Scientific workflows: Past, present and future

Malcolm Atkinson, Sandra Gesing, Johan Montagnat, Ian Taylor

Scientific Workflows: Past, Present and Future

Malcolm Atkinson[a], Sandra Gesing[b,], Johan Montagnat[c], Ian Taylor[b,d]

[a] *University of Edinburgh, School of Informatics, Edinburgh EH8 9AB, UK*
[b] *University of Notre Dame, Center for Research Computing, Notre Dame, IN 46556, US*
[c] *Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, FR*
[d] *Cardiff University, School of Computer Science & Informatics, 5 The Parade, Cardiff CF24 3AA, UK*

Abstract

This special issue and our editorial celebrate 10 years of progress with data-intensive or scientific workflows. There have been very substantial advances in the representation of workflows and in the engineering of workflow management systems (WMS). The creation and refinement stages are now well supported, with a significant improvement in usability. Improved abstraction supports cross-fertilisation between different workflow communities and consistent interpretation as WMS evolve. Through such re-engineering the WMS deliver much improved performance, significantly increased scale and sophisticated reliability mechanisms. Further improvement is anticipated from substantial advances in optimisation. We invited papers from those who have delivered these advances and selected 14 to represent today's achievements and representative plans for future progress. This editorial introduces those contributions with an overview and categorisation of the papers. Furthermore, it elucidates responses from a survey of major workflow systems, which provides evidence of substantial progress and a structured index of related papers. We conclude with suggestions on areas where further research and development is needed and offer a vision of future research directions.

Keywords: scientific workflows, scientific methods, optimisation, performance, usability

1. Introduction

Data-intensive Workflows (a.k.a. scientific workflows) are routinely used in the majority of data-driven research disciplines today, often exploiting rich and diverse data resources and parallel and distributed computing platforms. Workflows provide a systematic way of describing the methods needed and provide the interface between domain specialists and computing infrastructures. Workflow management systems (WMS) perform the complex analyses on a variety of distributed resources. With the dramatic increase of primary data volumes and diversity in every domain, workflows play an ever more significant role, enabling researchers to formulate processing and analysis methods to extract latent information from multiple data sources and to exploit a very broad range of data and computational platforms.

Email addresses: malcolm.atkinson@ed.ac.uk (Malcolm Atkinson), sandra.gesing@nd.edu (Sandra Gesing), johan.montagnat@cnrs.fr (Johan Montagnat), ian.j.taylor@nd.edu (Ian Taylor)

2. Highlights over the past 10 years

This special issue celebrates significant progress over the past ten years that has greatly increased the use of data-intensive workflows, built on substantial improvements in their usability, capabilities, architecture and reliability. Ten years ago there were diverse scientific workflow systems showing promise and early use [1]. We asked leaders of workflow groups, "*What was the most significant result from using your workflow system in the last 10 years?*" Their replies are collated in Table A.1; it offers an extensive body of evidence with comprehensive coverage and a structured index into the literature. Some highlights are presented here[1]. Pegasus played a key role in the detection of gravitational waves by offering sustained and reliable data handling and computation for the long-running research campaign[2]. Increased capacity, scale and reliability is reported in almost every case. The scientific workflow community spawned new technology developments in workflow-based data provenance. Kepler made early progress in this work and after a number of grand challenges, Gill started activity at W3C with an incubator[3] that proposed a core vocabulary[4] and which led to a working group (Moreau and Groth) who saw through the process all the way to W3C PROV standard[5]. Building on that standard, Taverna and WINGS have advanced workflow description significantly, initiating a foundation for reasoning about multiple workflow languages consistently. Building long-term relationships with communities, including tuning access via well-crafted science gateways and composing extensive libraries of workflows and workflow fragments has proved particularly productive, pioneered by Taverna but now almost universal. A striking example is Galaxy, with thousands of users on its public site and 4,300 publications citing Galaxy's contribution to their results in the last 10 years. That progress means that using workflows has become routine, *e.g.* KNIME. Table A.1 contains many examples of delivering success to others using the power of data-driven workflows.

There remains a diversity of systems, many with their own investments, culture and committed communities; some have remained leaders while others have been replaced, but the quality, support and maturity has increased across the board [2]. The scientific workflow community has educated the world; 10 years ago very few researchers had heard of workflows, today virtually every domain uses them. Some aspects of the consolidated progress are presented.

The tools for the whole workflow lifecycle have much improved, eliminating many impediments to use, and greatly improving the productivity of all those, from domain experts to data engineers, who work with workflows. Ten years ago the tooling was focused on authoring and adapting to distributed computing interfaces. It now extends to managing research campaigns using workflows, with automation of data identification, exploitation of provenance, support for curation and oversight of progress, processes and performance [3]. The combination of improved abstraction and full-lifecycle tools has made it much easier

[1] References to Table A.1 refer to the rows associated with the named workflow system.
[2] https://pegasus.isi.edu/application-showcase/ligo/
[3] https://www.w3.org/2005/Incubator/prov/charter
[4] https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/
[5] https://www.w3.org/TR/2013/REC-prov-dm-20130430/

for end users to understand, select, reproduce, adapt and use previously developed workflows. This builds on the sustained investment in building libraries of workflows, workflow fragments, re-usable components, accessible services, and established models for describing and cataloging these resources so they are easily found [4, 5].

The abstract definition and representation of workflows has advanced significantly, *e.g.* as reported in this issue by Garijo *et al.* [6]. This yields four benefits:

1. The meaning of scientific methods encoded in workflows is less dependent on implementations and therefore can be sustained as the digital environment evolves, thereby extending the benefits from investing in developing scientific methods as workflows.
2. This independent definition permits mappings to diverse platforms, enabling WMS to exploit the latest advances in hardware and software platforms.
3. Scientific workflows are more easily reused and re-purposed, lowering experimental design cost and speeding-up discoveries.
4. Abstract definitions also facilitate sharing ideas and effective methods across discipline boundaries.

The development, deployment and use of pioneering data-intensive workflows that cope with the scale of modern data, the rates of demand, and the diversity of enactment contexts continues to require effective alliances between innovating domain experts, adept data scientists, and experienced systems engineers. However, substantial advances have been made to accelerate their work and improve productivity. Practitioners can work with moderate scale test data sets on their personal devices or local facilities, and then use exactly the same workflows on production platforms. Several factors contribute to this achievement:

1. Improvement in virtualised infrastructures, abstractions, tools and monitoring introduced above.
2. Automation and optimisation of data handling, coping with diverse sources, eliminating unnecessary transfers, and discarding unneeded storage.
3. Planning and optimisation that automatically adapts to resource availability, scale and load.
4. Minimised recovery costs after partial failures.

The software stacks that provide parallelised multistage distributed heterogeneous target environments for workflows have increased capacity, elasticity and recoverability. These platforms have and will continue to evolve thereby increasing the importance of abstraction and automated mappings as a means of preserving the meaning of scientific methods. However, this can pose challenging set up requirements, to provide the initial enactment context or to rebuild an earlier enactment context for scientific reproducibility – some early experience tackling such issues has been reported [7, 8]. Many of these data-intensive platforms also have their own languages for creating data-driven methods, that are intimately integrated and well presented. These are emerging as new data-intensive workflow systems that appeal to communities who are not constrained by prior investments. The usability from such integrated systems has to be weighed against the potential for lock in.

Data distribution and data streaming are of growing significance. Over the past decade, the emergence of the Web of Data[6] and the Open Linked Data initiative[7] led to a much increased environment of remotely accessible and interoperable data sources for scientific analysis. In addition, data streaming occurs in the monitoring and exploitation of data from connected instruments, worn devices and the Internet of Things (IoT). Initially, it was the domain of signal processing communities and was treated differently; today the handling of data units flowing in streams from any source merges with the task-oriented traditional workflow approach. The Node-red system is a prime example[8]. The developers of several stream-based workflow enactment models show that this can also achieve substantial performance gains by reducing disk traffic [9].

There is a pervasive drive by funders, governments and researchers to increase the openness of research and the accessibility of data, motivated by the desire to increase equality, stimulate use and cross-fertilisation, and to improve quality by facilitating challenging review. This is captured as the FAIR principles [10]. The related workflows and their enactment context should be included, but workflows should also help implement those principles.

The diversity of data-intensive workflow languages will persist, partly because of prior investment but also because they are tuned to meet different requirements. Part of that investment is intellectual: learning, developing skills and understanding, and part is cultural, the interaction with other users. These factors, as well as wanting to continue practices that have proved effective, weigh against change. We may anticipate further languages emerging, from the interplay of programming language research, workflow system research and new application requirements. These will continue to yield benefits but the multiplicity of languages has inherent costs:

1. Research methods often draw on methods that have already been encoded. If these are not in the language a researcher is using, or differ in the language they use, they cannot be composed without adaptation.
2. Workflow management systems require substantial development, maintenance and support effort. The separate language communities partition this effort.

Several researchers are circumventing this issue by mapping methods in the individual languages into a common language. Garijo *et al.* [6] report one example in this issue as they address the first symptom, while Plankensteiner *et al.* [11] proposes an intermediate workflow language addressing the second one. Terstyanszky *et al.* [12] report a different approach with alternative formal foundations [13]. The Common Workflow Language (CWL) [14] has been developed by an informal, multi-vendor working group consisting of various organizations and individuals that have an interest in portability of data analysis workflows, e.g., Galaxy [15] and Taverna are participants in this group. Askalon, Taverna, Triana and Wings also report relevant contributions – see Table A.1. Tavaxy [16] follows a different approach and was built for interoperability between Galaxy and Taverna. It allows for the submission of workflows in both languages.

[6] https://www.w3.org/standards/semanticweb/data
[7] http://linkeddata.org/
[8] http://nodered.org

3. Special Issue Overview

This special issue focuses on key advances in workflow engineering and delivery over the ten years for which the WORKS conference has been held, drawing on work progressed during the first nine years, and the leading papers from the tenth conference, WORKS15[9]. A core focus of this special issues was to look ahead, identifying the critical issues that deserve attention in the future and how they may be addressed.

This special issue therefore covers a broad range of topics in the scientific workflow lifecycle that include:

- the notation for describing, refining and sharing workflows;
- workflow design and composition interfaces;
- workflow provenance;
- the underpinning management systems that enact workflows including the management of data;
- workflow mapping techniques that may optimise the execution of the workflow;
- workflow enactment strategies that need to deal with failures in the application and execution environment; and
- a number of computer science research challenges related to scientific workflows, such as semantic technologies, compiler methods, fault detection and tolerance.

As such, this special issue presents various insights into the field of scientific workflows, which we categorise into three areas:

1. *reviews* of the current state of the art and taxonomies for better classification of systems;
2. *usability*, including provenance, re-purposing, repeatability and validation; and
3. *performance and optimisation*.

These are presented in the following three sections.

3.1. Review Papers

In their paper, "A Characterization of Workflow Management Systems for Extreme-Scale Applications" Ferreira da Silva *et al.* [17] present workflows and their characteristics in extreme-scale computing, which will bring forth multiple challenges for the design of workflow applications and management systems. The authors define a novel characterisation of workflow management systems and classify 15 popular workflow management systems in terms of workflow execution models, heterogeneous computing environments, and data access methods.Their work elucidates gaps for future research on the road to extreme-scale workflows and management systems.

In the paper "Software architectures to integrate workflow engines in science" [18], Glatard provides a comparison of six different software architectures commonly used to integrate workflow engines into science gateways. By analysing these architectures, the authors

[9] http://users.cs.cf.ac.uk/Ian.J.Taylor/works15/

extract several different types of components and interactions and provide a set of metrics that can be used to rate the architectures, including complexity, robustness, extensibility, scalability, and the support for meta-workflows as well as fine-grained debugging.

3.2. Usability

Usability is split into three sub-categories: —emphre-use, *reproducibility* and *validation*. The first sub-category includes two papers that address re-use and re-purposing of workflows [19, 6]. The second sub-category addresses reproducibility, an area that has drawn much attention of late, and includes two papers [20, 21]. The third sub-category focuses on the validation of workflows [22].

3.2.1. Re-using and re-purposing workflows

In "Scientific Workflows in Data Analysis: Bridging Expertise Across Multiple Domains" by Sethi *et al.* [19], the authors focus on interdisciplinary research. They demonstrate the use of scientific workflows in this challenging context by re-purposing workflow fragments in the areas of text analysis, image analysis, and analysis of activities in videos. The paper highlights how the reuse of workflows allows scientists to link across disciplines and to save time as well as effort. In-depth studies of various tasks in multiple areas such as machine learning and computer vision elucidate the benefits for the researchers who use their system's assistance.

The second paper in reuse and repurposing category, entitled "Abstract, Link, Publish, Exploit: An End to End Framework for Workflow Sharing" by Garijo *et al.* [6] tackles an issue vital for the long-term health of scientific workflows by addressing the challenge of making workflows more easily understood and (re)used by typical users who are not expert in computing science notations and concepts, though they are adept at developing or employing data-driven and computational (scientific) methods. Typically, their principal area of expertise is in some application domain, and they need to use workflows fluently as a tool in their work (research, information production, decision support or policy formulation) without being distracted by having to learn strange technical dialects that may prove ephemeral. The authors publish and exploit workflows as data on the Web, with a representation that is independent from the workflow system used to create them. They achieve this goal by adopting Linked Data principles and by developing standard ontologies in order to publish workflow inputs, intermediate results, outputs and codes.

3.2.2. Reproducibility

By focusing on workflow reproducibility in "Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities" [20] Cohen-Boulakia aims to deliver a set of criteria for reproducibility-friendly workflow systems, which can be used as an evaluation framework for assessing the level of reproducibility of workflow systems and companion tools. The authors extract requirements from three use cases: The Phenome project[10], which uses the OpenAlea workflow [23]; the French health and science

[10] https://www.phenome-fppn.fr/

agency, National Cancer Institute[11] (INCa), which use several workflow systems (including Galaxy and KNIME); and Transcriptomics analysis, which uses the Galaxy workflow. The authors identify four levels of reproducibility: repeat, replicate, reproduce and reuse, and use the resulting taxonomy to compare four workflow systems: Taverna, Galaxy, OpenAlea, and NextFlow.

The second paper is entitled "Applications of Provenance in Performance Prediction and Data Storage Optimisation" by Woodman *et al.* [21]. Workflow environments are often instrumented to capture and deliver provenance information to end users and system administrators, as provenance can be critical to properly interpret scientific data derived from computations or help in identifying software or hardware problems during workflow execution for instance. In this paper, the authors show how provenance traces can be used to predict performance of future workloads and to determine data regeneration versus data storage costs. Both aspects are discussed in the context of workflows execution repeatability, an important feature for scientific worklow systems as illustrated in the medical data analysis domain.

3.2.3. Validation

Validation is addressed by the paper "Static Analysis of Taverna Workflows To Predict Provenance Patterns" by Alper *et al.* [22]. In this paper, the authors identify a class of potential problems when workflows depend on multiple inputs, with cross-product like operations in the graph, and forms of iteration over values in those inputs. This leads to potential imprecision in the provenance traces that may limit their value for diagnostics, replay, partial re-enactment and indexing the intermediate or result data.The paper very carefully develops a formal understanding of these issues and proposes a workflow design approach that avoids them. It then develops algorithms that statically verify that a Taverna workflow complies with these properties.

3.3. Performance and optimisation

This section is sub-divided into three complementary areas related to optimising workflows performance: parallel computing [24], scheduling and planning [25, 26, 27, 28], and data management [29, 30].

3.3.1. Parallel computing

Assessing the evolution of parallel computing architectures and the generalisation of multi-core systems, the paper on "A workflow runtime environment for many core architectures" by Janetschek *et al.* [24] describes a language, compiler and workflow enactor for workflow execution using a multi-core, shared-memory machine rather than a more traditional network-connected distributed computing infrastructure. This work tightens the links between the scientific workflow and parallel computing domains, recognizing the workflow abstraction as a generic formalism both suited to describe High Throughput Computing and High Performance Computing applications.

[11] http://en.e-cancer.fr/

3.3.2. Scheduling and planning

For scheduling and planning, we have four papers. In "Deadline constrained scientific workflow scheduling on dynamically provisioned cloud resources", Arabnejad *et al.* [25] tackle the important problem of cost-efficiently exploiting cloud resources, especially commercial cloud solutions. They present their algorithms Proportional Deadline Constrained (PDC) and Deadline Constrained Critical Path (DCCP) that address the workflow scheduling problem on dynamically provisioned cloud resources and prove that these algorithms achieve better cost efficiency and success rates than previous algorithms.

Time critical applications, on the other hand, are particularly challenging in cloud environments due to the difficulty of securing guaranteed performance from the underlying virtual infrastructures. In the paper "Planning Virtual Infrastructures for Time Critical Applications with Multiple Deadline Constraints", Wang *et al.* [26] propose a cloud infrastructure planning algorithm that accounts for multiple overlapping internal deadlines on sets of tasks within an application workflow. They evaluate their algorithm using a large set of workflows generated at different scales with different execution profiles and deadlines. Their proposed algorithm can satisfy all overlapping deadline constraints with a consistently lower host cost in comparison with the widely used algorithm IC-PCP.

In the paper by Chirkin *et al.*, "Execution time estimation for workflow scheduling" [27] the authors tackle the important problem of estimating a workflow's execution time. They propose a solution that takes into account the complexity and the stochastic aspects of the workflow components as well as their runtime. The proposed solution addresses the problems at different levels from tasks to workflows leading to reductions in the time to complete workflows. The authors have integrated their approach into the scheduling algorithm GAHEFT and elucidate the benefits via tests within the CLAVIRE platform.

Finally, in "Orchestration and Analysis of Decentralised Workflows within Heterogeneous Networking Infrastructures", Macker and Taylor [28] apply the Network Edge Workflow Tool (Newt) to workflows on dynamic heterogeneous wireless networks. They apply time-windowed conversational graphs to analyse and orchestrate distributed workflows. They demonstrate the power of their approach by showing how it would produce, *e.g.* schedule actors responding to one another, the Shakespearean play Hamlet with each actor on a different mobile device.

3.3.3. Data management

In the data management sub-category, we have two papers. The "Raw Data Queries during Data-intensive Parallel Workflow Execution" paper by Silva and co-authors [29] presents a timely work on massive raw data querying at workflow runtime and links with the workflow management system provenance component. At a time of ever increasing scientific data set volume and complexity, the components described offer an abstraction to index and query raw data files, so that data sets can be described independently from their file location and structure. This feature is important to describe data flows where structure is not tightly bound to technical aspects such as data representation.

Al-Kiswany *et al.* in "A Cross-Layer Optimized Storage System for Workflow Applications" [30], report developments improving distributed file storage of long-term value. Their

proposed modification to the interface with the file system better supports workflow systems by providing a non-disruptive two-way communication channel between the workflow and storage system, exploiting standard UNIX file interfaces. They demonstrate significant performance improvements for typical workflows when the file system has been informed of anticipated data access patterns. They carefully support incremental adoption. That is, any code exploiting their new communication with the file system still works in contexts where the file system has not changed, and code written for standard Unix file systems still runs in their new context. Their new storage system improves file usage in almost all cases.

4. Trends for the next 10 years

During the next ten years, we believe we will see advances in the following areas:

1. Much *consolidation* in this field, as existing advances are brought together and further developed.
2. Continuing improvement in *usability* for an expanding range of workflow users.
3. Substantial *engineering* investment to exploit new technologies leading to very significant improvements in efficiency, performance and sustainability.
4. Extension of support to encourage more extensive and sophisticated *data-powered collaboration.*
5. A move towards more *dynamic decentralised workflows* that meet the needs of emerging environments, such as IoT and Edge computing.

Each of these aspects of workflow research and development (R&D) are considered separately below, but in reality they will be intimately intertwined. For example, improvements in engineering will yield responsiveness that helps usability, whereas, the verifiable rule enforcement needed to achieve sustained collaboration across autonomous organisations will incur costs. Improved architectures will mitigate these effects, as described in one paper in this issue [18]. Improvements in theory underpinning workflow languages and systems, such as the work of Wadler[12] and Cheney [31], will become more essential as complexity and scale grow. It is to be hoped that relevant architectural and theoretical research will proceed in tandem. The separate lines of R&D introduced below, will then need to be properly influenced by that progress.

4.1. Consolidation

A critical factor in consolidation will be a steady improvement of abstraction to enhance the precision and conceptual clarity of formal encodings of data-driven and computationally intensive scientific methods. This will improve sustainability as engineering exploits new technologies, and will increase practitioners ability to find, share, reuse and compose workflows from other communities, thereby accelerating interdisciplinary research and pooling the intellectual effort needed to refine methods. It will lead to new architectures, where all of

[12] http://links-lang.org/

the user-facing tools and interfaces will be cast in terms of these abstractions, while the abstractions may be mapped to a variety of underlying implementations through standardised interfaces. This has the following advantages:

1. Fewer intellectual hurdles to be mastered before proficiency is attained in finding, using, revising, authoring or improving data-intensive workflows.
2. Much longer sustained value of that learning, of investment in data-driven methods and of their associated culture and collected libraries.
3. Domain experts should be able to take responsibility for methods, innovate and move to production, only calling on workflow experts for exceptional requirements.
4. Support for the creation and enactment of workflows composing fragments from multiple languages.
5. An opportunity to support more communities via one workflow platform development path, thereby amortising costs and sharing benefits.
6. Increased opportunities for sharing methods between communities and for exploiting shared methods and working practices.

What remains to be seen is the extent to which the application domains and communities converge on common solutions. The cultural momentum and the fundamentally different requirements will sustain at least presentational differences, but paper [6] in this issue shows the holistic potential of building abstract models of workflows, their description and their enactments.

4.2. Usability

The usability of interfaces for authoring and revising workflows has progressed substantially over the last ten years [32]. That will continue, and improved search for components, workflow fragments and data sources will amplify the benefits. The usability advances will broaden to include all aspects of workflow and data lifecycles. Better tools are emerging with improved interfaces and adaptable visualisations, with greater scope, and with increased power to select targets and apply operations to all members of a target. They will extract and use relevant information from standardised provenance records. This will facilitate diagnostics and validation, and lead to better understanding of the processes encoded and the ways in which they have been used. Thus, domain experts will be equipped to take responsibility for the quality of workflow-encoded methods and research campaigns using them, thereby improving the reliability and reproducibility of evidence from data; a key requirement if presented conclusions are to influence society's response to global challenges. The scale of the provenance data used by these tools poses performance and presentational challenges. These grow if we consider using these data, through appropriate tools to manage research campaigns involving millions of data sets and thousands of applications of methods. Without appropriate tools, practitioners will be inundated with the provenance information and swamped with management chores. Users need the power to specify rules that discover and inject their domain-related information into provenance streams, so that they can navigate and specify bulk operations in terms that have meaning in their domain. All of the

underlying systems will need to respond to the controls and bulk operations and supply provenance streams in a consistent form.

Usability will include presentation best adapted to meet different practitioners needs from visualisation to the routine use of packaged methods to innovation inspired by the power unleashed by the latest technical, algorithmic and hardware advances handling unprecedented quantities and varieties of data. Monitoring and logging will be enhanced with more interactive components for intermediate stages of active workflows. These interactions with workflows will be accessible through all media from handheld devices to immersive interaction studios. As the tools and operations to manage a whole workflow lifetime become more sophisticated, as the variety of users increases, and as the scale of communities and research campaigns grow, the measurement of usability with involvement of the communities will be essential. Diverse usability standards, metrics and guidelines exist [13] to measure usability and fundamental research on usability is well supported. In industry the importance has been recognised already years ago and considered in product design and software engineering and is evident in Apple products, for example. The scientific community will catch up via integrating usability measures into software engineering for scientific workflows. It will be necessary to consider novice users, citizen scientists and experts with different provisions for each and intellectual ramps to support users as tailored as possible to their needs but also encourage them to develop their expertise and fluency. At least four categories of expert need to be considered: (1) *Domain scientists*, who collect and organise required data, understand relevant phenomena and identify the questions to ask and strategies that lead to the relevant evidence; (2) *Research developers*, who formulate methods to implement those strategies drawing on extensions to the products from the next two expert groups; (3) *Data scientists*, who develop algorithms for handling uncertainty, variable data quality, data heterogeneity and statistical inference by tailored application of machine learning; and (4) *Systems engineers*, who deliver services on distributed e-Infrastructures able to handle a growing scale of data, very demanding computations and high rates of interaction and data transport. Usability will extend beyond supporting each category of user well. It will also need to consider how best to help them combine their efforts to build the data-driven research systems that are needed even for today's challenges. Consequently, research into how to define and measure relevant workflow usability, will be needed to inform the R&D improving the delivered usability.

4.3. Engineering

Many factors drive improvement in the engineering that supports workflow systems; these include:

1. Exploitation of the advances in computing platforms and their connected storage systems brought about by fast non-volatile memory that blurs the difference between storage and memory; by 3D memory addressing, providing massive memory bandwidth; by near-memory processing; and by chip-integrated photonic interfaces, that enable ultra-low-latency and high-bandwidth communication.

[13] http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/

2. Exploitation of the software, algorithms and provision models developed for data handling, databases and data science delivered by contemporary industrial and commercial R&D. These have been very dramatic in the past ten years, and we can be sure that the major drivers of digital economies will continue to dominate this digital ecosystem. Data-driven science is a minor player in comparison. If it does not adapt well to the imposed changes it will not thrive. However, it should also draw directly on technical and theoretical advances, *e.g.* the automatic parallelisation of graph algorithms [33] which will have wide application in scientific workflows.
3. The models and cost functions underpinning optimisation will progress to be more realistic and more sophisticated. They will exploit the recurrent patterns of usage to parameterise these models accurately to cope with growing scale and complexity – this will exploit machine learning both to discover and characterise critical patterns, and to recognise when workflows and their data are sufficiently similar that they benefit from the same optimisations. They will cover partitioning and distributing workflows so that the parts better match the platforms they run on, and the overall trade off between cost and performance is substantially improved. Cost functions will give increasing weight to total energy consumption.
4. The components, services and functions that scientific workflows can draw on will individually improve and grow in diversity, partly by importing capabilities automatically from other informatics contexts. They will be better described owing to the advances in abstraction described above, and better characterised by mining information from the growing population of previous runs. These components will cover the full range of data types and semantic categories, including graphs, time series, arrays and measures of uncertainty, that dominate scientific data, thereby enriching the creative context, *e.g.* the scientific databases: RASDAMAN[14], Monet-DB[15], Exareme [34] and SciDB[16] all parallelise user-defined functions taking account of data placement – something workflows will need to do.

4.4. Extending data-powered collaboration

The improved communication provided by better abstractions and descriptions will boost collaboration, but only within organisations, closely knit communities or where there are no rules, mores, privacy concerns or competition pressures that inhibit sharing. Many of the application domains require more sophisticated management of rules and attribution of credit. The collection of data, its lifetime management, and its exploitation is often organised by institutions, teams and individuals to meet priorities determined by their stakeholders, funders and research leaders. They commit to continuously improving the quality of their processes, delivered data and offered services. Their identity, justification for continued support, limited resources and changing priorities, all need to be respected when data from their work is brought into a data-intensive federation (DIF) to be combined with data from

[14] http://www.rasdaman.com/
[15] https://www.monetdb.org/Home
[16] http://www.paradigm4.com/

other sources to present a long-lived holistic research environment for a cluster of purposes, *e.g.* to help address a global, social or economic challenge. We give three examples of clusters of research campaigns each of which would motivate the formation and long-term support of a DIF:

1. Organising the response to natural hazard events, including advice to responders and follow up advice to communities and authorities for build-back-better campaigns. This draws on natural hazard models, local topological, geological, hydrological and land-use data, including the distribution of vulnerable people. It may draw on citizen data sources, social media, satellite images and rapidly deployed field instruments. Initial urgency is followed by several years of support actions. Paper [28] in this issue, considers a scenario contributing in this context. The DIF holds commonly required data, and has methods for assembling and revising the data needed for each event. NGOs and others would analyse the effectiveness of response and communication strategies using the data from multiple events.
2. Advising governments and agricultural organisations on appropriate crop choices taking into account the effects of global warming, current local farming practices, land exposure, slope orientation, altitude and surface geo-chemistry. Here the data includes similar topological, geological, hydrological and land use data. It will include models of the effects of climate change on pests and require climate model re-scaling. Here again, there will be sensitive data concerning current land use and farm-management practices.
3. Predicting and measuring the environmental impact of mineral resource exploitation. Again models of the Earth's surface and systems need to be linked with local data and combined with the commercially sensitive mineral extraction plans, *e.g.* to increase lithium extraction to support smart devices and green-energy storage. This would include both atmospheric and coastal marine chemical and ecological time series, as well as the commercial-in-confidence operational data from companies proposing or conducting the extraction.

These examples cover similar domains to illustrate engagement in multiple DIFs – see below. In Europe, all three of these could adopt the bundle of standards and rules incorporated in the INSPIRE directive [35] intended to facilitate geo-spatial data sharing, but each would need to extend this to handle specialist data, and add rules governing computation, optimisation, accounting, publication and attribution. Of course, there are a huge number of potential DIFs supporting distinct endeavours drawn from a wide range of fields; overlap may prove a rarity. How will we support those planning, formulating and steering a DIF, *e.g.* negotiating over rules in order to obtain access, that is not over restricted, to critical data? For example, they will need tools with underpinning workflows, that predict the consequences when rules meet or when one organisation changes its rules.

Frequently, requirements such as scale, access to urgently needed quality assured data products, rapid ingest of revised methods and honoring newly imposed rules, prevent co-location in a data warehouse. A DIF often spans disciplines, cultures, nations and communities. Provider participants in a DIF, often find they need to engage with many DIFs,

e.g. each national geological survey would contribute to all three of the above DIFs[17]. Each DIF has its own dynamics that include: (i) changes to agreements on the information content and services of the common view, (ii) changes in rules of conduct constraining DIF use or in a contributing organisation constraining its own activities, and (iii) revisions of the dynamic mapping algorithms that transform the latest data when it is needed. Maintaining support for just one of these, as it draws dynamically on an institution's data and services is hard. Dealing with many, each with different trajectories will become infeasible without improved tools to automate adaptation and incorporation of those changes, both in the workflows delivering the provider's contribution and in the workflows implementing the research methods using the DIF. When providers deliver the same data or provide the same service to multiple DIFs, they expect to do this with the same workflow in all cases. However, during execution that workflow will need to comply with the rules of the DIF from which the work was requested.

As we tackle more complex problems, the importance, size and number of DIFs will grow. Workflows will be key to both the resource providers and to researchers using the combined resources. Templates for both roles have been prototyped but a full repertoire of exemplars is needed. These workflows and their supporting systems must incorporate DIF requirements, *e.g.* working across and respecting multiple legal frameworks, or performing optimisation, such as caching, with associated approved accounting, so that the originators still get credited from cached hits. This requires the development of (standard) notations for describing DIF rules that are both comprehensible to the data diplomats negotiating such rules and capable of verifiable implementation across all of the platforms that data and workflows may visit. Such rules need to be comprehensible as individuals, institutions and organisations need to accept and respect them, as their formulation with sufficient flexibility and their implementation with sufficient reliability will depend on good will and ingenuity.

Robertson *et al.* [37] explore these issues in the limited context of sharing medical data between countries. The management systems then inspect the workflows for compliance, *e.g.* verifying that personal details are not passed across jurisdiction boundaries, and that statistical data comes from at least five subjects. The workflows may be automatically extended to achieve compliance, *e.g.* using the reasoning WINGs reports in Table A.1. As another example from a different context, when a facility has been used for a production run over a threshold, and the results should be announced and made available to others within a set time, the workflow would be extended to automate that scheduled publication. Such support for definable trustworthy rules that can be negotiated and validated will be necessary as expertise and data is harnessed to address our world's most pressing challenges. Technical, organisational and socio-political considerations will interlock to force radical rethinking. For example, suppose a collection of data has to be removed to comply with rules on consent or emergency conditions ending. How will a workflow be synthesised that cleans up *all* of the caches, registries and catalogues? How should this interact with archival and curation processes? What should provenance for reproducibility mean in such

[17]For example, one data scientist at BGS is responsible for its role in five DIFs, including aspects of the above three examples [36].

circumstances? How should we support the concerned citizen investigating the quality of the rules and their implementation in DIFs that are of legitimate concern? There are clearly many questions here where the strength and ingenuity of the workflow research community will be an essential input.

Today most DIFs are being constructed by *ad hoc* methods and the workflows they use have to be manually adapted by providers for each DIF they contribute to, and by users and developers to add or test the implementation of the DIF's rules, ethics and agreed practices. The workflow research community should rally to this cause as DIFs will grow in importance and number, and will become a predominant context for vital data-intensive applications in science, business and government.

4.5. Dynamic decentralized workflows

The data collected across the IoT landscape from devices is vast. According to Cisco, all of the people and things connected to the Internet will generate 507.5 zettabytes of data by 2019[18]. Sending all this information to centralised servers will lead to bottlenecks, increase latency and be inefficient, especially when some or all of processing can be computed near the source(s). As data from systems, devices and people become more intertwined, information will naturally become more distributed, and scattered across multiple locations. The *Edge computing paradigm* acknowledges this trend and attempts to push the frontier of computing applications, data, and services away from centralised nodes to the edges of the network.

Currently scientific workflows are not designed to operate with such environments due to their centralised management, and cannot effectively deal with dynamic environments that may consist of collections of edge nodes offering varying amounts of resources including bandwidth, storage, and computation. We envisage therefore a move towards workflows that become more capable of spontaneously aggregating services, computation and analytics, on collections of geographically distributed and potentially transient edge devices.

5. Conclusion

The creativity and ingenuity of the scientific workflow community will build on the substantial advances of the last 10 years in unpredictable ways, as well as those anticipated by the rest of the papers in this celebratory issue, and the five directions we envisage above. Workflows will continue to play a central role in the strategies adopted to address our global challenges and underpin data handling and knowledge discovery in the major research projects, *e.g.* DALiuGE an attempt to prepare for the data volumes (180 petabytes of ready-to-use archived science data per year starting 2023) from SKA1[19] [38]. Data-driven workflows will push the limits in three directions simultaneously; at the same time they will deliver stability upon which communities can safely build – see Figure 1. The challenges for scientific workflows are then:

[18] http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf

[19] http://skatelescope.org/

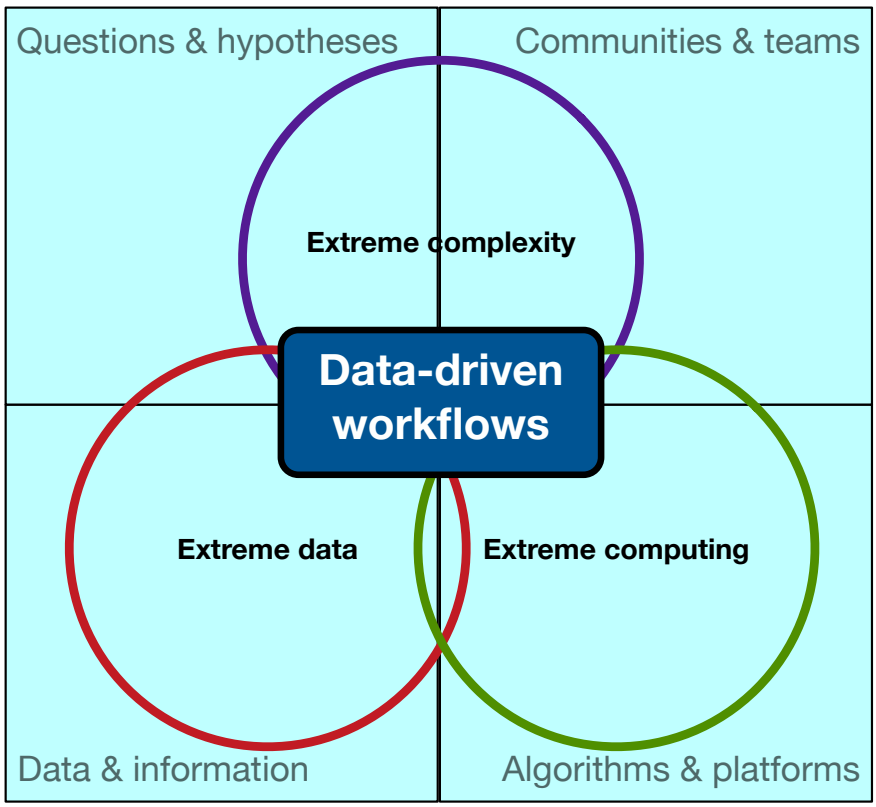

Figure 1: Scientific workflows simultaneously extend our capabilities in three critical directions: coping with more complexity, data and computation. They deliver an island of stability amid four seas of change. *Complexity* grows as more difficult issues are tackled and as more viewpoints contribute. *Data* capacity grows to meet demand and as society appreciates its value. *Computing* capacity grows to meet demand imposed by data and complexity. The *questions and hypotheses* evolve as understanding develops. The *communities and teams* expand, recruit new talent and revise priorities. The *data and information* grows and diversifies as a result of sustained campaigns. The *algorithms and platforms* capture new insights and incorporate new technologies. The scientific workflows retain the value of previous intellectual investment while reaping benefits from advances in all four seas.

- to provide the conceptual and practical framework that fosters synergy between experts contributing advances from all four seas of change, and
- to provide sustained precision and validation that enables them to deliver reliable, actionable evidence.

Acknowledgements

We thank Supercomputing for hosting WORKS – now extended to SC17. We thank the workshop chairs and reviewers of the Supercomputing Conference and the WORKS PC members, speakers and attendees. They have been and remain vital for WORKS success. We are grateful for the contributions of the authors and reviewers of this special issue and the editors of Future Generation Computing Systems.

We thank Aurora Constantin, Oscar Corcho, Rosa Filgueira, Alex Hardisty, Alessandro Spinuso and Luca Trani who commented on drafts of this paper.

We also thank those who helped us identify ten-year achievements, namely: Michael Berthold, Dave Clements, Ewa Deelman, Thomas Fahringer, Rosa Filgueira, Yolanda Gil, Carole Goble, Jeremy Goecks, Peter Kacsuk, Tamas Kiss, Bertram Ludäscher, Marlon Pierce, Rafael Ferreira da Silva, Stian Soiland-Reyes, James Taylor and Michael Wilde.

Atkinson was supported by the EU H2020 project ENVRI^{plus} No. 654182[20] and by SKA-Link[21]. Gesing was supported by the Science Gateways Community Institute NSF Award Number ACI-1547611[22], by the IMLS project "Planning a Research Data Software Preservation Quality Tool" Award Id LG-72-16-0122-16[23] and by the Center for Research Computing at the University of the University of Notre Dame[24]. Montagnat was partly funded by the French National Research Agency (ANR) through the "Investments for the Future" Program reference ANR-11-LABX-0031-01[25]. The core Triana workflow system (Taylor) was supported by PPARC (GridOneD and Geo 600) and we thank the EU for funding the SHIWA[26] project workflow interoperability work. Triana also received support for STFC (TRIACS[27] ST/F002033/1), OMII-UK (WHIP) and the Sintero[28] (Wellcome Trust).

Appendix A. Collation of survey responses

The table below collates the responses received from experts answering the question "*What was the most significant result from using your workflow system in the last 10 year?*"

Table A.1: Advances and successes in the past 10 years

[20] http://www.envri.eu
[21] http://amiga.iaa.es/p/330-SKA-Link.htm
[22] http://sciencegateways.org/
[23] http://presqt.crc.nd.edu/
[24] http://crc.nd.edu/
[25] http://www.ucnlab.eu
[26] http://www.shiwa-workflow.eu/
[27] http://triacs.cs.cf.ac.uk/
[28] http://www.openehr.org/news$_e$*vents*/*industry*$_n$*ews.php*?*id* = 35

WFS	Advances or major achievements
Airavata	Apache Airavata-based gateways [39] have supported over 100 scientific publications, primarily in biophysics and computational chemistry, since the project was founded in 2012. The top-cited paper that Airavata has supported, according to Google Scholar and Web of Science, is Harkness *et al.* [40].
ASKALON	The ASKALON workflow system was ported to a federated infrastructure of private and public Cloud providers using OpenStack middleware. The ASKALON scheduler was extended to optimise in a multi-objective space with conflicting parameters such as execution time, economic cost, energy consumption, reliability, and instability/variability. This yielded the Multi-Objective Heterogeneous Earliest Time algorithm (MO-HEFT – extending the HEFT algorithm). It achieves a good approximation to the Pareto frontier of trade-off solutions [41]. The ENTICE project, http://www.entice-project.eu/, enriched ASKALON with multi-objective optimised placement of fragmented VMs from a distributed repository with faster delivery, enhanced availability, and reduced storage costs [42]. ASKALON provides automatic composition of workflows using semantic descriptions of workflow activities and ontologies; delivering resilience by creating alternative workflows or sub-graphs [44].

continued on next page

continued from previous page

Advances and successes in the past 10 years

WFS	Advances or major achievements
dispel4py	The dispel4py system was developed to enable seismologists to exploit their growing wealth of data – two examples are: Seismic Ambient Noise Cross-Correlation and MISFIT calculation. The cross-correlation detects seismic-wave velocity variations to reveal stress-field changes – a key achievement was to cross-correlate traces from 1000 seismometers every hour on the geoscientists' shared machine before the next traces arrived [9]. The MISFIT takes synthetic seismograms from simulations and compares them with preprocessed real observations to improve Earth or earthquake models (Inversion). The VERCE Science Gateway exposes the MISFIT calculation workflow as a service, in combination with the simulation phase. Both phases can be configured, controlled and monitored by seismologists via a user interface that hides the complexity of accessing computational and data services. The system collects W3C-PROV provenance data.
Galaxy	In the past 10 years, Galaxy has grown to have three distinct and complementary components. First, it is a public web service at http://galaxyproject.org that is an installation (or instance) of the Galaxy software combined with many common tools, visualizations, and data sources. The site provides substantial CPU and disk space, making it possible to analyze large datasets. The public site supports thousands of users and hundreds of thousands of jobs per month (https://bit.ly/gxystats). Second, the Galaxy open-source application (distributed under the terms of the permissive Academic Free License; https://getgalaxy.org) that can be deployed on any Unix system. The Galaxy software is highly customizable and integrates with a wide variety of compute environments ranging from laptop computers to clusters to compute clouds. Third, the Galaxy Community, which has a global community of users, tool developers, bioinformaticians, and administrators who maintain Galaxy instances. The Galaxy Tool Shed (https://usegalaxy.org/toolshed) facilitates sharing tools developed by the community between instances of Galaxy by providing a central location where tool developers can upload both their tool configurations and 'recipes describing how to install necessary dependencies. Galaxy is now used in domains ranging from genomics to proteomics to areas as diverse as social science and natural language processing. Over the last decade Galaxy has been referenced, used, implemented, and or extended in over 4300 publications, including over 3500 journal articles.

continued on next page

continued from previous page

Advances and successes in the past 10 years

WFS	Advances or major achievements
gUse	Kacsuk *et al.* [45] and Kiss *et al.* [46] demonstrated, using WS-PGRADE, how scientific workflows could tackle the interoperability of computational and data resources of Grid computing architectures. A coarse-grained workflow interoperability approach that enables embedding non-native workflows into a hosting meta-workflow has been described by Terstysanszky *et al.* [47]. The importance of such meta-workflows has been demonstrated in Herres-Pawlis *et al.* [48] by combining heterogeneous workflow systems to build meta-workflows for computational chemistry. WS-PGRADE/gUSE is widely used via domain-specific science gateways – over 30 have been built in the SCI-BUS project [49]. Significant results are reported: life sciences (Kiss *et al.* [50] and Shahand *et al.* [51]), and astrophysics Costa *et al.* [52].
Kepler	One of the early and most-cited works on scientific workflows in general and Kepler in particular is [53], which provides an overview, vision, and desiderata of scientific workflow systems, introducing technical concepts such as "smart re-run", data provenance, different models of execution (via Ptolemy II's directors), and a comparison with business workflows. The scientific workflow community spawned new technology developments in workflow-based data provenance (*e.g.* [54, 55, 56, 57, 58] all employed Kepler in one form or another), leading first to the Open Provenance Model [59] and subsequently to the W3C PROV standard. Other technology advances include Kepler applications in Map-Reduce style processing [60], heterogeneous models of computation [61], and data curation workflows [62]. Kepler has been used in numerous science domains, e.g., ecological niche modeling [63], bioinformatics [64, 65] and particle physics [66], among many others.
KNIME	KNIME [67] has focused on usability for life sciences, chemistry, finance and media through integration (with other tools and multiple data types and sources) and on reproducibility (KNIME workflows from version 1.0 still run and produce exactly the same results in today's version). This has delivered convenience so that one customer automatically runs nightly training for tens of thousands of predictive models, but tuning to meet application community needs also yields discoveries [68, 69].
MOTEUR	MOTEUR [70, 71] is a scientific (data-driven) workflow manager designed to efficiently enact large data flow-based computations, leveraging the capacity of various distributed computing infrastructures. At its foundation lies the clear separation between the scientific process description, the parallel enactment engine, and the target execution infrastructure. The scientific process is described through the well-defined GWENDIA data-driven workflow language [72], which requires no parallel computing expertize, yet can represent very complex data flows. This language significantly inspired the design of the IWIR workflow language for interoperability [11]. Its data-driven nature makes the enactment engine able to automatically distribute computing tasks with the highest asynchronicity level. The computing tasks are abstracted so as to remain as independent as possible from the execution platform [73]. Provenance and reasoning on dense workflow execution traces bridges the gap between enactment technicalities and the (high-level) scientific process [74]. User-intelligible experiment summaries computed from these logs enable validation of the semantic coherence of workflows.

continued on next page

continued from previous page

Advances and successes in the past 10 years

WFS	Advances or major achievements
Pegasus	Pegasus [75, 76] builds on the foundation of proven abstractions (directed acyclic graphs, DAGs), fundamental constructs (recursion), and scalable algorithms (graph traversals, graph node clustering). Deployed locally by a scientist with no support from a site administrator and with no impact on the remote cyberinfrastructure, Pegasus allows scientists to submit locally and run globally. A collaboration can deploy Pegasus on a shared submit host to send jobs to campus clusters, high-throughput (HTC) and high-performance (HPC) resources and academic and commercial clouds [77]. The power of this Pegasus-enabled capability was evinced in early 2016 when the LIGO collaboration announced the first detection of gravitational waves from colliding black holes, and Pegasus was used to execute hundreds of thousands of tasks from the main LIGO data analysis pipeline (PyCBC) on HTC and HPC platforms [78, 79]. Pegasus is unique because of its features (portability in time and space, data reuse, and automated data management), scalability and reliability. Pegasus is agnostic of the jobs in the workflow, which can be MPI codes, use GPUs, invoke Map-Reduce, Storm, or Spark "big data" computations.
Swift	The Swift parallel scripting language is notable for its implicitly parallel dataflow-based execution model, which automatically parallelises a high-level functional description of a workflow [80], and for its availability in both a highly portable Java stack for cloud, cluster and grid environments, and a highly scalable MPI runtime for extreme scale environments [81]. Over two dozen diverse science groups have leveraged Swift and published work based on it, demonstrating its usability and generality. Notable examples of its benefits include the recent use of extreme-scale Swift/T to host the EMEWS model exploration system used in cancer classifications [82] and its use in scattering science at both light and neutron source shared facilities to help numerous experiments process data [83] and detect experimental errors right at the beam line, saving invaluable time and effort at the instrument in a true example of boosting scientific productivity [84]. Recently Swift was instrumental in automating the production of digital elevation maps of the Arctic at an unprecedented level of resolution on the Blue Waters supercomputer [85], including novel means to recover from node failure on that petascale system.

continued on next page

continued from previous page

Advances and successes in the past 10 years

WFS	Advances or major achievements
Taverna	Taverna (now in Apache Foundation Incubator) was one of the first open source WFMS and influenced the following generation of WFMS, particularly in the Life Sciences [86, 87, 88, 89]. Taverna has widespread use in other disciplines including Biodiversity [90], Social Sciences, Library Preservation and Astronomy [91]. Taverna pioneered provenance collection [92], orchestrating web services and web service description leading to the ELIXIR EDAM Ontology [93]. Taverna introduced packaging workflow description, discovery, preservation, sharing and metadata into Research Objects [4], leading to myExperiment [94, 95].
Triana	Over the past 10 years, Triana has focused on interoperability mechanisms to integrate with other systems, workflow engines and monitoring infrastructures. This pursues 3 main avenues. For fine grained interoperability, the Interoperable Workflow Intermediate Representation (IWIR) provides a common bridge for translating workflows between languages independent of the underlying computing infrastructure. The IWIR concept has been demonstrated by showing how a workflow could interoperate between Triana, Askalon and the Moteur [11]. For coarse grained interoperability, Triana has implemented the SHIWA bundling format for execution in a multi-workflow environment [96]. Using this approach Triana nests workflow executions using the Open Archives Initiative Object Reuse and Exchange (ORE) standard for workflows and their metadata. Finally, for workflow monitoring, Pegasus and Triana integrated the Stampede monitoring infrastructure in order to add generic interoperable real-time monitoring and troubleshooting across both systems [97]. Stampede provides monitoring via a three-layer model: (1) a common data model to describe workflow runs; (2) tools to load and store logs in that model; and (3) a common query interface.
WINGS	The major contribution from WINGS is the development of representations for semantic constraints for workflows, and the workflow reasoning algorithms that propagate them to generate and validate workflows [98]. This enables WINGS to use abstract descriptions of workflows, and search through possible specializations pruning out ones that are invalid because they violate constraints. This yields a significant contribution: the ability to do automated meta-reasoning about what workflows to retrieve and to test a user's hypothesis – automating scientific discovery [99]. WINGS also separates reasoning about data and components in a workflow from its execution. We can take the same high-level workflow and execute it in different execution engines – demonstrated by WINGS composing workflows for Pegasus and Apache OODT [98].

continued on next page

continued from previous page
Advances and successes in the past 10 years

WFS	Advances or major achievements

References

[1] E. Deelman, D. Gannon, M. Shields, I. Taylor, Workflows and e-Science: An overview of workflow system features and capabilities, Future Gener. Comput. Syst. 25 (5) (2009) 528–540. doi:10.1016/j.future.2008.06.012.

[2] C. S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, J. I. van Hemert, Scientific workflows: Moving across paradigms, ACM Comput. Surv. 49 (4) (2017) 66:1–66:39. doi:10.1145/3012429.

[3] S. Davidson, J. Freire, Provenance and scientific workflows: challenges and opportunities, Proceedings of the 2008 ACM SIGMOD international conference on Management of data (2008) 1–6.

[4] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gómez-Pérez, S. Bechhofer, et al., Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web 32 (2015) 16–42.

[5] N. Cerezo, J. Montagnat, M. Blay-Fornarino, Computer-Assisted Scientific Workflow Design, Journal of Grid Computing 11 (3) (2013) 585–610.

[6] D. Garijo, Y. Gil, O. Corcho, Abstract, Link, Publish, Exploit: An end-to-end framework for workflow sharing, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.

[7] I. S. Pérez, R. Ferreira da Silva, M. Rynge, E. Deelman, M. S. Pérez-Hernández, Ó. Corcho, Reproducibility of execution environments in computational science using semantics and clouds, Future Gener. Comput. Syst. 67 (2017) 354–367.

[8] R. Filgueira, R. Ferreira da Silva, A. Krause, E. Deelman, M. P. Atkinson, Asterism: Pegasus and dispel4py Hybrid Workflows for Data-Intensive Science, in: Seventh International Workshop on Data-Intensive Computing in the Clouds, DataCloud@SC 2016, Salt Lake, UT, USA, November 14, 2016, IEEE Computer Society, 2016, pp. 1–8. doi:10.1109/DataCloud.2016.004.

[9] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Moreno, dispel4py: A Python Framework for Data-Intensive Scientific Computing, Internat. Journal of High Performance Computing Applications.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 160018.

[11] K. Plankensteiner, R. Prodan, M. Janetschek, T. Fahringer, J. Montagnat, D. Rogers, I. Harvey, I. Taylor, Á. Balaskó, P. Kacsuk, Fine-Grain Interoperability of Scientific Workflows in Distributed Computing Infrastructures, Journal of Grid Computing 11 (3) (2013) 429–456.

[12] G. Terstyanszky, T. Kukla, T. Kiss, P. Kacsuk, A. Balasko, Z. Farkas, Enabling scientific workflow sharing through coarse-grained interoperability, Future Gener. Comput. Syst. (0) 46 – 59. doi:http://dx.doi.org/10.1016/j.future.2014.02.016.

[13] J. Arshad, G. Terstyánszky, T. Kiss, N. Weingarten, G. Taffoni, A formal approach to support interoperability in scientific meta-workflows, J. Grid Comput. 14 (4) (2016) 655–671.

[14] doi:10.6084/m9.figshare.3115156.v2.

[15] J. Goecks, A. Nekrutenko, J. Taylor, T. G. Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol 11 (8) (2010) R86.

[16] M. Abouelhoda, S. A. Issa, M. Ghanem, Tavaxy: Integrating taverna and galaxy workflows with cloud computing support, BMC Bioinformatics 13 (1) (2012) 77. doi:10.1186/1471-2105-13-77.

[17] R. Ferreira da Silva, R. Filgueira, I. Pietri, M. Jiang, R. Sakellariou, E. Deelman, A Characterization of Workflow Management Systems for Extreme-Scale Applications, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.

[18] T. Glatard, Software architectures to integrate workflow engines in science gateways, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.

[19] R. Sethi, Y. Gil, Scientific Workflows in Data Analysis: Bridging Expertise Across Multiple Domains, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[20] S. Cohen-Boulakia, Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[21] S. Woodman, H. Hiden, P. Watson, Applications of Provenance in Performance Prediction and Data Storage Optimisation, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[22] P. Alper, K. Belhajjame, C. A. Goble, Static analysis of Taverna workflows to predict provenance patterns, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[23] C. Pradal, C. Fournier, P. Valduriez, S. Cohen-Boulakia, OpenAlea: Scientific workflows combining data analysis and simulation, in: Proceedings of the 27th International Conference on Scientific and Statistical Database Management, SSDBM '15, 2015, pp. 11:1–11:6.
[24] M. Janetschek, R. Prodan, S. Benedict, A Workflow Runtime Environment for Manycore Parallel Architectures, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[25] V. Arabnejad, K. Bubendorfer, B. Ng, Deadline Constrained Scientific Workflow Scheduling on Dynamically Provisioned Cloud Resources, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[26] J. Wang, A. Taal, P. Martin, Y. Hu, H. Zhou, J. Pang, C. de Laat, Z. Zhao, Planning Virtual Infrastructures for Time Critical Applications with Multiple Deadline Constraints, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[27] A. M. Chirkin, A. S. Belloum, S. V. Kovalchuk, M. X. Makkes, M. A. Melnik, A. A. Visheratin, D. A. Nasonov, Execution Time Estimation for Workflow Scheduling, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[28] J. P. Macker, I. Taylor, Orchestration and analysis of decentralized worklows within heterogeneous networking infrastructures, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[29] V. Silva, J. Leite, J. J. Camata, D. de Oliveira, A. L. Coutinho, P. Valduriez, M. Mattoso, Raw Data Queries during Data-intensive Parallel Workflow Execution, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[30] S. Al-Kiswany, L. B. Costa, H. Yang, E. Vairavanathan, M. Ripeanu, A cross-layer optimized storage system for workflow applications, Future Gener. Comput. Syst. This special issue (N) (2017) pp–pp.
[31] S. Fehrenbach, J. Cheney, Language-integrated provenance, in: J. Cheney, G. Vidal (Eds.), Proceedings of the 18th International Symposium on Principles and Practice of Declarative Programming, Edinburgh, United Kingdom, September 5-7, 2016, ACM, 2016, pp. 214–227.
[32] S. Gesing, M. Atkinson, R. Filgueira, I. Taylor, A. Jones, V. Stankovski, C.-S. Liew, A. Spinuso, G. Terstyanszky, P. Kacsuk, Workflows in a Dashboard: A New Generation of Usability, in: Proc. of the 9th Workshop on Workflows in Support of Large-Scale Science, IEEE, 2014, pp. 82–93.
[33] W. Fan, Y. Wu, J. Xu, W. Yu, J. Jiang, Z. Zheng, B. Zhang, Y. Cao, C. Tian, Parallelizing sequential graph computations, in: ACM SIGMOD Conference on Management of Data (SIGMOD), 2017.
[34] Y. Chronis, Y. Foufoulas, V. Nikolopoulos, A. Papadopoulos, L. Stamatogiannakis, C. Svingos, Y. E. Ioannidis, A relational approach to complex dataflows, in: T. Palpanas, K. Stefanidis (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016., Vol. 1558 of CEUR Workshop Proceedings, CEUR-WS.org, 2016.
[35] EU Parliament, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Official Journal of the European Union 50 (L108).
[36] H. Glaves, Data-Intensive Federations in which the British Geological Survey (BGS) is involved, Personal communication at ENVRI week (May 2016).
[37] D. Robertson, F. Giunchiglia, S. Pavis, E. Turra, G. Bella, E. Elliot, A. Morris, M. Atkinson, G. McAllister, A. Manataki, P. Papapanagiotou, M. Parsons, Healthcare data safe havens: Towards a logical architecture and experiment automation, IET Journal.
[38] C. Wu, R. Tobar, K. Vinsen, A. Wicenec, D. Pallot, B. Lao, R. Wang, T. An, M. Boulton, I. Cooper, R. Dodson, M. Dolensky, Y. Mei, F. Wang, Daliuge: A graph execution framework for harnessing the

astronomical data deluge, CoRR abs/1702.07617.
[39] M. Pierce, S. Marru, S. Pamidighantam, B. Demeler, E. Brookes, C. Smith, M. Govindaraju, Apache airavata: Enabling science with science gateways.
[40] K. M. Harkness, Y. Tang, A. Dass, J. Pan, N. Kothalawala, V. J. Reddy, D. E. Cliffel, B. Demeler, F. Stellacci, O. M. Bakr, et al., Ag 44 (sr) 30 4-: a silver–thiolate superatom complex, Nanoscale 4 (14) (2012) 4269–4274.
[41] J. J. Durillo, V. Nae, R. Prodan, Multi-objective workflow scheduling: An analysis of the energy efficiency and makespan tradeoff, in: Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on, IEEE, 2013, pp. 203–210.
[42] D. Kimovski, N. Saurabh, S. Gec, P. Stefanic, G. Kecskemeti, V. Stankovski, R. Prodan, T. Fahringer, Towards an environment for efficient and transparent virtual machine operations: The ENTICE approach, in: 5th IEEE International Conference on Cloud Networking, Cloudnet 2016, Pisa, Italy, October 3-5, 2016 [43], pp. 242–247.
[43] 5th IEEE International Conference on Cloud Networking, Cloudnet 2016, Pisa, Italy, October 3-5, 2016, IEEE, 2016.
[44] J. Qin, T. Fahringer, R. Prodan, A novel graph based approach for automatic composition of high quality grid workflows, in: Proceedings of the 18th ACM international symposium on High performance distributed computing, ACM, 2009, pp. 167–176.
[45] P. Kacsuk, T. Kiss, G. Sipos, Solving the grid interoperability problem by P-GRADE portal at workflow level, Future Gener. Comput. Syst. 24 (7) (2008) 744–751.
[46] T. Kiss, T. Kukla, Achieving interoperation of grid data resources via workflow level integration, J. Grid Comput. 7 (3) (2009) 355–374.
[47] G. Terstyánszky, T. Kukla, T. Kiss, P. Kacsuk, Á. Balaskó, Z. Farkas, Enabling scientific workflow sharing through coarse-grained interoperability, Future Gener. Comput. Syst. 37 (2014) 46–59.
[48] J. Arshad, A. Hoffmann, S. Gesing, R. Grunzke, J. Krüger, T. Kiss, S. Herres-Pawlis, G. Terstyanszky, Multi-level meta-workflows: new concept for regularly occurring tasks in quantum chemistry, Journal of cheminformatics 8 (1) (2016) 58.
[49] P. Kacsuk (Ed.), Science Gateways for Distributed Computing Infrastructures, Development Framework and Exploitation by Scientific User Communities, Springer, 2014.
[50] T. Kiss, P. Greenwell, H. Heindl, G. Terstyánszky, N. Weingarten, Parameter sweep workflows for modelling carbohydrate recognition, J. Grid Comput. 8 (4) (2010) 587–601.
[51] S. Shahand, A. Benabdelkader, M. M. Jaghoori, M. al Mourabit, J. Huguet, M. W. A. Caan, A. H. C. van Kampen, S. D. Olabarriaga, A data-centric neuroscience gateway: design, implementation, and experiences, Concurrency and Computation: Practice and Experience 27 (2) (2015) 489–506.
[52] A. Costa, P. Massimino, M. Bandieramonte, U. Becciani, M. Krokos, C. Pistagna, S. Riggi, E. Sciacca, F. Vitello, An innovative science gateway for the cherenkov telescope array, J. Grid Comput. 13 (4) (2015) 547–559.
[53] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. B. Jones, E. A. Lee, J. Tao, Y. Zhao, Scientific workflow management and the kepler system, Concurrency and Computation: Practice and Experience 18 (10) (2006) 1039–1065.
[54] I. Altintas, O. Barney, E. Jaeger-Frank, Provenance Collection Support in the Kepler Scientific Workflow System, in: L. Moreau, I. T. Foster (Eds.), Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers, Vol. 4145 of Lecture Notes in Computer Science, Springer, 2006, pp. 118–132.
[55] S. B. Davidson, S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, J. Freire, Provenance in scientific workflow systems, IEEE Data Eng. Bull. 30 (4) (2007) 44–50.
[56] S. Bowers, T. M. McPhillips, S. Riddle, M. K. Anand, B. Ludäscher, Kepler/ppod: Scientific workflow and provenance support for assembling the tree of life, in: J. Freire, D. Koop, L. Moreau (Eds.), Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers, Vol. 5272 of Lecture Notes in Computer Science, Springer, 2008, pp. 70–77.

[57] P. Mouallem, R. Barreto, S. Klasky, N. Podhorszki, M. A. Vouk, Tracking files in the kepler provenance framework, in: M. Winslett (Ed.), Scientific and Statistical Database Management, 21st International Conference, SSDBM 2009, New Orleans, LA, USA, June 2-4, 2009, Proceedings, Vol. 5566 of Lecture Notes in Computer Science, Springer, 2009, pp. 273–282.

[58] L. Moreau, B. Ludäscher, I. Altintas, R. S. Barga, S. Bowers, S. P. Callahan, G. C. Jr., B. Clifford, S. Cohen, S. C. Boulakia, S. B. Davidson, E. Deelman, L. A. Digiampietri, I. T. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. A. Goble, J. Golbeck, P. T. Groth, D. A. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. M. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. E. Scheidegger, K. Schuchardt, M. I. Seltzer, Y. L. Simmhan, C. T. Silva, P. Slaughter, E. G. Stephan, R. Stevens, D. Turi, H. T. Vo, M. Wilde, J. Zhao, Y. Zhao, Special issue: The first provenance challenge, Concurrency and Computation: Practice and Experience 20 (5) (2008) 409–418.

[59] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. T. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. G. Stephan, J. V. den Bussche, The open provenance model core specification (v1.1), Future Gener. Comput. Syst. 27 (6) (2011) 743–756.

[60] J. Wang, D. Crawl, I. Altintas, Kepler + hadoop: a general architecture facilitating data-intensive applications in scientific workflow systems, in: E. Deelman, I. J. Taylor (Eds.), Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, WORKS 2009, November 16, 2009, Portland, Oregon, USA, ACM, 2009.

[61] A. Goderis, C. X. Brooks, I. Altintas, E. A. Lee, C. A. Goble, Heterogeneous composition of models of computation, Future Gener. Comp. Syst. 25 (5) (2009) 552–560.

[62] H. H. Ali, Y. Shi, D. Khazanchi, M. Lees, G. D. van Albada, J. Dongarra, P. M. A. Sloot (Eds.), Proceedings of the International Conference on Computational Science, ICCS 2012, Omaha, Nebraska, USA, 4-6 June, 2012, Vol. 9 of Procedia Computer Science, Elsevier, 2012.

[63] D. D. Pennington, D. Higgins, A. T. Peterson, M. B. Jones, B. Ludäscher, S. Bowers, Ecological niche modeling using the kepler workflow system, in: Workflows for e-Science, Springer, 2007, pp. 91–108.

[64] A. L. Hartman, S. Riddle, T. McPhillips, B. Ludäscher, J. A. Eisen, Introducing waters: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences, BMC bioinformatics 11 (1) (2010) 317.

[65] T. Stropp, T. M. McPhillips, B. Ludäscher, M. Bieda, Workflows for microarray data processing in the kepler environment, BMC Bioinformatics 13 (2012) 102.

[66] J. Cummings, A. Pankin, N. Podhosrzki, G. Park, S. Ku, R. Barreto, S. Klasky, C. Chang, H. Strauss, L. Sugiyama, et al., Plasma edge kinetic-mhd modeling in tokamaks using kepler workflow for code coupling, data management and visualization, Communications in Computational Physics 4 (3) (2008) 675–702.

[67] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, in: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

[68] S. Aiche, T. Sachsenberg, E. Kenar, M. Walzer, B. Wiswedel, T. Kristl, M. Boyles, A. Duschl, C. G. Huber, M. R. Berthold, K. Reinert, O. Kohlbacher, Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry, PROTEOMICS 15 (8).

[69] G. Nicola, M. R. Berthold, M. P. Hedrick, M. K. Gilsondoi:10.1093/database/bav087.

[70] T. Glatard, J. Montagnat, D. Lingrand, X. Pennec, Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR, International Journal of High Performance Computing Applications 22 (3) (2008) 347–360.

[71] T. Glatard, C. Lartizien, B. Gibaud, R. Ferreira Da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, N. Cerezo, P. Clarysse, A. Gaignard, P. Hugonnard, H. Liebgott, S. Marache, A. Marion, J. Montagnat, J. Tabary, D. Friboulet, A Virtual Imaging Platform for multi-modality medical image simulation, IEEE Transactions on Medical Imaging 32 (1) (2013) 110–118.

[72] J. Montagnat, B. Isnard, T. Glatard, K. Maheshwari, M. Blay-Fornarino, A data-driven workflow

language for grids based on array programming principles, in: Workshop on Workflows in Support of Large-Scale Science, ACM, Portland, USA, 2009, pp. 1–10.
[73] J. Rojas Balderrama, T. Truong Huu, J. Montagnat, Scalable and Resilient Workflow Executions on Production Distributed Computing Infrastructures, in: The 11th International Symposium on Parallel and Distributed Computing, IEEE Computer Society, Munich, Germany, 2012, pp. 119–126.
[74] A. Gaignard, J. Montagnat, B. Gibaud, G. Forestier, T. Glatard, Domain-specific summarisation of Life-Science e-experiments from provenance traces, Web Semantics: Science, Services and Agents on the World Wide Web 1 (2014) 17.
[75] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. Maechling, R. Mayani, W. Chen, R. F. da Silva, M. Livny, R. K. Wenger, Pegasus, a workflow management system for science automation, Future Gener. Comput. Syst. 46 (2015) 17–35.
[76] E. Deelman, G. Singh, M. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. C. Laity, J. C. Jacob, D. S. Katz, Pegasus: A framework for mapping complex scientific workflows onto distributed systems, Scientific Programming 13 (3) (2005) 219–237.
[77] E. Deelman, K. Vahi, M. Rynge, G. Juve, R. Mayani, R. F. da Silva, Pegasus in the cloud: Science automation through workflow technologies, IEEE Internet Computing 20 (1) (2016) 70–76.
[78] Pegasus team, Pegasus Powers LIGO Gravitational Wave Detection Analysis (2016).
URL https://pegasus.isi.edu/2016/02/11/pegasus-powers-ligo-gravitational-waves-detection-analysis/
[79] B. Abbott, R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, et al., Gw150914: First results from the search for binary black hole coalescence with advanced ligo, Physical Review D 93 (12) (2016) 122003.
[80] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, I. T. Foster, Swift: A language for distributed parallel scripting, Parallel Computing 37 (9) (2011) 633–652.
[81] T. G. Armstrong, J. M. Wozniak, M. Wilde, I. T. Foster, Compiler techniques for massively scalable implicit task parallelism, in: International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2014, New Orleans, LA, USA, November 16-21, 2014, 2014, pp. 299–310.
[82] J. Ozik, N. Collier, J. Wozniak, C. Macal, C. Cockrell, M. Stack, G. An, High performance model exploration of mutation patterns in an agent-based model of colorectal cancer.
[83] J. M. Wozniak, K. Chard, B. Blaiszik, R. Osborn, M. Wilde, I. Foster, Big data remote access interfaces for light source science, in: Big Data Computing (BDC), 2015 IEEE/ACM 2nd International Symposium on, IEEE, 2015, pp. 51–60.
[84] L. Wolf, et al., Boosting beamline performance (2014).
URL https://www.alcf.anl.gov/articles/boosting-beamlineperformance
[85] K. Williamson, Blue Waters supercomputer used to create 3D elevation models for White House Arctic Initiative.
URL http://www.ncsa.illinois.edu/news/story/blue_waters_supercomputer_used_to_create_3_d_elevation_models_for_white_house
[86] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, T. Oinn, Taverna: a tool for building and running workflows of services, Nucleic Acids Research (Web-Server-Issue) 729–732. doi:10.1093/nar/gkl320.
[87] T. M. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, R. M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, P. Li, Taverna: a tool for the composition and enactment of bioinformatics workflows, Bioinformatics 20 (17) (2004) 3045–3054.
[88] T. M. Oinn, R. M. Greenwood, M. Addis, M. N. Alpdemir, J. Ferris, K. Glover, C. A. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. W. Lord, M. R. Pocock, M. Senger, R. Stevens, A. Wipat, C. Wroe, Taverna: lessons in creating a workflow environment for the life sciences, Concurrency and Computation: Practice and Experience 18 (10) (2006) 1067–1100.
[89] K. Wolstencroft, R. Haines, D. Fellows, A. R. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. N. de la Hidalga, M. P. B. Vargas, S. Sufi, C. A. Goble, The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud, Nucleic Acids Research 41 (Webserver-Issue) (2013)

557–561.

[90] A. R. Hardisty, F. Bacall, N. Beard, M.-P. Balcázar-Vargas, B. Balech, Z. Barcza, S. J. Bourlat, R. Giovanni, Y. Jong, F. Leo, et al., Biovel: a virtual laboratory for data analysis and modelling in biodiversity science and ecology, BMC ecology 16 (1) (2016) 49.

[91] J. Ruiz, J. Garrido, J. Santander-Vela, S. Sánchez-Expósito, L. Verdes-Montenegro, Astrotavernabuilding workflows with virtual observatory services, Astronomy and Computing 7 (2014) 3–11.

[92] J. Zhao, C. Wroe, C. A. Goble, R. Stevens, D. Quan, R. M. Greenwood, Using semantic web technologies for representing e-science provenance, in: S. A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), The Semantic Web - ISWC 2004: Third International Semantic Web Conference,Hiroshima, Japan, November 7-11, 2004. Proceedings, Vol. 3298 of Lecture Notes in Computer Science, Springer, 2004, pp. 92–106.

[93] C. Wroe, R. Stevens, C. A. Goble, A. Roberts, R. M. Greenwood, A suite of daml+oil ontologies to describe bioinformatics web services and data, Int. J. Cooperative Inf. Syst. 12 (2) (2003) 197–224.

[94] D. de Roure, C. A. Goble, R. Stevens, The design and realisation of the $\text{my}_{\text{experiment}}$ virtual research environment for social sharing of workflows, Future Gener. Comput. Syst. 25 (5) (2009) 561–567.

[95] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. T. Michaelides, D. R. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, D. D. Roure, myExperiment: a repository and social network for the sharing of bioinformatics workflows, Nucleic Acids Research 38 (Web-Server-Issue) (2010) 677–682.

[96] D. Rogers, I. Harvey, T. T. Huu, K. Evans, T. Glatard, I. Kallel, I. Taylor, J. Montagnat, A. Jones, A. Harrison, Bundle and pool architecture for multi-language, robust, scalable workflow executions, Journal of grid computing 11 (3) (2013) 457–480.

[97] K. Vahi, I. Harvey, T. Samak, D. Gunter, K. Evans, D. Rogers, I. Taylor, M. Goode, F. Silva, E. Al-Shakarchi, et al., A case study into using common real-time workflow monitoring infrastructure for scientific workflows, Journal of grid computing 11 (3) (2013) 381–406.

[98] Y. Gil, P. A. González-Calero, J. Kim, J. Moody, V. Ratnakar, A semantic framework for automatic generation of computational workflows using distributed data and component catalogues, J. Exp. Theor. Artif. Intell. 23 (4) (2011) 389–467.

[99] Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce, A. Srivastava, P. Mallick, Towards continuous scientific data analysis and hypothesis evolution, in: S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., AAAI Press, 2017, pp. 4406–4414.