# ALAN TURING INSTITUTE SYMPOSIUM ON REPRODUCIBILITY FOR DATA-INTENSIVE RESEARCH – FINAL REPORT

## AUTHORS

**Lucie C. Burgess[1], David Crotty[2], David de Roure[1], Jeremy Gibbons[1], Carole Goble[3], Paolo Missier[4], Richard Mortier[5], Thomas E. Nichols[6], Richard O'Beirne[2]**

1. University of Oxford
2. Oxford University Press
3. University of Manchester
4. Newcastle University
5. University of Cambridge
6. University of Warwick

## TABLE OF CONTENTS



## ACKNOWLEDGEMENTS

The Organising Committee would like to thanks the speakers, session chairs and delegates to the symposium, and in particular all those who wrote notes during the sessions and provided feedback on the report.

This report is available online at:
**https://osf.io/bcef5/** (report and supplementary materials)
**https://dx.doi.org/10.6084/m9.figshare.3487382** (report only)

## 1. Symposium Overview

### 1.1 Objectives

The Alan Turing Institute Symposium on Reproducibility for Data-Intensive Research was held on 6th-7th April 2016 at the University of Oxford. It was organised by senior academics, publishers and library professionals representing the Alan Turing Institute (ATI) joint venture partners (the universities of Cambridge, Edinburgh, Oxford, UCL and Warwick), the University of Manchester, Newcastle University and the British Library. The key aim of the symposium was to address the challenges around reproducibility of data-intensive research in science, social science and the humanities. This report presents an overview of the discussions and makes some recommendations for the ATI to take forwards.

As the UK's leading data science institute, the ATI has key role in supporting and promoting the reproducibility of data-intensive research, from three perspectives. Firstly, reproducibility is a data science research area in its own right, requiring the development of novel analytical techniques, computational methods, technical architectures and other foundational research. Secondly, reproducibility is a technical-socio-cultural issue, requiring the implementation of new workflows, research working practices and policies in an increasingly open and transparent environment, enabled through supporting infrastructure. One area of focus of the workshop was the ATI's own outputs, which will include algorithms, computations, data and code, and the techniques that the ATI should use to ensure these are reproducible, cite-able and re-useable. Thirdly, the ATI has an important role to play as an advocate for reproducibility. A major goal of the symposium was to encourage researchers to coalesce around the topic, to exchange their expertise, and to maximise participation from the developing ATI community. The symposium was intended to help envision and articulate these objectives at a time when the ATI is in the early stages of its formation.

This final report is a key outcome of the symposium. It is intended both to inform the data science research programme; to inform researcher practices, for example in the practical application of reproducibility tools and techniques, structures for professional development, credit and reward; to enable the development of key policies, for example in data sharing, re-use and intellectual property; and to inform the development of the ATI's data and compute infrastructure. Although primarily intended for the ATI and its partners, we hope this report will be useful to the wider UK and international data science community and a broad spectrum of stakeholders in government, policy and industry.

### 1.2 Citing this report

This report, the symposium programme, speaker biographies and delegate list, slides and video recordings from the presentations and talks are publically available online at the Open Science Framework https://osf.io/bcef5/ and through Figshare

### 1.3 Organising Committee

The symposium organisers were (in alphabetical order, primary institutional affiliation shown):
Lucie Burgess, Associate Director for Digital Libraries, University of Oxford, http://orcid.org/0000-0001-6601-7196
Dr David Crotty, Editorial Director, Journals Policy, Oxford University Press, http://orcid.org/0000-0002-8610-6740
Prof David de Roure, Professor of e-Research, Director of the Oxford e-Research Centre, University of Oxford, http://orcid.org/0000-0001-9074-3016
Dr Adam Farquhar, Head of Digital Scholarship, British Library http://orcid.org/0000-0001-5331-6592
Prof Jeremy Gibbons, Professor of Computing, University of Oxford http://orcid.org/0000-0002-8426-9917
Prof Carole Goble CBE, Professor of Computer Science, University of Manchester, http://orcid.org/0000-0003-1219-2137
Dr Paolo Missier, Reader, School of Computing Science, Newcastle University http://orcid.org/0000-0002-0978-2446
Dr Richard Mortier, Lecturer, Computer Laboratory, University of Cambridge, https://orcid.org/0000-0001-5205-5992

Prof Thomas E. Nichols, Professor, Department of Statistics and Warwick Manufacturing Group, University of Warwick, http://orcid.org/0000-0002-4516-5103
Richard O'Beirne, Digital and Journals Strategy Manager, Oxford University Press, http://orcid.org/0000-0001-7398-1653

## 1.4 Format

The symposium format was an interactive workshop, hosted by the University of Oxford and held at the Dickson Poon China Centre, St. Hugh's College. The full programme is attached at Annex A. The symposium opened with a keynote presentation by Professor Carole Goble CBE, University of Manchester, who gave a definition of reproducibility in the context of computational data analytics. Professor Jared Tanner, University of Oxford, gave an introduction to the mission and objectives of the ATI. Members of the Organising Committee chaired five workshop sessions on the following topics, in which the research challenges, down-stream impacts and ATI priorities were discussed:

- Session 1: Data provenance to support reproducibility
- Session 2: Computational models and simulations
- Session 3: Reproducibility for real-time big data
- Session 4: Publication of data-intensive research
- Session 5: Novel architectures and infrastructures to support reproducibility.

Each workshop session opened with talks from expert speakers followed by three parallel breakout groups, each with a separate discussion topic and expert chair. Data science is inherently inter-disciplinary and the participation of non-domain-experts led by an expert in a specific domain brought a wide range of views to the discussion. Delegates contributed 16 lightning talks on a broad range of topics related to reproducibility, such 'Provenance in neuroimaging with NIDM - results'; 'Reproducible model development with the Cardiac Electrophysiology Web Lab'; 'Data sharing stories from Scientific Data (Nature Publishing Group)'; and 'The Skye project: bridging theory and practice for scientific data curation'. All symposium talks (slides and video recordings) are available at the links in section 1.2.

## 1.5 Delegates

The symposium convened an invited inter-disciplinary group of researchers who employ data-intensive computational methods in their research, from many areas of computer science including provenance, programme verification, statistics, mathematics, psychology, bioinformatics, behavioural social science, web science, climatology, musicology, history, and linguistics. Stakeholders from key institutions such as the Digital Curation Centre and the Software Sustainability Institute, and from publishers and data repositories such as Elsevier, GigaScience and F1000 research also attended. Although most delegates were recognised international experts in their fields, we also welcomed a small cohort of early career researchers from the Turing Institute joint venture partners. The diversity in requirements and perspectives amongst these stakeholders was intended to crystallise a broad range of research challenges and to maximise downstream impact across sectors. A delegate list is provided at Annex B.

## 1.6 Funding

We are pleased to acknowledge funding from the Alan Turing Institute, without which the event would not have been possible, and generous sponsorship by Oxford University Press.

## 2. Executive Summary

### 2.1 What is reproducibility and why is it important?

Professor Goble, in her keynote to the symposium on the "R* Brouhaha" offered a definition of reproducibility and its importance to the ATI. Reproducibility can be defined as the conclusion of one study confirmed independently in another [1], and has been described as the cornerstone of a cumulative science [2]. However, it is important to note that the language and conceptual framework of reproducibility are not standardised across the disciplines [3].

Numerous papers have been published on the topic of reproducibility in recent years. New tools and technologies, increased computing power, massive amounts of data, the open availability of large public databases, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts. "As full replication of studies on independently collected data is often not feasible, there has recently been a call for reproducible research as an attainable minimum standard for assessing the value of scientific claims. This requires that papers in experimental science describe the results and provide a sufficiently clear protocol to allow successful repetition and extension of analyses based on original data" [2], [4]. It has been suggested that the scientific community needs to develop a culture of reproducibility for computational science [4].

The challenge of reproducibility in data-intensive research is neatly characterised in a recent report from the January 2016 Dagstuhl seminar on the topic of *Reproducibility of data-oriented experiments in e-science* [5]: "In many subfields of computer science, experiments play an important role. Besides theoretic properties of algorithms or methods, their effectiveness and performance often can only be validated via experimentation. In most of these cases, the experimental results depend on the input data, settings for input parameters, and potentially on characteristics of the computational environment where the experiments were de- signed and run. Unfortunately, most computational experiments are specified only informally in papers, where experimental results are briefly described in figure captions; the code that produced the results is seldom available."

Reproducibility is fundamental to public trust in science, scientific discourse and transparency in the use of public funds. As well as being a moral obligation, for researchers, "good habits of reproducibility may actually turn out to be a time-saver in the longer run" [2]. An open question is the extent to which reproducibility may be important for data scientists outside the domain of scientific research, for example those computing analyses for actionable business intelligence or data journalism (see, for example [6]). As the work of the ATI will span the full realm of data science, we suggest that the broader context should be kept in mind.

The symposium did not attempt to re-create every aspect of the reproducibility debate but instead focused on five key areas, which we believe are important for reproducibility in interdisciplinary data-intensive research: **Data provenance to support reproducibility; Computational models and simulations; Reproducibility for real-time big data; Publication of data-intensive research; Novel architectures and infrastructures to support reproducibility.** The conclusions and recommendations of our discussions are summarised below.

### 2.2 Summary of recommendations for the Alan Turing Institute

As the UK's leading data science institute, the ATI has a powerful role to play in supporting reproducibility. There are three key ways in which this can be achieved. Firstly, through funding and conducting world-leading research into technical (computational and infrastructural), social and cultural aspects of reproducibility, an area of data science research in its own right; secondly, through implementing practical mechanisms that support reproducibility during the pursuit of interdisciplinary data science research, also both technical, social and cultural; and thirdly, through acting as an advocate, exemplar and champion of reproducible research, engaging with the community and developing partnerships with existing institutions in this area.

### 2.2.1 Key research questions related to reproducibility in data-intensive research

The symposium identified a number of key areas for further research, which participants believe are within the remit of the ATI and are described further in the corresponding sections of this report:

- Assessment of existing system-level, low-overhead, automated and scalable provenance capture systems and the evaluation of their role in supporting reproducibility (see, for example [7]);
- Further development of the PROV W3C standard data model and ontology in different domains (see, for example D-PROV [8]);
- Development of domain-specific languages (DSLs) in data-intensive domains (see, for example, Project Skye: a programming language bridging theory and practice for scientific data curation [9]), codifying the interfaces between generic data science implementations/ algorithms and domain-specific knowledge; and in particular development of DSLs for massive-scale simulations e.g. climate models, urban traffic flow models;
- Development of automated methods for capturing the verification of statistical inference over large real-time data sets;
- Integrating machine learning approaches with statistical methods in order to improve calibration or reduce bias;
- Establishment of research methodologies that can be applied to large-scale, real-time, automated systems;
- Postulating and evaluating mechanisms of transitive credit for re-use of data and software beyond current citation mechanisms;
- Development and deployment of novel technical architectures to support reproducibility, such as Unikernels[1] (see, for example [10]);
- Further research into containers and wrappers (such as ResearchObjects[2] and Docker[3] [11]) for encapsulating data and methods, and evaluation of barriers and enablers to their use.

### 2.2.2 Practical mechanisms to support reproducibility at the ATI

**Software engineering best practices.** Recent years have seen dramatic advances in computing infrastructure that support reproducibility. The ATI has the opportunity to establish a world-leading infrastructure for reproducible research, embedding end-to-end reproducibility in all the research its conducts in support of its data science objectives. Central to this vision is the adoption of software engineering best practices using modern tools and workflows, including:

- Systematic code management and sharing (e.g., Git and GitHub),
- Continuous integration and testing (e.g., Travis CI[4]),
- Virtualisation and portability by means of appropriate development and deployment environments (e.g., Docker, Unikernels)

Reproducibility can be further supported through integration of preservation and citation services within the data processing infrastructure, e.g. by exposing programmatic APIs into repositories for research outputs, data and code.

**Exemplary data governance.** As an international leader in data science, there is an opportunity for the ATI to act as an exemplar, and in that context it will be critical for the ATI to embed social, cultural, and technical mechanisms to support reproducibility. Regarding data governance, these include the establishment and implementation of:

- Policies for data and code sharing, dissemination and openness;
- Policies for transparent intellectual property management, particularly concerning the ownership and rights to benefit from data and code, with the default being openly-licensed;
- Mechanisms of credit and reward for both direct and transitive contributions, such as artifact evaluation.

[1] http://unikernel.org
[2] http://www.researchobject.org
[3] https://www.docker.com
[4] https://travis-ci.org

In particular we recommend that the ATI should ensure that research data and software are viewed as important scholarly outputs, and the ATI should reflect this ethos in its grant-giving guidelines, funding decisions, researcher selection and promotion decisions. We suggest that ATI should develop a 'five year plan for reproducibility', setting out the steps it intends to take, publishing measurable targets, and reviewing its progress.

The ATI has a responsibility to build trust in its research, and taking active steps to conduct reproducible research is one of the ways through which this aim can be achieved. An understanding of the provenance of data, a consideration of bias, the limits and impact of consent, ethics and a discussion of limitations around data sharing are necessary to ensure trust. We are confident that, as a leader in data science funded by both public and private funds, the ATI will embrace exemplary data governance to ensure that its research both meets the terms and conditions specified by data providers and is for the public good.

**Professional development and outreach.** The ATI has a unique opportunity to mediate between computer scientists and domain experts across a wide range of domains in developing and applying reproducible methods and tools, due to the wide range of use-cases and researchers working across many different disciplines underpinned or enabled by data science. We recommend that the ATI incentivises community participation through inviting and publicising use-cases which demonstrate the utility and benefits of reproducible research.

Recognition of the human-in-the-loop aspects of reproducibility is key. There is an opportunity for the ATI to foster collaboration amongst researchers using established methods with those using new methods, for example those working with data from real-time systems. There may be a need for training of data scientists in social science methodologies, and vice-versa, to improve the quality of conclusions from data analytics into real-time datasets.

In particular there is an urgent need to invest in researcher skills in support of reproducibility, providing training in data curation, citation and software sustainability. Although this may not be the remit of the ATI specifically, the ATI has an opportunity to work with organisations such as the Digital Curation Centre[5], the Digital Preservation Coalition[6], the Software Sustainability Institute[7] and the Research Data Alliance[8], who support skills development through professional training and communities of interest.

### 2.2.3 Advocacy, engagement and partnerships

Furthermore, the ATI can play an important role as an advocate for reproducibility in research institutions and universities, in government and industry, for example working with government to embed data management and software sustainability into the UK's science and innovation strategy. We suggest that the ATI should convene a group of eminent academics, practitioners, publishers and other stakeholders to develop a policy on enhancing the quality, transparency, accountability and communication of research, embracing reproducible methods and tools.

There was considerable interest from data scientists attending the symposium in developing partnerships with the ATI. There are strong data science communities across the UK and internationally in addition to the universities which are ATI joint venture partners. For example, from the UK, the universities of Bristol, Kings College London, Manchester, Nottingham, Newcastle and Southampton were represented at the symposium. The data science community also extends to organisations such as the Digital Curation Centre and the Software Sustainability Institute and others, to advocates of open data in government such as the National Archives and to the publishing community, which is adapting its business models to support reproducible methods. We suggest that the ATI should build diverse networks across these communities in order to fully leverage its investment.

[5] http://www.dcc.ac.uk - Digital Curation Centre website [accessed 28-May-2016]
[6] http://www.dpconline.org - Digital Preservation Coalition website [accessed 28-May-2016]
[7] http://www.software.ac.uk - Software Sustainability Institute website [accessed 28-May-2016]
[8] https://rd-alliance.org - Research Data Alliance website [accessed 28-May-2016]

## 3. Data Provenance to Support Reproducibility

*Dr Paolo Missier, Newcastle University; Prof Thomas Nichols, University of Warwick;*
*Prof Dorothy Bishop, University of Oxford; Prof Luc Moreau, University of Southampton*

### 3.1 Overview

The drive for greater reproducibility requires the capability to track and record every aspect of experimental design, data acquisition, pre-processing, analysis and results generation. The PROV data model [12] was published as a W3C standard in 2013 to help enable this goal. Complete provenance would then allow an independent investigator to understand exactly what was done to the data at each step, and attempt to reproduce the result with either the original data, ideally shared and open, or a new set of data. The ability to capture provenance is also important to support the exploration of alternative experimental designs, by making it possible to reason about, and explain, differences in outcomes produced by different versions of an experiment.

Realising this potential has been difficult, however. The goal of the session was to understand the reality of provenance management practices with respect to reproducibility. The session began with three presentations, which explored provenance issues in the neurosciences; gave a case study of open data and provenance in psychology research; summarised the provenance standards and provided an overview of experimental tools that leverage those standards to facilitate reproducibility. Following the talks, three breakout groups explored specific issues:

- motivations, challenges and limitations to exploiting provenance, particularly for open data,
- integration between provenance and other tools, and
- automated reasoning using provenance.

### 3.2 Recommendations

3.2.1 Provenance is pivotal to reproducibility of data-intensive research (see [5] for a general consideration, and [13] for a case-study in neuro-imaging, discussed by participants). There are many research challenges associated with provenance, a diverse area of data science, in which the ATI can play a key role as a world-leading research institution.

3.2.2 In particular, we recommend that ATI should investigate automated and low-overhead provenance collection systems, and evaluate their use in supporting reproducibility, with the understanding that those are going to deliver fine-grained provenance which will require higher level abstraction. We suggest that the ATI should select and deploy at least one such provenance collection system as a pilot, but at a scale significant enough to include a variety of its data science activities within 6 months, to embed provenance across its research activities and create a useful dataset.

3.2.3 The PROV data model [12] was adopted as a standard by the World Wide Web consortium in 2013 and since then further data models have been adapted from the standard (see, for example [8], [14]). The application and adaptation of PROV to different use cases and domains is a key area for further research.

3.2.3 There exist a number of social and cultural issues around the use of provenance to support reproducibility in which the ATI can play a powerful role as an advocate. In particular, establishing the appropriate research infrastructure to support reproducibility, both technical (in terms of data and compute infrastructure, as mentioned above) and socio-cultural (through establishing and implementing appropriate data-sharing policies and reward mechanisms) will be critical.

3.2.4 Practices to encourage reproducibility and re-use needed to be incorporated into the research process at the beginning, and this could be supported by the ATI through implementation of appropriate data and code-sharing policies, and through appropriate training of its research fellows and doctoral students in metadata management and data curation.

### 3.3 Key discussion points

3.3.1 Reproducibility and re-use of underlying data and software are key to doing better research, improving research dissemination and providing transparency for the use of public funding in research. Reproducibility requires a richness of the descriptions of the steps leading to the production of the data and derived results. Provenance plays a key role in recording repeatable experiments and scientific workflows (through, for example, recording acquisition parameters, instrument settings and other metadata) and is also pivotal where science is observation-based and thus not inherently reproducible, e.g. archaeological excavations, large-scale physics experiments or longitudinal observational data, where analysis can be repeated, but the experiment cannot. There are many motivations, challenges, and limitations in the recording and exploitation of provenance [13].

3.3.2 One such key challenge is the balance between cost and benefits associated with collecting provenance for a specific system or even a single application. One provocative starting question was whether it was worth collecting provenance at all and whether there were "success stories" that could act as exemplars for the value proposition of provenance management. Provenance is costly for humans to collect and a compelling case for this cost can be hard to make. However, automated tools can improve the economics: see, for example, use of provenance in the DataOne project [8], a large-scale federated repository for observational earth data; and the use of provenance in neuroimaging using the NeuroImaging Data Model [15].

3.3.3 A better understanding of data requires tracking of its context, captured through provenance information. Automated collection and storage can result in low-level detail that needs abstraction to create semantically meaningful information, but that can be very useful. The Fabric for Reproducible Computing (FRESCO)[9] Observed Provenance User Space (OPUS) project at the University of Cambridge [16] provides an exemplar for system-level provenance capture components that operate unobtrusively, transparently and with low overhead providing a low barrier to entry for making applications provenance-aware. The captured provenance is not PROV-compliant but clearly a useful starting point. Another example is Harvard's Provenance-Aware Storage System, PASS [17], and SPADE, an open-source software infrastructure for provenance data collection and management [18].

3.3.4 Provenance collection systems have the advantage of being domain-agnostic but are not necessarily immediately useful to users, as they require higher-level abstraction. This is a key area for further research. We recommend that the ATI should fund research into mechanisms to bridge the abstraction gap in order to make automatically collected provenance information understandable and exploitable at scale. Ideally, the research would both analyse and use the data collected and propose feasible methodologies for others to capture provenance data, with the aim of more easily answering complex research questions at the scientists' level (for example, can provenance be used to explain the different levels of confidence returned by two repetitions of a method testing the same hypothesis under slightly different configurations of the experiment?).

3.3.5 An interesting area for future research is in the use of provenance to support automated reasoning; to investigate machine learning/ inference techniques for mapping low-level patterns to high-level abstractions, as well as mechanisms for helping users exploit large provenance traces in a way that is intuitive and highly customisable, but removing the need for sophisticated programming. This research will require large-scale provenance datasets to be made available, which is a challenge in itself.

[9] https://www.cl.cam.ac.uk/research/dtg/fresco/, accessed 28 May 2016.

3.3.6 Research data, particularly open data, is dynamic in nature; there are different levels of aggregations, granularity and versions. Although provenance is useful to be able to track findings and conclusions from the original experimental data to the published dataset, there is a need for encapsulation, wrapping data and methods together. Frameworks such as Research Objects[19] have gone some way to making this encapsulation a reality, but needs investment into ease of use to become more widely adopted in the research community.

3.3.7 There is a need for harmonized metadata standards within disciplines and between disciplines to record provenance and enable reproducibility. The Bioinformatics community has been prescient in tackling these issues, through large-scale community efforts like the ELIXIR[10] project in bioscience but this is not always affordable for community-curated databases and datasets, or for projects with small research grants.

3.3.8 Recording provenance information, and cleaning, enriching, curating and sharing data can be a significant overhead. One participant commented "Scientists are not inherently excited about database/annotation 'stuff', so why is it worth for them doing more than the minimum?" Another participant commented: "What is the art of the possible? It's not about what's right or what's best. It's about what you can actually get done!" Key to overcoming this challenge is participation and buy-in from research funders, appropriate policy interventions by research institutions, and, most importantly, evidence that provenance management generates measurable added value.

3.3.9 To incentivise community buy-in and participation we need demonstrator/exemplar projects that showcase the benefits in implementing provenance and sharing data; greater rewards for researchers who implement reproducible research methods; a deeper recognition of, and support for publishing data sets and methods in data papers and for sharing of data and software; and support for data citation as a form of attribution and evidence of shared data to be part of the scholarly record.

3.3.10 A culture of change needs to be driven by funders. Following the example of biomedical funders requiring a statistician, every data science project should have the support of a research data manager and informatics expert. These individuals can lead efforts to incorporate techniques supporting reproducibility, including provenance management, from the inception of a research project.

3.3.11 However, there is a major training and skills gap between what the provenance research community knows how to do in principle, and what research support professionals are taught and able to support in practice. Although this may not be the remit of the ATI specifically, the ATI has an opportunity to work with organisations such as the Digital Curation Centre or the Research Data Alliance who support skills development through professional training and communities of interest.

[10] https://www.elixir-europe.org/ - a distributed infrastructure for life science information. [Accessed: 28-May-2016].

## 4. Computational models and simulations

*Prof Jeremy Gibbons, University of Oxford; Dr Nicola Botta, Potsdam Institute for Climate Impact Research; Prof Patrik Jansson, Chalmers University of Technology, Sweden; Dr Camil Demetrescu, Sapienza University of Rome*

### 4.1 Overview

The focus of this session was two-fold; firstly, participants considered computing techniques employed to yield correct computational simulations of abstract models (such as differential equations in economics), including Domain-Specific Languages (see [20], [21] for a general introduction to DSLs) and code verification techniques. We have the ability to simulate the evolution of coupled earth system models over thousands of years, create synthetic populations of millions of agents and analyse networks of material and energy flows on very fine scales. However, interpreting and communicating results in a rigorous, unequivocal way is challenging, in spite of the growth of available computing power. DSLs can play a crucial role in helping to close the gap between scientific computing and rigorous scientific advice. DSLs were discussed in the more general context of reproducibility, rigour and repeatability in software and systems research [22], [23], [24].

Secondly, the workshop considered artifact evaluation in computer science, a process used to validate that the code supplied with an academic paper gives the same results as those reported in the paper [25]. Participants discussed an evaluation process of supplementary materials (software, data etc.) that complement conference publications in computer science and has been widely tested in several mainstream conferences in computing since 2011 [26], [27]. We considered the motivations, implementation and outcome of artefact evaluation, its application to other domains and specifically to the ATI.

### 4.2 Recommendations

4.2.1 We suggest that the ATI has an opportunity to mediate between computer scientists and domain experts across a wide range of domains in developing or applying DSLs. A key challenge for the ATI is in developing or adapting DSLs for application in specific disciplines (see, for example [28] and [29] which survey the use of DSLs in machine learning and robotics respectively), and across a range of disciplines. The ATI has a unique position to enable development and use of DSLs due to the wide range of use-cases and researchers working across many different disciplines underpinned or enabled by data science. For example, it might be fruitful to develop DSLs to better understand the interaction between biological, ecological and environmental systems.

4.2.2 In terms of reproducibility, DSLs have the potential to enable improved specification of hypotheses and research methodologies in order to reproduce an experiment or re-use data in interdisciplinary research. The ATI could play a leadership role in supporting the development of DSLs aimed at answering research questions in specific domains or across domains. Furthermore the ATI could enable reproducibility through the research of the ATI community, designing use cases and developing best practices drawing on deep expertise in programming language design and implementation across the UK.

4.2.3 Data scientists and domain scientists have complementary skill sets, and identifying and codifying abstraction layers to help them communicate requires a third skill set, namely programming language design and domain specific language implementation. Applying these skills requires time-consuming collaboration to develop the much-needed understanding of both sides. The ATI could encourage reproducibility by encouraging "DSL phenomenology", that is, embedding researchers with programming language design expertise as part of teams involving data scientists and application scientists to help understand and codify the interfaces between generic data science algorithms/implementations and domain-specific knowledge. This could, for example, take the form of short-term visits for programming language researchers or students interested in DSL development to visit the ATI and work on such projects.

4.2.4 We suggest that artifact evaluations be addressed at workshops and conferences organised by the ATI, co-operating with publishers towards a better integration of traditional scholarly publications and the corresponding artefacts that were used in their preparation, and working with existing organisations in this space such as DataCite and the Software Sustainability Institute.

### 4.3 Key discussion points

4.3.1 'Domain Specific Languages' (DSLs) are custom-designed programming languages applied to model a particular domain. They allow a domain-expert to allow enough precision for translation of a complex model to code or software while minimising errors. Most participants in the workshop session were familiar with examples of DSLs such as Excel, R, SQL and database query languages, languages for ontologies, and mark-up in XML or HTML, although as most participants were non-DSL experts they were unfamiliar with the term DSL.

4.3.2 DSLs allow computer scientists and data scientists to 'speak the same language' as domain experts, for example in machine learning using big data [28], robotic systems [29], systems biology or quantitative finance. DSLs can reduce the distance of the abstraction from the domain specialist to the generality of the coder. For example, MLFi (modelling language for finance) is a programming language that enables description of complex financial contracts such as derivatives [30], and uses OCaml (a functional programming language) as its primary implementation. As a counter example, many non-DSLs such as Java use code libraries that are effectively 'black-box' implementations to the domain scientist and their use may impact on the results. There is no clear boundary between DSLs and tools or software implementations for specific applications or domains.

4.3.3 Custom-designed programming/ modelling languages for particular research domains have immense potential benefits, such as verification, portability, optimization and improvements in communication between data scientists/ computer scientists and domain experts.

4.3.4 Successful programming languages 'live' much longer than hardware. For example, FORTRAN was introduced almost 60 years ago and is still in wide use whereas the computers on which it was originally used have long since become museum pieces; C, C++ and Java have enjoyed considerably longevity. DSLs to support the work of the ATI (e.g. for statistical modelling, distributed computation, etc.) should be designed with longevity in mind, rather than as a reaction to short-term needs, and should be independent of architecture. The factors that underpin the success of practical DSLs, and the problems caused by some of their design flaws, need to be better understood.

4.3.5 There is a tension between the degree of specification needed to reproduce an experiment or methodology, and the agility required for successful research. In the absence of a specification, reproducibility is difficult or impossible, but the requirement to formally document a specification can act as a barrier to agile research methods.

4.3.6 Many researchers are unsure about what to document in a research environment which is dynamic and rapidly changing. Electronic lab notebooks are being used successfully in some areas. There is a need for high-level, descriptive languages which that can be used to implement methodological specifications across different disciplines. In neurosciences, for example, domain-specific languages such as R and Matlab meet the needs of the community, but are less useful in other domains.

4.3.7 The inter-disciplinary nature of data science requires accessible high-level languages which can be used to synthesise computational models for non-experts. For example, the Office for National Statistics is taking steps to share and describe the software used to develop its data in order to implement principles of open science and support reproducibility. In computational biology, documenting experiments and methodologies mostly relies on XML schemas, rather than formal specification languages that could be more adaptive and flexible. In the social sciences, there is a need for languages that can describe narratives and reasoning.

4.3.8 The workshop discussed also the use of DSLs in code verification and analysis applied to large-scale datasets and simulations. In massive-scale simulations (e.g. climate modelling, urban traffic flow models) validation of computations is challenging, because there is often no straightforward test or method of verification by which correct results can be identified. In those circumstances, it is all the more important that code is a faithful implementation of the abstract model and that the DSL prevents as many avoidable errors as possible.

4.3.9 Many scholarly articles in computational sciences report claims based on experiments with software and data artefacts. Unfortunately, their key role in validating and disseminating research is often overlooked. A recent artifact evaluation process was adopted at some major computer science conferences, recognising artifacts as first-class citizens in creating and disseminating research in computing. The evaluation process is fully open and transparent.

4.3.10 The group discussed this process as a step forward towards the reproducibility challenge. We argued that software and data play a major role in many fields, hence the idea could be successfully exported beyond computer science. Furthermore, in addition to conferences, a variant of the process could be applied to journals. While sustainable artefact evaluations processes exist, currently many publishers do not offer effective means to archive, index, search, and make them citable using the appropriate metadata. Collective pressure needs to be applied for this to change, working with publishers, research organisations, funders and other key stakeholders.

4.3.11 While a helpful step, artifact evaluation alone is not enough for reproducibility, because artifacts are evaluated in a contemporaneous technological context that may not be available indefinitely in the future. Artifact evaluation should be used as one of the key tools available in the reproducibility toolbox.

## 5. Reproducibility for Real-Time Big Data

*Professor David de Roure, University of Oxford; Dr Suzy Moat, University of Warwick; Dr Eric Meyer, University of Oxford*

### 5.1 Overview

Today we embrace the methodological challenges of big and real-time data, arising from science but also exemplified by new and emerging forms of data such as social media, which provides a new lens onto society, demands study in its own right, and is itself a research tool. The Internet of Things, deployed in our cities, cars, homes and bodies, brings yet more data, machine-to-machine. During this session we discussed reproducibility in the context of real-time big data and new forms of digital scholarship, characterised by machines and people operating together at scale. Widespread adoption of new technologies leads to massive data generation, while at the same time we have crowd-scale personal engagement with the data and its analysis, such as in citizen science projects. This democratisation and empowerment leads to entirely new social processes, and new challenges and opportunities for reproducibility in working with these new forms of data. We considered how we might reproduce data science research using social media analytics, which examines new social processes at the scale of the population and in real time.

The discussion was contextualised with examples of research projects which analyse data from everyday use of the Internet [31], [32], and ask whether sources such as Google, Wikipedia and Flickr can be used to measure and even predict human behaviour in the real world. We looked ahead to our increasingly automated future, asking whether it is meaningful to automate reproducibility, and if and how we should keep the human in the loop.

### 5.2 Recommendations

5.2.1 The ATI should take bias in large-scale online datasets very seriously, and consider both automated and human means of adjusting for bias in order to make data more meaningful. Some participants in the symposium had participated in ATI scientific scoping workshops on ethics of data science and there was consensus that the ethical issues related to bias are very important and must be factored into research conclusions.

5.2.2 There is an opportunity for the ATI to foster collaboration between researchers using established methods with those working with new methods, and in particular between social scientists and data scientists in order to understand better new and emerging forms of data, for example created through the Internet of Things. There may be a need for specific training of data scientists into social science methodologies, and vice versa, to improve the quality of conclusions from data analytics into real-time big data sets.

5.2.3 There are interesting opportunities for the ATI to develop automated methods for capturing the verification of statistical inference over large real-time data sets, or to bring together machine learning approaches with statistical methods in order to improve calibration or reduce bias.

5.2.4 Reproducibility is both more important and more challenging in social data science using repurposed data, proprietary data, closed commercial data or personal data. The ATI has a responsibility to build trust in its research, and reproducibility is one of the ways through which this aim can be achieved. Provenance of data, a consideration of bias and consent, ethics and a discussion of limitations around data sharing are necessary to ensure trust. As a leader in data science funded by both public and private funds, the ATI should embrace exemplary data governance to ensure that research meets the terms and conditions specified by data providers, but is also for the public good.

### 5.3 Key discussion points

5.3.1 "In the wild" methodologies, real-time working and increasing automation all mean we are working with/interfering with a system that is itself adapting (and interfering with us). Examples of automation today include bots and the targeted digital advertising ecosystem. We need to establish research methodologies that cope with this, including dynamic calibration of our models in real time. We should embrace rather than reject the challenging of existing assumptions.

5.3.2 The data available to social science research has changed irrevocably in recent years. Research is needed into automated methods that can more easily provide quality assurance, but there is also a need for a human in the loop. Issues affecting data quality and therefore scientific conclusions have always existed, such as choosing the right variables, survey participants and methods which perturb the data, as well as selective publication of data or over-reliance on statistics such as p-values. However these issues become increasingly challenging with the increasing size and complexity of datasets.

5.3.3 Dr Moat gave an example of Google flu trends algorithm which over-predicted an outbreak of seasonal flu, due to press reports which may have triggered online searches relating to flu by people who were not ill [33]. If we try to use online sources as more up-to-date/economical substitutes for official data, we need to be very careful about the populations we make inferences about. Real-time datasets often cannot be matched with a calibration sample, because the calibration sample does not exist.

5.3.4 Much social science data is inherently different to scientific data or data generated through simulations or computational models: repurposed data (e.g. Twitter, Wikipedia), 'data exhaust' (e.g. from mobile phones or loyalty cards), proprietary commercial data, data that comes with non-sharing or non-disclosure agreements, or data limited in distribution due to differing national legal frameworks (e.g. government or personal data). There are interesting issues around what reproducibility means in these contexts, such as a potential conflict between ethics and reproducibility: for example, is research of poor quality if it is collected within appropriate ethical framework, but then the results cannot be reproduced due to the constraints around access to the data? Ideally, privileged access to closed datasets should be as transparent as possible, with full provenance information provided in the absence of wider access to the underlying data, such that others with the same permissions could replicate.

5.3.5 For example, participants suggested that some datasets being used by policy makers in government might not be as well-documented as they should be. In the same way that code verification can find bugs, both automated and human methods can be utilised to establish the appropriate provenance of data, and similar methods should be developed for capturing the verification of statistical inference over large data sets. As a leader in the field, the ATI has a responsibility to establish appropriate benchmarks and standards through its own research, or point to existing well-established benchmarks (see, for example [34]).

5.3.6 The application of real-time data in data science research requires social scientists and computer scientists to work closely together, for the reasons given above. Participants questioned whether this is an area in which we need specific training, so that the inherent bias in online real-time datasets or ethical issues (for example, related to consent) can be understood and the impact evaluated.

5.3.7 There may be opportunities to bring together machine learning approaches with statistical methods in order to improve calibration or reduce bias. An example was given of prediction of the outcome of the 2012 US presidential elections and 2014 Scottish Referendum using data gathered from X-box users [35], which of course would produce a biased sample but is one of multiple sources.

5.3.8 Inter-disciplinarity is key to data science and its reproducibility, and there are good examples of data-oriented multidisciplinary research programmes such as the Trans-Atlantic Platform for the Social Sciences and Humanities *Digging into Data* challenge [36]. While we were drilling down into digital social research, we equally could focus on digital humanities. Also we need more critical thinking - some communities see benefit of constructive feedback, but elsewhere people are uncritical, and the scholarly ecosystem seems overly generous in publishing results uncritically, and not publishing negative results. This quality control issue, and the need for critical thinking, extends beyond the research itself into its application in the lifecycle of innovation, as new data science drives innovation in marketplace.

5.3.9 Participants suggested that computer scientists have a responsibility to inform researchers using online data as to how heavily phenomena are influenced by user interface design. There is an incorrect assumption in some research that online behaviour is the same as human behaviour, or that online datasets do not incorporate bias. This is a discussion to have with social science and humanities and may be a blind spot.

5.3.10 There are more philosophical questions arising about the increasing automation, for example to process data at scale and at speed, together with the use of machine learning in data analytics. As we automate the conduct of research, to what extent are we irrevocably embedding our current methods into the knowledge infrastructure, and how are they to be challenged?

5.3.11 There is a need to clarify what is meant by 'best practice' in relation to reproducibility, particularly for real-time datasets, and for the social sciences and humanities. In the sciences it may be enough to appropriately cite data and software, and store a Docker file or script in order to record an experimental methodology, but real-time online data sets are mixed-mode, involving both people and machines. Repeating or re-using crowd-sourced analyses, for example, may simply be impossible.

## 6. Publication of Data-Intensive Research

*Professor Carole Goble, University of Manchester; Dr David Crotty, Oxford University Press; Richard O'Beirne, Oxford University Press; Simon Hodson, CoData; Dr Laurie Goodman, GigaScience; Neil Chue Hong, Software Sustainability Institute*

### 6.1 Overview

The idea of data being a publishable entity has become a prominent concept in researcher and publisher conversations. This session explored the role and impact of the publication of data intensive research as the methods and outputs of research change. Participants discussed the role of data and software in achieving reproducibility, and the links to research institution and funder policies, the publishing ecosystem, and researcher behaviours and skills. Laurie Goodman, Editor-in-Chief of the respected data journal *GigaScience*[11], presented a publisher's perspective of data citation: lessons learned, issues needing more education and practical steps needed to encourage data re-use in research. Neil Chue Hong, Director of the Software Sustainability Institute, reminded us that, whilst open data is moving us forward, we risk being stalled by our software. Participants discussed ways to overcome the barriers to the availability and accessibility of data, software and other outputs, which are fundamental to good research, and made a number of recommendations for research policies and practices.

### 6.2 Recommendations

6.2.1. As an international leader in data science, there is an opportunity for the ATI to provide an exemplar, to ensure that data outputs and software are viewed as important scholarly outputs, and to reflect this ethos in its grant guidelines, funding decisions, researcher selection and promotion decisions. The ATI should develop a 'five year plan for reproducibility', setting out the steps it intends to take, publishing measurable targets, and reviewing its progress.

6.2.2. The ATI has an important role to play as an advocate of reproducibility in universities, government and industry, embedding data management and software sustainability into the UK's science and innovation strategy. The ATI should pursue partnerships with the many organisations already working in this sphere, such as the Software Sustainability Institute, in order to support and embed their work in the data science community.

6.2.3. The ATI should fund and pursue research into mechanisms of credit for re-use of data and software over and above citation, a complex computational problem when using datasets 'in the wild' on the web.

6.2.4. The ATI should develop a clear policy on research data and software curation and publication, including a policy on the ownership of intellectual property rights in data and software, and support its implementation through appropriate training for its researchers and support staff.

6.2.5. The ATI should convene a group of eminent academics, working with the national academies, to write a policy document on reproducibility with respect to the role of data and software publication to enhance the quality, transparency, accountability and communication of research, following on from the Royal Society's "Science as an Open Enterprise" report [37].

[11] http://gigascience.biomedcentral.com

### 6.3 Key discussion points

6.3.1. Action by UK research funders to improve data publication is necessary but currently insufficient. Beyond compliance with funder policies, we need realistic incentives for academics to publish their data, code and software, above and beyond the current Research Excellence Framework. One participant noted, "If I have my annual review coming up, the only thing of importance is number of publications; not open access or other altruistic measures".

6.3.2. Although publication of data and code is growing in adoption, levels are still low. There is an opportunity for the ATI to drive change by setting ambitious goals within the data science community. For example, the ATI could accelerate the recognition of the role of data and software in research through establishing an ambitious five-year plan in support of reproducibility with measurable objectives. In 5 years, what steps would need to be taken to ensure that, perhaps, 75% of articles published by ATI researchers cite the associated dataset? What might be a realistic target for citation of software or complete Research Objects?

6.3.3. There was consensus that to drive change, we need improved forms of credit for researchers resulting from data and code publication. Data citation is one form of credit; tracking transitive credit from research objects might be another. This is a data science research area in its own right and could form part of the ATI's research agenda.

6.3.4. In parallel to the conversations around ethics, the ATI should engage its researchers in a conversation around why reproducibility matters, and ensure that researchers in data science see reproducibility as critical to improving the quality of their research.

6.3.5. We use the term "software" to mean everything from a small fragment of code to a suite of programs. There are two orthogonal issues related to software sustainability – 'runnability' (making sure that software is resilient to future changes in technology and will continue to compute) and 'readability' (ensuring that the software used is well-documented and its role in the research can be interpreted and understood, as highlighted in Carole Goble's keynote). The ATI has a role as a beacon for best practice in software "readability" for data science research, drawing on, adapting, supporting and promoting to researchers.

6.3.6. The ATI should produce guidelines (drawing on and adapting existing best practice) as to how code should be written and reviewed to improve the "communication" of data-intensive research done by the ATI. This should be backed by practical support within the ATI, e.g. code review groups, using the in-house Research Software Engineer team to provide advice, and checklists to complete before submitting research including ensuring information about the software should be used is captured. We suggest that the ATI should form a partnership with the Software Sustainability Institute, which is undertaking similar work for RCUK and ELIXIR EU Research Infrastructure.

6.3.7. It has been discussed in many other fora that technology improvements, such as the development and adoption of well-documented open APIs or improvements in database queries, are necessary to make data and code sharing, annotation, submission and data curation much easier (see, for example [38]). Organisations like DataCite, the Digital Curation Centre, Jisc, the Software Sustainability Institute and Force 11 are doing excellent work in these areas. The ATI should support the advocacy work of these institutions and there is an opportunity to work with them to support technology enhancements for reproducibility in data-intensive research. There are also fora that are discipline specific; where appropriate these should be engaged with.

6.3.8. Modern web practices implicitly demand a 'loss of control' through the use of distributed, 'loosely coupled' services, perhaps owned by many different stakeholders. For example, a researcher's pre-print may be published on arXiv, her data in Dryad, her code in GitHub, peer review comments in Publons, her methods described in Protocols.io

and only a minimal narrative held by the research institution or publisher. Services like ORCID[12] (open identifiers which disambiguate authors and improve interoperability between data and code repositories) enable a transfer of control back to the researcher by holding all their personal data in a single store through which they can authenticate and authorize the use of third party web services. The ATI should encourage or mandate the use of ORCIDs by its researchers, following on from the example set by funders like the Wellcome Trust in support of reproducible data-intensive research. The practices to cross-link across stakeholders to give a metadata-led, integrated view (developed by the Research Object group) should be explored.

6.3.9. As an example of advocacy work, there is an opportunity for the ATI to work with publishers and learned societies to improve practices in support of reproducibility – for example, for ATI researchers to leverage their relationships with journals via editorial boards to widen adoption of data and software citation best practices; providing case studies of data-intensive research that have successfully used open methods; or providing analysis of citation/reuse data to demonstrate increase in credit from reproducibility. For example, the ACM (the scientific and educational computing society) is driving the development of best practice and policy in reproducibility through its Task Force on Data, Software and Reproducibility in Publication [39].

6.3.10. There are other important practical steps that the ATI can take to encourage data and software publication. There is much existing work in the community on data and software citation from which the ATI can draw and implement. For example, the ATI should adopt the Force 11 joint declaration of data citation principles and the forthcoming software citation principles, and foster engagement with this organisation. The ATI should make a clear policy statement around research data curation and software publication activities that underpin its research, building on the data management policies of its university partners. Some work has already been started in this area by the joint venture partner university libraries, but needs driving forwards.

6.3.11. In particular, the ATI should establish a clear policy on ownership of intellectual property rights in published data and unpublished research data, as a lack of clarity in the rights in data can be a barrier to sharing and re-use.

6.3.12. We suggest that the ATI should form an expert data management team to support researchers in data deposit, management and, crucially, software curation, providing exemplars for good practice. Both the data management team and policies must interoperate with research data curators and policies of partner institutions and work collaboratively to raise standards across the wider data science community. Working with institutions like the Digital Curation Centre, the ATI should offer options for professional development in data and software curation in areas where researchers find they need support. There is much excellent work in the community from which to draw; for example, the free online training provided by Edina at the University of Edinburgh[13]. We suggest that such skills should be considered part of the full range of capabilities and competencies required to be a successful, high-quality data scientist.

6.3.13. The Share Initiative[14], led by the Center for Open Science and the Association of Research Libraries in the US may be a good model to emulate, with its strong nucleus of key stakeholders that can effect change. There may be an opportunity for the ATI to lead a similar initiative in the UK, for example convening academics in data science from its member institutions alongside the British Computer Society, the ACM, major funders such as the EPSRC and major publishers to demand higher standards for openness and credit within their own field? This could be a more tractable problem and lead the way for others to follow.

[12]
[13] http://datalib.edina.ac.uk/mantra/
[14] http://www.share-research.org

## 7. Novel architectures and infrastructure to support reproducibility

*Dr Richard Mortier, University of Cambridge; Dr Adam Farquhar, British Library; Dr Kenji Takeda, Microsoft Research*

### 7.1 Overview

Recent years have seen dramatic advances in computing infrastructure that support reproducibility. Virtual machines, cloud computing and container technology such as Docker all provide means to capture and replicate the software environment in which code must run, to varying degrees of fidelity. This session explored how these infrastructure technologies are being used and extended, with case studies from Microsoft Research, use of its Azure platform and benchmark datasets including the Microsoft Academic Graph [40]; and Unikernels [10], developed at the University of Cambridge, a means to create entirely self-contained runnable software images by linking application code with necessary the platform libraries at build time. Participants considered current developments, future technical requirements and research questions arising from reproducibility in data intensive research. Practical issues such as incentives and business models were also discussed.

### 7.2 Recommendations

7.2.1 The ATI should ensure that end-to-end reproducibility is embedded in all the research that it carries out and supports, so as to become an exemplar of best practice.

7.2.2 The ATI should measure the cost, impact and worth of reproducible research, both in the research it carries out, and to capture a broader baseline among its partners and the data science community through surveys and other instruments. This baseline should feed into an understanding of how best to support reproducibility in the wider scientific community, and the resourcing implications for doing so.

7.2.3 The ATI should ensure that software development carried out under its aegis adheres to best software engineering practice using modern tools and workflows, including appropriate code management and sharing (e.g., Git and GitHub), continuous testing and integration (e.g., Travis CI), and management and sharing of development and deployment environments (e.g., Docker images and the Docker Hub).

7.2.4 To facilitate this, we recommend that the ATI supports a limited number of development and deployment configurations to assist in alleviating the versioning problem. Having a small number of standard images should reduce maintenance and administrative burden of ATI infrastructure, make it easier to reproduce application behaviour during development, make it easier to map from development to deployment, and greatly reduce the overheads in sharing code both between existing developers and when on-boarding new developers. For example, using the Docker platform, the ATI might develop their own images based off existing standard images, e.g., https://hub.docker.com/_/r-base/ provides a standard, official environment for the R data processing language for applications using R, or https://hub.docker.com/_/debian/ provides a standard, official environment for the Debian Linux distribution for more general development. These ATI images could then be encouraged (or perhaps mandated) for all development work carried out within the ATI, and could provide a useful (even commercial) platform for other data scientists, nationally and internationally.

7.2.5 The ATI should also stay abreast of and support further investigation into more research-oriented techniques and their applicability to reproducibility and data science generally. For example, could something like the OPUS toolchain [41] be used to collect provenance data for all data science carried out by the ATI, from the start? Could the increased use of modern language toolchains by unikernels simplify the process of mapping pre-existing codes onto emerging hardware architectures such as massively multi-core and distributed memory machines, providing greater scalability and future proofing?

7.2.6 Any processes the ATI puts in place for acquiring computational resource should be at least as straightforward to exercise as those provided by current cloud service offerings such as Microsoft Azure or Amazon AWS.

7.2.7 Legal (open source licensing and other intellectual property issues particularly), ethical, and privacy issues all impact reproducibility in data science, and the ATI must thus be cognisant of them. There are thus important inter-relationships with other social science interests already expressed within the ATI through, e.g., the scoping workshop on the Ethics of Big Data workshop, and the Responsible Innovation and Human-Data Interaction symposium.

### 7.3 Key discussion points

7.3.1 Reproducibility is a complex matter that places many design constraints on applications. For example: a researcher may want to re-run their own developed software in the future; or a researcher may want to apply another researcher's software using their own data. The technical discussion focused primarily on the repeatability aspects of reproducibility, though it was noted that this question itself may require research: is bit-wise repeatability always necessary or desirable, or might repeatability at different layers be of use (e.g., when simulating using random seeds).

7.3.2 Of particular interest are technologies such as Docker containers and unikernels, which make code dependencies explicit. Originally leveraging "Linux Container" technology but now expanding to cover "Windows Server Containers" and "Windows Hyper-V Containers", Docker is a platform that makes it straightforward to create container images by explicitly specifying environmental dependencies, to deploy and manage those images individually and in clusters (Docker Swarm), and to share those images with others (Docker Hub). A number of event participants already made extensive use of Docker to create repeatable, sharable environments in which to run their code: its use appears to be widespread and increasing in many scientific communities.

7.3.3 Techniques that support reproducibility through computer system design and software engineering practice include: *unit testing* and *test driven development*, where tests for subcomponents are developed in parallel with or even before those components, to ensure that regressions do not occur during future development; *continuous integration*, where suites of tests (often unit tests) are run on every "commit" or "check-in" of code; and *continuous deployment*, where a running service is updated frequently rather than waiting weeks or months for releases (e.g., Facebook is reported to update its site twice a day). Coupled with effective source code management workflows using tools such as Git, these techniques ensure developers are aware when key functionality has been changed. Through use of services such as Travis CI with Docker developers can also automatically ensure that code executes as expected in multiple configurations of environment.

7.3.4 Related to containers but currently more research oriented are unikernels, which leverage modern languages and their toolchains (that is, the sequence of tools that are applied to convert high-level source code into an executing application; typically this might include compilers, linkers, linters, optimisers, libraries and runtimes) to remove components while building code. This results in the build output capturing the minimal set of dependencies enabling it to run in the target environment (e.g., the application might run as a standard UNIX process during development, but might be retargeted to boot directly on the Xen hypervisor for more efficient and repeatable deployment). Other related tools, such as OPUS from the University of Cambridge, attempt to detect which libraries application code depends on to run, during the development and build processes.

7.3.5 Code developed in data-driven research is likely to cross multiple workflows, operating systems and software environments/ applications. Therefore, we may need to plan for the deployment of an ensemble. One participant

made an analogy of making the job of a 'software archaeologist'[15] in the future easier, through employing different strategies to implement reproducibility: "Level 0: archive all past major versions of the operating system, to create an environment where code can potentially be rerun; Level 1: explicitly document dependencies; Level 2: maintain your software publicly as a package that can be redeployed".

7.3.6 The costs and effectiveness of different reproducibility approaches were questioned and discussed. The lack of good baseline data was noted. The agility of current cloud services to provide the resources needed for research, quickly, flexibly according to need and at reasonable cost was praised.

7.3.7 The problem of handling proprietary or otherwise unavailable software or hardware was discussed, raising cases where Docker containers are not presently a good fit. In such cases the British Library (and others) have had some success using full virtualisation approaches. Comments were made that the ATI may wish to discourage such practices where possible, as use of standard toolsets is more likely to enable reproducibility.

[15] In his novel *The Fire Upon the Deep* (1992), the computer scientist and science-fiction writer Vernor Vinge wrote of a time when programmer-archaeologists maintained the fabric of civilization by diving into and modifying legacy code that ran the systems on which society depended.

## 8. References

[1] B. R. Jasny, G. Chin, L. Chong, and S. Vignieri, 'Again, and Again, and Again …', *Science (80-. ).*, vol. 334, no. December, p. 2011, 2011. DOI: 10.1126/science.334.6060.1225

[2] G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, 'Ten Simple Rules for Reproducible Computational Research', *PLoS Comput. Biol.*, vol. 9, no. 10, p. e1003285, Oct. 2013. DOI: 10.1371/journal.pcbi.1003285

[3] S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, 'What does research reproducibility mean?', *Sci. Transl. Med.*, vol. 8, no. 341, p. 341ps12, Jun. 2016. DOI: 10.1126/scitranslmed.aaf5027

[4] R. D. Peng, 'Reproducible Research in Computational Science', *Science (80-. ).*, vol. 334, no. 6060, pp. 1226–1227, 2011. DOI: 10.1126/science.1213847

[5] J. Freire, N. Fuhr and A. Rauber, 'Reproducibility of data-oriented experiments in e-science', in *Dagstuhl Reports, 6(1)*, 2016. Available: http://drops.dagstuhl.de/opus/institut_dagrep.php?fakultaet=07

[6] M. Broussard, 'Big Data in Practice: Enabling computational journalism through code-sharing and reproducible research methods', *Digit. Journal.*, vol. 4, no. 2, pp. 266–279, 2016. DOI: 10.1080/21670811.2015.1074863

[7] M. Stamatogiannakis, H. Kazmi, H. Sharif, R. Vermeulen, A. Gehani, H. Bos, and P. Groth, 'Trade-Offs in Automatic Provenance Capture', Springer International Publishing, 2016, pp. 29–41. Available: 10.1007/978-3-319-40593-3_3

[8] P. Missier, S. Dey, K. Belhajjame, V. Cuevas-Vicenttín, and B. Ludäscher, 'D-PROV: Extending the PROV Provenance Model with Workflow Structure', *Proc. 5th USENIX Work. Theory Pract. Proven.*, pp. 9:1–9:7, 2013. Available: http://dl.acm.org/citation.cfm?id=2482949.2482961

[9] J. Cheney, 'Project Skye: A programming language bridging theory and practice for scientific data curation'. [Online]. Available: http://homepages.inf.ed.ac.uk/jcheney/group/skye.html#project. [Accessed 28-May-2016]

[10] A. Madhavapeddy, R. Mortier, C. Rotsos, D. Scott, B. Singh, T. Gazagnaire, S. Smith, S. Hand, J. Crowcroft, 'Unikernels - library operating systems for the cloud', in *Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems - ASPLOS '13*, 2013, vol. 48, no. 4, p. 461. DOI: 10.1145/2451116.2451167

[11] C. Boettiger, 'An introduction to Docker for reproducible research', *ACM SIGOPS Oper. Syst. Rev.*, vol. 49, no. 1, pp. 71–79, 2015. DOI: 10.1145/2723872.2723882

[12] L. Moreau, P. Missier, Paolo; K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, C. Tilmes, 'PROV-DM: The PROV Data Model. W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium, April 2013.' 2013. Available: http://www.w3.org/TR/prov-dm/

[13] P. Missier, 'The lifecycle of provenance metadata and its associated challenges and opportunities', *Building Trust In Financial Information - Perspectives on the Frontiers of Provenance*, Springer, 2016. [Online]. Pre-print available: http://bibbase.org/network/publication/missier-thelifecycleofprovenancemetadataanditsassociatedchallengesandopportunities-2016. [Accessed: 28-May-2016].

[14] Y. Cao, C. Jones, V. Cuevas-Vicenttín, M. B. Jones, B. Ludäscher, T. McPhillips, P. Missier, C. Schwalm, P. Slaughter, D. Vieglais, L. Walker, and Y. Wei, 'DataONE: A Data Federation with Provenance Support', in *Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings*, M. Mattoso and B. Glavic, Eds. Cham: Springer International Publishing, 2016, pp. 230–234. DOI: 10.1007/978-3-319-40593-3_28

[15] D. B. Keator, K. Helmer, J. Steffener, J. A. Turner, T. G. M. Van Erp, S. Gadde, N. Ashish, G. A. Burns, and B. N. Nichols, 'Towards structured sharing of raw and derived neuroimaging data across existing resources', *NeuroImage*, vol. 82. pp. 647–661, 2013. DOI: 10.1016/j.neuroimage.2013.05.094

[16] N. Balakrishnan, T. Bytheway, Thomas, L. Carata, Lucian, O.R.A. Check, J. Snee, James; S. Akoush, R. Sohan, M. Seltzer, A. Hopper, 'Recent Advances in Computer Architecture: The Opportunities and Challenges for Provenance', *7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15), Edinburgh, Scotland, USENIX Association, Jul 2015*, 2015. [Online]. Available: https://www.usenix.org/system/files/tapp15-balakrishnan.pdf. [Accessed: 28-May-2016].

[17] D. A. Holland, M. I. Seltzer, U. Braun, and K. K. Muniswamy-Reddy, 'PASSing the provenance challenge', *Concurr. Comput. Pract. Exp.*, vol. 20, no. 5, pp. 531–540, 2008. DOI: 10.1002/cpe.1227

[18] A. Gehani and D. Tariq, 'SPADE: Support for Provenance Auditing in Distributed Environments', *Proc. ACM/IFIP/USENIX 13th Int. Middlew. Conf.*, pp. 101–120, 2012. DOI: 10.1007/978-3-642-35170-9_6

[19] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan, 'Research Objects: Towards Exchange and Reuse of Digital Knowledge', *Nat. Preced.*, Jul. 2010. DOI: 10.1038/npre.2010.4626.1

[20] P. Hudak, 'Building domain-specific embedded languages', *ACM Comput. Surv.*, vol. 28, no. 4es, p. 196–es, Dec. 1996. DOI: 10.1145/242224.242477

[21] W. Taha, 'Plenary talk III Domain-specific languages', in *2008 International Conference on Computer Engineering & Systems*, 2008, pp. xxiii–xxviii. DOI: 10.1109/ICCES.2008.4772953

[22] J. Vitek and T. Kalibera, 'Repeatability, reproducibility, and rigor in systems research', in *Proceedings of the ninth ACM international conference on Embedded software - EMSOFT '11*, 2011, p. 33. DOI: 10.1145/2038642.2038650

[23] S. Krishnamurthi and J. Vitek, 'The real software crisis: repeatability as a core value', *Commun. ACM*, vol. 58, no. 3, pp. 34–36, Feb. 2015. DOI: 10.1145/2658987

[24] J. Gibbons, 'Science relies on computer modelling - so what happens when it goes wrong?' *The Conversation* [online]. Available: https://theconversation.com/science-relies-on-computer-modelling-so-what-happens-when-it-goes-wrong-56859 [Accessed 28-May-2016].

[25] B. R. Childers, G. Fursin, S. Krishnamurthi, and A. Zeller, 'Artifact Evaluation for Publications: Dagstuhl Perspectives Workshop', 15452, 2015. DOI: 10.4230/DagRep.5.11.29

[26] S. Krishnamurthi, 'Artifact evaluation for software conferences', *ACM SIGPLAN Not.*, vol. 48, no. 4S, p. 17, Jul. 2013. DOI: 10.1145/2502508.2502518

[27] A. Bergel and L. Bettini, 'Artifact evaluation (summary)', in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2013*, 2013, p. 24. DOI: 10.1145/2491411.2508114

[28] I. Portugal, P. Alencar, and D. Cowan, 'A Survey on Domain-Specific Languages for Machine Learning in Big Data', Feb. 2016. arXiv preprint arXiv:1602.07637 (2016)

[29] C. Schlegel, U. P. Schultz, S. Stinckwich, and S. Wrede, 'Proceedings of the Sixth International Workshop on Domain-Specific Languages and Models for Robotic Systems (DSLRob 2015)', Jan. 2016. arXiv:1601.00877 (2016)

[30] S. P. Jones, J.-M. Eber, and J. Seward, 'Composing Contracts : An Adventure in Financial Engineering', *ACM SIGPLAN Not.*, vol. 35, no. 9, pp. 280–292, 2000. DOI: 10.1145/357766.351267

[31] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, 'Quantifying Wikipedia Usage Patterns Before Stock Market Moves', *Sci. Rep.*, vol. 3, May 2013. DOI: 10.1038/srep01801

[32] T. Preis, H. S. Moat, and H. E. Stanley, 'Quantifying Trading Behavior in Financial Markets Using Google Trends', *Sci. Rep.*, vol. 3, Apr. 2013. DOI: 10.1038/srep01684

[33] D. Butler, 'When Google got flu wrong', *Nature*, vol. 494, pp. 155–156, 2013. DOI: 10.1038/494155a

[34] 'Data science at Microsoft Research'. [Online]. Available: http://research.microsoft.com/en-us/projects/data-science-initiative/#datasets. [Accessed: 28-May-2016]

[35] W. Wang, D. Rothschild, S. Goel, and A. Gelman, 'Forecasting elections with non-representative polls', *Int. J. Forecast.*, vol. 31, pp. 980–991, 2015. DOI: 10.1016/j.ijforecast.2014.06.001

[36] 'Trans-Atlantic Platform Social Sciences and Humanities - Digging into Data Challenge'. [Online]. Available: http://www.transatlanticplatform.com/2016/02/29/trans-atlantic-platform-announces-the-2016-t-ap-digging-into-data-challenge/. [Accessed 28-May-2016].

[37] G. Boulton, P. Campbell, B. Collins, P. Elias, W. Hall, L. Graeme, O. O'Neill, M. Rawlins, J. Thornton, P. Vallance, and M. Walport, 'Science as an open enterprise', *Science (80-. ).*, no. June, pp. 1–104, 2012. [Online]. Available: https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

[38] P. Buneman, S. Davidson, and J. Frew, 'Why data citation is a computational problem', 2016. Pre-print, to appear, Communications of the ACM (2016). [Online]. Available: http://frew.eri.ucsb.edu/private/preprints/bdf-cacm-data-citation.pdf

[39] 'ACM Task Force on Data, Software and Reproducibility in Publication - web page'. [Online]. Available: https://www.acm.org/publications/task-force-on-data-software-and-reproducibility. [Accessed: 07-Jun-2016].

[40] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. (Paul) Hsu, and K. Wang, 'An Overview of Microsoft Academic Service (MAS) and Applications', in *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, 2015, pp. 243–246. DOI: 10.1145/2740908.2742839

[41] N. Balakrishnan, T. Bytheway, R. Sohan, and A. Hopper, 'OPUS: A Lightweight system for Observational Provenance in User Space', in *Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance*, 2013 (2013).[Online]. Available: https://www.usenix.org/conference/tapp13/technical-sessions/presentation/balakrishnan