Workspace Workflow Comparison

N. Oakes and D. Thomas

CSIRO Data61, Australia
Email: nerolie.oakes@data61.csiro.au

Abstract: Workspace is a Scientific Workflow System (SWS) that has been under development at the CSIRO since 2005. It is commonly used in a collaborative style with multiple people and or teams working on the same set of workflows. The workflows are serialized in XML format, and typically change periodically over time as the project develops. The main workflow development application is a graphical Workflow editor that helps the user to design and execute. Workspace also ships with Workflow comparison tool aimed at helping developers keep track of differences between multiple versions of the same workflow over time.

File comparison algorithms have a long history and are important components in fields as diverse as molecular biology, information processing, data retrieval and network security. There is always a trade-off between speed, breadth of application and development time. The Workspace workflow comparison tool is a highly customized XML comparison that parses two workflows, extracts semantically relevant information, compares the two sets of extracted information and produces an interactive graphical display that highlights relevant differences.

It presents differences in two different ways: a graphical display similar to the workspace editor with the extracted differences highlighted and a tree-based display that shows only the extracted differences.

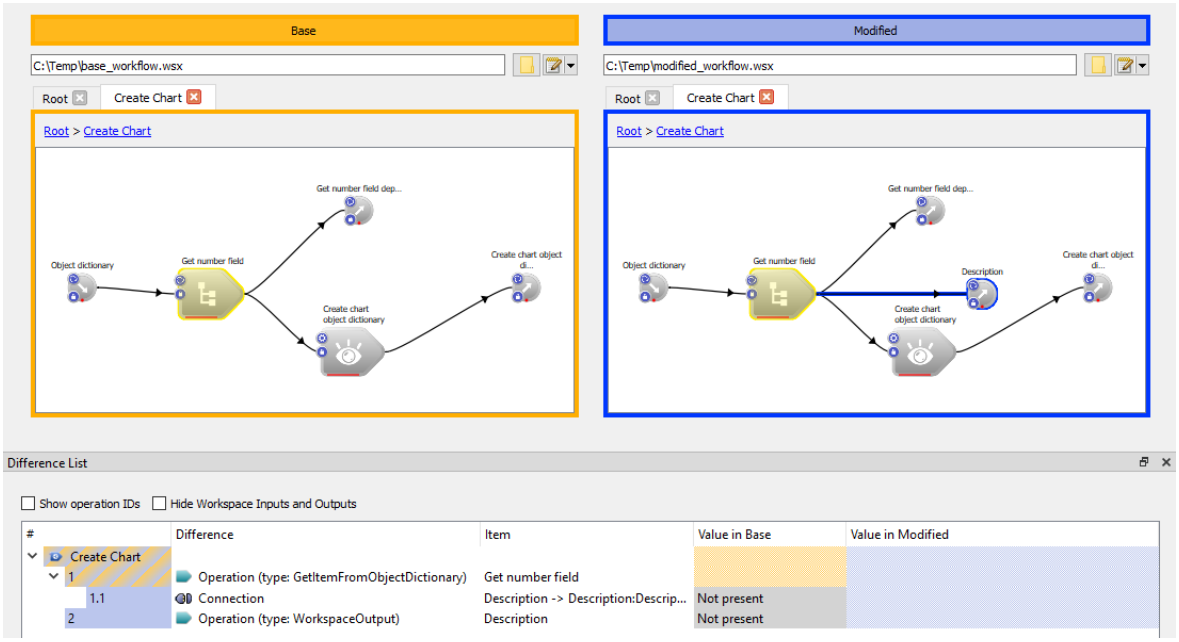

Figure 1. The Workspace Workflow Comparison Tool

In this paper, we discuss the types of workflow differences that are extracted, the difficulties of presenting this information using generic text-differencing applications, and how the workflow comparison tool helps overcome these. We also look a case study to study a set of workflows that were produced as part of a project stretching over eight years with workflow revisions saved to software versioning system.

Keywords: *Workspace, workflow, visualisation*

1. INTRODUCTION

Workspace is a Scientific Workflow System (SWS) developed by the CSIRO and released for free public usage in 2014. Initially conceived as providing a low-cost development conduit for IP, its progress is guided by four key themes: Analyse, Collaborate, Commercialise and Everywhere. (Bolger, M. *et al*. 2016; Watkins et al 2017)

An SWS can be described as a platform that enables users to construct their applications as a visual graph, connecting nodes together where each connection represents a connection between the output of one process to the input of the next. The key features of Workspace have been described by Watkins et al. (2017) A Workspace workflow primarily consists of a number of connected operations, each of which represents a self-contained task or algorithm. A workflow also contains visual and functional elements. Visual elements may include: the layout and colour of the operations, user-defined labels, connection anchor points, notes and display widgets. Functional elements include unique identifiers for operations and user-defined global names that can be used as settings, or to link a workflow to a GUI or command line.

Figure 2. A Workspace operation

Workspace workflows are represented in XML, and this standard format allows for significant variation in format and component ordering. Workflows develop over the lifetime of a project, and when multiple users and developers are collaborating on the same application, it is likely that multiple sets of changes will be made to the same workflow. Workflow documents are often so complex that even small functional differences can be hard to identify with generic text-comparison tools. In 2015 the Workspace team realised that the development of a domain-specific workflow differencing tool would make it much easier to work collaboratively and began to develop the graphical differencing tool *workspace-diff*. (Oakes et al, 2019)

The guiding principles behind this workflow comparison tool are:

- It should be designed to compare two workflows with a common ancestry
- It should identify the most relevant differences between the two workflows
- It should be easy for someone used to the Workspace editor to understand quickly
- It should be responsive to user feedback

This paper describes how the underlying workflow comparison algorithm works, and the custom graphical user interface that has been developed to help Workspace understand the differences between the two workflows under consideration. Finally, it looks at a set of workflows, developed over the course by a Workspace-based project that has been underway for several years, and saved to an SVN repository as they changed. It presents the results as a series of charts that show how workflows change over time: what kinds of differences are made, who makes them, how effective is the workflow comparison tool in pinpointing relevant changes, and how could it be improved. The project team were asked what changes they thought would immediately make the tool more useful, and the suggested changes discussed.

2. THE WORKFLOW COMPARISON TOOL

There are two main graphical components to the Workflow comparison tool: the Workflow Panels and a Difference list panel (Fig 3). The *Difference List Panel* shows an ordered list of operations that have been added, removed or changed (Fig 6 in section 3) while the *Workflow Panels* show the workflows graphically in tabs similar to those in the workflow editor (Fig. 4)..

The difference list panel presents the changes in tree widget form. Each top-level row represents an operation that has been added/removed or changed. In the Workflow Panel section, each workflow is represented graphically in a tabbed widget similarly to that of the canvas section of workspace editor (see below). There are some colour differences designed to highlight differences between the two workflows. This is explained in more detail below.

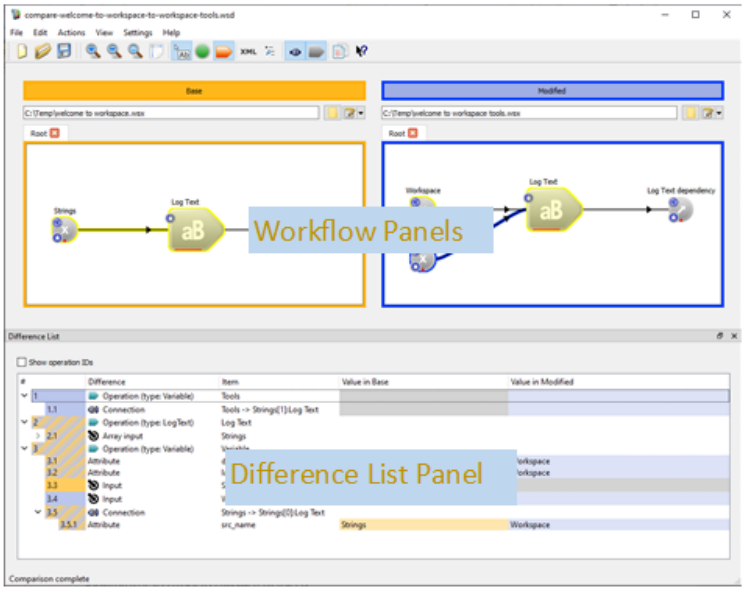

Figure 3. The Comparison Tool Layout

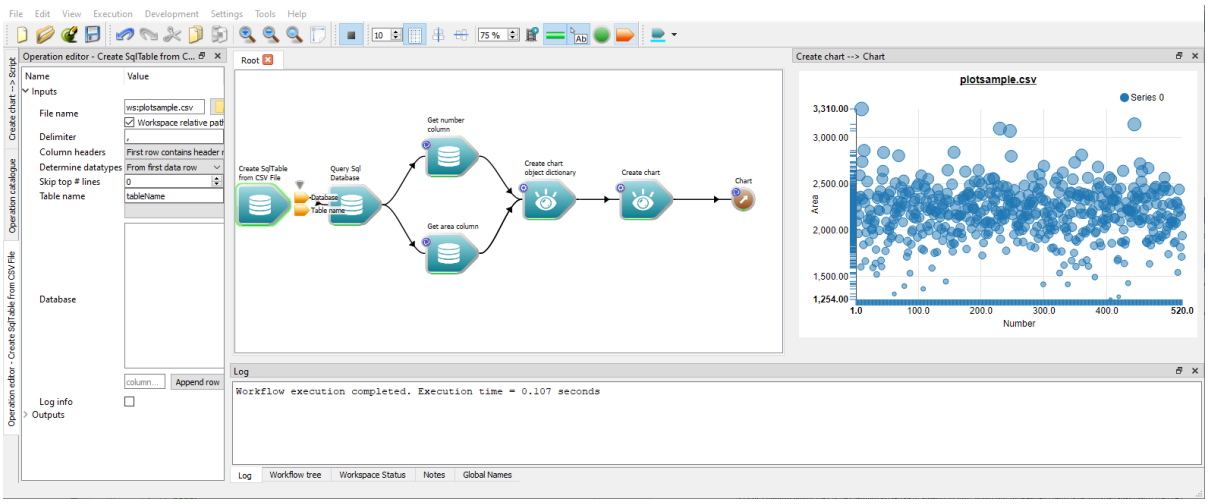

Figure 4. The Workspace Editor

3. WORKFLOW COMPARISON – EXTRACTING DIFFERENCES BETWEEN FILES

The comparison tool is itself a Workspace workflow-based application. The underlying operation is a CompareTwoWorkflows operation, which takes two workflow filenames as input and produces an array of Workflow differences as output. The application is linked to the workflow through the four inputs and one output with global identifiers (those marked with G)(see the Workspace Manual 2019). The two extra inputs are a reporting level, so that the user can control the level of logging, and an update barrier trigger which lets the application control when the workflow is run.

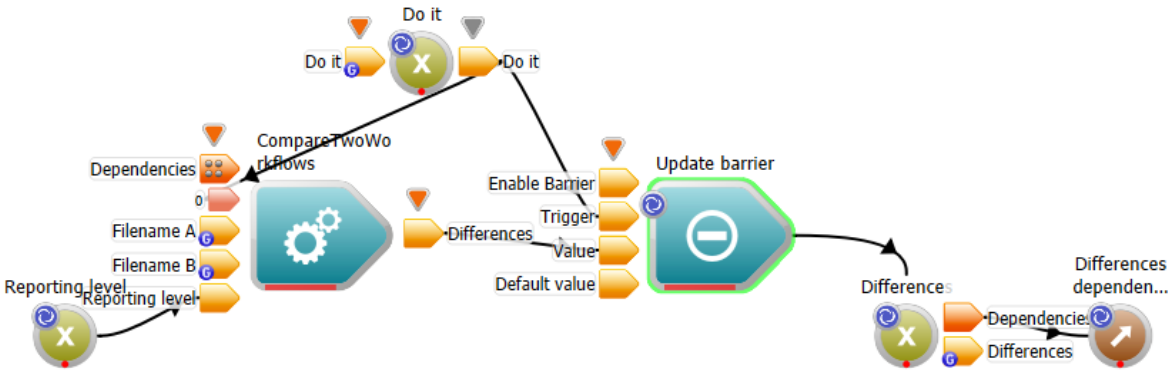

Figure 5. The Workspace workflow on which the Workflow Comparison tool is based

The application displays the workflow difference array as a custom tree widget, which is directly linked to the Operation differences array through the globally-identified output, *Differences*.

When a new operation is added to a workflow, a Universally Unique IDentifier (UUID) is generated. When two workflows are derived from a common ancestor, then any operations added before they branched will have the same UUIDs in both. Operations added after the versions branched will have newly-generated UUIDs. The parsing algorithm takes advantage of this when looking for changes. It takes two passes through each of the workflows, populating first a map of all the operation-UUID to data for each operation, and secondly a set of connections between operations. Next, it compares the two sets of maps that were generated, UUID by UUID, looking for non-trivial differences between them.

One of the most important advantages the custom comparison tool has over a generic differencing tool is the ability to define what is important, and what is not. It can completely ignore differences in node order where the XML code applies this randomly. It looks for elements that have been added or removed, as well as changes. Currently, the diff tool looks for the following kinds of differences: operations (and attributes such as labels, global names and colours); connections; input and outputs (names and/or values); notes; scheduling features; layout changes – such as operations moved significantly or connection anchors; display widgets; and Workspace and plugin versions. Anything NOT on this list is ignored,

For each UUID,:if it only exists in one map, then an OperationDifference is created and added to the Difference array; if it exists in both maps, then it looks for differences between the two sets of data, using rules specific to each type of difference. If it doesn't find any differences, then the operation is ignored; if it finds differences, then then an OperationDifference is created and added to the Difference array along with data about each of the differences found.

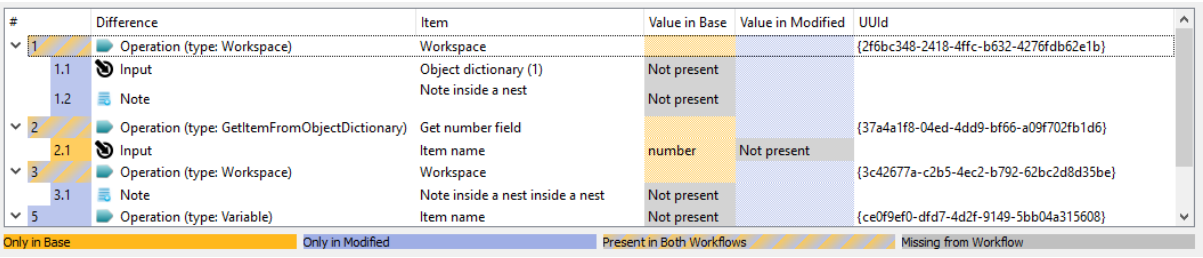

Figure 6. The Comparison Tool Difference List panel

4. GRAPHICAL COMPARISON OF WORKSPACE WORKFLOWS

One of the most important features of the workflow comparison tool is the ability to present differences in the familiar environment of the Workspace editor. The workflow component of the tool is based on the canvas used in the Workspace editor, so here is a brief explanation of how it works

4.1. The Workflow editor

To create instances of operations, users drag-and-drop them from the operation catalogue onto the Workspace canvas. Users can edit the values of unconnected inputs. The user can also select an alternative widget to view/edit the data for a given input. GUI widgets can be created to visualise data attached to any input or output in the workflow. Notes can be added. Users can also change the layout in order to make it easier to understand what the workflow is doing. Such layout elements are saved along with the Workflow so that the user can set up a development layout specific to each workflow. The Workspace editor also lets the user edit data and have a combination of configurable and dynamic control over what is displayed. For more details about how the Workspace workflow editor works, see (Oakes et al, 2019).

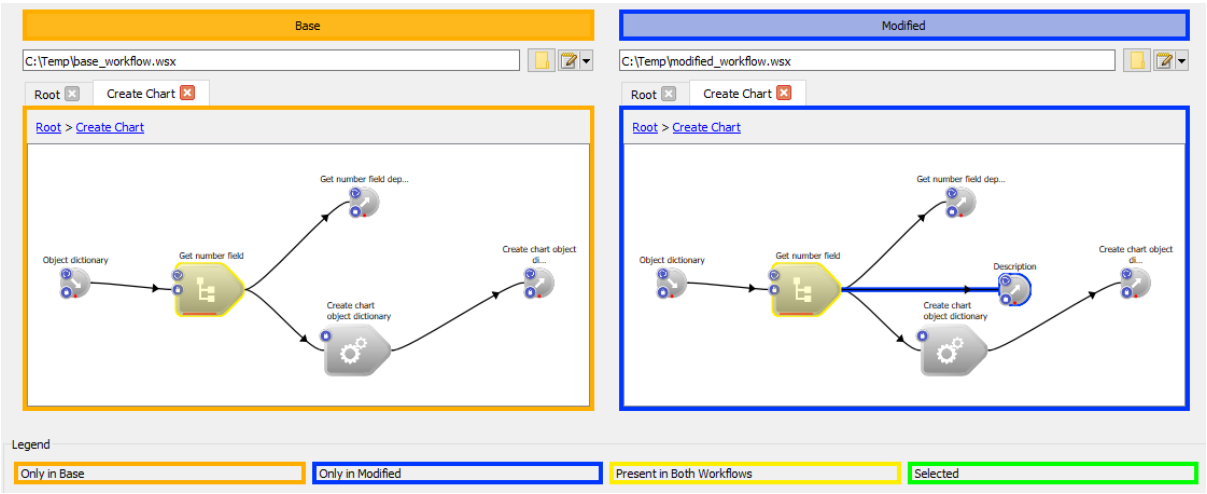

Figure 7. The Comparison Tool Workflow panel

In the comparison tool, each workflow is represented graphically in a tabbed widget similarly to that of the workspace editor. The two main differences are that (a) you cannot edit data and (b) There are some differences designed to highlight differences between the two workflows.

There are lots of features designed to make it familiar to Workspace-editor users, such as legends and tooltips (which can be turned off). Most of the user-configurable display options available in the Workspace editor are available here: you can zoom in and out, display or hide mini-operations, and view operation properties and the *Operation editor* in read-only mode.

There are also many features not found in the Workspace editor to aid comparison: most obviously the colour changes. In the figure, items that exist only in the Base workflow are highlighted in gold and items that exist only in the Modified workflow are highlighted in purple. Operations/connections that are part of both workflows are given a light yellow outline if something has changed. If you want to see the highlighted outlines more clearly, you can toggle between greyscale and coloured operations by clicking on the "desaturate operations" tool button. The colours used are user-configurable, meaning that vision-impaired users can select options that suit them.

Other features not in workspace-editor include options to show ghost elements (surrogates for a missing element), desaturate the operations (so that the annotations pop out), and work even without plugins (while a missing plugin is an error in the editor, it is not in the diff tool, which does not need to be able to run the workflow: the diff tool tries to make it look as normal as it can).

5. WORKFLOWS AS THEY CHANGE

The case study considers a set of 25 workflows that have been under ongoing development as part of a ten-year collaborative project. Firstly, the team was consulted about how the diff tool might be enhanced, and secondly, we looked for differences between the file versions saved to their SVN repository.

The team suggested the following enhancements: that changes to inputs and outputs be easier to see on the workflow panels; that all the data normally available in the Workspace editor, be available, not just changed data, that the significance of the colour annotation be more obvious, and that there be more user-control of the configuration of display elements. The SVN repository workflow revisions were scrutinised, first by running a single-workflow analyser (another Workspace workflow tool) over each to determine how large the workflow was (by operation count), and secondly by running differencing tools over every pair of revisions. First the generic text-difference supplied by TortoiseSVN was used to determine a rude sense of the scope of the change according to the number of lines added and removed; secondly, the Workspace diff tool was used to compare them.

Of the 25 workflows in the set, six were never revised after being saved initially, and a seventh workflow was revised just once to correct improper formatting. These have been excluded from the results. Only two of the workflows were revised by more than one author. The figure below shows how the Workspace diff tool reduces the number of differences shown compared to a generic text difference by filtering out insignificant changes.

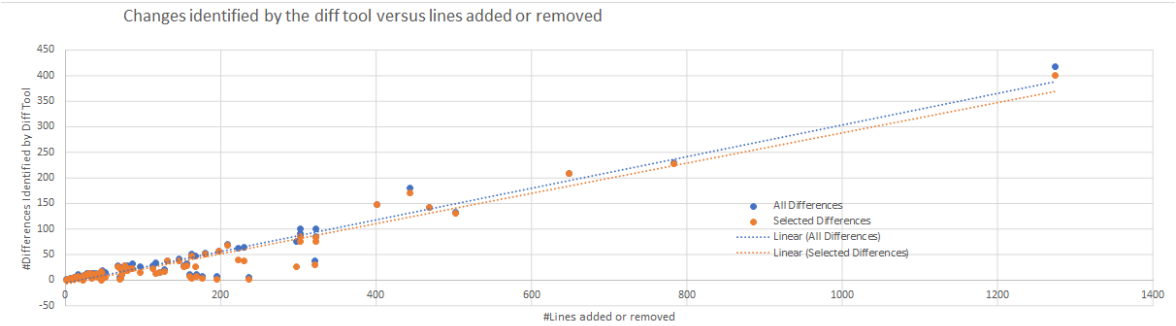

Figure 8. Plot of the number of significant changes as identified by the customised comparison tool compared to the unfiltered output of a generic text differencing tool

There is a substantial unit test suite behind the diff tool, and it is consequently unlikely to miss non-trivial changes. Hence the ratio of differences identified by the diff tool to those of a generic difference is a measure of how effective the tool is at identification. The current tool has a ratio of 0.3 for all changes (blue line). The orange dots represent functional changes (ignoring layout changes) and the ratio improves slightly to about 0.28.

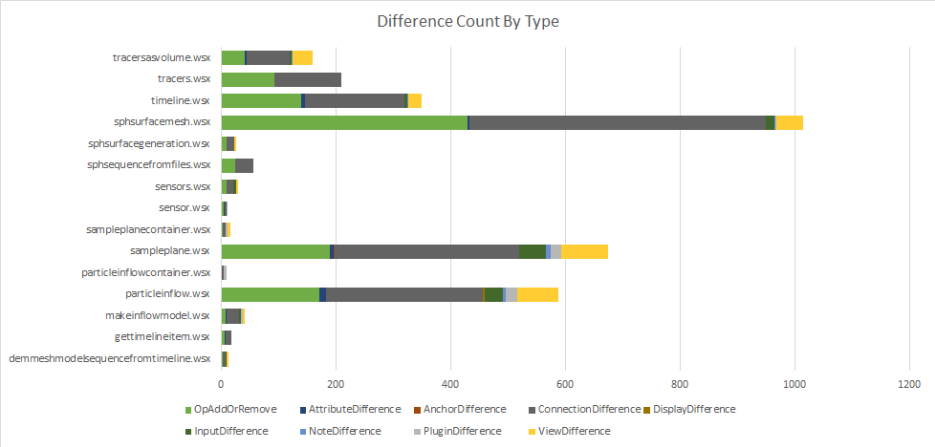

Figure 9. The types of changes typically made

Fig. 9 shows the distribution of the types of changes we detected. The most common type of change is to connections, and secondly to operations. The next figure shows how the types of differences as a function of the operations changed in the revision. There is a strong correlation between this and the number of connection changes made, but other types are only weakly correlated. The final chart shows the distribution of changes as a function of overall workflow size.

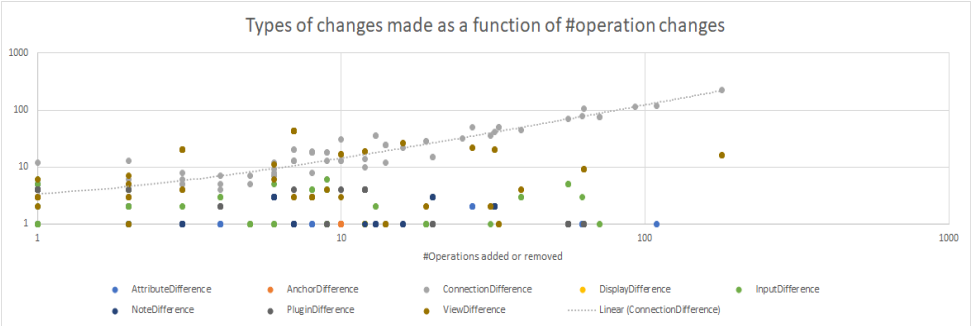

Figure 10. The types of changes as a function of how many operations were updated in the revision

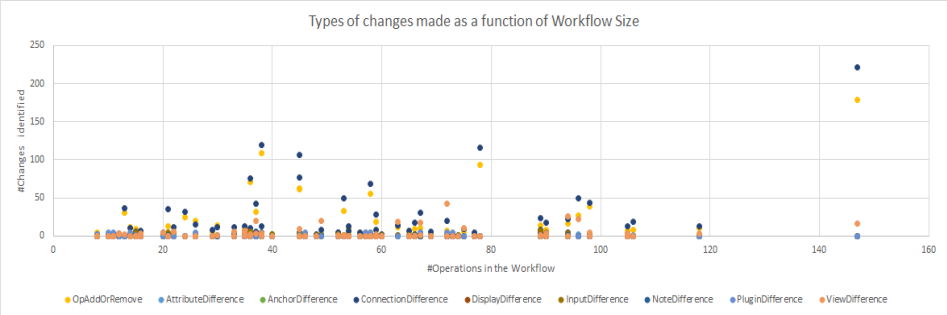

Figure 11. The types of changes as a function of the overall size of the workflow

6. DISCUSSION AND CONCLUSION

The Workspace workflow comparison tool is intended to make it as easy as possible for users to understand the differences between two workflows. The case study aids in understanding how the workflows change over their lifetime, and how best to capture this. The most significant difference, objectively, is the ratio of differences identified to lines of text changed, with lower ratios indicating better efficiency of the diff tool. Often, a single action inside a workflow will produce multiple correlated changes, and ideally, we would capture these as a unit. This ratio is currently standing at about 0.3, and the most obvious way of reducing this would be to capture correlated addition/removal of operations and the new or removed connection as a unit. There are other semantic changes that are created in a single mouse-click, including nesting and explosion of sub-workflows, adding chains of linked inputs and outputs. To date, no serious attempt has been made to optimise the speed of the differencing algorithm. This is not a current concern for small workflows, but an increase in speed would be noticeable for the more complex workflows created (see Mikhael 2011).

It is interesting to note that none of the enhancements suggested by the team concerned the types of differences detected, nor the speed of computation. This tends to confirm that the most important contribution that the differencing tool makes to productivity is the display of pertinent data in a way that is easy to understand. Consequently, this will continue to be an important focus of the development of the tool.

REFERENCES

Bolger, M. *et al*. (2016) ,Workspace: a fast and low cost methodology for delivering commercial applications based on Research IP, *eResearch Australasia*, 6–10.

Cleary, P, Watkins, D., Hetherton, L., Bolger, M. and Thomas D. (2017). Opportunities for workflow tools to improve translation of research into impact, *22nd Int. Congr. Modelling Simulation*, 1–7.

Mikhaiel, R. A. E. (2011), Comparing XML Documents as Reference-aware Labeled Ordered Trees, *PhD submitted to University of Alberta*

Oakes, N , Hetherton, L , Bolger, M , Thomas, D , Rucinski, C , Watkins, D , Cleary, P (2019). Workspace – a Scientific Workflow System with commercial impact, *23nd Int. Congr. Modelling Simulation,[1]*

Watkins, D., Thomas D., Hetherton, L., Bolger, M. and Cleary, P.W., (2017). Workspace – a Scientific Workflow System for enabling Research Impact, *22nd Int. Congr. Modelling Simulation.*

Workspace Manual (2021), https://research.csiro.au/static/workspace/docs