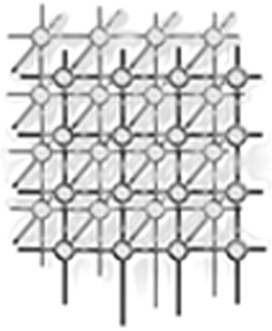
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2000; **00**:1–7 *Prepared using* cpeauth.cls *[Version: 2002/09/19 v2.02]*



Benchmarking Workflow Discovery: A Case Study From Bioinformatics

Antoon Goderis[1], Paul Fisher[1], Andrew Gibson[1,3], Franck Tanoh[1], Katy Wolstencroft[1], David De Roure[2], Carole Goble[1,*]

[1] *School of Computer Science*
The University of Manchester
Manchester M13 9PL, United Kingdom
[2] *School of Electronics and Computer Science*
University of Southampton
Southampton SO17 1BJ, United Kingdom
[3] *Swammerdam Institute for Life Sciences*
Universiteit van Amsterdam
Amsterdam, The Netherlands

SUMMARY

Automation in science is increasingly marked by the use of workflow technology. The *sharing* of workflows through repositories supports the verifiability, reproducibility and extensibility of computational experiments. However, the subsequent *discovery* of workflows remains a challenge, both from a sociological and technological viewpoint. Based on a survey with participants from 19 laboratories, we investigate current practices in workflow sharing, re-use and discovery amongst life scientists chiefly using the Taverna workflow management system. To address their perceived lack of effective workflow discovery tools, we go on to develop benchmarks for the evaluation of discovery tools, drawing on a series of practical exercises. We demonstrate the value of the benchmarks on two tools: one using graph matching, the other relying on text clustering.

KEY WORDS: Scientific Workflow, Bioinformatics, Discovery, Benchmark, Taverna, myExperiment

*Correspondence to: carole.goble@manchester.ac.uk

Received
Revised

1. Introduction

The process of scientific research has a crucial social element: it involves the sharing and publication of protocols and experimental procedures so that results can be reproduced and properly interpreted, and so that others may re-use, repurpose and extend protocols to support the advancement of science.

Scientific experiments conducted on the Web are increasingly being captured as Scientific Workflows, as workflow tools are adopted to exploit computational services and to deal systematically with the deluge of data generated by new experimental techniques [7]. An example of such a Scientific Workflow Management System is Taverna [19], which has been widely adopted across a range of disciplines and is particularly popular in the Life Sciences.

Mechanisms for publishing and sharing scientific workflows are beginning to emerge on the Web. For Taverna alone, we found more than 15 repositories, harboring over 700 workflows. The myExperiment sharing site [21] hosts over 500 workflows (`www.myexperiment.org`). However, it is not enough simply to publish workflows; faced with an increasing number of workflow systems and an increasing number of workflows, myExperiment users now ask for assistance in discovering them too.

As we will show later, the workflow literature contains multiple approaches to discovery which remain untested in practice. How many of these are relevant to scientific workflow systems? By understanding how scientists achieve workflow discovery, we can provide better support for effective finding over a growing body of workflows. This paper presents two main contributions:

1. *A study of re-use and discovery attitudes.* As a step towards achieving this understanding, we have worked with scientists to identify the prevalent attitudes to re-use and discovery. When do scientists care to share their workflows? Are they happy with the current discovery support? Under what conditions is re-use possible? Our user cohort is drawn from bioinformatics, a domain which makes very significant use of workflows. The study draws on a survey and a series of controlled exercises.

2. *Benchmarks established for workflow discovery tools.* To evaluate the effectiveness of discovery tools in solving scientists' problems, we have built a series of benchmarks. They consist of a set of representative discovery tasks and their solutions. The tasks were prepared by expert bioinformaticians using real-life Taverna workflows. The solutions were provided by a much larger group of bioinformaticians during controlled exercises. We assessed the practical value of the benchmarks by comparing the performance of two tools against them.

Our approach and the resulting benchmarks should be of interest to scientific workflow system developers wishing to test candidate discovery tools for their own system and domain. The survey, exercise materials, benchmarks and evaluation data are available on-line from `www.myexperiment.org/benchmarks`.

We proceed as follows. Section 2 defines workflow discovery. The results of a survey on workflow sharing, re-use and discovery are presented in Section 3. In Section 4 we contrast the state of workflow discovery tools in e-science with the diversity in approaches found in

the workflow literature. The gap motivates us to invest in the development of benchmarks in Section 5. We demonstrate them by comparing the performance of example tools. Section 6 relates our work to the literature, while Section 7 concludes and considers future work.

2. Defining workflow discovery

We introduce the following definition of **workflow discovery** and extend it to the case of scientific workflows below.

> Workflow discovery is the *process* of *retrieving orchestrations of services* to *satisfy user information need*.

Workflow discovery is a process Workflow discovery is a manual or automated process. Manual workflow discovery does not scale well for the individual faced with many workflows, but its observation potentially reveals problem-solving patterns that are useful to automated techniques. The benchmarks will capture the outcome of such problem solving behaviour.

Satisfy user information need Our target users are *scientists* looking for existing workflows that support their research. To be able to satisfy them, we need to document their information need and to evaluate how well retrieval techniques fulfill it.

In earlier work [27], we documented user information need based on several case studies of scientists recycling workflows created by others. We found it useful to draw a distinction between *workflow re-use*, where workflows and workflow fragments created by one scientist might be used as is, and the more sophisticated *workflow repurposing*, where they are used as a starting point by others.

- A user will **re-use** a workflow or workflow fragment that fits their purpose and could be customised with different parameter settings or data inputs to solve their particular scientific problem.
- A user will **repurpose** a workflow or workflow fragment by finding one that is *close enough* to be the basis of a new workflow for a different purpose and making small changes to its *structure* to fit it to its new purpose.

It is important to realise that the difference between supporting workflow re-use and repurposing leads to *different requirements for the discovery process*. Whereas re-use requires finding workflows that are similar to a given user query ("Find a workflow that produces protein sequence."), repurposing requires finding both similar workflows ("Find a workflow able to replace my faulty workflow fragment.") and complementary ones ("Find a workflow that extends my current annotation pipeline with a visualisation step."). We will develop benchmarks measuring how life scientists find both similar and complementary workflows.

Figure 1 provides an example of repurposing based on two Taverna dataflows. It shows the insertion of service **c** from Workflow 2 in between the previously connected services **a** and **b** of Workflow 1. In terms of the underlying bioinformatics, Workflow 1 is extended with the Transeq service, which changes the workflow from a pipeline for measuring similarity of DNA sequences into one that analyses similarity of peptide sequences. Observe how Workflow 1 provides useful input to locate Workflow 2 in a repository: one can concentrate the search on those service compositions that acccept service **a**'s output and produce service **b**'s input.

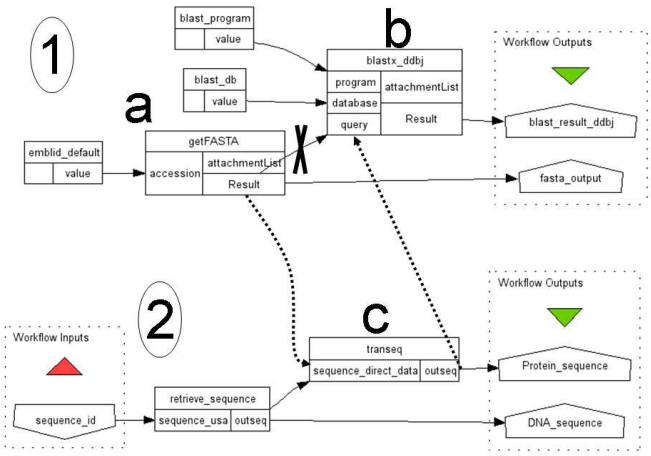


Figure 1. Example of an insertion based on two Taverna workflows

Finding compatible *insertions* is but one type of discovery that supports the repurposing of dataflows. The other types are the discovery of *replacements* and the discovery of *extensions* that append or prepend a workflow.

Orchestrations of services Multiple definitions of a workflow are in use in the scientific community [7] and in the business community [18]. What unifies these is the notion that a workflow orchestrates services. In terms of content, workflows vary on the following dimensions:

a) The kinds of domain and process they represent. For example, this paper draws computational workflows from the bioinformatics domain.
b) The granularity and type of services they orchestrate, *e.g.* local Beanshell scripts versus a Web Service that provides access to the National Grid Service in the U.K.
c) Their language, *e.g.* the BPEL standard or Scufl, Taverna's dataflow language [19]. A language relies on one or more models of computation to govern service interactions [9]. Our focus is on *dataflow models of computation*. Contrary to BPEL, dataflow has proven to be a popular paradigm with scientific workflows to capture data transformations [7].
d) The phase of the workflow lifecycle they reflect. The workflow lifecycle entails the following phases: (i) *during design*; (ii) *post design, pre-enactment*, as either a finished, concrete workflow where the required resources are known, or as a finished yet abstract workflow (also known as a template) whose resources will be decided during enactment; (iii) *during enactment*, when intermediary results come about and (iv) *post enactment*, when all results are available. Workflow discovery applies to representations capturing any of these phases.

Retrieving We limit the scope of the paper to the *retrieval of finished, concrete workflows based on existing workflows* ("finding workflows by example"). Elements of a *concrete* workflow can serve to find related concrete workflows: (i) a single service can act as a basis to retrieve relevant workflows; (ii) a selection of a subset of services, void of any control flow, can suffice;

and (iii) sometimes a workflow fragment or the complete concrete workflow will be relevant, including its control flow. Likewise, an *abstract* workflow could serve as input to find a cluster of related concrete workflows. Finally, we may also know the *evolution* of the design (or provenance) of the workflow, so that we can ask "Which workflows were derived from this one?" Our benchmarks only capture scenarios about the *retrieval of concrete workflows based on concrete workflows*. This is due to the practical constraint that abstract workflows and workflow evolution are not supported in the version of Taverna used (version 1.7).

3. A survey on workflow sharing, re-use and discovery practice

In order to better understand the practice and requirements of discovery we undertook a survey. The survey complements that of [10], which asked about the differences between workflow discovery and service discovery. Our survey provides insight on workflow sharing, re-use and discovery attitude among bioinformaticians closely familiar with workflows.

Participants and questionnaire From September to November 2007 a survey was published on `Keysurvey.com`. Twenty four bioinformaticians from 19 research laboratories participated. Of the 24 participants, 19 had built workflows before. Seventeen out of 19 were Taverna users. The survey was designed to document attitude towards workflow sharing, re-use and discovery in the world of bioinformatics, where services can be local, private and under control of the author (*e.g.* a local database of microarray results) as well as distributed, public and autonomous (*e.g.* the NCBI BLAST analysis service at `www.ncbi.nlm.nih.gov/blast`).

Insights into workflow sharing Survey participants were asked about their *reservations about sharing* their workflows for re-use by others. Two main concerns came forward:

1. Receiving proper *acknowledgements* for the work (36.8%) and
2. The workflow doing the job, but not being a piece of software they are proud of (31.6%).

Other factors were deemed less important:

- Being scooped (i.e. beat to obtaining results) by their own doing (15.8%);
- Sharing the data that is obtained from the workflows (15.8%);
- Sharing the data that feeds into the workflows (10.5%);
- The brittleness of shared workflows, either due to the use of non re-usable local services or due to the volatility of remote services (10.5%);
- Being able to share the workflow without others being able to see how it works (5.3%).

Our conclusion is that *participants are open to share quality workflows but want credit for doing so.* myExperiment caters for this attitude in part [21]. It is designed explicitly to provide users with proper acknowledgements for their work. Mechanisms for workflow attribution, rankings of popular downloads and a community-based star rating system are available. The scientist has fine control over visibility and sharing of workflows. Other mechanisms to ease the fear of attracting a reputation as a poor workflow builder, such as "work in progress categories" or anonymous publishing, are not provided at this time.

Insights into workflow re-use Polling participants about their concerns for workflow re-use, the following opinions surfaced:

- All respondents believed that in most cases there is not enough documentation to understand a workflow.

- For three quarters of respondents, some of the services in a workflow were (always or at least often) non-reusable due to the service being local to the original author. The same sentiment existed with respect to services being down.
- The majority of respondents believed there is no way of trusting the analysis performed by a workflow.
- Little under half of the respondents believed that often there are not enough workflows around, so they do not look for workflows.

Community-driven exchange platforms such as myExperiment could go a long way in meeting these concerns, through community-based annotation (description and tags), a repository of software code, workflow monitoring mechanisms and sharing best practice about building re-usable workflows.

One surprise finding, in light of the reported difficulty understanding workflows and trusting the analysis, is that of the 15 participants in the survey reporting re-use, seven had re-used workflows from third parties; other sources were fellow research group members (four mentions), project collaborators (four mentions) and a colleague at their institute (two mentions). The fact that *half of the re-users had adopted workflows from third parties* and not from people in their direct circle is an encouraging result for sites like myExperiment.

A second survey finding is that, again despite the difficulties understanding and trusting workflows, 15 out of the 19 workflow authors indicated having re-used workflows. The *high level of re-use activity* is remarkable. This may be due to the type of participant that volunteered to participate in the study – typically workflow enthusiasts and experts of Taverna.

Insights into workflow discovery *Ninety* percent of respondents believed there are *no effective search tools to find relevant workflows.* The most quoted current discovery mechanisms, in order of relevance, are: word of mouth, the myExperiment workflow sharing site and Google.

4. The state of the art in workflow discovery tools

Given the frustration of survey participants with the current support for workflow discovery, what kind of support exists within scientific workflow repositories and what specialised techniques exist in the workflow literature?

Discovery support in scientific workflow repositories We characterise the current situation in (finished concrete) workflow discovery in Web workflow repositories based on five academic systems (myExperiment, BioWep, INB, Sigenae and Kepler)[†] and two commercial ones (Inforsense and Pipeline Pilot).[‡] myExperiment, BioWep, INB and Sigenae offer Taverna workflows; the other systems their own type.

All provide basic discovery capabilities, by searching over workflow titles (Pipeline Pilot) or textual descriptions (myExperiment, BioWep, Sigenae, Kepler, Inforsense). Some systems

[†]Web sites: `www.myexperiment.org`, `bioinformatics.istge.it/biowep`, `www.inab.org/MOWServ`, `www.sigenae.org/index.php?id=84` and `library.kepler-project.org`
[‡]Web sites: `hub.inforsense.com` and `www.scitegic.com/products`

provide the possibility to search (Kepler, Inforsense) or browse (INB) semantic descriptions. All regard a workflow as an atomic entity, focussing on its overall inputs and outputs, and disregarding its internal structure. None of the systems supports finding workflows by example, *i.e.* finding similar or complementary workflows.

Specialised techniques for workflow discovery Multiple techniques exist to discover finished concrete workflows, of which we made a selection. For additional references see [8]. We know of only two techniques that report a user-based evaluation (covered in the Related Work section). The literature considers many workflow languages but each proposed technique limits itself to one language. Different data structures represent workflows, with graphs being a popular option, and different techniques work over these structures.

VisTrails [22] retrieves pipelines of Visualization Toolkit modules by example. The Chimera system [29] matches workflows available as Virtual Data Language specifications using untyped graphs. Corrales et al. [4] use error-correcting graph subisomorphism detection to match BPEL workflows. Bernstein and Klein [3] issue queries over an Entity Relationship diagram of the MIT Process Handbook. Kiefer [13] uses text similarity to retrieve Process Handbook entries modelled as Resource Description Framework graphs. Wroe and colleagues [26] use subsumption reasoning to detect whether the services in one Taverna workflow subsume the set of services in another. Mahleko and Wombacher [16] search over RosettaNet Partner Interface Processes using Finite State Machine matching.

Selected techniques from the Web service discovery and Web service composition literature are useful as they contribute methods for matching at either the atomic service level or at the workflow fragment level. Many techniques exist for Web service discovery, see *e.g.* [24] for a survey. A similar number of techniques exists for Web service composition, see *e.g.* [6]. Our interest is in those techniques capable of matching *groups* of services instead of those matching only two at a time. There are three types of approaches: those that assume the existence of a pre-defined template or framework, *e.g.* [14], those that compose unrelated sets of atomic components without relying on a given structure, *e.g.* [1] and those that combine both approaches, *e.g.* [17]. See [8] for details.

5. Benchmarking workflow discovery

The perception of survey participants that no effective discovery tools exist motivates experimentation with novel methods. We also established that virtually no work exists on evaluating workflow discovery techniques with end users. The work reported in this section helps address both issues. It aims to help evaluate the output of tools. Its outcomes could also be mined to identify promising metrics for a tool to then implement. It does *not* support measuring the responsiveness or scalability of a range of discovery tools. Rather, it documents practical tasks involving the discovery, re-use and repurposing of *finished concrete* workflows, where workflows are searched for by example.

The section starts by listing the key assumptions of the approach. The setup of the exercises is explained next, followed by a brief overview of participants, materials and procedure. The exercise results are then analysed. The resulting benchmarks are used to evaluate two tools. The section finishes with a discussion about the benchmarks' wider relevance.

5.1. Issues in benchmark-based evaluation

In designing the evaluation approach, we made three assumptions. The first is that discovery systems should attempt to discover those workflows in a corpus that are judged relevant by scientists to solve specific re-use and repurposing tasks. Such judgements are considered the *gold standard* for discovery systems. Reiter and Sripada [20] draw from their own experience in using corpora as gold standards to question the underlying assumption that a system should produce artifacts similar to those produced by humans for two main reasons. Firstly, human authors make mistakes, especially when they work hastily. Software systems should not imitate these mistakes. Secondly, there are substantial variations between individual people which reduces the effectiveness of corpus-based learning. To minimize the risks, we follow the line of Karamanis [12] and aim for a smaller corpus of high quality assessments instead of a large potentially problematic corpus. In addition, for some of the assessments we are able to verify their correctness technically.

The second assumption is that workflow discovery is inherently *task-driven*. In earlier work [10], we attempted in vain to capture universal metrics that bioinformaticians use to establish matches between workflows (*e.g.* the number of services shared *and* not shared between workflows). Such metrics could then be confidently incorporated in discovery tools to support a workflow by example discovery approach. In hindsight, we believe that the negative outcome was due to the fact that workflow discovery is inherently driven by a concrete information need. People are known to approach similarity based on multiple cognitive approaches [11]. Leaving the purpose for a similarity measurement unspecified leaves participants with many options to base similarity assessments on. We, instead, set practical tasks.

The final assumption is that *controlled exercises* are a good vehicle to build benchmarks. The alternative is to use empirical data from the daily search for workflows on sharing sites like myExperiment. The problem with log based approaches is that user information need is not explicitly captured.

5.2. Setup

We conducted three small-scale controlled exercises and one larger one. They rely on a corpus of real-world bioinformatics workflows, as generated by domain experts with the Taverna workbench. The exercises differ widely in their setup, reflecting different approaches for capturing how re-use occurs and different conditions under which re-use occurs. Table I provides an overview of the exercises. We discuss the details below.

5.2.1. Re-use and discovery tasks measured

The controlled exercises presented here fix the user information need by giving participants clear re-use and discovery goals. Exercises 2 and 3 aimed to record discovery behaviour, while exercises 1 and 4 also captured edit behaviour.

Exercise 1 asked for discovery of 20 workflows that do either a supertask or a subtask of a provided exemplar workflow, and to select relevant fragments in them.

Table I. Overview of the exercises.

	1	2	3	4
Setup				
Re-use subtasks	discovery, editing	discovery	discovery	discovery, editing
Discovery task	sub, super	earlier versions	sub, super, alternative	sub, super
Re-use directions	B2C	A2A	B2A	B2A, A2B, B2C
Participants				
Number	15	2	2	24
Expertise	++	++++	++++	+++
Materials				
Workflows	1+19	67+78	19+11	18
Documentation	++	++	++	++++
Procedure	embedded	independent	independent	independent
Results				
Average duration	20 min.	105 min.	30 min.	90 min.
Statistical support	-	N/A	+++	+++
Assessments	N/A	145	456	1848

Exercise 2 set the task as identifying all versions of a workflow authored by the participant within a wider set of his workflows.

Exercise 3 asked for boolean assessments of "usefulness." Useful was defined as meaning that one workflow either (i) provides an alternative to the other, (ii) provides an extension to it (supertask) or (iii) provides a useful fragment of it (subtask).

Exercise 4 presented participants with 12 practical repurposing tasks, involving workflow insertions, replacements and extensions. Example tasks included: "For the given workflow, connect/add a workflow fragment to generate genes in a region." and "Given this workflow which has a faulty service where indicated, fix the workflow to retrieve only a karyotype image." Given an exemplar workflow, a participant was asked to find and edit those workflows that would allow them to solve the task at hand. Each task presented up to five candidate workflows to choose from. The envisaged integration between the exemplar and candidate workflow was to be communicated through drawing. One example of such drawings is shown in Fig. 1.

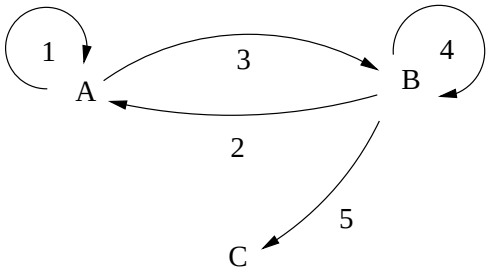


Figure 2. Types of workflow re-use from the perspective of the author of a set of workflows A.

5.2.2. Re-use directions

We took into account the fact that workflows are re-usable between several parties in several directions. It is important to make these distinctions because they influence the difficulty of a re-use task and the strategies to solve it, *i.e.* re-use of a familiar set of workflows is easier.

We distinguish between *personal re-use*, involving only workflows authored by the re-user, and *cross-author re-use*, involving other workflows. Assume the re-user has a number of workflow sources available: her own set of workflows **A**, a set of workflows **B** created by a second party and a set of workflows **C** created by a third party. Figure 2 summarizes the five possible ways a user can re-use her own as well as other people's workflows.

Table I summarises which exercises investigate which re-use directions. Interestingly, which direction is measured in a given re-use exercise is dependent on the combination of a particular re-use task and the participant in question. For example, the case where none of the workflows in the exercise are authored by a participant means either B to B ("B2B") and B to C ("B2C") is being measured. If the original author of one of the workflows were to solve that same task, one would be measuring A2A, A2B or B2A instead.

5.3. Participants

Between two and 24 bioinformaticians participated in any given exercise. In exercise 1 participants had no workflow experience. In exercises 2 and 3, we used the same two authors, both of whom had authored over 100 bioinformatics workflows. Exercise 4 drew on the survey participants, where 19 out of 24 had workflow experience.

5.4. Materials

Workflows were presented to participants on paper – A4-format for exercises 1-3, A1-format for exercise 4. Workflows were selected according to the re-use task to be measured. For details on the selection method see [8]. For exercises 1-3 the workflows stemmed from two authors. For exercise 4 workflows came from 12 authors. Conversely, the characteristics of the workflows influenced the setup of re-use tasks: for exercise 4 the annotation of only 18 workflows took two person months.

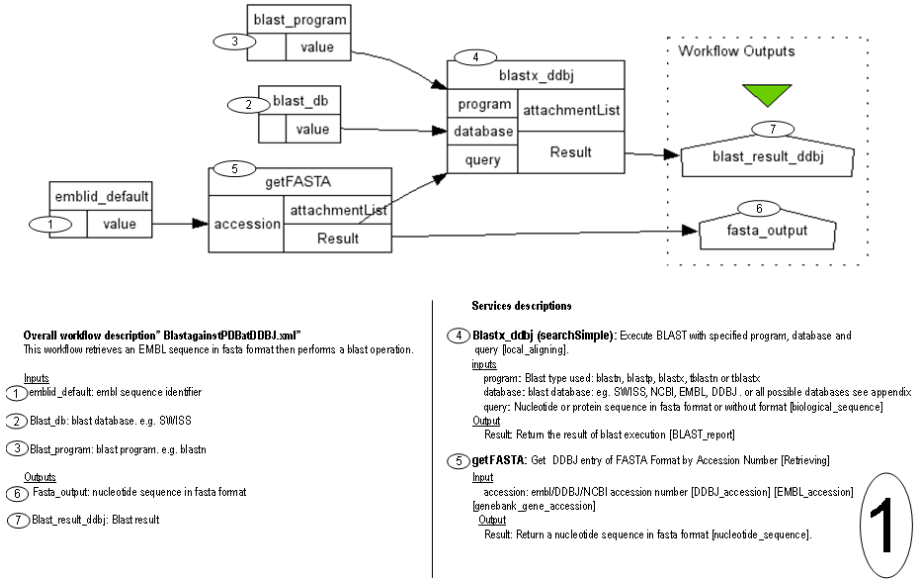


Figure 3. The annotated version of Workflow 1 in Figure 1.

The different exercises showed different amounts of workflow detail to users. In exercises 1 to 3, a workflow's name was shown as well as the orchestration of its services rendered as a diagram, showing only those inputs and outputs actively involved in the orchestration (as in Fig. 1). In exercise 4, more detail was provided with the inclusion of textual descriptions of the overall workflow task and of the services. In addition, semantic annotation was provided describing the task, inputs and outputs of the 98 services present in the 18 workflows, based on concepts selected from the ^{my}Grid bioinformatics service ontology (navigable at `www.mygrid.org.uk/ontology/OwlDoc/index.html`). The annotated version of Workflow 1 in Fig. 1 is shown in Fig. 3.

5.5. Procedure

Organisation-wise, the first exercise piggybacked on a training day for the Taverna workflow editor. Exercises 2 and 3 were run with local bioinformaticians. For exercise 4, the exercises were first double-checked in two pilot studies with two bioinformaticians, leading to extensive changes in the vocabulary used in the instructions, the curation and the re-use task descriptions. A third bioinformatician verified the validity of what consisted the correct solutions to the tasks, by creating and testing the corresponding workflows in Taverna for all tasks. A short feedback form was created. Thirty bundles of A1 posters were then printed and put in a tube and sent off to 19 research labs. Twenty four tubes were returned.

For exercises 3 and 4 two separate types of Kappa measures for inter-rater agreement were calculated [23]; see [8] for details. For exercise 4, per task data was gathered and analysed

Table II. Overview of the benchmarks

Benchmark	Exercise	Participants	Behaviour captured	Assessments	Agreement (Kappa)
PR2	2	2	Personal discovery	145	N/A
CA2	3	2	Cross-author discovery	456	Very good (0.678)
CA24	4	24	Cross-author repurposing	1848	Very good (0.666)

for (up to) seven variables: Difficulty, Confidence and Relevance of the (up to) 5 candidate workflows for solving the task.

5.6. Results analysis

The work of participants translated into a documented set of assessments. From their analysis we obtained figures about the content and validity of the benchmarks as well as insights into re-use and discovery practice that complement the survey.

5.6.1. From exercises to benchmarks

The outcome of an exercise was judged to be useful only when the results from the exercises showed a minimal level of agreement between participants and were confirmed by a bioinformatician as sensible. In exercise 1, the combination of inexperience with workflows and poor quality workflow descriptions resulted in unmotivated participants, who gave up rapidly on the task and yielded no useful answers. The three other user exercises had positive outcomes and produced benchmarks with different characteristics, listed in Table II. They are named after the type of re-use captured and the number of participants.

Benchmark **PR2** (exercise 2) collects similarity assessments made by a workflow author about pairs of his *own* workflows. In Benchmark **CA2** (exercise 3), a *collaborator* made similarity assessments on those same workflows. Benchmark **CA24** (exercise 4) contains the assessments made regarding the relevance of candidate workflows to solve tasks. In addition, the edit operations made by participants through drawings, document how integration between workflows is done, and can be formalised through graph edit operations [22].

5.6.2. Benchmark validity

The value of a benchmark depends on how well it represents the optimal performance of domain experts on representative tasks. To what extent can the exercise results be trusted?

In terms of the *methodological approach*, we took the following steps to ensure benchmark validity: (i) we designed the tasks based on real workflows and in collaboration with domain

experts, (ii) we cross-linked a participant's solution to participant confidence during the task and to inter-participant agreement on the solution and (iii) where possible, a participant solution was compared with a solution known to be correct.

In terms of the actual *results*, we found that all benchmarks were created by participants who felt confident while creating them.

For **CA2**, both participants felt confident during the exercise and although agreement was never perfect, they agreed strongly on the assessments made, as shown by the Kappa statistic for inter-rater agreement (see Table II).

The same is true for **CA24** (exercise 4). Analysis of ratings shows participants in general had high confidence and found the exercises to be of easy to moderate difficulty. Surprisingly, analysis of inter-rater agreement showed that they did not agree which exercises were easy, moderate or difficult. Similarly, they did not agree when they had high, medium or low confidence. An explanation for this apparent paradox is either that participants come from very different backgrounds and thus find different tasks challenging, or they use a different internal scale to assess confidence and difficulty. Their results on relevance assessments suggest the latter is true. Participants of **CA24** agreed on the relevance assessments made (Kappa value of 0.666).

For benchmark **CA24**, because known correct answers were field-tested beforehand, disagreement on relevance assessments could be measured in terms of technical correctness. Contrasting participant relevance assessments with the correct solution shows that they were correct in, on average, 83% to 91% of all cases, depending on the scheme adopted to assess a given answer. The schemes vary on whether they count a "maybe" answer as a correct one (leading to better scores) and whether blank answers count as negative answers (leading to better scores given that most candidate workflows are irrelevant to a task).

The main sources of error for participants in **CA24** were: (i) incomplete exercises because of a natural "*blind spot*" in the exercise material; (ii) incomplete or ambiguous *descriptions of data items*; (iii) assumptions made on the required *generality of a solution* across species and (iv) assumptions made on the *admissibility of additional "shim" or glue services* which were not available in the presented workflows.

5.6.3. Insights into workflow re-use and discovery practice

The successful outcome of exercises 2 and 3 suggests that *A2A and B2A type re-use* (see Fig. 2) is *feasible* when motivation is high and workflow documentation is at Fig. 1's low level.

Results from exercise 4 show that *B2C type re-use* of workflows is *feasible* when motivation is high and workflow documentation is at the high level of Fig. 3. We also analysed whether the amount of expertise building workflows or the time taken to complete an exercise correlates with the correctness of a solution. Neither factor proved to be a determinant. This indicates that people in general with a good bioinformatics background and very good documentation can muster the tasks of editing workflow diagrams and that some people simply work fast.

Exercise 4 also reveals that *relevance assessment and editing* are done in two *distinct phases*. During both phases, the workflow *diagram* is the first and most used point of recourse for finding information, despite its low detail and ambiguity. Textual workflow and service

inputs and outputs are also used eagerly, but less so than the diagram. The overall workflow description and workflow name are deemed useful for relevance assessment only.

5.7. Evaluation of the benchmarks

To evaluate the worth of the benchmark data in assessing real tools, we selected two techniques specifically developed for Taverna Scufl workflows (details of Taverna's language are in [19]).

5.7.1. Tool descriptions

The first tool is an existing graph matching based tool [10]. The second tool is new and consists of an adaptation of Woogle, a search engine for Web services [5].

Graph matching over Taverna workflows Graph matchers assume graphs of a certain kind as input; in the case of [10], they are attribute-less graphs of nodes and directed, attribute-less edges. To produce results, the graph matcher relies on sub-isomorphism detection over a graph repository. See [10] for details on the translation of workflows into graphs.

Text clustering over Taverna workflows Besides software specifications, workflows are also documents which contain natural language. Our adaptation of Woogle relies on its information retrieval based similarity search for Web services. To adapt the tool to workflows, we abstract a workflow to be a bag of services. Essentially a lossy translation of a workflow into the format of a Web service is established. A parser translates Scufl workflows into the Woogle WSDL service input format by regarding each workflow as a WSDL service and each constituent workflow service as a WSDL operation. The technique takes in a collection of Scufl workflows, clusters the available textual descriptions in an off-line step, and then, when given an input workflow, produces rankings of workflows from the collection.

In addition to the raw performance of these two techniques, we also consider the **combination hypothesis** – the idea that further advances in search technology will be based on a cross-disciplinary approach. In our context, we consider the impact of combining the results of the graph matching and text clustering techniques, either (i) using the *logical intersection* of results (when both techniques agree) or (ii) using their *logical union*.

5.7.2. Evaluation results

We tested the graph matcher and the text clustering tool on benchmarks **PR2** and **CA2**. Table III summarises the performance of the 2+2 techniques for personal re-use. It shows the average precision and recall for performing the two versioning tasks. The table also provides the average precision and recall for cross-author discovery on 11 tasks, where each task compares a different workflow with the other author's 19 workflows (*cfr.* Table I). The performance with respect to the top x results is shown (in percent).

The figures bring out the trade-off between precision and recall, in that an increase in precision means a decrease in recall. The different classes of discovery techniques come with their own strengths and weaknesses. The *text clustering* technique performs well on cross-author discovery, but does poorly when it comes to versioning. The *graph matcher* does well in comparison on the versioning task. When applying the graph matcher for cross-author

Table III. Average recall and precision on benchmarks **PR2** and **CA2**.

	PR2						**CA2**				
Measure	Top x results	Graph matcher	Text clustering	Intersection	Union		Top x results	Graph matcher	Text clustering	Intersection	Union
Precision	25	65	34	51	44						
Recall	25	50	24	17	57						
Precision	10	65	35	90	48		11	-	60	-	-
Recall	10	21	9	7	25		11	-	74	-	-
Precision	5	70	40	83	56		5	-	50	-	-
Recall	5	12	6	2	16		5	-	36	-	-

discovery, however, no results are returned in any of the cases. Inspection of results revealed its lack of a lexical component is to blame. As a result, the application of the combination hypothesis turns out to be sensible only in the case of versioning, where both techniques yield results. The *intersection* technique has good precision on the versioning task compared to the other techniques, but displays a drop in recall, whereas the *union* technique displays a converse pattern. The combination hypothesis idea does not improve the quality of search results overall in our exercise; one has to choose between either bettering precision or bettering recall. Comparing the results of the techniques and the experts, we found multiple matches which were only identified by the experts based on background knowledge of biology or bioinformatics. In this case, the addition of additional machine interpretable information, be it through full text descriptions or semantic annotation would help performance. In conclusion, by using the benchmarks we were able to establish that the tools do not approach what humans would achieve or might expect of a discovery system.

5.7.3. Relevance to other workflow systems

How useful are the benchmarks for other workflow systems? The developed suite of benchmarks captures re-use behaviour involving data flows. Translating the Scufl workflows into a canonical workflow representation would make the results more accessible. Unfortunately, there is no such representation (for a critique of for example XPDL or BPEL, see [10]). However, each scientific workflow system capable of modelling data flows should be able to re-model the Scufl workflows into its own language. It could then test its own discovery system with respect to the benchmarks. First experiences converting Scufl workflows to VisTrails VTK pipelines suggest that the exercise is achievable. The main difficulty lies in reconciling the different service parametrisation schemes (*e.g.* the way default values are assigned to services) and finding equivalent service types.

Another constraint is that the benchmarks are grounded in the bioinformatics domain. As long as the discovery systems under review are domain independent though, this does not prohibit a relative comparison of tools. For instance, the Provenance Challenge, an initiative to compare provenance management between workflow systems, picked a single domain [28].[§]

6. Related work

Software re-use and the associated problem of finding relevant components have been a theme in software engineering for a long time [15]. The sharing of workflows is a theme in (business) workflow management systems [18], but so far only basic search functionality has made it into practice. More advanced query facilities (allowing for control-flow based queries) can be found in enterprise architecture systems, which specialise in managing the business process repository for very large enterprises, *e.g.* the Mega tool[¶] comes with its own query language.

Neither the workflow literature nor the Web services literature contain many examples of user evaluation work. Fields with a tradition in using human-generated benchmarks to assess the performance of automated techniques include Information Extraction (*cfr.* the TREC competition series) and Natural Language Generation [20][12]. There is little work in the *workflow literature* on building human benchmarks. Recent work in the area has aimed to uncover the metrics people use for establishing workflow similarity. Bernstein and colleagues [2] look for the best semantic similarity measures to rank business processes from the MIT Process Handbook; the processes are not computational. The work of Wombacher [25] seeks to elicit the similarity metrics used by workflow researchers when performing the task of comparing the control flow complexity of workflows described by Finite State Machines (FSMs); data flow is left outside the scope. In the *service discovery literature*, most papers ignore how humans go about discovery and focus instead on a technical evaluation, demonstrating how expressive a technique is, or how scalable. An exception is the work by Dong et al. [5], who built a small human benchmark based on real Web services to test the performance of the Woogle tool. We know of two community initiatives to compare Web service discovery techniques: the Semantic Web Services Challenge and the Web Service Challenge.[‖] Both have limited involvement from users. In the former, a challenging scenario is put forward involving fully automated discovery and invocation. In the latter, techniques are evaluated by a subjective score issued by the organizers on the system design as well as on performance and accuracy.

7. Conclusion and future work

Workflows are proving successful in automating scientific experiments conducted on the Web. Public repositories are appearing to enable their re-use and repurposing into new experiments.

[§]Web site: `twiki.pasoa.ecs.soton.ac.uk/bin/view/Challenge`
[¶]Web site: `www.mega.com`
[‖]Web sites: `ws-challenge.org` and `www.sws-challenge.org`

We investigated current practices in workflow sharing, re-use and discovery amongst life scientists chiefly using the Taverna workflow management system. We found that sharing is possible once author credits and reputation are safeguarded. Under specific conditions, re-use happens even with poor documentation or between third parties. The perception is that no effective discovery tools exist. Given the range of untested workflow discovery tools in the literature, we developed user-driven benchmarks for evaluating discovery tools. We showcased the benchmarks on two Taverna-based tools.

We hope that our work will stimulate the adoption of existing tools from the workflow community and enable their comparison. Additional benchmarks can be devised, for example to test tool scalability or to record how contextual workflow information such as authorship and popularity drives discovery. In this respect, myExperiment provides a promising resource.

Acknowledgment

The authors are grateful to the Taverna training days participants, the myGrid team, the myExperiment team and Luna Dong for the Woogle source code. We also thank the participants of the survey and Exercise 4 for their time and effort: Anonymous (LLNL, USA), Adam Barker (NeSC, UK), Anika Joecker (MPI for Plant Breeding Research, Germany), Arnaud Kerhornou, Peter Rice (EBI, UK), Bela Tiwari, Tim Booth (NEBC, UK), Ben Mahy (VIB, Belgium), Benjamin Good, Mark Wilkinson (University of British Columbia, Canada), Mike Cornell, Cornelia Hedeler, Duncan Hull, Helen Hulme, Peter Li (University of Manchester, UK), Francois Moreews (INRA, France), Giovanni Dall'Olio (IMIM, Spain), Hannah Tipney (UCHSC, USA), Lin Ching - Fong, Wu I-Ching (National Yang-Ming University, Taiwan), Marco Roos (Universiteit van Amsterdam, The Netherlands), Pieter Neerincx (Wageningen Universiteit, The Netherlands), Nathan Nicely (RENCI, USA) and Paolo Romano (ISTGE, Italy).

REFERENCES

1. Daniela Berardi, Giuseppe De Giacomo, Maurizio Lenzerini, Massimo Mecella, and Diego Calvanese. Synthesis of underspecified composite e-services based on automated reasoning. In *2nd International Conference on Service Oriented Computing ICSOC*, pages 105–114. ACM Press, 2004.
2. A. Bernstein, E. Kaufmann, C. Brki, and M. Klein. How similar is it? Towards personalized similarity measures in ontologies. In 7 *Internationale Tagung Wirtschaftsinformatik*, February 2005.
3. Abraham Bernstein and Mark Klein. Towards high-precision service retrieval. In *Proceedings of the First International Semantic Web Conference (ISWC)*, Sardinia, Italy, 2002. Springer.
4. J. C. Corrales, D. Grigori, and M. Bouzeghoub. BPEL Processes Matchmaking for Service Discovery. In *Conference on Cooperative Information Systems (COOPIS)*, LNCS 4275, pages 237–254, Montpellier, France, 2006.
5. X. Dong, A. Halevy, J. Madhavan, E. Nemes, and J. Zhang. Similarity search for web services. In *Proc. of the 30th VLDB Conference*, Toronto, Canada, 2004.
6. Schahram Dustdar and Wolfgang Schreiner. A survey on Web services composition. *Int. J. Web and Grid Services*, 1(1), 2005.
7. Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, December 2007.

8. Antoon Goderis. *Workflow re-use and discovery in bioinformatics*. PhD thesis, School of Computer Science, The University of Manchester, 2008.
9. Antoon Goderis, Christopher Brooks, Ilkay Altintas, Edward A. Lee, and Carole Goble. Heterogeneous Composition of Models of Computation. *Future Generation Computer Systems (FGCS)*, Accepted for publication.
10. Antoon Goderis, Peter Li, and Carole Goble. Workflow discovery: requirements from e-science and a graph-based solution. *International Journal of Web Services Research (JWSR)*, 5(4), 2008.
11. R. L. Goldstone. *MIT encyclopedia of the cognitive sciences*, chapter Similarity, pages 757–759. MIT Press, Cambridge, MA, 2001.
12. Nikiforos Karamanis. *Entity Coherence for Descriptive Text Structuring*. Phd thesis, School of Informatics, University of Edinburgh, 2003.
13. Christoph Kiefer, Abraham Bernstein, Hong Joo Lee, Mark Klein, and Markus Stocker. Semantic process retrieval with iSPARQL. In *European Semantic Web Conference (ESWC)*, pages 609–623, 2007.
14. J. Kim, Y. Gil, and V. Ratnakar. Semantic metadata generation for large scientific workflows. In *Int. Semantic Web Conference (ISWC)*, Athens, USA, November 5-9 2006.
15. Charles W. Krueger. Software reuse. *ACM Comput. Surv.*, 24(2), 1992.
16. Bendick Mahleko and Andreas Wombacher. Indexing business processes based on annotated finite state automata. In *ICWS*, pages 303–311, 2006.
17. Brahim Medjahed, Athman Bouguettaya, and Ahmed K. Elmagarmid. Composing Web services on the Semantic Web. *VLDB J.*, 12(4), 2003.
18. D. Miers, P. Harmon, and C. Hall. The 2007 BPM suites report. http://www.bptrends.com.
19. Tom Oinn, Mark Greenwood, Matthew Addis, Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew Pocock, Martin Senger, Robert Stevens, Anil Wipat, and Chris Wroe. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience: Special Issue on Scientific Workflows*, 18(10):1067–1100, 2005.
20. E. Reiter and S. Sripada. Should corpora texts be gold standards for NLG? In *INLG*, pages 97–104, New York, USA, 2002. Harriman.
21. David De Roure, Carole Goble, and Robert Stevens. Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows. In *Third IEEE International Conference on e-Science and Grid Computing*, pages 603–610, Bangalore, India, December 10-13 2007.
22. Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, and Claudio T. Silva. Querying and creating visualizations by analogy. *IEEE Trans. Vis. Comp. Graph.*, 13(6):1560–1567, 2007.
23. S. Siegel and J. N. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.
24. Ioan Toma, Kashif Iqbal, Dumitru Roman, Thomas Strang, Dieter Fensel, Brahmananda Sapkota, Matthew Moran, and Juan Miguel Gomez. Discovery in Grid and Web services environments: A survey and evaluation. *Multiagent and Grid Systems Special Issue on Advances in Grid services Engineering and Management*, 3(3):341–352, 2007.
25. A. Wombacher. Evaluation of technical measures for workflow similarity based on a pilot study. In *CoopIS*, Montpellier, France, November 1-3 2006.
26. C. Wroe, R. Stevens, C. Goble, A. Roberts, and M. Greenwood. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. *Intl. J. of Cooperative Information Systems*, 12(2):197–224, 2003.
27. Chris Wroe, Carole Goble, Antoon Goderis, Phillip Lord, Simon Miles, Juri Papay, Pinar Alper, and Luc Moreau. Recycling workflows and services through discovery and reuse. *Concurrency and Computation: Practice and Experience*, 19(2):181–194, 2007.
28. Jun Zhao, Carole Goble, Robert Stevens, and Daniele Turi. Mining Taverna's semantic web of provenance. *Concurrency and Computation: Practice and Experience*, 20(5):463–472, 2008.
29. Y. Zhao, M. Wilde, and I. Foster. Applying the virtual data provenance model. In *Int. Provenance and Annotation Workshop (IPAW)*, Chicago, USA, May 3-5 2006.