Ways of Looking at Data: MESSAGE (Mobile Environmental Sensing System Across Grid Environments)

DCC e-Science Longitudinal Study #2

Martin Donnelly
Digital Curation Centre
University of Edinburgh

October 2010
Ways of Looking at Data: MESSAGE (Mobile Environmental Sensing System Across Grid Environments)

MESSAGE was a three-year research project, running from October 2006 to September 2009, with two interlinked aims: to investigate, develop, harness and demonstrate the potential of diverse, low-cost and semi-ubiquitous sensors to provide data for the planning, management and control of the environmental impacts of transport activity at urban, regional and national levels, and to report pollution levels to end users in real-time; and to develop a flexible, scalable and reusable e-Science infrastructure for transferring and processing the data captured by the sensors. This study by the DCC provides a description of the project's data-related aspects, together with a contextual analysis of themes which emerged from an observation of the project during its final year.
EXECUTIVE SUMMARY

Mobile Environmental Sensing System Across Grid Environments (MESSAGE) was a three-year project funded jointly by the Engineering and Physical Sciences Research Council (EPSRC) and the UK Department for Transport (DfT). The rationale behind this mix of funders was "to demonstrate some e-Science concepts that could be applied in a real world scenario." The project ran from October 2006 to September 2009.

During the project's final year, a representative of the UK’s Digital Curation Centre (DCC) interviewed key personnel within the project team, with a view to identifying and recording the data-related practices and attitudes of a variety of research stakeholders, particularly the variance in perspectives which researchers bring to bear on the same dataset, and the agreements and rights – the sensitivities, in other words – which pertain to the creation-collection of research data during and beyond the project lifecycle.

In broad conclusion, the report:

- Highlights the need for user-centred strategies for dealing with the creation/collection, use and longer term management of data;
- Discusses a fundamental problem for longer term preservation of data, i.e. that the period of a dataset's usefulness is likely to be considerably longer than the period for which the capturer-creator of that data is funded to maintain it;
- Identifies differences in attitude between disciplines, partner institutions, and stakeholder groups; in short, 'ways of looking at data';
- Investigates a range of sensitivities pertaining to various data types created/captured/reused by MESSAGE, stressing the need for data-sharing policies to give these appropriate consideration;
- Notes the effects of time on aspects of the data management endeavour;
- Supports the view that research funders should encourage researchers to develop and maintain data management plans (and other formalised agreements) from the grant application stage onwards.

Recommendations:

- That techniques such as use cases, storyboarding and profiling be investigated and exploited with a view to understanding and meeting user needs in this area;
- That training and guidance continue to be developed with sensitivity to the needs and perceptions of researchers, paying particular attention to issues of appraisal and selection which currently appear to be largely overlooked;
- That funding bodies provide increased support to grant-holders and applicants in the production of data management plans and consortium agreements, and place more value on their production;
- That a relevant organisation commission a subsequent study to concentrate on issues of formalised agreements, data sharing, contracts, etc.
Contents

Legal Notices ........................................................................................................................................... 2
Catalogue Entry ......................................................................................................................................... 2
Disclaimer .................................................................................................................................................. 2
EXECUTIVE SUMMARY .......................................................................................................................... 3
Contents ....................................................................................................................................................... 4
Part 1: The MESSAGE Project .................................................................................................................. 5
  Background to the case study .................................................................................................................. 5
  Methodology and scope .......................................................................................................................... 5
  Focus and purpose .................................................................................................................................. 5
  Precursor projects / antecedents ............................................................................................................ 6
  Outline and aims of the MESSAGE project ............................................................................................. 6
  Strand I: Sensors ...................................................................................................................................... 7
  Strand II: e-Science infrastructure ......................................................................................................... 8
  Data specifics: volumes, data transfer .................................................................................................... 11
  Outcomes ............................................................................................................................................... 12
Part 2: Overarching Data-Related Themes .............................................................................................. 13
  Ways of looking at a dataset ................................................................................................................... 13
  The time-sensitivity of data .................................................................................................................... 15
Themes related to Roles and Responsibilities .......................................................................................... 17
  Interactions ........................................................................................................................................... 17
    (i) Between computer scientists and domain specialists ................................................................. 17
    (ii) Between the project partners ....................................................................................................... 18
    (iii) Between the project and third-parties (internal/external) .......................................................... 18
  Division of labour .................................................................................................................................. 18
  Human factors: the importance of domain specialists and direct intervention in the curation process.. 20
Themes related to Value and Benefits of Data Management .................................................................... 21
  Pay it forward: standards and good practice ......................................................................................... 21
Themes related to Dealing with Sensitive Data ....................................................................................... 23
  Data trade within the research field, and beyond it .............................................................................. 23
  Security and sensitivity: barriers to openness and sharing ................................................................... 24
  Agreements: formal and informal ........................................................................................................... 26
  Data management plans ......................................................................................................................... 28
Further Themes ......................................................................................................................................... 28
  Problems with the short term project funding/organisational model .................................................. 28
  The 'keep everything' mentality ............................................................................................................. 30
  The user's data experience ..................................................................................................................... 31
Part 3: Conclusion ..................................................................................................................................... 32
  Broad conclusion: On the comparative difficulties of human and technological data-related issues .. 32
  Specific conclusion 1: On the visibility of advice ................................................................................ 32
  Specific conclusion 2: On the usefulness and appropriate scale of data management planning ......... 32
  Specific conclusion 3: On the need for further study .......................................................................... 33
Afterword ................................................................................................................................................... 34
Acknowledgements ..................................................................................................................................... 34
Annexes ....................................................................................................................................................... 35
  Annex I: List of interviews .................................................................................................................... 35
  Annex II: Interview question framework ............................................................................................. 36
It was evening all afternoon.
It was snowing
And it was going to snow.

- Wallace Stevens, 'Thirteen Ways of Looking at a Blackbird'

**Part 1: The MESSAGE Project**

**Background to the case study**

MESSAGE (Mobile Environmental Sensing System Across Grid Environments) was a three-year research project, running from October 2006 to September 2009 and funded jointly by the Engineering and Physical Sciences Research Council (EPSRC), as part of the e-Science initiative, and the UK Department for Transport (DfT).

This study, which is largely confined to the project's data-related aspects, constitutes the second DCC/e-Science longitudinal case study, and stands as a sequel to Graham Pryor's analysis of the CARMEN (Code, Analysis, Repository & Modelling for e-Neuroscience) project.¹

Undertaken between December 2008 and November 2009, the study tracks the emergence of technological and organisational solutions to a range of discrete data handling problems spanning a variety of domains, including Transport Studies, Computer Science, Environmental Science, and Engineering.

Additionally, while MESSAGE was a large, multi-site collaboration between computer and environmental scientists and transport researchers, the identification of generic/transferable infrastructural and workflow commonalities across the wider e-Science community forms an integral part of the engagement.

**Methodology and scope**

The primary source for this study was a series of nine interviews carried out with key data-related members of the MESSAGE team, spanning researchers, evaluators, and dedicated computing staff, as well as the project coordinator, Professor Neil Hoose.²

Interviews were semi-structured and most were carried out in person, usually at the subject's home institution. Where this was not practicable, interviews were carried out via telephone. The interviews were recorded, and transcripts were provided to the subjects in order to give participants an opportunity to correct any errors.

In addition to the interviews, the author was granted access to the project's internal wiki, and was invited to attend (as a silent observer) several meetings of the project's Scientific Committee, and a major public demonstration day held in London towards the project's conclusion.

**Focus and purpose**

MESSAGE was selected as the subject of this study as an exemplar of an innovative, large scale, interdisciplinary, multi-partner project, with a strong data component covering a variety of sources and stakeholder types. It is important to note that data and data-related issues were not the project's primary focus, but they were the study's primary focus, the selection made since the project dealt with data that were voluminous, heterogeneous, and had multiple owners across diverse business sectors.

The aim was to track the development of attitudinal responses to data-related issues throughout the project's final year, interviewing at least one person from each of the project's academic partners. As the interviews progressed, many data-related issues which arose seemed to originate from a lack of formalised agreement at the project's outset. This led in part to the DCC developing a Data Management Planning template and tool, and engaging with UK Research Councils to encourage the recommendation (or mandate) and take-up of these resources.

¹ [http://www.dcc.ac.uk/docs/publications/case-studies/carmen.pdf](http://www.dcc.ac.uk/docs/publications/case-studies/carmen.pdf)

² Annex I provides a complete list of interviewees, and the question framework which underpinned each of the semi-structured interviews is given as Annex II.
While a number of technical issues are covered here, the study's primary focus is on the human factors that influence curation and data trade within and beyond the research project environment. The study also offers some reflections on the roots of the threats faced by data collected or created via short term research funding.

Although MESSAGE did not in itself create or capture large-scale volumes of data, the project has laid the foundations for future large-scale data capture. This case study focuses, therefore, on the approaches and attitudes towards data demonstrated by the enablers of these future datasets – the researchers, in other words – at a critical moment that comes comparatively early in the data lifecycle. It examines their familiarity with data management techniques, and the extent to which the transfer and sharing of project-related data were made explicit/formalised within the project.

Precursor projects / antecedents

The MESSAGE project finds its roots partially in an earlier project: the National Transport Data Framework (NTDF). This was funded by the DfT, and was a collaboration between the Cambridge e-Science Centre and the Transport Studies group at Imperial College. The NTDF collected various transport-related data resources, including traffic flows, events on the transport network (such as accidents, scheduled engineering work), and railway timetables, and put together a search engine which enabled people to discover these sorts of data more easily. The data were annotated to work in conjunction with ontologies, making searches more intelligent and enabling users to discover sources of data based on keywords.

MESSAGE grew out of that collaboration, with the goal of building an infrastructure to collect a different sort of data, namely pollution data, on much more fine-grained temporal and spatial scales than had previously been done.

Outline and aims of the MESSAGE project

MESSAGE had two inter-linked aims: (i) to investigate, develop, harness and demonstrate the potential of diverse, low-cost and semi-ubiquitous sensors to provide data for the planning, management and control of the environmental impacts of transport activity at urban, regional and national levels, and to report pollution levels to end users in real-time; and (ii) to develop a flexible, scalable and reusable Grid-based e-Science infrastructure for transferring and processing the data captured by this heterogeneous sensor network, and supporting a wide range of scientific, policy-related and commercial applications. This infrastructure involved linking middleware applications to process incoming data (both from the MESSAGE sensors themselves, deployed in one way or another by application scientists, and from third-party data sources) to provide greater value to potential users of the system.

The team aimed to demonstrate how low-cost sensors – both mobile and static – could be deployed in high densities and linked wirelessly, sharing information and thereby minimising data logging and data handling issues. The integration of sensors with mobile devices that people generally carry with them most of the time (i.e. mobile phones) effectively turns their bearers into a shifting network of roaming sensors. This lays the groundwork for the creation of a larger archive dataset of pollution levels which could act as input to pollution models to help predict more accurately what future pollution levels might be like based on their current and past values, local weather conditions, and other factors. (Figure 1, below.)

The differences in focus and objectives between the partner institutions, disciplinary affiliations, and indeed the project's twin aims (sensors and infrastructure), emerged as a sub-theme for this study, as reflected in the study's title and the general emphasis on perspectival difference.4

3 http://www.dft.gov.uk/rmd/project.asp?intProjectID=12065
Figure 1. Data Modelling, Mining and Evaluation to Support Environmental Management (J. Cohen)

Roles

The core project consortium comprised five academic partners: the Universities of Cambridge, Leeds, Newcastle and Southampton, and Imperial College London, the last of which was the lead partner: Professor John Polak, Professor of Transport Demand and Head of the Centre for Transport Studies, was the Principal Investigator. The project also had the support of nineteen non-academic organisations, spanning public sector transport operations, commercial equipment providers, systems integrators, technology suppliers and local authorities.\(^5\)\(^6\)

Each of the academic partners had its own particular focus. Three of the partners each developed a distinct sensor platform: Cambridge investigated the potential for personal devices (mobile phones) to support a sensing system; Newcastle developed a "smart-dust" network using ZigBee (IEEE 802.15.4) motes; and Imperial devised a network utilising Wi-Fi (IEEE 802.11.g) and WiMAX (IEEE 802.16) technologies for communications and positioning, together with a set of novel sensor designs. Southampton's main involvement was in the area of evaluation and user studies, while Leeds' role was to quality assure the data captured by the experimental mobile sensors against their own well-established, static sensor technology.

Strand I: Sensors

The mobile sensors were variously mounted on vehicles or carried by (walking and cycling) humans in order to act as mobile, real-time environmental probes, sensing transport and non-transport related

---

\(^5\) Professor Neil Hoose, an external consultant, was briefly engaged prior to the bid preparation stage, and then brought back in as part-time project coordinator for the duration of the funding period. His independent, objective status enabled him to bring the disparate strands and partners together, but in his own words led to a lack of perceived authority.

\(^6\) Most (although not all) of the non-academic partners provided general guidance. A couple of partners provided technical support, notably O2 who provided some data SIMS. The major non-academic contributors were Leicester City Council and Gateshead City Council, who provided test sites for Motes; Cambridge City Council, who provided access to existing pollution sensor systems for comparative measurement with the Cambridge handheld devices; and Transport for London, who provided traffic model data and access to a bus to act as a mobile platform for the DUVAS Sensors. These four partners also provided access to existing datasets and data collection platforms.
pollutants and hazards. Each of the sensor platforms was designed to integrate with the e-Science architecture, and the differences in data structure were ironed out using a shared, project-wide schema.

Within the sensor strand of the project there were two sub-areas of concentration, exemplified here via sample use cases:

i) **Longer term data.** Transport researchers, local authorities and transport companies wish to examine emissions from traffic over periods of time, and adjust their policies, strategies or models either to study, mitigate or minimise emissions, or to inform changes in routing;

ii) **Real-time data.** Someone with a bronchial or other breathing condition wants to go into London from Brighton. Before travelling they want to know what the local pollution forecast is for that day, and whether they should walk down a particular street or take a different route to the same destination via not too many more steps, thereby encountering significantly less pollution, e.g. walking through traffic-calmed Soho rather than along the bus route of Oxford Street.

The team at Cambridge encountered difficulties in obtaining sensors at the required price/ performance point, and were therefore obliged to engage an independent consultant to package off-the-shelf components and GPS/GPRS communications hardware into custom-made sensors. From an optimistic initial ambition to roll out approximately 100 sensors in continuous operation for "a large period of time," the system was scaled back to "thirty or forty sensors over perhaps a couple of weeks," thus leading to lower data volumes than were originally anticipated. Nonetheless this represents a sizeable amount of data, capable of underpinning significant and innovative discovery.

As is to be expected, other issues were experienced during the sensor development. GPS sensors typically have an accuracy of ten metres, but the Cambridge team found this to be less precise when tracking the mobile sensors. The sensors also report their speed and bearing, so calibrations can be made to compensate for this. Additionally, the GPRS network coverage is not 100%, and occasionally mobile users would stray into dead zones. When this happened, real-time data flow suffered, but no data was lost: the sensors simply transmitted the backlog when the user returned to an area with good coverage.

**Strand II: e-Science infrastructure**

The other primary aim of the project was to develop a flexible e-Science infrastructure to support the heterogeneous sensor deployments. This strand of work was led by Dr Jeremy Cohen of the Internet Centre at Imperial College, and involved broad collaboration across the project partners, the other e-Science groups (Newcastle and Cambridge), and with external organisations; a key interaction was with the transport groups.

The e-Science infrastructure was created to handle the data captured by the heterogeneous sensor deployments at multiple sites. Data imported from the various sensors carries similar properties: all sensors record time, location/position and readings for various pollutant species – the mobile sensors additionally record their speed and bearing – and other properties such as noise and calibration settings may be recorded.

---

7 At this stage, it is perhaps worth a brief examination of the shifting perceptions of the term ‘e-Science.’ In his CARMEN study *(op. cit.)*, Pryor identifies "what have come to be regarded as the key characteristics" of e-Science:

- cross-domain collaboration in which the conduct of research is dependent upon the electronic sharing of data;
- a research environment enabled by the interconnection of computers;
- the use of ubiquitous networks to provide high-speed data transmission across a geographically distributed community.

The definition of e-Science put forward by one of the interviewees, Mark Hayes of Cambridge e-Science Centre, is simply "applying innovative IT to do new science.” The term has been heard with less frequency recently, as increasing numbers of projects and research groups apply innovative IT techniques in their daily scientific endeavours. It may be that what was previously thought of as e-Science increasingly becomes thought of simply as ‘science.’
This data enters the MESSAGE data management infrastructure, which is the layer between the sensors and the databases. (See Figure 2, below.) Pre-processing is carried out at this stage to transform the data from various native capture formats to a standard XML storage format: Urban Traffic Management and Control (UTMC)\(^8\).

Early in the project, a decision was taken whereby each of the partners would follow their own data capture, processing and storage workflow, using their own version of the UTMC schema, and that the data management infrastructure would deal with the differences and encourage efficiencies. This held the additional benefit of trying out multiple approaches to determine which would be the most efficient when it came time for a much larger scale, future deployment, with potentially millions of sensors online at once nationally, or even internationally. The system developed at Imperial is therefore built on Web services, comprising multiple gateways that accept data from the sensors and pass it into the central management area.\(^9\)

Gateways can only handle a limited amount of simultaneous information input, so the numbers of gateways required at any given time is a function of the number of sensors feeding into the system, and the frequency with which they are reporting. This has the potential to vary dramatically over the course of a day, from large numbers of mobile sensors online during (e.g.) morning rush hour, to fewer sensors in the middle of the night.\(^10\) One of the goals of the infrastructure work was to ensure that the system was capable of scaling to meet these peaks in demand without having large amounts of expensive computing resource sitting idle during the longer, quieter periods.

**Cloud Computing**

The proposed solution was to utilise cloud computing to leverage an on-demand resource from a third party supplier, in this case Amazon's Elastic Cloud Compute service,\(^11\) to bolster the in-house compute resources. Functionality was put in place for a future central management node which would accept calls from the sensors, identify the best server according to current capacity demands and availabilities, and link the sensor to that resource. The management node would also close down resources when there was no load on them. The sensor gateway nodes could then be used to carry out whatever data processing was required, e.g. reformatting, averaging values, calculating standard deviations, maxima, minima, etc. From there the data would be passed to another service which would act as a buffer, holding the data until such times as it could be inserted into the database.\(^12\)

**Technical interactions between the sensor-developing partners**

The Imperial team worked primarily with an Oracle database, with comparative speed tests run against a MySQL database. Newcastle used an Oracle Edge Server product, specifically designed to handle sensor data from large numbers of sensors. Similarly, Cambridge's system accepts data from mobile phones via a GPRS channel, which forwards the data towards a stack of services to populate the database.

Due to the different database platforms and services utilised across the academic partners, OGSA-DAI and OGSA-DQP were used to provide a single point of access and control across multiple distributed databases (Figure 3, above), and Cohen in particular worked closely with the OGSA developers in Edinburgh to identify and help solve issues which arose on both sides.

\(^8\) http://www.dft.gov.uk/pgr/roads/tpm/utmc/

\(^9\) The data is stored only once the metadata has been described. Metadata elements include the nature or subject of the data; its provenance or origin; its location; its quality including the reliability of the source providing it. The metadata assignment process is carried out manually, owing to the complexity of integrating heterogeneous data sources. It is therefore necessary for there to be a human involved in the process, somebody who has a familiarity with the data and with the datasets in order to ensure that what's recorded is appropriate, in order to carry out quality assurance.

\(^10\) An international roll-out would, of course, reduce to a certain extent these peaks in demand.

\(^11\) http://aws.amazon.com/ec2/

\(^12\) Jeremy Cohen: "I should say at the moment, we're not actually doing that; we're basically using the nodes as general handling for the data, but the functionality is in there to do that."
Figure 2. The MESSAGE e-Science Architecture (J. Cohen)

Figure 3. Federated Database Infrastructure (J. Cohen)
Testing and Deployment

While much of the infrastructure was tested using simulated datasets derived by Cohen in the lab, the final system was tested via live deployments at the sensor development sites towards the end of the project. When Cambridge succeeded in running fifty pedestrian-mounted and cycle-mounted sensors concurrently in Autumn 2009, Cohen and his colleagues were able to gain valuable insight as to what the non-cloud system could handle, and where and when the bottlenecks started to emerge as data was transferred between the sites.

Similarly, the consortium demonstrated the live system at their launch/demo day in London in July 2009, and at a conference in Sweden in September 2009 where the Newcastle team organised a small network of motes operating throughout the conference hall.

Limitations due to shortage of financial/compute resources

_We were working with a machine which did start to struggle when we got up to fifty sensors, but we were well aware that we could have had a machine that was significantly more powerful and handled a much higher load._

- Dr Jeremy Cohen

Given the relatively small-scale deployments that the MESSAGE team worked with during the project's lifespan, the use of local computing resources was generally sufficient for their needs. However, during the test deployments, the team encountered an issue where, when the user moved the map interface (the 'view' on to the underlying data) to a new location, the database had to process all the data again; processing two or three million database rows took around a minute for the calculations to be adjusted, resulting in a disjointed user experience. When these unacceptable time lags were encountered using a relatively low-powered server, Cohen was able to demonstrate the effectiveness of the system by moving temporarily to a more powerful machine, which cut the delays from around one minute to something in the order of a couple of seconds.

However, the system's ambitions extended beyond local, smaller scale deployments, and in order to cope with the additional load brought on by the addition of hundreds, thousands or millions more sensors, a more flexible solution would be required. Due to the financial constraints inherent in grant-based funding, Cohen was limited as to what he could realistically achieve in his use of on-demand, cloud-based resources. He found that systems such as Amazon's work best (and demonstrate the best return on investment) when they are used to host what he called "a stable product": since the MESSAGE system was constantly being tweaked as part of the iterative design process, this was made more difficult. Also, when machines are running around the clock, third-party services tend to become expensive quite quickly: it is when they are only used occasionally that their cost benefits become most clear.

MESSAGE project coordinator Professor Neil Hoose shares Cohen's enthusiasm for the potential benefits of cloud computing to a larger-scale MESSAGE deployment. One of the most interesting things in the project, he said, was "the exploration and use of utility computing... buying computer power as a resource, as a service on demand, not as a capital item." This, he reckoned, held the potential for far-reaching consequences.

Data specifics: volumes, data transfer

The commercial telecommunications partners provided the project team with guidance in making efficient use of the GPRS mobile phone network by compressing the data transported across it, and by optimising the frequency of readings from the sensors in order to avoid using excessive amounts of valuable bandwidth. This points to something of a dichotomy in data management: there is a widespread perception that storing data is no longer a problem, that storage is now cheap and reliable enough to keep everything forever; however, we are rapidly approaching a situation, if we have not already reached it, where we run the risk of collecting greater volumes of data than we can readily move around via networks. (As

---

Biotechnology and Biological Science Research Council (BBSRC) Chief Executive Douglas Kell noted in his IDCC09 keynote, the quickest way to move a petabyte of data may be by truck.14

That said, the MESSAGE data-handling activities are more proof-of-concept than large-scale roll-out, and did not deal with extremely large datasets. At Cambridge, the sensors transmit plain text, which is then stored in a relational Postgres database, with one large table holding archived data and one smaller table holding real-time data (around 500,000 rows in total at time of interview, but this was expected to grow considerably as more sensors were brought online.) Outputs from the database are produced in Comma Separated Value (CSV) or Keyhole Markup Language (KML) formats, which then go through to the visualisation/interface layer.

The Newcastle team members were helpful in describing the data specifics that they were handling and stewarding, which led to a few surprises. The log files, for example, were markedly larger than the real-time data: having collected data for six months, their Oracle database was between 80 and 100 GB in size, with archive log files twice as large. The Newcastle team reported that these volumes were not a great challenge to manage, but that being responsible for the entire data lifecycle marked a difference to the working practices they had been used to outwith the university environment, where each individual is generally responsible for only a subset of these tasks. This unity of responsibility (or lack of division of labour, depending how you look at it) is considered both a strength and a weakness of the research environment compared to the private sector.

Outcomes

Successes of the project

MD: Are there any particular key successes which you would want to highlight, again from your own perspective?

BW: I think some of the key things that we need to put forward are the work that's been done, varied work between each group individually about actually presenting some of the data in actual human readable forms. Everybody's been working on it slightly differently, on slightly different datasets and different ways of doing it, but there's been a constant dialogue and exchange of information about how different people are doing it, both directly and indirectly through talking to each other and showing each other the latest results.

- Interview with Dr Ben Waterson, Southampton

Failing to achieve a balance between the e-Science and sensor strands of the project was one of the risks identified at the very outset, and while in Neil Hoose's words "in reality the sensor development always tends to dominate because it's very practical, it's very physical and you can see it", the project team and the two primary funders appear satisfied that a good middle-ground was achieved. A summary of the final report to EPSRC is available via their Website.15

Indeed, all of the stakeholders with whom I spoke felt that the project had succeeded in its aims. Each of the groups reported success in the development of their respective sensor types, and the system as a whole (sensors plus e-Science architecture) was successfully demonstrated at a high-profile event in London on 30 June 2009. Due to delays encountered in some of the sensor development strands, the bulk of the infrastructure development and testing, and indeed the visualisation/proof of usefulness work, was carried out on simulated datasets, which was a source of frustration to some of the interviewees, and one expressed a view that the project "did pretty well. We deployed far more [sensors] than most other pilots or research projects ever do, but I'd still say we were an order of magnitude lower than was needed to test an e-Science concept fully."

14 http://www.dcc.ac.uk/sites/default/files/documents/events/dcc-2009/programme/Presentations/0915kell.ppt
15 http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/E002102/1 There is also a report to the DfT, which according to the project coordinator is not published and is not likely to be in the foreseeable future. For further reading, Traffic Engineering and Control Vol. 50 No. 11 (December 2009) contains three articles covering the entire project.
At the London demo day, real data was used for the route planner, although it was collected at a lower granularity than the synthesised datasets, and the resolution was artificially increased. Some discrepancies were experienced between the pollution levels reported by the mobile sensors and those recorded by the static units when these were compared in real time, although this may have been reconciled since the interviews took place.

Delays in the collection of the data and having to scale back the numbers of sensors led inevitably to adverse knock-on effects for those hoping to carry out analysis on the datasets during the project's lifetime. As Neil Hoose notes, "we haven't gone as far with the data analysis and interpretation as a lot of people would like. We've now shown that we can collect interesting datasets with some interesting characteristics. What they [the datasets] mean was never the goal of the project particularly. We have basically demonstrated that we could get the data and that... this data could be used for various purposes, but we haven't linked the two together saying it should be used for various purposes, and "Oh look, here's a problem that you weren't aware of..." This, in fact, is the direction in which the domain scientists are intending to take this in as the next step.

In Cohen's words, the project has "really been able to give a lot of new ideas and new models and new ways of handling data to people who currently are working in a very different way I think. [...] We now have devices where we can press a button on the front, the device comes on and, you know, it starts sending data [...] It's packaged now in a manner that it is available for us to put on a vehicle on the roof or inside with suitable ducting or whatever, and it can be started up and it takes a bit of time I think to warm up and then once the sensor is running we can send it out and we can capture data on demand. So... we are at that stage where we are able now to sort of put those sensors on vehicles and take them round when we want to do demo's or when we want to just capture specific events."

**Part 2: Overarching Data-Related Themes**

Having thus far given a brief and broad outline of the project from its conception to conclusion, the remainder of this study goes on to identify and analyse the data-related themes that emerged over the course of the observation. Where relevant, metaphors and external quotations have occasionally been attached to the themes in order to communicate them in a more transferable and abstract way.

**Ways of looking at a dataset**

[We were] trying to set up ways in which the end user could potentially interact with the data without being exposed to the full horrors of the actual data itself.

--- Dr Ben Waterson

In preservation circles, there is a tendency to think of a dataset as a simple, objective fact, as a series or matrix of numbers and words which is fixed, immutable, carved in stone, and which must be protected precisely in that form. But, as with linguistic signifiers, no dataset exists in isolation, it exists within a galaxy of other datasets which may overlap, nuance, contextualise and shed different quantities and qualities of light on it.

Additionally, attitudes and trends change over time, and the way in which we view a dataset will undergo a shift as new discoveries and re-evaluations are made, new connections identified. So a dataset's interpretation, and consequently its usefulness, must undergo periodic or continual shifts.

With this in mind, certain approaches to long term preservation and curation may benefit from a rethink. There is no single 'correct' way to curate a dataset, but rather a multiplicity of potentially useful 'views' on to it, which will shift over time. Using Ben Waterson's terminology, above, we tend to concern

---

16 This was achieved by taking rolling averages, and then interpolating points according to traffic speeds and demands and levels. I was told that six units were running throughout the afternoon, with four producing "reasonable data" and two producing "patchy data."
ourselves primarily with 'the full horrors' to the detriment of the potential interactions. This, to my mind at least, is where the nexus between blunt preservation and fuller curation lies.

The phenomenon is highlighted by an example developed in one of the case study interviews, where we discussed a particular air pollution dataset being "useful to Researcher A for one thing, it's useful for Politician B for another thing, it's useful for Bus Driver C for another thing, it's useful to Asthma Sufferer D for another thing." Each stakeholder group has a different relationship with the same data: the university researcher is interested in what she can do with the pollution data in terms of modelling, while the Politician wants to make sure the relevant local, national and European emissions regulations are seen to be met. The bus driver (or more likely the controller of the buses) wants to avoid congested areas that would slow him down and prevent him from meeting his schedule, while the asthma sufferer, on foot, tries to avoid areas of poor air quality which could trigger an attack.

Time-wise, the politician and researcher may primarily be interested in long collections of legacy data which show how the situation has changed, while the bus driver and asthma sufferer are only interested in what's happening at that particular time. So it must be seen as a case of 'horses for courses': that a dataset will be differently useful according to context and user needs. Real-time data becomes legacy data with the passage of time.

Jeremy Cohen acknowledged that these multiple perceptions/ points of view / voyages-in drove the development of the MESSAGE system's different data interfaces: "There have been several points during the project that there have been questions: how do we store the data? What format do we store the data in? Where do we store it? What's the right way to handle the data? And there are so many different ways to do that [...] there are many different ways it can be of use to people. [...] We have many different visualisations: for people who want to access the data at that sort of high level visual view of the data, and at a lower level people who want to get down and write SQL queries will be able to. [...] And then even people who want to go to a lower level and who want to get into the database and actually start writing much more advanced database queries, or stored procedures or whatever it may be to do things with the data. And ultimately the system will allow any of those and it's just a matter of [...] policy: who do you allow access to?"

Researchers generally want to use data as soon as possible, to analyse and "do science" on it. But that's not what MESSAGE was really about, to the occasional frustration of the domain scientists. So different cohorts brought different expectations and goals to the project as a whole, just as different communities have different ideas about what a given dataset will do for them. And caught in the middle is the data manager. As Newcastle's Lakshmi Suresh says, "It is very, very tricky to work interdisciplinarily and inter-institution where everyone's requirements and needs are always different. There is always something, as a data manager, you need to accomplish which is always hidden behind the agenda of others."

One way of expressing the project's end-goal (beyond the technology) was "to get an understanding of what's going on pollution-wise in a network [i.e. an urban area of particular scale] in order to present a coherent, consistent picture to travellers." At the London demo day, Waterson and his team reused traffic level data from the Imperial College archives to get a base idea as to levels of congestion within the network, combining these with air quality data to create a pseudo 'time and spatially variant' pollution field map of South Kensington. This then enabled the creation of custom information tools on top of it, including a mock journey planner that enabled users to choose a minimum pollution exposure route as well as the traditional fastest or shortest routes. This offers different views on to the data, with only the layers and granularity that are of interest to the user being displayed. These then become value-added services on top of the data, and people tend to see it from within their own view, i.e. what can this data do for me?

The MESSAGE datasets are primarily observational, i.e. non-repeatable, and therefore a strong candidate for preservation and curation. Ancillary materials which may complement (or be crucial to the understanding of) the datasets include instrument calibration settings, experiment parameters, lab notebooks
and other experimental records. It is not clear how these ancillary materials might be preserved alongside (and linked to) the datasets.

We note the potential value of this data (and future datasets collected via the MESSAGE sensor infrastructure) to the public policy decision-making process. But the datasets derived from MESSAGE are of interest to a broad and diffuse set of stakeholders, spanning research scientists, transport companies, public sector and government authorities, and the general populace. Each of these groups has different wants and needs, making the job of preserving – and granting appropriate levels of real-time access to – these materials all the more difficult.

While no data from the project is currently available publicly, it is acknowledged that – particularly from the transport side – this is an important issue. When asked about depositing the datasets, for example in a data centre, Lakshmi Suresh told me that the Newcastle datasets will be stored in a new 'transport observatory' which will allow access to transport and air quality researchers across the North-East region, as well as local authorities. Indeed, it is interesting to consider the potential variety of planned future projects into which MESSAGE will feed. Each partner will take their respective strands of work in different directions, and the datasets will become differently useful in time.

**The time-sensitivity of data**

> I think we’re realising that that database probably serves more purposes than I originally had thought.
>  
>  
> – Dr Jeremy Cohen

There is a temptation to think of digital preservation as a kind of cryogenic process, a freezing in time: "this is the dataset, it's exactly as it was originally, we have preserved it." But while the bitstream may have been preserved, and the software needed to read and manipulate the data may still be running (directly or via an emulator), the nature of a dataset will change inevitably over time, from different perspectives and as its context and position and importance shifts, even if the data itself does not change. Embalmed data is of limited use...

Clearly there's a lot of potential opportunities to make data available in a useful way to people, and likewise much longer term historic data, looking at data over periods of months – or even years potentially – and being able to look at how trends are changing and how things have altered, and then being able to correlate that with changes in road patterns, or changes in other sort of urban landscapes or whatever. So I think from my point of view, from the technical side of the project, it's really just been a matter of trying to provide a platform that will offer ways to access data in many different manners and show what opportunities there are to actually get to that data, and really the potential uses of it are clearly very, very wide ranging and potentially very great. (Dr Jeremy Cohen)

More than one interviewee noted frustration in the delays they experienced receiving data from other project partners, and indeed some confessed that they had been responsible for delays owing to a variety of internal and external factors. Consequently, Southampton's usability study depended more on analysed data (or 'results') than raw data, which took time to produce and was, in turn, impacted by delays further up the chain.

Perceptions of the role and position of the datasets did indeed change over the course of the project. In our first interview, Jeremy Cohen noted that his original view had been that the real-time data would be the most important and the most valuable resource or facet of the system, but he had in time come to the conclusion that the historical analysis of large quantities of older, legacy data was actually just as important, and perhaps even more so.
Figure 4. Shifting perception of a dataset I (pessimistic)

Figure 5. Shifting perception of a dataset II (optimistic)
In Cohen's words, "The bulk of effort […] has been put into what we called at the time the real-time database, but I think we're realising that that database probably serves more purposes than I originally had thought." The analysis and longer term historic processing possibilities became more important as the project developed, which led to the database being split into efficient high-speed caches which can provide current data, and the larger, slower legacy database, into which the real-time data will shortly be ingested. The relationship, therefore, is not binary, and either-or. It is a continuum, and this is a fundamental characteristic of live data.

The relationship between current behavior and future possibilities is always at stake: it is modified by what has been done in the past and is always modifying what can and will be done in the future. In formal economic terms, preservation decisions are path-dependent.

- Blue Ribbon Task Force Final Report, p28

The Blue Ribbon report acknowledges that 'preservation is temporally dynamic – it takes place over time.' (p31) But that is not all: data itself is temporally dynamic – a dataset is very likely to change over time, and moreover its place in the world will almost inevitably change in time as the sum of human knowledge expands, and new connections are drawn between previously unlinked models or phenomena.

Of course, hindsight is a wonderful thing, and this development of consciousness is by no means unique, but it highlights the need for longer term preservation, description and management strategies to be put in place early in the project's genesis, and continually revisited over its lifetime (and beyond).

**Themes related to Roles and Responsibilities**

In each preservation scenario we can distinguish three pivotal roles—users or beneficiaries who demand preserved digital objects, owners of digital assets, and archives which together supply preserved digital objects. Preservation works best when the interests and actions of these users, owners, and archives can be aligned in an economic strategy and operationalized in a business model. When the interests (perceptions of value) and actions (incentives) are in alignment, sorting out the roles and responsibilities of the three entities can be straightforward, even if there is jockeying among the three about who pays "the most," that is, more than they want to. By the same token, long term preservation is more difficult when the roles are diffuse or fragmented. Preservation is most difficult when these three roles are diffuse and change over time.

- Blue Ribbon Task Force Final Report, pp29-30

**Interactions**

(i) **Between computer scientists and domain specialists**

Given the two-pronged nature of the project (sensors and system architecture), the first type of interaction probed during this case study was between the domain specialists and the computing-side people. Each cohort had to learn from the other about the specifics of the data they were managing, and Jeremy Cohen in particular noted that he both anticipated and experienced a steep learning curve when it came to picking up the work done by the transport and environmental scientists. Providing support for domain specialists, that is those people with an intimate and expert connection to the data, is not simply a matter of reworking standardised procedures and principles in a new environment: it's more complex than that.

Despite this, relationships between the sensor scientists and the architects have been fairly modular. As Cohen put it, "my role [as an e-Scientist] is to take the data and do something with it using e-Science tools, and not really to have to spend a lot of time becoming the person that knows all about the data. It's useful to have input from the domain scientists but not actually to become a domain scientist yourself in that area." The role becomes one of "facilitation," and "giving the opportunities to the people who produce the data to do something with it."

With regard to the possibilities opened up by this approach, the system architects were initially surprised by how warmly their expertise was received by the domain specialists. What they felt to be quite basic functionality was in fact considered a great step forward, and tremendously powerful in advancing the discipline.18

(ii) Between the project partners

_It became pretty clear, pretty early on, that MESSAGE as a project had an incredibly ambitious scope and it was under-staffed, and therefore it was important to have a sort of liaison role in the middle that was kind of running round in circles trying to knit things together._

— Dr Robin North

MESSAGE was organised with a two-committee structure. The **Advisory Board**, which met every six months, was a mechanism for the funders to come together to learn about the project, to offer advice, and to liaise amongst themselves. The **Scientific Committee** was a working forum for the academics, and intended to foster interpersonal relationships. As project coordinator Neil Hoose put it, the quality of human, face-to-face contact differentiated this forum of information transfer from the others. Overlap between the two committees was achieved through the senior PIs attending both. The general consensus seemed to be that this worked quite well, with the Advisory Board people kept up to speed and engaged with the project's scientific developments.

Internal communication was initially expected to take place via a project wiki and internal mailing lists, with built-in document management. It was hoped that this would provide a central space for the industrial and non-academic partners to interact with the project, but according to North this did not pan out.

In terms of external/non-academic partners, liaison with Cambridge City Council enabled the Cambridge team to carry out real-time comparisons between the data that they collect via mobile sensors and the data from the static sensors. The Council also holds eleven years' worth of data from fixed air quality monitoring sites. At Leeds, James Tate noted that they had also provided their data for benchmarking/quality-assurance purposes, but had had to wait a long time to receive data from the other partners, which appeared to be something of a sore point. (The initial intention had been that they would have the opportunity to 'do some science' with the data collected from the new sensors, but various delays stymied this ambition.)

(iii) Between the project and third-parties (internal/external)

The major non-partner organisations with whom the team noted good working relationship was the OGSA-DAI project, a partner in OMII-UK. These relationships, like most relationships, appear to have developed as a result of necessity and mutual interest. Cohen had originally considered using OGSA-DAI to power the architecture's front-end, but soon realised that the compute time overheads would be prohibitive. Identifying and discussing this issue 'kicked off the really detailed discussions with OMII-UK and with the OGSA-DAI team specifically.' The support and assistance offered by this EPSRC-funded organisation appears to have been very valuable to the project.

Division of labour

One of the DCC resources discussed with the case study interviewees was the diagram below, *Core Skills for Data Management*. This resulted from a Research Data Management Forum event in November 2008, and it attempts to identify and situate different skills in relation to each other and to four key roles: Data Manager, Data Creator, Data Librarian and Data Scientist.

---

18 It is, of course, unsurprising that domain specialists may lag somewhat behind the systems experts in the perception and uptake of new IT-driven possibilities.
Some of the interviewees were shown the diagram and invited to consider which of the roles corresponded most closely to their own role within the project. Some choices were obvious – the database administrator and systems architect, for example, were straight data managers, responsible for (in Suresh's words) "the complete lifecycle of the data."

The core task for a data manager is to capture the data, analyse the data… to do the basic analysis and processing on the data, so that it can be visualised and disseminated to the users in the right way, in the right format as needed [and followed by storage and archiving]. (Mrs Lakshmi Suresh)

Often it is the boundary/fringe cases which prove most interesting. Among the domain researchers, Mark Hayes at Cambridge and James Tate at Leeds both situated themselves somewhere between data manager and data creator, while Ben Waterson at Southampton (and, to a lesser extent, Robin North at Imperial) tended more towards the data scientist category, both stressing the reuse aspects of their roles.

North sees his role as partly interpretative: "I do a lot of data curation, and that's probably one of my biggest jobs in MESSAGE: trying to figure out what the hell it all means, and put it into context." North, who had significant responsibility for inter-partner liaison, was at pains to note that his role in the project has been "a combination of several different things. Officially it was to do various bits of modelling. In practice it's been a lot to do with project management, to do with field unit design, to do with field trial design, liaison with stakeholders, scenario development and so on. I guess one of the main functions has been integrating the different bits of research within Imperial and also to some extent with the other departments." This underlines the importance of two key activities: flexibility and liaison.
Human factors: the importance of domain specialists and direct intervention in the curation process

"It's really been, very much I think, a joint effort of everybody looking at stuff together. And that's been very important."

– Dr Jeremy Cohen

JT: **Data is quality assured... it depends on the measurement.** We follow the sort of certified procedures for some measurements. So air quality measurement is the main one. To make good quality measurements of say nitrogen dioxide which is a key pollutant, you need to calibrate your analyser, do checks nightly, automatically, do manual checks every two weeks and then you need to ratify that data. And we follow those UK and EU ratification procedures. And then the data is shared as ratified. We're very careful... we can share the raw data internally between ourselves but we only ever share ratified data because that's to ensure data quality. Because if we put some measurements out that haven't been ratified and you're not confident in them, then...

MD: There's a risk attached to that.
JT: There's a risk attached with it. And then once it's in somebody's analysis and PowerPoint presentation you're never getting it back. [laughs]

- Interview with Dr James Tate

A critical link in Leeds' data management chain is the technician who, in James Tate's words, can "eyeball a graph on a screen and [say]: That's broken." Direct human intervention is something that tends to be sidelined, because it's expensive, inaccurate, and absolutely crucial...

Every day he comes in he just checks that everything looks sensible. If there is a fault he'll generally go out, try and work out what the problem is before calling one of our service providers. We've got service agreements with the instrument manufacturers. [...] Some of the information streams don't require ratification other than "they're working". Some need to be ratified by the technician, or myself, or one of the other researchers. [...] If you're measuring traffic, all these things, you can only better understand them when you've got all ten measurements, information streams, coming in at the same time. (Dr James Tate)

Some sensors will automatically flag up serious faults, but smaller faults require human intervention to identify them. 19 This is currently the only way of identifying what, if missed, would lead to a serious shortfall in data quality, hence the human part remains vital. Cohen underlines this view from his own, non-domain scientist perspective:

JC: [It's] difficult to know from somebody who's not an expert in all the values that come off the GPS receiver or all the things that come from, you know, various different types of sensing equipment that's available, what values matter, what values we need, what are the domains of those values and so on. And so really the main input has been from the transport side and I think there are so many properties with this sort of data, that so many different types of data, it's very difficult to put together storing all these properties reliably and knowing that this will be able to provide a sufficiently efficient structure to query the data out at a very high rate or whatever. And a lot of this, I would have to say, is probably things we don't know until we start testing it. Yeah, we can use general common sense and ideas about how the database will handle certain structures and certain types of values and how quickly it will be able to get certain data out, but until we start really putting lots of data through it's difficult to know. And while we've started doing tests on that, there's still a fair way to go in the last few months of the project in really testing that, sort of, infrastructure. So we've, you know, we have now a UTMC-based schema that we're working with here which is very, very similar to the Newcastle schema. And it's really a matter of now getting that, you know, getting the data running through that as more and more sensors become available to test and also again using simulated data to really boost up the number of sensors.

19 Appraisal of data was closely linked to quality assurance, with outliers manually filtered out in order to satisfy the researchers and data managers that the data met quality standards.
MD: I'm getting a picture of the value that's added by the domain specialists in terms of the interpretation and the 'doing things with the data', it is very much something that's... the domain and the discipline knowledge is crucial to that?

JC: I can't say enough about how important it's been to have both those people there and actually working on things because it's… you know, people can tell you this value is this, this value is that, this value is something else and when you actually look at the data there're always things that stand out as, why is this, this, why is that like that. And so, while there has been a sharing of knowledge on both sides about what our systems mean and what their systems mean and what their data is and vice versa, it's really been, very much I think, a joint effort of everybody looking at stuff together. And that's been very important.

- Interview with Dr Jeremy Cohen

Robin North describes his working process for the curation of data thus:

With each field trial there will be a set of either notes or more recently a video recording of what we've been doing, which is effectively my running log book of what we did during that experiment, so some of this is interim type of work because the full MESSAGE system is nearly there but not quite, so in order to get early datasets what we've been doing is effectively running a series of different sub systems in parallel, so independently but at the same time and then dragging the data off each one and then archiving it in some kind of folder relating to that test run. Typically that consists of...stop me if this isn't the sort of thing you're after, but typically that consists of a measurement of GPS directories, a video looking through the front of the vehicle from the camera lens eye view if you're looking at it with a mobile. On the sound, the audio trace on the video recording, I would tend to make voice notes as I'm going round about the experimental set up and any other observations about particular features that we might want to put through the data. We tend to record the time synchronisation information between the different datasets that way as well. So a lot of these things are about ways of linking together, a lot of them are things that shouldn't be necessary in the final MESSAGE system, but they're about making it relatively easy for me to link together the separate datasets when they get back.

It is noteworthy that this process covers the curation of fairly disparate types of data, from video recordings to pollution measurements on the street, and that the scientists still apply very manual workflows/ techniques to data. There is as yet no single/simple substitute for direct human intervention, but increased automation of (e.g.) technical metadata frees the valuable research resource to give attention to the more high-level and complex metadata required to enable comprehension and reuse by third parties.

The validation and assurance process was not automated, because in North's words, "we haven't had a system generating a lot of data in a routine manner that it's worth automating the process of recording that side of things." But this is more than just a question of economies of scale: the expert user's direct attention and intervention remains a necessary component of this process. It is therefore clear that a specialist's work in interpretation or annotation of data makes it markedly more valuable, reusable and useful than a simple: "Here are the numbers..." Indeed, there is much evidence that scientists who do not share data are reluctant to do so because of fears that their findings may be misinterpreted. This fear may be linked to a lack of confidence in the organisation and/or annotation of the data, as the recent 'Climategate' scandal showed.20

Themes related to Value and Benefits of Data Management

The past is never dead. It's not even past.

– William Faulkner, Requiem for a Nun (1951)

Pay it forward: standards and good practice

If, as it is sometimes described, the research environment is a gift economy, good data management practice is about 'paying it forward' as opposed to paying it back; taking the time and the effort to structure, describe, format and deposit data in a manner that facilitates future reuse is perceived as an act of altruism.

I asked about the issues that the researchers had experienced in their attempts to use others' data, both internally within the MESSAGE consortium, and beyond it, and how this had influenced the ways in which they organised, documented and managed their own data. Robin North categorises himself as a Data Creator in terms of designing experiments and developing systems that can measure things; and a reuse of external data (specifically weather data, traffic flow data and speed compositions) in his modelling activities. North acknowledges that he is a beneficiary of third parties' data management practices, particularly when data is organised and structured clearly and explicitly in order to facilitate its reuse. He has encountered frustration in trying to utilise datasets which, in his own words "clearly weren't created or archived in a manner that was designed to be used by others" (e.g. in the more automated context of e-Science), and so there is often a process of adapting and re-factoring it in some way to make it fit into work for which it was not expressly created.

Within the project, the team adopted a shared UTMC schema to facilitate interoperability between project members, together with an XML schema called JSON. Following the adoption of this approach, the groups found that there were differences between the partner institutions with regards to what they needed the schema to hold, and so different versions of the standard schema were created and then tweaked to meet their respective needs. At the system architecture/infrastructure level, conversations were held to establish how these disparate requirements could be met while at the same time ensuring interoperability and system efficiency; the management of this proved more time-consuming than had initially been forecast.

The integration of self-collected or self-created data with third party data complicates matters, in that it is generally received in the format that the third party dictates. Third party data utilised within MESSAGE included live and archived traffic information from Transport for London and the local authorities, real-time weather information from the Met Office, air pollution data from the National Air Pollution Archives, and live data from the Department for Environment, Food and Rural Affairs (DEFRA). Each of these data-streams had to be pre-processed in a different way, with algorithms and protocols set up based on the individual properties of the data. After pre-processing, the data are integrated into the common UTMC structure.

When asked whether their experiences with using third party data had impacted or influenced the ways in which they have structured (or encouraged colleagues to structure) the data which they are capturing, researchers noted that trying to carry out what seemed (on the surface at least) to be relatively straightforward manipulations of data was found to be 'more difficult than it feels like it should be' as a result of variance in the third parties' data management practices.21

It is expected (although not formalised) that after the project's funding ends, the project consortium partners will have access to each others' data. In Lakshmi Suresh's words, "everything works on a bit of an understanding between the institutions, so that shouldn't be a problem." More widely, as has been found elsewhere, researchers appeared more keen to be able to find and reuse other people's data – interviewees

---

21 One barrier to data sharing can be the use of closed or incompatible standards and formats. I learned that proprietary standards and file types may find their way into the lab by default, rather than as a result of any conscious decision. Instead of building bespoke systems from scratch, instruments are often purchased as 'turnkey packages' which may include everything from the instrument itself, to analysis, data logging, communications, a local operator terminal, and a central database that deals with all the messaging, data collection, visualisation and has a back database behind it. These out-of-the-box solutions often sit atop the vendor's preferred proprietary database management system, which is used as a kind of staging post: the data comes in, is stored there for a while, and then is imported into the lab's own databases which are stored in an open format such as CSV.
noted not only that they had they no wish to 'reinvent the wheel', they simply did not have the time to do so – than to spend time and effort in making their own data more widely available and documented. The general view was that the data would be available to interested parties by one means or another, but there was no unified project approach to this, and none of the interviewees appeared certain as to the extent of their own specific responsibilities. So while the researchers generally appeared content for others to access their data, at the same time they were not falling over themselves to facilitate it...

MD: In terms of discoverability, you said that there were certain data streams that you would be happy to make open to interested peers anyway, if not every man and his dog or whatever. [...] How would the people who might be interested in this and benefit from it discover the existence of the dataset?

JT: I would be happy [to] have a MESSAGE contact, whatever data is available from whatever MESSAGE demonstrations there have been. There's a central repository, people can... everything is documented to a reasonable amount, you know, [but] you could go over the top with it, and then nothing ever becomes available.

- Interview with Dr James Tate

This quote seems to support the perception that some researchers are warily supportive of more open access to their data, but that they have concerns about the standard to which their data is documented, and therefore about how useful it would be to third party users.

Similarly, Ben Waterson says: "I have no problem with other people using my data, but [...] I suppose my approach is that I will publish things on the data and if that gets people's interest up as to the kind of things I've got then they will get in touch with me and I'll do what I can." Waterson's language is similar to Tate's ("I'm happy for X to happen...", "I've no problem with Y...", but there is a clear sense that this is a passive process on their side, and underlines the view that expecting researchers to manage data for future use without reward (or even funding) is an unsustainable strategy: it will always play second fiddler to the tasks that gain reward and recognition, as the quote from the Blue Ribbon Task Force which introduces the next section illustrates.22

Themes related to Dealing with Sensitive Data

Data trade within the research field, and beyond it

Preserved digital assets are nonrival in consumption because once one party preserves the assets, they are for all intents and purposes preserved for all. In these circumstances, the incentive for any single party to incur the cost of preservation is weakened, since the other parties can free ride on the benefits.

- Blue Ribbon Task Force Final Report, p28

The data trade is schizophrenic: a zero-sum gift economy within the academy, but a capital resource economy when it comes to dealing with industry and knowledge transfer. There are, of course, different levels of exchange, and different strata of access permission and restriction: intra-research group, intra-institution, intra-discipline, intra-domain, and so on.

In terms of gaining access to datasets created by others, researchers reported that their major problem was in identifying and then making contact with the appropriate contact that could provide access to the data in an appropriate format or level of granularity. (Robin North: "most of the datasets that are publicly available are much more aggregated than I can use, so [I] tend to end up being reliant on persuading people to give me either raw or unprocessed, or high resolution datasets from somewhere or other, and that then relies on knowing who to ask, and getting on with them well enough to be able to get the datasets.")

The process for gaining access may be made via a formal or an informal approach, or indeed on an entirely impersonal "right-click, save-as" level. The prevailing perception is that within the academic

---

22 For his part, James Tate was adamant that data collectors should be recognised for their part in the scientific process: "we collect all the data; that's science as well as just doing some data analysis and interpretation. [...] The facility, and the people that are associated – co-authors, et cetera – should be rightfully acknowledges as appropriate."
Security and sensitivity: barriers to openness and sharing

**RN:** A lot of the things that we are measuring are regulated air pollutants.

**MD:** So there are EU regulations, et cetera?

**RN:** Exactly, there are regulations and the member states are obliged to try and hit those regulations, and that cascades down to the local authorities being responsible for the air quality in their area. And there are then a set of cascaded processes about declaring air quality management areas, hypothesizing funds to deal with it, showing progress against objectives and so on, so there's a whole sort of political machine and administrative machine associated with dealing with this type of data. The way these data are recorded historically is very different to the way we are measuring them, although nominally the units and the quantities we're measuring are the same. So we're measuring NO$_2$ in parts per billion or whatever, and the regulation is specified as NO$_2$ in parts per billion, but there it's specified in terms of an annual average and we're measuring it second by second. So the numbers that we see, in particular the peak numbers that we see, are enormous by comparison with the annual average values. And it's not actually necessarily that they're in breach of anything, but it looks awful, and so there's a lot of sensitivity from the public authority partners about how data is presented, particularly publicly.]

- Interview with Dr Robin North

**JT:** They don't very often monitor at sets of traffic lights. When do cars emit? When they're accelerating. Where do vehicles accelerate? At traffic junctions. Do they ever monitor at traffic junctions? Oh no. Where are people exposed to pollution? When they're standing at a traffic junction to cross a road.
Is this because they don't want to get... they don't want these readings? Because they're particularly high and they might be, you know...

You could say that some authorities might not be looking for the worst locations, shall we say.

- Interview with Dr James Tate

As North and Tate note, one of the hot topics covered by MESSAGE was the political sensitivity of the data collected by the sensors they developed. Pollution levels are subject to European legislation as regards maximum permissible levels, and there was concern that the mobile sensor data would bring to light areas where pollution levels measured above the maximum levels. This is not to say that any of the local authorities were in breach of the regulations – the maximum levels are measured over a period of time, and readings are taken as an average of the inevitable spikes and troughs – but high pollution readings never look good, hence there were real political sensitivities over what were, on the surface at least, fairly mundane datasets:

[T]he demo day was kind of a risk day, because that's when we actually had data on show. We've been very careful to show data as relative colours, not absolute values, so as far as I am aware there is virtually no data in the public domain that has values on it that could be set against European targets. We've also been very clear to say that the data we've collected is largely uncalibrated and unchecked and therefore has to be interpreted. (Professor Neil Hoose)

While the external viewer may be forgiven for initially viewing this data as fairly mundane, harmless stuff, working on the project made Cohen realise that the political sensitivities that accompany research in this field extend far beyond the real data to include synthesised / 'non-data' as "something that could cause issues."

As a data manager, Cohen is aware of the potential pitfalls accompanying a large-scale, multi-partner project, particularly when the system that he has built uses data captured from sensors that are still under development. Simple and seemingly innocuous data are more than the sum of their parts, for political and other reasons. These sensitivities even extend to simulated datasets, and the researchers had to be very careful not to let 'non-data' appear on (e.g.) presentation slides without clear explanations that they were not real values...

I'm really realising now, only probably in the last month or two months, how important the data is that we're capturing. Even if it's just prototype systems generating basic development data for testing the systems that we have, and so on. From that point of view I think data... I realise now data potentially has very, very significant values but also could cause serious problems. […] Even if our sensor is not properly calibrated and what it's providing is a load of rubbish, if somebody gets hold of that data there can be a problem. […] And from that point of view I realised then that you have to be very, very careful about what you do with your data even if it's not data that […] you class as real data. (Dr Jeremy Cohen)

Neil Hoose summarises the issue succinctly: "The degree of sensitivity is not to the data itself, it's to the interpretations that could be put on the data..." Naturally, this influenced the degree to which data could be shared within and beyond the project boundaries, with the emergence of 'data corrals'. It was noted that each of the academic partners was constrained in terms of exactly what they could share, and when. As North puts it: "Anyone that's running a test site is working in conjunction with their local authority, and that will include both the council and the transport authorities in that regard, and depending on what arrangement and what deal was struck with the council about access to data and sharing of information and so on, and publicity of this information, there are different constraints on the data, and different constraints on the extent to which people are willing to share." So we can see at once how the different perspectives impact on the ability to share data, and make it open to users: what appears to the system architect (or case study author) a fairly mundane set of factual readings, can be a political 'hot potato' for a regulatory compliance officer at the local authority.

---

23 During the project's lifetime, little or no data appeared to be shared beyond the project consortium, although Leeds allow their post-graduate students to use real data for analysis subject to strictly observed terms-of-use agreements.
MD: I don't want to stir or play devil's advocate, but I'm anticipating that there is the possibility that a set of circumstances could emerge whereby...

JT: Newcastle don't want to share their information?

MD: No, whereby you do share the data, Newcastle University don't mind, but Gateshead Council go nuts because they don't like, you know, what's in the data. And that's not a political problem for you, you know, in Leeds, but it could conceivably be one for them in Newcastle. Is that a possible... and I'm not saying a likely, is that a possible scenario?

JT: It could be for some information streams. Probably not most. But for some it could be, possibly, but it's pretty temenous.

MD: I'm asking because I'm interested in the degree to which agreements are formalised, as you might have... you know, I do tend to bang on about it, and it's incredibly tedious but...

JT: No, it's important. [...] I'm an engineer, I don't like any mucking around, I know how I want things to be, this is how it is, it's got to be formalised. You get this information, you don't get that. The reason why you get this is that. Let's discuss it, agree it, write it down, move on, because I want to do some research, not get bogged down in a load of imprecise discussions.

MD: And politics, you know.

JT: And politics, yeah.

- Interview with Dr James Tate

What is clear is that barriers to data sharing may emerge from many sources: there are self-imposed barriers (relating to confidence or lack thereof in the quality of data and documentation); internal influences such as research group agreements and institutional ethics policies; external influences such as legal and regulatory issues; and perhaps even technical barriers.

**Agreements: formal and informal**

*Even the strongest incentives to preserve will be ineffective without explicit agreement on the roles and responsibilities of all the actors —those who create the information, those who own it, those who preserve it, and those who make it available for use. There are many different ways of allocating those responsibilities among stakeholders, and often the allocations occur naturally. This is the case in much of higher education, for example, where universities support scholars to produce scholarship, and they support libraries to preserve that scholarship and make it accessible over generations.*

- Blue Ribbon Task Force Final Report, p22

When asked about the extent to which agreements regarding access to and ownership over data had been formalised within the project, researchers frequently used terms like "gentleman's agreement" and "unwritten understanding", and references to "the proper thing to do" and "the British way" of doing things. There are obvious difficulties with this, and similarly there are difficulties with taking a more formal/rigorous approach with regard to granting (and monitoring) permissions.

Added to this, when we spoke about the difference between the public and private sectors, Neil Hoose notes that the "academic environment is much freer, people feel more empowered in many ways. They have fewer rules and regulations to continually fight with." From some interviews I got the sense that there was a desire towards increasingly formalised collaboration agreements with specific regard to data, but that the public sector (specifically academic) environment/atmosphere acted as a barrier. Robin North reported that the "local councils are very sensitive about anybody having access to [their] databases, including other partners within MESSAGE." However, the process for gaining access was not, in his view, adequately formalised, which was the root of some problems. James Tate also reported friction between the various project partners...

JT: [T]here has been some friction, you know, I would be lying if [I said] there wasn't, but I try to formalise things, agree things with good reason, be open and honest and then... and do the best job you can, and I think that's all you can do. [...] I think quite often you need to agree these things... let's write it down. Let's agree it, let's debate it, let's agree this is the basis to move on from. And quite often it's not.

---

24 Hoose also noted that his role as project coordinator was hampered by an inability to be prescriptive; the academic consortium is a much looser arrangement than its private sector equivalent, and the project manager has few tools at his disposal beyond cajoling and persuasion, particularly when it comes to academic researchers based in different institutions.
MD: The difficulty is for many projects is that this is the first time that these two people from these two institutions have even met each other and it's difficult to... nobody wants to give any ground. So gentlemen's agreements tend to emerge just because nobody actually wants to go there at that point.

- Interview with Dr James Tate

So there appears to be a widespread resistance within academia towards making things particularly explicit, especially early in a project when the partners may be tentatively getting to know one another. However, for a project that was, in the project coordinator's own words, "so disparate across so many institutions," and which involved not only academic but private sector partners too, it might have been fruitful to have pinned down data-related agreements early in the development of the work. Indeed a memorandum of agreement covering confidentiality and intellectual property rights was insisted upon by some of the industrial partners, early in the project's lifespan. This took the entire duration of the project to pull together, and was finally signed the day before the project ended, highlighting another area of difference between the formal, agreement-based world of private companies, and the (often over-) relaxed academic sphere. Whether this was a worthwhile use of Hoose's time, or an unenforceable and inflexible bit of bureaucracy, is debatable. It might have been better included as a condition of grant, with an agreement drafted and mediated by the funder(s).25

Early stage formalisation, of course, is not a panacea. There is a risk that attempting to fix agreement too early in a relationship may lead to argument and possibly nip a promising project in the bud. As Robin North puts it, "in an ideal world it would have helped quite a lot to have had that as an open discussion and have our principles right from the beginning. I'm not sure how possible it would have been. I think, as you say, it's quite difficult to have those sort of cards-on-the-table, open discussions with people that you maybe don't know so well or haven't got such a strong relationship with. And I'd say even people who do know each other well, there's a lot of stuff... they're kind of, in other projects they're competitors, so they're not necessarily going to reveal other aspects of things."26

And there are multiple levels of agreement, some of which span multiple research projects and other items of work. The fixed monitoring sites in Leeds are operated in conjunction with the local authority, with a two-way information exchange between the University and the Council. The arrangement is not explicitly formalised at the moment (Tate describes it as 'a gentleman's research partnership agreement'), but more formal partnerships and arrangements sharing of costs for joint facilities are being investigated. At Newcastle, Lakshmi Suresh reported that agreements with third party data providers are formalised via mechanisms such as signed data protection forms, and members of the research group at the university sign up individually. These procedures cover both the raw data and aggregated/derived datasets, and their subsequent storage.

The project also had a formal agreement with the Department for Transport, that project outputs and MESSAGE-related publications would be provided to them in advance of publication, in order to allow ministers and civil servants to prepare for any questions that might remerge as a result. The project would have broken their agreement with the Department for Transport if they had failed to make them aware that something was going to be put in the public domain; however, it was stressed that the DfT had no right of veto and no right of embargo. The purpose of the agreement was straightforwardly 'forewarned-is-forearmed,' and it gave the civil servants and politicians time to prepare answers to any questions that might emerge as a result of the project's findings.27

25 Neil Hoose: "[I]t's kind of a waste, because no real IPR was ever discussed at [the Advisory Board] meetings. It was protection... [...] They needed to have it a certain way because otherwise they couldn't be part of the project. So you've got academics, you've got corporate and you've got public sector [...] All of whom have got different drivers, and therefore different things they will sign up to." In a later email, Hoose added that "this sort of pre-condition can delay the start of a project significantly with consequent problems for staff continuity, access to resources and maintenance of a 'leading edge'. The academic partners did have in place a consortium agreement at the start; it was the NDA [non-disclosure agreement] covering the external parties that took the time."

26 One wonders to what extent the unwritten (or at the very least dispersed) nature of the UK Constitution is an influential factor in this area, but this may be an extrapolation too far...

27 Similarly, Leeds allow research students at the department access to the real data, provided they sign an agreement governing its use. Publications citing such data are checked internally and shared with the local authority to avert political problems.
Data management plans

*I suppose it's fair to say that we had what might be classed an abstract plan.*

- Dr Jeremy Cohen

At the grant application and post-funding stage, there was no pressure from the project's funders to come up with a data management strategy or plan. As I was at the time developing a checklist for Data Management Plans,28 I was very interested in learning researchers' views on the potential benefits of developing these at the early project stage.

With hindsight, Hoose acknowledged that the data management issue is an area of greater complexity than he initially thought, and that as the project (and the case study) developed his awareness increased. "I think with hindsight [we] might have given it more consideration then [at a consortium level], but we sort of never got that far [in the creation of larger datasets] that we got to the issue of it."29

Jeremy Cohen concurs: "I think I realise that you can make things much more efficient by thinking about a lot of these things beforehand. […] When you're putting together a big plan for a very large project that's going to build a large system, you've got so many things going through your mind about how this thing is going to work, and what it is that you're proposing to do, that sometimes it's easy to forget some of the things which may actually seem quite obvious when you look through and see them written down."

Robin North is more guarded. When asked whether he felt that some of the data-sharing issues experienced between the academic partners might have been leavened by the early introduction of data management plans covering (for example) issues such as data sharing, access, intellectual property right, embargoes, and third parties, North conceded that "there are aspects of that that we could definitely have done better," but when I spoke to him the second time he felt that some of the frustrations expressed in the first interview had been assuaged in the interim. "I think with the multi-site set up it's difficult to ensure that each of the locations does things in a compatible way." That said, he does welcome the introduction of data management plans at the application stage, particularly as in his view this would aid cross comparability between environmental datasets, which are often impossible to replicate, thereby assisting scientists in deriving increased value from multiple meta analyses.

It is reasonable to conclude that a lack of precise allocation of responsibility for data management for the longer term leads to a risk that the datasets will be lost, or inaccessible or incomprehensible, which may be worse. When I asked who was ultimately responsible for the data over the longer term, one respondent laughed and said: "I suppose I assume it's not me." A culture of assumptions and lack of clear responsibility unquestionably leads to risk, which can be managed to a considerable degree by the development and adoption of data management strategies early in a project's lifecycle.

Further Themes

Problems with the short term project funding/organisational model

*…all the instability that's inherent in the whole thing.*

- Professor Neil Hoose

The academic research environment tends towards devolved decision making, and independence of practice, which makes prescriptive policies difficult to monitor and enforce.30 This is, of course, a further reaching problem than any single body can hope to address, and goes to the heart of the short term, project-

29 At the same time, Hoose worried that adding extra stages to the grant application process detracted from the amount of time that already busy researchers were able to spent actually doing their research. This is indicative of a challenge that the DCC and other bodies face, namely encouraging researchers to view good data management as integral to good science, rather than as an additional step or burden.
30 There are areas of exception, such as ethics, where the central University or Faculty/College/Department will usually take a strong position with censures for scholars who breach its rules.
based research funding model. This case study offers no solution to the problem, but only wishes to remind readers of its existence.\(^{31}\)

The classic research data scenario may be succinctly simplified thus:

- a researcher creates a dataset for the purposes of research;
- that dataset is analysed and a research paper is written which describes the researcher's findings;
- the findings are published in a journal, and the dataset which underpins those findings is left to atrophy and one day disappear altogether.

This situation has been remedied to a significant extent in recent years with the development of research council policies on data management, and the creation of specialist bodies to support best practice; indeed, the encouragement of good data management practices for the longer term is more or less the Digital Curation Centre's core mission. However, the changes wrought by time are not something covered or addressed by the short project set-up that funders fund. In short, the crux of the problem is a straightforward paradox, i.e. that long term preservation activities are generally funded by short term funding, but that project plans often do not incorporate data management costs (short or long term), and future workloads seldom include looking after things from the past...

I'm interested in how we can keep a long term store of this data and be able to add things to it over time. So I think that there must be a lot of opportunities for adding metadata either because we can identify new information in the future which we can then back process data and add metadata into what's stored, or derived data or whatever. I think there's a lot of scope there, but that's probably not within the scope of MESSAGE. (Dr Jeremy Cohen)

We've been focusing so hard on making it happen and making it work at all, that the question of, "So are we all going to agree to keep this running thereafter, and is there going to be funding available to keep this running thereafter?" hasn't really been addressed. (Dr Robin North)

A related issue is that of sustainability. While project budgets do not explicitly cover the costs associated with longer term management and stewardship of data, the data remain at risk. Anything which is not explicitly supported is implicitly expendable, as the Blue Ribbon Task Force report notes:

[T]he incentive to preserve diminishes as the decision-making unit becomes more granular—the most granular being the individual researcher. Generally, and in grant-funded research in particular, preservation is framed as a zero-sum activity. Time and money spent on preservation activities are deducted from the total budget for research. This has major impact on any preservation incentive, whether mandated or not.

- Blue Ribbon Task Force Final Report, p58

In addition to financial sustainability, there is also the issue of sustaining knowledge and skills, a goal which is problematised to a huge degree by the prevalence of the short term project model. Robin North speaks of the "nature of post-docs" as being "a transient position: people come through that and do something interesting for a few years, and then generally what happens is the expertise leaves and they become institutional, they don't necessarily propagate that to their successors, so [...] actually it's really a good idea to design [the data management infrastructure] properly in the first place."

When it comes to skills development, the general perception is that non-specialist data managers are crying out for training, but one interviewee expressed a view that "with lots of academics there's an arrogance associated [...] 'We don't need to go and talk to other people, we'll just do it our way.'" If this is in fact a widespread attitude, the endeavours of the DCC and other bodies to bring data training to frontline researchers, especially more experienced researchers with their own established data management habits,

\(^{31}\) If, to paraphrase Paul Valéry, a research project is never finished, only abandoned, there is a problem reconciling short term project funding with the increasingly prevalent Web 2.0 idea of perpetual beta. It is clear that these are not readily compatible models.
may be doomed to failure. It is clear that increased service visibility will be necessary to establish whether it is indeed the case.

**The 'keep everything' mentality**

*I'll keep the datasets for… ad infinitum, personally. I'm a statistician, so no data is ever obsolete.*

- Dr Ben Waterson

We have reached a stage where storage space is sufficiently cheap that researchers expect to be able to keep all of the data they capture or create. Examining Waterson's language in the quotation above – 'I will keep the datasets…' not 'I would keep the datasets…' – it is clear that the eternal availability of data is an expectation, its preservation and interpretability perceived more or less a given. The researchers interviewed intend to keep everything that's real (i.e. everything except the simulated dataset), or at least, they want to keep it. (N.B Cohen doesn't keep simulated data for long, but it's sensitive during its short lifetime; Waterson holds on to everything.)

Other researchers expressed the same viewpoint, as the following three quotes illustrate...

JT:    We don't ditch anything, we keep all the event data. [...] It's quite cheap to store, you know, it's just numeric and text so you can keep that kind of... that's not a problem anymore.

    - Interview with Dr James Tate

RN:    The sensor stores the very raw… the rawest possible information and then there's a client programme that sits on top and allows you to display whichever parameters you want.

MD:    Pull out what you're interested in, and presumably you only store the parts that you're interested in?

RN:    Well I archive all of it because actually the raw file's not that much bigger anyway.

    - Interview with Dr Robin North

MH:    [A]t the moment the idea is that we don't throw anything away because we don't necessarily know what's going to be valuable and what isn't. So that's in a nutshell our current approach.

MD:    So keep everything?

MH:    Exactly.

    - Interview with Mr Mark Hayes (Cambridge)

There are numerous problems with this, not least the fact that bitstream preservation (the simplest form of digital preservation) makes no contribution towards the understandability of the information contained within. Additionally, the more data you have, the harder it is to find what you want. While it may be true that 'no data is ever obsolete,' the flipside of this is that data may become inaccessible, obfuscatory, incomprehensible, or worse, misleading. It follows that there is a need for advocates of data disposal as well as data preservation, and that appraisal and selection need to be built into data management training programmes.

Among the interviewees, only Jeremy Cohen, a data manager rather than a domain scientist, expressed any disquiet about this...

[T]here's the more subtle aspect of, you know, you get data that's perfectly capable of being stored, but is it the right data to be stored? Actually, those issues have been much more interesting because I suppose they're the bits that make your system useful or not useful, if you store too much data that's not of any use you can invalidate the rest of the good data, you know, you can really spoil things[.]

Finally, contrary to general belief, we cannot, in fact, store everything indefinitely: a recent IDC White Paper highlights the growing gulf between the amount of digital content being created globally and the storage available to house it.\(^\text{32}\)

---

The user's data experience

A minor theme which emerged through conversations with Jeremy Cohen in particular was the importance of ease of use when it came to engaging with domain scientists and the advantages of good presentation, organisation and formatting of datasets and the interfaces used to query and manipulate them. The general perception was that researchers are willing to take time to learn new systems if the benefit of doing so is made apparent, but that it was preferable to make the experience as painless as possible...

One of the things that we did a few weeks ago was to build a simple Web-based interface that will allow the researchers on the project to type straight SQL queries into a Web page and then get the data back in CSV format that they can put straight into Excel. Which is pretty simplistic, but in fact that's just the sort of thing that people want to do and the sort of thing that's not terribly simple for people to do otherwise. [...] People are willing to learn these things, and we could spend time teaching them, but in fact it's a whole lot easier if we say go to this Web address, type this username and password and then just type your query in the box... which they do. [...] We've got the ability to pull out metadata from the database, so they can view the table structures, the schema, on the same webpage. You select which database, you click a button and then the schema comes up. [...] The main thing is that they just want to get files in a certain format, and they can easily do that. (Dr Jeremy Cohen)

Interestingly, one of the principal drivers for internal data sharing and ease of use was the need to have direct, common access to "all the data on a central server that we can all access in a format that we can all read easily" for "more campaign-driven information". Here we see impact measurement begin to emerge as a force majeur influence on data sharing.
**Part 3: Conclusion**

**Broad conclusion: On the comparative difficulties of human and technological data-related issues**

This study's major themes are the identification and development of skills needed to manage data during the live project phase, and to set it up for a long life post-project; the importance of the human infrastructure in managing data; and the passive attitude that prevails among researchers when it comes to data-sharing. For researchers, data itself is not the be-all-and-end-all: it is a means to an end.

The conclusion therefore supports the view that the focus of bodies such as the Digital Curation Centre needs to remain on human data-related issues, which are more pressing, more messy, and more difficult than their technological counterparts, and the observations detailed herein (and subsequent feedback from the researcher-interviewees) support the view expressed widely in recent years (notably in the Blue Ribbon Task Force report, ibid) that increased effort must be made to embed best data management practice from the outset: doing so ensures longer term access to the data that underpins the records of scientific research.

To achieve this we need to understand our targets better, and couch guidance in terms that stand a better chance of being readily accepted by its intended audience.

**RECOMMENDATION:** That techniques such as use cases, storyboarding and profiling be investigated and exploited with a view to understanding and meeting user needs in this area.

**Specific conclusion 1: On the visibility of advice**

The interviews also indicate that training and advisory bodies, including research support services within HEIs, need to make themselves more visible to researchers in order to be in a position to offer guidance at early stages in both projects and careers...

> [T]here's a lot of lessons I've learned in doing it that would probably have been things that I've learned the hard way and it would have been a lot easier to make use of some of the expertise you guys have outside. [...] My attitude towards it at least is that we need to learn from what we've done this time round and do it better next time, so hopefully this is an important part of the process. [...] There are quite a few things we could have certainly improved upon. But I don't know if given the same decision points each time I would have done anything differently. (Robin North)

While North notes that he "would certainly echo the need to build in data curation concepts earlier in the project lifecycle," he also emphasises "the need for external support in doing this, as the nature of research projects is that very often the personnel involved have not worked on anything similar before. This may be especially true when looking at large inter-disciplinary projects."

**RECOMMENDATION:** That training and guidance continue to be developed with sensitivity to the needs and perceptions of researchers, paying particular attention to issues of appraisal and selection which currently appear to be largely overlooked.

**Specific conclusion 2: On the usefulness and appropriate scale of data management planning**

The interviews and observations support the view that good data management planning at best prevents and at worst mitigates problems that emerge during the research project lifetime. There is a fundamental disconnect here between the formalisation of tightly bound data-related agreements and operational management plans, and the looseness/fluidity which characterises the research environment, particularly in older universities. The general feeling seems to be that top-down, dictatorial efforts will be resisted or simply ignored, and that consortium agreements are 'a waste of time' and 'utterly unenforceable.'
Until such times as higher education changes its prevailing culture, the topic will need to be tackled sensitively and via consensus building, which is always a slow process...

One learning point I guess is that it's quite time consuming to build a good structure for the datasets and to build [...] a feasible, practical way of recording the relevant metadata. It's time well-spent though, but in some cases it's time that you can't… you have to try and do it once and get it wrong before you can do it again and actually get something that's right and robust, so maybe having ways of evolving a data curation process become very important. So I think it's something we've talked about a bit before is this sort of data curation cycle, data cycle, and I think that's... I have more of an awareness of that type of process now, which is... I think will be important next time I end up in a similar situation. (Dr Robin North)

It is always difficult to conceive of how data will be used until you actually come to try and use it. However the idea of paying more attention to this aspect up front has to be right. (Professor Neil Hoose)

While supportive of increased attention to data management planning, Neil Hoose was wary about the potential for introducing an additional layer of bureaucracy that could slow down the actual starting of projects, so there is clearly a job to be done here in convincing researchers (and project managers) that time spent planning data management more than pays for itself in the long run. Some amount of this persuasive work will be taken care of by research funders increasingly making solid data management planning a grant prerequisite: the Economic and Social Research Council has recently announced a new policy which gives increased weight to this activity.33

Hoose also suggested that model data management plans/agreements may "solve the issue in an efficient way"; my objection to this would be that providing sample plans detracts from the researchers' actual engagement with the issues, and runs a very real risk of leading to a 'tickbox culture', precisely the additional layer of unengaging bureaucracy that we wish to avoid. (At the same time, Hoose noted the importance of making data discoverable, although MESSAGE data was not made externally (nor always internally) discoverable during the project's lifetime, and no plans appear to have been in place to share it more widely post-project.)

RECOMMENDATION: That funding bodies provide increased support to grant-holders and applicants in the production of data management plans and consortium agreements, and place more value on their production.

Specific conclusion 3: On the need for further study

This study covered a broad range of topics, and in writing it (and re-reading the interview transcripts) one of the major frustrations was that I did not have the opportunity to go into more depth asking detailed questions about consortium agreements, data sharing arrangements, formal contracts, and so on. This emerged as an unforeseen secondary focus of the study, and would perhaps make for an interesting case study in its own right.

RECOMMENDATION: That a relevant organisation commission a subsequent study to concentrate on these issues.34

33 http://www.esrc.ac.uk/ESRCInfoCentre/Images/ESRC_Research_Data_Policy_2010_tcm6-37350.pdf
34 N.B. I should stress that I am not suggesting that such as study be on the same project (or its successors), nor that it have the same author as this.
Afterword

This report starts with a somewhat gnomic quotation from Wallace Stevens, which I should probably explain before it concludes. The rapid recent increases in data production and storage are commonly and collectively referred to as a 'deluge', which brings to mind images of rain, the flood, and so on. I feel that snow is the better metaphor: rain drains away quickly, while snow has a tendency to remain somewhat longer, yet still impermanently. Capturing rain is a one-chance-only activity: if you miss your chance, it's gone. The act of shovelling snow is less frantic: provided the temperature remains cold enough, you can take a little time to develop a strategy. Most of the data we have is not at immediate risk: as the Blue Ribbon Task Force report notes, minimal intervention will ensure the opportunity to delay decisions until such a strategy can be arrived at. But the nature of an individual dataset is itself snow-like: the original data is deepened and compounded with each snowfall via additional metadata, ancillary materials, and the simple yet inexorable passage of time, which changes… if not everything we see… then certainly the ways in which we see it.

You could not step twice into the same river; for other waters are ever flowing on to you.
- Heraclitus

Time present and time past
Are both perhaps present in time future,
And time future contained in time past.
- T.S. Eliot, Four Quartets 1 ('Burnt Norton')

Sometimes the recompense arrives
so far ahead of what you'll give
that you will fail to recognise
the reciprocity,
the love

that circles in the universe:
this life a grace advanced, its knack
to meet requital with its cause -
the offering up, the giving back.
- Kona Macphee, 'The Gift'

Acknowledgements

This case study is part of the Digital Curation Centre Phase 2 work funded by the Joint Information Systems Committee (JISC), to whom the author would like to express his gratitude. On the MESSAGE side, particular thanks are due to Dr Robin North and Dr Jeremy Cohen, and to project coordinator Professor Neil Hoose, all of Imperial College London, who were instrumental in setting up the study and maintaining its momentum.
## Annexes

### Annex I: List of interviews

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Medium</th>
<th>Role in project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mr Mark Hayes</td>
<td>Cambridge</td>
<td>Telephone</td>
<td>Distributed Computing</td>
</tr>
<tr>
<td>2 Dr Jeremy Cohen</td>
<td>Imperial</td>
<td>Face-to-face</td>
<td>e-Science, on-demand utility computing, middleware</td>
</tr>
<tr>
<td>3 Dr Robin North</td>
<td>Imperial</td>
<td>Face-to-face</td>
<td>Vehicle emissions monitoring, transport technologies</td>
</tr>
<tr>
<td>4 Dr James Tate</td>
<td>Leeds</td>
<td>Face-to-face</td>
<td>Transport and the Environment. Traffic flow and congestion, Vehicle emissions, Dispersing air flows and roadside air pollution using both monitoring and modelling tools</td>
</tr>
<tr>
<td>5 Mrs Visalakshmi Suresh</td>
<td>Newcastle</td>
<td>Telephone</td>
<td>Databases and middleware</td>
</tr>
<tr>
<td>6 Dr Ben Waterson</td>
<td>Southampton</td>
<td>Telephone</td>
<td>Transportation Research. Quantifying traffic conditions based solely on combinations of detector data streams</td>
</tr>
<tr>
<td>7 Dr Jeremy Cohen</td>
<td>Imperial</td>
<td>Face-to-face</td>
<td>e-Science, on-demand utility computing, middleware</td>
</tr>
<tr>
<td>8 Professor Neil Hoose</td>
<td>Imperial</td>
<td>Face-to-face</td>
<td>Project coordinator</td>
</tr>
<tr>
<td>9 Dr Robin North</td>
<td>Imperial</td>
<td>Telephone</td>
<td>Vehicle emissions monitoring, transport technologies</td>
</tr>
</tbody>
</table>
Annex II: Interview question framework

Case Study Interviews

Interviewee #:

Interviewee name(s):

Interview date:

Institution(s):

Discipline(s)/specialism(s):

Interview medium:

Location (if applicable):

Please note that interviews are intended to be semi-structured. This guide is not a script, and so questions will vary from case to case.

PART 1 – Introductions

1. Introductions
   a. Background to the case study
   b. Case study methodology
   c. Your background: institutional role and research interests

PART 2 – The MESSAGE project

2. Project rationale
   a. Aim of the project
      i. What was your understanding at the outset?
      ii. What is your understanding now?
   b. Aims and expectations of individuals (or, "What will MESSAGE do for me?")
3. Your role within the project (data creator, data manager, data re-user and/or data scientist?)
4. Project methodology and dynamics (inc. views on project management, priorities, policies, cohesion)
   a. Key drivers, influences, biases
   b. How were decisions taken, and how were these communicated? Shared best practice.
   c. Main obstacles
   d. Specific successes

PART 3 – Data, curation and preservation

5. Overview of your data-related activities
6. Key characteristics of the data created/used
   a. File formats
   b. Data volumes
   c. Software, tools and infrastructure (present and desired future, inc. storage arrangements)
   d. Metadata: methods of annotation for added-value/usability/sharing (automated or manual)
   e. Third-party data (provenance, trust, authenticity)
7. Curation issues, drivers and constraints
8. Metadata requirements and solutions
9. Quality management (calibration, sampling, validation)
10. Security
11. Criteria and mechanism for preservation selection/deposit
12. User knowledge/training requirements (for you; for user community)
13. Ownership/licensing/reuse issues (inc. IPR, funding body and institutional policies)

PART 4 – Implementation
14. Assessment of the MESSAGE pilot: positives and negatives
15. Services: which are essential, and are they sustainable?
16. What are/should be the immediate priorities? To what extent is data a factor in current and ongoing work?

PART 5 – Future work
17. Implications/opportunities for sustaining and building upon data curation in your area
18. What has your work on MESSAGE taught you about data and/or data curation?

Thank you.

Martin Donnelly
Digital Curation Centre
University of Edinburgh
martin.donnelly@ed.ac.uk
www.dcc.ac.uk