Accelerating Scientists' Knowledge Turns

Carole Goble[1], David De Roure[2], and Sean Bechhofer[1]

[1] School of Computer Science, The University of Manchester, Manchester, UK
{carole.goble,sean.bechhofer}@manchester.ac.uk
[2] Oxford e-Research Centre, University of Oxford, UK
david.deroure@oerc.ox.ac.uk

Abstract. A "knowledge turn" is a cycle of a process by a professional, including the learning generated by the experience, deriving more good and leading to advance. The majority of scientific advances in the public domain result from collective efforts that depend on rapid exchange and effective reuse of results. We have powerful computational instruments, such as scientific workflows, coupled with widespread online information dissemination to accelerate knowledge cycles. However, turns between researchers continue to lag. In particular method obfuscation obstructs reproducibility. The exchange of "Research Objects" rather than articles proposes a technical solution; however the obstacles are mainly social ones that require the scientific community to rethink its current value systems for scholarship, data, methods and software.

Keywords: Reproducible Research, Scientific Workflow, Research Object, Digital Scholarship, Open Science

1 Introduction

A "knowledge turn" in manufacturing enterprises is the cycle of a process that derives more good leading to new or better products and competitive advantage [43]. One turn corresponds to a single trial-and-error cycle by a professional in a focused area of knowledge, and includes the learning generated by the experience. Long-term sustainable competitive advantage is advanced if a business learns faster than its competitors: the experiences of its people recorded, efficiently harnessed, communicated and built upon. Scientific research is the business of knowledge turning. The classical scientific method turns observations and hypothesis – through experimentation, comparison, and analysis – into new knowledge and improved "scientific products" such as more observations and hypotheses, experimental methods, models, techniques, data, protocols, and publications. A "Hypothesis-Prediction-Observation-Analysis" cycle operates over a knowledge pool of "know-what, know-how, know-why and know-who" populated by the publication of peer-reviewed articles, the gathering of scientists at symposia, sharing of skills and experience in collaborations, and the exchange of datasets, methods and, increasingly, computational tools and software.

As an example of knowledge turns and clear motivation for accelerating them, Josh Sommer is an impressive young man suffering from Chordoma, a form of bone can-

cer. His condition is uncommon and the research scattered and piecemeal. A cure was being held back by lack of funding, but it was also being hindered by the scattered researchers getting hold of resources, restrictions on information flow between the researchers and poor coordination and collaboration [54].

Three labs are researching the condition (Fig. 1). Lab1 produces experimental results that Lab3 uses to test a hypothesis. Data produced by Lab3 generates a hypothesis that when combined with a locally innovative technique in Lab2 produces new insights that, unexpectedly, is just the missing piece that Lab1 needs to confirm an earlier experiment. Each step is a knowledge turn, to turn prior results into new results. Restricted flows between the labs slow down the whole process. Years can pass between the emergence of a result and its availability to another researcher, if it ever becomes available, wasting time and wasting opportunities.

Josh co-founded the Chordoma Foundation (www.chordomafoundation.org) as a resource and knowledge broker, to remove the barriers on information flow, particularly addressing the social influences operating on scientific communication practices that make them slow and ineffective: e.g. the jealous guarding of pre-published results and ignorance of other research. The foundation has maximized productivity and reduced lags by providing resources, by creating, collecting, storing, and distributing information and biological materials, and by facilitating information exchange and collaboration among researchers.

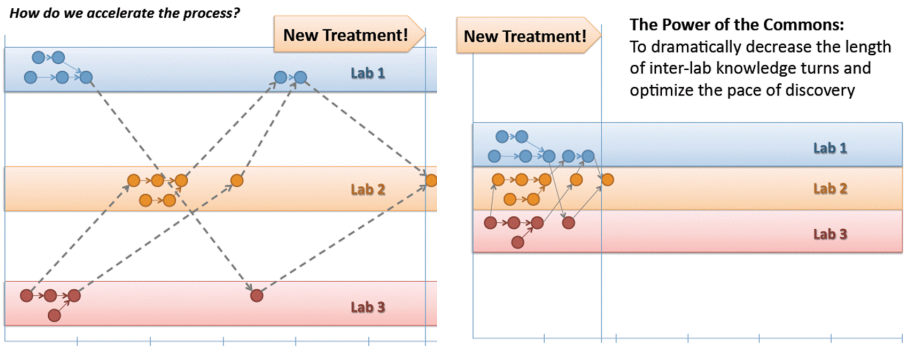

Fig. 1. Rapidly contributing results in a commons increases knowledge flow, decreases the length of inter-lab knowledge turns and optimizes the pace of discovery [54].

In this paper we will discuss what affects knowledge turns, the practices and challenges of science in the web era, and how we might accelerate the turns of scientific knowledge. In Section 2 we introduce today's scientific research practice through digital instruments, information technology and community. In Section 3 we consider reproducibility and the primacy of method, and offer the important social perspective in Section 4. In Section 5 we introduce *Research Objects* and our current work in the "Workflow Forever" project, and we conclude in Section 6.

2 The Mechanics of Fact Making

Shapin [49] highlights three key aspects in the "mechanics of fact-making" that have revolutionized scientific research: (i) The *instrument technologies/experimental equipment* used to undertake science, and produce, process and analyze knowledge; (ii) the *information technologies* used for recording and spreading knowledge; and (iii) the composition and *social organization of scientific communities*, and their conventions in cooperation, collaboration and dealing with scientific claims. How are advances in these components revolutionizing science in the web era?

2.1 Digital Scientific Instruments

Scientific instruments are devices for observing, measuring and handling natural or experimentally produced phenomena. Instrument technology relates to their design, production and use. We are in an age of digital instruments; observations and measurements are born digitally or the instruments themselves are *in silico*. To paraphrase McLuhan: "scientists shape tools and thereafter the tools shape science" [35]. Gray defines four paradigms of science: *empirical*, observing and describing natural phenomena; *theoretical*, using models and forming generalizations; *computational*, simulating complex phenomena; and *data exploration*, unifying theory, experimentation, and simulation [24]. How are these shaped by some of today's instrument technologies?

Revolutions in technology have increased the rate of empirical observations and decreased the cost of ownership. Genomes can now be sequenced ~50,000 times faster than in 2000 yet Next Generation Genome Sequencing machines are cheap enough to be local laboratory commodities potentially generating terabytes of data [40]. Acceleration in data scales is accompanied by a proliferation in the varieties, with derived, distilled and predictive datasets alongside raw data, and online accessibility. The Molecular Biology Database Collection lists 1380 databases [17] ranging across genomes, proteomes, structures, pathways, etc. PubMed Central (a citation index for the biomedical literature, www.ncbi.nlm.nih.gov/pubmed) indexes over 20 million published articles, a new one every 30 seconds. Generating data is no longer the bottleneck – rather, it is what we do with it, and scientists can now turn to data first where they used to turn to an instrument.

Today software is an instrument – it makes it possible to predict, simulate, and generate/confirm a hypothesis on the basis of datasets and models alone [29]. Increased automation copes with the scale, repetition, accuracy and complexity of processing. Our bottlenecks are curation, comparison, validation, filtering, mining, integration, visualization and analytics; the issue is not the $1000 genome but the $100,000 analysis. Publishing software and data as Web Services opens up an ecosystem of interacting services; i.e. "service-oriented science". Commodity based elastic compute and data commons cloud platforms opens up the prospect of a pay-as-you-go "science as a service" delivery model: a lab generates the data and the questions; a public or commercial service manages the data, processing and even the processing pipeline, spreading the cost and outsourcing routine but scarce analysis expertise [3].

Example: Science as a Service using Taverna Workflows

Some cattle breeds in sub-Saharan Africa are highly tolerant of sleeping sickness infection (African bovine trypanosomiasis), but the potentially more productive breeds are more susceptible. Increasing their tolerance could have a major impact on food production. A multi-institution, multi-disciplinary team sampled many cattle specimens and repeatedly used computational analysis techniques that accessed online tools, drew data from public data sets, and compared to findings in the literature. They succeeded in finding the candidate genes that could hold the key to the differences between the cattle [42]. To accelerate the analysis the team automated it using the Taverna Workflow Management system [45] (Fig. 2). A scientific workflow combines data and processes into a configurable, structured sequence of steps. Taverna provides software to combine data sets systematically and automatically, and integrate analytical software tools and services even when they are independently made and incompatible [26].

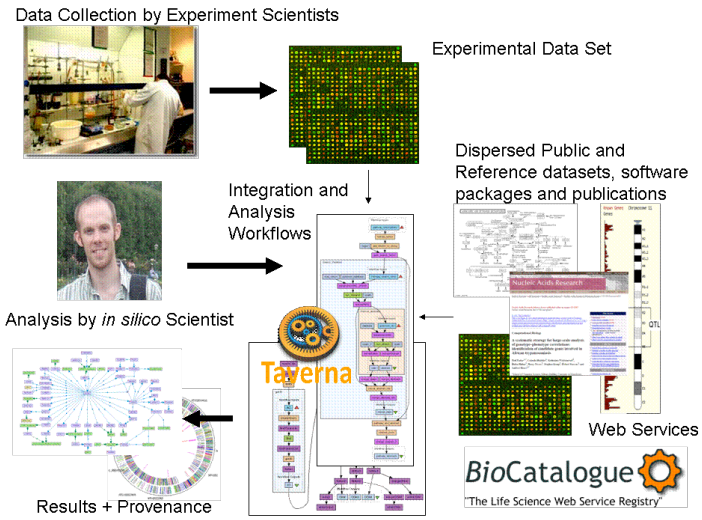

Fig. 2. Improving research productivity using automated Taverna workflows

A visual interface enables skilled computational scientists to assemble pipelines and access services shielded from low-level programming concerns. The execution platform accesses the scattered datasets and tools remotely hosted by providers so consumers do not need to download resources or learn their codes [15]. The Taverna workflows ensure that the process is comprehensive, systematic and unbiased, important for protocols often repeated. They represent an accurate record of the computational method and automate recording the provenance of the computed results which is crucial for experimental and data reproducibility (discussed in Section 3). They liberate the scientists from routine informatics, reducing the manual analysis task from weeks to hours, so the scientists can concentrate on scientific discovery.

Computational workflows are effective because the data are in machine-readable formats and the databases and tools are accessible through machine-processable interfaces (APIs) such as Web Services. For example, the BioCatalogue [5], a public catalogue of Web Services in the life sciences, is integrated into Taverna. It currently registers 2000+ services from 160+ providers. Next Generation Sequencing makes it possible for this consortium to sequence a genome every few days. By putting the analysis workflows on Amazon Cloud, alongside reference datasets and tools, we create an "Analysis as a Service", enabling non-informatics skilled experimentalists to rerun "pre-cooked" computational protocols repeatedly against new samples and changes in the reference services, without having to manage complex infrastructure. This opens up the processes and data resources to a wider group of scientists and scientific application developers. Elastic compute, pay-as-you-go Clouds enable scientists to pay on-demand. The costs are spread by making the service available to the whole community when the analytical process is routine.

2.2 Digital Scientific Information Technology

Information technology is the means used to accumulate knowledge; to record, transmit, analyze, verify and discuss claims; and to "virtually witness" experiments and observations. The traditional method is the peer-reviewed scientific publication, the traditional technology is print, and the traditional archive is the library. The Web, originated as a research communication tool, is *the* information technology that has most changed the dispersal of scholarship: publications and most results and analyses are "born digital", and anyone can be a publisher or a librarian.

However, mainstream online publishing is chiefly putting facsimiles of print on the Web. Alternatives to written articles, like videos, are seen as exotic and time-consuming with little benefit for the author (though this may change as tools improve and new digitally native researchers gain influence). Data and methods are frequently unavailable for peers to test claims [34], or inaccessibly embedded in documents in unprocessable forms (tables, graphs, pictures, text) [44]. Journals with data policies that demand that data be available in public archives or supplementary material have accelerated discovery in the Life Sciences, but frequently turn out to neither enforce their own policies nor ensure long-term compliance [1]. Open Access enables the flow of research – if we count citations as a metric [18] – and are actively promoted within disciplines (e.g. ArXive.org in Physics) and by funders (e.g. the US NIH publicaccess.nih.gov). However, Open Access does not guarantee reusability [52], and though it is free for consumers it is not for providers, restricting publication to those who can afford it. As the majority of publishers remain subscription-based and legislative efforts in the USA aim to block open access mandates, open access for all content is some way off.

The Web supports the technological and social innovations to circumvent publishers and libraries. Scientists can now assemble their own publishing environment, organize agile peer review and create their own scholarly research productivity platforms [27]. Such agility is enabled by:

- *Mass Market resources*: Free publishing platforms (e.g. wikis and blogs) and cloud-based services (LinkedIn, YouTube, SlideShare, Twitter, Wikipedia, GoogleDocs, DropBox, GitHub, though not so much Facebook);
- *Specialised resources*: data commons (FigShare, Sage Bionetworks); methods (nanoHUB, myExperiment, OpenWetWare); models (BioModels), reference management (Mendeley, CiteULike, ReadCube); publishing analysis (Google Citations); review management (EasyChair); LIMS (YourLabData); Lab management (LabGuru); social networking (ResearchGate, NatureNetworks, SciLink etc);
- *Cross-platform search and indexing*: free and powerful search tools (Google, Yahoo) and services (Google Scholar) enable cross-library, cross-publisher and cross-repository discovery, impact metrics, and science mappings;
- *Open standards*: open licenses, non-proprietary standards, open data formats and APIs that enable platforms to interoperate, and added-value services to proliferate. Monolithic publishing systems are giving way to ecosystems of services.
- *Open metadata*: common reporting standards and ontologies allow different systems to represent, exchange and link information on communities (e.g. SIOC), citations (e.g. CiTO [51]), and artefacts (e.g. Dublin Core). Ontologies representing claims and discourse enable recorded discussions across publications, datasets and communities [8].

These resources improve the flow of results and provide the specialized management services needed for different data types, but they also act to fragment information.

Embedded Science and Scattered Science

The scholarly article as a monolithic, single document is no longer fit for purpose, especially for online scholars using digital instruments. Readers need (i) easy access to underlying data, software, and methods and (ii) the ability to consume content on a variety of electronic devices and by digital instruments [44]. From one perspective articles are compound "collages", incorporating data (in tables and charts) and algorithms/methods (in incomplete written sketches). For articles to be actionable and verifiable scholarship, these need to be disinterred from computationally inaccessible forms into links to the actual data and executable codes – to re-link the scholarship with the instruments. Besides, the data and methods are crucial scholarly results that should be published, cited and credited in their own right.

From another viewpoint, articles are structured reports of claims related to evidence and prior work. But these claims are deeply embedded and computationally inaccessible. To surface such content requires authors to structure their claims more systematically, expert readers to explicitly recover and structure them [9], and text mining to assist discovery and recovery. For example, components of the workflow in Fig. 2 mine the literature available in PubMed to surface links to genes that are then tracked in the datasets. Semantic Publishing [50] and Semantic Publishing systems such as Utopia Documents [44] are attempts to tackle computational accessibility through explicit metadata.

Accelerating knowledge turns between laboratories is not just about flow and availability of results. Researchers must cope with the piecemeal and fragmentary

nature of publishing, gathering results before they can begin to use them. Scattering occurs in several forms:

- *Stewardship*: On one hand we have massive centralization in the various general data centers, specialist data collections and publishers' supplementary repositories. On the other we have massive decentralization in private lab books, group wikis, project web sites and institutional repositories. The original experimental data is on a disk under a post-doc's desk whereas the analysis results are embedded in a table in PDF supplementary file held by the publisher behind a pay-wall.
- *Asset type*: For one experiment the analysis software is on GitHub, the workflow is on myExperiment, and the tutorial is on Slideshare. Scattering results onto specialist service providers delegates their stewardship to experts, but it also risks the demise of a "rented" resource from which content is hard to recover; for example, Google shut down its medical and scientific data vault. On the other hand research labs are even less likely to maintain accessible archives.
- *Multi-part datasets*: Although biological science is integrative its data ecosystem is "siloed" and dispersed. Separate resources hold information on gene sequences, structures, interactions, proteins, etc. These scattered complex data sets can only be reassembled by consumers equipped to navigate the various reporting standards and technologies used. Unified frameworks for metadata that ensure biomedical research datasets become interoperable are urgently needed [47].

Value-added data integration platforms like Galaxy [21] and topic-specific *commons* like Sage Commons (sagebase.org/commons) and Pathway Commons (www.pathwaycommons.org) attempt to improve productivity for the hard-pressed researchers who otherwise have to track down the diverse components of an experiment for themselves.

Example: the myExperiment e-Laboratory for Workflow Exchange

An e-Laboratory is a set of (usually) scattered online components – workflows, resources, data, algorithms, texts, queries – that circulate in a collaborative space used by (usually) scattered scientists. They enable the planning, execution, analysis and publication of *in silico* experiments and their results. As we have already seen, scientific workflows capture a computational process so that results can be reproduced, the method accurately reviewed and validated and know-how shared, reused and adapted.

myExperiment [11] is an e-Laboratory for the circulation, conservation, preservation and publishing of workflows. Communication flow is encouraged by social tagging, comments, ratings and recommendations, social network analysis and reuse mining (what is used with what, for what, and by whom). myExperiment encourages the flow and reuse of methods, and supports workflow citation and "altmetrics" for its contributors. The open public myExperiment.org currently has 2000+ workflows from 21+ workflow systems and 5000+ registered members. Workflow design is a skill often beyond a researcher. By establishing shared collections of workflows that contain reusable processing pipelines we help avoid the reinvention of techniques and propagate best practices: some workflows have access statistics of 1000+. The work-

flows developed by Fisher for the cattle investigation (Fig. 3), have been successfully reused for chronic colitis [31] and bacterial responses to oxygen [33].

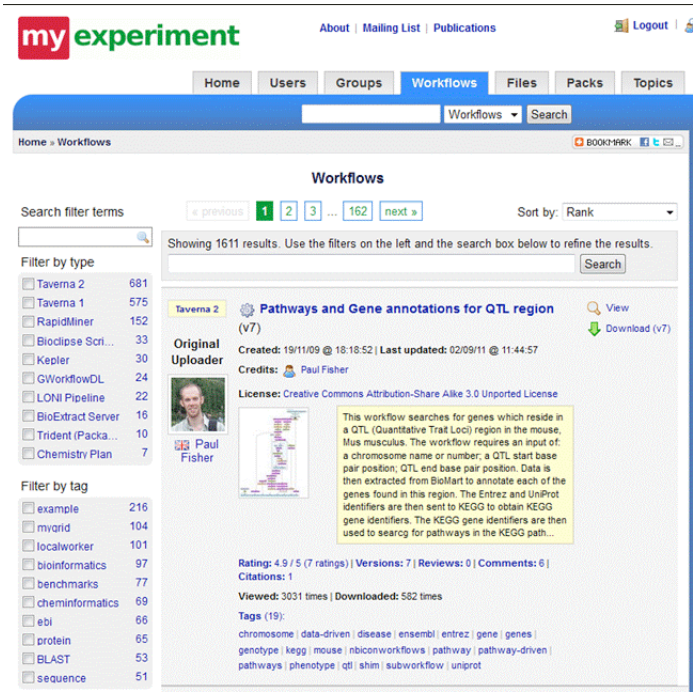

Fig. 3. myExperiment showing an entry from the cattle workflow suite by its author

myExperiment directly addresses scattering through the notion of "packs". Essentially a pack is a single sharable collection of items which could include anything from workflows and data to slides and papers, and may be on the myExperiment site or elsewhere. Packs are one of the "social objects" of myExperiment, but equally they are a technical object in that they can be accessed programmatically though myExperiment's Linked Data interface, a semantic interface using the Object Reuse and Exchange representation. Packs form the basis of the Research Objects discussed in Section 5.

2.3 The Online Scientific Community – Collaboration and the Crowd

Scientific investigation is a social activity. Each scientific community and subcommunity operates within conventions shaped by the roles of its members, their activities and the perceived value of their contributions. We classify the "crowd" into:

- *Suppliers/consumers of instruments*: Experimental, theoretical scientists and modelers are usually thought of as *consumers*. *Suppliers* include scientific informaticians, computational scientists, specialist tool developers, service and resource

providers, content curators, infrastructure developers and system administrators. A bioinformatician is a consumer of software from a programmer who in turn is a consumer of a dataset supplied by an experimental scientist.

- *Suppliers/consumers of information*: All players consume. Librarians, publishers, data centres, the Web, supply technology. Scientists supply content, i.e. *result*s.
- *Governance regulators*: funders, employers, and peer-reviewers define expectations, confer value on scientific investigations and evaluate the performance of members of all the scientific community through metrics such as paper citations, software download figures or scales of datasets.

How does this social "ego-system" (i) collaborate and (ii) peer-produce? Networks of people construct scientific instruments (e.g. Uniprot Protein Sequence Database, the Large Hadron Collider), conduct and discuss experiments, and record and publish experimental results. Almost all original natural science papers have multiple authors; in January-October 2008, Nature published only six single-author papers, out of a total of 700 reports [58]. Traditionally networks may be organized scientific societies or projects. Thanks to the Web virtual communities are easy to set up and can span geographical, institutional, discipline and time zone boundaries. Across a spectrum of formal, coordinated and funded collaborations to informal self-organizing, volunteer-based networks, we are in an era of "Networked Science".

The use of online tools has the potential to dramatically speed up the rate of scientific discovery across all of science but also to amplify the *collective* intelligence of science, and expand the range of scientific problems which can be tackled [36]. Virtual communities of scholars in all fields and all team sizes produce higher impact and more highly cited work than comparable co-located teams or solo scientists [61]. Moreover, increasing specialization of all fields and the need to tackle problems beyond the scope of a single discipline or research practice necessitates interdisciplinary team research. Thus collaborative research needs *shared understanding* in addition to shared expertise and shared resources, and hence clear and regular communication.

Although face-to-face networking is essential, all scientists work in online cooperative spaces. Just using email makes us part of an "invisible college" on a scale that was not feasible 30 years ago. This completely changes the speed and range of conversation. First, cheap and instant communication – instant twitter broadcasting, instant message conversations and virtual Skype meetings – make complex cooperation feasible across a scattered community. Word of mouth recommendation has been raised to web-scale by micro-blogging with twitter, feeds, blogs, social bookmarking and auto citation tracking: to cope with an information deluge, science gossip is now electronic. Papers that are tweeted appear to be more highly cited [14]. Second, "limited focus" social networking improves the flow of expertise between people who already collaborate and between people in the "long tail" of small but expert research labs and singleton graduate students. They create "hang outs" for isolated and/or distributed researchers to crowd around a focus: people (e.g. ResearchGate); resources, such as citation recommendations (e.g. CiteULike, Mendeley); digital instruments such as workflows (myExperiment) or simulations (nanoHUB); and scientific problems such as SysMO-SEEK for Systems Biology in MicroOrganisms and Scratchpads for BioDiversity data (scratchpads.eu). Finally, collaboration tools enable two-way

communication and collaboration with practitioners and beneficiaries, such as patients and policy makers [9].

Science has always relied on crowd-sourcing, building upon a cumulative pool of other's results. Traditionally, publishing houses crowd-source experimental results, peer-review harnesses the wisdom of the crowd and citations are crowd-voting metrics. Public datasets like Uniprot, and data commons like Sage Bionetwork, are repositories for community-contributed results. Now we have the means to widen the crowd to scientists in different disciplines, to scientists in the "long tail" including amateurs, and to citizens too (e.g. the Galaxy Zoo citizen contributions to Astronomy [32]). Curated public databases like Chemspider (www.chemspider.org) for chemistry, and wikis such as Specieswiki (species.wikimedia.org), are a game changer. Inspired by Wikipedia they are also mechanisms for concentrating information, for quality-improving mass-curated distillations of scientific knowledge otherwise dispersed [57]. Wikipedia itself is used as a way of improving "official" datasets by mass curation [10]. The same phenomenon holds with open source software: it is popular because it is free and collective ownership gives confidence in its durability, quality and continued availability.

myExperiment supports a mixed crowd including a "long tail" of isolated computational scientists who are strangers working independently. The site allows them to find, share and brag about expertise and get/offer help. Interestingly, tagging and reviewing other's content is rare, and collaborations sparked by the site are continued outside it privately. Groups, typically from specific projects or disciplines, establish their own collaboration and sharing norms. SysMO-SEEK (www.sysmo-db.org) is a more elaborate version of a group-based e-Laboratory: a specialist, private resource supporting 15 multi-partner consortiums (around 300 scientists) for a European-wide research program in Systems Biology. The collaborations are within established groups and between the groups, focused on one discipline, and focused on exchanging data, models, and procedures. The social dynamics of the two systems have commonalities and differences, and we will refer to them both in this paper.

3 Reproducibility and the Primacy of Method

A scientific communication has two prime goals: to announce a result so it can pass into use by the reader and to convince the reader that the result is correct [34]. Reproducibility underpins the scientific method and is fundamental to the flow of knowledge between researchers [55]. Productivity is strongly linked to the ease with which a result can be reproduced and sufficiently validated to be reused. Consequently, a communication must present its experimental materials and methods. We are now familiar with the notion that the data (materials) described should be available for inspection. However, to truly reproduce results we also need methodological transparency. The methods need to be precisely described to stand up to critical examination and reproduction, and the provenance of the data needs to be precisely reported so that it can be accurately cited, safely compared and correctly interpreted. This is not just for the consumers' benefit but also the producers'. Misunderstanding

the subtle contexts of data and methods leads to misuse and misinterpretation of results when used by other disciplines, policy makers, journalists and the public.

In experimental science, methods include laboratory protocols and Standard Operating Procedures (SOP). These are detailed, written instructions to achieve uniformity of the methodology used and without them it is impossible to compare or validate results. SOP repositories in biology include Nature Protocols, OpenWetWare.org and MolMeth.org. The earliest request for SysMO-SEEK by its users was for a Standard Operating Procedures registry. Media publishing sites like Jove and SciVee publish videos of experimental methods that can be linked to written protocols or used in teaching. In computational science methods include: (i) the *in silico* instrument – algorithms, models, workflows, scripts, web services, tools, software, and so on; (ii) the configuration of variables, parameters, error bounds and thresholds; and (iii) the protocols used to choose and apply them. Digital method public repositories include: BioModels for System Biology models; myExperiment for computational workflows; NanoHUB for simulation codes; BioCatalogue for cataloguing Web Services; MethodBox.org for statistical scripts over social science surveys.

Access to (and preservation of) software is a particular issue with the reproducibility of computational experiments. Open software, foundations and open software repositories such as GitHub are technical mechanisms to try to make software source codes available and sustainable. Virtualization platforms such as SHARE [22] try to preserve the software binaries so they can be rerun as they were when a result was announced. SHARE was a prizewinner at the Elsevier's Executable Paper Grand Challenge 2011 (www.executablepapers.com) which set out to stimulate innovations that make it easy to rerun and reuse methods from publications. Other prizewinners adapted existing publication models to include data and analyses, by embedding executable code in papers [41], and facilitated the validation, citation and tracking information through identification schemes for verifiable computational results [19]. The challenge reflects a rather incremental view of scholarly communication in that the hub is still the written paper. Similarly, the *Open Research Computation* journal (www.openresearchcomputation.com) uses traditional reviewed articles to describe software in much the same way as data journals such as GigaScience use traditional articles to publish data. This traditionalism is largely due to credit being tied up with articles, which we discuss in the next section.

The absence of explicit, comprehensive and accurate descriptions of methods is recognized as a serious problem, leading to "black box" science [37], difficulties in peer review and exposure to allegations of fraud. Scientific papers conform to conventions for presenting results rather than accurately and comprehensively describing what really happened. Tacit knowledge and shorthand references are prevalent. Furthermore software might not be available to be executed or examined to accurately assess the algorithm it encodes. We will address the various social reasons for this and why this obfuscation of method is sometimes unintended but often intentional.

Workflows give an accurate and transparent record of the method used to comprehend, justify and compare resultant data products. For example a Taverna workflow uses a mixture of data and software instruments which may be code embedded in the workflow logic, calls to locally provisioned data sets, or calls to services supplied by

public providers. The workflow is deposited in myExperiment so it has a unique identifier for citations. As a freestanding method it is an asset that can be reused – retrieved, referenced, and repeatedly used on its original data or new data. For a specific execution bundled in a pack with the configuration of input data and variable settings, the provenance of its run can be examined to "replay" its execution.

However, even if a computational method is recorded and shared it might cease to be useful. "Reproducibility by re-run" is sensitive to the stewardship of components, especially web services hosted by third parties. Components alter, become incompatible or unavailable. Step decay leads to workflow decay, equivalent to an instrument becoming obsolete. Techniques are available to reduce this decay; e.g. Taverna workflows deposited in myExperiment that use web services deposited in BioCatalogue are monitored for such circumstances.

Where actual reproducibility cannot be achieved, partial reproducibility is a means to play back workflow execution based on the provenance of previous executions [13]. Even in cases where the workflow cannot be executed and no provenance trace is included, the details of the workflow description may be enough to justify and explain a result. To review research requires retrieval of workflows which are transparent but not necessarily executable. However, to re-run these workflows as "black boxes" they could be executable but not transparent. It is useful to distinguish the former case as *preservation* of a workflow and the latter as *conservation,* whereby a method is restored or repaired so as to be re-executable. Workflows also lend themselves to *repurposing*; i.e. reuse of a method against different settings, for example different datasets or parameter configurations, but also *structural reuse*, for example substituting or reordering steps; and *fragment reuse*, using parts or combinations of parts of the workflow [60]. The descriptions of the relationships between these parts and how they are assembled inform their re-use. As the majority of experimental designs are variants, reusing scientific workflows leads to better procedures and spreading best practice. Reproducibility in computational science turns out to have many subtle aspects that warrant further investigation: the Wf4Ever project (www.wf4ever-project.org) is investigating workflow reproducibility, preservation and conservation, proposing a framework based on "Research Objects" which we discuss in Section 5.

4 Open Knowledge Flow: The Common Good vs. Self-Interest

In general, we are moving towards an era of greater transparency in all of these topics (methodology, data, communication and collaboration). We have the instrument and information technology to deliver, but do we have the necessary social conventions? Mismatched motivations, value placed on knowledge and social capital, reward schemes, poor reciprocity and distrust together conspire to block the circulation of knowledge [38] [6].

Open Science [46][48] encompasses the ideals of transparency in experimental methodology, observation, and collection of data coupled with the public availability and reusability of scientific data and public accessibility and transparency of scientific communication. Openness is a means of achieving accelerated knowledge transfer

and networked science [36]. Underlying this open ideal is a notion of voluntary sharing of methods, results, and scholarship, and the objects of scholarship belonging not to the individual scientist but to the larger community [55]. However, scientists are people working within their social norms and as self-interested as any other group of people. Their prime motivations include funding, building reputation, and getting sufficient time, space, and resources do their research. Sharing results is not a motivation in itself, so has to be placed within a context of maximizing reward, minimizing risk and optimizing costs:

- *Reward for Sharing*: to gain competitive advantage over rivals by establishing a claim on priority of a result; to establish public reputation and recognition through credit; to accelerate the widespread adoption/acceptance of a result; to gain access to otherwise unavailable instruments, data, techniques or expertise.
- *Risk of Sharing*: the threat of rivals gaining a competitive advantage; damage to public reputation through scrutiny or misinterpretation; not getting credit; a sense that others will get a "free-ride".
- *Cost of Sharing*: the time and resources needed to prepare; the inconvenience and/or difficulty in preparing to share or sharing; potential long-term sustainability obligations.

Stodden [55] gives an excellent sociological account of scientific sharing practices, broadly summed up as: scientists do not like to share and when they have to they prefer to share with those they trust and be rewarded for it. Although they do not always trust data other than their own, they like to be shared *with*. This mismatch leads to "Data-Mine-ing" – your data is mine and my data is mine. Liebowitz also argues that the intrinsic worth of knowledge is a sharing factor related to the scarcity of prized commodity in local asset economies. Researchers will trade when there is a local unavailability of an asset such as specialist data. If assets are expensive and have to be collectively obtained, like the Large Hadron Collider or a telescope, the consortium typically obliges the results to be collectively shared. However, if assets are local investments, like a Next Generation Genome Sequencer, and data is locally available, then there is no need to share to acquire, and if the data is scarce and prized it will be protected.

The behavior of the members of the SysMO-SEEK consortium highlights the value placed on knowledge capital and the distrust that lies between rivals, manifested as incremental sharing that widens the availability of content as its local value proposition changes. At first (or perhaps only) an individual or laboratory uses the e-Laboratory as a private, preserved repository. This is useful when scientists are mobile, moving from grant to grant and institution to institution. Next, *trusted* collaborators within each project may exchange pre-published content. Results shared outside a trusted group prior to publication are rare. When a scientific article is finally published publicly we could expect its associated data/method/model to be deposited publicly. However, if the investigator thinks they can wring another paper out of some data they will not share it even if it is the basis of an announced result. Data are only made widely available when their local capital is exhausted. We also observe that (i) models, procedures and workflows are more likely to be openly shared than data,

suggesting that the scientific community places greater value on data than experimental method; (ii) formal consortia are less likely to publicly share than individuals; and (iii) young researchers and very senior well established researchers are more willing to share than mid-career researchers in the midst of establishing their reputations.

The e-Laboratories we have built deal exclusively with non-personalized data. In the social and clinical sciences a common objection to data commons is the risk to personal privacy. Although of course important, in practice the Chordoma Foundation and other patient groups have found that patients are far more positive towards data sharing than their self-appointed guardians in the clinical profession. Yakowitz argues that risks from anonymized data rarely materialize, and current privacy policies overtax valuable research without reducing any realistic risks [62]. A discussion about the throttling of clinical knowledge exchange by well meaning but ill-informed ethics committees is a topic for another paper.

4.1 Reciprocity and Flirting

The tendency to protect data is sometimes called "data hugging". In myExperiment and SysMO-SEEK we observe "data flirting" where scientists strategically (or maybe tacitly) hold back information, communicating *just enough* to interest their community and publish *just enough* to preserve their claim for priority on the findings but *not enough* in practice for competitors to be able to take advantage. Specialized knowledge on experimental details is withheld. Scientific jargon is used to frustrate competitors. This "counterfeit sharing" is prevalent when funders make data sharing directives. The data is deposited, and thus "shared", but it is hard to find and impossible to reuse or reproduce [38].

Borgman [6] highlights the underlying mismatches in the motivations to share by data producers and data consumers. myExperiment highlights some examples. Providers of workflows want credit and, sometimes, control of who benefits from their work or how it is used. Consumers tend to follow a "search-download-logoff" pattern, wanting to easily reuse the workflow without constraint. However, they often fail to credit the provider, or contribute comment or review. They do not feedback results that arose from using the workflow and it is hard to track the workflow from published research unless it was explicitly cited. This lack of feedback fosters a sense of "free riding".

Reciprocity is fundamental to effective knowledge exchange. Where myExperiment is used by an organized group, good citizenship is more governed. However, when it is used by individuals who do not even have to be registered members the social pressure for reciprocity is absent or at best tacit. To close the reciprocity feedback loop, and secure/preserve their reputation, workflows are described by their producers such that the consumer must enter into a dialogue to reuse them. The author can then secure credit or negotiate a beneficial collaboration. They can also protect their reputation by guarding against misuse, or maybe even withhold access altogether. Interestingly, this dialogue is usually conducted outside myExperiment making it invisible to other consumers (and the service provider).

4.2 Reusability: The Burden of Curation

To be reusable an asset must be sufficiently well described that it can be understood, so that the consumers and producers have a shared understanding [30]. However, the cost of preparing metadata (known as *curation* and *annotation*) combined with data flirting conspires against data being sufficiently well described to be reusable [38]. The workflow in Fig. 2 integrates several datasets. Its complexity is due to mapping between various identifiers, nomenclatures, schemas, structures, formats and terminologies. Reuse of data is hard. Consumers need assets described as well as possible by understandable and explicit reporting standards: insufficient contextual information about methods means they cannot really be trusted or validated, and if results cannot be understood they will be reinvented rather than reused. Furthermore if they cannot be redone then there is a risk of misunderstanding and misuse. Reuse is correlated with familiarity (we find this in myExperiment) but this reduces the opportunity for innovation and cross-discipline sharing. To overcome the many esoteric formats and nonstandard terminologies that face consumers, biology has developed a range of *reporting standards:* 150+ minimum reporting guidelines, 260+ terminologies and 50+ exchange formats listed by the BioSharing.org initiative, mostly targeted for specific data types.

For providers, curating is a burden. The range of standards is bewildering and many are difficult to adopt. A survey of 160 major biology data providers revealed that although 74% used controlled vocabularies only 26% used community standards, and although 31% used minimum reporting checklists; only 8% used those recognized by the community [53]. It takes knowledge, skill, effort and time to curate, especially when using new or combinations of technologies [47], and consistent description – particularly when contributors self-curate – is hard, especially in the absence of tool support. In reality, quality curation is only evident when professional, dedicated curators, such as members of the International Society for Biocuration, are paid to do it.

The general view is that the best time to describe something is at the time it comes into being. For data that would be when it is collected. For a workflow it would be when it is designed. However, experimental data may be acquired with the expectation it be thrown away or never published. A workflow may be created for private or temporary use, and it is uneconomic to curate results that are anticipated to be disposable or private. So curation is left until later when the reporting information is difficult to recover retrospectively. Online, public e-Laboratories are particularly vulnerable to poor curation. Self-curation is commonly sketchy as it is usually only intended for the author or close colleagues. Contributions by non-authors, for example reviews and ratings, require effort. Our experiences indicate that the technical incorporation of reporting standards is relatively straightforward but getting contributors to curate against them is very difficult.

4.3 Easing the Curation Burden: Ramps for Knowledge Transfer

Convenient knowledge mechanisms are a crucial component of successful knowledge transfer between people [30], and hence knowledge turning. Providers, and consum-

ers, need convenient *automated and manual curation ramps*. A "ramp" is a mechanism embedded in a routine practice or familiar tool that eases a user to use a technology. For example, a common instrument for data collection is the Microsoft Excel spreadsheet. The RightField tool [59] wires acceptable terms from community controlled vocabularies into Excel templates. Data is thus collected using selections against the correct terminologies by the experimentalist without changing their work practices or tools. *Stealth ramps* attempt to gather metadata in the appropriate tool at the right time in a familiar work practice, a kind of knowledge acquisition by stealth. *Automated ramps* use instrumentation for knowledge collection, for example the Taverna Workbench is instrumented to record usage and responses of workflows (from myExperiment) and services (from BioCatalogue) to feedback on operational and usage profiles. *Collective ramps* (re)assemble the component parts of an investigation scattered across databases. Commons, wikis, our e-Laboratories and special programs like ENCODE in genomics, are examples.

Collective ramps are vulnerable to a kind of "tragedy of the commons". People use them but do not contribute to or maintain them, instead relying on other curators to integrate and validate descriptions, and other data providers to submit and check data. Recent work in identification of co-author groups and formally declared consortia are first steps in establishing responsibilities for stewardship over complex datasets spanning multiple institutions, journals, databases and funders. The suggestion is that more complete and granular information about the people who generate knowledge will contribute to sustainable access to the datasets in perpetuity [39]; however, fundamentally we need to ensure contributors are rewarded.

4.4 Credit where Credit is Due

Liebowitz [30] highlights reciprocity, the intrinsic worth of knowledge, and interpersonal trust and respect as factors for successful knowledge sharing. Scientific reputation is the key measure of worth and respect, and the giving and gaining of credit is the way we express reciprocity and measure reputation. Incentivizing through credit is needed to accelerate the sharing and adoption of results and reward the burden of curation. Scientists fight to get their names onto papers because currently credit is based on peer-citation of articles. However, now web-scale information technology can build credit and attribution networks at the *article* (not journal) level and for *all* digital instruments. Altmetrics (altmetrics.org) and Scientometrics widen credit metrics to all the commodities of science essential for communication (blogs, wikis), and reproducibility (software, data, models, methods), the better for measuring impact. Downloads and views, service calls, expert opinion in public peer review, links, bookmarks, citations in twitter or blogs all contribute to a richer picture. myExperiment download and visit statistics, cross-attributions and trackable derivations, ratings and favorite bookmarking, and references in papers are all bragging material for their authors. Technologies for data citation, like Datacite.org, and for tracking the attribution and acknowledgement of researchers, like Orcid.org, needs to be wired into our data commons.

Technically, we have to rethink what we mean when we cite data and methods. Published articles do not change but database entries improve and software evolves. Citing resources that are both archival and active is an open question. Citing web pages that are in flux is similarly challenging [56]. Provenance tracking transparently records where results came from, but we also need accurate propagation of attributions on method variation and data derivation using shared knowledge models for citation [51] and provenance, such as that proposed by the W3C's Provenance Working Group (www.w3.org/TR/prov-dm).

Sociologically, we need community governance regulators to recognize the value of alternative metrics and to build reputation and asset economies for data, method and code. For example, software, because it can be copied and distributed at essentially no cost, opens the door to unprecedented levels of sharing and collaborative innovation. However reputation economies of software production are not well rewarded through the traditional reputation economy of science leading to over-production of independent scientific software packages, and the under-production of collaborative projects based on pre-existing codes [25]. Paid service professionals, scientists who produce intrinsic software as a by-product of their work, and scientists who are already well established produce software because they do not need academic reward. Where the software *is* the academic credit then recognition, respect and sustainability is essential or else it will not happen. Similar findings are made for data [6]. Data and software journals are temporary measures between old and new credit systems.

Curators are skilled people, required to be experts in both instrument and information technologies, and are motivated by many different drivers that are not just financial reward. The most significant incentives are reputation building, altruistic improvement of the quality of information for their field and the not-so-altruistic promotion of their ideas. However, curators are undervalued and low worth is applied to their contribution, even the professional ones. Curators, and scientific informaticians, need to be first class citizens rather than the "blue collar" scientists as they are often viewed, without a recognized career path. Crowd-sourced curation is needed to match the scales of data but we know that voluntary peer review does not just happen [2]. Reviewers, commentators and curators of third party, openly published data need tangible recognition and social kudos. An excellent review should be citable.

5 Research Objects and Workflow Preservation

We have argued that scientific workflows are computational *instrument technologies* for executing methods and *information technologies* for recording and disseminating computational method. However, they are not only "technical objects" which can be interpreted and executed by machines. They are simultaneously "social objects" that are shared by people as part of the flow in scientific knowledge turns in collaborative environments like myExperiment. As *information objects* they have reusable and repurposable knowledge capital in their own right. They encode scientific methods and know-how and are thus worth exchanging and preserving, subject to the same stresses as data for sharing, credit, attribution and curation. As *digital instruments*

they are components in the reproducibility of computational experiments, components of a greater experimental design, linked with publications, more data and other computational and experimental methods, and aggregations combining data, configurations and executable steps (services, codes, tools, databases).

We need workflows to be both embedded and non-embedded within scholarly communications, to behave as independently and collectively, and to be both a record and an executable. We need a form of information currency that allows workflows to: (i) be, and be part of, an aggregated and richly annotated complex scholarly communication; (ii) be an accurate, citable preserved record of method; and (iii) be an actively conserved executable method.

The printing press gave us the academic paper, but a picture of a workflow in a PDF is not going to do the job. Force11 [16] calls for "a new, enriched form of scholarly publication that enables the creation and management of relationships between knowledge, claims and data" and the need for "a full record of the research undertaken requires preservation of these processing steps and software tools employed, in addition to the datasets upon which they acted".

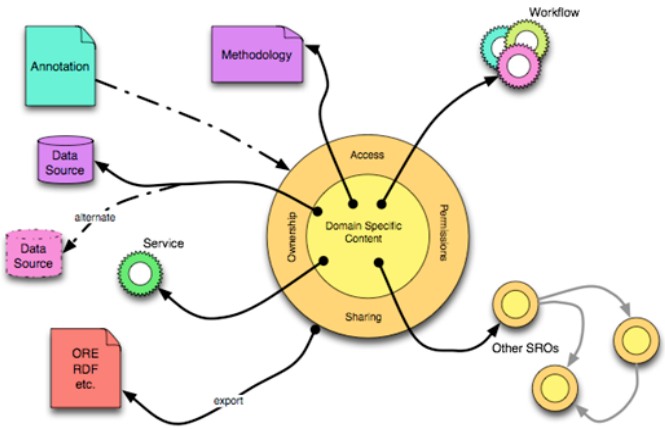

Fig. 4. A notional schematic of a Research Object

The Research Object (RO) [4], depicted in Fig. 4, aggregates potentially scattered resources that are collected in order to support research activity, investigation and experimentation. A Workflow Research Objects bundles a workflow, together with provenance traces obtained by its enactment, and annotations that semantically describe the domains of the workflow parameters, its operational semantics, its version history, author attributions, citation credit, license etc. An Experiment Research Object bundles a Workflow Research Object with others and with data, ideas, people, a description of the experiment, publications and so on, and the semantic interrelationships between them.

The notion of collecting or aggregating resources is not new. Hunter proposes the idea of Scientific Publication Packages (SPP) to describe "the selective encapsulation of raw data, derived products, algorithms, software and textual publications" [28].

SPPs are motivated primarily by the need to create archives for the variety of artifacts produced during the course of a scientific investigation; they ideally contain data, methods, software and documents, but also their provenance as well. The LiquidPub project introduces Scientific Knowledge Objects [20], which describe aggregation structures intended to describe scientific papers, books and journals. A key aspect of this approach is consideration of the *lifecycle* of publications in terms of "states": Gas, Liquid and Solid, which represent early, tentative and finalized work respectively.

Wf4Ever (www.wf4ever-project.org) is an EU-funded STREP project that aims to develop technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows, in particular through definition of suitable Research Objects and the services that support their creation and management. The focus of the project is on workflows as concrete descriptions of *method*, and of workflow-centric Research Objects, particularly examining the duality of both preserving computational workflows and conserving them in order to support reproducibility. The specific user domains targeted in the project are Astrophysics and Genomics.

From a social perspective ROs must support different roles. For the creator and contributor they should gather credit metrics and provide citation mechanisms; for the reader they should gather and provide quality measures; for the re-user they should provide execution and repair mechanisms; for the reviewer they should provide a means to compare reruns and for the curator they should provide mechanisms to gather, maintain and validate annotations. From a technical point of view ROs that use external, specific or fragile services, data or execution platform are susceptible to *decay*. Changes to, or unavailability of, external resources may compromise the possibility of re-executing a workflow and "reproducibility by rerun" is likely to be problematic. ROs need to carry sufficient provenance information in order to support the replay of an execution [13] or an examination of the processes enacted in order to allow validation.

The aggregated structure of the RO also supports the repurposing of constituent parts. The RO provides the container within which information relating to the workflow and its use or execution can be maintained. If this information is of sufficient detail to allow a refactoring of the process (e.g. via the substitution of an alternate appropriate service or data set), method *conservation* can be achieved. In terms of the social role of Research Objects, *understanding* is the key to useful preservation. Again, the inclusion of information regarding the provenance of results is facilitating the understanding of those results that then supports reusability.

The Workflow Research Object approach in Wf4Ever defines (i) an abstract RO data model and concrete encodings; (ii) core RO services (such as credit management and execution capability management) and added-value RO services (such as recommendation systems and quality control); and (iii) the protocols for interoperating services and managing the model. Using standardized web infrastructure makes the approach backwards compatible, adoptable by publishers and libraries and future enabled. We specifically propose encoding ROs into Linked Data [23]. Linked Data has uptake in scientific domains [7] and neatly fits scholarly and semantic publishing [50]. We intend that the RO model itself be small and will extensively reuse community ontologies for citation, discourse, provenance and so on. Similarly we aim to

leverage standard models for aggregation and harvesting, such as OAI-ORE and OAI-PMH. Linked Data publication is complemented and enriched by ROs [4]. The ROs not only aggregate resources, but add additional annotations and metadata supporting both the technical and social roles that those objects play.

The view of workflows (and associated aggregations) as social objects has also been observed in the myExperiment platform [12] with packs demonstrating a role in workflow reuse and curation. The Wf4Ever RO reference implementation of services is being built using myExperiment, Taverna, BioCatalogue and the dLibra Digital Repository. However the aim is not to create a monolithic system but lightweight components that can be incorporated into data services such as DataVerse (www.thedata.org), digital repositories and platforms such as Galaxy.

6 Conclusions

A "knowledge turn" is the cycle of a process that derives more good leading to an advance. The majority of scientific advances in the public domain result from collective efforts that depend on rapid exchange and effective reuse of results. We have powerful computational instruments, such as scientific workflows, coupled with widespread online information dissemination to accelerate knowledge cycles.

However knowledge turns between researchers continue to lag. Open science and open data are still movements in their infancy and method obfuscation continues to obstruct reproducibility. An ecosystem of "Research Objects" is a possible technical approach towards reusable results. In this paper we have suggested that the shared objects of scientific practice, which underlie knowledge turns, are both technical and social – and indeed that the real obstacles are social.

Strategically, we want to do our best to circulate results and methods so we can attack the big goals of science, like curing Chordoma and protecting the cattle of Africa. But operationally the metrics, processes and norms developed over the past 50 years need serious revision to meet these goals. Until transparent, open science is rewarded it will remain elusive and safer to hug than share. Until curation is recognized as a necessity rather than a luxury results will remain un-reusable. The whole scientific community – from the lab to the publisher and policy makers – needs to rethink and re-implement its value systems for scholarship, data, methods and software. Otherwise we are seriously letting down Josh Sommer.

Acknowledgements. We acknowledge the many members of the myGrid team. We thank Tim Clark for introducing us to Shapin, and acknowledge Tim's insights along with those of Marco Roos, Jose Enrique Ruiz, Josh Sommer, Chris Taylor, Sweitze Roffel, Robert Stevens, Andy Brass, Paul Fisher, Katy Wolstencroft, Jay Liebowitz, Dawn Field, James Howison, Heather Piwowar, Victoria Stodden, Susanna-Assunta Sansone, Phil Bourne, Scott Edmunds, Anita De Waard and Chris Borgman. We thank Ian Cottam for his unswerving support. The work was supported by EU FP7 270192 Wf4Ever, EPSRC EP/G026238/1 myGrid Platform Grant and the BBSRC BB/I004637/1 SysMO-DB2.

References

1. Adie, E. Commenting on scientific articles (PLoS edition), http://blogs.nature.com/nascent/2009/02/commenting_on_scientific_artic.html Last accessed 11 Feb 2009.
2. Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.A. Public Availability of Published Research Data in High-Impact Journals. PLoS ONE 6(9) (2011)
3. Baker, M. Next-generation sequencing: adjusting to data overload. *Nature Methods* 7 495 - 499 (2010)
4. Bechhofer, S., Buchan, I. et al. Why linked data is not enough for scientists. Future Generation Computer Systems. 2012 *in press* doi:10.1016/j.future.2011.08.004 (2012)
5. Bhagat, J., Tanoh, F. et al. BioCatalogue: a universal catalogue of web services for the life sciences *Nucleic Acids Research* 38(suppl 2): W689-W694 (2010)
6. Borgman, C.L. The Conundrum of Sharing Research. *Journal of the American Society for Information Science and Technology,*1-40 (2011)
7. Chen, B. et al. Chem2Bio2RDF: a semantic framework for linking data and mining chemogenomic and systems chemical biology data BMC Bioinformatics 11, 255 (2010)
8. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T. The SWAN biomedical discourse ontology. *J Biomed Inform.* 41(5):739-51 (2008)
9. Clark, T. and Kinoshita, J. Alzforum and SWAN: The Present and Future of Scientific Web Communities**.** Briefings in Bioinformatics 8(3):163-171 (2007)
10. Daub, J., Gardner, P.P., Tate, J. et al. The RNA WikiProject: Community annotation of RNA families 14 (12): 2462 (2008)
11. De Roure, D., Goble, C. and Stevens, R. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems, 25(5) pp. 561-567 (2009)
12. De Roure, D., Bechhofer, S., Goble, C. and Newman, D. Scientific Social Objects: The Social Objects and Multidimensional Network of the myExperiment Website. *1st Intl Workshop on Social Object Networks* (2011)
13. De Roure, D., Belhajjame, K. et al. Towards the Preservation of Scientific Workflows. In Proc 8th Intl Conf on Preservation of Digital Objects (2011)
14. Eysenbach, G. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact *J Med Internet Res* 2011;13(4):e123 (2012)
15. Fisher, P. et al. A systematic strategy for large-scale analysis of genotype–phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Research*, 35(16) 5625-5633 (2007)
16. Future of Research Communications and e-Scholarship (FORCE 11) Force11 Manifesto http://force11.org/white_paper Last Accessed 2 Feb 2012 (2011)
17. Galperin, M.Y., Fernandez-Suarez, X.M. The 2012 Database Issue and the online Molecular Biology Database Collection *Nucleic Acids Research* 40 D1 D1-D8 (2012)
18. Gargouri, Y., Hajjem, C., Larivière, V. et al. Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE 5(10) (2010)
19. Gavish, M., Donoho, D. A universal identifier for computational results, Procedia Computer Science 4 Proc of the Intl Conf on Comp Science, 637-647 (2011)
20. Giunchiglia, F., ChenuAbente, R. Scientific Knowledge Objects V.1, Technical Report DISI-09-006, University of Trento (2009)

21. Goecks, J., Nekrutenko, A., Taylor, J. and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. Aug 25;11(8):R86 (2010)
22. Gorp, P.V., Mazanek, S., Share: a web portal for creating and sharing executable research papers, Procedia Computer Science 4, Proc Intl Conf on Comp Sci, 589-597 (2011)
23. Heath, T. and Bizer, C. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136 (2011)
24. Hey, T., Tansley, S., Tolle, K. (eds) The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft (2009)
25. Howison, J., Herbsleb, J.D. Scientific software production: incentives and collaboration, Proc ACM 2011 Conf Computer Supported Cooperative Work, 513-522 (2011)
26. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M., Li, P., Oinn, T. Taverna: A tool for building and running workflows of services. *Nucleic Acids Research* 34:W729-W732 (Web Server Issue) (2006)
27. Hull, D. et al. Defrosting the digital library: bibliographic tools for the next generation web PLoS Comput Biol 4 (10) e1000204 (2008)
28. Hunter, J. Scientific Publication Packages – A Selective Approach to the Communication and Archival of Scientific Output, Intl J of Digital Curation 1 (1) (2006)
29. Kell DB, Oliver SG, Here is the evidence, now what is the hypothesis? *BioEssays* 26(1) 99-105 (2004)
30. Liebowitz, J., Ayyavoo, N., Nguyen, H., Carran, D., Simien, J. Cross-generational knowledge flows in edge organizations, Industrial Management & Data Systems, 107(8) 1123-1153 (2007)
31. Levison, S. E. et al. Colonic transcriptional profiling in resistance and susceptibility to Trichuriasis: phenotyping a chronic colitis and lessons for iatrogenic helminthosis. *Inflammatory Bowel Diseases* 16(12):2065-79 (2010)
32. Lintott, C. J., Schawinski, et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389: 1179–1189 (2008)
33. Maleki-Dizaji S., Rolfe M., Fisher P., Holcombe M. A Systematic Approach to Understanding Bacterial Responses to Oxygen Using Taverna and Webservices. *13th International Conference on Biomedical Engineering. 2009:* 77-80 (2009)
34. Mesirov, J. Accessible Reproducible Research *Science* 327(5964), 415-416 (2010)
35. McLuhan, M. Understanding Media: The Extensions of Man. McGraw Hill (1964)
36. Nielson, M. Reinventing Discovery: The New Era of Networked Science. Princeton University Press (2011)
37. Nature Editorial “Illuminating the black box”. *Nature* 442, 1 (2006)
38. Nature Editorial “Data's shameful neglect”. *Nature* 461, 145 (2009)
39. Nature Genetics Editorial “It's not about the data”, *Nature Genetics* 44, 111 (2012)
40. Nature Special Issue on Big Data, *Nature* 455 (2008)
41. Nowakowski, P., Ciepiela, E. et al. The collage authoring environment, Procedia Computer Science 4. Proc Intl Conf on Comp Science, 608-617 (2011)
42. Noyes, H. et al. Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. PNAS *108(22)* 9304-9309 (2011)
43. Orr, J. One Good Turn. http://cofes.com/About/OneGoodTurn/tabid/57/Default.aspx Last accessed 31 Jan 2012.

44. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A. and Attwood, T.K. Ceci n'est pas un hamburger: modelling and representing the scholarly article. Learned Publishing, 24 (3): 207-220 (2011)
45. Rennie, C., Hulme, H., Fisher, P., Halp, L., Agaba, M., Noyes, H., Kemp, S., Brass, A. A systematic, data-driven approach to the combined analysis of microarray and QTL data. *Developments in biologicals*, 132:293-299 (2008)
46. The Research Information Network. Open science case studies. http://www.rin.ac.uk/ (2010)
47. Sansone, S-A. et. al. Toward interoperable bioscience data. *Nat Gen* 44, 121–126 (2012)
48. Schroeder, R. e-Research Infrastructures and Open Science: Towards a New System of Knowledge Production? Prometheus: Critical Studies in Innovation, 25:1, 1-17 (2007)
49. Shapin, S. Pump and Circumstance: Robert Boyle's Literary Technology Social Studies of Science 14(4) pp: 481-520 (1984)
50. Shotton, D., Portwin, K., Klyne, G. and Miles, A. Adventures in semantic publishing: exemplar semantic enhancement of a research article. PLoS Comp Bio 5 (4) (2009)
51. Shotton, D. CiTO, the Citation Typing Ontology, J of Biomed Sem 2010, 1(Suppl 1):S6 (2010)
52. Shotton, D. The Five Stars of Online Journal Articles — a Framework for Article Evaluation. D-Lib Magazine January/February 2012 18(1/2) (2012)
53. Southan, C. Elixir Database Provider Survey http://www.elixir-europe.org/prep/bcms/elixir/Documents/reports/WP2_Annex-Provider_Survey_Report.pdf (2009)
54. Sommer, J. "Sage Commons: Josh Sommer, Chordoma Foundation". Video available on http://fora.tv/2010/04/23/Sage_Commons_Josh_Sommer_Chordoma_Foundation
55. Stodden, V. The Scientific Method in Practice: Reproducibility in the Computational Sciences. MIT Sloan Research Paper No. 4773-10. Available at SSRN: http://ssrn.com/abstract=1550193 or doi:10.2139/ssrn.1550193 (2010)
56. Van de Sompel, H. et al. An HTTP-Based Versioning Mechanism for Linked Data. *Proc of Linked Data on the Web (LDOW2010)*, http://arxiv.org/abs/1003.3661v1 (2010)
57. Waldrop, M. Big data: Wikiomics *Nature* 455, 22-25 (2008)
58. Whitfield, J. Collaboration: Group Theory. *Nature* 455, 720-723 (2008)
59. Wolstencroft, K., Owen, S. et al RightField: Embedding ontology annotation in spreadsheets. *Bioinformatics* 27(14) 2012-2022 (2011)
60. Wroe, C., Goble, C. et al. Recycling workflows and services through discovery and reuse. *Concurrency and Computation: Practice and Experience* 19(2) 181-194 (2007)
61. Wuchty, S., Jones, B.F., Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* May 18;316(5827):1036-9 (2007)
62. Yakowitz, J. Tragedy of the Data Commons. Harvard J of Law and Tech, Vol. 25 (2011)