Workflow Reuse in Practice:
A Study of Neuroimaging Pipeline Users

Daniel Garijo and Oscar Corcho
Dpto. Inteligencia Artificial,
Facultad de Informática
Universidad Politécnica
de Madrid
{dgarijo, ocorcho}@fi.upm.es

Yolanda Gil
Information Sciences Institute
and Department of Computer
Science
University of Southern
California
gil@isi.edu

Meredith N. Braskie, Derrek Hibar,
Xue Hua, Neda Jahanshad, Paul
Thompson, and Arthur W. Toga
Laboratory of Neuro Imaging
Institute for Neuroimaging and
Informatics
Keck School of Medicine
University of Southern California

Abstract—**Workflow reuse is a major benefit of workflow systems and shared workflow repositories, but there are barely any studies that quantify the degree of reuse of workflows or the practical barriers that may stand in the way of successful reuse. In our own work, we hypothesize that defining workflow fragments improves reuse, since end-to-end workflows may be very specific and only partially reusable by others. This paper reports on a study of the current use of workflows and workflow fragments in labs that use the LONI Pipeline, a popular workflow system used mainly for neuroimaging research that enables users to define and reuse workflow fragments. We present an overview of the benefits of workflows and workflow fragments reported by users in informal discussions. We also report on a survey of researchers in a lab that has the LONI Pipeline installed, asking them about their experiences with reuse of workflow fragments and the actual benefits they perceive. This leads to quantifiable indicators of the reuse of workflows and workflow fragments in practice. Finally, we discuss barriers to further adoption of workflow fragments and workflow reuse that motivate further work.**

Keywords— scientific workflows; workflow fragments; workflow reuse; LONI Pipeline

I. Introduction

Workflows have many benefits to scientists managing complex data analysis [8] [7] [9] [20]. They make it easier to reuse expert-grade methods and the software that implements them, helping newcomers understand complex multi-step data analysis methods, and can track provenance and facilitate reproducibility. Workflow reuse is often cited as a major benefit of workflows, and has been studied in repositories of workflows [19]. However, there are no studies on the level of reuse of workflows in practice in research laboratories. We are also particularly interested in whether workflow fragments are more reusable than entire workflows [5].

This paper reports on a study on workflow reuse in labs that use a particular workflow system, the LONI Pipeline [3] [4]. The LONI Pipeline includes facilities for users to define subsets of workflows as "groupings" that may be reused by themselves and with others in new workflows. The community of the LONI Pipeline users provides a unique opportunity to study how workflow fragments are used in practice, whether they improve reuse, and the barriers that users find in reusing workflow fragments.

The main contributions of this paper are twofold. First, it articulates the benefits of workflows and workflow fragments reported by users in a neuroscience research lab. Although many of these benefits such as reuse and time savings have been discussed in the context of workflows, others are not commonly highlighted, such as promoting standards, facilitating debugging, and teaching newcomers to the lab. The second contribution of the paper is a survey of workflow users which provides a useful quantitative perspective on the relative importance to them of the benefits that we had identified. This leads us to identify and prioritize areas of research in workflow frameworks.

After discussing related work, we give an overview of the benefits of workflows and workflow fragments reported by users in informal discussions. We report on a survey of researchers in a lab that has the LONI Pipeline installed, which helps us quantify the relative adoption of workflows and the actual benefits of using workflow fragments. Finally, we discuss further work on workflow fragment detection to promote reuse motivated by this work.

II. Related Work

There have been reports of requirements on workflow reuse [1] [11]. Other work discusses technical bottlenecks for workflow reuse [10] and common practices and barriers to software reuse in general [12]. In this paper we discuss workflow reuse through the benefits for the authors of the workflows, instead of analyzing the technical difficulties that they might encounter when trying to reuse someone else's work.

There are shared repositories of workflows to promote sharing and reuse [2] [14], as well as standards and extensions for sharing workflows across workflow systems [15] [16] [6]. In our own work, we are investigating whether defining workflow fragments can improve reuse, since end-to-end workflows as a whole are too specific to be applied to new projects [5].

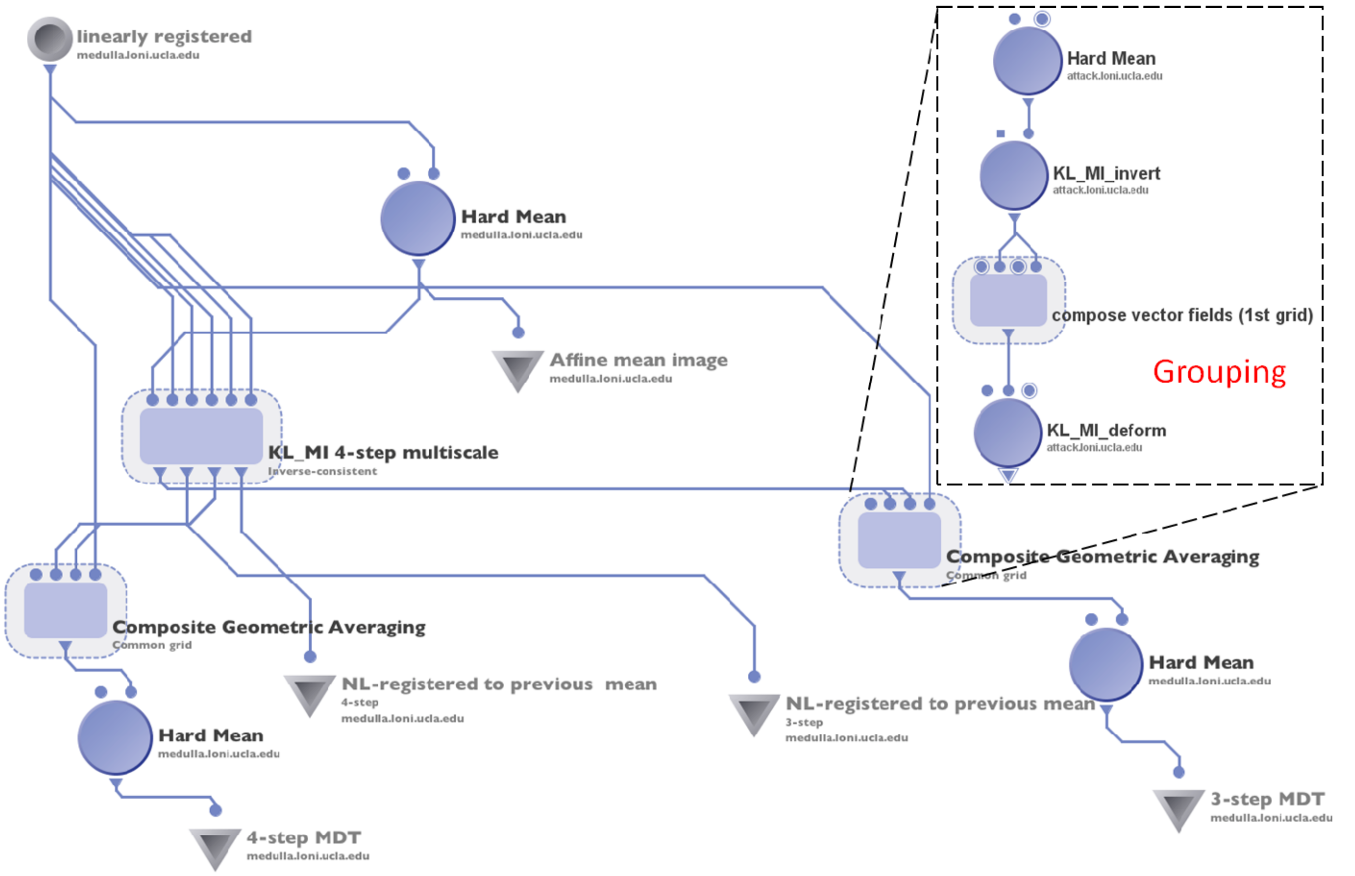


Figure 1. An example of a workflow in the LONI Pipeline, with workflow steps (components) shown as circles. Outputs are shown as triangles while the input (linearly registered) is a smaller circle. The connections among steps represent the dataflow. Users can select subworkflows to create "groupings" of components (shown with dashed lines), which can be reused in the same workflow and in others (shown as rectangular components).

Several approaches have been developed to facilitate workflow reuse through workflow matching [11] [1] and workflow completion [13]. The LONI Pipeline is another example, which we introduce briefly in the next section.

Finally, in [19] the authors present a statistic analysis on how workflows, subworkflow and steps are reused in the myExperiment public repository [2] by different authors. However, they do not study workflow reuse from the point of view of the scientists, and how they perceive the utility of workflows.

III. The LONI Pipeline

The LONI Pipeline is a workflow system developed by the Laboratory of Neuro Imaging (LONI) mainly for neuroimaging applications [3] [4]. It provides an efficient distributed computing solution to address common challenges in neuroimaging research, enabling investigators to share, integrate, collaborate and expand resources including data, computing platforms, and analytic algorithms. Using its graphical interface, users can connect components that implement algorithms. The LONI Pipeline is mostly used for complex neuroimaging analysis, which often requires knowledge about the input/output requirements of algorithms, data format conversions, optimal parameter settings, and a unique running environment since imaging studies tend to produce large amounts of data.

We are particularly interested in the LONI Pipeline because it includes some capabilities for defining and reusing workflow fragments. These capabilities are:

- **Grouping Tools**: Grouping Tools allow users to define a "group" of components in a workflow, which they can copy/paste in different workflows. Although they have been adopted by many users, these tools have still very basic functionality. For example, new changes to a grouping are not propagated through the workflows where the grouping was pasted in the past.
- **Workflow Miner**: The Workflow Miner[1] allows users to browse the dependencies among different workflow components based on their use in different workflows. It uses a probability network to detect workflow fragments and displays to the user how those workflow fragments appear in different workflows.

Figure 1 shows an example of a workflow with groupings defined by scientists using the LONI Pipeline. The figure shows a minimal deformation target (MDT) pipeline to serve

[1] http://pipeline.bmap.ucla.edu/products-services/workflow-miner/

as an unbiased average brain template for neuroimaging studies.

The fact that the LONI Pipeline includes tools to create and view workflow groupings is an indication that users and/or developers have found a need for workflow fragments. We set out to understand the current level of adoption, the perceived benefits, and the barriers regarding reuse of workflow fragments.

IV. Benefits of Reuse of Worfklows and Workflow Fragments

We conducted several discussions with a small group of scientists to understand what their motivations to use the workflow system were. This section presents the perceived benefits of workflows and workflow fragments, including both current and potential benefits. These benefits drove the design of our survey with a large number of participants that we report below.

1) Sharing Workflows with Collaborators

Workflows are shared often among lab researchers. Workflow fragments are also shared, but to a lesser extent.

Non-programmers find a barrier to running complex neuroimaging analyses as they cannot create components or code to that level of complexity. Reusing workflows that others have created enable them to do tasks that they would not otherwise do.

Personal documents are sometimes used to annotate how workflows and components are used instead of including this documentation in the workflow system, which provides facilities for doing so.

2) Time Savings

Individual users save time when they define workflows, as the software for each step is well encapsulated in a workflow component that has clear inputs and outputs and can be run independently, and similar experiments can be repeated with minimal efforts. A lot of time is saved by being able to copy and paste a subworkflow into a new workflow being created. Other users save time as well when they reuse a workflow created by someone else, since they do not have to re-implement or re-install the codes. Workflow fragments can also save time in similar ways, and have an additional feature of being easier to find based on their frequency [5].

The alternative to reusing workflows is sharing documents with “protocols”, which are extremely detailed instructions about how to run end-to-end analysis. This is one approach adopted by the ENIGMA Consortium [21].

3) Teaching

Neuroimaging concepts, such as observing brain differences in disease or tracking changes in brain structure throughout development, are easily understood among students. However, the various steps involved in image processing are not always intuitive. Critical steps can be accidentally left out or reordered, and wrong inputs may be used for some points. In the best case scenarios, the mistakes will be obvious and quality control of the workflow results will allow students to see where something went wrong. However, in some cases, forgotten steps may not manifest themselves clearly in the final product. Pipelines can be used as an effective way to teach students about the workflow and the sequence of steps involved for processing. Breakpoints are often placed throughout the pipeline to serve as checkpoints and make sure that execution was performed correctly. These breaks in execution as opposed to an extended workflow allow for novices to learn the expected output of each step so that they too may help teach future generations.

4) Visualization

The ability to use a visual interface to manage the many steps involved in an analysis was considered important. It is easier to track how the overall method is structured, as well as the algorithms used in each step.

In the case of the LONI Pipeline, users specify workflows using a visual interface. Functions receive input, perform some task, and then output data. They are represented as big circles. Required input sources for a given function are represented with smaller circles directly above the function circle. Similarly, function outputs are represented as upside-down triangles directly below the function circle. Workflow inputs/outputs are connected with solid lines and upside-down triangles showing the direction of flow. Beyond the basic interface, workflows in the LONI Pipeline have a hierarchical organization. Users can select functions in a workflow into a grouping, which gets condensed at higher level into a single function circle. When the user double-clicks on the grouping the workflow expands to show each of the functions within it, similar to a file browser. The hierarchical organization can be used to group functionally related tasks into a single visual element. This allows workflow developers to group complex tasks with highly-fragmented code into a single visual unit that other users can incorporate into their workflows.

5) Design for Modularity

Defining workflow components makes the scientists more aware of the need to design their code in a modular manner. Workflows provide a high-level view of the major steps involved in an analysis, and exposing those major steps drives the design of the code in a modular fashion.

6) Design for Understandability

Workflows may be organized in many ways. Standard scripts are written in an enumerated format, listing out steps sequentially. In contrast, visualizing the organization of the workflow allows users to know what steps are prerequisites to future steps. For example, intensity normalization of images may be a step that a few completely different analyses have in common. To be able to perform a step and visualize that the analysis can branch off into one of 3 ways, allows users to understand where data processing may have gone wrong and help speed up backtracking and debugging. Otherwise, an unfamiliar value (following the example in Figure 1) may be at the “KL_MI_deform” step of the written protocol when users realize something is wrong. The workflow can let them know how to debug, and the best way to backtrack. They can easily tell that they do not need to check the files for the last "Hard Mean".

This level of understanding also improves the efficiency of collaborations. If all users need to perform the "KL_MI 4-step multiscale" for a variety of works, that will only need to be run once across a large group, and outputs can be used as input for other modules and fragments of workflows. One user can run steps "Hard Mean", and another one can use the outputs of that and "KL_MI 4-step multiscale" to run the full workflow described in Figure 1. The ability of these workflows to be broken down into fragments allows users to re-use inputs and outputs of previous workflows for more advanced levels of analysis. Workflows are designed with re-use in mind. Standard processing tools and image formats are critical to such workflows and allow for the exchangeability of inputs, from various image datasets.

7) Design for Standardization

Scientists reported that in creating workflows and intending to share them, they found themselves adopting the practices that others adopted in the lab. Before the workflow system was adopted, different researchers used different platforms and it was difficult to combine different codes. Using a common workflow system allows researchers to see how others process certain kinds of data, what software packages they use, and what formats are more common in the lab. This leads to workflows that effectively capture emerging standards in the ways that data is formatted and processed, based on common practices adopted in the lab.

This is particularly useful for newcomers. In the past, they had to struggle with many formats, code bases, and platforms. Now there are fewer things to learn, and what is used is more compatible and easier to learn and to integrate.

8) Debugging

A workflow execution might fail due to incorrect setup, problems in the underlying code, missing files, incompatible file types, or server-related issues.

Programmers use the workflow system's environment to debug errors in the workflow. A log viewer displays execution information, including server information, command string that was submitted to the server, output stream, error stream, and output files. This unified system allows easy reporting of server related issues and debugging issues with the pipeline setup. The entire pipeline can be submitted to the pipeline support team or to an expert for evaluation as the pipeline captures the entire processing workflow as well as input/output specifications.

For non-programmers, debugging is more challenging. When there is a failure, they cannot easily tell whether the error resides in the data or in the code, or if there is a failure on the server (e.g., a failed node). The workflow system allows users to report the unique ID of their workflow run, so they can request help with an execution failure.

9) Paper Writing

In addition to discussions, we did a preliminary analysis of ten articles by the group. There were clear commonalities in the "methods" section of the papers, indicating room for "groupings" and reuse across different projects in the lab. Further work is required to determine how the papers correspond with the pipelines that are created by the lab.

A single paper typically includes several pipelines. This might indicate that the pipelines themselves are fragments of a larger workflow that would represent the analysis reported in a paper.

Linking papers to pipelines was acknowledged as desirable by scientists, but this is not a common practice. Some papers include an "implementation section" that cites the pipelines and describes them.

10) Reproducibility and Inspectability

Reproducibility has been recognized as an important concern in science [17] [18], including in neuroimaging. Reproducibility in neuroimaging studies may be difficult to achieve between laboratories as journal space constraints may limit the ability of researchers to report the occurrence and ordering of complex analysis steps with sufficient detail to allow a new user to execute the analysis in exactly the same way. Deficits in the ability to reproduce analyses using new data add variability to results among labs, making interpretability of results more difficult.

The workflow framework allows users to track what they executed and record provenance for new results. In addition, it allows them to inspect what others have done to check whether any errors were made or there is anything unusual with intermediate results that might indicate problems in the pipeline setup or the data pre-processing. These kinds of checks and inspections are particularly important when a new person joins the lab and runs workflows so that correct use of methods and data is enforced.

V. Types of Users

We identified three major categories of users in the workflow system:

1. Developers: These users write code and componentize it. Their codes use sophisticated algorithms for image processing that have to be written to slice and dice the datasets into layers, tiles, voxels, and pixels, and with efficiency in mind. They create workflows and use them to run analyses themselves. They also share their codes and workflows. They are typically bioinformaticians and engineers.

2. Beginner programmers: These users can write small scripts and also program spreadsheets. This allows them to do some minor data reformatting and preparation so that their data fits a workflow that they want to run. They reuse workflows that others have created. They are typically neuroscientists.

3. Non-programmers: These users cannot write any code. They reuse the workflows that others create. They are typically students.

We found that in order to use a workflow framework for carrying out research it is important to have at least some basic programming skills. Otherwise, it is hard to reuse workflows previously defined by others.

VI. User Survey Report

To check whether other users would agree or disagree with the benefits above, we created a survey and sent it to the mailing list of users of the LONI Pipeline. We included users who use the LONI Pipeline system installed at USC, but did not include many other users that have downloaded the system and run it themselves in their own servers. The survey was conducted on-line, and the responses were anonymous.

The survey included two kinds of questions. Some questions presented a choice of answers using a five-level Likert scale. For example, for the question "Is reusing workflows from others useful?" we offered five answers: very often, often, sometimes, occasionally, and never. Other questions offered a list of possible answers and allowed users to provide their own answers. For example, for the question "Why is reusing previously created workflows useful?" the list of possible answers included "It saves time", "Workflows give a high-level diagram that helps remember what was done", and "Other". If the latter was chosen, respondents could provide text with their own reasons. Respondents could do more than one selection.

We received 21 responses. We discuss the results of the survey below, highlighting in boldface our findings.

Writing and Sharing Code

We wanted to have some reference for comparing the responses about workflow sharing, so the survey included some questions about code sharing.

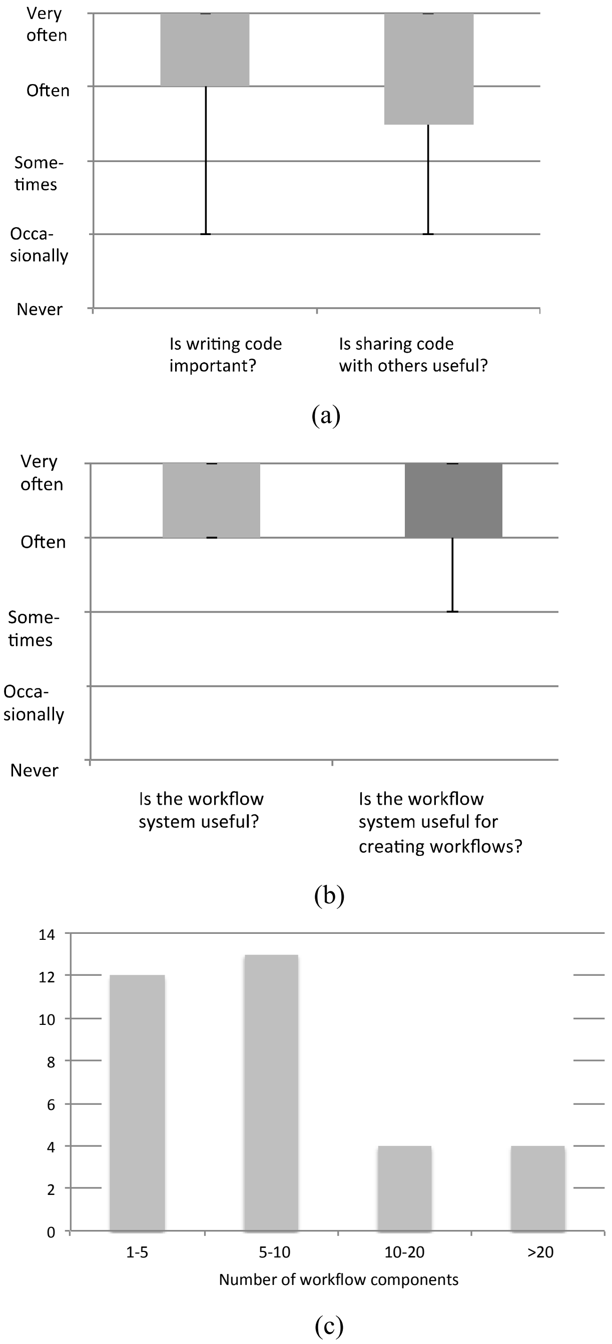
Figure 2(a) shows responses regarding the importance of writing code and reusing code. **Writing code is considered very important for this area of research. Sharing code is not considered to be as important.** These answers imply that the responders are well aware of the importance and value of their software.

Adopting a Workflow System

Figure 2(b) shows responses regarding the workflow system basic utility in creating workflows for their work. **The overwhelming majority of responders found the workflow system useful.** This perhaps reflects a self-selection bias of the user population that responded, but is nevertheless useful to put in perspective the survey responses and the conclusions of this study.

Figure 2(c) shows the most usual sizes of workflows according to the respondents. Workflows of fewer than 10 steps seem to be the most common.

We asked for the reasons not to use the workflow system. We assumed that even users of the workflow system may not use it for all their analyses. We offered one choice and then free text answers. Two respondents selected the given choice of "It takes time to learn to create workflows". Free-form answers included "Minor changes to underlying scripts or tuning of parameters may require more work than just editing scripts themselves," and "Sometimes it is easier to run a certain command in loops or batches or to edit the various input/output parameters (file names, paths, options, etc) on the command line, rather than clicking through the workflow GUI."



(a)

(b)

(c)

Figure 2. Survey results concerning (a) the utility of writing and sharing code, (b) the utility of creating workflows, (c) the size of workflows created. The distribution of the responses of the first two figures is presented as a "box and whisker" diagram. The whiskers represent the range of the responses. The dark grey box represents the distribution on the second quartile (Q1-Q2), while the light grey box represents the distribution on the third quartile (Q2-Q3). The median is represented by the bottom line of the dark grey box (or the top line of the light grey box).

Overall, all respondents seem to find utility in the workflow system.

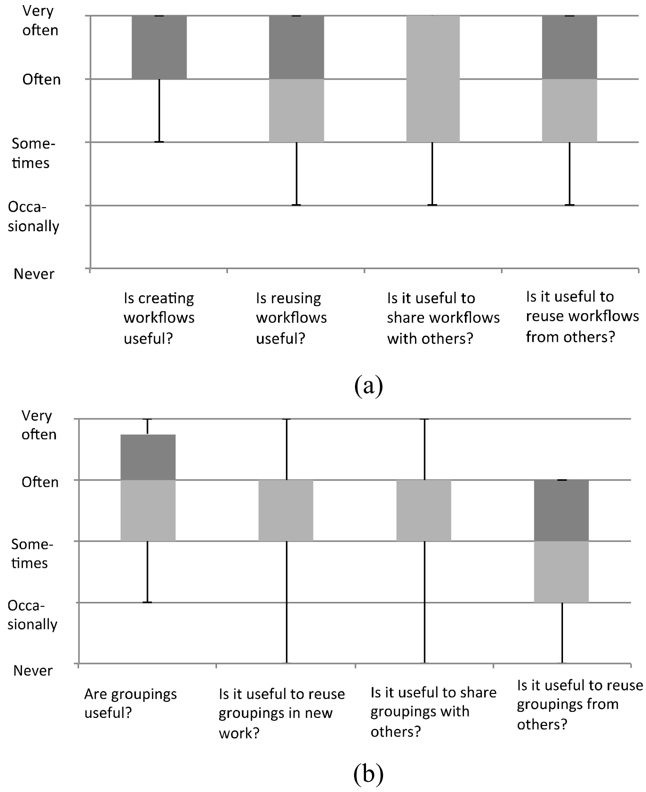


Figure 3. Survey results concerning (a) the utility of sharing and reusing workflows, and (b) the utility of sharing and reusing groupings.

Using Workflows

The survey included questions about reuse of workflows and about reuse of groupings. We discuss the results for each.

Figure 3(a) shows the survey answers regarding the utility of creating and sharing workflows. **Respondents responded overwhelmingly that creating workflows is very useful, but the reuse of workflows was seen as less useful.** Therefore, reuse is not the only reason why workflows are created. **Reusing workflows from a user's prior work is considered as useful as reusing workflows from others**.

When asked "Why is creating workflows useful?", respondents were given the choices shown in Table 1(a). The number of respondents that selected each choice is also shown in that table. **The benefits of workflows that the majority of respondents agreed with include time savings, organizing and storing code, having a visualization of the overall analysis, and facilitating reproducibility. Many respondents agreed to other benefits that included debugging complex code, and encouraging the adoption of standard ways to do things.** Free-form responses included: "Workflows are mainly used for population studies so that you can run many subjects in the same time, and it is easy to pass around to someone who doesn't know how to code," "The main reason is that it is easy to send a prepared pipeline to another researcher and they can usually figure out how to use it, regardless of their programming knowledge," "It's a really intuitive visualization of the underlying code. Sort of brings the code 'to life'!" and "Parallelizing without having to use the Sun Grid Engine script."

Table 1(b) shows responses for the question of why is it useful to reuse workflows in new analyses. Overwhelmingly, **users found that using workflows saves them time**. **They also found the visualization of the workflow useful**. Free form answers included: "We often re-run the exact same or very similar analysis steps on our data (e.g., pre-processing, statistical tests), so often we only need to change the inputs and outputs (and maybe some parameters)."

Table 1(c) shows responses for why it is useful to share workflows with others. **The overwhelming majority of respondents said workflows are useful for both non-programmers and for teaching new students.** It also saves them time because they do not need to re-implement code. No free form answers were specified.

Table 1(d) shows answers for why are workflows not shared. **Respondents did not offer very overwhelming reasons for not sharing workflows**. Free form answers included "The best pipelines to share are the ones that have all the kinks worked out, so we can explain how to edit the input and output filenames and then the person can just run it."

Table 1(e) shows responses for why it is not useful to reuse workflows from others. **Respondents did not offer very overwhelming reasons for not reusing workflows from others**. Free form answers included "Documentation can be easily fixed by adding comments or providing a verbal/written explanation along with the pipeline."

Using Groupings

We asked the same questions about groupings. Figure 3(b) shows the answers regarding the utility of sharing and reusing groupings. As with workflows, **reuse is not the only reason why groupings are created**. Unlike workflows, **reusing groupings from one's own work is more useful than reusing groupings from others.**

Table 2 shows the results for the multiple-choice questions about groupings. **Most respondents agreed that groupings help simplify workflows. Groupings also make workflows more understandable by others**. Like with workflows, **groupings save time. Groupings also make code more modular and more understandable, more so than workflows. Groupings are seen as useful to non-programmers and students. Very few respondents gave any reasons for not sharing groupings and not reusing groupings from others.** A free-form answer for why groupings are not used was "It is a pain to dissect when debugging to know where things failed. For why are groupings not shared, one respondent selected that it is hard to explain what they do, and a free form answer was "Others want a finished product, not pieces that they have to put together on their own."

If we compare the responses in Figures 3(a) and 3(b) and Tables 1 and 2, **workflows are considered generally more useful than groupings.** On the other hand, **more respondents said that groupings help make their code more modular and understandable.**

TABLE 1. SURVEY RESULTS WITH MULTIPLE CHOICE ANSWERS CONCERNING BENEFITS OF SHARING WORKFLOWS.

(a) Why is creating workflows useful?

Workflows save time	13
Easier to track and debug complex code	9
Convenient way to organize/store code	11
Help write more organized code	6
Help make code more modular/reusable	4
Help make methods more understandable	8
Visualization of overall analysis	11
Workflows facilitate reproducibility	10

(b) Is reusing workflows in new analyses useful?

Saves time	19
Gives a diagram of what was done	13

(c) Why is it useful to share workflows with others?

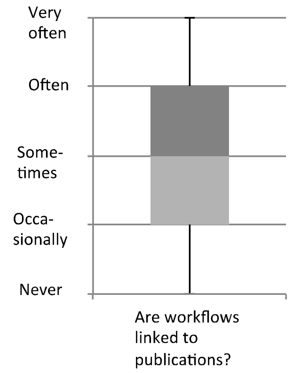
Non-programmers can use them	20
New students can easily learn	19
No need for others to re-implement code	14
Adoption of standard ways to do things	9

(d) Why are workflows not shared?

Others would not want to use them	1
Others ask too many questions of the creators	2
Workflows from others are difficult to understand	3
It is difficult to understand how to prepare data for a workflow	3

(e) Why is it not useful to reuse workflows from others?

Workflows from others are difficult to understand	4
It is difficult to understand how to prepare data for a workflow	2
Workflows created by others are too specific	1
It is hard to take workflows created by others and make them work	2



(a) How are workflows linked to publications?

Keeping lab notes about what workflows were used in a paper	11
Posting links publicly in a project or personal web site	7

(b) When workflows are not linked to publications, why is that?

People do not know how to do it	6
People do not find it useful	3

Figure 4. Survey results regarding how workflows are linked to publications.

TABLE 2. SURVEY RESULTS WITH MULTIPLE CHOICE ANSWERS CONCERNING BENEFITS OF SHARING GROUPINGS.

(a) Why is creating groupings useful?

Visualization of the analysis	10
To simplify workflows that are complex overall	12
To make workflows more understandable to others	12

(b) Is reusing groupings in new analyses useful?

Groupings save time	12
Help make code more modular/reusable	10
Help make methods more understandable	7

(c) Why is it useful to share groupings with others?

Non-programmers can use them	12
New students can easily learn	11
No need for others to re-implement code	9
Adoption of standard ways to do things	6

(d) Why are groupings not shared?

Others would not want to use them	0
Others ask too many questions of the creators	1
Workflows from others are difficult to understand	4
It is difficult to understand how to prepare data for a grouping	1

(e) Why is it not useful to reuse groupings from others?

Groupings from others are difficult to understand	2
It is difficult to understand how to prepare data for a grouping	3
Groupings created by others are too specific	1
It is hard to take groupings created by others and make them work	4

Paper Writing

We asked whether papers are linked to the workflows used in the analyses reported. Figure 4 shows the responses. **Workflows are not systematically linked to publications.** We also show that most responders believe that **the link between a workflow and a publication is kept in private laboratory notes,** rather than in a publicly accessible manner.

VII. DISCUSSION

Workflows have a clear benefit to the lab. Although a one benefit of using workflows is to easily submit jobs to the cluster shared by the research group, researchers clearly see the most benefit from sharing and reusing workflows (in Figure 3(a) medians are “very often” and “often” respectively). There are several important directions of future research suggested by this work.

One important area is to improve the use of groupings. Groupings were seeing as important to making workflows more modular and easier to understand (median is “often” in Figure 3(b)). If users had more assistance in specifying and finding groupings, it is possible that workflows and fragments would be more reused.

Another area is debugging and checking results. Currently, when workflows are large they are broken down into smaller pieces so that they can each be submitted for execution separately. Each piece is checked before the next one is submitted, which saves effort when something goes wrong. Better mechanisms to handle checking intermediate execution results would allow users to define larger workflows.

Another area of further work is better documentation of workflows. Documentation of workflows tends to be private and scattered, and not usually linked to papers. Two kinds of documentation are useful depending on the user: how to use a workflow without going into details of how it works, and details about a workflow's implementation and methods. Documentation does not necessarily imply text; it could include more sophisticated forms of interactive assistance to users based on representing explicitly the use constraints of the workflow and its steps. This approach could help in checking results and ensuring proper use of the workflows discussed in the prior point.

Finally, an important area is making it very easy to publish workflows and link them to papers. Papers provide important context and documentation for workflows. Since a single paper typically uses several workflows, users need an appropriate mechanism for linking the workflows to the paper and for specifying how the workflows relate to one another. In addition, even before a paper is being written, researchers should be able to give others access to a workflow for inspectability and analysis, particularly when researchers are using a new workflow that they may be unfamiliar with.

VIII. Conclusions

We analyzed the benefits of sharing workflows and workflow fragments from a user's perspective with a population of users of neuroimaging workflows that use the LONI Pipeline. Workflows had clear utility to users in the lab, saving time, helping researchers organize their code, helping debug complex analyses, and facilitating reproducibility. Our work can be expanded by validating our findings with more respondents, reflecting their experience level on the questionnaire and including statistics of the groupings usage on the workflows they create. Other important areas of future work include improving documentation of workflows, better support for debugging large workflows and facilitating the publication of workflows when papers are prepared. Finally, there are clear opportunities to develop best practices for designing workflow components and modularizing code, encouraging standards adoption, and facilitating understanding by other users.

Acknowledgments

This research was supported in part by the US National Science Foundation with awards IIS-1344272 and ICER-1343800 (to YG), and by the Spanish Science and Innovation Ministry (MICINN) with an FPU grant (Formación de Profesorado Universitario) and the MEC grant TIN2013-46238-C4-2-R. Additional support was provided by NIH grants R01 AG040060, R01 NS080655, R01 EB008432, R01 MH097268 (to PT) and P41 EB015922 (to AWT).

References

[1] Bergmann, R. and Gil, Y. Similarity Assessment and Efficient Retrieval of Semantic Workflows. Information Systems Journal, 40. 2014.

[2] De Roure, D; Goble, C. and Stevens, R. "The design and realizations of the myExperiment Virtual Research Environment for social sharing of workflows". Future Generation Computer Systems, 25 (561-567), 2009

[3] Dinov ID, Van Horn JD, Lozev KM, Magsipoc R, et al. Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline. Front Neuroinform. 2009;3:22. doi: 10.3389/neuro.11.022.2009. PubMed PMID: 19649168; PubMed Central PMCID: PMC2718780.

[4] Dinov ID, Torri F, Macciardi F, Petrosyan P, et al. Applications of the pipeline environment for visual informatics and genomics computations. BMC Bioinformatics. 2011;12:304. doi: 10.1186/1471-2105-12-304.

[5] Garijo, D.; Corcho, O. and Gil, Y. Detecting Common Scientific Workflow Fragments Using Templates and Execution Provenance. Seventh ACM International Conference on Knowledge Capture (K-CAP), Banff, Canada, 2013.

[6] Garijo D. and Y. Gil. A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science (WORKS-11), held in conjunction with SC-11, Seattle, WA, Nov. 12-18 2011.

[7] Gil, Y. From Data to Knowledge to Discoveries: Artificial Intelligence and Scientific Workflows. Scientific Programming, 17(3):231-246. 2009.

[8] Gil, Y.; Deelman, E.; Ellisman, M. H.; Fahringer et al. "Examining the Challenges of Scientific Workflows." IEEE Computer, 40(12):24-32. 2007.

[9] Gil, Y. Intelligent Workflow Systems and Provenance-Aware Software. In Proceedings of the Seventh International Congress on Environmental Modeling and Software, San Diego, CA, 2014.

[10] Goderis, A.; Sattler, U.; Lord, P. and Goble, C. Seven bottlenecks to workflow reuse and repurposing, in: International Semantic Web Conference, Springer, pp. 323–337. 2005.

[11] Goderis, A.; Fisher, P.; Gibson,A.; Tanoh, F. et al: Benchmarking workflow discovery: a case study from bioinformatics. Concurrency and Computation: Practice and Experience 21(16): 2052-2069 (2009).

[12] Howison, J., and Herbsleb, J. D., "Scientific software production: incentives and collaboration" ACM Conference on Computer-Supported Collaborative Work, 2011.

[13] Leake, D.B. and Kendall-Morwick, J.: Towards Case-Based Support for e-Science Workflow Generation by Mining Provenance. ECCBR'08, LNCS 5239, pp. 269-283, Springer 2008

[14] Mates, P.; Santos, E.; Freire, J. and Silva, C. T. Crowdlabs:social analysis and visualization for the sciences, in: 23rd InternationalConference on Scientific and Statistical Database Management, SSDBM, Springer, 2011, pp. 555–564

[15] Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J. et al. The Open Provenance Model Core Specification (v1.1) Future Generation Computer Systems, 27(6). 2011.

[16] Moreau, L.; Missier, P.; Belhajjame, K.; B'Far, R et al. "PROV-DM: The PROV Data Model." World Wide Web (W3C), 2013.

[17] Nature Special issue on Challenges in Irreproducible Research, Nature 496, 25 April 2013.

[18] Science Special Issue on Data Replication and Reproducibility, Science 334, 2 December 2011.

[19] Starlinger, J., S. Cohen-Boulakia and U. Leser. (Re)Use in Public Scientific Workflow Repositories. Proceedings of the 24th international conference on Scientific and Statistical Database Management, p 361-378. 2012.

[20] Taylor, I., Deelman, E., Gannon, D. and Shields, M., (Eds). "Workflows for e-Science", Springer Verlag, 2007.

[21] Thompson, P.; Stein, J.; Medland, Sarah; Hibar, D. et al. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain imaging and behavior vol 8(2).p 153-182. 2014.