Machines, Methods and Music: On the Evolution of e-Research

David De Roure
Oxford e-Research Centre
University of Oxford, UK
david.deroure@oerc.ox.ac.uk

ABSTRACT

Against a backdrop of increasing computational capability we are seeing acceleration of research through broader adoption and sharing of tools, techniques and resources, both for 'big science' and the 'long tail scientist'. This paper discusses the evolution of e-Research, focusing on a web-scale computational musicology project as an illustration of emerging methodology and using the myExperiment social website as a lens to glimpse future research practice.

KEYWORDS: data deluge, e-Science, e-Research, scientific workflow, computational musicology, linked data.

1. INTRODUCTION

Ten years ago we saw that new experimental techniques, including lab automation, DNA microarrays, combinatorial chemistry, online instruments, sensor networks and earth observation, were set to produce more data than individual researchers could process using established tools and methods [1]. Partial processing would mean that results hidden in the detail would be missed, and it would be impossible to spot the patterns in the big picture. The challenges of extreme data capture, distribution and processing were exemplified at that time in the computing infrastructure being designed to support the Large Hadron Collider.

This data deluge continues today, with more data from more sources. Facebook is not so much a Large Hadron Collider as a 'Large People Collider', we report our lives for public analysis through twitter and corporate analysis through every electronic transaction, our homes are instrumented with smart electricity meters and our streets sense society in motion. Old data is being reborn-digital in digitisation projects like Google Books and the crowdsourced transcriptions of Old Weather[1]. Hidden data is being liberated by opening up government data[2] and remote access to secure data[3] provides yet another new source. Some data is specifically collected for reuse by researchers, but much is collected for a specific purpose, and much sits in silos.

At its outset ten years ago the UK e-Science Programme was very much predicated on dealing with this deluge. John Taylor, then the director of the UK research councils, defined e-Science to be "global collaboration in key areas of science and the next generation of infrastructure that will enable it" [2]. Significantly this definition understands that progress in science is not just about technology but about people working together and being empowered by technology – and the emphasis on science reminds us that ultimately success is measured by new research outcomes.

Researchers in several disciplines, from computational sciences to digital humanities, were already sophisticated users of advanced computing techniques. The programme kick-started a broader set of collaborations between computer scientists and domain scientists, and established a wider notion of e-infrastructure to support this. It facilitated co-evolution: researchers and technologists together creating and harnessing innovative technology to achieve new research outcomes. The data deluge is caused by, and needs to be handled by, innovation in automation and by the new scale of participation of scientists in the digital world.

This is the digital research ecosystem and in it we can observe three phases of e-Science: the early adopters of new tools, followed by a phase of embedding and re-use and then, building upon this new sociotechnical platform, a world of open science and radical sharing. Importantly this is against a backdrop of increasing computational capability and increasing everyday participation in the digital world.

The three generations are discussed in the following three sections, then in section 5 we look at a computational musicology project as an example of emerging

[1] Old Weather Zooniverse project, http://www.oldweather.org/

[2] Opening Up Government, http://data.gov.uk/

[3] Secure Data Service, http://securedata.data-archive.ac.uk/

methodology and practice. We close in section 6 with some discussion points about research in ten years time.

2. First Generation: Isolated adoption

This generation of e-Science is characterised by researchers using tools within their particular problem area, with some reuse of tools, data and methods within the discipline. Traditional publishing is supplemented by publication of some digital artefacts like supplementary data. Science is accelerated and practice is beginning to shift to emphasise *in silico* work. For some this was circa 2001-2005 but the timing is domain-dependent.

These tools help with the local data deluge but more is needed. When a scientist conducts a series of experiments in a lab they know how to interpret the measurements, but when we have petabytes of data flowing around in our increasingly automated 'distributed global collaborations' we need to make sure we capture context crucial to support interpretation and reuse – preferably in a machine-processable form and using automated techniques too, so that we can handle the scale. While we may know the intended use of the data, we must equally plan for unanticipated reuse, which is inevitably challenging.

Chemistry researchers at the University of Southampton recognised this challenge early on and introduced the "publication @ source" initiative which sought to create a complete digital chain of knowledge from the scientific laboratory through to the scholarly output [3], so that the scientist reading a number in a paper or on the web page could chase back and see exactly where it came from and how. This provenance is crucial to interpretation and reuse but also to trust. At a time when projects tended to "warehouse" data, this ethos of publishing was also significant: augmenting the Web rather than building data silos [4].

This particular collaboration between chemists and computer scientists illustrates co-evolution in action: the computer science team did not capture requirements then go away and engineer a solution, but instead the chemists were empowered to harness the technology – in this case Semantic Web. This led to a rich set of outcomes including semantic lab books, web sites and blogs, and semantically enhanced publications, all supporting data capture and reuse [5].

While the chemistry work emphasised the smart laboratory, in bioinformatics the research was *in silico*, using data analysis pipelines – data flowing through a series of online processing steps, providing the automation essential to relieve manual drudgery [6]. These in silico workflows have become a key automation technique for systematically handling the data deluge, and they have given us the workflow as a new sharable artefact of digital science – to record, repeat, reproduce and repurpose an experiment.

Observers of the digital research ecosystem may note just how many of these workflow systems there are [7]. This is because each one comes prepackaged to solve particular problems for particular research communities, familiar when they open the box. This is another example of co-evolution in action, and demonstrates that adoption of a technology comes by focusing on the specific before the generic [8].

4. Second Generation: Investing in re-use

The second Generation is characterised by facilitated reuse of the increasing pool of tools, data and methods across areas/disciplines. We see some freestanding, recombinant, reproducible "research objects". New scientific practices are established and, through sharing, opportunities arise for completely new scientific investigations. For some this phase was around 2006-2010.

Paul Fisher, a bioinformatician in Manchester, had a data deluge in his research on sleeping sickness in cattle (trypanosomiasis). Manually 'triaging' the data to work on promising subsets left results undiscovered, but using a scientific workflow system to process all the data systematically he found the result he was seeking [9]. Then Paul shared his workflow with a colleague who was working in mouse, and she used it to find a result that she had also been seeking for some time. Thus a new scientific outcome resulted from re-use, and we see one of the benefits of sharing workflows.

In 2006 we knew that teenagers were using new collaboration tools like mySpace, but it was not clear that scientists would share in that way – after all, "It's not e-science, it's me-science" as Carole Goble pointed out [10]. We built the myExperiment social website for sharing workflows as an experiment to explore this [11]. Like photos on flickr or movies on youtube, we set out to produce *the* site for workflows: New Scientist called it mySpace for the dudes in lab coats [12].

Today myExperiment has the largest collection of scientific workflows publicly available and has inspired other sites. A key feature is its attention to "social metadata": credits, attribution and licensing, without which it would be unacceptable to researchers. Scientists do share – and they do it globally, as in Taylor's definition of e-Science – but the sharing cultures vary from discipline to discipline and also between methods

and data. We can usefully distinguish between "can't share" and "won't share" – sites like myExperiment help with the former and illustrate the benefits of overcoming other barriers.

myExperiment is distinctive for its focus on sharing methods. If there is a data deluge, surely we must be developing new data analysis methods too and these could be usefully shared: although data is receiving lots of well-deserved attention, there is perhaps a neglect of methods as first class citizens. The combination promotes reproducibility, re-use and the sharing of know-how. We tend to nail down data and run methods over it, but perhaps we should also think of methods as first class citizens with data flowing through [12].

myExperiment supports the sharing of not just single items but aggregations or bundles of things, that we call packs. For example, Paul Fisher's pack has workflows, PDFs of papers, slides and service invocation logs – this makes his work repeatable, reproducible, repurposeable and referenceable [14]. Packs enable us to test workflows at a later date, an important approach to dealing with "workflow decay" as the external environment of workflow execution remains in flux.

Packs are making an interesting contribution to the important debate on the future of scholarly publishing. We see packs (co)evolving into scholarly knowledge objects that can be shared and dropped into the tooling of e-Research: they are prototypical examples of Research Objects, a notion due to Iain Buchan of University of Manchester [15].

myExperiment is just one of many examples of data and method exchange and re-use through numerous portals, repositories and Virtual Research Environments (VREs). At the same time the traditional scholarly knowledge cycle is being supplemented with new modes of scientific discourse, with blogs and wikis becoming mainstream. These collaboration-centered socio-technical systems are typically web-based and they are characteristic of Science 2.0 [16].

In the early adopter phase the effort went into effective tools to tackle specific problems, but projects like myExperiment show the effort also going into tools and environments which facilitate sharing and adoption. An important species in the research ecosystem is the "intellectual access ramp" which enables incremental engagement with new technologies rather than dropping people in the deep end (or fast lane): workflow systems are a ramp for users of the data and computational infrastructure, and myExperiment is a ramp for workflow users.

4. Third Generation: Radical Sharing

This new sociotechnical situation means we are better equipped to cope with the data deluge that predicated the e-Science programme, and with adoption of new tooling we also see the emergence of new practice. The Third Generation is the solutions we are developing now, characterised by global reuse of tools, data and methods across any discipline, and surfacing the right levels of complexity for the researcher. Research is significantly data driven and we see increasing automation and decision-support for the researcher as the environment becomes assistive with the growing digital record.

Many commentators have observed the change to data-centric practice, at least in the sciences (one might note that the arts and humanities communities have always been data-centric). The Fourth Paradigm book [17] provides a broad look into the "rapidly emerging field of data-intensive science". Doug Kell and Stephen Oliver's "BioEssay" provides an account of changing practice under the title "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis driven science in the post genomic era." [18]. More controversially, Wired magazine gave us the headline "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" [19].

myExperiment is also a fascinating social probe into research practice. We see researchers beginning to work at a higher level of abstraction; for example, acquiring data from multiple sources, building a temporary repository, querying it and publishing the results. This is made increasingly possible by the improving circumstances of reuse in the digital research ecosystem and particularly by the adoption of Linked Data[4]. As well as publishing Linked Data workflows, myExperiment itself is part of that fabric, nicely exemplified in Roos and Marshall's proof-of-concept mashup in which they demonstrate an assembly of resources to answer a research question and in doing so make a point about the future methods section of papers [20].

The increasing participation of researchers in the digital world enables sharing and "network effects", and it empowers the "long tail". Whether or not researchers upload content, their use of resources provides a basis for recommendation and trust, i.e. it adds value. Meanwhile the increasing participation of citizens promotes public awareness and understanding and also citizen science. For example, Galaxy Zoo, in which people classify galaxies, was the first in a number of successful citizen science projects jointly called the Zooniverse [21]. In his talks

[4] Linked Data, http://linkeddata.org/

Arfon Smith, one of the creators, suggests three important principles: treat people as collaborators not users, contribute to real research and don't waste people's time.

Some of the generics of the new practices are now becoming evident: Astronomers built telescopes, gained insights and a new understanding of our place in the universe; e-Researchers build *datascopes* – telescopes for the 'naked mind'. We construct the socio-technical apparatus that takes us from a signal in the real world through to understanding, whether in medicine, music or papyrology [22].

6. A Case Study in Computational Musicology

The *Structural Analysis of Large Amounts of Music Information* (SALAMI) project is a Digging into Data project[5] to support musicologists, with three international partners conducting an analysis of digital music recordings on an unprecedented scale. In its approach SALAMI exemplifies several aspects of the e-Research of 2010: high performance computation, data publication, workflows, crowdsourcing, community software development and attention to sustainability.

SALAMI is a datascope: the project inputs are masses of "signal", some 350,000 digital recordings of music including a major collection of live performances in the Internet Archive. We also have the community-maintained and annually-evaluated algorithms of the MIREX music information retrieval evaluation exchange [23]. The project outputs are Linked Data repositories of music analyses for use by musicologists, students and citizens alike. Over 1000 pieces of music have been annotated by music students at McGill and Southampton, UK: this is the crowdsourced "ground truth" that enables us to calibrate our algorithms, and is a unique and powerful resource in itself.

SALAMI is also an example of *computational thinking* [24]. The problem solving skills of the computer scientists are being applied within the research domain of musicology; i.e. the computer scientists are not just in a service role. Perhaps this synergy may result from some common issues in dealing with multiple representations and in working between hierarchical symbolic and serialised forms

Finally SALAMI is also a Ramp – the user interface is the web browser and the user sees a task-specific interface for their research, concealing the Semantic Web machinery behind the scenes. While many "semantic web browsers" are in development, we might suggest that this is actually what such a browser should be like.

Third generation projects also think about sustainability. Our sustainability model is essentially that of the Web; we are not building a warehouse that will close at the end of the project. This approach is illustrated in Figure 1 in which digital music collections provide the signal which is analysed by experts (1) and, based on this, by machine (2) in order to publish new Linked Data resources that are used to support musicologists (3). New signal and new results can be added by the community.

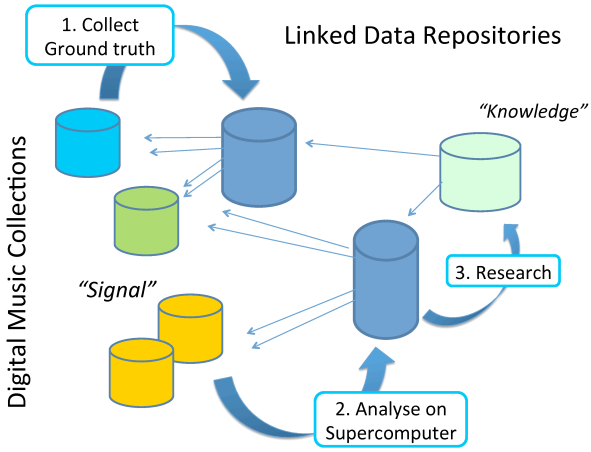

Figure 1. The SALAMI approach

Our prototype of this approach is nicknamed "country/country": it uses genre classification to measure how much country music comes from different countries. as a rehearsal for the structural analysis in SALAMI. Hidden behind the scenes we bring together multiple public Linked Data sources and use Meandre workflows [25] on myExperiment. Country/country provides the demonstration of SALAMI's web-based methodology for web-scale research [25].

10. Discussion

New data and computational resources provide huge potential to do accelerated or, perhaps more importantly, entirely new research. However, they are only as useful as the extent to which they harnessed by researchers.

With the interdisciplinary nature of grand challenges it was important to move beyond isolated successes of the first generation to more pervasive and more collaborative adoption. Second generation e-Science then saw an emphasis on the collaborative infrastructure and the ramps to widen adoption. This enables the third

[5] Digging into Data Challenge, http://www.diggingintodata.org/

generation to move to the next level on the new sociotechnical platform.

Different communities are in different phases of their "computational turn" but what might we anticipate in the next ten years? Here are four observations extrapolating from the analysis in this paper:

1. *New methodologies, new research*. The last ten years have seen the phase change from the data warehousing mindset to publication and reuse, supported increasingly by data sharing policies. The techniques to deal with this wealth of data, with its scale and imperfections, are now set to be established. With this has come the application of problem solving methods from one discipline in another, and the creation of new disciplines.

2. *Assistance and Automation*. Through co-evolution we might anticipate a balance between researcher and machine in which humans are empowered to do what they do best – the creative process of research – and machines support them by dealing with what can be automated. The richness of the digital 'footprint' of researchers enables machines to be more assistive. It also, incidentally, makes it easer for them to behave indistinguishably from humans, and perhaps it is time to revisit the Turing test.

3. *New shared digital artefacts*. The academic paper is very much ingrained as a unit of discourse, sharing and analysis, but it is a legacy of the publishing process of old. New Research Objects, used computationally as well as by humans, are set to emerge through co-evolution. In these we hope to see the primacy of method and with it reproducibility and assisted sharing of know-how.

4. *New research spaces*. In physical days the research environment was a separately equipped space. Now we have specially equipped digital spaces and they are accessible flexibly from the physical world. As well as the VRE in the Web browser, these physical and digital worlds intersect in the 'Internet of Things'. We may anticipate new means of conducting research in the digital world but also the physical space of the 'laboratory'.

Will we be doing a different kind of research? Capabilities aside, some would argue that our research training and environment blinkers our work, so with such radical change we seem set to see things differently. With not just new data but new thinking we are set for new research outcomes that we cannot even anticipate today.

ACKNOWLEDGEMENTS

I am indebted to Professor Iain Buchan of University of Manchester, not only for introducing me to his notion of Research Objects but also for the 'three generations' framework which he first suggested in the context of e-Laboratories. Thanks are due to all who have shaped and explored the e-Research ecosystem, including Malcolm Atkinson, Stephen Downie, Jeremy Frey, Carole Goble, Jim Hendler, Tony Hey, Anne Trefethen and all their teams.

REFERENCES

[1] Hey, T. & Trefethen, A. (2003) "The data deluge: an e-science perspective" in Grid computing: making the global infrastructure a reality, ed. Berman, F., Fox, G. and Hey, T. pp. 809–824. John Wiley & Sons, Ltd. doi: 10.1002/0470867167.ch36

[2] Hey, T. and Trefethen, A. (2002) "The UK e-science core programme and the grid", *Future Gener. Comput. Syst.* 18, 8 (October 2002), pp. 1017-1031. doi: 10.1016/S0167-739X(02)00082-1

[3] Taylor, K.R, Essex, J.W., Frey, J.G., Mills, H.R., Hughes, G. and Zaluska, E.J. (2006) "The Semantic Grid and chemistry: Experiences with CombeChem", *Web Semant.* 4, 2 (June 2006), pp. 84-101. Elsevier Science. doi: 10.1016/j.websem.2006.03.003

[4] De Roure, D. and Hendler, J.A. (2004) "E-Science: The Grid and the Semantic Web", *IEEE Intelligent Systems*, vol. 19, no. 1, pp. 65-71, Jan./Feb. 2004, doi: 10.1109/MIS.2004.1265888

[5] Hughes, G., Mills, H., De Roure, D., Frey, J., Moreau, L., schraefel, m. c., Smith, G. and Zaluska, E. (2004) "The semantic smart laboratory: a system for supporting the chemical eScientist", *Organic and Biomolecular Chemistry*, 2. pp. 3284-3293. doi: 10.1039/B410075A

[6] Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M., Li, P. and Oinn, T. (2006) "Taverna: a tool for building and running workflows of services", *Nucleic Acids Research*, vol. 34, iss. Web Server issue, pp. 729-732, 2006. doi: 10.1093/nar/gkl320

[7] Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., and Myers, J. (2007) "Examining the Challenges of Scientific Workflows", *Computer*, vol. 40, no. 12, pp. 24-32, Dec. 2007, doi: 10.1109/MC.2007.421

[8] De Roure, D. and Goble, C. (2009) "Software Design for Empowering Scientists", *IEEE Software* 26, 1 (January 2009), pp. 88-95. doi: 10.1109/MS.2009.22 http://dx.doi.org/10.1109/MS.2009.22

[9] Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. and Brass, A. 2007. "A

systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis", *Nucleic Acids Research* 35(16). pp. 5625-33. doi: 10.1093/nar/gkm623

[10] Thomson, H. (2006) "'Me-Science' the New e-Science", Science Grid this Week, October 11, 2006. Available on http://www.interactions.org/sgtw/2006/1011/mescience_more.html

[11] De Roure, D., Goble, C. and Stevens, R. (2009) The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems, 25. pp. 561-567. doi: 0.1016/j.future.2008.06.010

[12] "MySpace for the dudes in lab coats", New Scientist magazine, issue number 2574, 21 October 2006, p. 29. http://www.newscientist.com/article/mg19225745.500-myspace-for-the-dudes-in-lab-coats.html

[13] De Roure, D. and Goble, C. (2010) "Anchors in Shifting Sand: the Primacy of Method in the Web of Data", in Proceedings of WebSci10: Extending the Frontiers of Society On-Line, 26-27 April, 2010, Raleigh, NC, US. Available on http://journal.webscience.org/325/

[14] De Roure, D. (2010) "Replacing the Paper: The Twelve Rs of the e-Research Record", Nature Network eResearch blog, article posted November 27, 2010. Available on http://blogs.nature.com/eresearch/2010/11/27/replacing-the-paper-the-twelve-rs-of-the-e-research-record

[15] Bechhofer, S., Buchan, I., De Roure, D. et al (2011) "Why Linked Data is Not Enough for Scientists", *Future Gener. Comput. Syst* (to appear).

[16] Shneiderman, B. (2008) "Science 2.0", Science 7 March 2008: Vol. 319 no. 5868 pp. 1349-1350. doi: 10.1126/science.1153539

[17] Hey, T., Tansley, S. and Tolle, K. (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Available on http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[18] Kell, D.B. and Oliver, S.G.(2004) "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era", Bioessays, Vol. 26, No. 1. January 2004, pp. 99-105. doi:10.1002/bies.10385

[19] Anderson, C. (2008) "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine 16.07, 23 June 2008. Available on http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[20] Roos, M., Bechhofer, S., Zhao, J., Missier, P., Newman, D., De Roure, D. and Marshall, M. S. (2010) "A Linked Data Approach to Sharing Workflows and Workflow Results", in Proceedings of the 4th international conference on Leveraging applications of formal methods, verification, and validation - Part I (ISoLA'10), Tiziana Margaria and Bernhard Steffen (Eds.), Springer-Verlag, Berlin, Heidelberg, pp. 340-354.

[21] Raddick, M.J., Bracey, G., Gay, P et al (2010). "Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers", Astronomy Education Review, 9 (1), 010103, doi: 10.3847/AER2009036

[22] Tarte, S.M. (2011) "Papyrological investigations: transferring perception and interpretation into the digital world", Lit Linguist Computing. doi: 10.1093/llc/fqr010

[23] Downie, J.S. (2006) "The Music Information Retrieval Evaluation eXchange (MIREX)", D-Lib Magazine, Volume 12 Number 12, December 2006. Available on http://dlib.org/dlib/december06/downie/12downie.html

[24] Wing, J.M. (2006) "Computational thinking", Communications of the ACM 49, 3. March 2006, pp. 33-35. doi: 10.1145/1118178.1118215

[25] Llora, X., Acs, B., Auvil, L.S., Capitanu, B., Welge, M.E. and Goldberg, D.E. (2008) "Meandre: Semantic-Driven Data-Intensive Flows in the Clouds" IEEE Fourth International Conference on eScience, 7-12 Dec. 2008 238 – 245. doi: 10.1109/eScience.2008.172

[26] De Roure, D. Page, K.R., Fields, B., Crawford, T., Downie, J.S. and Fujinaga, I. (2011) "An e-Research Approach to Web-Scale Music Analysis", Philosophical Transactions of the Royal Society Series A (to appear).