other, who write bespoke code in response to the dynamic research agendas of their local domain scientists.

The group activities explored four main data flow tropes: metrics of speed, cost and size; workflows and pipelines; and genome assemblies that were explored in break outs.

[1] Section drafted by Tomomitsu and Hui
[2] http://www.ncbi.nlm.nih.gov/About/news/09may2011.html

Participants were given images of maps, flowcharts and pipelines and were asked to re-map and re-visualize NGS data in specific areas. These exercises tackled questions such as: where does the data flow and where do they themselves fit on the map? What is flowing? What is rendered invisible? Is there a standard way of representing workflows? Are there other visual forms? We also facilitated group discussions based on the results of the shared work on visual forms.

The documents analysed, the workshop presentations, the annotated posters, and transcriptions of the discussions formed a digital data archive which was later analysed and used to inform the design of the focus group (see below).

ii) Workshop: Data Flows in Environmental Networked Sensors

Venue: E-Science Institute, Edinburgh
Date: 18-19 May 2011
Event URL**:** http://www.nesc.ac.uk/esi/events/1198/
Event wiki: http://wiki.esi.ac.uk/Data_Flows_in_Environmental_Networked_Sensors
Presentations: http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=1198 (not yet uploaded)
OR use http://www.nesc.ac.uk/esi/events/1198/
Number attended: 23
Number of speakers: 8
Description: Making and using data from environmental sensor networks requires collaboration between members of the public as well as domain experts and technical experts. Participants and presenters at this workshop were invited from a range of public and private institutions, including the eBird project (Cornell), Centre for Embedded Network Sensing (CENS, UCLA), Lancaster Environment Centre and the Scottish Environment Protection Agency . Eight invited speakers spoke about and stimulated discussion on the challenges and potential of data flows.
The agenda interspersed expert presentations with linked group activities. Preparation for the workshop began with analysis of documents for the recent literature on ENS. This analysis was used to inform the selection of speakers and participants, the organisation of the agenda and the focus and design of the group activities.

Several speakers spoke about their involvement in collaboratively developing and deploying sensor networks. While these networks produced data supporting new understandings of the natural world, many difficulties arose when dealing with these data. There were challenges of data integration – bringing together often large sets of data from different sensors – and data conflation, which occurs when dataset are in different formats. Just what data are measuring can also pose a challenge, because while sensors can productively measure proxies of natural phenomena, they can also produce data that are an artefact of the sensors themselves. The maintenance and management of sensor networks therefore becomes a key challenge, which requires its own technical tools. In addition to detecting faults within networks, determining data priorities and cleaning data to eliminate 'noise' are crucial processes. As speakers acknowledged, the complicated and collaborative process of data production can lead to unanswered questions about who has ownership and responsibility for data, as well as to significant practical challenges surrounding the sharing of data. The interface of domain and technical expertise was shown to be both crucial to the success of ENS projects and potentially precarious and contested. Technical and domain researchers often work within different paradigms, and can have different assessments of why networked sensors are important and what they should look like in the future.

The group activities took place in breakouts where the participants were asked to engage collaboratively and visually with issues surrounding data flows. As in the NGS workshop, participants were presented with posters containing visualisations from journal articles on ENS, and asked to discuss and annotate how the visualisations spoke to key questions: Where are sensor networks? How can we think about relevant geographies of data? What shapes visualisations of revolutions, generations and metrics in ENS data? How are the instruments of sensor networks represented? What is an instrument and what is a network? How do work flows represent data integration in ENS? How is data transformed and aligned? These breakout exercises highlighted

the lack of realism in many visualisations of ENS data, as well as questions regarding the role of different parties (domain experts, technical experts, funders) in shaping the past and future of ENS data flows.

As for the NGS workshop, the documents analysed, the workshop presentations, the annotated posters, and transcriptions of the discussions formed a digital data archive which was later analysed and used to inform the design of the focus group (see below).

iii) Focus Group: Data Flows in NGS & ENS

Venue: Lancaster University
Date: June 22, 2011
Number attended: 11
Description: The objective of the focus group was to explore data flows with those who are interested but less involved in NGS and ENS. Drawing from data collected at the previous NGS and ENS workshops, two possible scenarios of data flows were presented using a montage of images from some of the workshop presentations. These scenarios were simplifications that ignored problems or contradictions, and instead presented in a somewhat evangelical way, a picture of large-scale change. The focus group consisted of biologists and environmental scientists, none of whom are very involved in high-throughput or large-scale data intensive research. Participants were asked to first individually annotate slides, and then the discussion was opened up to a general focus group-style discussion on the implications of the scenarios for their work. The discussions from scientists in different domains was very useful because it brought to light more general issues about how uncertainties are dealt with, what counts as good data, and the role of models and instruments. There was also keen interest from some scientists to hear how problems were dealt with in other domains.

iv) Public Lecture: "Data Flows in Genomic and Environmental Science: Durability, Replicability and Metrology"

Venue: e-Science Institute, Edinburgh
Date: June 28, 2011
Presented by: Ruth McNally & Adrian Mackenzie
Event URL: http://www.nesc.ac.uk/esi/events/1214/
Podcast: http://www.esi.ac.uk/meetings/1214/videos/4830

Research Outputs

i) Research paper submitted for publication. Adrian Mackenzie, Ruth McNally, Jennifer Tomomitsu, Allison Hui (2011) Understanding the 'intensive' in 'data intensive research': data flows in Next Generation Sequencing and Environmental Networked Sensors. *International Journal of Digital Curation*. Under peer review.
ii) Research paper submitted for oral presentation. Adrian Mackenzie, Ruth McNally, Jennifer Tomomitsu, Allison Hui (2011) Understanding the 'intensive' in 'data intensive research': data flows in Next Generation Sequencing and Environmental Networked Sensors. At "7th International Digital Curation Conference, Public? Private? Personal? Navigating the open data landscape", 5-7 December 2011, Bristol, UK.
iii) Invited presentation. Adrian Mackenzie, Ruth McNally. 'This is not a heatmap', *Data diversities Conference*, Max Planck Institute for History of Science, Humboldt University, Berlin, November 2011
iv) Collaboration. July 2011. Adrian Mackenzie and Ruth McNally in teleconference with Jenny Reardon, Associate Professor, University of California Santa Cruz, about collaboration around big data in genomic science, especially in the context of UCSC Genome Browser and EMBL-EBI/Sanger Centre Ensembl Genome Browser. Further meetings are planned that will also include participation by scientists.

v) Research proposal writing workshop. July 2011. Adrian Mackenzie and Ruth McNally are taking the research on data flows in Next Generation Sequencing forward as Co-Investigators in a multinational research proposal on data practices, to be submitted to the European Research Council 'Open Research Area in Europe for the Social Sciences' funding programme with partners in the UK, Netherlands and France.

Main Section

Executive Summary

Addressing the Data-Intensive Research (DIR) Workshop at the e-Science Institute (e-SI) in Edinburgh in March 2010, Alex Szalay called for research towards a 'sociology of data'. A year later, we took up this challenge when, from January until July we ran an e-SI mini-theme on data flows in genomic and environmental science. Our focus on the 'flow' of data picks up on the prevalence of watery metaphors (deluge, tsunami, drowning, sea) attached to the movement of Big Data. Our chosen case studies epitomise two very different data flow topographies: in Next Generation Sequencing (NGS) for genomics, highly parallel sequencing facilities generate large quantities of sequence data; in Environmental Networked Sensors (ENS) loosely networked remote and field sensors produce streams of different data types. Drawing on the sociological field of mobilities studies, we conceptualise data flow as being relational and performed. Our starting premise was that whatever is happening in data flow, it is not a single change happening at just one point in time, but that changes in the movement of data have duration, they have uneven dynamics and they work on many scales. We devised methods for sensing and make sense of the qualities and relations of people, things, places and ideas that impel altered modalities of data flow. Our key methodological innovation was the staging of events for practitioners with different levels and kinds of expertise in data intensive research to participate in the 'collective annotation' of visual forms. We built a substantial digital archive of data based on scientific papers, workshop and focus group notes and presentations, coded transcripts, photographs and videos. We analysed our data with respect to three related traits, or intensive qualities, of data flow: durability (the timing, temporalities and coordination of data flows); replicability (how data flows propagate), and metrology (how durability and replicability and other traits of data flow become measurable). The findings from a preliminary analysis of our data are that studying data flow with respect to these three traits provides better insight into how doing DIR involves the coordination of many different people, things, places, knowledge and institutions. These disparate elements are the features and forms that shape the topographies of data flow and condition when and whether non-linear changes take place. We argue that whilst much attention is given to phenomena such as the scale, volume and speed of data in DIR, these are metrics of what we call 'extensive' changes in data flow rather than its intensive ones. Our conclusions are 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. The key role that metrology plays is the provision of metrics for sensing and making sense of the relations between these disparate elements. Thus, data metrics are not only a measure of change in data flow, but are also instruments that impel altered modalities of data movement. The making of metrics is a change-making process. However, available maps and metrics of data flow lack sufficient detail. 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 enriching the topographic maps and developing the missing metrics of data flow across different groups and settings as a way to lead to changes in e-Science

Motivation for the Research: Towards a Sociology of Data

There are undoubtedly very interesting changes going on around bulk movements of data, whether this is called the data deluge, data driven science, democratising data, or open data. These changes have major implications for science, government, industry and popular culture, at every scale from individuals to global civil society and global climate. The Fourth Paradigm in DIR is

bringing fundamental changes to the organisation and practices of scientific research and the epistemologies of scientific knowledge (Anderson, 2008, Atkinson and De Roure, 2010; Atkinson et al. 2010; Hey, Tansley, & Tolle, 2009). Data intensive research (DIR) with Next Generation Sequencing (NGS) and environmental networked sensors (ENS) takes place in the highly charged and volatile context of consciousness of climate and environmental change, personalisation of medicine, growth of global civil society and citizen science, and the emergence of a bioeconomy. NGS and ENS data flows are caught up in, and indeed crucial to, many of these developments. More broadly, DIR figures as a significant part of a data economy that extends well beyond life sciences, or indeed sciences in general, to include industry, business, media and states. The ways in which people value data, the tensions they experience between different facets of data, and the ongoing development of ways of making sense of and resolving those tensions should be matters of serious attention.

Against this background we took up Alex Szalay's challenge for research on a 'sociology' of data in DIR. [3] In selecting data 'flow' as our focus, we took our cue from the dominant tropes in data talk where there are floods of data, data deluges and tsunamis, and people are drowning in a sea of data. In other words, data is invoked as something that is both wet and on the move.

Conceptual approach: Mobilities studies

Much of sociology and geography today is concerned with flows of people and things, and how to make sense of them. In conceptualising the notion of data flow we draw on the burgeoning sociological field of mobilities studies (Urry, 2000). Mobilities studies take a 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 studies means thinking about how such flows are relational and performed.

Extensive and intensive changes in data flows

The dominant ways of comprehending and expressing the changes associated with data deluges are expressed in terms of 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 (for instance, see the recent IDC report on the zettabyte age (Gantz & Reinsel, 2011). These are what we are calling the 'extensive' changes of data deluges.

However, we are interested in sensing and making sense of the qualities and relations of people, instruments, infrastructures, conventions and institutions that impel altered modalities of data movement. We are interested in changes in the mobility of data as phenomena to be mapped and understood. To draw on a metaphor from physics, our objective is to explore ways of sensing data flows that are derived more from their 'intensive' properties rather than from their extensive quantities. In physics, intensive properties are 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, 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, in undertaking our research for this mini-theme, we developed research methods that allow relational properties of data flows to be studied, and we analyse and discuss *replicability, durability* and *metrology* as intensive properties of data flows.

[3] For example, Alex Szalay, addressing the Data-Intensive Research Workshop at the e-Science Institute in Edinburgh, 15-19 March 2010.

Selecting the cases

Our cases are next generation sequencing (NGS) for genomics and environmental networked sensors (ENS). In both settings, there is said to be a 'data deluge'. (Indeed the term itself has roots in the early 1990s in association with the Human Genome Project (HGP).) Both cases can also be said to epitomise DIR in the life sciences. However, they are very different examples of it as illustrated by the simplified scenarios derived from their respective literatures in Boxes 1 and 2. (below). 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 on nodes) for its sole use. They thus epitomise very different data flow 'topographies', albeit with increasing connections.

Box 1: NGS Scenario

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 (Figure 1). In June 2011, 1622 NGS instruments were recorded globally, including 712 in USA, 199 in China and 132 in the UK.[4] 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 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).

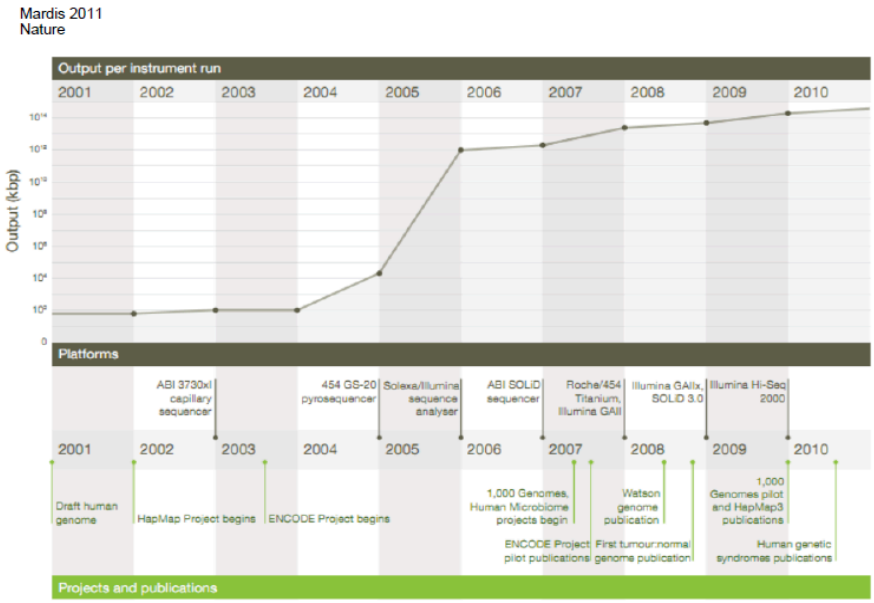

Figure 1 A decade's perspective on DNA sequencing technology (Mardis, 2011).

[4] Next Generation Genomics - World Map of High Throughput Sequencers: http://pathogenomics.bham.ac.uk/hts/ accessed 27 June 2011.

Box 2: ENS Scenario

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 (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, moisture, humidity, respiration and nutrient flux (see. e.g. Figure 2). 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 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).

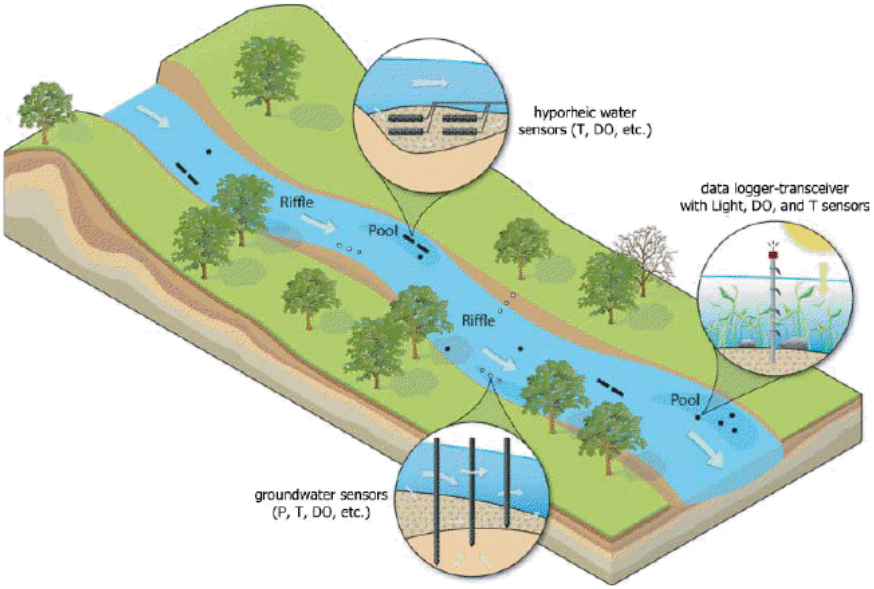

Figure 2 Hypothetical deployment of an aquatic sensor network sampling pool and riffle patterns along a stream course (Rundel et al., 2009).

Research Design

The idea of data flows deeply influenced how we shaped the planning and conduct of this mini-theme. To determine how data moves in NGS and ENS, we began with an analysis of documents from the recent scientific literature in each of these fields (see Appendices 1 and 2). This allowed us to identify main ways of talking about data, delineate key dynamics and problems in the area, generate lists of relevant actors (individual, institutional, commercial) and categorise visual forms (i.e. graphics, diagrams, images) associated with data talk.

Figure 3 Collective annotation (ENS, Tomomitsu)

Our investigation of data flows in NGS and ENS consisted of a mixture of document analysis, observation, and exercises using visualisations as provocations. Our objective was to stage events for the re-mapping, re-measuring and re-visualisation of data flow in NGS and ENS.
In keeping with the theme-based events at e-SI, we organised two workshops in Edinburgh and ran a focus group in Lancaster (see chronology of events for more details). Both workshops were designed around models of data flow derived from the documents analysed. The workshop agendas themselves were organised to reflect pipelines or workflows specific to the field, and we chose contributors who could speak to specific parts of the flow, either because they were involved in building systems, or because they were closely involved with their design and use.

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 *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. During the breakout sessions we organised participants in small groups and supplied them with posters of common graphics and marker pens (see Figures 3 and 4). 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.Our emphasis on practice stems from scholarship in STS 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 on data encourages a more fluid approach to data gathering with the understanding that methods (ours and those of DIR) *produce* realities at the same time that they attempt to describe them.

Figure 4 Collective annotation (NGS, Tomomitsu)

The workshops were specifically utilized as *social science research instruments* that created data. Talks were recorded and partially transcribed, presentations were collected, and breakout groups after each session were organised around the annotation of data flow visualisations. The annotated documents, as well as speaker presentations, were collected and subsequently photographed for analysis. Thus, through our research practices we built a substantial electronic archive of presentations, scientific papers, workshop notes, coded transcripts (see Figure 5), photographs, annotated visuals and videos which became our data.

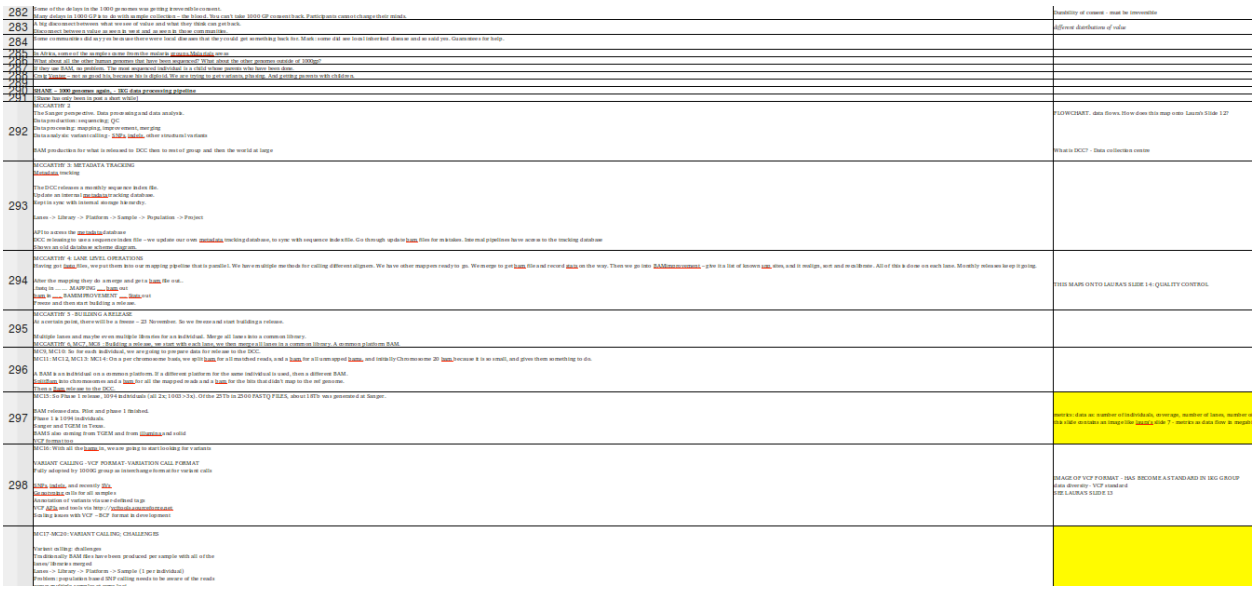

Figure 5 Coded transcript of workshop data

Analytical Framework: Durability, Replicability, Metrology

Our analysis of data flow in NGS and ENS focused on three related traits: durability (the timing of data flows and their temporalities and coordination); replicability (how data flows repeat and propagate), and metrology (how durability and replicability and other intensive traits become measurable).

The changes identified as most desirable for realising the power of DIR are often premised on replicability and durability. That is, 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. Durability is about the temporalities of data flows, how they exist in time. 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 is about when data come into being and the timing of this in relation to other events and temporalities. Durability is also about the architectures of data flow and how they endure amidst constant change (for example, in methods, techniques, infrastructures, funding and commercial environments, global collaboration and competition). Durability is about how data flows change over time; how they endure, not by remaining the same, but by being flexible and adaptive. Durability is also about ephemerality or transience; for instance, when flow ceases because data are deleted, abandoned or become inaccessible.

Replicability is about how the practices and architectures of data flows repeat and multiply, and how they increase in number. Growth and expansion of data flows entails a chain of propagation. What is taken into account; what has to be fixed, stabilised or remain the same for something elso to propagate and grow? These facets are not reducible to standard measures of experimental replication.

Metrology can be understood as the aspect of data flows that is rendered in terms of metrics, in terms of measures and quantities. Metrics are threaded through almost any work and discussion of DIR. 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. Metrology helps shape a sense of 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 open ways that differences of scale, cost, time and various forms of scientific and practical value are brought together. In a certain sense, metrology makes data flow.

Research Results

We have not yet completed analysing our extensive digital archive, but here we present our preliminary results.

Durability

Durability is about how data flows exist in time, the timing of their coming into existence, and how they endure. Looking upstream, we found that when data could start to flow was contingent upon the temporalities of other events and processes, and that these were different in NGS and ENS. 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. 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' (focus group) taking advice on correcting experimental redesign. When they involve time series data, biological and environmental studies are vexed by the timing of sampling, and in ENS the use of time stamps can be problematic. 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."(ENS)

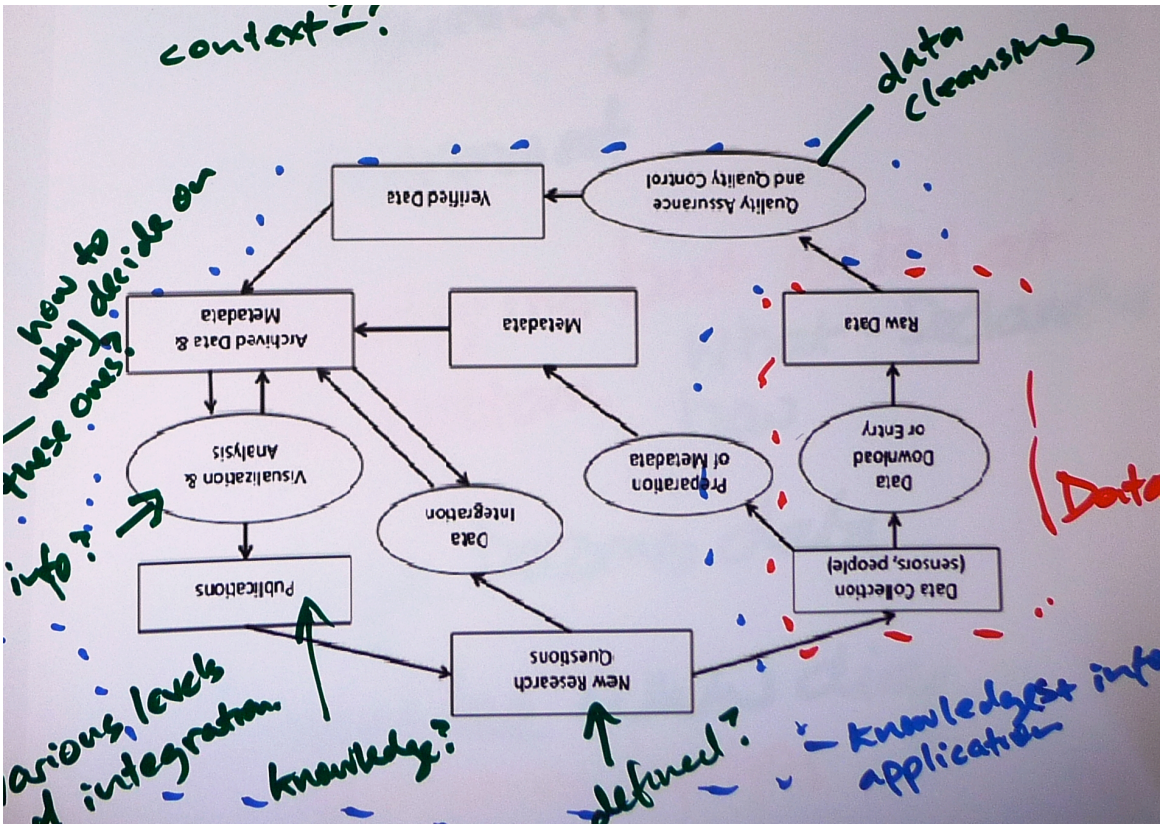

Figure 6 Annotated dataflow (ENS)

Although collaborating on the same DIR project, technical and domain experts were found to occupy different 'time zones' in relation to data flow. In one ENS project there was only a limited period of time, approximately 2 years in a 10 year project, when the network yielded data of use to the environmental scientists. Related to this, the timing 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 2 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." (ENS)

Another finding was that projects themselves change over time. 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." (NGS)

> "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." (ENS)

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. In summary, durability of data flow in NGS and ENS is conditioned by the disparate temporalities of people, things, places and ideas, and maintaining and optimising the flow is about synchronisation and adaptation to change and difference.

Replicability

We found that the meaning a value of the trait of replicability in NGS and ENS was not self-evident, and there were domain specific differences between NGS and ENS. In ENS, the temporal and spatial specificities of the environmental settings pose severe limits on the replicability of data infrastructures and data flows. Replicability is almost a practical impossibility. In one ENS case, the chronic risk of missing unique data events led to the creation of a fault detection group to monitor irreplicable data flows in real time. In NGS, by contrast, replicability of data flow is almost too easy, and can undermine the value and hence the durability of existing data.

> "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."(NGS)

Replicability of data flows is not just about high throughput instruments, it is also about infrastructures and practices and standardisation. 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." (NGS)

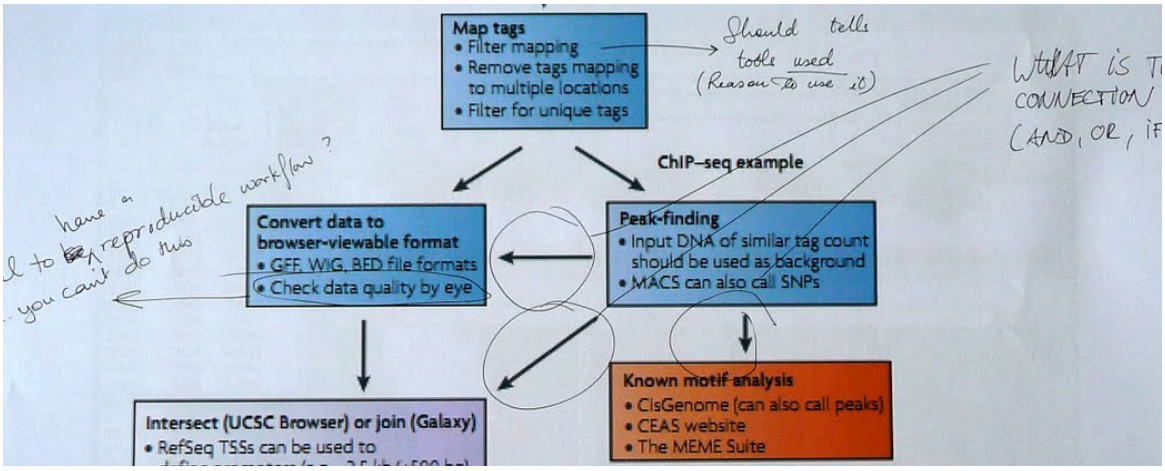

Figure 7 Annotated workflow diagram (NGS)

At the NGS workshop, several presentations described workflow systems and data analysis pipelines designed to capture and propagate good laboratory and analytical practices. However, discussion pointed out the tensions between this kind of replicability and the inherently innovative nature of research practices:

> "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." (NGS)

Replicability *per se* is not always a good thing. It matters which particular practices are being replicated. There was concern about how to monitor the quality of the particular workflows and pipelines to guard against replicating the 'wrong' ones. Moreover, in practice propagation of what

is identified as a good standard to adopt was found to lead to a proliferation of variants. Replicability across different settings always results in difference:

> "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." (NGS)

The graphs and charts of metrics prevalent in NGS associate step-wise increases in data flow with the proliferation of instruments, and more high-throughput instrument (see Figure 1). However, even on relatively small projects, DIR is 'collaboration-intensive' as well as data-intensive.

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

Thus distinct shifts in data flow are not just about adding more instruments, or more efficient instruments, but about intensified 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, have to clearly demonstrate it works as well as previous methods or better and then wait for acceptance from discipline before go too far." (ENS)

Replicability across domains is in tension with the difficulty of synchronising work between different domains, with different priorities, different rates of development and different funding cycles and even different epistemologies. Bringing these disparate things together may require systemic changes in order for the collective effort to mesh. An example of what this entails came from e-Bird, 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." (ENS)

In summary, the meaning and value of replicability in both NGS and ENS is not self-evident. What is too much replicability and too little, and what should and should not be standardised, are questions that have to be negotiated and re-negotiated. And replicability entails change. Moreover, the relationship between replicability and enhanced data flow is not straightforward. Whilst distinctive shifts in scale are related to changes in instrumentation, they are also related to changes in the nature of collaboration. The dramatic increase in data flow 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 in data flow may require qualitative, systemic change, for example, in the reconfiguration of the network, the forms of collaboration, and epistemic cultures. Finally, durability and replicability interact in complex ways, sometimes reinforcing and sometimes undermining one another.

Metrology

Metrology is about how data flow and related things are measured, and how these metrics affect what people do. 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. We also found metrics being used to evaluate sensors, and convey 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." (ENS)
>
> "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." (NGS)

Analysis of metric talk illustrated how metrics not only measure data flow, they play a role in how data flow. Domain and technical scientists in both fields were aware of growth curves (of costs, time, work, storage, 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 (see, for example, Figure 8).

Figure 8 Strategies for handling data growth (Cochrane, 2011)

At the same time, many discussions, interventions and presentations at the workshops and focus group demonstrated a complicated awareness of metrics that were *missing*, as illustrated in the annotations in Figure 9. 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 and other factors that should also be taken into account when planning DIR whether at a laboratory level or on a science policy level.

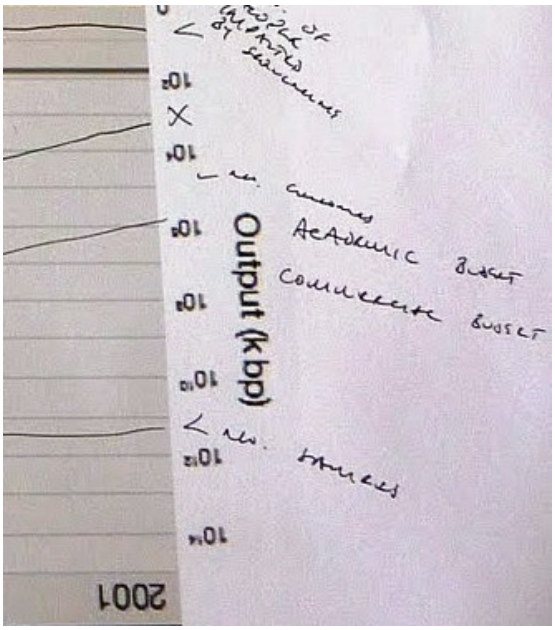

Figure 9 Collective annotation and adding missing metrics (NGS)

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, of communicating a shared sense of change of flow, but are often invoked as guides or instruments for change.

Main Conclusions

1. **The importance of the topographies of data flow**

Our starting premise was that, whatever is happening to data flow, however it is changing, 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. Analysing data flow with respect to durability, replicability and metrology provides better insight into the disparate elements that are the features and forms that give data flow its topographies and condition how it moves.

Features of the data topographies we have described that make a difference to data flow in NGS and ENS include the different types of instruments and their distributions (sequencing centres vs. sensor networks), the different environmental settings and temporalities, the different patterns of coordination and compositions of collaboration, and the relative openness and closure of to economic, civic and political forces

2. **Extensive shifts in data flow consist of multiple changes coming into conjunction**

In addition to differences in topographies between fields, a given data topography is itself an entanglement of differences. 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.

3. **Extensive shifts in data flow may require systemic changes**

Successful DIR stages a transition or a 'phase change' of some kind. Such changes result in significantly non-linear changes in metrics as they enrol new groupings and associations of people and t hings. 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.

4. **Metrology is a change making process**

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. 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 rather than providing a measure of change, the making of metrics through metrology is a change-making process. Making and seeing metrics allows one to see what kinds of transformations and changes are involved in marshalling and federating disparate things.

Research Outcomes

1. **Can you see a way to propagate these outcomes through the community and influence the uptake and adoption of standards, practices and/or e-Science/Grid technologies?**

The use of the word propagation could imply moving out uniformly. However, movement entails change; things change as they move to different settings. The implication of our research is that standards, practices, technologies and infrastructures move unevenly. The uneven and unpredictable uptake and adoption of standards, practices and/or e-Science/Grid technologies is because of the uneven terrain that shapes their data flow topographies, and because different groups relate to different metrics.

2. **In your opinion, what direction(s) should future investigations on this topic take?**

The available surveys, measures and maps, even for a relatively 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. More research is needed to further our understanding of how data does and does not flow and to develop better metrics and maps with which to navigate.

3. **How could it lead to changes in the ways in which we do e-Science?**

Transitions come about through confluence, by bringing differences and disparate things together. Moreover, propagation implies change, and change requires methodological innovation and intervention. A good example of both of these points is the e-Bird project discussed above.

One way to proceed would be to undertake more ethnographic studies of situated practices to learn more about how to bring about change in specific e-science sites.

Another approach is the one we have adopted, and indeed one that has been characteristic of the many events facilitated at the e-SI over the years, which is to stage and experiment with new formats for bringing different types of e-science practitioners from different types of e-science together

Our key contribution to future investigations with the potential to lead to changes is our methodological innovation of organising events for practitioners with different kinds of expertise in data intensive research to participate in the collective annotation of visual forms. The goal of

this approach is to enrich our knowledge about the features that condition the flow of data, and add to the repertoire of metrics that relate these features and aid navigation.

Future Activities

We plan to continue using and developing our methodological approach of collective annotation of visual forms with practitioners with different kinds and levels of expertise and experience in DIR, and in particular with the inclusion of more domain scientists, especially for NGS.

We are in the process of developing some collaborative research with Jenny Reardon, Associate Professor, University of California Santa Cruz. The focus of this will be big data in genomic science, especially in the context of UCSC Genome Browser and the EMBL-EBI / Sanger Institute Ensembl Genome Browser.

Another interest we have is in working on models of data flow that allows the contrasts between ENS and NGS to be grasped, and looking at domains that seem to be bringing ENS and NGS together e.g., NGS as a 'biosensor' for detecting and identifying organisms for environmental metagenomics.

We are keen to widen the notion of data flows into an account of the 'data economy' by studying a spectrum of related domains, some in the biosciences and some outside of them, and especially in fields of DIR that lie across the boundaries between research on the one hand, and providing support for decision-making and intervening on the other.

We are taking the research on data flows in Next Generation Sequencing forward as Co-Investigators in a multinational research proposal on data practices, to be submitted to the European Research Council 'Open Research Area in Europe for the Social Sciences' funding programme with partners in the UK, Netherlands and France.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired,* 16(7). Retrieved July 22, 2011, 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, National e-Science Centre: Edinburgh, UK.

Atkinson, M., De Roure, D., van Hemert, J., Jha, S., McNally, R., Mann, B., Viglas, S., & Williams, C. (Eds.) (2010). *Data-intensive research workshop report*. National e-Science Centre: Edinburgh, UK. Retrieved August 2, 2011, 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: Duke University Press.

BBSRC. (2011). *BBSRC review of Next Generation Sequencing*. Retrieved July 28, 2011, 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. *ACM/IEEE Joint Conference on Digital Libraries 2007*. Vancouver, BC.

Cochrane, G. (2011). *Next generation sequence data archiving in the European Nucleotide Archive.* Paper given at the Data Flows in Next Generation Sequencing: Replication, Durability and Metrology Workshop, March, 16 2011, e-Science Institute, Edinburgh.

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), pp.402-407

Gantz, J.F. & Reinsel, D. (2011). *Extracting value from chaos. IDC digital universe survey.* Retrieved July 25, 2011, 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), pp.192-204.

Hart, J. & Martinez, K. (2006). Environmental Sensor Networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3-4), pp.177-191.

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.

Hawkins, R.D., Hon, G.C., & Ren, B. (2010). Next-generation genomics: An integrative approach. *Nature Reviews Genetics*, 11(7), pp.476–486.

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), pp.75-87.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), pp.198-203.

Metzker, M.L. (2009). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1), pp.31-46.

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 July 28, 2011, from http://www.oecd.org/document/48/0,3343,en_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), pp.385–397.

Rundel, PW et al., 2009. Environmental sensor networks in ecological research. *NEW PHYTOLOGIST*, 182(3), pp.589-607.

Sheller, M. & Urry, J. (2006). The new mobilities paradigm. *Environment and Planning A*, 38, pp.207-226.

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), pp.1-3.

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

Appendix 1: Next Generation Sequencing Literature

Aleksic, J. & Russell, S., 2009. Chiping away at the genome: the new frontier travel guide. *MOLECULAR BIOSYSTEMS*, 5(12), pp.1421-1428.

Anon, Genome.gov | DNA Sequencing Costs. Available at: http://www.genome.gov/sequencingcosts/ [Accessed March 10, 2011b].

Anon, Genome.gov | DNA Sequencing Costs. Available at: http://www.genome.gov/sequencingcosts/ [Accessed March 15, 2011c].

Anon, 2010. The human genome at ten. *Nature*, 464(7289), pp.649-650.

Anon, 2011. The next generation: Using new sequencing technologies to analyse gene regulation - CULLUM - 2011 - Respirology - Wiley Online Library. Available at:

http://onlinelibrary.wiley.com.ezproxy.lancs.ac.uk/doi/10.1111/j.1440-1843.2010.01899.x/pdf [Accessed March 2, 2011].
Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. *NEW BIOTECHNOLOGY*, 25(4), pp.195-203.
BBSRC 2011. 'BBSRC review of next generation sequencing'. February. http://www.bbsrc.ac.uk/sequencingreview/
Braeutigam, A. & Gowik, U., 2010. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *PLANT BIOLOGY*, 12(6), pp.831-841.
Brion, M. et al., 2010. New technologies in the genetic approach to sudden cardiac death in the young. *Forensic Science International*, 203(1-3), pp.15-24.
Costa, V. et al., 2010. Uncovering the Complexity of Transcriptomes with RNA-Seq. *JOURNAL OF BIOMEDICINE AND BIOTECHNOLOGY*.
Courtney, E. et al., 2010. Transcriptome profiling in neurodegenerative disease. *Journal of Neuroscience Methods*, 193(2), pp.189-202.
Ekblom, R & Galindo, J, 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*. Available at: http://www.nature.com/doifinder/10.1038/hdy.2010.152 [Accessed March 3, 2011].
Engstrand, L., 2009. How will next-generation sequencing contribute to the knowledge concerning Helicobacter pylori? *CLINICAL MICROBIOLOGY AND INFECTION*, 15(9), pp.823-828.
Flicek, P. & Birney, E., 2009. Sense from sequence reads: methods for alignment and assembly. *NATURE METHODS*, 6(11, Suppl. S), p.S6-S12.
Fouse, S.D., Nagarajan, R.P. & Costello, J.F., 2010. Genome-scale DNA methylation analysis. *EPIGENOMICS*, 2(1), pp.105-117.
Hawkins, R.D., Hon, G.C. & Ren, B., 2010. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7), pp.476–486.
Johansen, S.D. et al., 2010. Approaching marine bioprospecting in hexacorals by RNA deep sequencing. *NEW BIOTECHNOLOGY*, 27(3, Sp. Iss. SI), pp.267-275.
Kim, S. & Kim, J.-S., 2010. Targeted genome engineering via zinc finger nucleases. *Plant Biotechnology Reports*, 5(1), pp.9-17.
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), pp.75-87.
Mardis, E.R., 2008. Next-generation DNA sequencing methods. *ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS*, 9, pp.387-402.
Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), pp.198-203.
Marguerat, S. & Baehler, J., 2010. RNA-seq: from technology to biology. *CELLULAR AND MOLECULAR LIFE SCIENCES*, 67(4), pp.569-579.
Medina, M. & Sachs, J.L., 2010. Symbiont genomics, our new tangled bank. *GENOMICS*, 95(3), pp.129-137.
Metzker, M.L., 2009. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46.
Meyerson, M., Gabriel, S. & Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *NATURE REVIEWS GENETICS*, 11(10), pp.685-696.
Miller, J.R., Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *GENOMICS*, 95(6), pp.315-327.
Milne, I. et al., 2010. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3), pp.401 -402.
Morozova, O., Hirst, M. & Marra, M.A., 2009. Applications of New Sequencing Technologies for Transcriptome Analysis. *ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS*, 10, pp.135-151.
Neale, D.B. & Kremer, A., 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, 12(2), pp.111-122.

Nowrousian, M., 2010. Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems. *EUKARYOTIC CELL*, 9(9), pp.1300-1310.
Pérez-Enciso, M. & Ferretti, L., 2010. Massive parallel sequencing in animal genetics: wherefroms and wheretos. *Animal Genetics*, 41(6), pp.561-569.
Petty, N.K., 2010. Genome annotation: man versus machine. *Nature Reviews Microbiology*, 8(11), pp.762-762.
Plenge, R., 2010. GWASs and the age of human as the model organism for autoimmune genetic research. *GENOME BIOLOGY*, 11(5).
Raymond, F.L. et al., 2010. Molecular prenatal diagnosis: the impact of modern technologies. *PRENATAL DIAGNOSIS*, 30(7), pp.674-681.
Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24-26.
Roh, S.W. et al., 2010. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *TRENDS IN BIOTECHNOLOGY*, 28(6), pp.291-299.
Singleton, A.B. et al., 2010. Towards a complete resolution of the genetic architecture of disease. *TRENDS IN GENETICS*, 26(10), pp.438-442.
Snyder, L.A.S. et al., 2009. Next-Generation Sequencing-the Promise and Perils of Charting the Great Microbial Unknown. *MICROBIAL ECOLOGY*, 57(1), pp.1-3.
Stapley, J. et al., 2010. Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, 25(12), pp.705-712.
Stein, L.D., 2010. The case for cloud computing in genome informatics. *Genome Biology*, 11(5), p.207.
Tucker, T., Marra, M. & Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics*, 85(2), pp.142-154.
Varshney, R.K. et al., 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. *TRENDS IN BIOTECHNOLOGY*, 27(9), pp.522-530.
van Vliet, A.H.M., 2010. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS MICROBIOLOGY LETTERS*, 302(1), pp.1-7.

Appendix 2: Environmental Networked Sensors Literature

Allen, M.F. et al., 2007. Soil sensor technology: life within a pixel. *Bioscience*, 57(10), pp.859–867.
Barseghian, D. et al., 2010. Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *ECOLOGICAL INFORMATICS*, 5(1), pp.42-50.
Benson, B. et al., 2010. Perspectives on next-generation technology for environmental sensor networks. *FRONTIERS IN ECOLOGY AND THE ENVIRONMENT*, 8(4), pp.193-200.
Broring, A. et al., 2011. New Generation Sensor Web Enablement. *SENSORS*, 11(3), pp.2652-2699.
Buratti, C. et al., 2009. An Overview on Wireless Sensor Networks Technology and Evolution. *SENSORS*, 9(9), pp.6869-6896.
Collins, S. et al., 2006. New opportunities in ecological sensing using wireless sensor networks. *FRONTIERS IN ECOLOGY AND THE ENVIRONMENT*, 4(8), pp.402-407.
Dardari, D. et al., 2007. Mathematical evaluation of environmental monitoring estimation error through energy-efficient wireless sensor networks. *IEEE TRANSACTIONS ON MOBILE COMPUTING*, 6(7), pp.790-802.
Farre, M. et al., 2009. Sensors and biosensors in support of EU Directives. *TRAC-TRENDS IN ANALYTICAL CHEMISTRY*, 28(2), pp.170-185.
Hamilton, MP et al., 2007. New approaches in embedded networked sensing for terrestrial ecological observatories. *ENVIRONMENTAL ENGINEERING SCIENCE*, 24(2), pp.192-204.
Hart, J. & Martinez, K., 2006. Environmental Sensor Networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3-4), pp.177-191.

Hasselquist, N.J., Vargas, R. & Allen, M.F., 2010. Using soil sensing technology to examine interactions and controls between ectomycorrhizal growth and environmental factors on soil CO 2 dynamics. *Plant and Soil*, 331(1), pp.17–29.

Ingelrest, F. et al., 2010. SensorScope: Application-specific sensor network for environmental monitoring. *ACM Transactions on Sensor Networks (TOSN)*, 6, pp.17:1–17:32.

Majidi, V. & Hassell, C., 2004. Miniaturized instrumentation for field applications: General considerations for environmental sensor networks. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL ANALYTICAL CHEMISTRY*, 84(14-15), pp.1111-1121.

Maurice, P. & Harmon, T., 2007. Environmental embedded sensor networks. *ENVIRONMENTAL ENGINEERING SCIENCE*, 24(2), pp.149-150.

Navarro, D. et al., 2010. Hardware and Software System-Level Simulator for Wireless Sensor Networks. *EUROSENSOR XXIV CONFERENCE*, 5, pp.228-231.

Porter, J.H. et al., 2009. New eyes on the world: advanced sensors for ecology. *BioScience*, 59(5), pp.385–397.

Quinn, N. et al., 2010. Use of environmental sensors and sensor networks to develop water and salinity budgets for seasonal wetland real-time water quality management. *ENVIRONMENTAL MODELLING & SOFTWARE*, 25(9), pp.1045-1058.

Rodriguez-Mozaz, S., Lopez de Alda, M.J. & Barceló, D., 2006. Biosensors as useful tools for environmental analysis and monitoring. *Analytical and Bioanalytical Chemistry*, 386(4), pp.1025-1041.

Rundel, PW et al., 2009. Environmental sensor networks in ecological research. *NEW PHYTOLOGIST*, 182(3), pp.589-607.

Tong, B. et al., 2010. Towards Reliable Scheduling Schemes for Long-lived Replaceable Sensor Networks. *2010 PROCEEDINGS IEEE INFOCOM*. Available at: http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=5&SID=W1aF@ibjoeb1C5iobc3&page=1&doc=10 [Accessed April 5, 2011].

Wang, B., Lim, H. & Ma, D., 2009. A survey of movement strategies for improving network coverage in wireless sensor networks. *COMPUTER COMMUNICATIONS*, 32(13-14), pp.1427-1436.

Yick, J., Mukherjee, B. & Ghosal, D., 2008. Wireless sensor network survey. *COMPUTER NETWORKS*, 52(12), pp.2292-2330.

Zerger, A. et al., 2010. Environmental sensor networks for vegetation, animal and soil sciences. *INTERNATIONAL JOURNAL OF APPLIED EARTH OBSERVATION AND GEOINFORMATION*, 12(5), pp.303-316.