Towards Computational Research Objects

David De Roure
University of Oxford
Oxford e-Research Centre
Oxford OX1 3QG
+44 1865 610703
david.deroure@oerc.ox.ac.uk

ABSTRACT
Research Objects are bundles of the digital bits and pieces that make up the reusable record of a piece of research; they are identifiable, citable and sharable. The evolution of this idea within digital research practice has led to the development of *workflow-centric* Research Objects with executable components. To address the evolving requirements of research we propose a further step, towards objects that are composable and executable by machine: Computational Research Objects – a vision in which the content of our digital libraries is autonomously conducting research.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection.* D3.3 [**Programming Languages**]: Language Classifications – *Applicative (functional) languages*.

General Terms
Languages.

Keywords
Research Object, Computational Workflow, Lisp.

1. INTRODUCTION
Academic papers have successfully supported scholarly communication since the Royal Society's introduction of this revolutionary open science model some 350 years ago. However, research practice is changing dramatically, particularly with the advent of digital techniques and new data sources. This raises an important question: are papers still fit for purpose as we move forwards? At the same time it is important to understand why this model has worked so very well. Such insights will help us ensure that future research communication is effective – an important exercise because failure to address requirements will result in restriction on discovery, innovation and insight.

In this short paper we take a look at changing research practice (section 2) to consider the emerging requirements of scholarly communication (section 3). This is not a rejection of papers *per se*, rather we suggest that papers no longer sufficient – a concern which has already motivated the notion of "research objects" (section 4). However in this paper we look to address the requirements more fully, emphasising that "machines are users too": we extend the Research Object idea to *Computational Research Objects* (section 5) and illustrate this by a conceptual analogy to the Scheme programming language (section 6).

DPRMA '13, July 25 - 26 2013, Indianapolis, IN, USA

2. EVOLVING RESEARCH PRACTICE
Three decades years ago researchers commanded only small numbers of computers. The advent of high performance computing, and investment in e-Science, e-Infrastructure and cyberinfrastructure, has enabled scientists to harness considerable computational resource in the pursuit of 'big science'. Over the same period the "long tail" of scientists and citizens alike has engaged in the digital world, largely by harnessing the Web and utilising new devices. Digital content and practice reaches across all disciplines.

Figure 1 below is a simple depiction of these technological and social dimensions in characterising research practice. The vertical axis measures infrastructural capability, such as the increasing numbers of processors and increasing volume of data storage as time progresses. The horizontal axis reflects the increasing number of people – researchers and citizens alike – who are participating in the online and digital world.

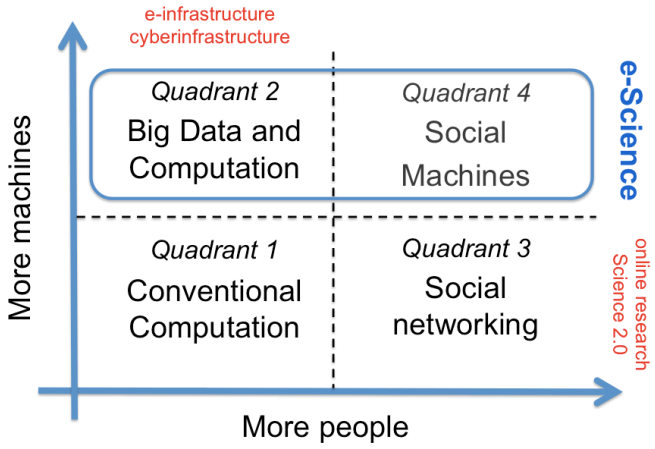

Figure 1. A simple depiction of our knowledge infrastructure

Investment in infrastructure, for example in High Performance Computing (HPC) and Grid Computing [1], initially resulted in "big science" solutions with relatively small numbers of people involved (quadrant 2). Meanwhile researchers in all disciplines have increasingly worked online but not necessarily using advanced infrastructure capabilities (quadrant 3), sometimes called Science 2.0 [2].

New developments in our knowledge infrastructure are very much in the "fourth quadrant" – lots of computers, lots of people – from purposeful online communities to citizen science. This is a complex and rapidly evolving socio-technical space in which computers are used by people – as in online communities constructed for some collective purpose – but also people act to be part of the "computation", as exemplified by aspects of citizen science projects like Galaxy Zoo [3]. In this quadrant we gain new forms of data for analysis, with the familiar characteristics of "big data", and also new analytic capability.

There is a discussion of futurology to be had around our advancement along the axes of this figure. Moore's Law seems to promise continual vertical progress, whereas there is a limit to the number of people – but the rate of progress along the people axis continues to be very significant. Either way, it is indeed clear that we are moving into a world which takes us further up the picture; i.e. where the data analysis burden is increasingly on the computers.

One way of characterising this space is as "Social Machines", a term introduced by Tim Berners-Lee [4] and which is now an area of study underpinning the Fourth Quadrant:

> Real life is and must be full of all kinds of social constraint – the very processes from which society arises. Computers can help if we use them to create abstract social machines on the Web: processes in which the people do the creative work and the machine does the administration… The stage is set for an evolutionary growth of new social engines.

Rather than the automation that comes with increasing computerisation, and which may imply replacement of humans, the Social Machines perspective articulates empowerment of humans; we could say "assistance" rather than "automation". In crowdsourcing we also see humans assisting machines, but still with the ultimate goal to assist humans. We can anticipate a powerful hybrid future.

3. FOURTH QUADRANT RESEARCH

Methodological changes in research practice towards data-centric and data-intensive study are well documented, notably in Microsoft's "Fourth Paradigm" book [5] (the previous paradigms were experimental and theoretical research and then computer simulation). New scientific outcomes result from these new methods (e.g. [6]). In the extreme there is a "data fundamentalist" approach (e.g. [7]) which challenges traditional methodology and favours a correlation rather than causation approach in the analysis of "big data".

Everyday digital research practice now has more of a sense of a "research system", where experiments and analytics are ongoing and we can think in terms of dataflows as well as datasets. These systems may involve many people – online communities interacting asynchronously with the research processes and artefacts, and with each other, forming a sensemaking network that takes us from data to signal, understanding and knowledge. This practice is evident in sciences and also digital humanities.

The field of Music Information Retrieval is a good example, where a vibrant multidisciplinary community has created an effective sociotechnical infrastructure to facilitate their analysis of the increasing number of digital music recordings. New algorithms for feature extraction are evaluated automatically in an annual "evaluation exchange" [8] which brings the community together for information sharing and critical discussion – and they are engaged throughout the cycle, with features identified democratically and machine evaluations requiring human-generated "ground truth". Hence the data, code, ground truth and evaluations emerge continuously.

Citizen Science provides another fourth quadrant example. Many "Citizen Scientists" in Galaxy Zoo are performing image classification tasks, performing data reduction that is better achieved by the human cognitive apparatus than by machine (at this time). This assists scientists, but furthermore there are citizen scientists engaging with parts of the scientific process, interacting via online talk fora where scientists can also engage – and this has led to new astronomical discoveries. Again we have a system, and indeed Galaxy Zoo projects are now delivered via the Zooniverse platform, which is essentially a factory for citizen science projects with a social process for selecting them.

These examples illustrates knowledge infrastructure that we can view as "social machines" where there is a good deal of automation and where people and computers are sharing artefacts like images, recordings, annotations and results. The products of research are distributed and continuously evolving, challenging our existing scholarly communication practice in which we exchange paper-sized chunks of knowledge human-to-human, sometimes with supplementary sharing of data and software for use by machines. The description of the construction and configuration of our research systems themselves – *blueprints* as it were – should now also be a reconstructable part of the scientific record. Practices in community software development may provide valuable insights into these challenges.

4. RESEARCH OBJECTS

Papers are "social objects" – they are units of knowledge but also the subjects of discourse which bring people together and form social networks. They also have an important function in our social system in terms of credit and attribution – indeed they are how our research is measured. Today we can identify new sharable objects – like data and software – which also flow in the sensemaking network of researchers and machines, and are stored for reuse. In Web 2.0 tradition, there are social websites which focus on being the best place to go for a particular kind of object (books, movies, photos etc) and similarly in research we see websites for software and data repositories.

Research Objects[1] are also a new social object: they are bundles of the digital bits and pieces that together provide the record of a piece of research; i.e. the evidence for a research outcome [9]. By aggregating the multiple digital pieces into one object with one identifier we achieve a new sharable, citable social object which drops into the tooling of digital research. It is in some ways like a paper, but crucially the content may be distributed and the objects may be exchanged with computers as well as humans.

An early example of a Research Object is the "pack" in myExperiment[2], a social website for sharing computational workflows [10]. Conceived with workflows as the social object, myExperiment users soon requested the ability to attach data, logs, papers, presentations etc to their workflows. This led to the notion of packs as essentially bundles of URLs pointing at the distributed content. Since it would not be possible (or necessarily desirable) to guarantee the immutability of the distributed

[1] http://www.researchobject.org/
[2] http://www.myexperiment.org/

contents, packs were designed with a weaker guarantee in mind; i.e. that the system could advise the user if the content has changed. Hence packs carry metadata that can include alternative URLs, version numbers or checksums.

myExperiment packs are represented using OAI-ORE and available as linked data and hence semantically-described for ease of discovery and reuse. Although they typically contain computational workflows this is not essential – the bundles are useful anyway, and this is demonstrated by the uptake of the OAI-ORE representation in very many projects. myExperiment's particular notion of workflow-centric Research Objects has been much more fully developed in the Wf4Ever[3] project [11].

Our analysis of myExperiment packs [12] led to a reflection on the nature and purpose of Research Objects, known informally as "the R Dimensions" [13]. This teases apart various aspects of Research Objects. One important aspect is that they should be reusable, another is repurposeable (e.g. self-describing). More generally they enable people or machines to *reconstruct* a piece or research. This all helps with reproducibility, but note that a Research Object is not "reproducible research" by itself: reproducibility means reusing a Research Object with a change to some circumstances, inputs, resources or components in order to see if the same results are achieved independent of those changes – and hence Research Objects should be amenable to such tests.

5. COMPUTATIONAL RESEARCH OBJECTS

Today workflow-centric Research Objects are typically shared by humans and executed by hand, e.g. using the appropriate workflow workbench. When new data is available people may choose to rerun experiments to achieve new results, generating new Research Objects. However there is no real need for humans to press the button – the objects can be executed automatically. We see some of this today in the automatic execution of workflows to check they that still run, i.e. part of autonomic support for curation.

This makes Research Objects part of the automated information circuits that characterise today's research practice – and given the move towards greater automation, this appears to be an important direction in which to develop Research Objects. If Research Objects are semantically described and programmatically accessible then in some ways they are ready for machine use. But there are many challenges: how will they be selected and executed by machine? How are they composed automatically? Will Research Objects consume and produce other Research Objects? How are systems of Research Objects described by Research Objects? How are errors and exceptions handled, when is the human in the loop, and how do we ensure integrity?

The proposal then is that we should define a model that enables machines to assemble and execute systems of Research Objects. This is essentially a computational model for Research Objects, indeed in generality it is a distributed computational model. This leads to a definition of a *Computational Research Object* (CRO). CROs describe process (method) for machine enactment/execution and the associated digital resources, and are:

- Social Objects, designed to facilitate human interpretation (e.g. containing narratives) and shared as part of a (hybrid) sensemaking network;
- "Machine Objects", semantically described and programmatically accessible, designed for automation, scale and heterogeneity
- Composable with a distributed computational model, such that a Computational Research Object can itself assemble systems of objects, and these systems may consume and produce Computational Research Objects.

We can illustrate this with a short scenario in which the CROs are first class citizens and are shared by human and machine:

1. Using an interactive environment I take a digital audio recording and perform a series of analysis tasks leading to a result dataset. The environment captures the history of my analysis in a CRO, which contains descriptions of the input data, the analysis history (workflow) including the software that was used, the output data, together with a narrative description of what has been done.
2. Another researcher then finds this CRO (because I cited it in social media), tests that it is still valid by running it with the existing data, and then reruns it with different audio data. They capture their work as a CRO which refers to mine and/or which uses resources described by mine.
3. A third researcher (a data scientist) then decides to configure the system so that this CRO is run automatically when new data arrives, by registering it for automatic execution on new data arrival.
4. They also create a post-process so that they are notified if the new results meet certain criteria. This common pattern of installing multiple CROs with a post-processor is captured for reuse as a CRO.

Over time multiple users create such information circuits, for example by installing CROs which are triggered by notifications. Meanwhile the execution and repository system monitors their success and performance, and there is ongoing automated testing and maintenance. It records which CROs are used with which data sources, and this provides a basis for recommendation to users. It also enables the system to try automatically a series of analysis tasks and notify users of the outcomes.

6. A COMPUTATIONAL ANALOGY

It is clearly possible to develop new abstract models for distributed computation involving Computational Research Objects. Here we draw a comparison with a very well established model in order to suggest "proof of concept" and illustrate some ideas which we hope will motivate further work.

Our comparison is with the Scheme language, a dialect of the Lisp programming language. While Scheme owes much to functional languages it also supports mutable state. We make the following observations to the reader familiar with Lisp or Scheme:

1. A *closure* in Scheme captures a description of computation which is given parameters when it is applied; it also carries some local state from the time it was defined. Hence the computation sees a (dynamic) global environment (think Web), local define-time values, and runtime values. These create and re-run moments correspond to a CRO.
2. When a closure is applied to parameters, a computation is constructed (by expanding descriptions) and then executed (in Lisp terminology these are called *apply* and *eval*). This is equivalent to instantiating and then enacting (executing) a Research Object.

[3] http://www.wf4ever-project.org/

3. Closures can also be inputs and outputs of computations. In the same way, execution of CROs may consume and produce other CROs.
4. Closures can also be thought of as stateful objects responding to messages. For example, a CRO might respond to a "test" message to validate it.
5. Identifiers are resolved by looking in "environment frames", and where there are no local frames then they are resolved globally. Hence these frames override global values with local state captured at other moments. In the CRO context they could be thought of as alias tables handling versioning.
6. Common patterns (abstractions) can be defined and reused, by higher order functions or also by macro mechanisms – for example to farm out analysis CROs and filter results, or even to set up crowd-sourcing activities.
7. Exception-handling is achieved through the use of *continuations*. More generally these enable computations to be installed for (repeated) completion on arrival of input data, for example.

A Scheme interpreter is easily written in Scheme itself, and it is not difficult to realise a "Computational Research Object interpreter" along the above lines. Prior research on distributed Scheme illustrates that a distributed computational model could also be established [14].

7. CLOSING REMARKS

Our future scholarly communication system will surely be co-constructed, and a vision like Computational Research Objects is just one glimpse of our possible futures – one that is intentionally more computationally-oriented than many of the discussions in this area. We offer it to encourage discussion in and across multiple communities, including computer and information science, digital libraries and socio-technical systems, and as a contribution to the future of research communications[4] debate.

Distributed enactment of information circuits requires a computational model, but computational issues aside it means that the content of our digital libraries will be autonomously conducting research, which raises many questions of ethics, quality and critical review. It may mean that Research Objects are only incidentally human-readable – serialisable as a narrative, like a paper – indeed a hypertext. It is also about the design of new Social Machines for Research Objects as part of our evolving knowledge infrastructure, which we may view as an ecosystem of Social Machines for scholarship.

8. ACKNOWLEDGMENTS

This paper is based on a talk given at the Microsoft e-Science workshop in Stockholm in December 2011. The author's work is supported by Wf4Ever (FP7-ICT ICT-2009.4 project 270192), e-Research South (EPSRC EP/F05811X/1), Digital Social Research (ESRC RES-149-34-0001-A), SOCIAM: The Theory and Practice of Social Machines (EPSRC EP/J017728/1) and Smart Society (FP7-ICT ICT-2011.9.10 project 600854). The idea of Research Objects is due to Iain Buchan at University of Manchester and the original implementation as packs in myExperiment was due to Sean Bechhofer, Jiten Bhagat, Don Cruickshank and David Newman (supported by JISC and Microsoft). I am grateful to Carole Goble, Jun Zhao and my myGrid, myExperiment and Wf4Ever colleagues for further developing workflow-centric Research Objects.

[4] http://www.force11.org/

REFERENCES

[1] AJG Hey and Anne E. Trefethen. (2003) The Data Deluge: An e-Science Perspective, in Grid Computing: Making the Global Infrastructure a Reality (eds F. Berman, G. Fox and T. Hey), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/0470867167.ch36

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

[3] Carol Christian, Chris Lintott, Arfon Smith, Lucy Fortson and Steven Bamford (2012). Citizen Science: Contributions to Astronomy Research. arXiv:1202.2577 [astro-ph.IM].

[4] Tim Berners-Lee and Mark Fischetti (2000). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. HarperCollins Publishers. 2000.

[5] The Fourth Paradigm: Data-Intensive Scientific Discovery (2009). Edited by Tony Hey, Stewart Tansley and Kristin Tolle. Microsoft Research. ISBN 978-0982544204.

[6] Douglas B. Kell and Stephen G. Oliver (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. 2004 Jan;26(1):99-105.

[7] Chris Anderson (2007). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine 16.07.

[8] J. Stephen Downie, Andreas F. Ehmann, Mert Bay and M. Cameron Jones (2010). The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. Advances in Music Information Retrieval Vol. 274, pp. 93-115

[9] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier *et al* (2012). Why Linked Data is Not Enough for Scientists, Future Generation Computer Systems, Vol. 29, No. 2. (February 2013), pp. 599-611, doi:10.1016/j.future.2011.08.004

[10] David De Roure, Carole Goble and Robert Stevens (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. Future Generation Computer Systems 25 (5): 561–567. doi:10.1016/j.future.2008.06.010

[11] Khalid Belhajjame, Oscar Corcho, Daniel Garijo, Jun Zhao *et al* (2012). Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica 2012), Heraklion, Greece, May 2012.

[12] David De Roure, Sean Bechhofer, Carole Goble and David Newman (2011). Scientific Social Objects: The Social Objects and Multidimensional Network of the myExperiment Website. In 1st International Workshop on Social Object Networks (SocialObjects 2011), Boston, MA, US.

[13] David De Roure (2010). Replacing the Paper: The Twelve Rs of the e-Research Record. e-Research, Nature blogs. http://www.scilogs.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/

[14] Christian Queinnec and David De Roure (1992). Design of a Concurrent and Distributed Language. In Proceedings of the US/Japan Workshop on Parallel Symbolic Computing: Languages, Systems, and Applications, Robert H. Halstead, Jr. and Takayasu Ito (Eds.). Springer-Verlag, London, UK, UK, 234-259.