NeuroProv: Provenance Data Visualisation for Neuroimaging Analyses

Bilal Arshad , Kamran Munir , Richard McClatchey ,
Jetendr Shamdasani , Zaheer Khan

PII: S1045-926X(16)30052-0
DOI: https://doi.org/10.1016/j.cola.2019.04.004
Reference: COLA 899

To appear in: *Journal of Computer Languages*

NeuroProv: Provenance Data Visualisation for Neuroimaging Analyses

Bilal Arshad, Kamran Munir*, Richard McClatchey, Jetendr Shamdasani & Zaheer Khan

Centre for Complex Cooperative Systems (CSCT), Department of Computer Science and Creative Technologies (CSCT), University of the West of England, Bristol, BS16 1QY, United Kingdom
{Bilal.Arshad, Kamran2.Munir, Richard.McClatchey, Jetendr2.Shamdasani, Zaheer2.Khan} @uwe.ac.uk

Abstract- Visualisation underpins the understanding of scientific data both through exploration and explanation of analysed data. Provenance strengthens the understanding of data by showing the process of how a result has been achieved. With the significant increase in data volumes and algorithm complexity, clinical researchers are struggling with information tracking, analysis reproducibility and the verification of scientific output. In addition, data coming from various heterogeneous sources with varying levels of trust in a collaborative environment adds to the uncertainty of the scientific outputs. This provides the motivation for provenance data capture and visualisation support for analyses. In this paper a system, NeuroProv is presented, to visualise provenance data in order to aid in the process of verification of scientific outputs, comparison of analyses, progression and evolution of results for neuroimaging analyses. The experimental results show the effectiveness of visualising provenance data for neuroimaging analyses.

Keywords: Provenance; Scientific Workflows; Biomedical Analysis; Neuroimaging; Visualisation

1. Introduction

E-Science platforms are growing at pace but they still lack in providing provenance support, such as provenance capture, storage and its usage to support analyses [1]. There are systems such as Prototype Lineage Server [2] that can adequately capture provenance data and store it in several formats, but, as yet, few researchers use it. Neuroimaging community requires the means to access and understand provenance data in order to support clinical analyses. This helps researchers, for example, in the study of Magnetic Resonance Imaging (MRI) to determine biomarkers for the onset of Alzheimer's disease [3, 4].

The domain of neuroimaging is complex - it includes multiplicities (versions) of datasets and versions of algorithms operating upon these datasets following specific workflow patterns. Breakthroughs in large-scale data analysis for neuroimaging are few and one of the major contributing factors is the lack of provenance data support. Without provenance data, researchers do not have the context of the analysis being performed. Furthermore, chances are that the absence of provenance data support increases the risk of an error(s) in data analysis. Therefore, since the analysis context is missing it is difficult to ascertain the authenticity of results. Since neuroimaging contains multiplicities (versions) of data; an error in an earlier step could percolate to the next stages of analysis and could alter the end result. The researcher might get an inaccurate result at a later stage that may lead to inappropriate results getting published. Thus the researcher might end up with an altogether different set of results compared to the anticipated results at the beginning of the analysis.

Since the analysis is repeatedly conducted in a collaborative research environment it is imperative to retain a track of who did what, when, on what data, using which algorithms, and why? All this information needs to be traced and logged so that the results can be visualised for easy understanding and analyses can be reproduced or amended as part of a rigorous research process. Typically, provenance data is represented in files and/or tabular format which, for complex workflows, is not straightforward for analysis by the scientific community and/or practitioners. Visualisation and provenance techniques, although used rarely in combination, may further help to increase the scientist's ability to understand scientific results. The scientist may be able to use a single tool in order to: a) evaluate final results; b) the derivation process; and c) any intermediate results produced

* Corresponding author
Email address: kamran2.munir@uwe.ac.uk (Kamran Munir)

during the experiment. In order to aid the researchers in the exploration process, there is consequently a need to apply suitable visualisation techniques on the provenance data.

However, the existing state of the art workflow systems are not completely generic and reconfigurable. Most workflow provenance management systems are designed for data-flow oriented workflows and researchers are now realising that tracking data alone is insufficient to support the scientific process (for example, see [5]). In this regard, this paper presents a system named as NeuroProv, to visualise provenance data for neuroimaging analyses. NeuroProv can present large amounts of provenance data in a visual format that is both intuitive and easy to understand by clinicians and neuroscientists. As a starting point, user requirements have been taken based on the scenarios defined in the N4U [6] project, which aimed to provide computing and storage infrastructure, and services to store neuroimages and to facilitate neuroscientists in defining and executing neuro-analysis on stored images.

The remainder of this paper is structured as follows: Section 2 presents related work, with emphasis on neuroimaging analysis, provenance visualisation and provenance/workflow systems; Section 3 elaborates requirements for provenance visualisation; Section 4 describes use-cases; Section 5 introduces NeuroProv system architecture; Section 6 highlights evaluation, results and discussion and Section 7 presents conclusions and future direction.

2. Related Work

Scientists often rely on visualisations to aid in data exploration that can be a complex process requiring close collaboration among domain scientists, computer scientists and visualisation experts. This section summarises related work and literature review in the context of neuroimaging analysis, provenance visualisation, and provenance/workflow systems.

2.1 Need for Provenance Data in Different Scientific Domains

Neuroimaging is an essential means for research and clinical neuroscience [7]. Provenance can be used for data interpretation, assessing data quality, and data interoperability [8] [9]. For example, in order to assist research into various neuro-degenerative diseases, such as Alzheimer's, researchers need to process brain scans for various biomarkers [4] [10] [11]. These biomarkers include the cortical thickness of the brain, thinning of which has been linked to the onset of Alzheimer's disease. In biological sciences and neuroimaging, in particular, it is imperative to keep track of provenance to assess the quality of data being gathered and to enable clinical researchers to verify results. With the advent of current techniques, one the prime challenge faced alike in all biological sciences is the management of the vast amount of data being generated. Coupled with the need for collaborations for scientific innovation and discovery, the need to share data over multiple locations, ensuring its availability and usefulness to the scientific community adds to the challenge manifolds.

Similarly, scientific experiments such as those at the Large Hadron Collider (LHC) at CERN [12] and projects such as N4U [6] generate extremely large amounts of data. These communities use scientific workflows [13] to orchestrate the complex processing of data for their analyses. During the computation of this large pool of data, scientists end up creating an even larger pool of data representing intermediate results and associated metadata. An important consideration during data processing is to understand the intermediate results and processes. This helps to derive final results, to verify the authenticity of those results and provide insight. Several systems have been recently proposed to capture provenance such as [14] for script executions using noWorkflow and YesWorkflow [15]. Other systems use provenance of 'Research Objects' (i.e. gathering of digital artefacts to enable knowledge sharing and reproducibility [16] [17]) for sharing knowledge about computational experiments [18]. Other systems include the use of graphs for representing provenance such as SGProv etc. [19] [20] [21] by providing summarization mechanisms and clustering of views. Here, views refer to the graphical representation of provenance data in a format understandable to researchers and users. Summarisation of views is essential for neuroimaging since it allows researchers to have a high-level summary of the provenance in addition to the fine-grained provenance for inspection. Several works have been undertaken to define various models for scientific workflow provenance such as PROV data model [23], its extension ProvOne data model [24] and formal model of provenance by El-Jaick et al. [22]. Most of these models are an extension of the original PROV model by Luc et al. [25] and contain elements that have been improvised over the period of time to meet the needs of provenance based solutions.

2.2 Provenance Visualisation

Effective visualisation of provenance data is necessary to understand the history of events and to evaluate data. The conventional visual encodings for provenance data are derived from the fields of network and graph visualisation. The most common visualisation strategy for provenance data is the node-link diagram (e.g. graphs) and is employed by common provenance tools such as Probe-It [26], Haystack [27] and Orbiter [28]. With this visual encoding, graphs are represented as nodes and edges (connections between nodes are represented as lines or curves). These tools utilise a variety of different visual encoding techniques including directed node-link diagrams [26] [27] and collapsible summary nodes [28]. Node-link diagrams are effective for representing provenance data for understanding local-activity, but they do not offer a high-level summary of activity and relationships within them. What is required is a system that provides detailed inspection of workflow elements alongside a high-level summary of the workflow.

Prior visualisation systems, deal with either the data product or the process, but not both. Specifically, Taverna [29] uses Haystack [27], which is a visualisation tool to help answer questions that establish how experimental results were obtained. However, Haystack does not offer its users to compare visualisations essential for detecting anomalies. VisTrails [30] allows users to navigate workflow versions in an intuitive way, to visually compare different workflows and their results, and to examine the actions that led to the result. The supplied visualisation portrays each workflow as a node, with each change to workflow as an edge from the original workflow to the modified result. In other words, VisTrails manages and displays the provenance of visualisation (keeps track of how workflows evolve), while NeuroProv and other provenance visualisation systems described in this section display visualisation of provenance. While not designed primarily as a visualisation tool, VisTrails workflow tracking feature could be useful in a provenance visualisation system if it allowed users to see how they had customised the configuration of the display, and how their interaction with the display evolved over time.

Other systems such as Probe-It! [26] enables scientists to move the visualisation focus from intermediate and final results to provenance, back and forth. However, it does not allow encompassing the ability to track the progression of data products or their evolution over the course of time. This is essential in deducing how workflows and data products have evolved over the course of an experiment. The Prototype Lineage Server [1] allows users to browse lineage information by navigating through sets of metadata that provide useful details about the data products and transformations in a workflow invocation. Prototype Lineage Server restricts the users to exploring parent and child metadata objects. Thus limiting its ability to be used for neuroimaging analysis where users need to view a multitude of metadata objects those extending beyond parent and child relationship. What is required is a system that allows users to drill down multiple levels of detail to access fine-grained provenance.

Provenance Explorer [31] provides users with personalised views of visualised provenance data, based on a combination of user requirements, semantic reasoning and access policies. Major drawbacks include the support to drill down to only one level of detail limiting users to the coarser level of provenance, for verification and authentication purposes fine-grained provenance is required to inspect elements of the workflow. Prov-O-Viz [32] is a web-based visualisation tool for PROV [25] based provenance tracks coming from various sources that leverage Sankey diagrams [33]. The aim of this system is to focus on a visualisation approach to identify the important activities within a provenance graph based on data flows. However, Prov-O-Viz only allows users a limited interaction with the visualisation (e.g. the ability to only move nodes vertically), whereas, for verification purposes, it is essential for end users to be able to fully interact and inspect elements of a workflow.

Above related work clearly indicates that there have been attempts for visualising provenance data of scientific workflows. However, existing approaches lack a comprehensive level of provenance details and/or the ability to have a high level of interaction with visualisations enabling scientists to explore and investigate workflow provenance. NeuroProv attempts to handle these limitations by providing users with the ability to fully interact with the visualisation of provenance data, inspect individual elements of the workflow (activities/entities), compare multiple workflows and visually see how a workflow has evolved over the period of time.

3. Generic Requirements for Provenance Visualisation

Kunde *et al.* [16] derive abstract user requirements for provenance visualisation, including: 1) process: the sequence of process steps is the centre of inspection; 2) results: the intermediate or end results of interactions are the centre of users view; 3) relationship: the relationship between actors is important; 4) timeline: the time is important to observe; 5) participation: the correctness of the participants is important; 6) compare: the comparison of subjects shows the difference between them, and 7) interpretation: an individual visualisation view depending upon end-user's requirement.

The goal of the visualisation presented in this research is to serve both the broadly and narrowly focused audiences in the domain of neuroimaging analysis. Consequently, it addresses each of the above seven requirements as follows: Requirements 1-3) NeuroProv is based upon an accepted model for provenance representation, namely, the PROV-XML [25], which denotes entities, activities and agents as nodes, and the relationship between them as edges in a graph. It is able to show complete graphs with both the process steps and intermediate (final) results, or abstract graphs focusing on either one of them; Requirement 4) the PROV-XML is capable of representing time information for nodes and edges; Requirement 5) participation is represented by agents through a "wasControlledBy" relationship in the PROV-XML, so NeuroProv helps the user visually evaluate the correctness of participation; Requirement 6) users can compare attributes of nodes and even compare multiple visualisations using it; and Requirement 7) for the last type of user requirement (interpretation), the proposed solution allows advanced users, in-depth inspection of workflows by drilling down to reveal further fine-grained provenance information about sub-activities (for a more detailed summary on PROV-XML, refer to [25]).

Goble et al., [34] define the seven W's (Who, What, Where, Why, When, Which, hoW) in order to capture aspects of provenance. This includes features of provenance fundamentals for the use of provenance data such as the person involved in the experiment (who); the material and methods used in the experiment (what and how); the conditions and timings at the time of the experiment (when and where); the purpose of running the experiment (why); as well as the results and the conclusions of the experiment (what). For neuro-imaging scientists, both the intention and the results of experiments are of crucial importance as well as the understanding the "how to" of experiments.

The coupling between the results and the associated provenance is inherent thus justifying the development of techniques to facilitate easy viewing of both. Kunde et al. requirements [16] do not encompass all aspects of the research primarily due to the fact that these requirements are very generic. This section presented generic requirements for provenance visualisation whilst the following section describes domain specific requirements for provenance visualisation for neuroimaging analysis. The following section presents use-cases based on our case study N4U and highlight requirements at the end of each use-case.

4. NeuroProv Use Cases & Domain Specific Requirements

Based on Goble's seven W's, generic requirements defined in [34] and using the N4U as a case study, the following assertions can be made: Scientists can use the visualised provenance for the following purposes: Verification - to verify a result or an intermediate result during the course of an experiment; Comparison - compare a certain result against an existing result; Progression - analysis of origin of results of an experiment and Evolution - following the natural course of exploration during an experiment.

This work has been conducted within the context of the N4U project [2]. N4U provides neuroscientists and clinicians with the ability to perform high-throughput imaging research and provides neurologists automated diagnostic imaging markers for neurodegenerative diseases such as Alzheimer's for individual patient diagnosis. The experiment also allows users to securely upload, use and share brain scans paired with access to computational power, large image datasets and specialised support and training for conducting neuroimaging analysis. The intended benefit of the N4U project was to enable the discovery of biomarkers for Alzheimer's disease that can improve diagnosis and help speed the development of innovative drugs. Within the context of N4U, its end-user community had identified a vital need for provenance.

In order to specify requirements for provenance visualisation, this section presents a potential end-to-end example scenario of the use of NeuroProv System, see Figure 1. This sets up the stage for presenting specific use-cases that contribute to generating additional domain-specific requirements for our proposed system NeuroProv. In N4U, various users such as Research Leaders, Researchers, Pipeline Developers, Image/Data Input Managers and System Administrators use provenance data for numerous purposes. For example, suppose a workflow yields some surprising and possibly significant results. A researcher may wish to confirm that the results are accurate and identify any mistakes that may have been made. Visualisation of provenance data for the workflow provides the means to analyse all the intermediary image sets and results to verify that the results and their analyses were incorrect. It may be found that the error was due to a specific group of images interacting badly within the workflow. The user can then annotate the workflow so that other users are warned if they attempt a similar analysis.

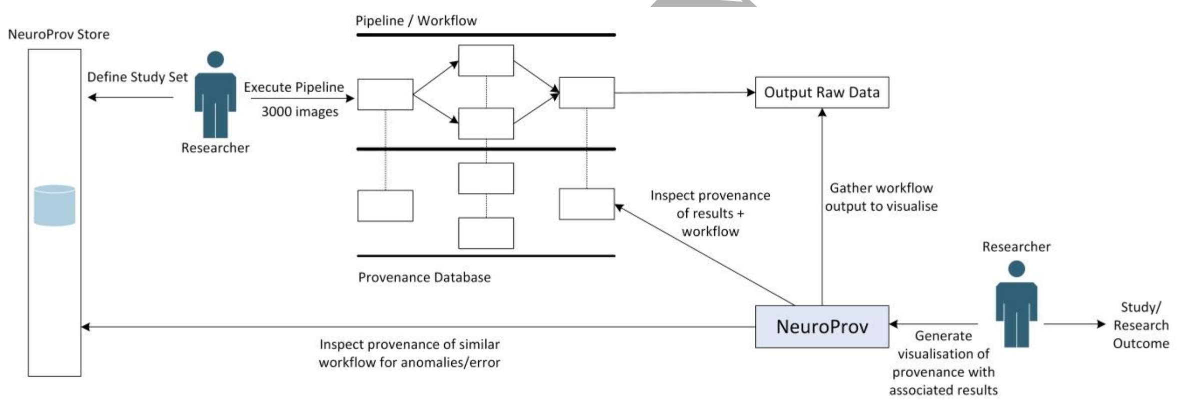

Figure 1 - NeuroProv end-to-end Example

Sometimes it may not be sufficient to simply reproduce the analysis results since it may also be necessary to validate and, if required, reproduce the workflow that has been used to obtain the results. This makes users confident not only in the results that have been produced but also in the process that led them to generate those results. For example, a user may create a new workflow and runs it on a test dataset. At each stage in the execution of the workflow, intermediary images or data are stored and a full track of provenance is kept. After results have been produced, the user can examine the visualised provenance to check that each stage of the analysis was completed correctly. The raw results can then be exported into the user's preferred analysis tool and the whole process can be added to the researcher's history for future reference. Initially, the new workflow may produce some poor results during testing. The researcher, therefore, can inspect the visualised provenance of the workflow execution and locate the problem. The user can then interact with the system to make changes to the relevant settings and re-run the test study. This time the process may run correctly and meaningful results may be produced. Without the mechanism to validate workflows, it would not be possible to correct the process and generate accurate results. In this case, the visualisation of provenance data clearly helps the researcher to validate results and workflows.

Each segment of the user requirements begins with a Use-Case story. The relevant user-requirements that are contained within it are described below it. The requirements are denoted by the R prefix. The prioritisation scheme focuses on Essential (E), Desirable (D) and Optional (O) requirements and is based on the variation of MoSCoW technique [35]. MoSCoW technique is an acronym of the first letter of each of the four prioritisation categories (Must have, Should have, Could have and Won't have) [35]. Essential requirements are those that are absolutely vital to the production of a functional system. Desirable requirements are those that whilst not vital, would provide important functionality to users and a reasonable proportion of these should be implemented. Optional requirements are those that might be useful but don't fit into the previous two categories and will probably be the last to be implemented if time allows. The individual use-cases and requirements have been prioritised using this scheme. The aim of this is to relate the priorities of finer-grained requirements within the context of the broader use-cases. This is not always easy to achieve and there are bound to be some conflicting demands. It was felt, however, that this provides an insight into how users think about and assess the priority that should be given to the various components of N4U. In this regard the following four use-cases are considered:

Use-Case 1:

'Verify results using visualised provenance data'. A workflow yields some surprising and possibly significant results. A researcher wishes to confirm that the results are accurate and identify any mistake that has been made, refer to Figure 2. By analysing the visualisation of associated provenance data the user is able to verify that the results were incorrect. It is found that the error was due to a specific group of images interacting badly within the workflow. The user annotates the workflow so that other users are warned if they attempt a similar analysis. Furthermore, consider the user creates a new workflow and runs a test dataset using it. At each stage in the execution of the workflow, the intermediary images or data are stored and a full provenance track is kept. After results are produced, the user examines the provenance to check that each stage of the analysis was completed correctly, refer to Figure 2. The raw results are then exported into the user's preferred analysis tool and the whole process is added to the researcher's history for future reference. For example, initially, the new workflow produces some poor results during testing. The researcher, therefore, looks at the associated visualisation of provenance data and locates the problem. The user then interacts with the system to make changes to the relevant settings and re-runs the test study. This time the process runs correctly and meaningful results are produced.

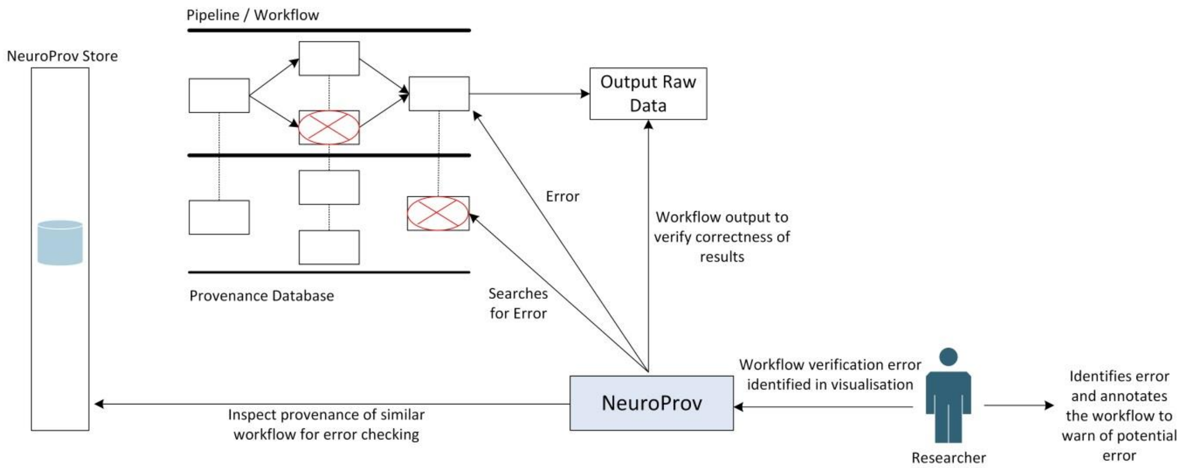

Figure 2 - Use-Case 1 Verification of Results

User Requirements:

Req No.	Description	Priority
R1.1	Carry out verification of all the stages that have been processed during workflow execution using the associated visualised provenance data.	E
R1.2	Perform a statistical analysis on provenance data.	O
R1.3	Annotate a workflow with information regarding potential errors and incompatibilities.	O
R1.4	Search a list of common errors that are known to affect a given workflow.	D
R1.5	Validate a workflow using visualised provenance to locate points of failure in it.	E
R1.6	Report errors in workflow execution.	E
R1.7	Annotate workflows with version information and complete change history.	D
R1.8	Provide users with workflow details and annotation present in the database.	E
R1.9	Provide users with the workflow execution timeline.	E

Use-Case 2:

A new workflow has been developed and verified. A user decides that it might be useful to compare it with an existing workflow designed for a similar study set. The user wants to compare the results of newly designed workflow against the past analyses, see Figure 3. The user requests to generate a visualisation of both the workflows to give a better understanding of the working of the two workflows under consideration thus yielding further insight into the study.

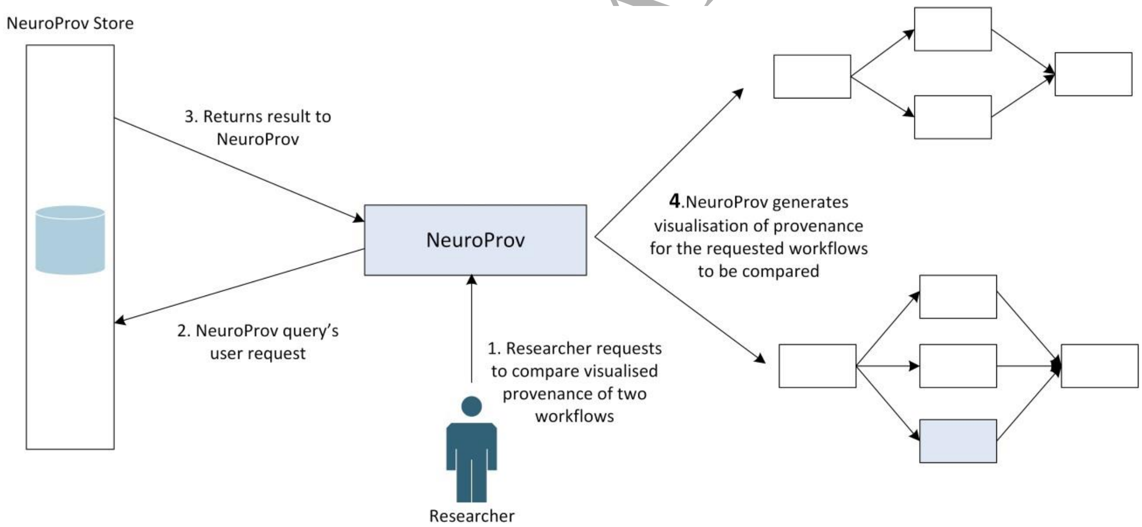

Figure 3 - Use-Case 2 Comparison of Workflows

Req No.	Description	Priority
R2.1	See how a workflow/dataset compares to an existing workflow/dataset.	E
R2.2	See how a workflow compares to multiple workflows.	E
R2.3	Visualise comparison of activities/entities of one or more workflows.	E

Use-Case 3:

The third use-case 'Evolution of Workflows'. A user wishes to view how a workflow or dataset has evolved over period of time for a particular type of analysis, see Figure 4. The study includes how a workflow has been used by owners and later on, edited by other users to carry out their respective studies. This provides the user with the essential know-how of why a particular workflow or dataset was used for a particular analysis and provide a basis to conduct his/her own analysis. The study also includes inspection of how a particular workflow has evolved over a period of time in order to determine what changes have occurred; who brought the changes and for what purposes.

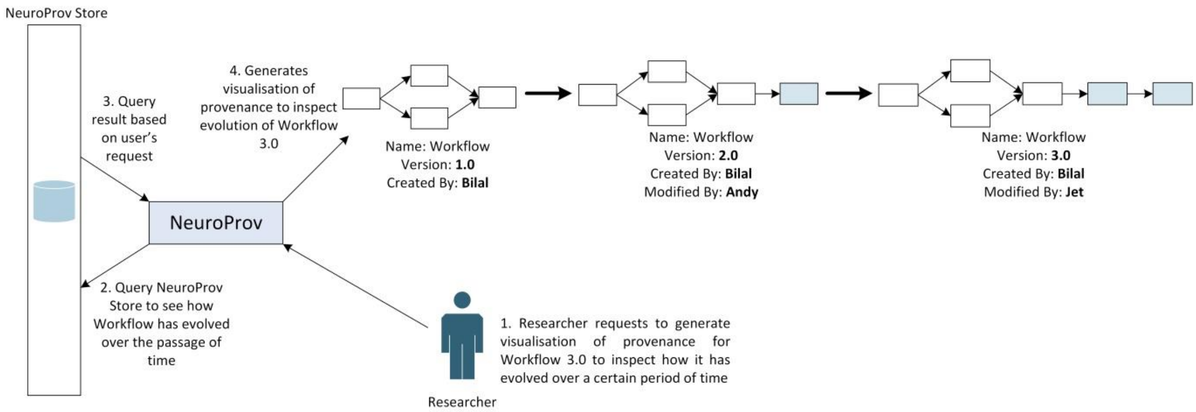

Figure 4 - Use-Case 3 Evolution of Workflows

Req No.	Description	Priority
R3.1	See how a workflow/dataset evolves over a period of time using visualised provenance data.	E
R3.2	Annotate workflows/datasets with useful information for future use.	E

Use-Case 4:

The fourth use-case 'Progression of Workflows'. The user wishes to conduct an analysis and wants to see if any other research team has already conducted a similar experiment. This will save the researcher some time and effort. The research that is already produced acknowledges the contribution of the workflow/dataset it becomes an established research method more quickly than would have been possible otherwise. The user will search for a particular workflow/dataset and the system will provide visualisation of provenance to be examined and if appropriate use the workflow/dataset for the researcher's further analysis, see Figure 5.

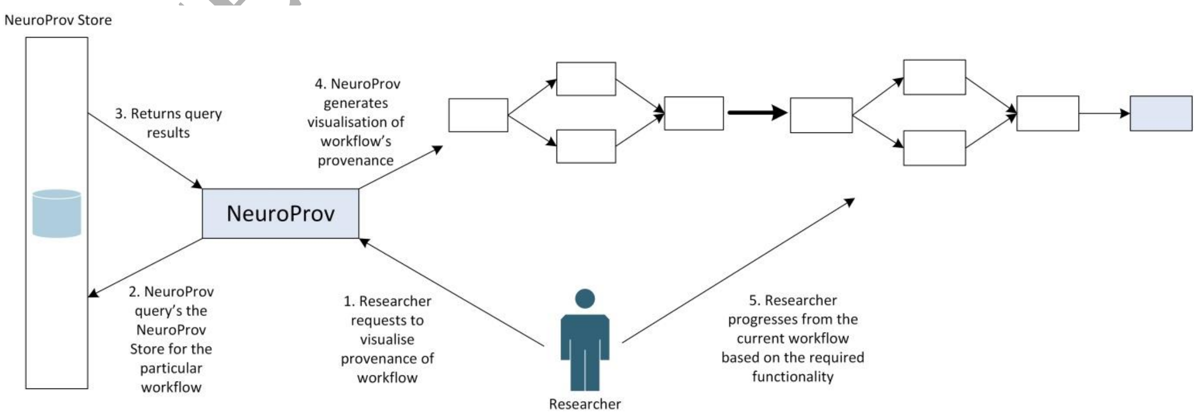

Figure 5 - Use-Case 4 Progression of Workflow

Req No.	Description	Priority
R4.1	Analysis of origin of results, to see how a workflow/dataset came into being so further analysis can be conducted upon it.	E
R4.2	Annotate the workflow/dataset with appropriate information to progress with the researcher's desired analysis.	E

Based on the requirements analysis as detailed in Section 3 and the Use-Cases presented, the following set of assessment metrics has been defined that are used to evaluate the results generated from using NeuroProv to visualise provenance data for neuroimaging analysis:

– Display workflows as nodes, edges and relationships with annotations and execution timeline;
– Compare provenance for given workflow(s);
– Highlight the trace for a particular artefact or process;
– Provide the ability for users to drill down and tracking of expanded stages;
– Examine a workflows' or an artefact's/process's attributes
– Enable search features e.g. to identify workflows, datasets, nodes etc.;
– View workflows in a separate view for Progression and Evolution use-cases.

5. NeuroProv Architecture

As NeuroProv System is developed in the context of N4U, Figure 6 illustrates how NeuroProv fits into the context of N4U project. It attempts to fulfill the use cases described in Section 4. The user selects data and the respective workflow to run and sends the request to NeuroProv, which then queries the NeuroProv Store containing data such as clinical variables, images, pipelines, project data etc. These are returned as a queried response to NeuroProv, which then enables to generate a visualisation to aid the researchers in the analyses.

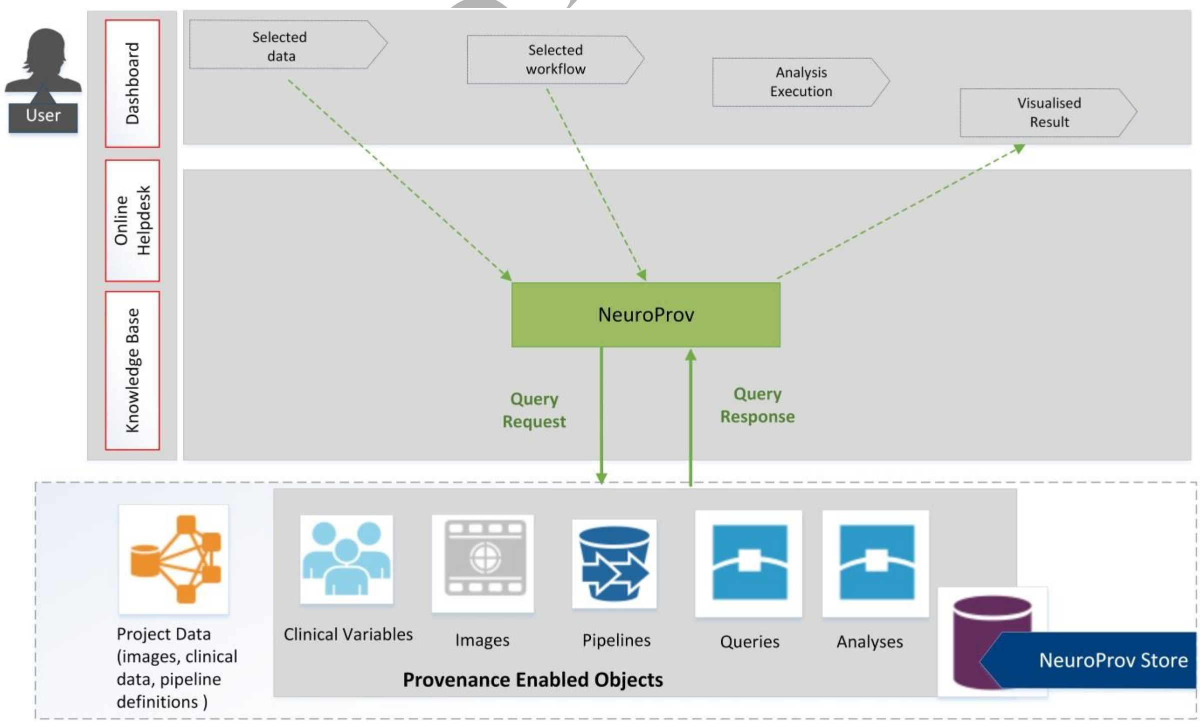

Figure 6 - NeuroProv in context of N4U project

Figure 7 presents the system architecture of NeuroProv, a visualisation system that has been developed based on the above requirements (Section 3 and Section 4). The NeuroProv architecture contains three following basic elements:

1. **NeuroProv Store**: The NeuroProv Store is a repository of MRI scans/images, associated metadata, workflow execution information, datasets and related provenance data of the above-mentioned items. Currently, the NeuroProv store contains provenance data and associated metadata from workflows being run in Pegasus [36].

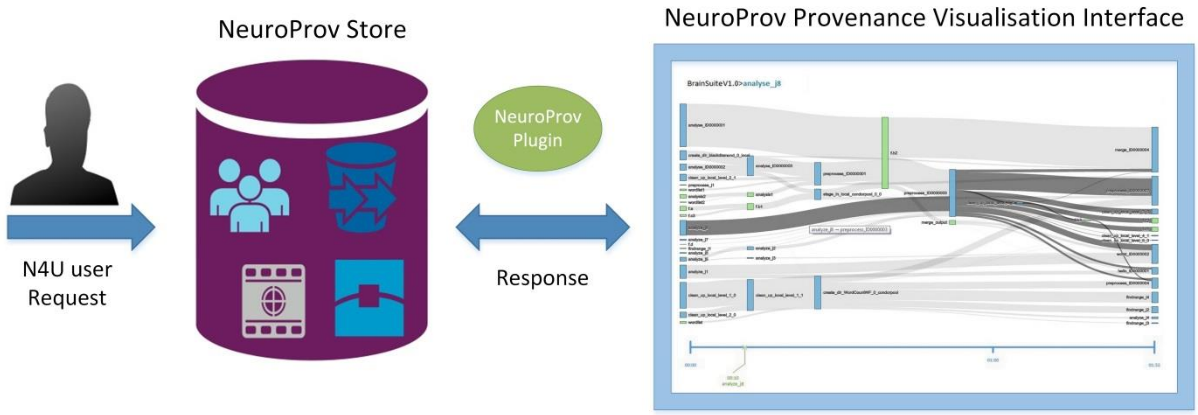

Figure 7 - NeuroProv Architecture

2. **NeuroProv Plugin**: Sankey diagrams [33] are particularly useful to represent the flow of data and to present temporal ordering inherent in provenance graphs. One of the earliest and famous examples of Sankey diagrams is Charles Minard's Map of Napoleon's Russian Campaign of 1812 [37]. Sankey diagrams enable researchers to view provenance as a network of activities where data flows through and between activities. Providing clinicians and researchers alike with a view that enables an understanding of how data flows through a selected activity/entity and its lineage to verify sources and identify choke points/anomalies. Other layout approaches for example 'radial layouts' tend to focus on the interconnectivity of data or activities and do not leverage the temporal ordering inherent to provenance [32]. The plugin ingests provenance data sent to it as an input in a JSON format and uses JavaScript to generate Sankey diagram based on the provenance data, metadata and annotations (if stored in the NeuroProv store).

3. **Provenance Visualisation**: Since D3 [38] is a JavaScript-based library it allows visualised Sankey diagrams to be presented in browsers. NeuroProv has been evaluated on Firefox (version 64) and Google Chrome (version 71) for the purpose of the research study. NeuroProv supports all major browsers except for Internet Explorer since d3.js lacks support for it.

Figure 8 shows the flow of activities in NeuroProv, the user generates a request to visualise provenance for a named workflow and sends the request to NeuroProv. NeuroProv plugin generates the appropriate query in order to retrieve provenance, metadata information and annotations associated with the workflow requested by the user. The query results from the NeuroProv Store are returned to the NeuroProv plugin in a JSON response. NeuroProv plugin then utilises the JSON response to generate Sankey diagrams to be visualised in the browser, thus providing users with visualisation to perform the analysis.

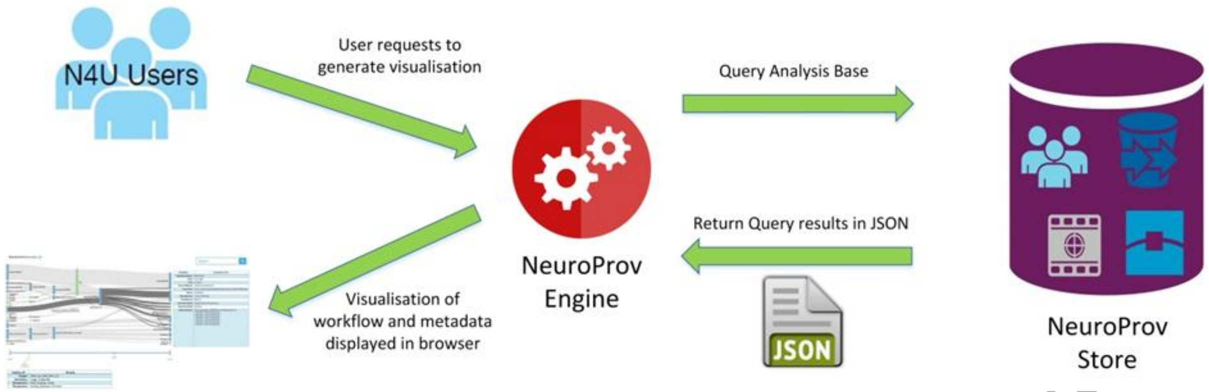

Figure 8 - Flow of activities in NeuroProv

Sankey diagrams have been particularly chosen to visualise provenance data for NeuroProv since neuroimaging provenance is viewed as a network of activities where data flows through and between activities. The aim is to provide a view that allows researchers to understand how data flows through a selected activity/entity and its lineage to verify sources and identify choke points/anomalies. Sankey diagrams enable users to highlight edges between different nodes by varying the link opacity. Without link opacity, the meaning of these relationships will be obscured. It is significant for clinicians to be able to identify relationships between different elements of the workflow. Furthermore, NeuroProv allows Sankey diagrams to be interactive, thus enabling users to move around nodes to better understand the relationship between important nodes.

NeuroProv visually represents activities as blue rectangles and entities as green rectangles. The width of the link between the rectangles represents the amount of information flow between the two nodes (activities and entities). Initially, once the visualisation has been generated, NeuroProv provides users with a basic view of the workflow provenance with each link's opacity set to 0.2. The opacity of 0.2 makes the links easily identifiable in the workflow. Once the user selects a particular node for inspection (be it an entity or an activity) by clicking on a node the concerned link(s) opacity changes to 0.5 while the opacity of the rest of the link remains at 0.2. This helps the users to visually differentiate between the highlighted and non-highlighted links in the visualisation; making it visually possible to differentiate and examine sources and/or target nodes.

NeuroProv's visualisation provides the ability to drill down into a workflow to view complete and detailed information allowing users to view a detailed summary of the workflow in the first stage and the user can progressively drill down as they proceed to view further details. The high-level summary provides naïve users with a basic level of understanding of the specific workflow. Experienced users such as Research leaders and Pipeline developers can further drill down to view detailed information that will help to completely understand provenance data. NeuroProv also provides tracking of the expanded stages to users, giving them an overview of what stage they have drilled down to so that they can click on previous stages when and if required to view a high-level summary of the workflow.

Additionally, the 'search' feature in NeuroProv allows users to discover information about a particular workflow or dataset to be administered whilst verifying an experiment. This feature enhances the ability of the user by providing the capability to look-up the required workflow while verifying the result. The request is sent as a query to the NeuroProv store which returns the appropriate query results to generate a visualisation of the searched workflow. This sets the scene for the following Section in which the evaluation results of the research conducted are presented along with the discussion on how the results are deduced.

6. Evaluation, Results and Discussion

For the purpose of the research study, Windows 7 machine with 4 GB RAM has been used. The visualised provenance is evaluated on Firefox (Version 64) and Google Chrome (Version 71). There are currently 131 workflows residing in the NeuroProv store for the purpose of the research study. Twenty workflows of varying size and data were selected, each containing multiple processes and images. Since it was not possible to obtain a truly random distribution within a huge set of workflows, the following strategy has been adopted for the selection of workflows. These workflows are divided into case-studies based on the type of use-cases namely verification, comparison, progression and evolution. The following table maps use-cases with the selected workflows for our research. For simplicity of understanding of the paper, we have presented Case-Study One for each of these use-cases. Each cell in the table represents the workflow id (wf_id) for the workflows visualised for that use-case in the case-study.

Table 1 - Describing Case Study and Use Case Mapping

Case-Study and Use-Case Mapping	Use-Case 1 Verification	Use-Case 2 Comparison	Use-Case 3 Progression	Use-Case 4 Evolution
Case- Study 1	Wf_id: 39, 94	Wf_id: 38, 39	Wf_id: 94, 96, 105, 106, 110, 123, 131	Wf_id: 94, 95, 96, 104, 105, 105, 110, 111, 113,123, 131
Case-Study 2	Wf_id: 1, 90	Wf_id: 94, 96, 131	Wf_id: 8, 14, 25, 29, 31	Wf_id: 36, 38, 39, 41, 44, 46
Case-Study 3	Wf_id: 91, 93	Wf_id: 2, 7, 13	Wf_id: 38, 39, 44, 46, 49	Wf_id: 9, 11, 17, 19, 21
Case-Study 4	Wf_id: 47, 53	Wf_id: 67, 71	Wf_id: 90, 91, 93	Wf_id: 68, 75, 77, 78, 81

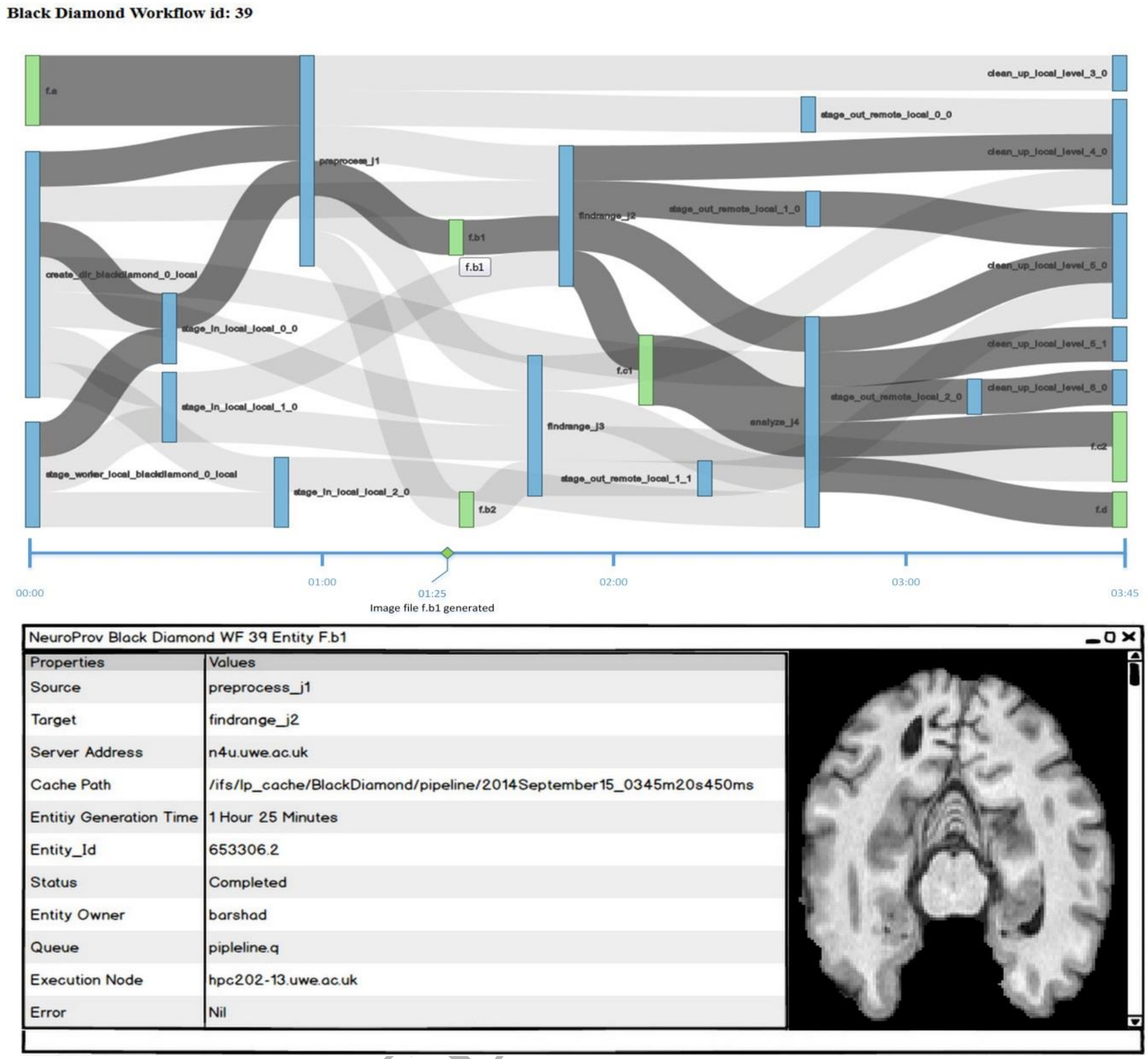

Figure 9 - Workflow Verification and Annotation

Use-Case 2 - Comparison: Insight often comes from comparing different provenance visualisations. Users can perform a comparison with the same attributes using different workflows to compare the results. This will provide further insights into the experiment. In figure 10, the first case-study from the use-case 'Comparison of Workflows', NeuroProv plugin generates a visualisation that compares two workflows (wf_id 38 with 39) and the differences are highlighted in red prompting users with the changes in the two versions of the workflow. Visually comprehending the differences can enable researchers to investigate the cause of the error and rectify it before it can percolate to a later stage and affect the final results. In Figure 10 NeuroProv highlights that the difference between workflow 38 and 39 is due to the generation/consumption of three files namely 'f.c1' (generated at two hours and seven minutes of execution), while 'f.c2' and 'f.d' (generated at three hours and forty four minutes of execution respectively) that caused changes in the workflow behaviour. In workflow 38 the files f.b1 and f.b2 (shown green in figure 10) were generated as a result of the workflow execution towards the end of execution around three hours and forty-five minutes. While similar files were generated around one hour, twenty-five minutes and the other at one hour and thirty minutes approximately in workflow 39 as can be seen on the execution timeline underneath the Sankey diagram.

Advanced users can annotate the workflow with the reasons that brought the change to help future users save valuable time and to ensure the validity of the changes that occurred. Furthermore, annotations can provide a basis for authentication to publish results in a journal/conference. NeuroProv plugin allows multiple workflow comparisons in addition to the case study provided where advanced users can compare more than two workflows against a single workflow in order to find anomalies and to highlight the changes occurred. This is a unique and novel feature in the NeuroProv system.

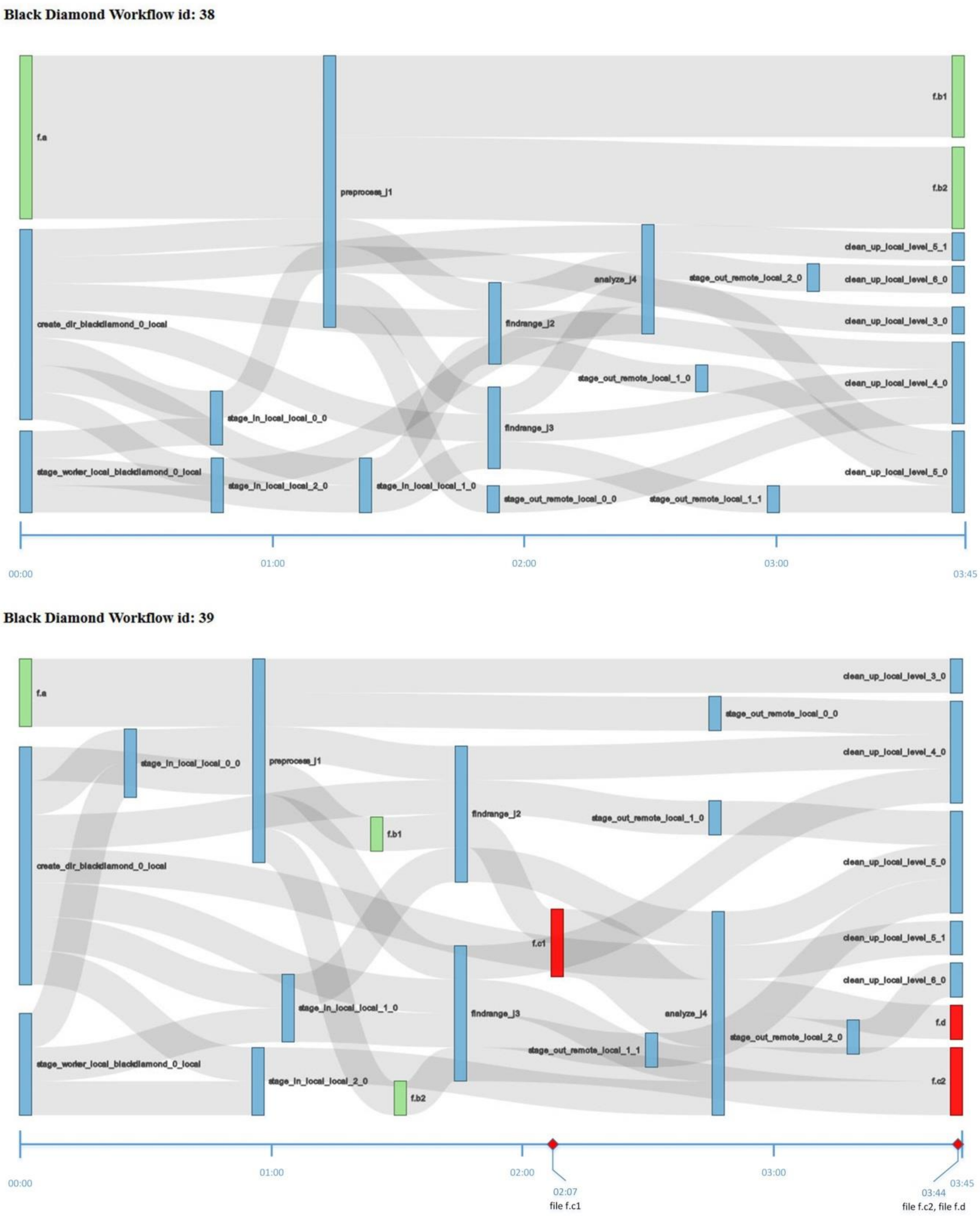

Figure 10 - Workflow Comparison

Use-Case 3 - Progression: Researchers can view the visualised provenance in order to determine the analysis and origin of results. In a collaborative environment such as neuGRID scientists frequently work with data that has been collected or processed by other groups or organizations. In order to verify the results for correctness, scientists need to view the progression of the data with the help of visualisation. Figure 11 shows the results from the first case-study from the use-case 'Progression of Workflows' for the 'WordCount' workflow. WordCount workflow contains file names for MRI scans. Over a period of time, the 'WordCount' workflow has progressed into several versions, each node representing a different version of the workflow. When a user clicks on a particular version of the workflow, it opens up a complete visualisation of the selected workflow version in a separate window allowing users to verify the workflow and its results. For instance, in the following scenario, if the user clicks on a link between 'WordCountWF 2.0' and 'WordCountWF 2.1' in Figure 11, the link is highlighted and the changes that led to the change in its version is shown in a table underneath the Sankey diagram. In this scenario, the changes brought about the consumption of file ‘wordlist’ and the generation of files ‘wordlist1’ and ‘wordlist2’ respectively. In the progression view of NeuroProv clicking on the nodes enables users to inspect complete visualisations of the workflow versions in separate windows while clicking on a link between the nodes provides users with the ability to highlight the changes that caused the new version to emerge. The ability to be able to visualise the progression of a workflow enables researchers to inspect changes that led to the generation of a particular version so that they can perform further experimentation on the existing version rather than starting from scratch for a similar version.

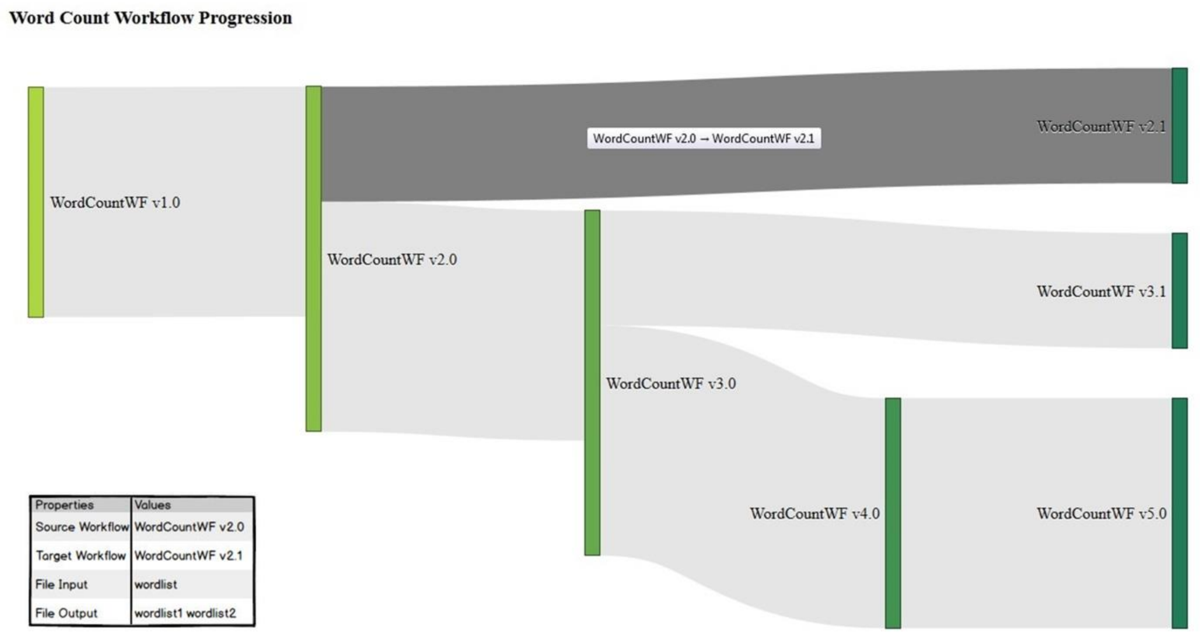

Figure 11 - Workflow Progression and Annotation

Use-Case 4 - Evolution: In order to determine how a certain data product has evolved during the course of an experiment, researchers need to view the visualised provenance. Clicking on a link provides users with a visualisation that shows how the workflow has evolved from one version to another, whilst clicking on a workflow version allows users to inspect the workflow in a separate window. Figure 12 shows the first case-study from the ‘Evolution of Workflows’ use-case showing the ‘BrainSuite’ workflow into multiple versions and stemming to become completely new workflows based on their functionality. If the user clicks on the link between ‘BrainSuitev1.0’ workflow and ‘BLASTv1.0’, NeuroProv highlights the trace between them and shows the changes that led to the transformation of ‘BrainSuite’ workflow to the ‘BLAST’ workflow at the bottom of the screen. Furthermore, users can view the visualised provenance of individual workflows such as ‘Dirac v1.0’, ‘PCARegistrationv1.0’ and ‘BrainSuiteCorticalSurfaceExtractionv1.0’ workflows in separate windows to determine the activities and entities involved that led to the generation of results.

In summary, for all of the above-mentioned provenance visualisation use cases, NeuroProv allows performing various provenance exploration activities such as:

1. Generate complete visualisation of a workflow, along with presenting annotations associated with the workflow and its elements;
2. Allow users to highlight a trace for a particular entity/activity, along with its sources and targets (which is essential to understand how an entity is generated and by which activity(-ies));
3. Allow users to drill down (provide high-level summary view and to avoid visual clutter) and keep track of expanded stages (this gives the user an overview of how farther they have drilled down in a workflow);
4. Have annotation support with workflows, entities, activities and links (for a better understanding of provenance data for the naive user);

5. Navigate a timeline to provide users with the ability to see when a particular activity/entity was generated and its associated details;
6. Have the ability to compare workflows and inspect visually encoded differences to allow users to identify anomalies;
7. Visualise how a workflow has evolved and progressed over the period of time and what changes have occurred.

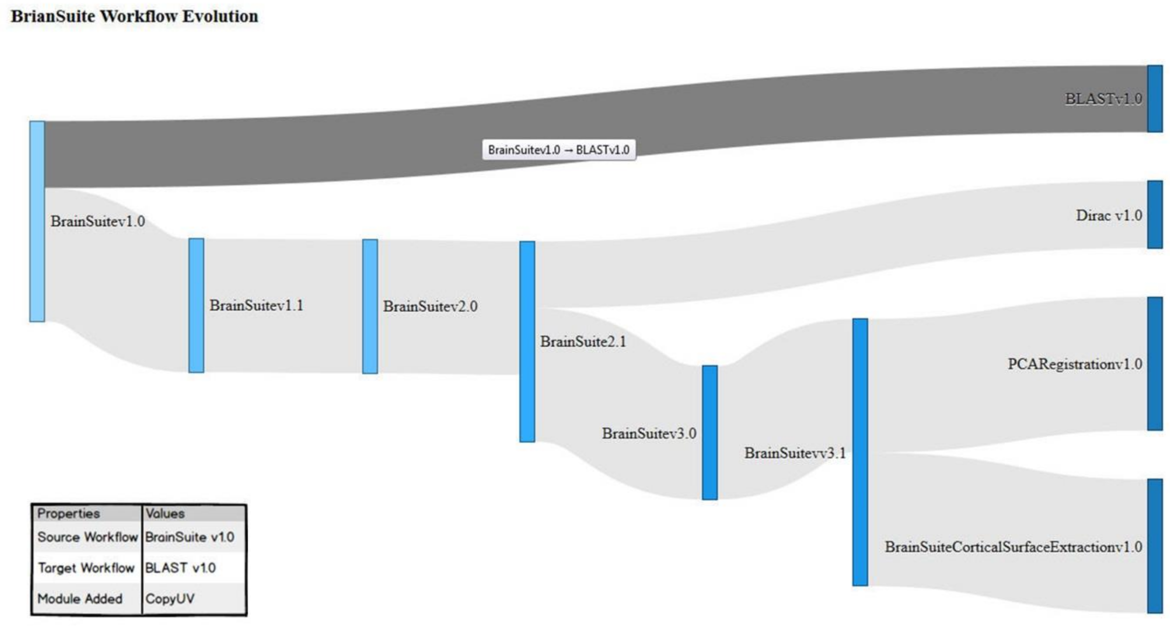

Figure 12 - Workflow Evolution

The evaluation approach taken in this research to derive results is qualitative in nature rather than quantitative. For visualisation aspects, there is no absolute measure to determine whether the approach followed works in all cases. This is a natural consequence of any visualisation exercise; measures to determine whether visualisation software is 'fit-for-purpose' are always subjective [39]. Table 2 maps the metrics to the use-cases for traceability of results.

Table 2- Mapping Metrics to Use-Cases

Metrics	Use-Case 1	Use-Case 2	Use-Case 3	Use-Case 4
Display workflows as nodes, edges and relationships with annotations and execution timeline;	✓	✓	✓	✓
Compare provenance for given workflow(s);		✓		
Highlight the trace for a particular artefact or process;	✓	✓	✓	✓
Provide the ability for users to drill down and tracking of expanded stages;	✓	✓	✓	✓
Examine a workflows' or an artefact's/process's attributes;	✓	✓		
Enable search features e.g. to identify workflows, datasets, nodes etc;	✓	✓	✓	✓
View workflows in a separate view for Progression and Evolution use-cases.			✓	✓

Moreover, in order to do a comparative analysis with other provenance visualisation approaches we have conducted thorough literature survey and in this regard, Table 3 provides a summary of how NeuroProv compares to other visualisation systems based on the set of metrics defined. Other visualisation approaches do not leverage the ability to highlight trace for a particular entity/activity and its associated sources/targets essential to verify results. The highlighted trace provides users with the ability to visually access the source, target activities and processes involved for a particular node. Systems such as Prototype Lineage Server [2] allow users to navigate metadata through clickable links bringing the focus of the item under inspection as shown in Figure 13. Views are restricted to only parent and child objects, in contrast, NeuroProv allows the user to view drill down multiple levels of details to view fine-grained provenance. Prototype Lineage Server is designed for Earth and Life Sciences data and caters to systems in similar domains.

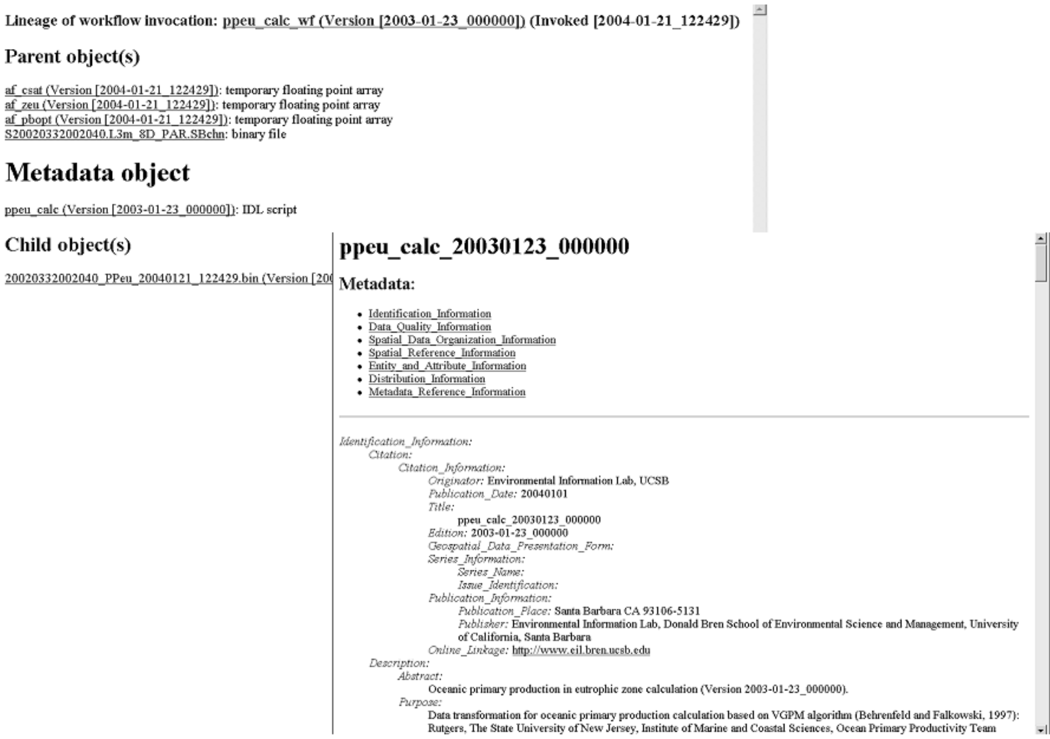

Figure 13 - Example screens for Lineage Server Web Application (Prototype Lineage Server) [2]

Furthermore fewer systems namely Provenance Explorer, Pedigree Graph and Probe-It! provide users with the ability to drill down and view fine-grained provenance meaningful for advanced users to inspect elements of provenance to determine the authenticity of results. Provenance Explorer [31] for instance shown in Figure 14, provides a Graphical User Interface (GUI) allowing permitted users based on their access privileges to drill down and to expand links between nodes (input states, processes and output states) to expose fine-grained information about particular sub-events or intermediate products similar to NeuroProv. One of the major drawbacks of Provenance Explorer is that the underlying model for visualising provenance and inference rules defined are specifically for processing events in a laboratory or manufacturing/processing plant. This is very different from neuroimaging analysis use-cases, in which workflows run on images and associated data.

Table 3 - NeuroProv Vs other Provenance Visualisation Systems

Metrics	Prototype Lineage Server	myGrid	VisTrails	Probe-It!	Pedigree Graph	PROV-O-Viz	Provenance Explorer	ESSW	Karma	CI-Browse-It!	NeuroProv
Display complete start to end provenance for a workflow	●	●	●	●	●	●	●	●	●	●	●
Allow users to compare two workflows				●							●
Allow users to compare multiple workflows											●
Display workflows as nodes, edges and relationships		●	●	●		●	●	●	●	●	●
Highlight trace for a particular data or process			●								●
The ability for users to drill down			●	●	●		●				●
Provide annotation with workflow, nodes and edges	●	●	●				●	●	●	●	●
Examine object history	●		●	●	●		●		●	●	●
Provide tracking of expanded stages											●
Search feature e.g. workflow, nodes, dataset			●	●							●
Visually view attributes (visually encoded)	●	●	●		●		●	●		●	●
Workflow execution timeline			●						●		●

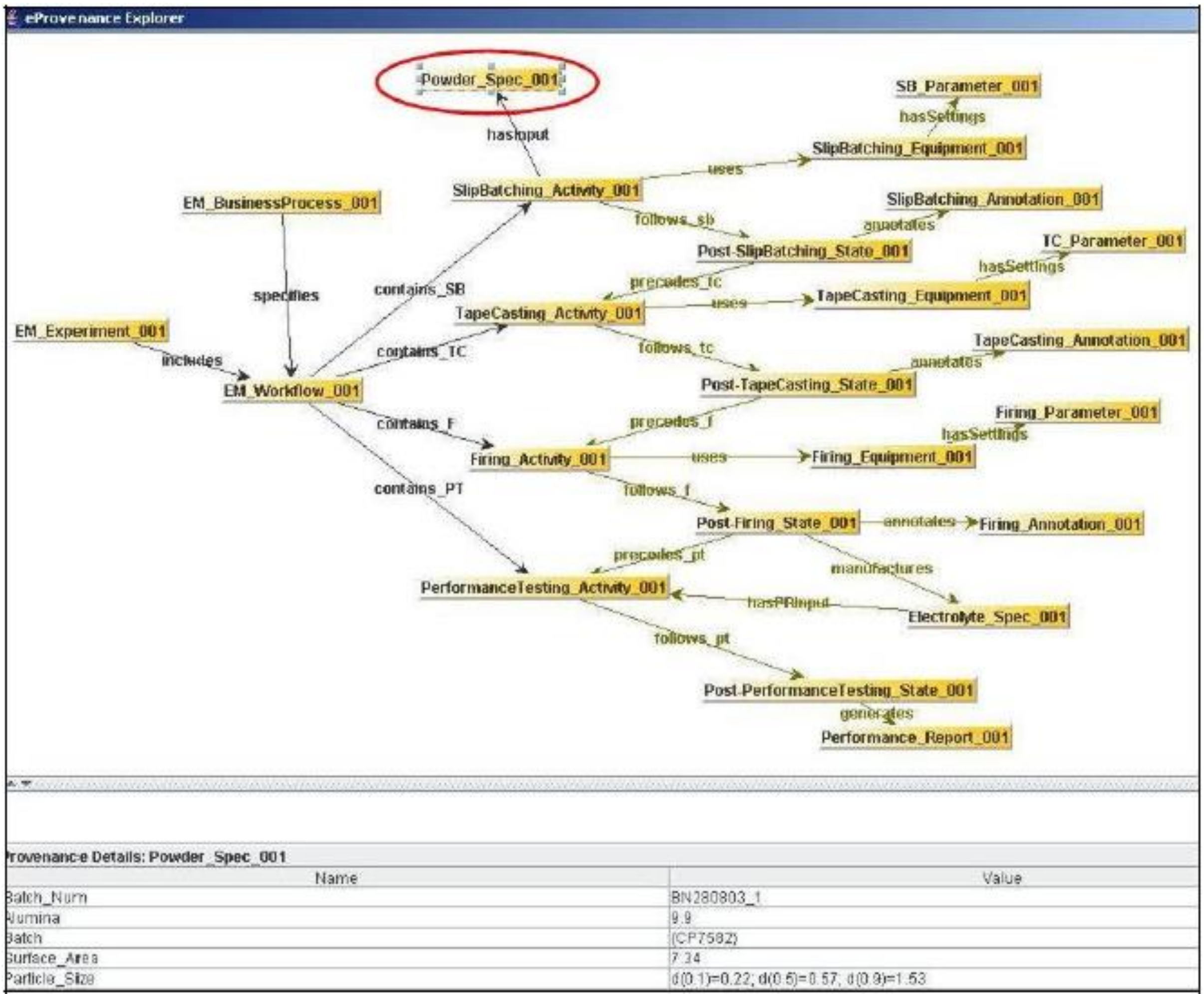

Figure 14 - Snapshot of Provenance Explorer's expanded provenance view [31]

In addition systems like VisTrails [30] shown in Figure 15, is a workflow and provenance management system that provides for scientific exploration and visualisation. VisTrails records modifications applied to the workflow while users are editing. In the context of this system, provenance refers to the history of changes made to a particular workflow in order to derive a new workflow; changes may include, adding, deleting or altering workflow processes. VisTrails provides a novel way to render this history of changes. A tree-like structure provides a representation for provenance where nodes represent a version of some workflow while edges represent changes applied to a workflow in order to derive a new workflow. However, VisTrails generates visualisation with the help of provenance data rather than visualising the provenance data. Although VisTrails provides the ability to compare different versions of the workflow or even a data product over the course of an experiment it fails to generate a visualisation of provenance data itself. It only provides associated provenance in a textual form which is of little use for neuroimaging since it is difficult to make sense of the sheer volume of provenance data associated with an analysis.

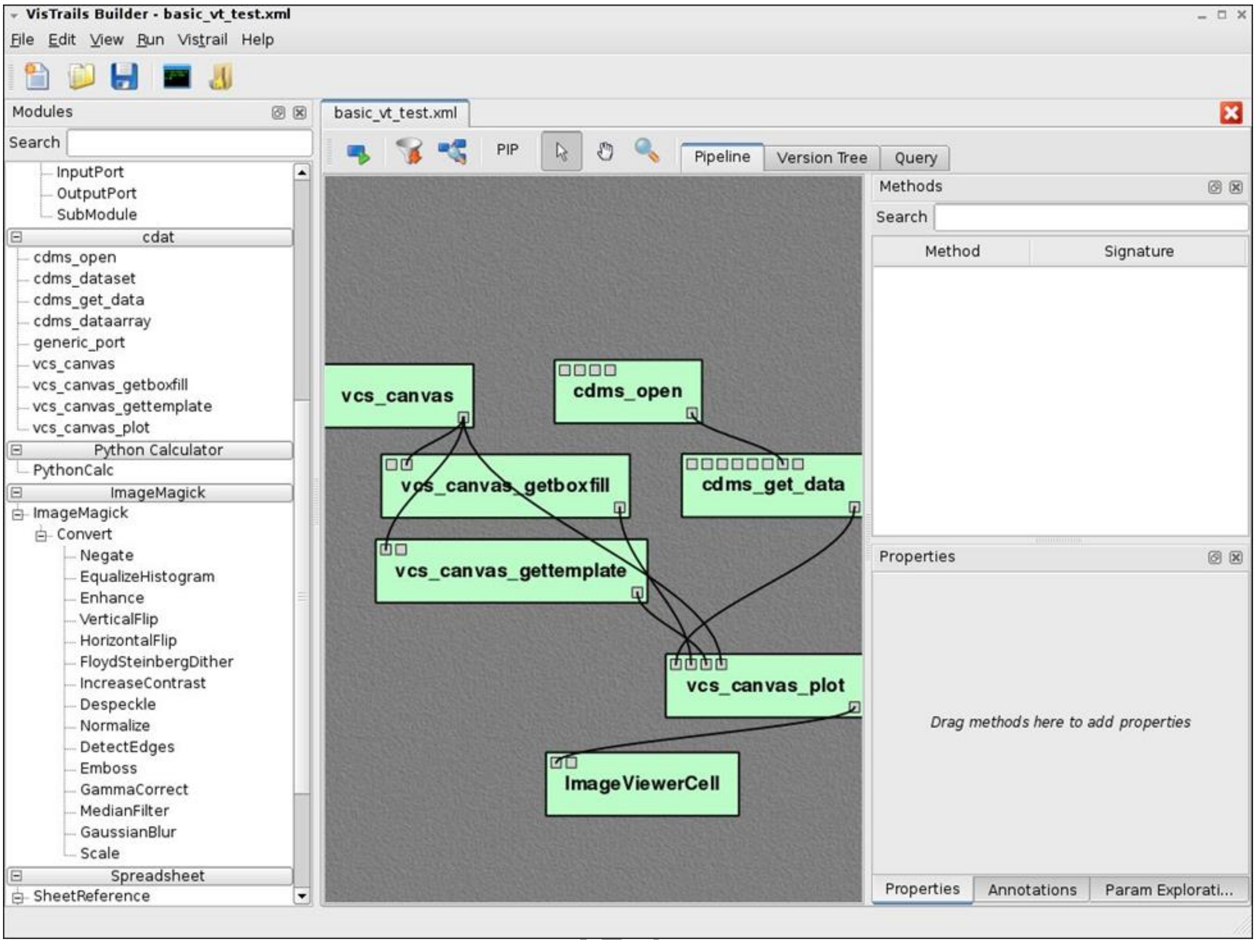

Figure 15 - VisTrails Screenshot of the main window [30]

Execution timeline provided by NeuroProv allows users to determine the causality between entities and activities to ensure the correctness of results. The ability to compare two or more workflows enables users to visually evaluate differences between multiple workflows and determine any anomalies. The next section concludes the research and identifies potential future directions for the research work.

7. Conclusions and Future Directions

This research concludes that visualisation techniques can enhance the utility of provenance data for neuroimaging analyses. Scientists and researchers need to visualise provenance data in order to aid them in the exploration process by providing means to understand complex data and processes. NeuroProv allows clinical researchers to visually represent provenance data in an intuitive manner thus allowing clinicians and users to exploit the true potential of provenance data. The use of Sankey diagrams for representing provenance data for neuroimaging analysis opens up new avenues for using such techniques for visualisation of provenance data, thus broadening the domain of data visualisation.

NeuroProv provides researchers both with a high-level summary of the analysis and the ability to drill down to examine details that might be helpful whilst verifying an analysis. A high-level summary might be effective when a 'Basic User' would want to inspect the provenance for understanding and authenticating a workflow or result. 'Intermediate' and 'Advanced Users' may wish to inspect detailed visualisation of provenance to determine errors/anomalies.

Over the course of the study, requirements for provenance visualisation have been defined (as elicited in Sections 3 and 4), derived from the use-cases designed for neuroimaging analysis based on the case study N4U. The use-cases encompass aspects of provenance visualisation for the domain of neuroimaging analysis. Particular focus has been made on the usage of provenance data in order to determine the requirements. A detailed study of the state-of-the-art reveals that current visualisation systems partially lack support for provenance visualisation. Since provenance of neuroimaging analyses involves images, datasets, pipelines, algorithms and arguments it is essential that provenance is completely and correctly visualised in order to verify a result or to reproduce an experiment. One of the contributing factors of this work is to allow researchers and clinicians to be aware of the requirements for a visualisation system for this domain. The other main contribution is the results of the qualitative study, devising metrics to evaluate visualisation for provenance data.

One major future direction for the continuation of this work is generalising the system to perform in other domains. In this regard, the results from this research could be validated externally to assess whether they can be generalised for other domains (in addition to neuroimaging related scientific analysis). The current research takes neuGRID and N4U as a case study but NeuroProv could be modified in future in order to accommodate other provenance systems to visualise data. Another future direction can be to research and integrate, within NeuroProv, a module that will enable applications to learn from their past executions and improve and optimise new studies and processes based on previous executions. This will include models that will inform researchers of missing processing stages, suggest available and verified processing modules and warn users of incompatible data types. One essential future direction could be provenance interoperability. Currently NeuroProv uses PROV [25] interoperability standard. This will allow N4U users to share their provenance data in other PROV-compliant systems.

Another possible addition to the system could be a readily searchable database of commonly used (and also rarely used) workflows that will greatly aid researchers in re-creating the conditions of a particular analysis, reproducing previous results and re-running a selected analysis with limited modification. Models could be formulated that could derive the best possible optimisation strategies by learning from past executions of experiments and processes. These models would gradually evolve over time and would facilitate decision support in generating a visualisation of future processes and workflows in a domain. Furthermore, another future direction can be to develop useful applications based on the work presented in this research, gaining access to real-world infrastructures for system deployment and engaging users in using provenance to inform neuroimaging research, encouraging subsequent data and provenance sharing, enhanced peer-reviewed publications and supporting multi-centre collaboration.

Acknowledgements

This work is funded by the EU 7th Framework Programme under the N4U project (reference 283562).

References

[1] Y. Gil, E. Deelman and M. Ellisiman , “Examining the challenges of scientific workflows,” *Computer,* vol. 40, no. 12, pp. 24-32, 2007.

[2] R. Bose and J. Frew, “Composing Lineage Metadata with XML for Custom Satellite-derived Data Products,” in *Scientific and Statistical Database Management,* 2004.

[3] R. McClatchey, J. Shamdasani, A. Branson, K. Munir, Z. Kovacs and G. Frisoni, “Traceability and provenance in big data medical systems,” in *IEEE 28th International Symposium on Computer-Based Medical Systems,* Sao Carlos and Ribeirao Perto - Brazil, 2015.

[4] A. S. Fleisher, W. S. Houston, L. T. Eyler, S. Frye, C. Jenkins, L. J. Thal and M. W. Bondi, “Identification of Alzheimer disease risk by functional magnetic resonance imaging,” *Archieves of Neurology,* vol. 62, no. 12, pp. 1881-1888, 2005.

[5] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat and P. Couch, “Why

linked data is not enough for scientists," *Future Generation Computer Systems,* vol. 29, no. 2, pp. 599-611, 2013.

[6] K. Munir, K. H. Ahmad and R. McClatchey, "Development of a large-scale neuroimages and clinical variables data atlas in the neuGRID4You (N4U) project.," *Journal of biomedical informatics,* vol. 57, pp. 245-262, 2015.

[7] A. J. MacKenzie-Graham, J. D. Van Horn, R. P. Woods, K. L. Crawford and A. W. Toga, "Provenance in neuroimaging," *Neuroimage,* vol. 42, no. 1, pp. 178-195, 2008.

[8] L. Moreau, B. Ludäscher, I. Altintas, R. S. Barga, S. Bowers, S. Callahan and G. Chin Jr , "Special issue: The first provenance challenge," *Concurrency and computation: practice and experience,* vol. 20, no. 5, pp. 409-418, 2008.

[9] Y. Zhao, M. Wilde and I. Foster, "Applying the virtual data provenance model.," *International Provenance and Annotation Workshop,* pp. 148-161, 2006.

[10] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, . W. Jagust, J. Q. Trojanowski, A. W. Toga and L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI).," *Alzheimer's & Dementia,* vol. 1, no. 1, pp. 55-66, 2005.

[11] H. Rusinek, S. D. Santi, . D. Frid, W.-H. Tsui, C. Y. Tarshish, A. Convit and M. J. de Leon. , "Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging," *Radiology,* vol. 229, no. 3, pp. 691-696, 2003.

[12] A. Dolgert, L. Gibbons, C. D. Jones, V. Kuznetsov, M. Riedewald, D. Riley, G. J. Sharp and P. Wittich, "Provenance in High-Energy Physics Workflows," *Computing in Science and Engineering,* vol. 10, no. 3, pp. 22-29, 2008.

[13] E. Deelman, D. Gannon, M. Shields and I. Taylor, "Workflows and e-Science: an overview of workflow system features and capabilities," *Future Generations Computer Systems,* vol. 25, no. 5, pp. 528-540, May 2008.

[14] S. Dey, K. Belhajjame, D. Koop, M. Raul and B. Ludäscher, "Linking prospective and retrospective provenance in scripts.," in *Theory and Practice of Provenance (TaPP).,* Berkeley, CA, USA, 2015.

[15] T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, K. Bocinsky and Y. Cao, "YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts," in *arXiv preprint arXiv:1502.02403,* 2015.

[16] T. Miksa and A. Rauber, "Using ontologies for verification and validation of workflow-based experiments.," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 43, pp. 24-45, 2017.

[17] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma and E. Mina, "Using a suite of ontologies for preserving workflow-centric research objects.," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 32, pp. 16-42, 2015.

[18] Z. Yuan, D. H. Ton That, S. Kothari, G. Fils and T. Malik, "Utilizing Provenance in Reusable Research Objects," *Informatics,* vol. 5, no. 1, p. 14, 2018.

[19] P. Macko, D. Margo and M. Seltzer, "Local clustering in provenance graphs," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management,* 2013.

[20] D. El-Jaick, M. Mattoso and A. A. Lima, "SGProv: Summarization Mechanism for Multiple Provenance Graphs," *Journal of Information and Data Management,* vol. 5, no. 1, p. 16, 2014.

[21] Y. Tian, R. A. Hankins and J. M. Patel, “Efficient aggregation for graph summarization.,” in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008.

[22] S. Cohen, S. Cohen-Boulakia and S. Davidson, “Towards a model of provenance and user views in scientific workflows,” in *International Workshop on Data Integration in Life Sciences*, Berlin, Heidelberg,, 2006.

[23] L. Moreau and P. Missier, “PROV-DM: The PROV Data Model,” W3C, 30 April 2013. [Online]. Available: http://www.w3.org/TR/prov-dm/. [Accessed 2018].

[24] . V. Cuevas-Vicenttín, B. Ludäscher, P. Missier, K. Belhajjame, F. Chirigati, Y. Wei, D. Koop, S. Bowers, I. Altintas, C. Jones, M. B. Jones, L. Walker, P. Slaughter and B. Leinfelder, “Provone: A prov extension data model for scientific workflow provenance.,” 2015.

[25] P. Groth and L. Moreau, “An Overview of the PROV Family of Documents,” 2013. [Online]. Available: www.w3.org/TR/2013/NOTE-prov-overview-20130430/. [Accessed December 2018].

[26] N. Del Rio and P. P. Da Silva, “Probe-it! visualization support for provenance,” in *International Symposium on Visual Computing*, Berlin, Heidelberg, 2007.

[27] D. Quan, D. Huynh and D. R. Karger, “Haystack: A Platform for Authoring End User Semantic Web Applications,” in *International Semantic Web Conference (ISWC)*, Berlin, Heidelberg, 2003.

[28] P. Macko and M. Seltzer, “Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs,” in *TaPP*, 2011.

[29] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger and M. Greenwood, “Taverna: a tool for the composition and enactment of bioinformatics workflows,” *Bioinformatics*, vol. 20, no. 17, pp. 3045-3054, 2004.

[30] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger and H. T. Vo, “Managing Rapidly Evolving Scientific Workflows,” in *Provenance and Annotation of Data*, Berlin, Heidelberg, 2006.

[31] K. Cheung and J. Hunter, “Provenance Explorer – Customized Provenance Views Using Semantic Inferencing,” in *5th International Semantic Web Conference (ISWC '06)*, 2006.

[32] C. Kwok and J. Hunter, “PROV-O-Viz Understanding the Role of Activities in Provenance,” in *International Provenance and Annotation Workshop (IPAW)*, Cologne, Germany, 2014.

[33] M. Bostock, “Sankey Diagrams,” 2012. [Online]. Available: http://bost.ocks.org/mike/sankey/ . [Accessed November 2018].

[34] C. Goble, “Position Statement: Musing on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics,” in *Workshop on Data Derivation and Provenance*, Chicago, 2002.

[35] “MoSCoW : Requirements Prioritization Technique,” 05 March 2013. [Online]. Available: http://businessanalystlearnings.com/ba-techniques/2013/3/5/moscow-technique-requirements-prioritization . [Accessed November 2018].

[36] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman and G. Mehta, “Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems,” *Scientific Programming*, vol. 13, no. 3, pp. 219-237, 2005.

[37] J. Corbett, Charles Joseph Minard, Mapping Napoleon's March, 1861. CSISS Classics, 2001.

[38] M. Bostock, “D3 - Data Driven Documents,” [Online]. Available: https://d3js.org/.

[39] M. Tory and S. Staub-French, “Qualitative Analysis of Visualization: A Building Design Field Study,” in *BELIV Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*.

[40] M. Kunde, H. Bergmeyer and A. Schreiber, “Requirements for a Provenance Visualisation Component,” in *International Provenance and Annotation Workshop*, Berlin Heiderlberg, 2008.

[41] D. H. Ton That, G. Fils, . Z. Yuan and T. Malik, “Sciunits: Reusable Research Objects,” in *Proceedings of the IEEE eScience*, Auckland, New Zealand , 2017.

[42] Y. Simman, B. Plale and D. Gannon, “A framework for Collecting Provenance in Data-Centric Scientific Workflows,” in *International Conference on Web Services*, Chicago, 2006.

[43] E. G. Kehoe, J. P. McNulty, P. G. Mullins and A. L. Bokde, “Advances in MRI biomarkers for the diagnosis of Alzheimer's disease,” *Biomarkers in Medicine*, vol. 8, no. 9, pp. 1151-1169, 2014.