NeuroProv: A Visualisation System to enhance the utility of Provenance Data for Neuroimaging Analysis

Bilal Arshad

A dissertation submitted in partial fulfilment of the requirements of the University of the West of England, Bristol

For the degree of Master of Philosophy

Faculty of Environment and Technology, University of the West of England

October 2015

To my beloved Parents, Wife,
Siblings and Son!

Acknowledgment

I would like to acknowledge many people without whom I wouldn't be able to complete this MPhil. First of all, I would like to thank Almighty Allah who has been Gracious and Merciful upon me throughout my life.

I really feel indebted to my supervision team which includes Richard McClatchey (Director of Studies), Kamran Munir (Co-Supervisor), Jetendr Shamdasani (Co-Supervisor) and Saad Liaquat Kiani (previous Co-Supervisor). Without their guidance, it would not have been possible for me to do this work. I am grateful to Richard McClatchey who has been supportive in the funding arrangements and the university administrative procedures throughout this research work. His kind words and suggestions have always been there to encourage me. I would like to appreciate Kamran Munir for being tolerant, and always helpful. His guidance in general and technical suggestions in particular really helped me during the course of the study. I am thankful to Jetendr Shamdasani who provided guidance and suggestions for this research study. I would also like to express my heartfelt gratitude to Saad Liaquat for his guidance and suggestions in the beginning of the research studies.

Apart from my supervision team, I would like to highlight and appreciate efforts and support provided by the people at UWE. I would like to thank Zaheer Abbas Khan, Khawar Ahmad, Kamran Somroo and Rawad Hammad who have always guided me with their suggestions throughout this research work.

I am extremely grateful to my father Arshad Ali, my mother Rizwana Arshad, my beloved wife Humna Asif and my siblings for their love, care and support. Without their support, it would have been very difficult for me to complete this work. I would like to take the opportunity to thank my son Ashar for the love and support for his presence and absence during the course of the research.

At the end I would like to acknowledge the support provided by European Commission in funding this work. I would like to acknowledge the support I have received from all my colleagues at University of the West of England (UWE).

Table of Contents

List of Figures

List of Tables

Abstract

E-Science platforms such as myGRID and NeuGRID for Users are growing at an amazing rate. One of the key barriers to their widespread use in practice is the lack of provenance data to support the reasoning and verification of experimental or analysis results. Clinical researchers use workflows to orchestrate the data present in e-science platforms in order to facilitate processing. Even though most systems capture provenance data and store it, systems rarely make use of it, thus limiting the exploitation of the true potential of such provenance. This thesis investigates mechanisms to visualise provenance data for neuroimaging analysis and to provide means to exploit the true potential of provenance data. In order to achieve this, a visualisation system has been implemented based on the use-cases that have been designed following requirements elicited for neuroimaging analysis. In this research a technique has been used to address the requirements of provenance visualisation for neuroimaging analysis. The prototype system has been tested against the provenance generated by NeuGRID for Users (N4U) as a proof of concept for our research. Different workflows have been visualised to study the efficacy of the proposed solution. Furthermore, evaluation metrics have been defined to determine whether the proposed solution is suitable for the purpose of the research conducted. The results show that the proposed visualisation system enhances the utility of provenance data for neuroimaging analysis and therefore the proposed research can be used to provide value to provenance data for neuroimaging analyses.

Contribution

This research study has been carried out in collaboration with NeuGrid for Users (N4U). The UWE's N4U research team and mainly my supervision team have helped me in understanding the internal workings of the N4U system with particular focus on provenance. I designed and developed the visualisation system for provenance data and generated results at UWE.

Chapter 1

Introduction

This chapter identifies the problem domain and describes the motivation behind the work carried out in this research. The research hypothesis, research questions and the methodology that will be used to evaluate the hypothesis in this thesis is also presented. At the end of this chapter, a structure of this thesis is outlined.

1.1 Background

The practice of representing information visually is not new. From students to scientists and analysts to politicians, all over the world researchers have been using data visualisation to track everything from DNA sampling to stock prices. According to Friedman [1], data visualisation's main goal is to communicate information clearly and effectively through graphical means. It does not mean that data visualisation needs to look primitive to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex dataset by communicating its key aspects in a more intuitive way.

Scientists rely on visualisations to aid in data exploration, which is often a complex process that requires close collaboration among domain scientists, computer scientists and visualisation experts. The ability to collaboratively explore data is one key to the scientific discovery process. The domain of neuroimaging analysis is no such exception. Neuroimaging is a crucial tool for both research and clinical neuroscience. In order to assist research into various neurodegenerative diseases such as Alzheimer's researchers require to process brain scans for various biomarkers. These biomarkers include the cortical thickness of the brain, thinning of which has been linked to the onset of Alzheimer's disease. A significant challenge in neuroimaging and in fact all biological sciences, concerns devising ways to manage the enormous amounts of data generated using current techniques. This challenge is compounded by the expansion of collaborative efforts in recent years and the necessity of not only sharing data across multiple sites, but making that data available and useful to the scientific community at large.

Scientific experiments such as CMS [2], DNA Analysis [3] or projects such as neuGRID [4] produce huge amounts of data. For instance CMS produces an estimated raw data of size up to 5

Peta Bytes (PB) per year [2]. Similarly neuGRID's initial setup [5] estimated generating 20 Tera-Bytes (TB) of neuroimages and associated provenance. This data is then processed and analysed using different tools and algorithms, such as neuroimaging algorithms in the case of neuGRID. These analyses require complex workflow orchestration, which is performed using workflows. A workflow can be defined as a collection of tasks orchestrated in order to achieve an overall goal of a workflow [6]. These tasks are arranged on the basis of their data or control dependency to perform complex analyses. These workflows enable researchers to collaboratively design, manage and obtain results that require hundreds of jobs accessing gigabytes of data. Since these workflows involve a huge amount of data, provenance storage is essential in order to understand the complex workflow, to verify an experimental result, to re-run an experiment etc.

Scientists use workflows in order to orchestrate the complex processing of data required for neuroimaging analysis. Clinical researchers and scientists need to understand results from workflows in order to accept them. In order to understand these results scientists require to inspect the associated *provenance*, which can be defined as the derivation of the history of an object [7]. Provenance information provides insight about sources and methods used to derive results that can increase the understanding and acceptance of results by scientists. Visualisation and provenance techniques, although used rarely in combination, may further help to increase a scientist's ability to understand results since the scientist may be able to use a single tool to evaluate final results, the derivation processes and any intermediate results produced during the experiment.

E-Science platforms are growing at an amazing rate; one of the major barriers is the lack of provenance support and its usage. There are systems such as Prototype Lineage Server [20] that can adequately capture provenance data and store it in one format or the other, but that is practically of no great benefit since few use it. The scientific community and in particular neuroimaging requires means to access and understand provenance data in order to support clinical analyses. This will help researchers in the study of MRI scans to determine biomarkers for the onset of Alzheimer's disease, as is being carried out in the N4U project. The domain of neuroimaging is complex including multiplicities of data-sets and versions of algorithms operating upon these data-sets. Since the analysis is repeatedly conducted in a collaborative research environment it is imperative to retain a track of who did what, when and for what purpose. All this information needs to be traced and logged so that the analyses can be reproduced or amended and repeated as part of a rigorous research process.

There are no major breakthroughs happening in neuroimaging 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; the chances are that lack of provenance support increases the risk of error. Since neuroimaging contains multiplicities of data; an error in an earlier step can percolate over to the next stages and 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 all together a different set of results compared to the anticipated results at the beginning of the analysis. Therefore the researcher is missing the essential elements of correct diagnosis of the results. This leads to the need to visualise provenance data to support clinical analyses in neuroimaging.

1.2 Motivation

With the increasingly digital nature of biomedical data and as the complexity of analyses in medical research increases, the need for accurate information capture, traceability and accessibility has been crucial to medical researchers in the pursuance of their research goals. For neuroscientific analyses, especially those centred on complex image analysis, the traceability of processes and datasets is essential for verification and reproducibility of results. Visualisation is a way of understanding scientific results obtained from complex workflows. In order to understand these results scientists require inspection of the associated provenance. Provenance information provides insight about sources and methods used to derive the results that can increase the understanding and acceptance of results. In other words, a researcher cannot know the reason behind a particular result unless they see a picture of the complete cause and effect. In fact, a researcher needs to visually see the combination of analysis output along with the associated provenance. This provides a researcher with the complete knowledge about a result and how it was generated. Clinical researchers and scientists can use provenance to identify event causality; enable broader forms of sharing, reuse and long-term preservation of scientific data; to attribute ownership and determine the quality of a particular dataset [8]. The coupling between scientific result and associated provenance is inherent, thus justifying the development of a system to facilitate the easy and intuitive viewing of both.

As an example of the research process in neuroimaging, consider a clinical researcher working in neuroscience using a research infrastructure such as neuGRID. The team of researchers includes a room full of doctors and a computer scientist. The leader of the group is a scientific director of neurology. The researcher needs to prove a particular hypothesis, perhaps concerning whether a particular range of cortical thicknesses might indicate Alzheimer's disease. The user selects multiple brain scans from the system and runs a particular workflow that is composed of various algorithms. This results in thousands of intermediary images which are created when data feeds through from algorithm to algorithm. The final result will be numerical values which can be

analysed using statistical techniques in order to prove or disprove a hypothesis. The researcher needs to verify that the experiment has been successfully carried out before the results can be published in a scientific journal. In this example, provenance data would have been acquired in order to keep track of the intermediary steps, to check whether each step has been performed as instructed and to validate it. Any errors that occurred will have been captured during the execution of the workflow. If at some point the researcher needs to go back and check how a particular step was performed or the conditions that were prevailing at the time of experiment, such data can be used. Provenance is therefore an essential element of clinical research infrastructures. Current visualisation systems partially or completely lack support for provenance visualisation thus there exists a need to develop a system that can address the requirements of visualising provenance for neuroimaging analysis.

1.3 Research Context

This research study is conducted within the context of N4U [9]. N4U or simply neuGRID for Users provides neuroscientists and clinician with the ability to perform high-throughput imaging research, and provide neurologists with automated diagnostic imaging markers for neurodegenerative diseases such as Alzheimer's for individual patient diagnosis. N4U 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 this project is to enable the discovery of biomarkers for Alzheimer's disease that will improve diagnosis and help speed the development of innovative drugs. Within the context of N4U, its end-user community has identified a vital need for provenance. Visualisation of provenance data will allow clinicians and researchers to understand and interpret results from these experiments and provide insight for future research. This leads to our ~~hypothesis and~~ research questions as mentioned in the next section that will drive our research work.

1.4 Research Hypothesis and Questions

The aim of this research is to assess and verify the following hypothesis:

'Visualisation techniques can enhance the utility of provenance data for neuroimaging analyses'

The research questions that will be explored and answered to evaluate the above hypothesis during this research study are given below:

1. What are the functional requirements for provenance visualisation in neuroimaging analysis?

2. Which visualisation technique(s) would be suitable for provenance visualisation based on the set of requirements identified?
3. What metrics will be used to evaluate the utility of provenance data for neuroimaging analysis?

1.5 Research Scope and Limitations

As discussed in the previous sections, the research study will attempt to achieve the following objectives:

1. Define requirements for effective representation of provenance data for neuroimaging analysis;
2. Devise representational technique to represent the visualised provenance, and
3. Visualise provenance for different classes of users based on requirements defined by use-cases.

The following is not part of this research study:

1. The security of the visualised provenance information is outside the scope of this study. It is assumed that the security mechanism (authorisation and authentication) provided by the underlined system is adequate for the purpose of this study.

As elaborated earlier, the specific aim of this research study is to evaluate the utility of provenance data being visualised for neuroimaging analyses. For this, a visualisation component will be developed based on its suitability and requirements for N4U [9] for evaluation purposes.

1.6 Research Methodology

In order to answer the hypothesis and research questions, a combination of qualitative and experimental research methodology [10] will be adopted i.e. a subset of the problem is investigated qualitatively through literature review and comparison with existing systems while another subset is analysed empirically through experiments. Further details on the methodology will be described in Chapter 4.

1.7 Thesis structure

- Chapter 2 – Background and Related Work

  This chapter will provide a survey of state-of-the-art projects related to provenance capturing approaches in general and visualisation projects in particular. It will then outline the shortcomings of the existing visualisation systems. This chapter will also highlight the need for provenance visualisation system for neuroimaging analysis.

- Chapter 3 – Proposed Research and Requirements Analysis

  This chapter will provide a detailed description of the actors in N4U, based on use-cases identified and elicit functional requirements for NeuroProv. This chapter then explains provenance in neuroimaging analysis in the context of N4U.

- Chapter 4 – Research Methodology

  This chapter will explain the research methodology in detail along with the evaluation procedure to analyse the results.

- Chapter 5 – NeuroProv Architecture & Experimental Setup

  This chapter will explain the NeuroProv's Architecture in detail within the context of N4U. The chapter will then enlist the experimental setup to evaluate the efficacy of the proposed system.

- Chapter 6 – Results and Analysis

  The results obtained from the experiments will be presented and discussed in this chapter. This will help in understanding the functioning of the proposed approach along with the pros and cons.

- Chapter 7 – Conclusions, Contribution to Knowledge and Future Directions

  This chapter discusses the contribution NeuroProv will make to the neuroimaging community. The chapter will highlight which research questions have been answered and discuss the significance of our research work. The conclusion of the research work is presented in this chapter along with the key findings of the research study. Future work will be highlighted that can be carried out to extend this work.

1.8 Research Publications

The following is a research publication in relation to this thesis:

Bilal Arshad, Kamran Munir, Richard McClatchey & Saad Liaquat, "**Position Paper: Provenance Data Visualisation for Neuroimaging Analysis**", in IEEE Xplore - Frontiers of Information Technology (FIT), 18 December 2014. (http://arxiv.org/abs/1502.01556)

1.9 Contribution to Knowledge

NeuroProv provides a clinical researcher in the domain of neuroimaging means to visually represent provenance data in an intuitive manner allowing users to exploit the true potential of provenance data. Researchers can benefit from the use of NeuroProv to understand provenance data for neuroimaging analysis; authenticate an experimental result; compare two or more workflows to figure out anomalies and ways to rectify it; visually see how workflows evolve over the period of time and last but not the least see the progression of a workflow over due course of time. The use of the Sankey Diagram 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.

Chapter 2

Background and Related work

2.1 Introduction

An introduction to scientific workflows, neuroimaging analysis, provenance and visualisation is provided in this chapter. This will establish the basis for understanding the terminologies and characteristics of scientific workflows for neuroimaging analysis and provenance visualisation. This chapter then provides a survey of the state-of-the-art in the light of requirements for provenance visualisation as identified in Section 2.2. Based on these requirements this chapter outlines the need for provenance visualisation system for neuroimaging analysis.

2.1.1 Scientific Workflow

The concept of a workflow was coined first in late 1970s and early 1980s [11] and it is defined in business community [12] as "*The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules*". In simple terms, a workflow is a collection of tasks defined in an order to achieve an overall goal of the workflow and these tasks can exhibit certain dependencies on each other. A scientific workflow can be defined as *a collection of tasks arranged in an order on the basis of their data dependency or control dependency to perform complex tasks such as the analysis of human brain images, analyses of physics particles or weather forecasting*.

Scientific experiments such as the neuGRID [4] and LHC [49] produce huge amounts of data and require high performance computations to analyse these data. For this, scientists have been using workflows which enable researchers to collaboratively design, manage, and obtain results. Complex workflows (such as ones used in neuGRID) involve many hundreds or thousands of steps and access gigabytes of data in order to generate results. Scientists require understanding these complex workflows and the results generated for insight into domains such as neuroimaging. This is done by inspecting the associated provenance of a workflow. Provenance information includes intermediary outputs and the final output produced during the execution of a workflow. This information is becoming more and more important for scientists since it helps them to understand what has happened in each step of a workflow and if need be modify their workflow accordingly.

This information can be used in workflow composition stage by learning from the past executions of a particular algorithm on particular datasets or even older workflows.

Workflows enable automation of tasks allowing capturing processes in an explicit manner, this saves repetitive time and effort. Workflows allow modification, maintenance, substitution and personalisation which are essential to carry out analysis. They are easy to share, explain, relocate, reuse and build [6]. Furthermore they release scientists/clinical researchers of worrying of managing tasks and let them focus on other work such as analysing trends and identifying biomarkers. The use of workflows and provenance for neuroimaging analysis is described in the following section.

2.1.2 Neuroimaging Provenance

Projects such as N4U and neuGRID enable clinical researchers to perform neuroimaging analysis using complex workflows to determine biomarkers for neurodegenerative diseases such as Alzheimer's. State of the art scientific workflows for neuroimaging analysis are increasing both in terms of computations they can perform and the size of data they consume/produce [13]. Due to this increase in complexity, it is essential to ensure the reproducibility of an analysis and also to confirm the correctness of resulting outcomes [14]. Additionally, the knowledge required to author any scientific analysis must also be captured. In a collaborative research environment such as neuroimaging, where the researchers use each other's' results and methods, traceability of the data generated, stored and used must also be maintained. All these forms of knowledge are collectively referred to as forms of 'provenance' information.

In any analysis system where there are multiplicities of datasets, and versions of workflows operating upon those datasets, particularly when the analysis is carried out repetitively and/or in collaborative teams, it is imperative to retain a record of who did what, to which data, on which dates, as well as recording the outcome(s) of the analysis. This information enables researchers to query analysis information, automatically generate workflows and to detect error and exceptional behaviour in past analyses. For these reasons this information needs to be logged as records of particular user's analysis so they can be reproduced or amended or repeated as part of robust research process. All of this information, normally generated through scientific workflows can be termed provenance data and it enables the traceability of origins of data (and processes) and, perhaps more important, their evolution between different stages of their usage. In neuroimaging analysis, the provenance about how a data is obtained is crucial for accessing the quality and usefulness of information, as well as enabling data analysis in an appropriate context [15].

Provenance can be divided into two subtypes, data provenance and processing provenance. Data provenance is the metadata that describes the subject (subject being a brain image in case of neuroimaging), how an image of that subject was collected, who acquired the image, what instrument was used, what settings were used, and how the sample was prepared. However, most scientific image data is not obtained directly from such measurements, rather derived from other data by application of computational processes. Processing provenance is the metadata that defines what processing an image has undergone; for example how the image was skull-stripped; how it was aligned to a standard space, etc. Even data that is presented as "raw" often has been subjected to reconstruction software or converted from scanner's format to more commonly used and easily shared file format [16]. For a complete picture it is essential that the history of a data product and process is transparent, enabling the free sharing of data across neuroimaging community.

2.1.3 Classification of Visualisations

In the broadest sense of classification, there are two categories of visualisation: exploration and explanation [17]. Exploratory visualisations support researchers with large volumes of data who are not certain what constitutes the data or what they are looking for. In contrast to exploratory visualisation, explanatory visualisations are designed when the researcher knows what the data has to say, and is telling a story to someone else. It is worth knowing that there is another kind of hybrid category involving a set of curated data that is nonetheless presented with the intention of allowing some exploration on the reader's part.

Visualisation of provenance is still largely an unexplored area, most of the systems focus on capturing and managing provenance information, while most of the visualisation systems focus only on providing an accurate rendering of some product, but not provenance. Providing accurate and complete visualisation of provenance data is one aspect of the novelty of the current research. As far as our research is concerned we will largely be dealing with exploratory visualisation, since we will be dealing with huge amounts of provenance data being generated as a result of (complex) workflow invocation.

2.2 Provenance Visualisation Requirements

Kunde *et al.* [18] derive abstract types of user requirements for provenance visualisation. For our research purposes we will be taking these requirements into account along with some other specific requirements based on our case-study on N4U. The requirements derived by Kunde include: 1) process: the sequence of process steps is the centre of inspection; 2) results: the intermediate or end results of interactions are in the centre of user view; 3) relationship: the relationship between interactions or actors is important; 4) timeline: the time is important to

observe; 5) participation: the correctness of participants is important; 6) compare: the comparison of two subjects shows the difference between them; 7) interpretation: an individual visualisation view depending upon end-user's requirements.

The goal of our visualisation research is to serve both the broad and narrowly focused audiences in the domain of neuroimaging, so it addresses each of the above mentioned requirements as follows: 1-3) our visualisation tool is based on an accepted model for provenance representation, namely, the PROV-XML [19], which denotes entities, activities and agents as nodes, and relationships as edges in a graph. It is able to show a complete graph with both the process steps and intermediate (final) results, or abstracts graphs focusing on either one of them; 4) the PROV-XML is capable of representing time information to nodes and edges; 5) participation is represented by agents through "wasControlledBy" relationship in the PROV-XML, so our tool helps the user visually evaluate the correctness of participation; 6) users can compare attributes of nodes or even compare multiple visualisations using our tool; 7) the last type of user requirement (interpretation), we aim to show that the proposed solution allows advanced users, in-depth inspection of workflow by drilling down to reveal further fine-grained provenance information about sub-activities. (For a more detailed summary on PROV-XML we refer our readers to [19]).

Kunde's requirements fail to encompass all aspects of our research primarily due to the fact that these requirements are very generic. Our research includes other requirements such as ways to display provenance information based on the use of provenance data for neuroimaging analysis. For instance verification, workflow composition allows the users to verify the correctness of a result or an intermediate result. Furthermore, insight often comes from viewing the analysis of origins of results (progression). In a collaborative research environment such as neuGRID scientists and clinicians frequently work with data that has been collected or processed by other groups or organisation. In order to understand how a particular set of users have progressed with a certain dataset or analysis in neuroimaging it is essential to view visualised provenance to ensure that the data has not been tampered with. Another essential requirement for provenance visualisation for neuroimaging analysis is evolution; in order to determine how a certain data product has evolved during the course of an experiment, researchers need to view visualised provenance. One of the major challenges in neuroimaging is the sheer volume of provenance data gathered over the period of an analysis. Some of the current workflow execution systems e.g. [26] use provenance to generate visualisation rather than generating visualisation of provenance data.

2.3 Provenance Visualisation Systems

Much research has been done in providing mechanisms to track and store provenance information in workflow systems but most of these systems lack visualisation support [20]. Provenance information can be considerably greater in volume when compared to the usual data stored by workflow systems. This is due to the fact that provenance information can contain details about the data and the processes involved in deriving that data product. The intermediate products that were generated during the process of workflow execution, along with other details about the system such as execution start and end times. One of the challenges of visualising provenance data is that the user might be overwhelmed by the amount of data with which he will be presented. It is important to note here that it is crucial to realise what and how information should be represented to the user so it may be beneficial for him. Presenting too much information on the screen will create a visual clutter making it hard for the user to realise the data and make sense of it. Too little data and we might be missing essentials that might give us insights. There needs to be a balance reached between the amount of information being displayed and its utility.

There are only a few systems that provide an all-in-one approach for visualising scientific products and their associated provenance. Provenance Explorer [31] provides visualisation support for data products and their associated metadata while VisTrails [26] provides visualisation of data products with the help of provenance rather than visualising provenance. Most systems focus on capturing and managing provenance information, while most visualisation systems focus only on providing an accurate rendering of some product, but not its provenance data.

The conventional visual encodings for provenance data are derived from the fields of network and graph visualisation. Having effective visualisation of provenance data is necessary to understand and evaluate the data. The most common visualisation strategy for provenance data is the Node-link diagram and is employed by common provenance tools such as Haystack [24], Probe-It [27] and Orbiter [38]. With this visual encoding, nodes are represented as graphs and edges or connections between nodes are represented as lines or curves. These tools utilise a variety of different visual encoding techniques including directed node-link diagrams [25, 27] and collapsible summary nodes [38]. Node-link diagrams are effective for representing provenance information for understanding local-activity, but they fail to offer a high-level summary of activity and relationships within them. The following sections provide a brief account of the projects that enable scientists to visualise provenance data along with the results of scientific applications.

2.3.1 Prototype Lineage Server

The Prototype Lineage Server [20] allows users to browse lineage information by navigating through the sets of metadata that provide useful details about the data products and the transformations it has undergone in a workflow invocation. Web server scripts on the lineage server query the lineage database and provide a Web browser interface that allows navigation via HTML links. Figure 2.1 below shows example screens of lineage server Web application. Views are restricted to parent and children metadata objects. Clicking on a parent object will move that link to the centre of the screen and show that object's parents. Clicking on the metadata object link in the centre of the screen will bring up the XML metadata for an object.

Since the system allows complete lineage for each system invocation, it is possible to navigate the transformation to and from any specific data product or transformation that contributed to a data product of interest. This helps the researcher in navigating and exploring which particular resources/processes contributed in a data product under consideration. Metadata requirements for the system are derived from the U.S. Content Standard for Digital Geospatial Metadata [36], coupled with Extensions for Remote Data Sensing Metadata [37] (CSDGM+ERSM). Since the requirements for the domain of custom satellite derived data products are derived from the above stated projects, Prototype Lineage Server caters to provenance being generated for this specific project, making it specific for Earth and environmental sciences.

Prototype Lineage Server allows tracking evolution of datasets and data products essential for determining how data has evolved over a period of time. It also provides the ability to focus on the results and processes. The functionality of the system to focus on parent and children metadata objects limits its ability to be used for neuroimaging analysis. Since neuroimaging analysis requires to drill down multiple levels of detail to inspect elements rather than just focusing on parents and children metadata objects. Furthermore the system only provides links that can be clicked instead of interactive visualisation that can be used for the purpose of our research.

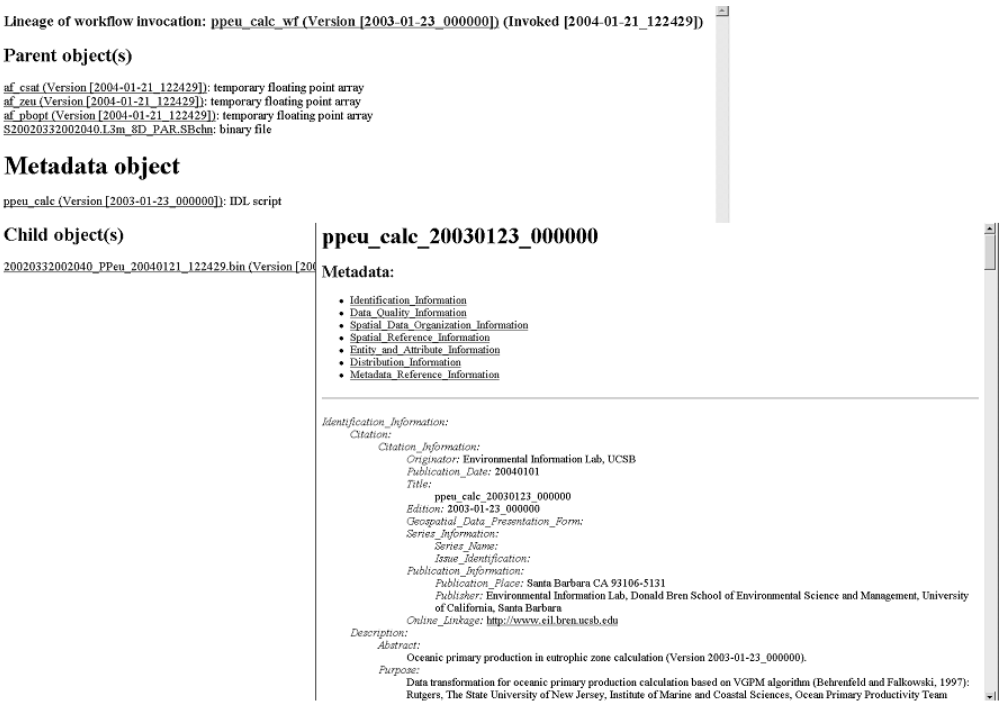


Figure 2.1 Example screens for Lineage Server Web Application [20]

2.3.2 myGrid

The myGrid Project [21], is a pilot e-science project in the UK, ~~has been~~ designed to provide middleware services for biological experiments represented as workflows [22]. myGrid is service oriented and uses the Taverna workflow system [23]. myGrid records all service invocations including parameters such as start and end times and the data items i.e. products consumed or produced. This log information can be used to derive the provenance of the data item produced by a workflow. This provenance information can be useful when reproducing and verifying an experiment since the researcher will have a clear idea of which data products or algorithms were used for a particular experiment.

Furthermore it also maintains contextual and organisation metadata such as experiment information, project and owner. This information is particularly useful in providing context to the provenance data of the workflow. myGrid encodes captured workflow provenance in RDF (Resource Description Framework); RDF is a mechanism used to relate provenance resources such as the data and services accessed during workflow execution. RDF graphs can be projected in to a range of views suitable for a particular user role. Based on a particular user role, the appropriate dimension of provenance is presented, knowledge, organisation, data or process level.

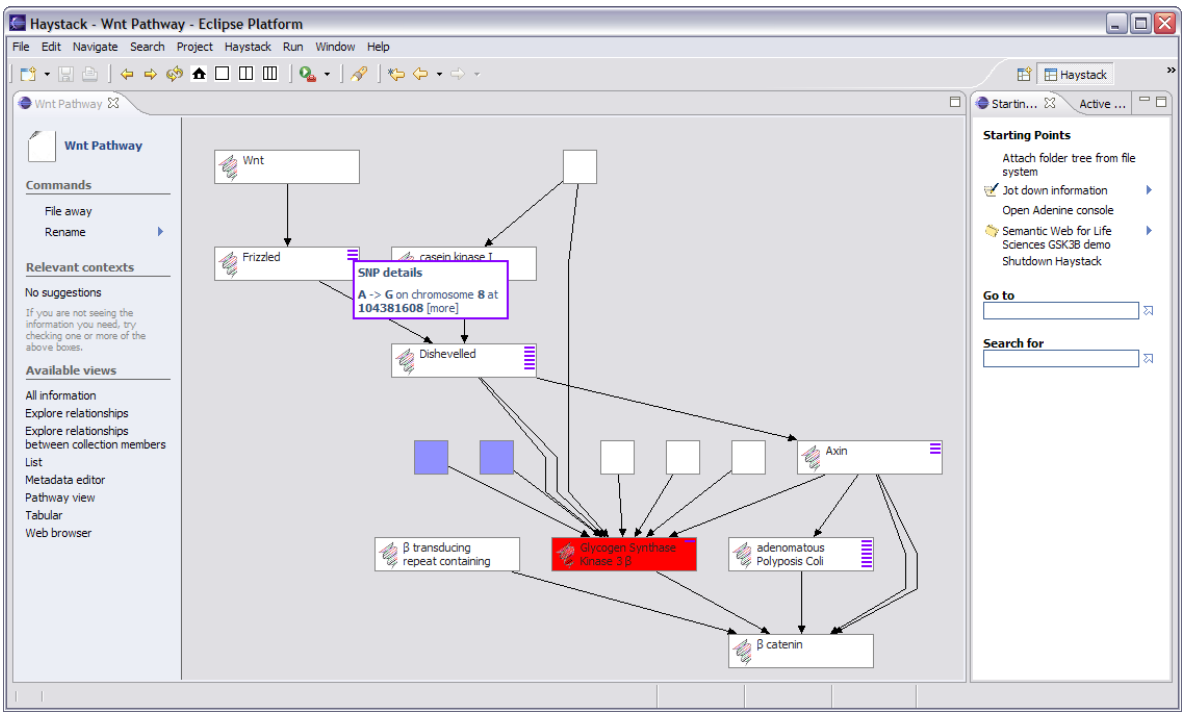


Figure 2.2 Haystack screenshot of visualising provenance log [25]

myGrid renders graph-based views of RDF encoded provenance using Haystack [24]. This is used to visualise network of semantic relationships among provenance resources associated with the experiment. Haystack is a semantic web browser that enables developers to provide tailored views over RDF metadata. Figure 2.2 shows a screenshot of Haystack visualising provenance log. Even though myGrid provides the capability to record and, visualise provenance, it is tightly integrated with the workflow execution environment and does not provide interoperable means for collecting and using provenance. What is needed is a system that allows users to compare multiple visualisations; the ability to generate an individual visualisation view based on end-user's requirements and the functionality to track the progression of data products over the course of the experiment.

2.3.3 VisTrails

VisTrails [26] is a workflow and provenance management system that provides support 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. Figure 2.3 shows screenshot of the main VisTrails window.

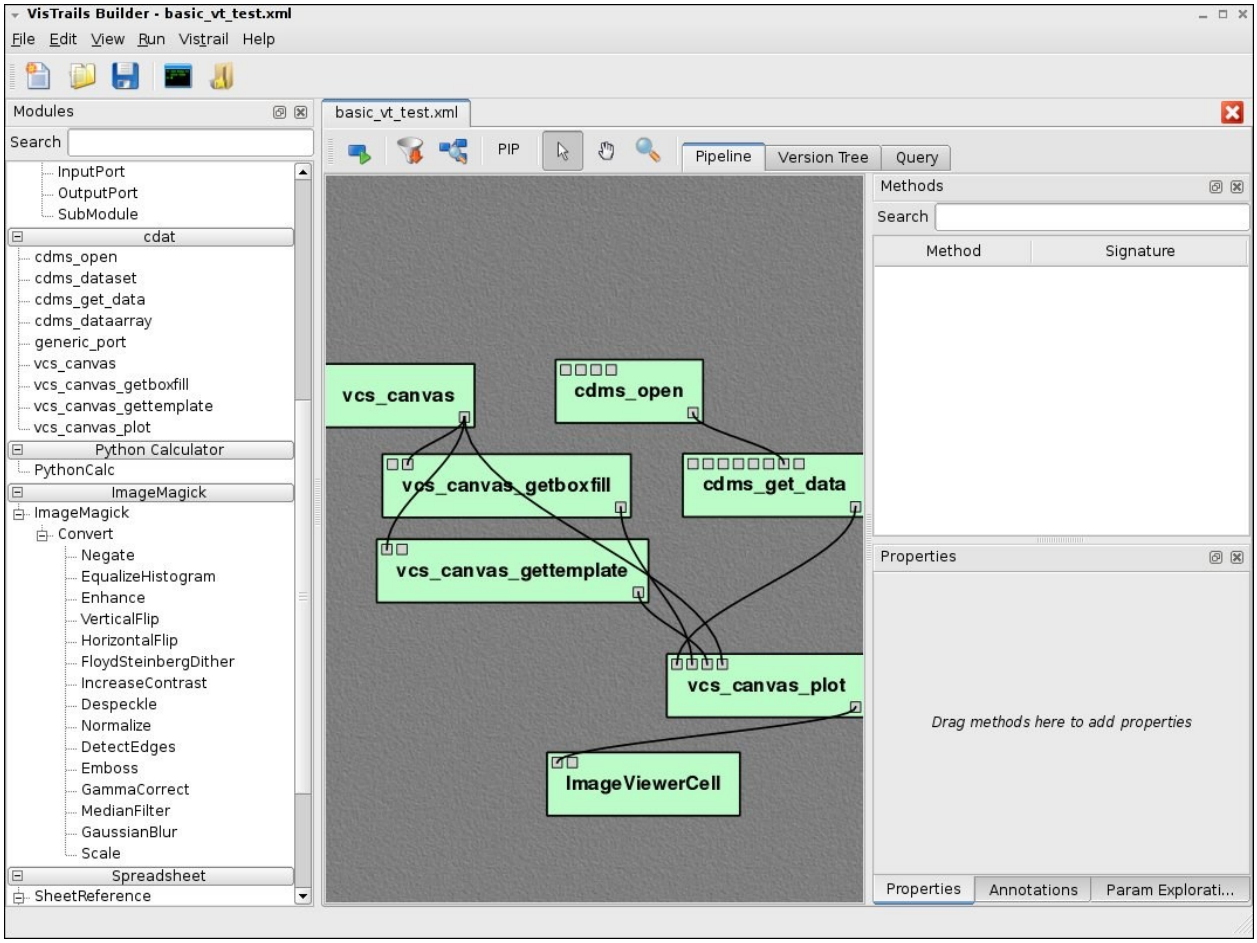


Figure 2.3 VisTrails Screenshot of main window [26]

VisTrails uses action-based provenance model that uniformly captures parameter values and pipeline definitions by unconstructively tracking all changes that users make to pipelines in exploration tasks. Action-based provenance greatly simplifies the exploration process and could reduce time for insight. Furthermore it only stores changes made to a workflow rather that storing a complete version of a workflow. By doing so, redundancy is removed and it results in better space utilisation. 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 visualisation of provenance data itself. In other words, VisTrails manages and displays provenance of visualisation rather than visualisation of provenance, which is a primary need for our research endeavour.

2.3.4 Probe-It!

Probe-It [27] is a tool designed to assist scientists in the understanding of scientific data by exploiting the useful features of both visualisation and provenance. It is a general purpose tool that has been used to visualise both logical proofs generated by inference engines and workflow execution traces generated by Kepler [28] (a workflow system). It is a browser suited for graphically rendering provenance information associated with results coming from inference engines and workflows. The explanations are rendered as directed acyclic graphs (DAGs), but are encoded via the Proof Markup language (PML) [29]. PML is an OWL based language for representing justifications of computationally derived results.

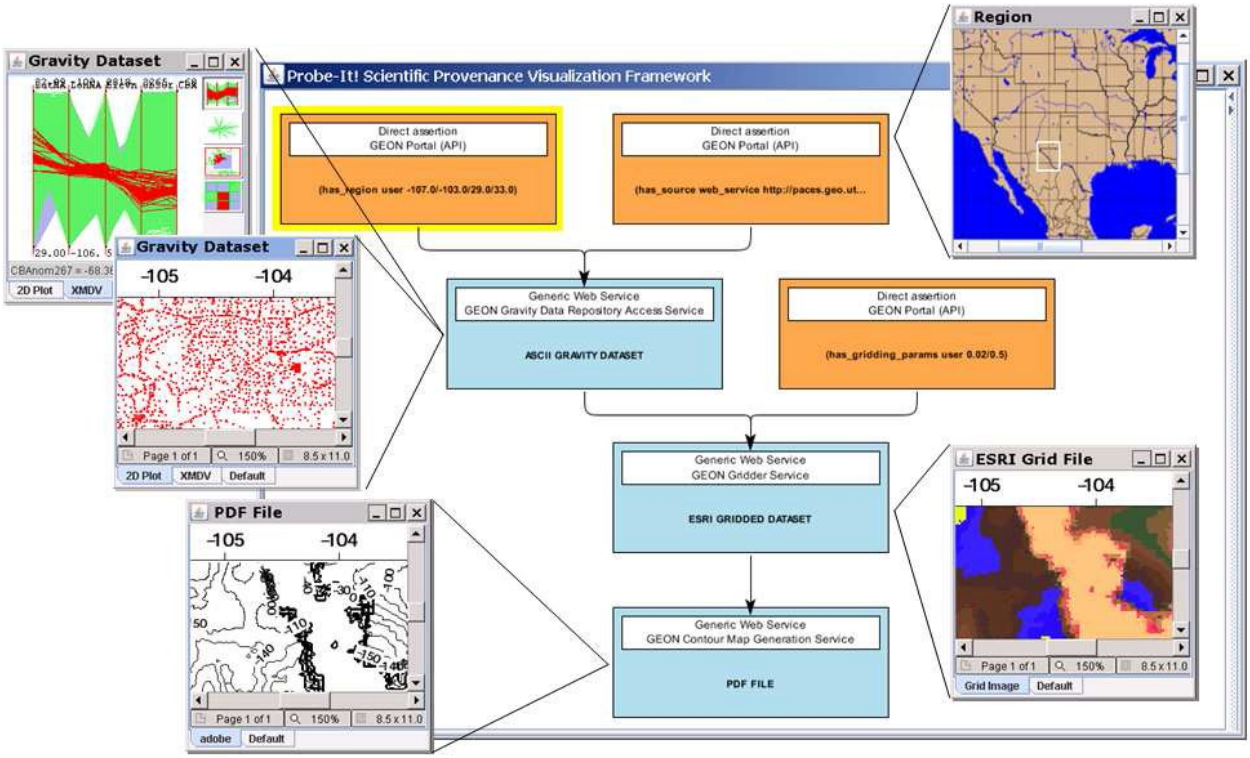


Figure 2.4 Probe-It! Snapshot [27]

In [27] an example is provided where gravity maps are used, contour maps are generated by geophysicists to identify subterranean features such as water table or oil reserves. Probe-It! provides the scientists with visualisation of provenance to help them both identify and explain map imperfections. Probe-It! moves the visualisation focus from intermediate and final results to provenance back and forth. It consists of three primary views to accommodate different kinds of provenance information: results, justifications and provenance which refer to final and intermediate data, description of the general process (i.e. execution traces), and information about the sources

respectively. As shown in Figure 2.4 provenance associated with the Gravity dataset can be visualised in two different views, XMDV and 2D plot.

Probe-It! allows scientists to help understand complex scientific products (e.g., datasets, reports, graphs, maps) derived by complex software system (e.g., applications and services) deployed on a distributed and heterogeneous environments such as a cyber-infrastructure. Probe-It! allows users to focus on processes, results and the relationship between interactions. It also provides the ability to compare multiple visualisations. However it fails to encompass the ability to track progression of data products or evolution over the due course of time. Furthermore Probe-It! functionality is currently restricted to the field of Earth Sciences hence limiting its ability to be used for neuroimaging analysis since the type of provenance data (images, algorithms and workflows) differs from that of Earth Sciences.

2.3.5 Pedigree Graph

Pedigree Graph [30], one of the tools in Multi-Scale Chemistry (MCS) portal from the Collaboratory for Multi-Scale Chemical Science (CMCS), is designed to enable users to view multi-scale data provenance. With the graphical view CMCS users can see relationships generated by multiple independent applications and discern connections that would not be evident within a single application. The portlet provides scientists with a two-dimensional visualisation of a data object or file or all of its scientific pedigree relationships. The view is static, and rendered straight from GXL (Graphical eXchange Language) files but users are able to traverse the tree by clicking on its links. Latest version allows display of formatted values for selected properties. Figure 2.5 shows CMCS Pedigree browser showing the metadata and relationships of the selected dataset.

Figure 2.5 CMCS Pedigree Browser showing the metadata and relationships of the selected dataset [30]

Pedigree graph exhibits limitations of a browsing metaphor for understanding long and complex relationships. It is difficult to follow and keep track of long chains of links while exploring for insight. Furthermore, moving back and forth from provenance visualisation to metadata or vice versa is not feasible and may result in error because the system is incapable of supporting this functionality. This transition is essential part of provenance visualisation for neuroimaging as the data product or process under consideration may need to be inspected often during the course of an exploration. The tracking of data or evolution in neuroimaging is essential for understanding how a dataset or data product has reached to its current state over the course of an experiment. This is essential to ascertain quality of a particular dataset and to ensure it has not been tampered with.

2.3.6 Provenance Explorer

Hunter and Cheung [31] developed a system at the University of Queensland, Australia with the ability to dynamically generate personalized views of provenance data based on a combination of user requirements, semantic reasoning and access policies. Provenance information is extracted from workflow systems (such as Kepler or Taverna) and user views are created based on the interfaces available. The system initially presents users with a coarse-grained view of the provenance data. However the GUI allows 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. Figure 2.6 shows a snapshot of Provenance Explorer's expanded provenance views. Based on user access/privileges Provenance Explorer allows users to drill down certain edges to view complete provenance as

shown in Figure 2.6. Along with the expanded view, Provenance Explorer also provides users with associated provenance in the textual form found below the graph in Figure 2.6.

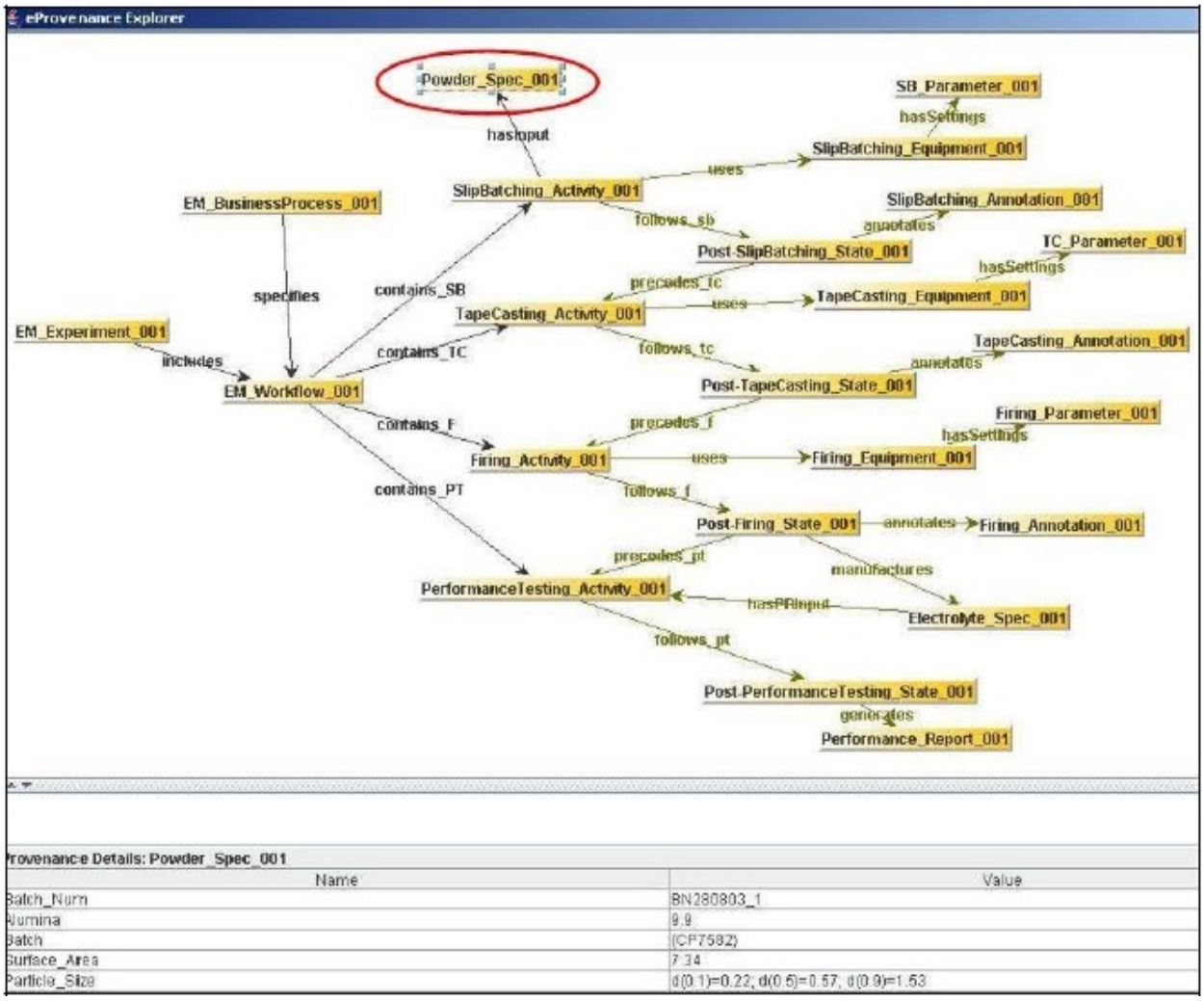


Figure 2.6 Snapshot of Provenance Explorer's expanded provenance view [31]

One of the major drawbacks of Provenance Visualizer is that it only supports expansion to one level of detail. Ideally users should be able to incrementally drill down to multiple levels of detail in order to gain insight. Furthermore 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. Our system will provide the ability to compare different visualisations for error detection and understanding the differences. This often provides insight by identifying anomalies in a given analysis or dataset. Furthermore our system will provide the ability to generate individual visualisation view depending upon end-user's requirement rather

than limiting the ability of a user to explore the workflow based on his/her access privileges. Often useful information is hidden underneath the details and is unearthed upon detailed inspection. If examination of provenance data is limited based on user access, the researcher will be deprived of useful information for analysis.

2.3.7 ESSW

The Earth Science System Workbench (ESSW) [32] is another system of capturing and presenting provenance information to users. ESSW is a client/server architecture in which a workflow (client) transmits provenance information to the services which manage the provenance. Upon user requests ESSW leverages a suite of Notebook tools that can display both the scientific product and the associated provenance, see figure 2.7. Stored scientific visualisations such as swaths [32] are rendered in HTML upon requests (in the form of query). Recoded provenance is rendered by Graphviz [33] in the form of a directed graph, where nodes are data objects and edges define relationships between objects.

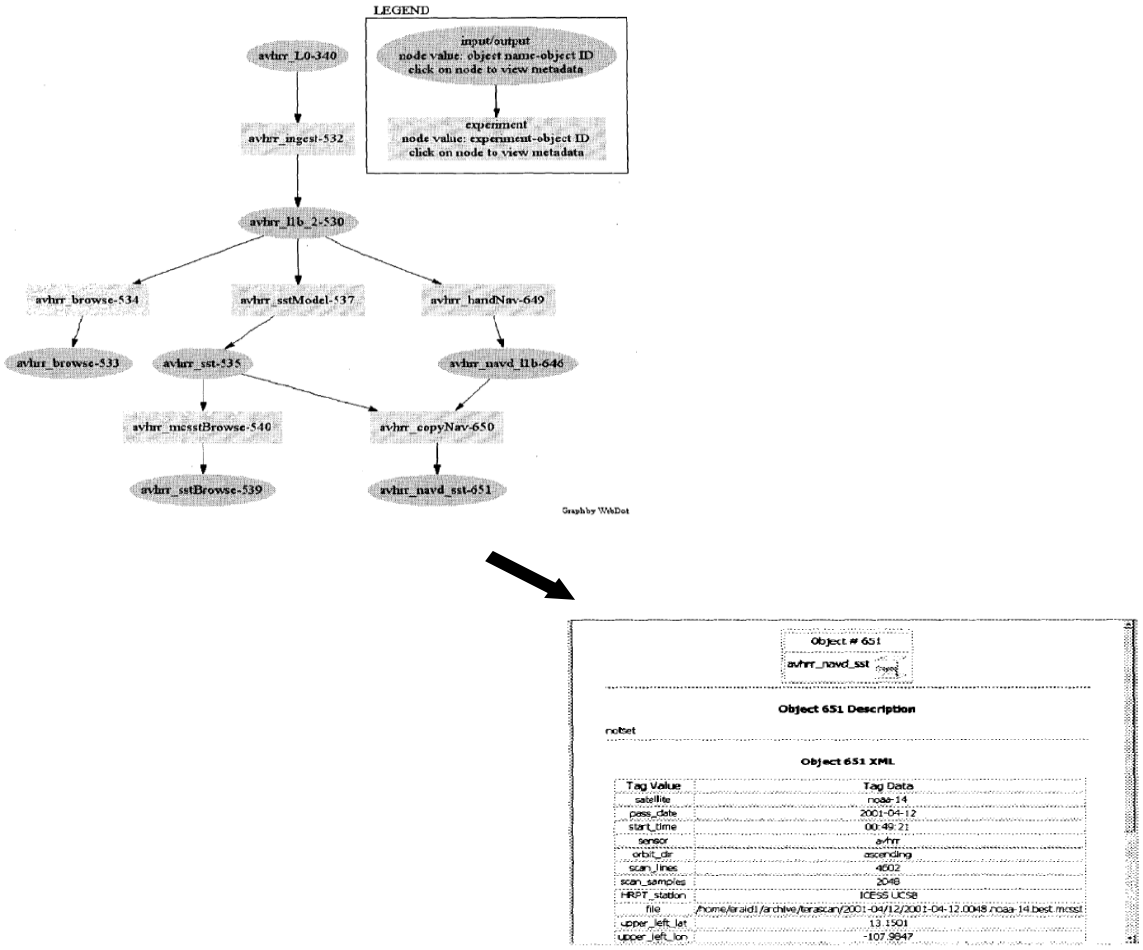


Figure 2.7 Using a Notebook tool to view science object lineage (top left), Viewing metadata for a science object in a lineage diagram (bottom right) [32]

Currently ESSW's capabilities are restricted to the domain of Earth sciences. Furthermore ESSW does not take into account user views depending upon end-user's requirement when displaying provenance. Since ESSW collects lineage as simple parent-child links between services and files, recording provenance is tightly coupled to its lineage database. Lacking support of logical data products hinders tracking data across virtual organisation. This limits the ability of ESSW to drill down details of provenance information essential for neuroimaging analysis. ESSW does not provide support for comparison of visualisations and also the ability to track progression and evolution of workflows.

2.3.8 Karma

Karma [34], developed at the Indiana University, is a non-obstructive provenance tracker for scientific results. Karma unlike ESSW provides an in-house approach for rendering provenance; an algorithm accurately pieces together a directed acyclic graph that describes the data or process provenance. Karma is particularly targeted at capturing provenance associated with service oriented workflows, thus rich provenance associated with Web service invocations are captured by the system. However since the system only takes into account network data it limits its ability to store and visualise provenance for neuroimaging analysis.

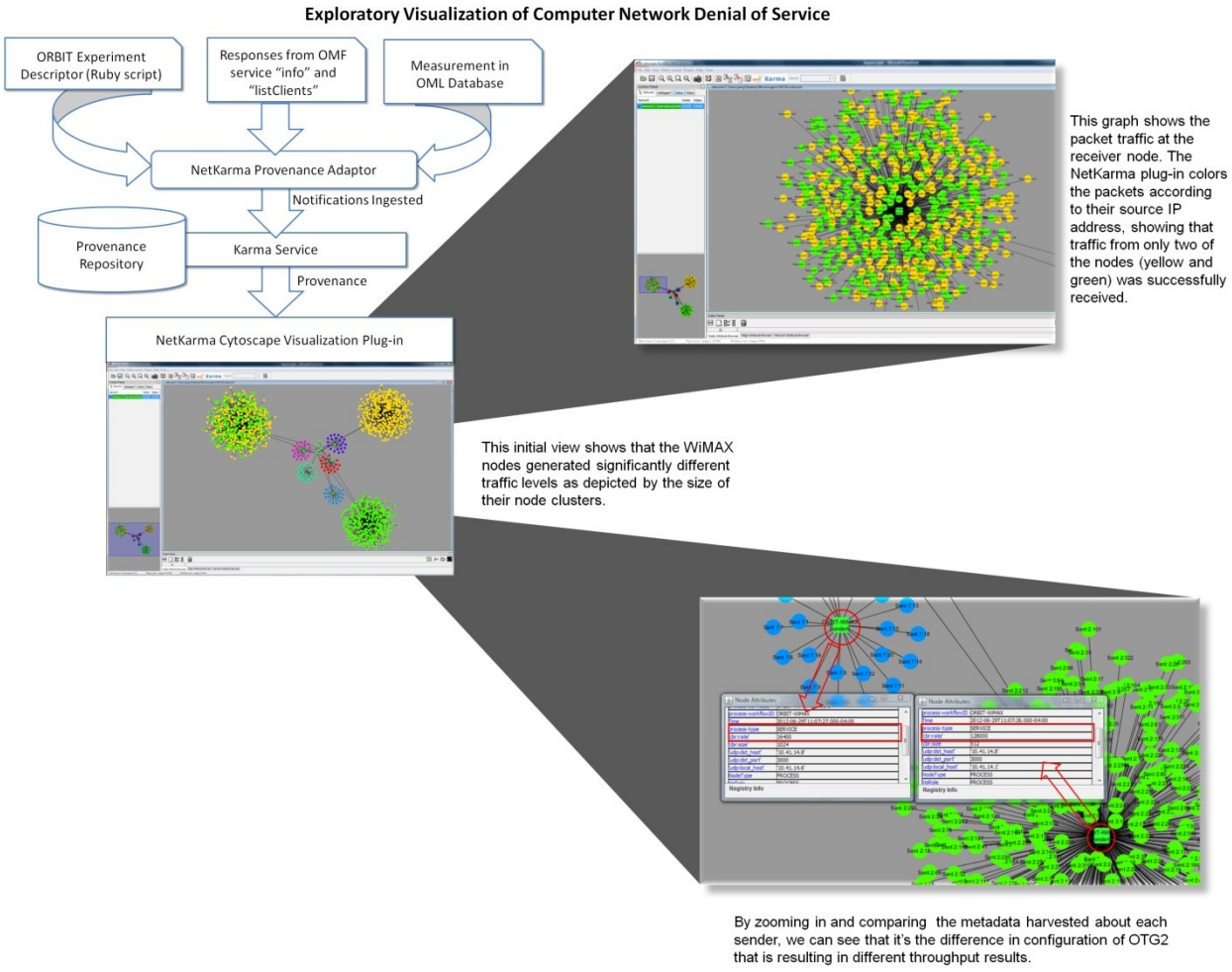


Figure 2.8 NetKarma visualising computer network denial of service [34]

The NetKarma (Figure 2.8) tool allows researchers to see the exact state of the network and store configuration of the computer network's denial of service experiment and its slice. NetKarma generates visualisation of provenance data stored as log files. There are several challenges to using

logs for provenance information [34]. One of the major challenges is that it if the system fails to capture log for a particular data it is impossible to capture its provenance. Another challenge of using log files in a distributed environment is the mechanism of consolidating the logs from different modules and hosts. Moreover an important consideration for recording provenance by using log files is the quality of the provenance itself. Since provenance is a quality artefact the quality should be quantified and logging in files is not suitable for doing so. NetKarma lacks support for comparative visualisations and tracking of provenance data over the course of the experiment. NetKarma has been superseded by Komadu [42], with additional support for PROV [19] compliant provenance data. One of the drawbacks of Karma is that its APIs are tightly coupled with workflows.

2.3.9 CI-Browse-It!

CI-Browse-It [35] developed at the university of Texas at El Paso is a graphical tool used for processing and browsing PML (Proof Markup Language) documents describing scientific workflow conclusions in a cyber-infrastructure environment. Furthermore it provides a framework for installing tools and techniques to visualise node set conclusions, whether the conclusions are the main or intermediate results of complex workflows. CI-Browse-It also provides rendering for an arbitrary number of formats for scientific results including PostScript, Portable Document Format and raw binary data.

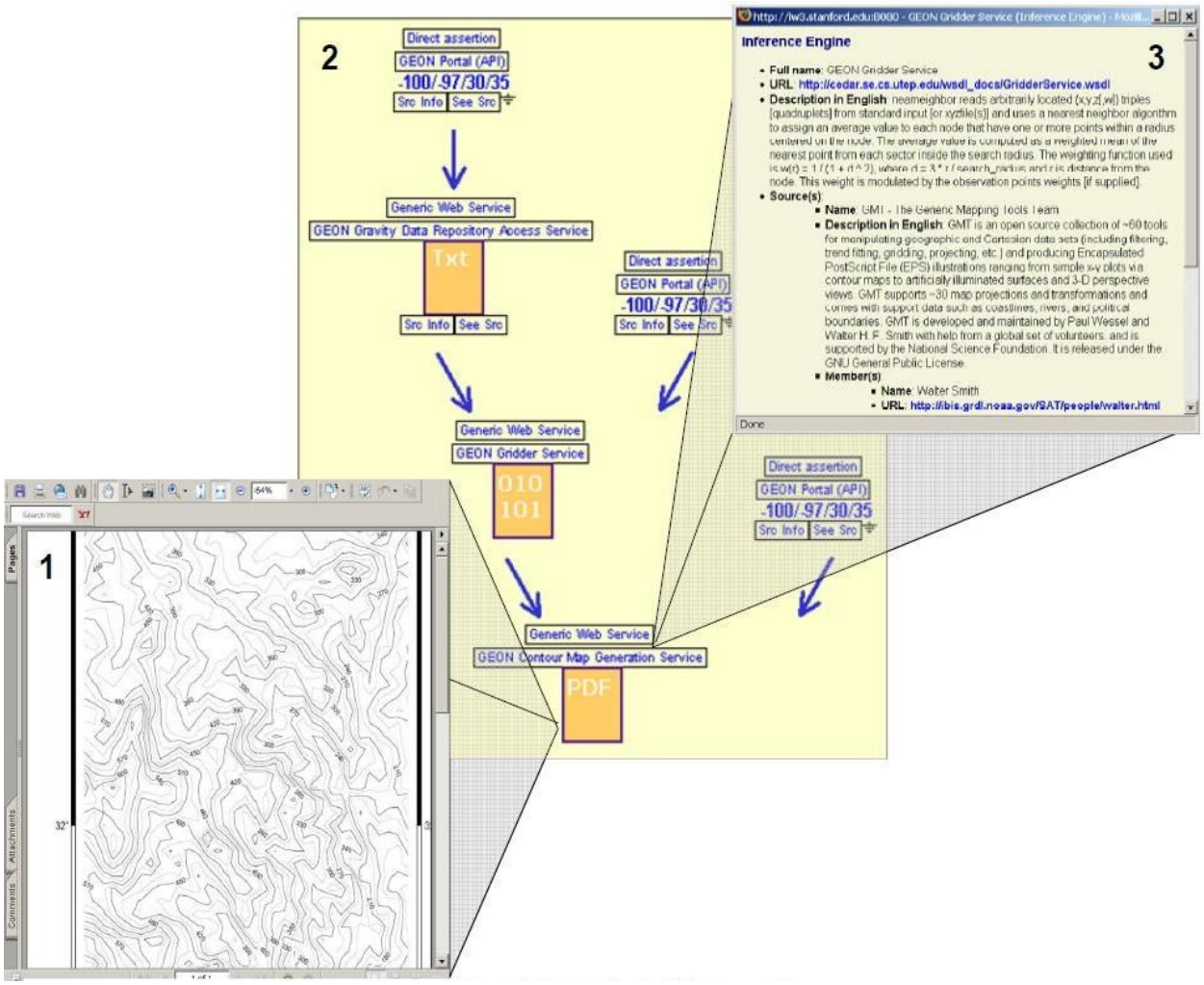


Figure 2.9 CI-Browse-It! Snapshot [35]

CI-Browse-It!'s Navigation model is composed of two primary views: content and provenance views. The Content view provides custom rendering of node set conclusions (data products) while provenance view displays the associated inference steps (provenance). In the context of gravity map scenario, the final result's node set conclusion is a contour map of gravity values as shown in Figure 2.9, label 1. Additionally users are presented with a link to provenance displayed in the form of DAG Figure 2.9, Label 2. Users can access metadata related to any web service consumed by the workflow as shown in Figure 2.9, label 3. Multiple representations of provenance data, as done in CI-Browse-It!, can help researchers decide which view can be best suited for analysis and insight. CI-Browse-It! lacks support of incorporating the element of time which is essential to determine the order of steps followed/performed in a neuroimaging workflow. This is essential to determine causality claims, e.g. a cause has to occur in order for an effect to exist. If time is not taken into account verification of events can be misjudged and this is essential

for neuroimaging analysis since this might tamper the data we are working with and in effect the whole analysis.

2.3.10 Prov-O-Viz

PROV-O-Viz [41], developed by VUE in Amsterdam represents a Web-based visualisation tool for PROV-based provenance traces coming from various sources that leverages Sankey Diagrams [40] to reflect the flow of information through activities. The aim of the system is to focus on a visualisation approach to identify important activities within a provenance graph based on data flow and link those activities together. Additionally it provides users the means to understand how data flows through a selected activity. Figure 2.10 displays Prov-O-Viz in action displaying provenance trail for Experiment 1 with various entities and activities.

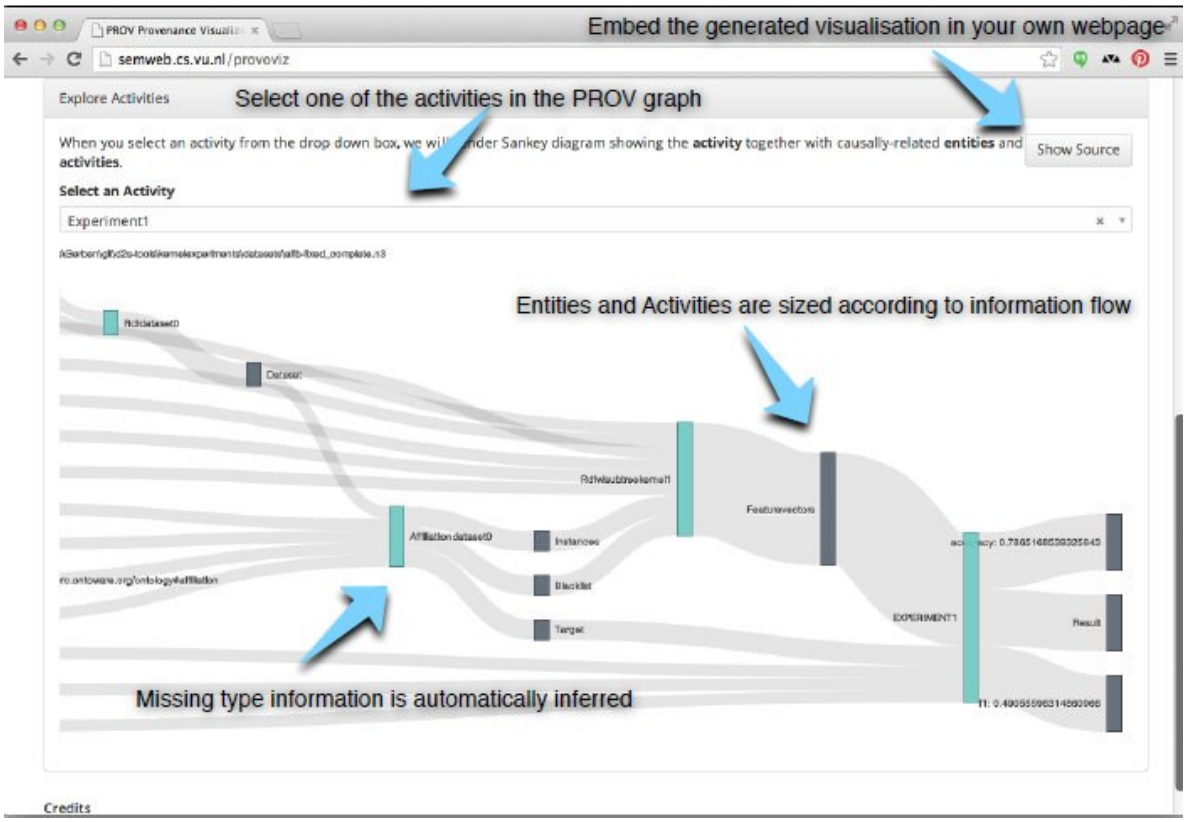


Figure 2.10 PROV-O-Viz - Visualisation of provenance trace generated by Ducktape [41]

One of the major drawbacks of the system is that it does not provide users with the ability to generate entity centric diagrams. Entity centric diagrams are a useful way of interpreting a particular entity under focus and its whereabouts. All of this information is helpful to validate results. Furthermore the system does not provide any browsing feature to find datasets, workflows etc. that the user might be interested to view along with a visualisation. The users do not have the

ability to click through the various parts of the provenance graph hence they cannot inspect individual workflow elements for verification and authentication purposes.

2.4 Conclusion

In this chapter, some of the key concepts related to workflows, neuroimaging analysis and provenance visualisation have been defined. Provenance visualisation is affected by the huge amount of data being produced as a result of workflow. In particular, neuroimaging analysis generates enormous amounts of data in the form of image files that needs to be accurately captured, rendered and readily available. Existing systems do not exhibit the characteristics that can accommodate this overarching requirement.

Based on the requirements identified for provenance visualisation in context of neuroimaging analysis, most of the current visualisation systems partially lack support for incorporating the requirements for the domain neuroimaging analysis. For instance Provenance Explorer uses a visualisation approach that renders provenance data in a user friendly environment but lacks the ability to further explore by only allowing/providing support to drill down one level in DAG. The ability to drill down to finer levels of detail is essential for exploratory purposes in neuroimaging analysis. Pedigree Graph attempts to visualise provenance in detail, however it lacks to provide support for comparison which often brings insight to experiments in the neuroimaging analysis domain. Similar is the case with ESSW, Karma and myGrid.

It is clear that there is a need for visualising provenance for neuroimaging analysis since the existing approaches are short of incorporating and fully representing provenance for neuroimaging analysis. Therefore there is a need for a new provenance visualisation approach to effectively visualise and represent provenance information for neuroimaging analysis. It is clear from the survey provided in [27] that there is some level of understanding of scientific needs for provenance visualisation. However, there is no systematic investigation that shows which representational techniques are better for provenance visualisation. The following chapter will identify functional requirements for the proposed system NeuroProv based on the use-cases to provide users with a better understanding of the system's working and representation of visualised provenance data.

Chapter 3

Proposed Research and Requirements Analysis

Based on the literature review this chapter aims to formulate the hypothesis that will drive our research study. Since existing provenance visualisation systems are short of fully representing provenance data for neuroimaging analysis this chapter then aims to identify the functional requirements for provenance visualisation for neuroimaging analysis based on the needs of the users. This will provide a brief account of the experience the users need for analysis. Furthermore this chapter will present use-cases for NeuroProv, the system we will be developing to visualise provenance data for neuroimaging analysis. This chapter also provides a brief description of the users of N4U in order to allow readers to better understand how users in N4U might interact with the system. Furthermore in conjecture with the use-cases we will highlight which set of users will be relevant for use-cases in order to provide better understandability of NeuroProv in context of N4U.

Current provenance visualisation systems are not sufficient for visualising provenance data for neuroimaging analysis. VisTrails [26] for instance, manages and displays provenance of visualisation rather than generating visualisation of provenance which is a basic requirement for authentication and verification of results in neuroimaging. Provenance Explorer [31] lacks the ability to drill down more than one level of detail whilst inspecting the provenance for a workflow. Limiting the system's ability to drill down to only one level of detail may hide essential details from the user's perspective to determine the correctness of the intermediate results. Prov-O-Viz [41] does not provide the ability to interact with the visualisation generated thus users cannot inspect workflow elements to authenticate the results and the processes involved.

Visualisation of provenance data is essential for N4U since it allows clinical researchers to verify the results of existing analyses. This can be done by reproducing the analyses with the given parameters and configurations in order to prove if they yield the same result. Similarly N4U users can compare multiple visualisations in order to detect anomaly in the analyses and rectify them, providing source of potential error for future users. NeuroProv will allow N4U users to compare how a workflow has evolved over the passage of time, determining what changes occurred during the course of evolution and what caused the changes. All this information is essential to determine

the quality of the workflows and also to administer the sources for verification of results. Furthermore N4U users can visually see how a workflow has progressed during the course of the experiment to analyse the previous versions and use them as a basis of future experimentation. The following section describes the actors that will interact with our proposed provenance visualisation system. A set of comprehensive functional requirements for provenance visualisation of neuroimaging analysis will be presented in the later sections of this chapter based on the use-cases derived from the N4U case-study.

3.1 Actors in N4U

A key part in analysing the requirements for any system is identifying the types of users that will make use of it in some way. This allows to ensure that no features are missed out that a small number of members within a wider user community may desire. By modelling the ways in which the actors interact with the system that is being designed, a range of important conclusions can be drawn. Practically speaking, this may mean ensuring that representative members from each group of actors are present during requirements elicitation sessions and that they review any specifications that are produced. This section briefly describes the actors that have been identified in N4U and gives some profiles of project members from within the N4U consortium that are members of these groups.

Research Leaders

Team leaders who need to monitor the progress, resource usage and perhaps distribute research studies to a research team.

Example Profile

Francesco is a Neurologist and Vice-Scientific Director of IRCCS-Fatebenefratelli Hospital Brescia (Italy). His main research interests are focussed on the exploitation of intensive computational neuroanatomy algorithms in translational neuroscientific research and in the dissemination of new brain image analysis tools to clinical neuroscientists and clinical physicians. He works with his team to carry out research and communicate findings to the wider community through publications and other scholarly activities.

Researchers

Individual members of the research team who will use N4U during their day-to-day research work. These may interact with the system in different ways depending on their experience

and the nature of the research that they are carrying out. Broadly speaking the following groups of users has been identified:

Basic User

This group represents users who have a certain level of computing expertise, but are mainly content to use software as it was installed and are not inclined to customise environments to their needs. They expect a reasonably straightforward user interface through which they can carry out their day to day tasks.

Example Profile:

Antonio is a PhD student at K.I. with Professor Lorenzo. His research area is the anatomy and volumetry of the frontal lobe. His main research project involves frontal lobe dementia, which can be investigated by the shrinking of various small structures in the brain such as the putamen and caudate. A typical day at SMILE (Stockholm Medical Imaging Laboratory and Education) for Antonio involves using the Hermes system to manually trace the 3D outline of the brain structures of interest, sometimes importing more images into the system (the material consists of 600 patients being scanned at intervals of a year or so) to work on. Even though Antonio has studied some "computer science", he knows very little of computer programming and more complex operations. He can navigate inside a Windows system (but not add a printer, for instance) and do some basic tasks on a Linux system (cd, ls -- grep is the limit of his knowledge). The Hermes system has GUIs with buttons (and a Unix platform which the average user needs not bother with, usually), which he handles expertly. Antonio also knows how to run FSL and FreeSurfer, but cannot write scripts at all, on any platform.

Intermediate User

This user is similar to the Basic user but requires a little more flexibility in the way that they work and want to have more control over their environment. They may wish to extend existing workflows or make some changes to settings or the way in which they are configured.

Example Profile:

Simone has been a PhD student at IRCCS-FBF with Dr. Alberto since 2004. Her research area is the control, pre-processing and post-processing of diffusion tensor images (DTI) with specific tools for the analysis of the weighted images. A typical day for Simone involves the usage of the FDT (FMRIB's Diffusion Toolbox that is part of FSL system) to perform scanner pre-processing (e.g. averaging of multiple acquisitions, removal of images affected by large artefacts).

These initial steps are usually done manually by Simone. Then, in order to correct stretches and shears due to current distortion in the images she runs different command line utilities. A probabilistic diffusion model of the corrected data is generated and finally a probabilistic tractography map is outputted for each image. Simone is an end user that is able to run programs from the command line shell and knows how to write bash scripts or simple programs in a language such as Matlab, Perl or Python on a range of different platforms.

Advanced User

This group of users wants full control over their work environment. They may wish to construct new tools or adapt existing ones for other purposes. It is likely that such users have a high degree of experience and probably a good understanding of computing techniques. The flexibility to do what they want is paramount to this group of users and they do not wish to be constrained in their work by the system. They may also perform tasks that are covered by the Basic and Intermediate user roles from time to time.

Example Profile:

Researchers at VUmc performed volume changes over time for the brain and the hippocampi of MCI patients. For this reason the hippocampi where manually outlined at baseline. The mask for the hippocampi was converted into Analyse to combine them with the original images. For the brain volume change and the change in hippocampi volume the brains on follow up were registered to BL. The Fluid algorithm of the dementia research group of London (DRG) was separated from its surrounding GUI and executed on the hippocampi and the whole brain. This resulted in a relational comparison between the brain volume change, the hippocampi volume change and the MMSE values of the used data. To perform this analysis a number of scripts were used. Some existing programs of other research centre were slightly modified and used in a fashion that better matched the used data.

Pipeline Developers

The developers of new research pipelines need to integrate them within the system in order to provide facilities to researchers. These are very technical users and share many similarities with the Advanced User. Given the cutting-edge nature of their work, it is likely that they may go beyond this profile and may require access to development and debugging tools. They will also require a good degree of flexibility from the system.

Example Profile:

Thomas is a researcher with a long track record in the development and validation of image processing algorithms and pipelines for the quantitative analysis of brain MRIs. Typically the development of novel algorithms relies on rapid prototyping and testing cycles, in which algorithm or parameter changes are implemented, executed, and their results observed. This requires relatively low-level, “hands-on” access to the system, with the ability to rapidly modify modules in a pipeline and/or modify the pipeline itself, and execute immediate tests. For thorough testing and validation though, an algorithm or pipeline may need to be run on tens or hundreds (or even more) cases; and/or collection of scans may need to be processed using a range of parameter values in order to establish optimal parameter values. This latter case requires the ability to process large numbers of scans and/or a set of scans with a possibly large range of pipeline configurations.

Image / Data Input Managers

Managers and administrators that work to upload and manage the data stored within the system.

Example Profile:

Olivia has been a PhD student at IRCCS-FBF with Dr Frisoni since 2004. She is a key figure in the neuGRID Data Archiving and Computation Centres (DACS). Her main task will be to ensure the correct uploading of both images and data from the data collecting sites (DCS). She will have to maintain contact with the "data input managers" in the other neuGRID core labs in order to adopt procedures for standard data handling. Before the upload of each dataset she will perform a quality control procedure. A key aspect of the data input Manager is to organise the available data for use by the neuGRID community providing different levels of access and maintaining data integrity. She will ensure proper data management and the saving of local mirror copies of data. Finally, she has a deep knowledge of MySQL because the data management will be conducted through the LORIS relational database system.

The data managers at VUmc are collecting data from various sites. From each site firstly a dummy run is requested. This dummy run is checked for image quality and commitment to the scan protocol (both by the data manager and a Radiologist). After one or more rounds to establish the best acquisition parameters, scan parameters are frozen. After the successful dummy procedure the site can send images to VUmc. Each scan is checked whether it fulfils the parameters agreed on at the dummy run, whether the image quality is good enough, whether the required patient information (random codes) are in the file header and for other quality indicators. After these checks the data can further be anonymised and will be sent to an image archive.

System Administrators

Technical support operators are responsible for installing, monitoring and generally administrating the system.

Example Profile:

Daniel is a graduate in Mathematics and started his PhD at IRCCS-FBF with Dr Frisoni. Daniel will maintain and operate the neuGRID computer system and its network. He is usually charged with installing, supporting, and maintaining servers or other computer systems. This entails a good knowledge of operating systems and applications, as well as hardware and software troubleshooting. An important thing is that he must also have a detailed knowledge of the purposes for which people use the neuGRID platform and most importantly, he has strong problem solving skills. Daniel has already demonstrated a blend of technical skills and responsibility.

The following section provides detailed description of an end-to-end example for NeuroProv and later several use-cases to identify specific requirements pertaining to the visualisation system.

3.2 Description of an End-to-End Example

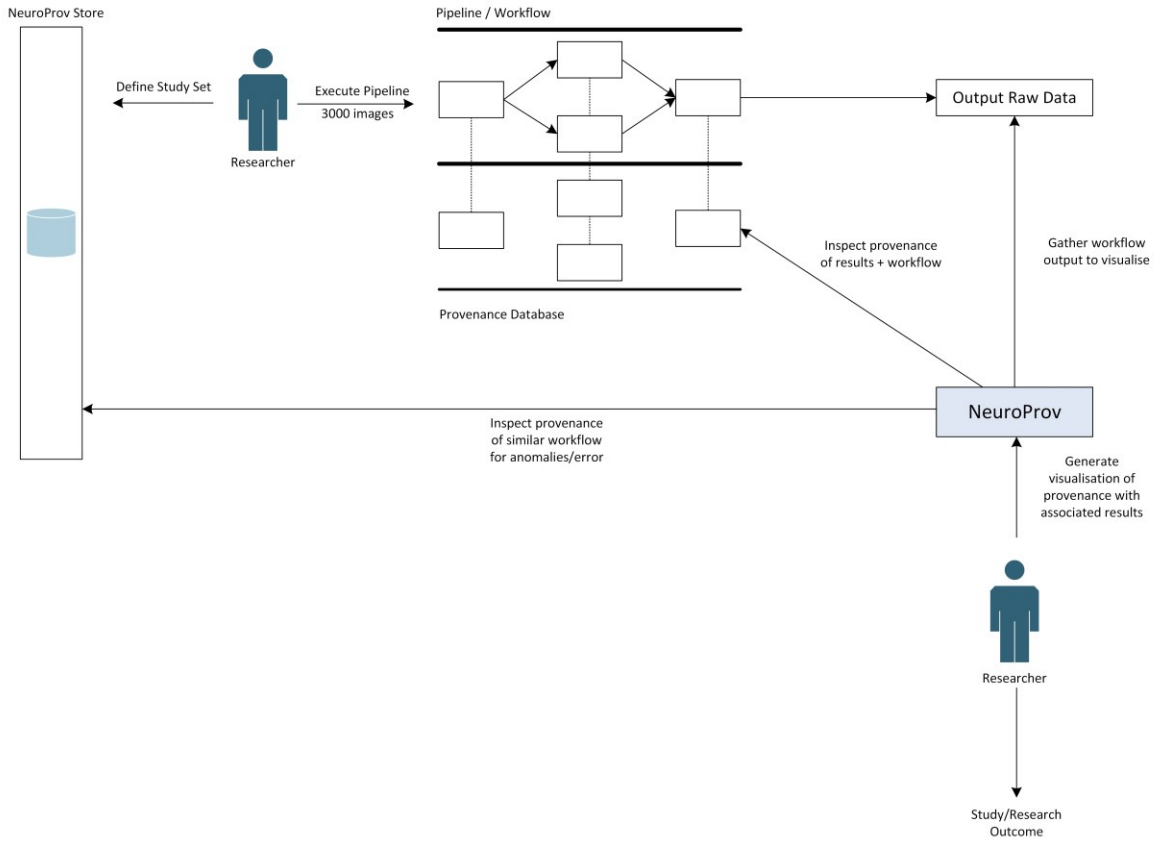


Figure 3.1 End to End Example of the use of NeuroProv

This section describes a potential end-to-end example scenario of the use of NeuroProv System, see figure 3.1. This sets the scene for the following section in which the requirements are specified, by demonstrating the use-cases for NeuroProv. In N4U, various users (as mentioned in the previous section) use provenance data for numerous purposes. For example 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 means to analyse all the intermediary image sets and results to confirm that the results 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.

Sometimes it may not be enough to reproduce the results. 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 these results. For example, a user may create a new workflow and run it on a test dataset. At each stage in the execution of the workflow, the intermediary images or data are stored and a full provenance track is kept. After the 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. Therefore visualisation of provenance data helps the researcher to validate results and workflows. The following section illustrates the use-cases defined for NeuroProv and the associated requirements for visualisation of provenance data. Actors have been added to use-cases to highlight the set of users relevant to appropriate use-cases. Primarily researchers have been targeted to interact with the system thus the three types of researchers namely basic, intermediate and advanced are incorporated in the use-case diagrams and highlighted next to use-cases requirements.

3.3 Use-Cases NeuroProv

Each segment of the user requirements specification begins with a Story. The relevant use-cases that are contained within it are described and then broken down further to form individual functional user requirements. The numbering scheme allows the hierarchical relationships between Stories, Use-cases and Requirements to be easily traced. The high-level Stories are indicated by the S prefix and Use-cases are given the prefix U. Individual requirements are denoted by the R prefix. The prioritisation scheme focuses on Essential, Desirable and Optional requirements and is based on the variation of the MoSCoW technique [39]. In contrast to functional requirements, non-functional requirements commonly describe system attributes such as security, reliability, maintainability, scalability etc. Non-functional requirements detail constraints, targets or control mechanisms for a system. They describe how, how well or to what standard a function should be provided [52]. For example, levels of required service such as response times; security and access requirements; technical constraints; required interfacing with users and other systems; and project constraints such as implementation on the organisation's hardware/software platform. This research study visualises provenance data with emphasis on functional requirements rather than the non-

functional requirements because we are interested in the capabilities the system provides based on the requirements gathered from use-cases.

Essential requirements are those which 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.

Where E = Essential D = Desirable O = Optional.

S1. 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 3.2. 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.

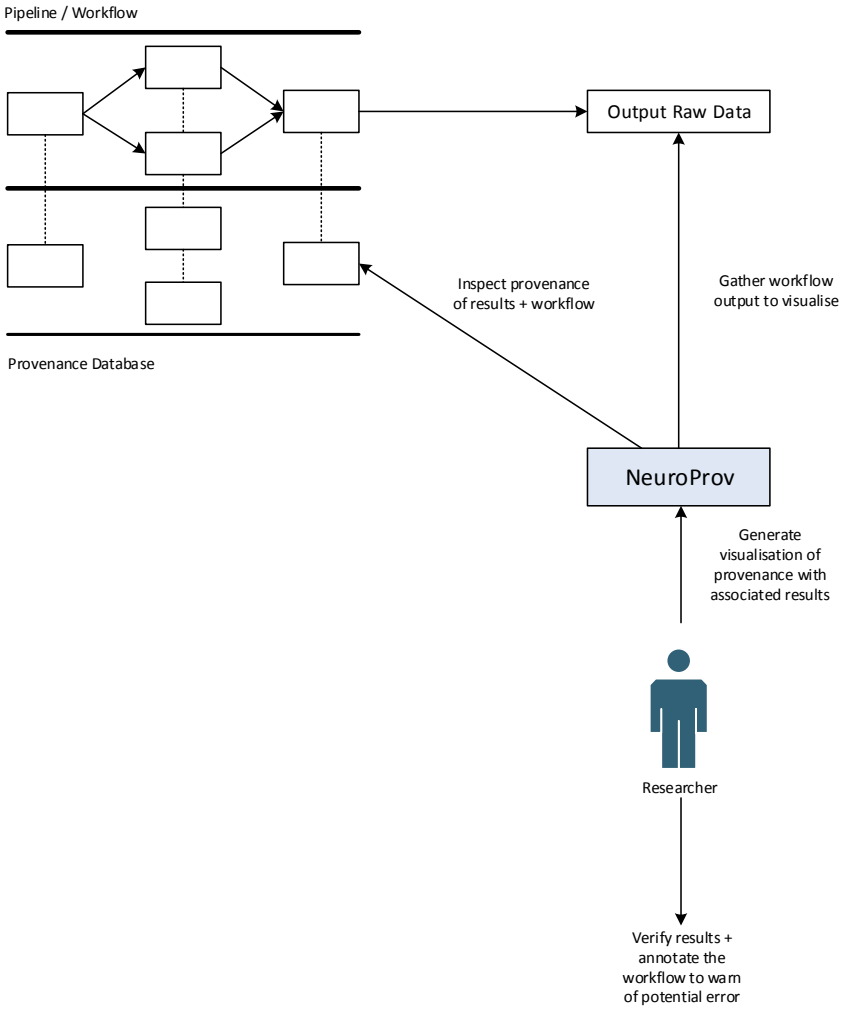


Figure 3.2 Verification of Results Use-Case

Indicative Use-Cases:

U1.1 Carry out verification of all the stages that have been processed during workflow execution using the associated visualised provenance data. E (Relevant to Basic, Intermediate and Advanced Researchers)

Functional User Requirements

R1.1.1	The possibility to import selected files from provenance database into the appropriate step in a given workflow using the GUI and then to analyse the results using NeuroProv.	D
R1.1.2	The ability to take the output from a single step in a workflow and look at it through a viewer/full text output	D
R1.1.3	Provide the user with visualised provenance to browse a completely executed workflow, process by process, and enable user to view all relevant intermediary output and logging information.	E

U1.2 Perform statistical analysis on the provenance data. O (Relevant to Intermediate and Advanced Researchers)

Functional User Requirements

R1.2.1	Check for additional abnormalities passed over in silence (weak field inhomogeneities, ringing entities etc.)	O
R1.2.2	Compare the results obtained with reference images.	D
R1.2.3	Allow a user to export/download provenance data to their computer system and perform statistical analysis on it subject to neuGRID usage policies.	D
R1.2.4	Results should be saved as a property of the provenance dataset. Files may go into a directory structure such as: Run number/Activity number/User-selected analysis set name/files.	D
R1.2.5	Provide any necessary format conversion tools.	O

U1.3 Annotate a workflow with information regarding potential errors and incompatibilities. O (Relevant to Intermediate and Advanced Researchers)

Functional User Requirements

R1.3.1	The workflow comments should not be unstructured text inputs but sorted into categories (General, Errors, Inconsistencies and Comment made by <name>.)	O
R1.3.2	When an error occurs a red colour could be used to depict that the workflow has a problem.	O
R1.3.3	Provide a user with the capability to annotate an item in the provenance database.	O

U1.4 Search a list of common errors that are known to affect a given workflow. D (Relevant to Intermediate and Advanced Users)

Functional User Requirements

R1.4.1	Search and display workflow comments regarding errors. Also, automatically save and compile statistics on which errors crop up during the run of a certain workflow.	D
R1.4.2	Post problems on the neuGRID technical forum	O
R1.4.3	Create a frequently asked question sections for each workflows.	O
R1.4.4	Provide the user with information about common errors that severely affect a workflow.	D

S2. Validation of Workflows

A 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 3.3. 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. 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.

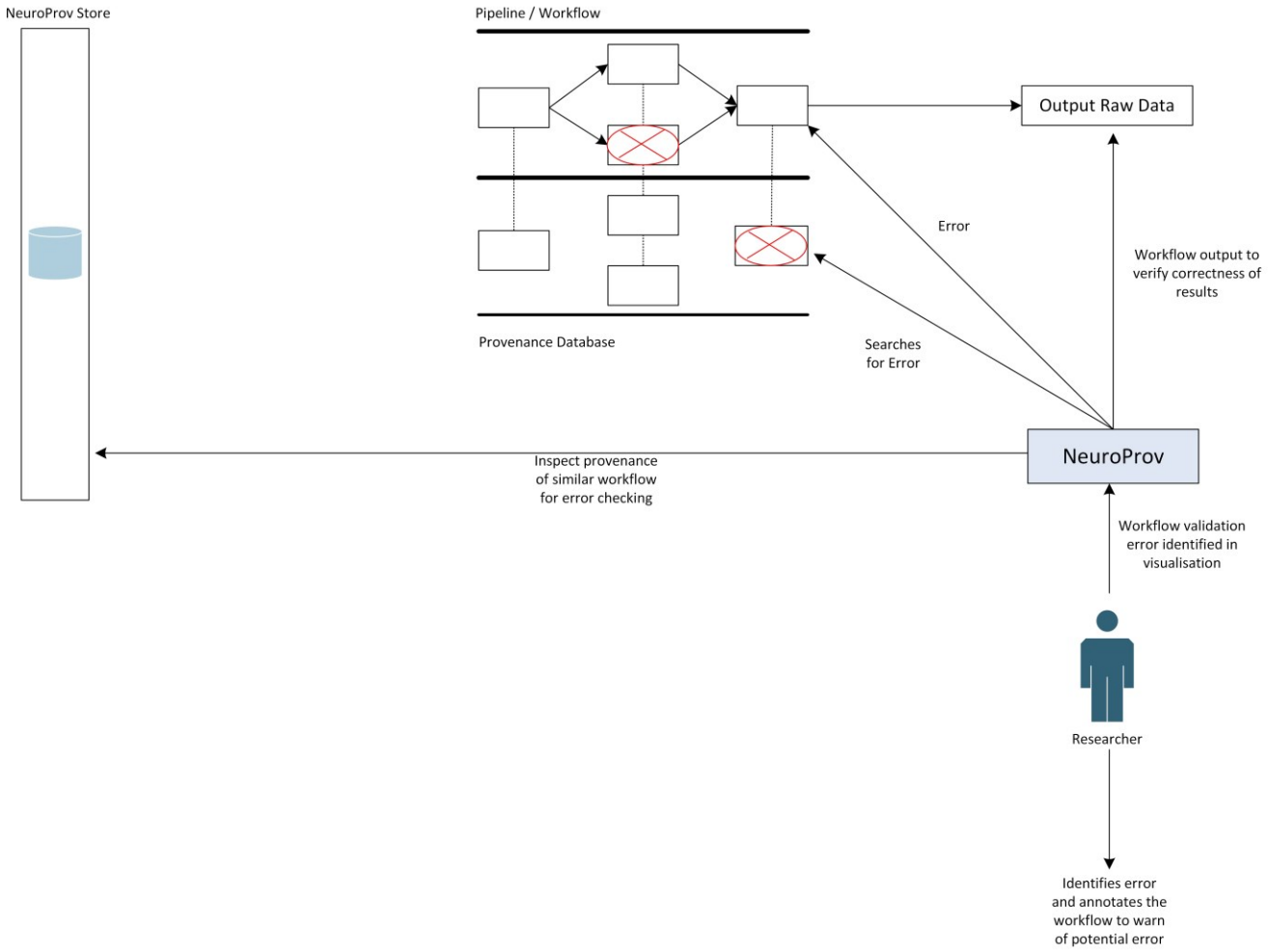


Figure 3.3 Workflow Validation Use-Case

Indicative Use-Cases:

U2.1 Validate a workflow using visualised provenance to locate points of failure in it. E (Relevant to Intermediate and Advanced Researchers)

Functional User Requirements

R2.1.1	Load the workflow into NeuroProv. The order of the boxes and layout of the workflow cannot be changed, but by clicking on each box the appropriate set of provenance data can be viewed: lists of images that can be put into the viewer (possibly to compare images, from different provenance sets and within sets) and numerical output data. Also the workflow setup can be viewed.	D
R2.1.2	Provide users with the capability to browse through visualisation generated from the execution of workflow	E
R2.1.3	The interface should be user friendly, and allow for browsing process by process provenance data	D
R2.1.4	Visualisation should allow intermediary output and associated provenance to be displayed	D

U2.2 Report errors in workflow execution. E (Relevant to Intermediate and Advanced Users)

Functional User Requirements

R2.2.1	An error report button should be included within NeuroProv GUI. It should send an email to the appropriate place with information regarding workflow setup, workflow name and dataset properties. It should also generate an error number for convenience and easy follow up.	D
R2.2.2	Some instances of a module could fail from time to time. In this case, it could be useful to have a viewer box in which all the failed instances of the module could be shown. With this information neuGRID users could diagnose the problems encountered during the execution of a workflow and hopefully solve them.	O
R2.2.3	Provide notification for critical events during an execution of a workflow.	E

U2.3Annotate workflows with version information and a full change history. D (Relevant to Intermediate and Advanced users)

Functional User Requirements

R2.3.1	Add a comment to a workflow.	D
R2.3.2	The possibility to make an analysis of the different usage patterns for each workflow that is available in the infrastructure. It would be useful to understand which data values are most commonly used by the scientific community and to analyse different types of acquisitions through different workflows.	O

U2.4 Provide users with workflow details and annotation present in the database. E (Relevant to Basic, Intermediate and Advanced users)

Functional User Requirements

R2.4.1	Search and display workflow details along with the visualised provenance	E
R2.4.2	Provide support to display annotations associated with images/activities	E

U2.5 Provide users with workflow execution timeline. E (Relevant to Basic, Intermediate and Advanced users)

Functional User Requirements

R2.5.1	Display workflow execution timeline along with the visualised workflow	E
R2.5.2	Provide users with the ability to view when an activity was generated or consumed	E

S3. Comparison of Workflows

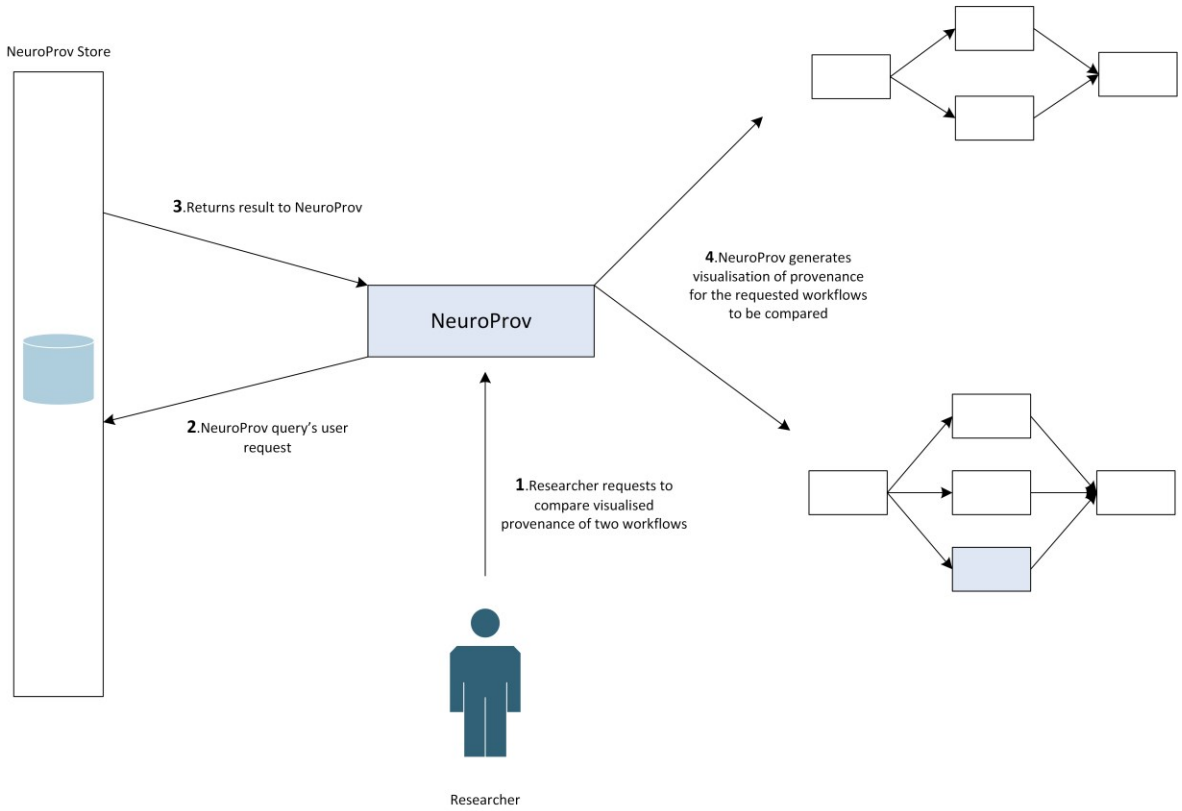


Figure 3.4 Comparison Use-Case

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 3.4. The user requests to generate 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.

Indicative Use-Cases:

U3.1 See how a workflow/dataset compares to an existing workflow/dataset. E (Relevant to Basic, Intermediate and Advanced Researchers)

Functional User Requirements

R3.1.1	Generate visualisation of two workflow/datasets under consideration side by side.	D
R3.1.2	Provide users with the capability to browse through visualisation of the workflows/datasets under consideration and do comparative analysis	E
R3.1.3	The interface should be user friendly, and allow for browsing process by process provenance data	D
R3.1.4	Visualisation should allow intermediary output and associated provenance to be displayed	D
R3.1.5	Show changes between the workflows in red colour	E

U3.2 See how a workflow compares to multiple workflows. E (Relevant to Intermediate and Advanced Researchers)

Functional User Requirements

R3.2.1	Generate visualisation of more than two workflow under consideration	D
R3.2.2	Provide users with the capability to browse through visualisation of the workflows under consideration and do comparative analysis	E
R3.2.3	The interface should be user friendly, and allow for browsing process by process provenance data	D
R3.2.4	Visualisation should allow intermediary output and associated provenance to be displayed	D
R3.2.5	Show changes between the workflows in red colour	E

S4. Evolution of Workflows/Datasets

A user wishes to view how a workflow or dataset has evolved over the period of time for a particular type of analysis, see figure 3.5. 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 basis to conduct his/her own analysis. The study also includes inspection of how a particular workflow has evolved over a period of time.

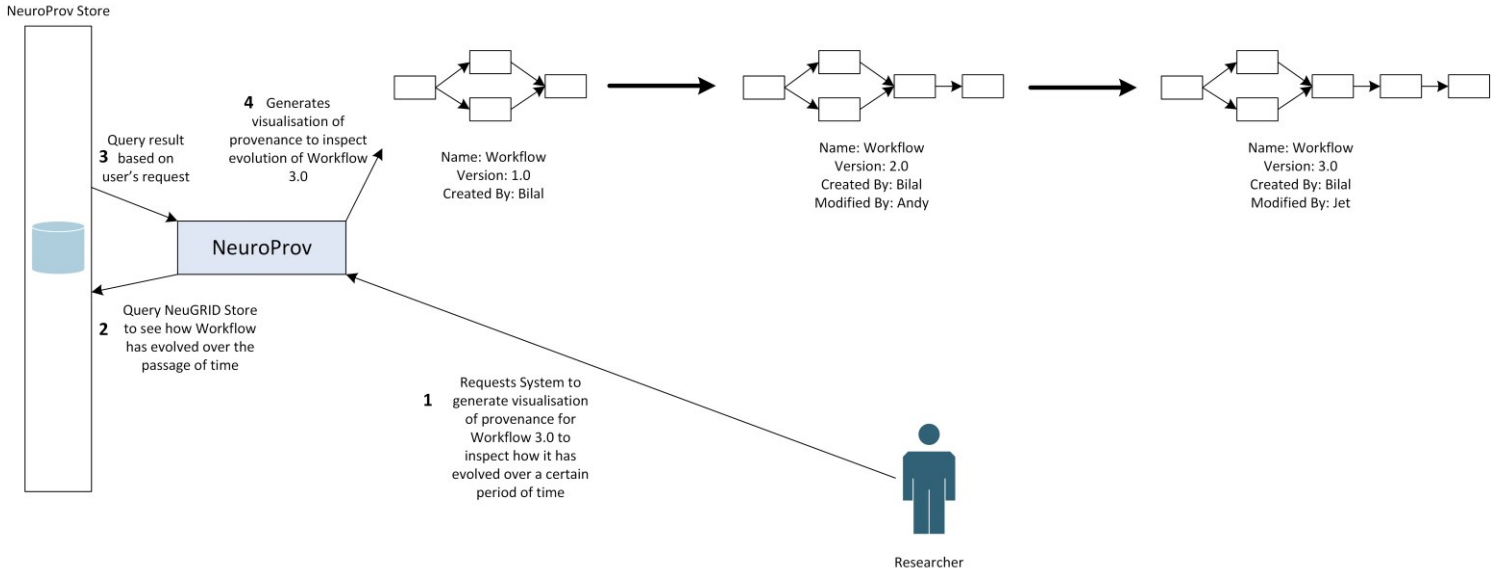


Figure 3.5 Evolution Use-Case

Indicative Use-Cases:

U4.1 See how a workflow/dataset evolve over a period of time using visualised provenance data. E (Relevant to Intermediate and Advanced users)

Functional User Requirements

R4.1.1	Allow for workflow setup to be viewed.	D
R4.1.2	Provide users with the capability to browse through multiple versions of workflow/dataset visualised.	E
R4.1.3	The interface should be user friendly, and allow for browsing process by process provenance data	D
R4.1.4	Visualisation should allow intermediary output and associated provenance to be displayed	D

Indicative Use-Cases:

U4.2 Annotate workflows/datasets with useful information for future use. E (Relevant to Intermediate and Advanced users)

Functional User Requirements

R4.2.1	Allow to query and browse through different versions of a particular	D

	workflow/dataset.	
R4.2.2	Provide users with the capability to annotate any useful information for future use.	E
R4.2.3	The interface should be user friendly, and allow for users to identify patterns/trends within various versions of a workflow/dataset.	D
R4.2.4	Allow users to inspect provenance for each version of the workflow separately	E
R4.2.5	Provide users with the ability to view annotations between different versions of the workflows to understand the changes occurred over subsequent versions	E

S5. Progression of Workflows/Datasets

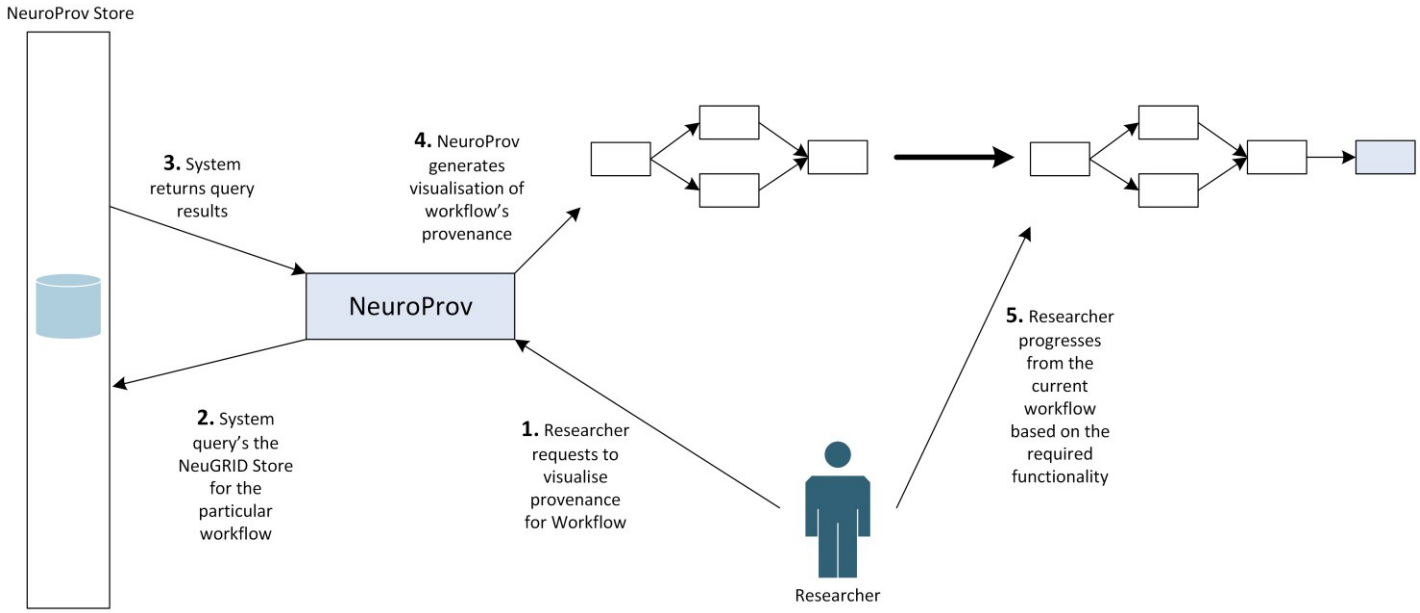


Figure 3.6 Progression Use-Case

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 3.6.

Indicative Use-Cases:

U5.1 Analysis of origin of results, to see how a workflow/dataset came into being so further analysis can be conducted upon it. E (Relevant to Intermediate and Advanced users)

Functional User Requirements

R5.1.1	Provide users with the ability to query and select a particular workflow/dataset.	D
R5.1.2	Provide users with the capability to browse through visualisation of workflow/dataset selected.	E
R5.1.3	The interface should be user friendly, and allow for browsing process by process provenance data	D
R5.1.4	Visualisation should allow intermediary output and associated provenance to be displayed	D

U5.2 Annotate the workflow/dataset with appropriate information to progress with the researcher's desired analysis. E (Relevant to Intermediate and Advanced users)

Functional User Requirements

R5.2.1	Provide users with the capability to annotate the workflow with useful information that might be helpful in future.	D
R5.2.2	The system should be able to provide researchers with useful information while progressing with analysis.	E
R5.2.3	Provide users with the ability to view annotations associated with the workflow to better understand the changes occurred	E
R5.2.4	Allow users to individually inspect complete provenance for workflows and associated details	E

3.4 Conclusion

This chapter has provided a detailed analysis of the requirements for visualisation of provenance data and how they are influenced by the needs of the users. The use-cases presented in this chapter are based on the experience the users require for analysis. In accordance with the use-cases actors have been identified and presented in detail along with an example user that will interact with our system. The requirements for NeuroProv have been traced out, establishing the basis for our research. Actors include research leaders, researchers, pipeline developers, image/data input managers and system administrators. Based on the users' knowledge and expertise certain features of NeuroProv will provide additional functionalities such as a fine grained view of provenance to an advanced clinical researcher. The relevant sets of users have been elicited along with the use-case requirements for better understandability of NeuroProv. Furthermore the following user scenarios have been identified keeping in mind the users elicited above. The scenarios included verification of results, validation of workflow, comparison of workflows/datasets, evolution of workflows/datasets and progression of workflows/datasets. Each set of requirements for use-cases have been marked with a prioritisation scheme that has helped us

to identify essential requirements for our visualisation system highlighting desirable features for our system.

The next chapter begins with the research methodology the research endeavour will follow in order to generate meaningful results and later the chapter presents users with the evaluation criteria to determine the efficacy of our proposed system NeuroProv.

Chapter 4

Research Methodology

Our research deals with visualising provenance data for neuroimaging analysis using N4U as a case-study. This chapter aims to explain the research methodology used for the purpose of our research work in detail along with the evaluation criteria used to determine the efficacy of our proposed system NeuroProv.

4.1 Research Methodology

This section describes the methodology undertaken for the purpose of our research endeavor. The following diagram explains the different elements of the methodology and how they interact with each other over the course of the research.

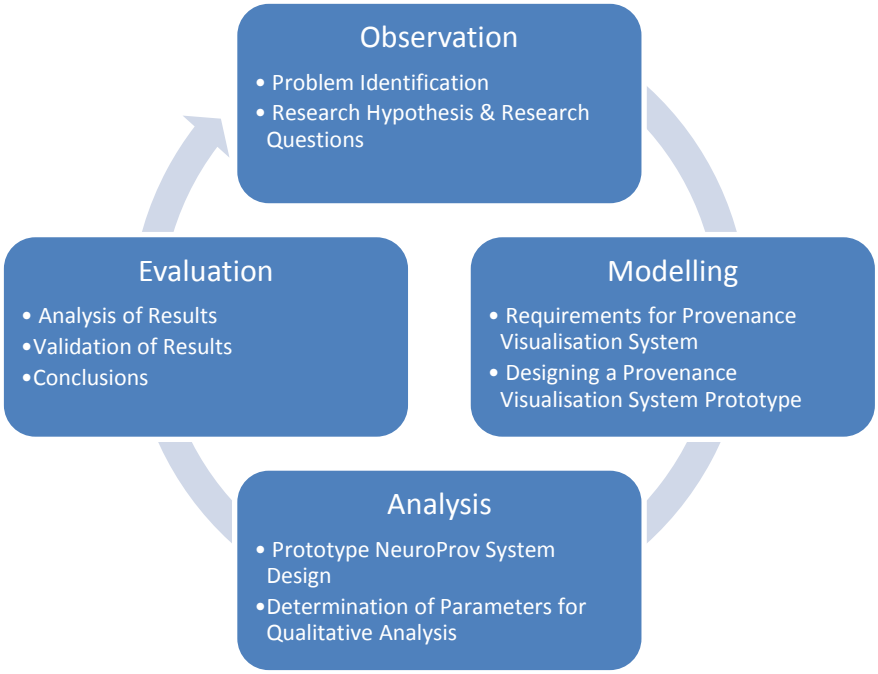


Figure 4.1 Research Methodology

Figure 4.1 shows that in the Observation phase, the shortcomings of existing provenance visualisation systems are identified from the aspects of collaborative analysis. This step is aided by a thorough literature review leading to the formation of our hypothesis and research questions.

Based on the outcome of the observation phase the modelling phase is carried out, which includes determining the requirements of the system that can overcome the limitations identified in the observation phase. The requirements are formulated based on how the users of our case-study N4U will interact with the system. A set of use-cases will be defined that will help generate functional requirements for visualisation of provenance data for neuroimaging analysis. The outcome of this phase is a proposed system for visualising provenance data. A prototype system will be developed in the Analysis phase and a set of experiments will be defined, which will be used to test the suitability of the proposed system against the elements of the hypothesis. The criteria for qualitative analysis will also be determined in this phase e.g. characteristics that may define the usability of our provenance visualisation system to a user. The Modelling and Analysis phases will benefit from our interaction with the potential users of the system i.e. scientists from the N4U community.

The results of the analysis phase will be analysed in the evaluation phase and will inform the answers to our research questions and test the validity of our hypothesis. We will also validate the results externally to assess if the results of this research can be generalised for other domains (in addition to neuroimaging related scientific analysis). This reflection on our research results requires iteration of earlier research phases and repetition of experiment steps. Finally, conclusions are drawn from the experimental results and qualitative analysis. The research will conclude by identifying and discussing future directions.

This thesis takes into account provenance data provided by the N4U system for the purposes of the research study. The use of the N4U as a case-study is a viable option for this research because of the accessibility of provenance data. N4U is an ongoing research project that provides real-time provenance data, generated for the purposes of neuroimaging analysis. The presence of ready-made provenance data being collected and stored for the objective of neuroimaging analysis is one motivation for the use of this data in this research. Workflows and associated provenance data have been chosen and verified to ensure the correctness of the resulting visualisations. One of the major goals of using case-studies is exploration [51] allowing researchers to observe how users interact with the system; use the system provided and respond to problematic situations.

The techniques involved in order to evaluate visualisations can be broadly classified into two main categories i.e. analytical evaluation and empirical evaluation. Analytical evaluation is based on formal models and is conducted by experts while empirical evaluation is realised through experiments with user test. Empirical evaluations can further be distinguished between quantitative and qualitative evaluations. While quantitative evaluation deals with analysis of determinate

hypotheses tested through direct measurements. In respect to that qualitative evaluation deals with analysis of qualitative data, which may be obtained through interaction with the potential users of the system to understand and explain social phenomena. For the purpose of our research study we will be dealing with qualitative analysis with a focus on testing the functional requirements elicited for the purpose of our research in Chapter 3. The following section describes the evaluation procedure followed to test that the functional requirements have been fulfilled and identifies set of metrics to test the suitability of our proposed system NeuroProv against our use-cases.

4.2 Evaluation Procedure

The approach we have taken 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 [44]. Visualisation approaches can thus be very individualistic and can vary from person to person based on their experiences. Consequently it is largely a matter of opinion whether one form of visualisation might work perfectly for one user whilst at the same time it may not even be appropriate for another user. Since the visualisation approaches are not testable in terms of purely quantitative measures we will be taking a qualitative approach to determine the efficacy of the approach taken for our research.

We will define a set of metrics that we will be using to evaluate the results generated from using NeuroProv to visualise provenance data for neuroimaging analysis. These metrics have been devised based on the requirements analysis carried out in Chapter 3. For the purpose of our research, a metric is defined as a means to assess the completion of our proposed system. Furthermore these metrics ensure designing of experiments, mentioned in the following chapter. In order to ensure that we have fulfilled the functional requirements elicited for provenance visualisation the following set of metrics help us to determine if these metrics hold true for the use-cases defined for our research study:

- Display the complete provenance for a given workflow
- Allow users to compare two workflows
- Allow users to compare more than two workflows at a time
- Display workflows as nodes, edges and relationships
- Highlight the trace for a particular entity or activity
- Provide the ability for users to drill down

- Provide annotations with workflows, nodes and edges
- Examine an entity's/activity's history
- Provide tracking of expanded stages
- Enable a search feature e.g. to identify workflows, datasets, nodes etc.
- View workflow and individual node's attributes
- Monitor a workflow's execution timeline
- View workflows in separate view for Progression and Evolution use-cases

NeuroProv is designed to allow users to:

1. Generate complete visualisation of a workflow, along with the ability to present metadata associated with the workflow such as workflow name, id, logical file name (LFN), owner, creation date etc.;
2. Allow users to highlight a trace for a particular activity/entity, along with its sources and targets (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);
4. Annotations support with workflows, entities, activities and links (for better understanding of provenance data for a naïve user);
5. Inspect an activity or an entity's history (information such as when was an entity is generated, using which input data etc.);
6. Providing users with the ability to track expanded stages when a user drills down (this gives users an overview of how farther they have drilled down in a workflow);
7. Timeline provides users with the ability to see when a particular entity/activity was generated and its associated details;
8. The ability to compare workflows and inspect visually encoded differences allow users to identify anomalies;
9. Users can visually see how a workflow has evolved over the period of time and what changes have occurred and
10. See how a user has progressed with a particular workflow and make changes to it if necessary.

NeuroProv provides users with comprehensive visualisation of neuroimaging workflows along with fundamental details such as the owner of the workflow, the date and time when the workflow was executed, its running time, input arguments, server address, host, directory, URL etc. All of these details are essential for verification and reproduction of each workflow. These details

are presented in tabular form next to any visualised workflow so the user can access the information when required.

NeuroProv allows users to visualise provenance and provides additional capabilities that help understand provenance data in a meaningful manner. For instance, insight often comes from comparing different workflows; unlike [24] NeuroProv allows users to compare multiple workflows. Workflows changes occurring over subsequent versions of a particular workflow are highlighted in a red colour. This capability enhances a user's experience by allowing him/her to identify difference between two or more workflows in a single inspection. Furthermore annotations provided with the comparisons allow users to determine what has caused the change to occur, at what instance and by whom. Highlighting changes in red provides users with the capability to investigate and verify the cause of a change that has occurred in different workflow versions and between similar workflows.

The ability to highlight a trace of a particular entity or an activity allows users to visually see what other activity or entity may have contributed to generate the output. Users can highlight the trace of a particular entity/activity by simply clicking on the entity/activity node. The opacity of the concerned links will be increased once a node has been clicked allowing better readability of the concerned activities and/or entities. This enhances the user experience by allowing them to trace backwards or forwards about all the contributing factors to a result. For neuroimaging analysis this is vital since it is essential to determine what caused an entity to be generated or consumed. Not only that for certain cases it is imperative to determine which activities were involved. Furthermore appropriate annotations provided with the workflow add to the understandability of the workflow components and the result is essential for verification and reproducing an experiment.

NeuroProv's ability to drill down into a workflow to view complete and detailed information allows users to view a high-level summary of the workflow in the first stage and then the user can progressively drill down as he/she proceeds to view further detail. The high level summary provides naïve users with a basic level understanding of the specific workflow. Experienced users such as Research Leaders, Researchers and Pipeline Developers (the roles identified in Chapter 3) can further drill down to view detailed information that will help to completely understand the 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 view a high level summary of the workflow.

The 'search' feature in NeuroProv allows users to discover information about a particular workflow or a 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 whilst verifying the result with the search bar provided on the top of the screen. The request is sent to the NeuroProv Store which will provide appropriate query results to generate a visualisation of the 'searched' workflow.

The NeuroProv workflow execution timeline feature provides users with the ability to determine when a workflow started, ended or when a particular entity/activity was generated. Other layout approaches do not leverage temporal ordering inherent to provenance graphs. This unique feature is useful in NeuroProv whilst verifying a workflow since the time-worthiness of an entity's usage or production is essential to determine if the resultant entity is appropriate or not. Inspection of start and end times reveals whether the workflow was reproduced in the same amount of time as the original workflow since the execution time is one of the measures for execution performance.

4.3 Conclusion

The research methodology followed in this research study is explained in this chapter along with the evaluation procedure for the results generated in Chapter 6. The evaluation procedure identified and explained in the previous section will be used to evaluate the results generated using our proposed visualisation system NeuroProv. The next chapter describes the architecture of our proposed system NeuroProv and later explains the experimental setup used to generate the results for our research.

Chapter 5

NeuroProv Architecture & Experimental Setup

The purpose of this chapter is to provide detailed information about the NeuroProv's architecture, test environment and the framework used to perform experiments in order to validate the work carried out in this thesis. These experiments visualise provenance of actual neuroimaging workflows from Pegasus [50], generate provenance visualisations of workflows based on use-cases and execute several tests to evaluate the utility of visualising provenance data for neuroimaging analysis.

5.1 NeuroProv Architecture

NeuroProv System is developed in context of N4U. Some of the basic elements in N4U Virtual Laboratory are shown in Figure 4.1.

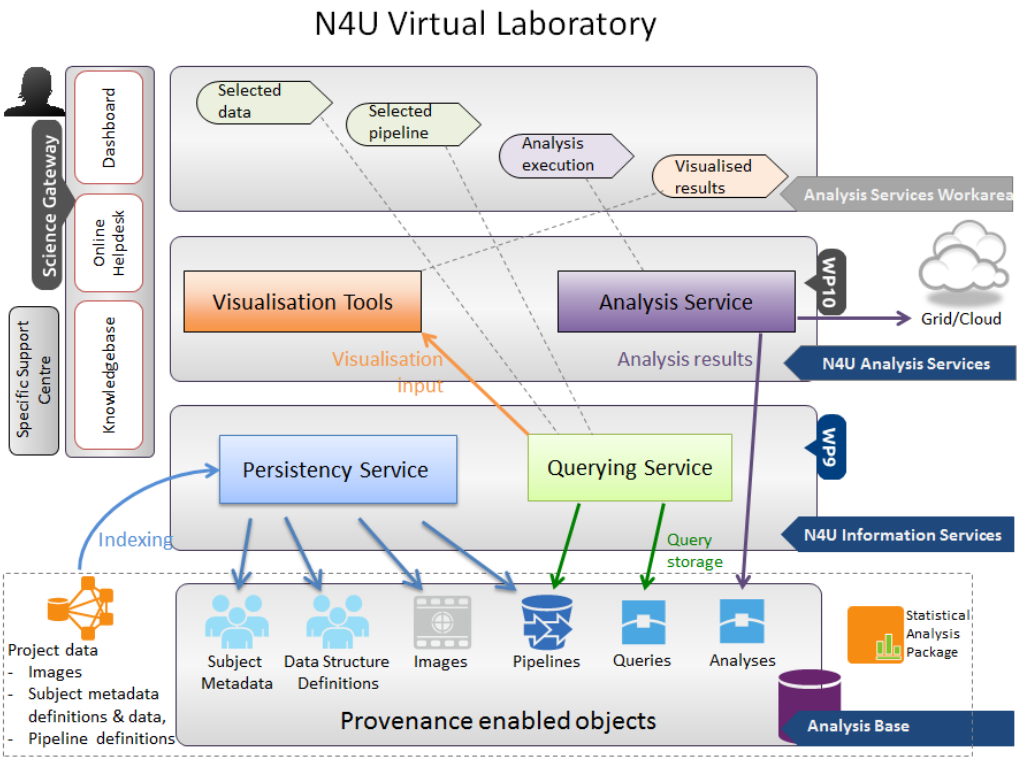


Figure 5.1 N4U Virtual Laboratory [43]

The N4U virtual laboratory provides services and tools to perform analyses over the Grid. As mentioned in section 1.4 N4U provides computing and storage infrastructure and services on top of stored neuroimages (Analysis Bases as shown in Figure 4.1) and to facilitate neuro-researchers in defining and executing their neuro analysis on stored images. Analysis Service Work area provides users with visualised provenance of workflows and other related information based on user's requirements. N4U provides visualisation capability with the support of NeuroProv. NeuroProv generates visualisation of provenance data by querying the NeuroProv Store for the desired results (shown in Figure 4.2). The query results are generated based on user's requirements to verify a workflow, compare two or more workflows, analyse how a workflow has evolved over a period of time or how a workflow has progressed so the user can make use of it. NeuroProv then generates appropriate visualisation based on the query results.

Figure 4.2 provides simplified version of the N4U Virtual Laboratory to understand the working on NeuroProv in context of N4U.

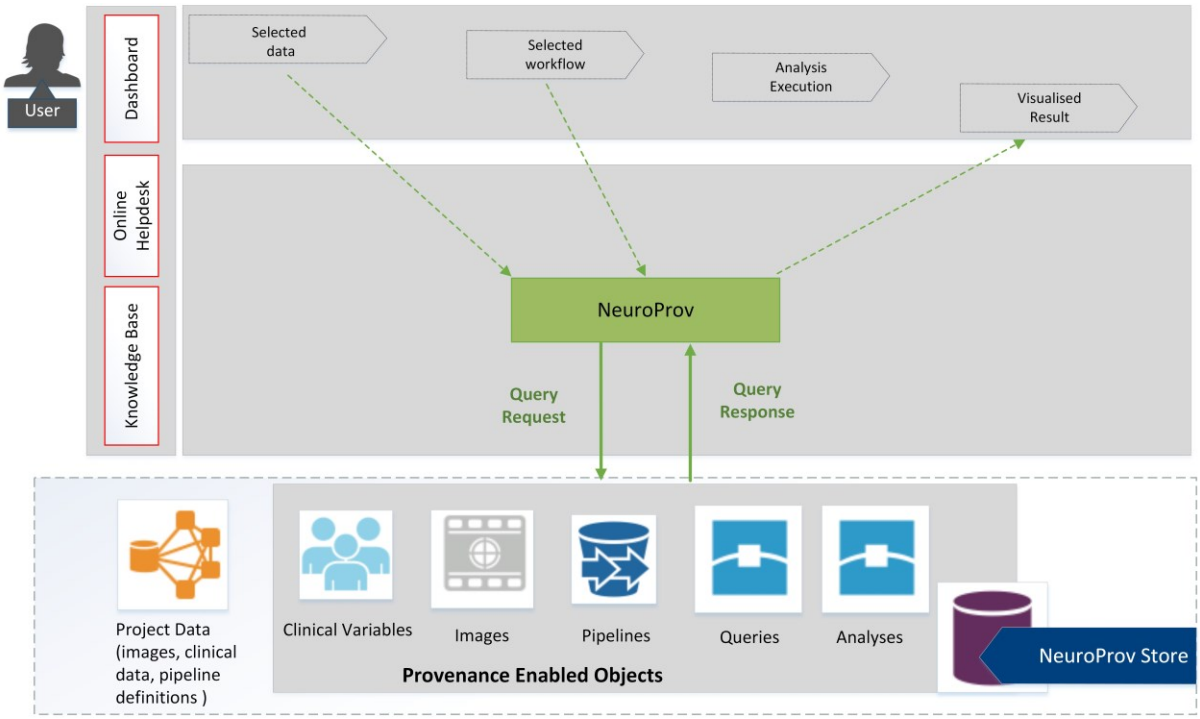


Figure 5.2 NeuroProv in context of N4U

NeuroProv allows researchers to select the purpose to generate visualisation which can be from the following; select a workflow to validate its result; generate visualisation to compare two or more workflows to identify difference and anomalies (if any); verify and analyse how a particular workflow has evolved over the subsequent versions and changes occurred; and analyse how a set of

workflow progress over a period of time that allow users a base to start experimentation. NeuroProv uses the Sankey toolkit provided by Mike Bostock [40] in order to generate Sankey Diagrams for the purpose of this research (see chapter 5). The toolkit provides the ability to generate Sankey diagrams from provenance data generated and stored for the purpose of our research study. The NeuroProv Store (a subset of Analysis Base containing provenance for the purpose of our research) provides us with the provenance data, associated metadata and annotations being stored for workflows.

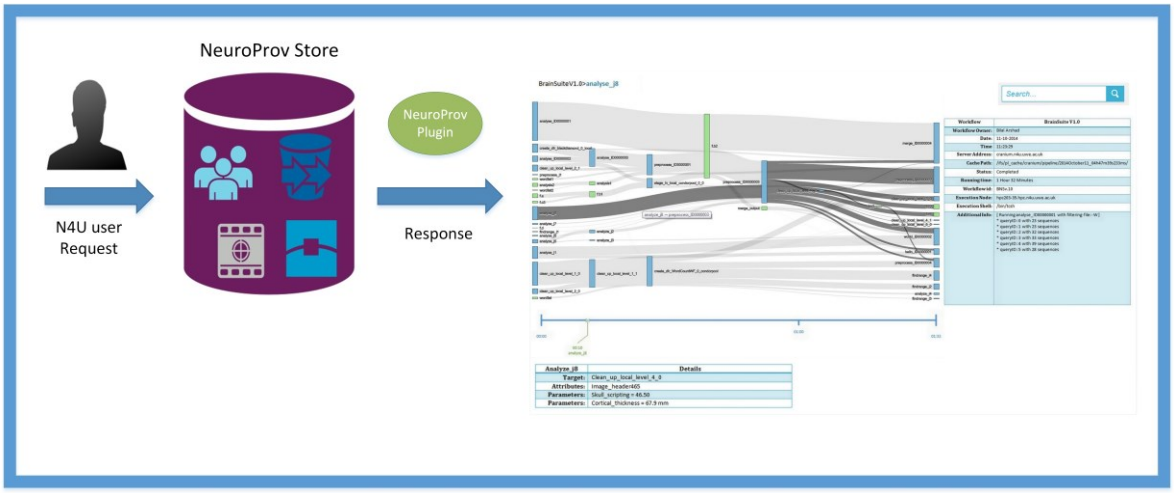


Figure 5.3 NeuroProv Architecture

Figure 4.3 describes NeuroProv's Architecture in detail.

NeuroProv Store: Currently NeuroProv Store contains provenance information and associated metadata from workflows being run in Pegasus [50]. NeuroProv Store provides access to previously performed analyses, already defined datasets and/or images produced over the past analyses. This allows users to generate visualisation based on provenance data available.

NeuroProv Plugin: The NeuroProv plugin ingests provenance information sent to it as 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). NeuroProv in a nutshell works with the NeuroProv Store in N4U in order to generate user queries and provide response in a JSON format to be consumed by NeuroProv.

Provenance Visualisation: Since D3 is a JavaScript based library it allows visualised Sankey diagram to be presented in browsers. We have used Firefox (Version 33.0.3) and Google Chrome

(Version 38.0.2125.111 m) for displaying visualisation of provenance. NeuroProv supports all major browsers except for Internet Explorer since d3.js lacks support for it.

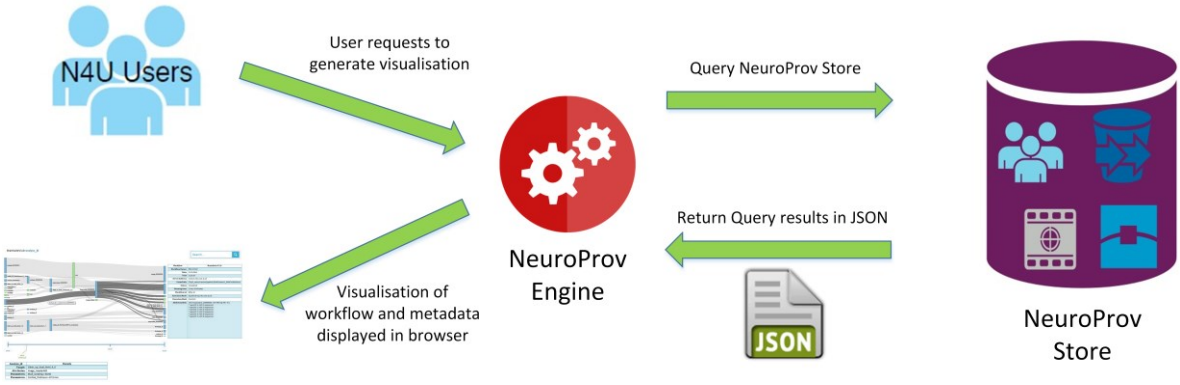


Figure 5.4 Flow of activities in NeuroProv

Figure 4.4 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 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 engine in a JSON response. NeuroProv then utilises the JSON response to generate Sankey diagrams to be visualised in the browser, thus providing users with visualisation to perform analysis.

5.2 Experimental Setup

Visualisations were generated using NeuroProv over multiple workflows from the NeuroProv Store to analyse and understand if the required functionality were met. Twenty workflows of varying size and data were selected, each workflow contained multiple processes and images. Since it is not possible to obtain a truly random distribution within a huge set of workflows, the following strategy was adopted for selection of workflows. The workflows were characterised based on their functionality and type of workflows. There are currently five major types of workflows in the NeuroProv Store. There are multiple versions of these workflows i.e. about 131 workflows currently residing in the NeuroProv Store.

In order to generate results we have divided the workflows into 4 case studies, each containing multiple workflows based on their resultant visualisation. The following table provides a matrix of use-cases plotted against the case-studies. The numbers in the cells correspond to the number of workflows visualised during the experimentation phase while the numbers inside the brackets correspond to the workflow id of the equivalent workflows. This provided us the basis of

experimentation, while the results of these analyses and visualisations are provided in the next chapter. Some of the visualisations are presented in the next chapter of the thesis while the rest of the corresponding visualisations are presented in the appendix.

Table 1 Experimentation Matrix

	Case-Study 1 (CS1)	Case-Study 2 (CS2)	Case-Study 3 (CS3)	Case-Study 4 (CS4)
Verification - Use-Case 1 (UC1)	2 (Workflow id: 39, 94)	2 (Workflow id: 1, 90)	3 (Workflow id: 90, 91, 93)	1(Workflow id: 1)
Comparison - Use-Case 2 (UC2)	6 (Workflow id: 35, 38, 39, 94, 96, 132)	3 (Workflow id: 35, 38, 39)	3 (Workflow id: 90, 91, 93)	-
Evolution - Use-Case 3 (UC3)	10 (Workflow id: 94, 95, 96, 105, 106, 110, 111, 113 123, 132)	4 (Workflow id: 35, 38, 39, 41)	3 (Workflow id: 90, 91, 93)	-
Progression - Use-Case 4 (UC4)	7 (Workflow id: 94, 96, 105, 106, 110, 123, 132)	4 (Workflow id: 35, 38, 39, 41)	3 (Workflow id: 90, 91, 93)	-

Provenance for each of these experiments was manually verified before visualising it. This is being done to ensure that appropriate provenance is being visualised for the associated workflow. Furthermore additional provenance data is being used in order to generate visualisation to provide appropriate detail alongside the visualisation.

5.3 Conclusion

This chapter provided a detailed summary of NeuroProv and its associated elements. It then provided a systematic diagram of NeuroProv's working in context of N4U Virtual Laboratory along with NeuroProv's architecture. The chapter then presented the experimental setup and the experimentation matrix within which it has been designed to generate results and analyse the results in order to enhance the utility of provenance data for neuroimaging analysis. The following chapter provides results and analysis based on the workings of NeuroProv within the context of N4U. These results are generated based on the case studies defined and number of workflows being visualised in order to fulfil the requirements for neuroimaging analysis. Furthermore the results are tested against the metrics defined in Chapter 4.

Chapter 6

Results and Analysis

This chapter provides results based on an evaluation of NeuroProv, our visualisation system, using N4U as the case-study. We will first introduce the 'Sankey Diagram' [40], which is a technique we will be using to visualise provenance data for neuroimaging analysis. This provides us with a basis for testing our hypothesis that 'Visualisation techniques can enhance the utility of provenance data for neuroimaging analyses'. This chapter will then provide an analysis based on the results in order to establish if the hypothesis defined for our study holds true or false. The results of the experimentation will be presented based on the experimentation matrix defined in Chapter 4 (Table 1) followed by the detailed analysis of results of the experimentation carried out. The outcomes of the research are based on the use-cases defined in Chapter 3.

6.1 Sankey Diagram

NeuroProv uses a visualisation technique known as the Sankey Diagram to visualise provenance data for neuroimaging analysis [40]. Going back to our research question, we will demonstrate that this approach enables us to answer our second research question i.e. 'which visualisation technique will be suitable for visualisation of provenance data for neuroimaging analyses?'. Several works have focused on the visualisation of provenance using a number of presentation paradigms including network, data flow diagrams and radial layouts [45] [46] [47] however Sankey Diagram is an appropriate choice for provenance visualisation due to the following reason. A Sankey diagram is a type of flow diagram where the 'flow' is represented by arrows of varying thickness depending on the quantity of flow. The aim of the research is to visualise provenance in an intuitive manner to highlight complete trace of provenance which is possible with Sankey Diagrams. They are often used to visualise energy, material or cost transfers and they are especially useful in demonstrating proportionality to a flow where different parts of the diagram represent different quantities in a system. Probably the most famous example of a Sankey diagram is Charles Minard's Map of Napoleon's Russian Campaign of 1812 (figure 5.1 below).

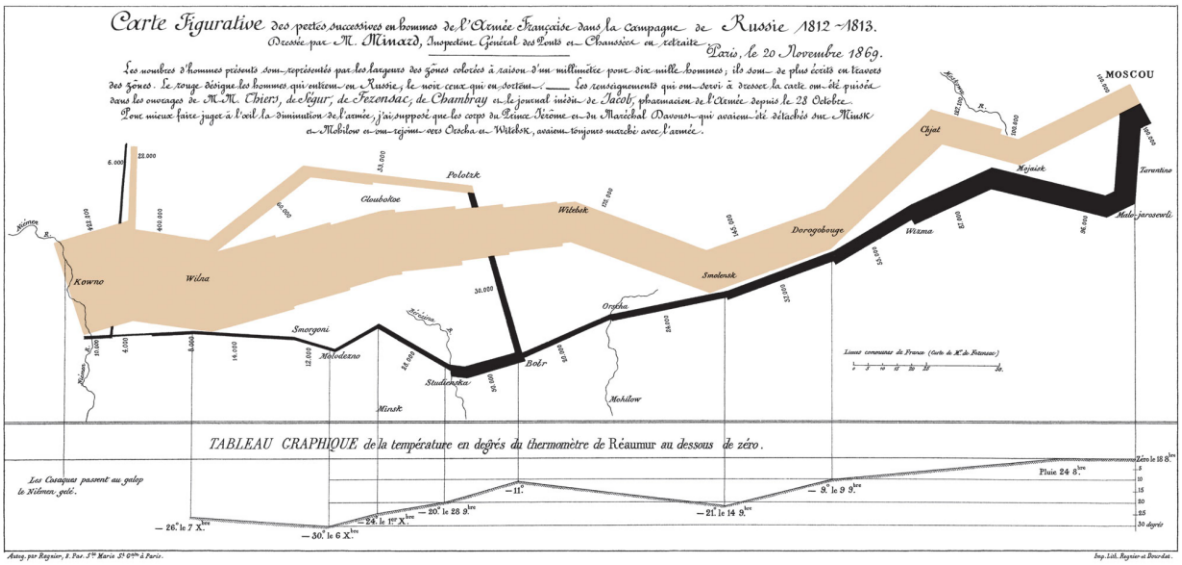


Figure 6.1 Napoleon's Russian March [48]

"*Etienne-Jules Marey first called notice of this dramatic depiction of the fate of the Napoleon's army in the Russian campaign, saying it all defines the pen of the historian in its brutal eloquence. Edward Tufte says it "may well be the best statistical graphic ever drawn" and uses it as a prime example in The Visual Display of Quantitative Information*" [48]

Sankey Diagram was named after the Irish Captain Mathew Henry Phineas Riall Sankey who used this diagram in 1898 showing energy efficiency of a steam engine. Sankey Diagram enables users to highlight edges between different nodes by varying link opacity. Without link opacity, the meaning of these relationships will be obscured. Furthermore Sankey Diagrams are interactive allowing users to move around nodes to understand the relationships between important nodes. One of the major drawbacks of using Sankey Diagram is that it does not support cycles. Since provenance in neuroimaging is non-cyclical it provides accurate rendering of provenance data for the purpose of our research.

The Sankey Diagram has been chosen to visualise provenance data for NeuroProv since we view a provenance graph as a network of activities where data flows through and between activities. Our 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. In a standard, directed acyclic graph (DAG) rendering, the flow of information can get easily lost in a large network. Other layouts, for example radial layouts, tend to focus on the interconnectivity of data or activities. Furthermore, other layout approaches do not leverage the temporal ordering

inherent in provenance graphs [41]. The next section presents the results of the research work carried out.

6.2 Results & Analysis

This section will present the results of the experimentation being done in order to generate a suitable visualisation of provenance data for neuroimaging analysis. We will be using the experimentation matrix (Chapter 4, Table 1) as a reference to present our results along with details related to the visualisation that will provide readers with appropriate information to interpret the results and understand the workings of NeuroProv. We will provide readers with a few visualisations from each of the Use-Cases pertaining to the Case Studies in this chapter along with their explanation. The rest of the visualisations will go in the Appendix A. Evaluation metrics will be highlighted for each set of results in order to evaluate the capability of NeuroProv for the particular use-case.

As a starting point, when looking at the Sankey Diagram representations (see for example figure 5.2), the blue rectangles represent the various activities in the workflow while the green rectangles represent input/output entities consumed or generated over the course of the workflow. The width of a link between the rectangles represents the amount of information flow between the two nodes (activities and entities). Initially, once the visualisation of a workflow has been generated, NeuroProv provides users with a basic view of the workflow provenance with link opacity of each link between the nodes set to 0.2. Once the user selects a particular node (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 rest of the links remains 0.2. This helps the users to visually differentiate between the highlighted and the non-highlighted links in the visualisation; making it visually observable to differentiate and examine the source and/or target nodes.

The following sections represent visualisations generated from NeuroProv based on individual case-studies along with detailed information to interpret results and understand features of NeuroProv that will help to prove the research hypothesis (from Chapter 1). The first Use-Case ‘Verification of Workflow’ is presented below:

6.2.1 Verification - Use-Case 1

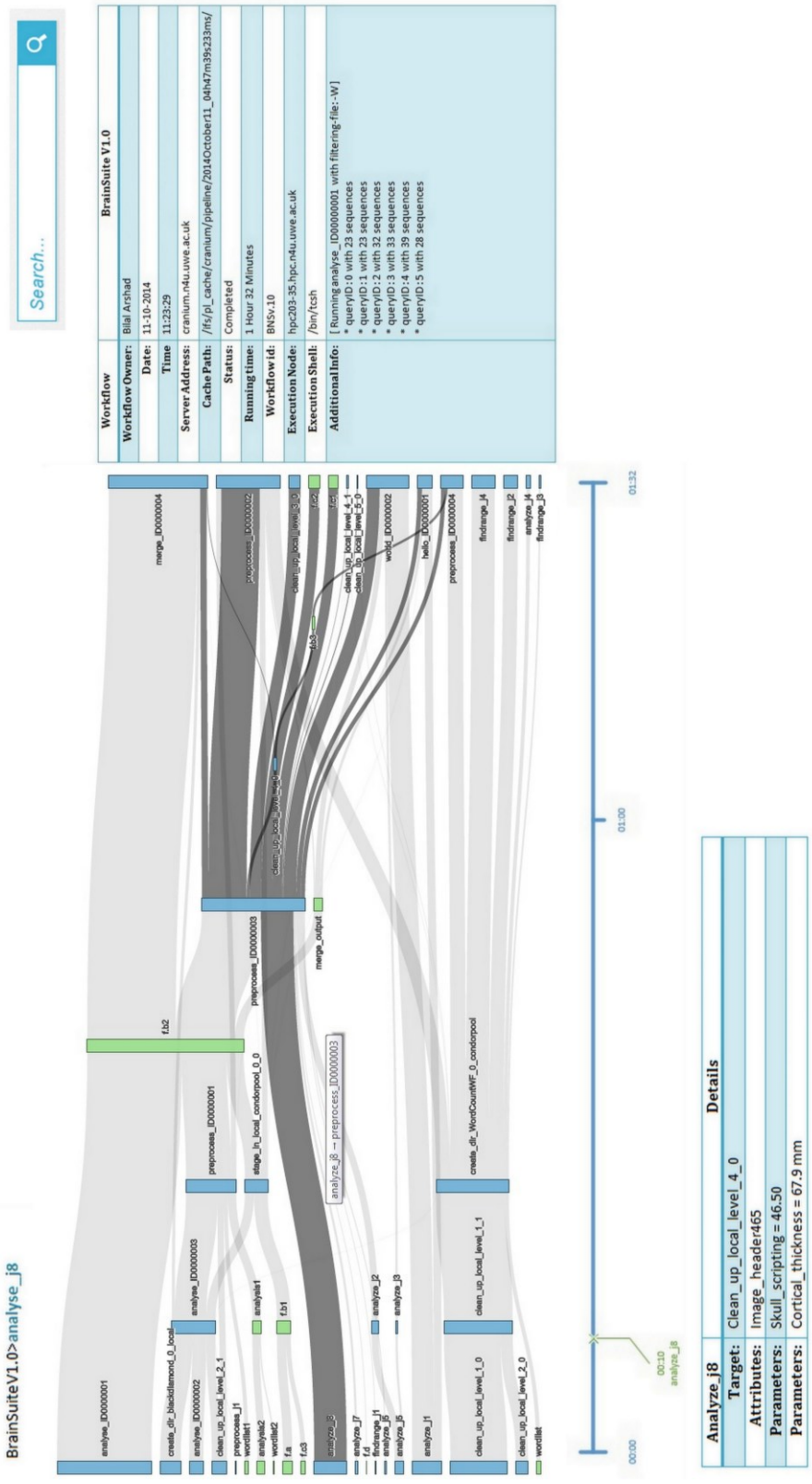


Figure 6.2 NeuroProv Verification Use-Case Screenshot

This section will present the first use-case i.e. the 'Verification of Workflows' with focus on Case-Study 1, the results of Case-Study 2 will be presented in Appendix A. Visualisations generated by NeuroProv for the purposes of verification will be presented along with other details essential for an understanding of the workings of the systems and details of features essential for provenance visualisation. Figure 6.2 shows the BrainSuite workflow along with the workflow details in the table accompanying the workflow. Essential information for workflows such as the owner of the workflow, the date and time, when the workflow was executed, the server address, the status and other details are listed. These details are essential to verify that the workflow has taken the exact amount of execution time, for each run, if provided with the same environment for the workflow to be executed. Furthermore provenance data is required to ensure that the user has been able to generate exactly the same results as anticipated when running the same workflow repeatedly with the same parameters. All this information helps the user to identify anomalies, to predict sources of error and to rectify wherever and whenever possible.

Based on the intial view provided by NeuroProv the user can visually identify the activities and entities (files in the case of this workflow) by the Sankey Diagram generated. This basic view presents the flow of data between appropriate activities and entities present in the workflow. The user then selected the 'analyze_j8' activity whose details are presented alongside with its attributes. NeuroProv's ability to provide users with detailed metadata information regarding a particular node (an entity or an activity) makes it highly desirable since this information is used to determine the correctness of a workflow on the whole. Below that NeuroProv provides the timeline feature which allows users to view when the activity was generated/consumed within the context of the workflow execution time. By clicking the 'analyze_j8' activity node the user can view which activity consumed it and which activities/entities then led to generation of further entities/activities. This is shown by the highlighted paths starting from the 'analyze_j8' activity to the subsequent activities/entities in the workflow. In the following sections relevant evaluation metrics will be highlighted in a table at the end of each section for the readers to view how NeuroProv has been evaluated in terms of use-case requirements. This will help readers to evaluate the efficacy of NeuroProv in relation to the requirements essential for neuroimaging analysis.

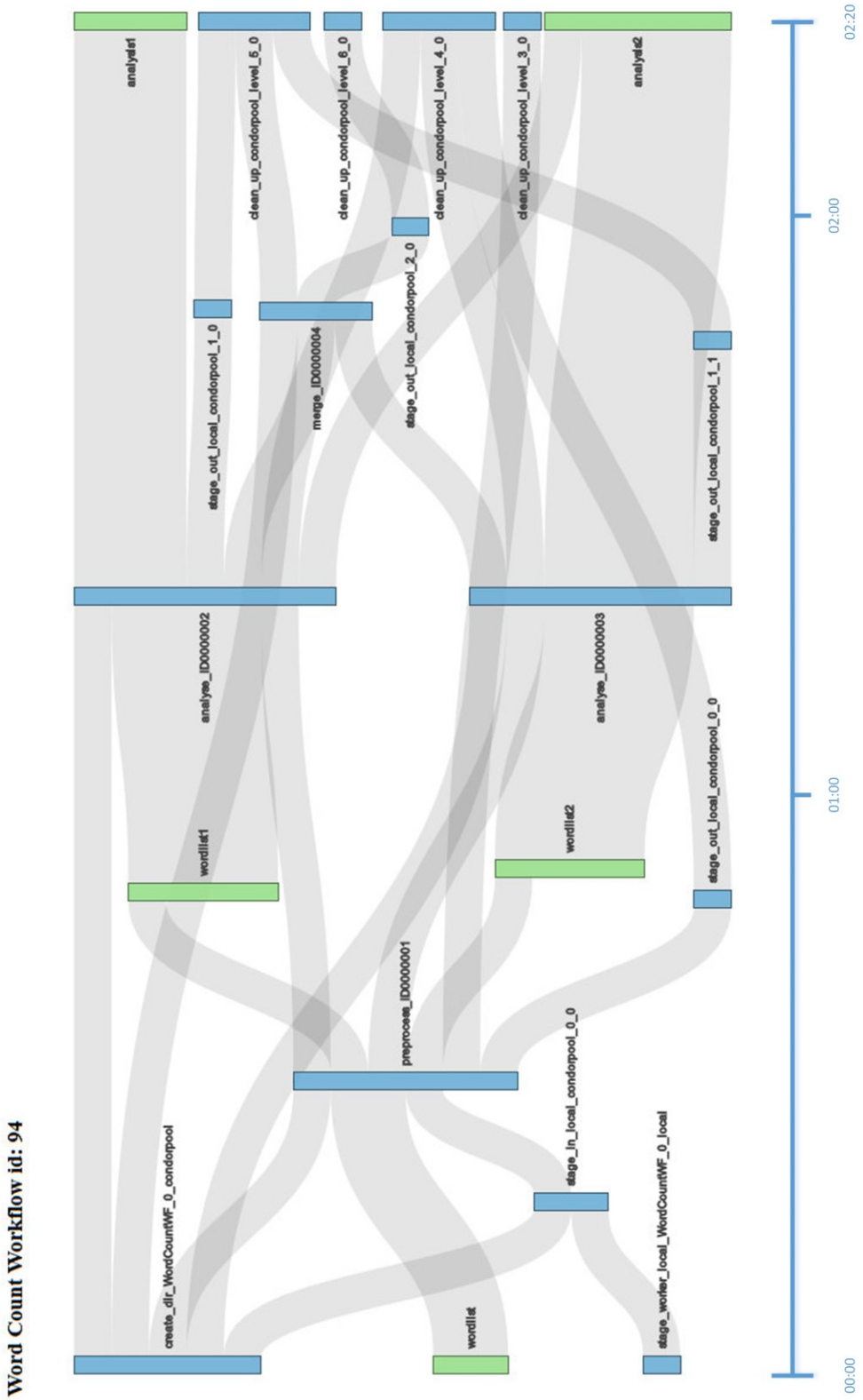


Figure 6.3 Workflow Verification Wf_id=94

Figure 6.3 shows the visualisation of provenance data for Workflow id 94. This workflow contains 15 different activities and five different entities that are either consumed or generated over the course of the workflow. Figure 5.4 shows that a user can click on one of the nodes for instance to inspect how an entity is produced and later consumed by different activities. The entire path is highlighted providing users with the ability to identify which activities took part in the generation or consumption of that entity and all the intermediary entities that are generated as a result of an entity under consideration. This is a useful technique for neuroimaging analysis since it allows researchers to verify how an entity came into being and how it has been consumed during the course of the workflow. Clicking on the node also allows users to inspect any metadata available with that particular entity/activity. Intermediate and advanced users can add metadata information based on their experience to provide an improved understanding of the workflow for future potential users.

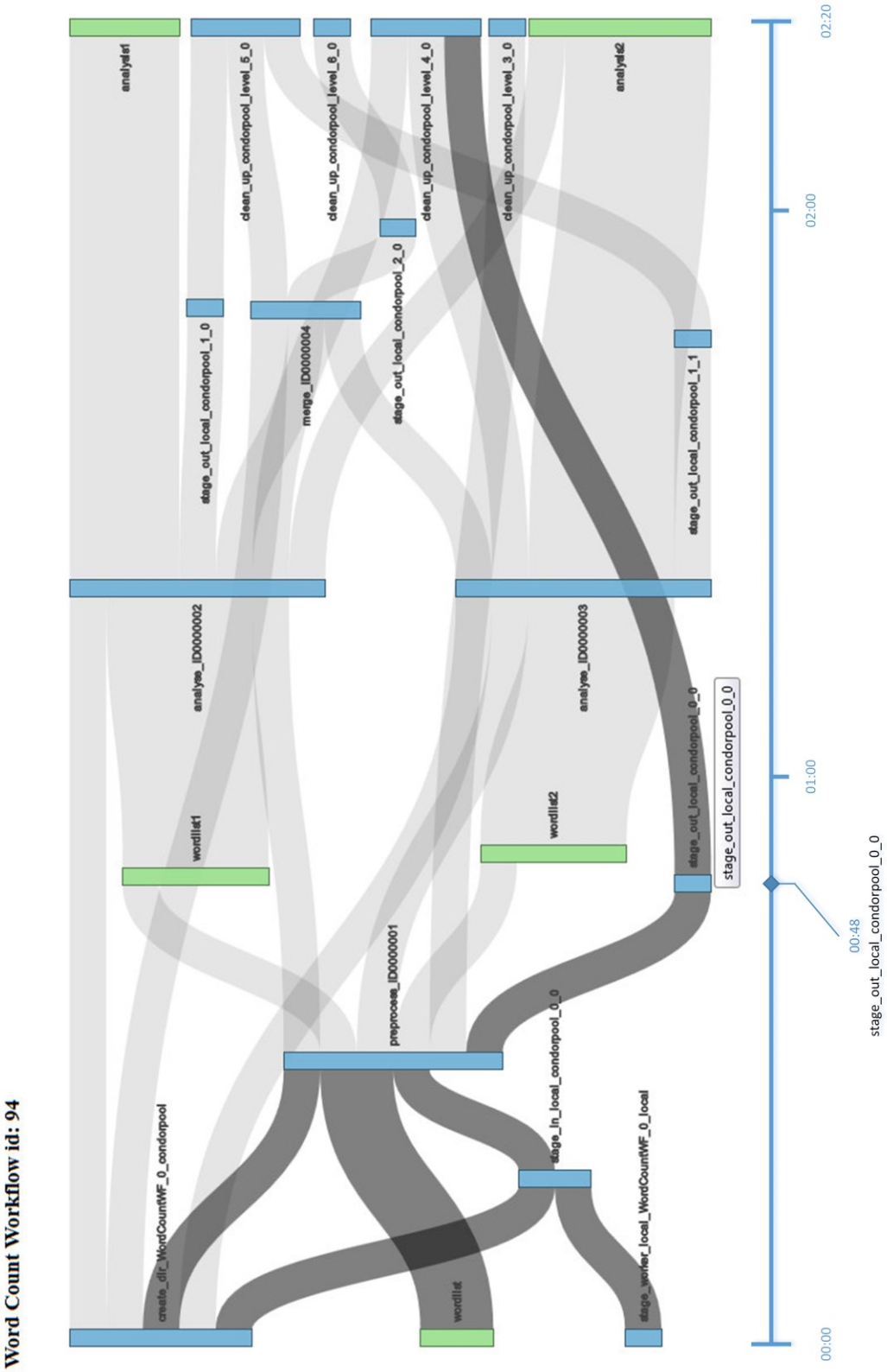


Figure 6.4 Workflow Verification with activity under inspection

As an example, in the WordCount Workflow id number 94, the user has clicked on the 'stage_out_local_condorpool_0_0' activity and NeuroProv highlights the trace for that activity. Figure 6.4 shows the trace highlighted for the activity under inspection. The activity is generated by 'preprocess_ID0000001' and led to the generation of another activity i.e. 'clean_up_condorpool_level_4_0'. During the execution of the workflow one file was also consumed for 'stage_out_condorpool_0_0' i.e. 'wordlist' highlighted in green on the extreme left of the visualisation. The following activities led to the generation of the activity under inspection namely 'create_dir_WordCountWF_0_condorpool'; 'stage_worker_local_WordCountWF_0_local'; 'stage_in_local_condorpool_0_0' and 'preprocess_ID0000001'. All the activities that are highlighted in the trace are in blue. Furthermore NeuroProv also provides the user with the ability to view when an entity or an activity was generated in the context of the workflow execution timeline. Our activity 'stage_out_local_condorpool_0_0' was generated at the 48th minute of the execution time whilst the entire workflow took 2 Hours and 20 minutes to successfully execute. The following image (Figure 6.5) provides a screen capture of the annotations provided by NeuroProv once the user clicks on the 'stage_out_local_condorpool_0_0' activity.

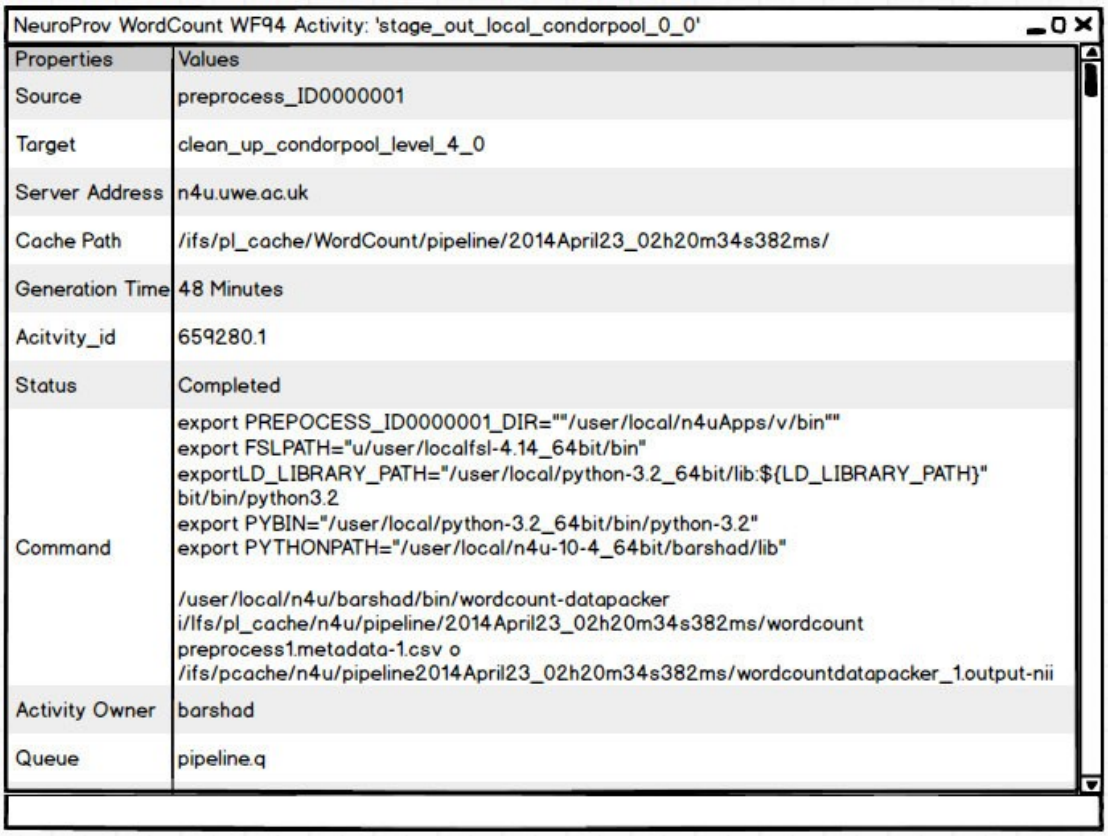


Figure 6.5 stage_out_local_condorpool_0_0 details NeuroProv

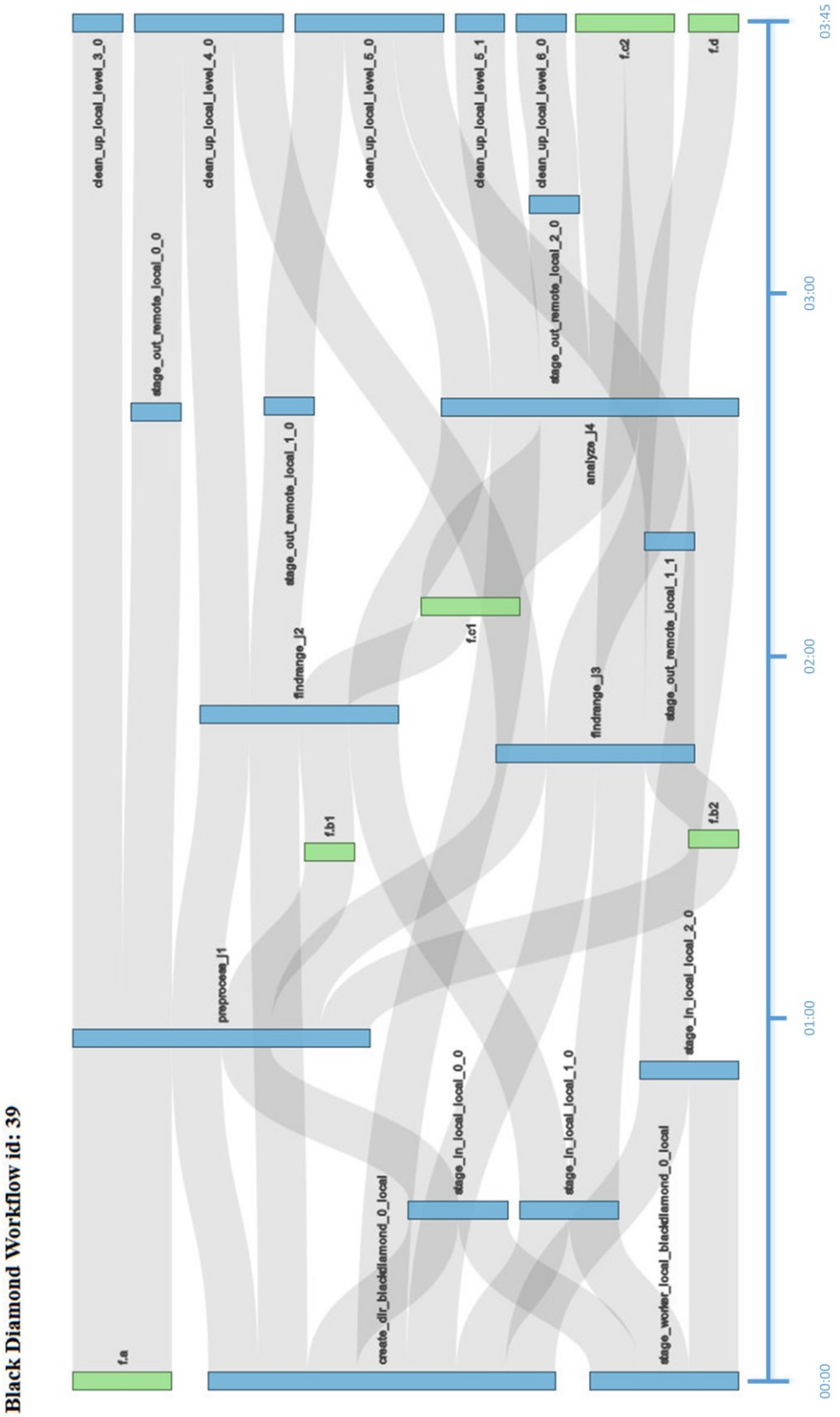


Figure 6.6 Workflow Verification Wf_id=39

Figure 6.6 shows the visualisation of provenance data for the Black Diamond workflow with id 39. The workflow contains 18 different activities and six different entities that are either generated or consumed during the course of the workflow. Clicking on one of the activities for instance 'stage_out_remote_local_0_0' allows us to inspect which activities have led to the generation of this activity and which entities are generated and consumed in order to get to this activity. Figure 6.7 below reveals the path highlighted for the intermediate image entity 'f.b1' and the subsequent activities that led to the generation of this entity (coloured green in the centre of the figure). Figure 6.7 shows one of the entities clicked by the user to inspect provenance. The entire trace for that particular entity has been highlighted to allow users to completely inspect the provenance for that entity. Furthermore annotations added by advanced users are shown by NeuroProv in a separate window (screen capture presented following the highlighted image on the next page).

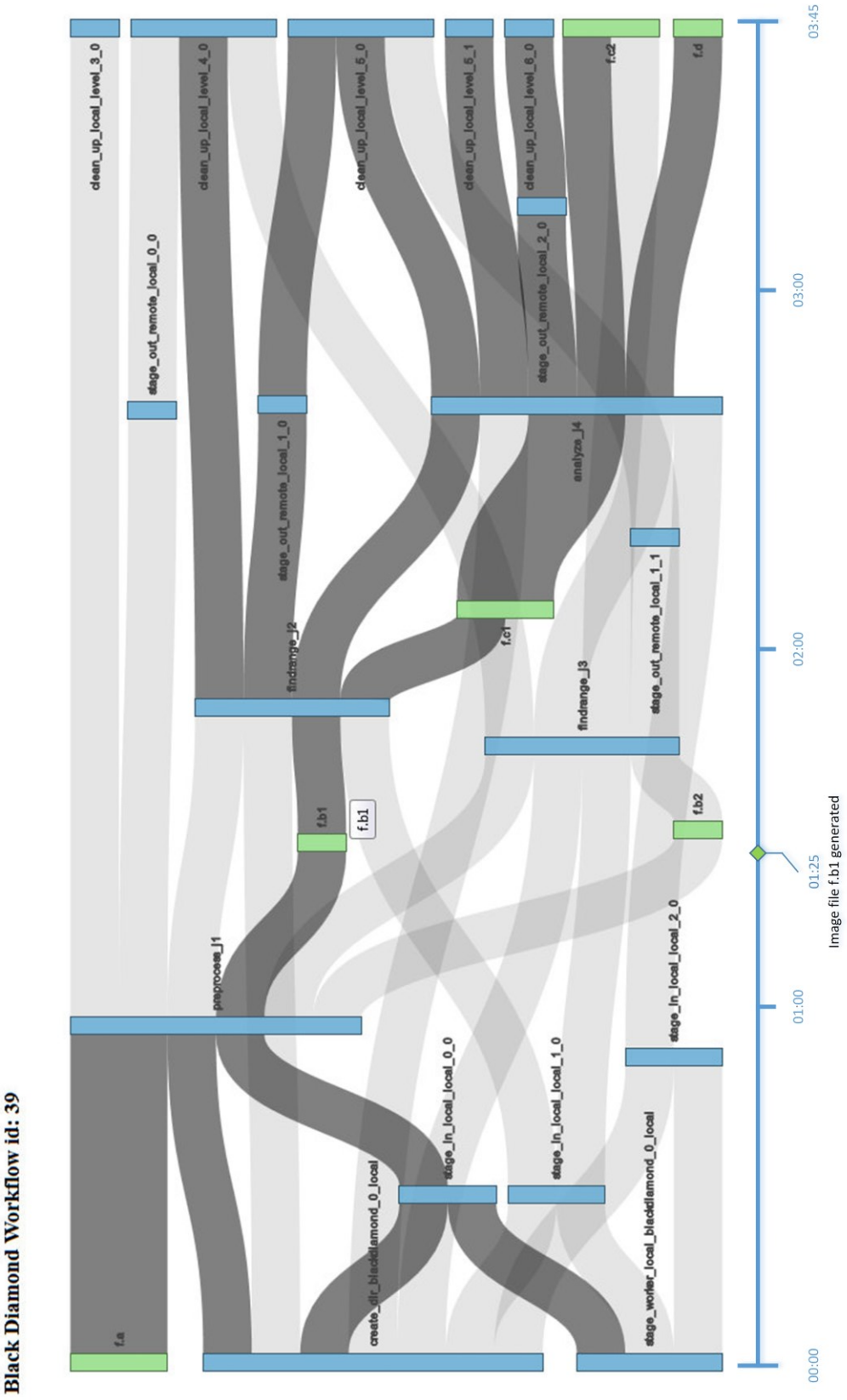


Figure 6.7 Workflow Verification Wf_id =39 with selected activity in inspection

The user intends to inspect the provenance for entity f.b1 (intermediate image file). Once the user clicks f.b1 in Figure 6.7 the trace is highlighted and any annotations provided are shown on the screen. In Figure 6.8 the annotations associated with f.b1 are shown, details such as when the activity was running, the source and target of the entity etc. are described in a table along with other metadata information present in the provenance store that has been added by an advanced user. The intermediate image of a brain scan that is generated as a result of workflow execution is also presented.

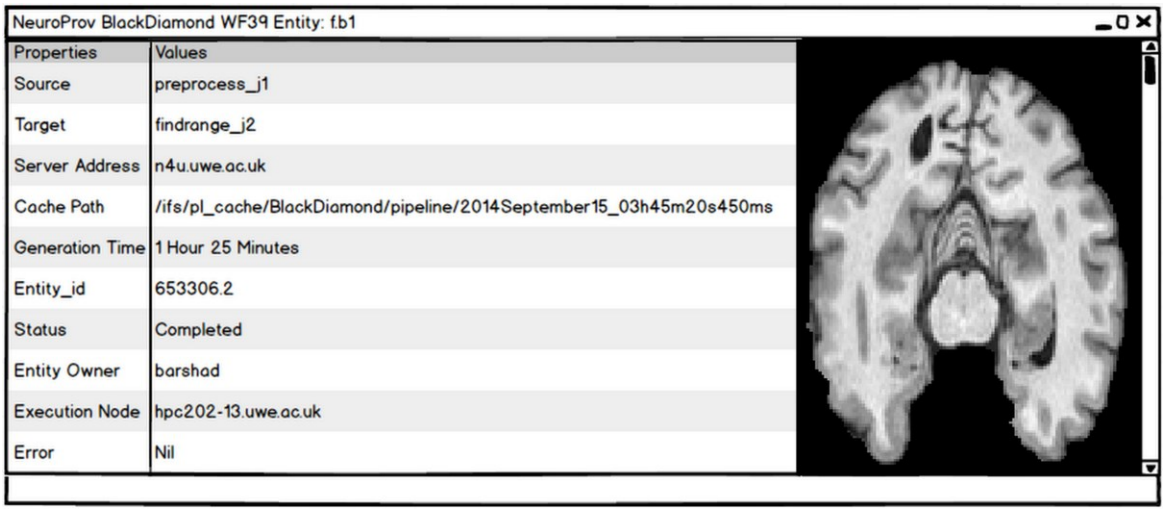


Figure 6.8 Image file f.b1 annotation provided with intermediate image

In table 2 (below) all the relevant metrics for the Verification of Workflows Use-Case have been enlisted and checked upon to conclude that NeuroProv exhibits the functionality required for the verification of workflows.

Table 2 –Metrics for Verification of Workflows

Sr No	Metrics	Verification of Workflows
1	Display complete provenance for workflows	✓
2	Display workflows as nodes, edges and relationships	✓
3	Highlight trace for a particular entity/activity	✓
4	Ability for users to drill down	✓
5	Provide annotations with workflow, nodes and edges	✓
6	Examine entity/activity' history	✓
7	Provide tracking of expanded stages	✓
8	Search feature e.g. workflow, dataset, nodes etc.	✓
9	View node attributes	✓
10	Workflow execution timeline	✓

6.2.2 Comparison – Use-Case

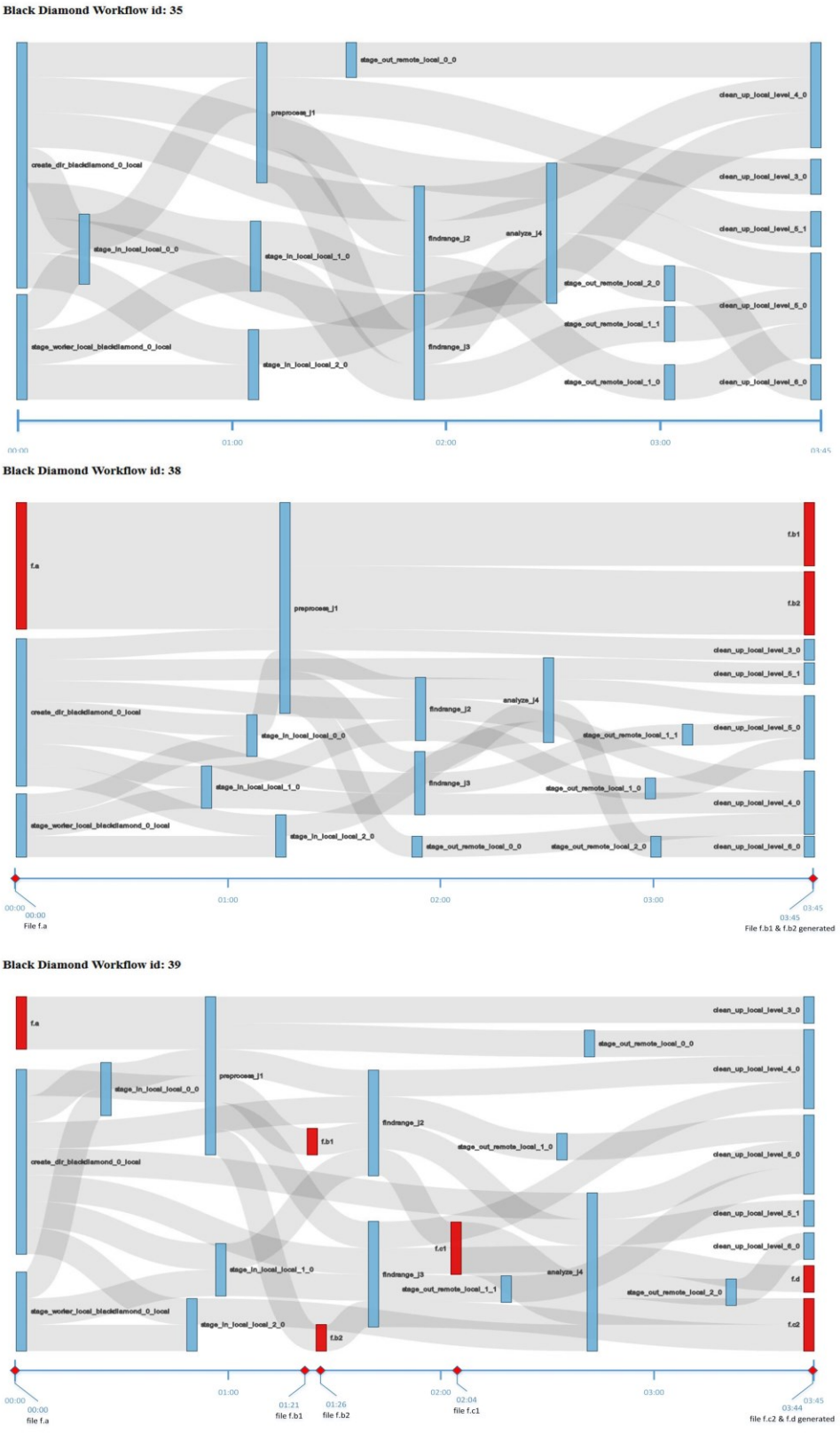


Figure 6.9 Multiple Workflow Comparison Black Diamond Workflow id 35, 38 & 39

Figure 6.9 shows a comparison of multiple workflows: workflow id 38 and 39 compared against workflow id 35. NeuroProv allows intermediate and advanced researches to compare multiple workflows to identify the differences and anomalies within the workflows. Figure 6.9 show's that comparing workflow 38 and 39 against workflow 35 identifies the addition of one or several different activities and entities. All the different entities and activities are highlighted in red to allow users to visually compare the different workflows and to record their opinions along with the other metadata information thereby helping other users over time. In case some other users want to generate a comparison amongst the same workflows the annotated information will be helpful for him/her providing useful information regarding the comparison thus saving valuable time for other tasks. Furthermore the annotated information provides a basis for authentication for publishing results in a journal/conference paper. Multiple workflow comparisons allow researchers to inspect the differences within multiple workflows in a single view thereby relieving the user from the inconvenience of generating multiple workflow visualisations to compare against a single workflow. Metadata information and annotations provided with the workflows further enhance the user's experience by providing enriching information helpful for researchers making a comparison.

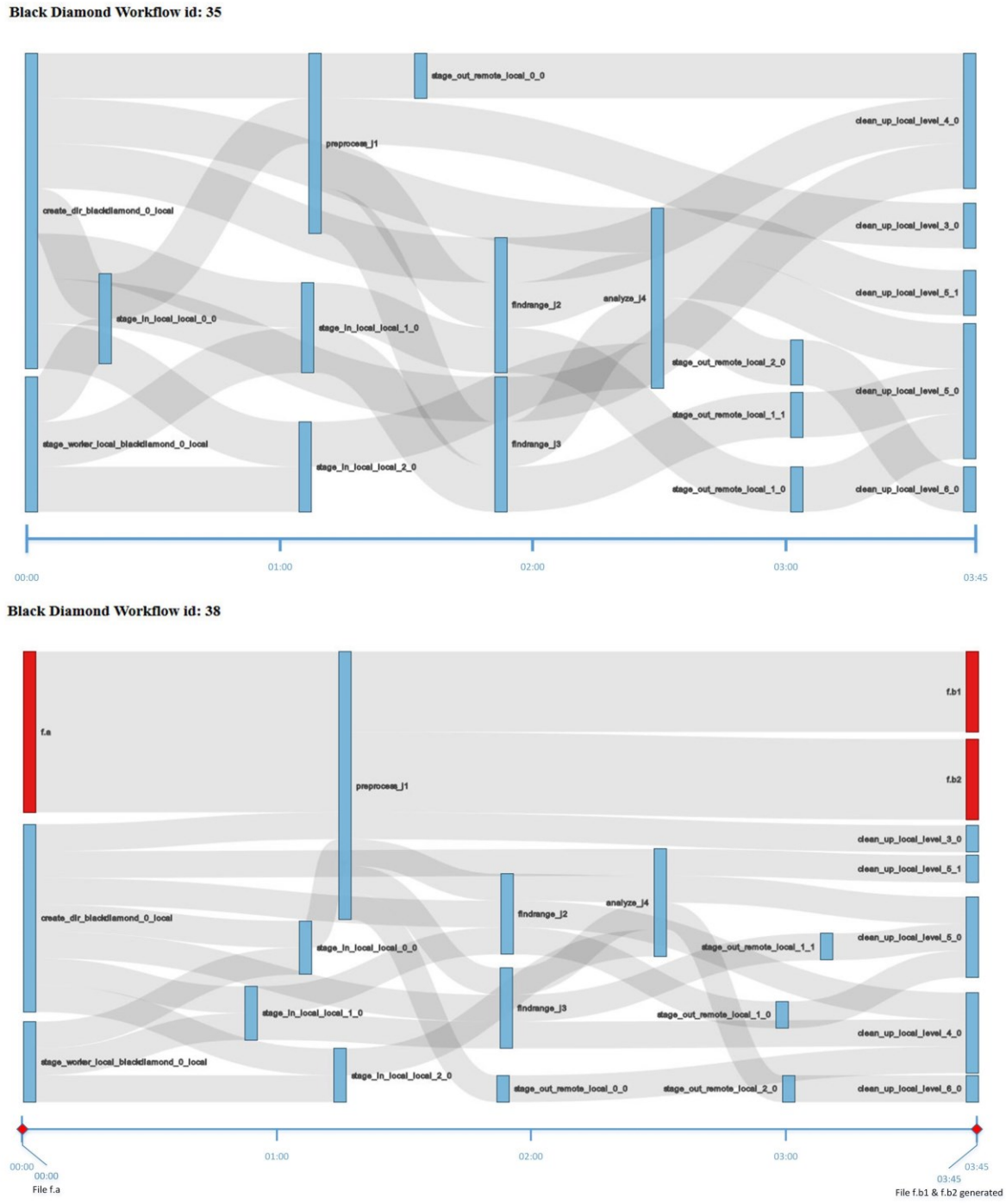


Figure 6.10 Workflow Comparison Workflow 35 & 38

Figure 6.10 allows basic users (along with intermediate & advanced users) to inspect the visualisation of provenance being compared for two different workflows i.e. workflow id 35 and workflow id 38. The entities highlighted in red in the diagram above are ‘f.a’, ‘f.b1’ and ‘f.b2’. Clicking on the ‘f.a’ entity allows users to determine that the entity has been consumed by several

activities and has eventually led to the generation of entities ‘f.b1’ and ‘f.b2’. NeuroProv allows users to highlight the trace for these entities in order to inspect the sources and results being generated over the course of the workflow. Comparison within the two workflows reveals that the earlier workflow (i.e. workflow 35) does not contain either of the entities (‘f.a’, ‘f.b1’ and ‘f.b2’). Metadata information present along with these workflows allows intermediate and advanced users to add essential information regarding the comparison performed which might be helpful for other potential users.

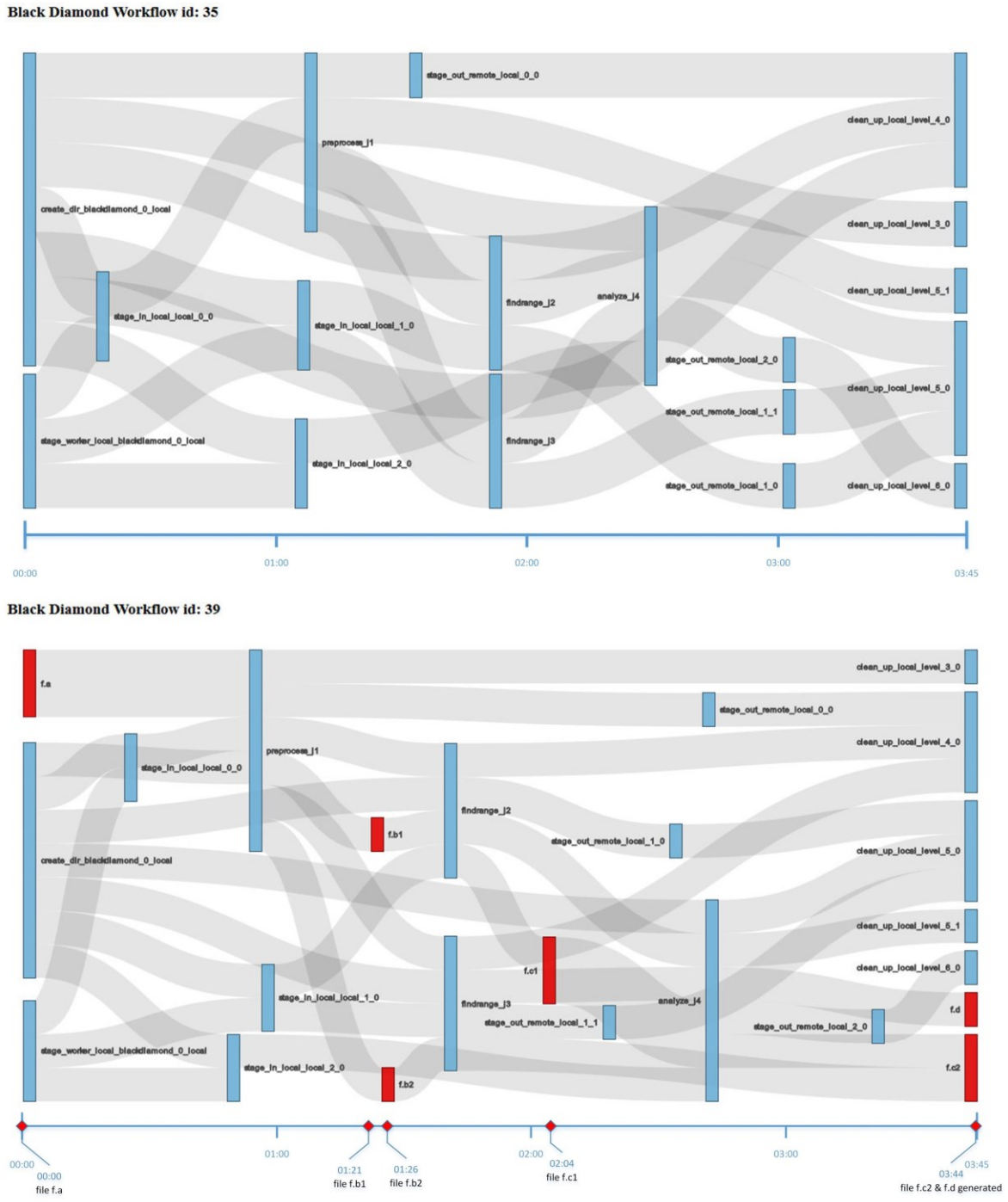


Figure 6.11 Workflow Comparison 35 & 39

Figure 6.11 shows a comparison of two workflows i.e. workflow 35 and workflow 39. Upon first inspection the user is able to figure out the differences highlighted in red. There are six different entities that are generated or consumed over the course of the workflow. These entities are ‘f.a’, ’f.b1’, ‘f.b2’, ‘f.c1’, ‘f.c2’ and ‘f.d’. Entity ‘f.a’ was provided as in input to the workflow

while entities ‘f.b1’, ‘f.b2’ and ‘f.c1’ are generated as intermediary results and entities ‘f.c2’ and ‘f.d’ are generated as output entities for workflow 39. Clicking on any of the entities allows users to inspect the activities involved and to see which entities have contributed towards the generation of which entities. Metadata information presented along with comparison allows researchers to inspect already provided information. Intermediate and advanced users can add further information based upon their detailed analysis of the comparison amongst the workflow. Furthermore the user might want to generate a comparative study between workflows 38 and 39. The following figure provides a comparison visualisation generated for workflows 38 and 39.

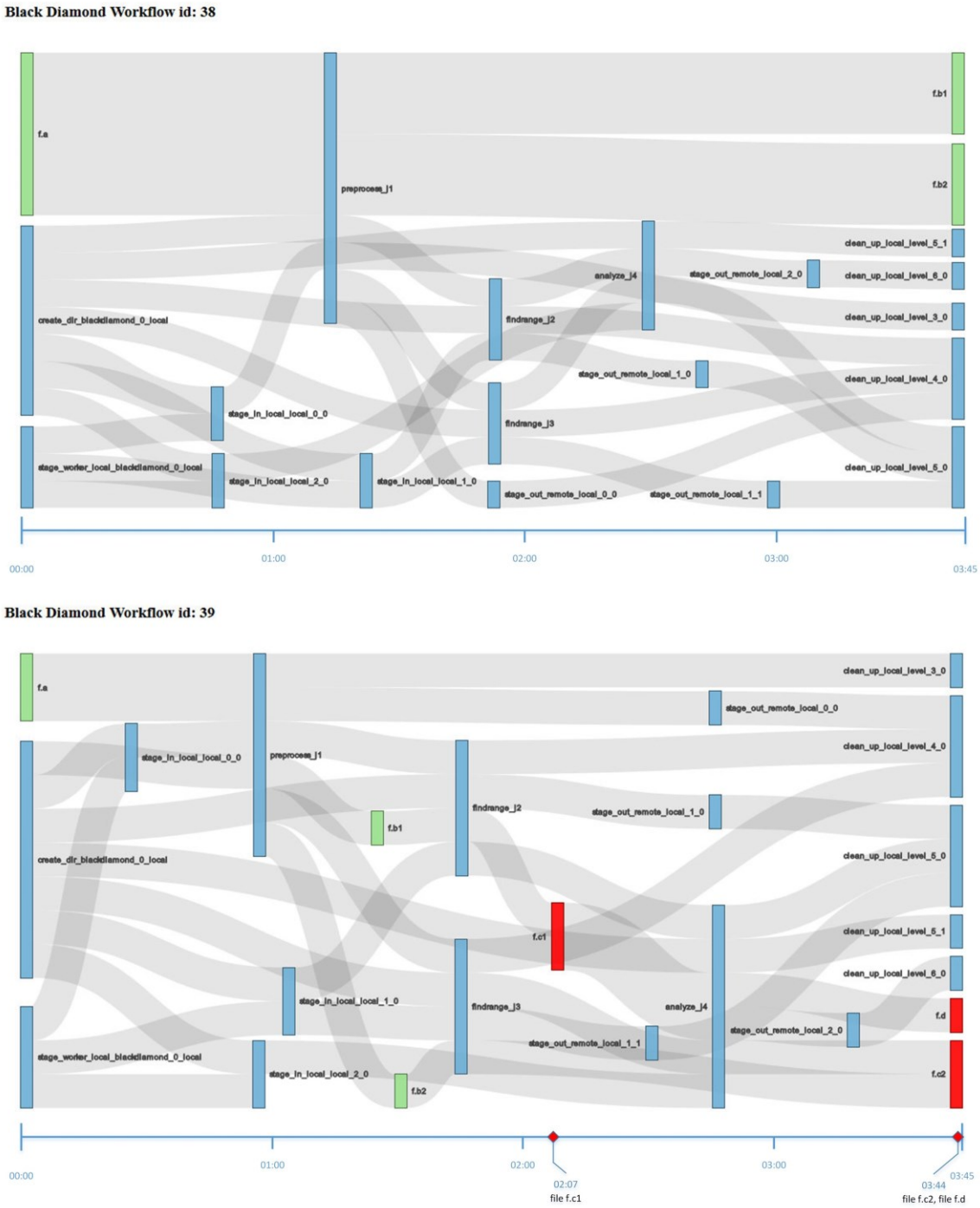


Figure 6.12 Workflow Comparison Workflow id 38 & 39

Figure 6.12 shows a further workflow comparison between workflows 38 and 39. The entities highlighted in red i.e. ‘f.c1’, ‘f.c2’ and ‘f.d’ are generated only within workflow 39 but not within workflow 38. The user, once clicking on the entity ‘f.c1’, can inspect how entities present in workflow 38 have contributed to the consumption of entities ‘f.c2’ and the generation of entity ‘f.d’.

Metadata information provided along with the workflow allows users to inspect when the particular entity was generated and consumed subsequently. Furthermore a comparative inspection of similar entities present in both the workflows can also be inspected to ensure how they have been generated, at what time and what activities generated them over the course of the workflow.

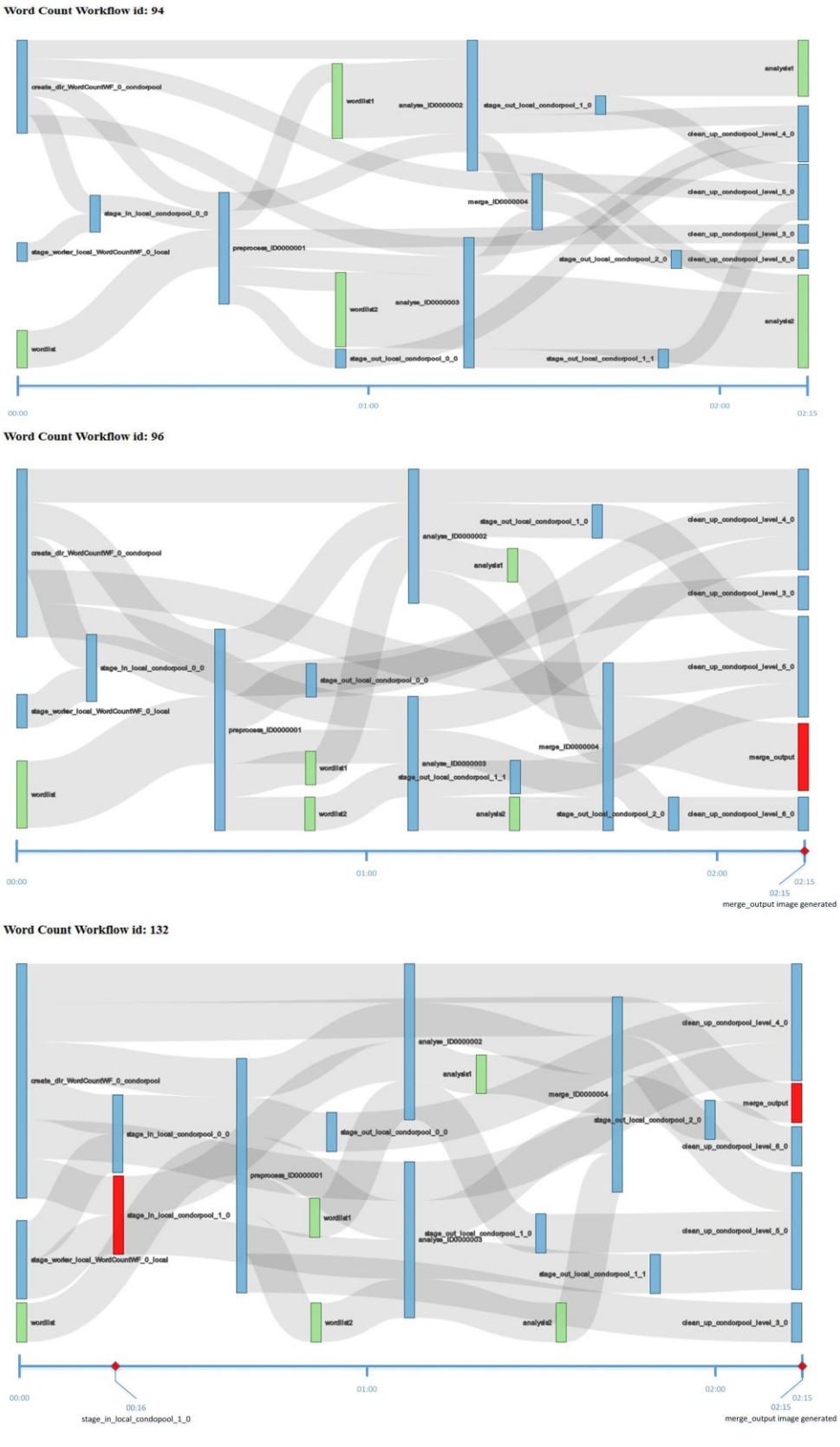


Figure 6.13 Multiple Workflow Comparison, Workflow Id's 94, 96 & 132

Figure 6.13 shows a comparison of workflows 96 and 132 against workflow 94. A complete comparison amongst the three workflows reveals major similarities and some minor differences. The metadata information represented along with the workflows present an improved understanding of these workflows by providing information regarding timeline and intermediary results. The multiple comparison workflow view of NeuroProv enables users to compare workflows 96 and 132 against workflow 94. The following set of results presents a visualisation comparison of workflow 96 against workflow 94 and workflow 132 against workflow 94 separately. This will provide users with separate views for comparing against a specific workflow. The combined comparison view provides users with the ease to view multiple visualisations in a single window with differences highlighted in red.

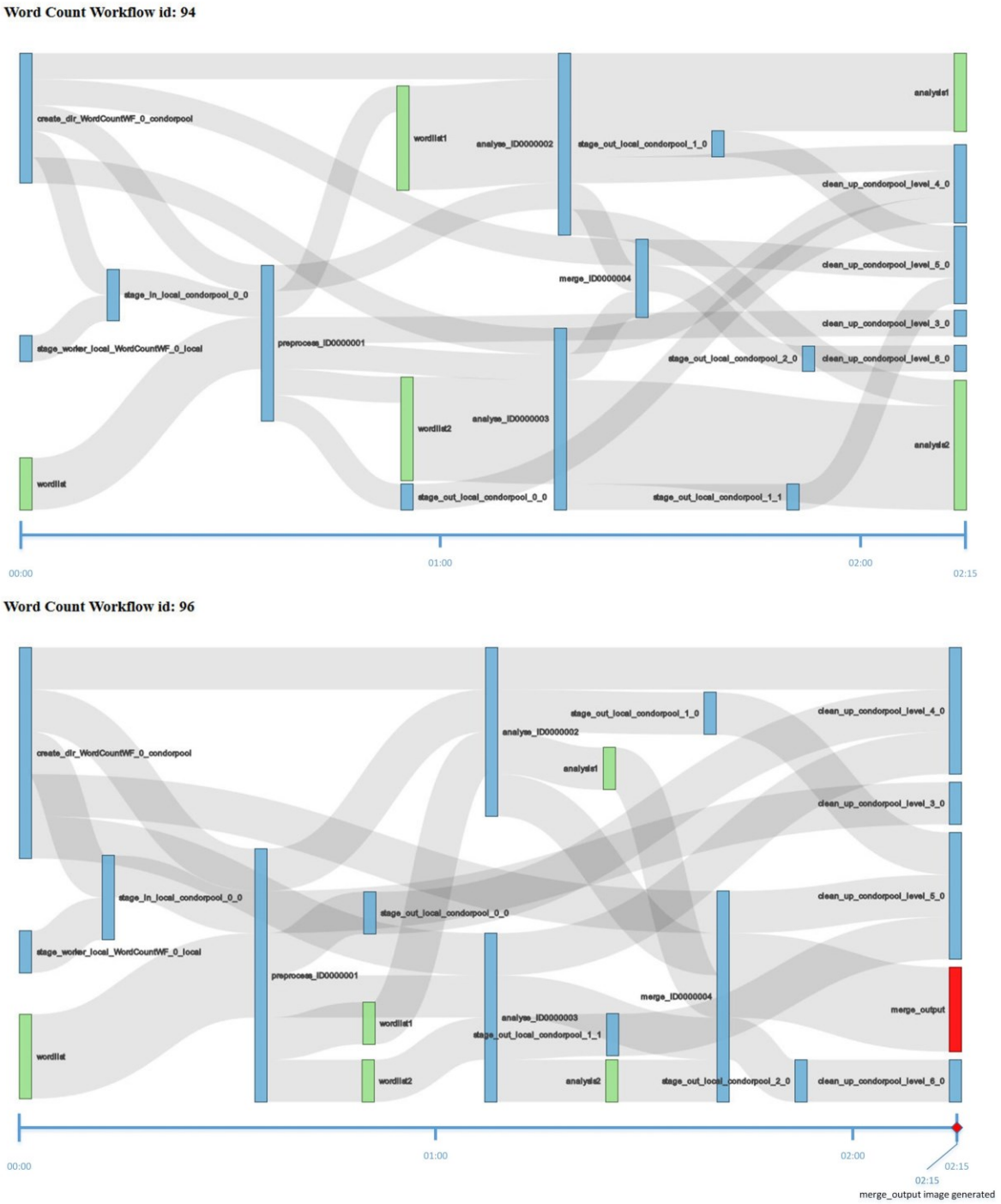


Figure 6.14 Workflow Comparison wf_id 94 & 96

Figure 6.14 provides a visualisation comparison between workflows 94 and 96. One of the prime differences between the two is the addition of a 'merge_output' entity as an output entity (as highlighted red in the figure above). The annotation provided with the workflows reveals the cause

of the ‘merge_output’ generated as an output entity. Clicking on the ‘merge_output’ entity allows users to highlight the sources of generating this entity along with other contributing activities.

Figure 6.15 Workflow Comparison wf_id 94 & 132

Figure 6.15 shows a visualisation comparison of workflow 132 against workflow 94. One of the prime differences revealed using the NeuroProv comparison window is the addition of activity ‘stage_in_local_condorpool_1_0’ and the resultant entity ‘merge_output’. The comparison window provides a brief account of the reason for the addition of both the activity and the entity. Intermediate and advanced users have the ability to annotate the comparison to provide further information that might be helpful for future potential users. Table 3 (below) provides an account of metrics taken under consideration for NeuroProv’s Comparison of Workflows Use-Case and the metrics that the system fulfils are ticked below.

Table 3 –Metrics for Workflow(s) Comparison

Sr No	Metrics	Workflow(s) Comparison
1	Display complete provenance for workflows	✓
2	Display workflows as nodes, edges and relationships	✓
3	Highlight trace for a particular entity/activity	✓
4	Allow users to compare two workflows	✓
5	Allows users to compare two or more workflows	✓
6	Ability for users to drill down	✓
7	Provide annotations with workflow, nodes and edges	✓
8	Examine entity/activity’ history	✓
9	Provide tracking of expanded stages	✓
10	Search feature e.g. workflow, dataset, nodes etc.	✓
11	View node attributes	✓
12	Workflow execution timeline	✓

6.2.3 Evolution – Use-Case 3

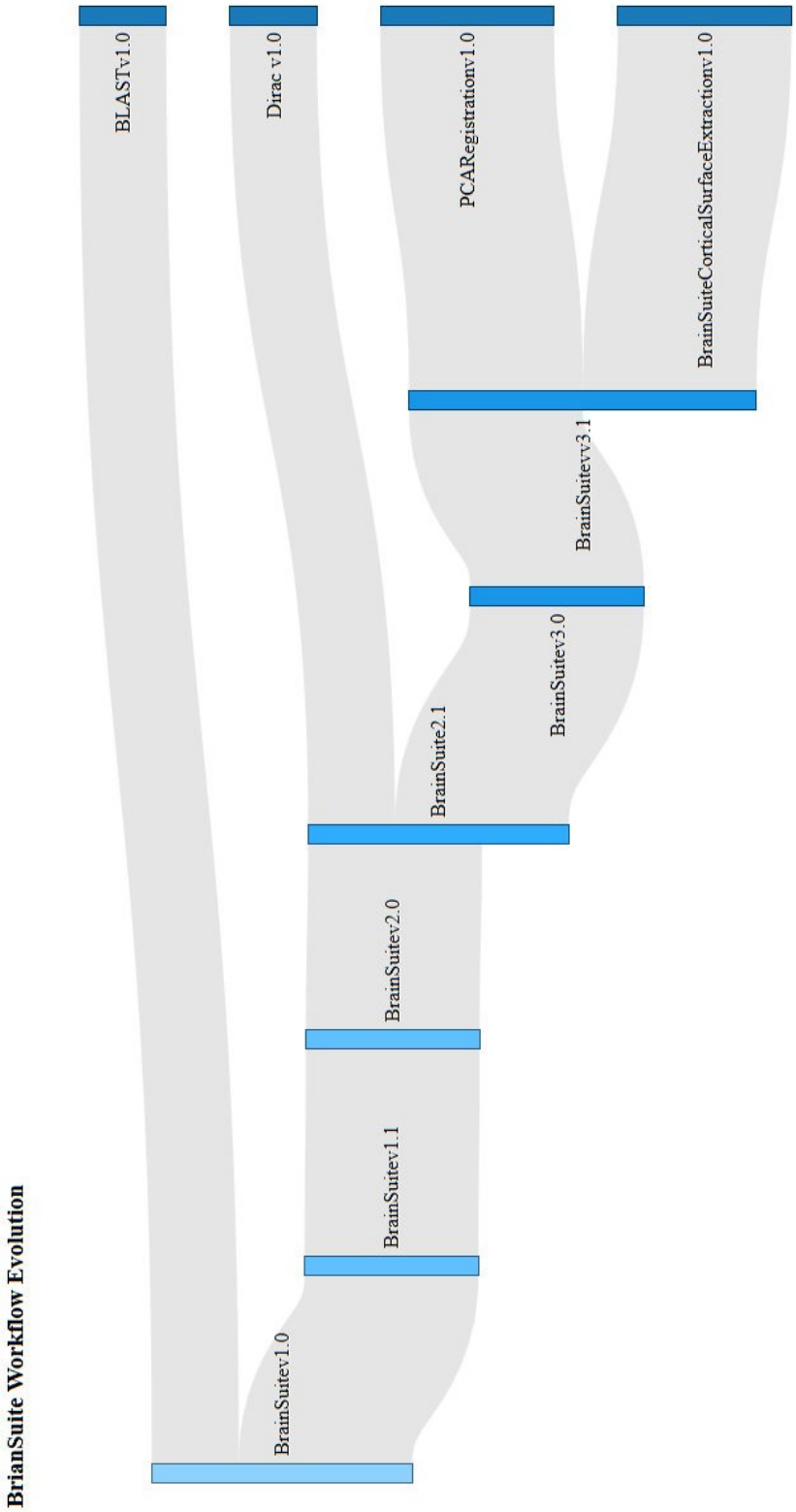


Figure 6.16 BrainSuite Workflow Evolution

Figure 6.16 shows the evolution of workflow ‘BrainSuitev1.0’ over the period of time with the addition/deletion of various activities and entities. The workflow on the left is the one that evolves over the period of time while the workflows on the right are the ones that have been transformed into different workflows. We started with ‘BrainSuitev1.0’ workflow that generated into four different workflows namely ‘BLASTv1.0’, ‘Diracv1.0’, ‘PCARegistrationv1.0’ and ‘BrainSuiteCorticalSurfaceExtrationv1.0’. Clicking on any of the workflow or its version allows users to inspect how a workflow came into existence by highlighting its entire trace to examine its sources and targets. By double clicking the workflow or its version under inspection NeuroProv provides user with a complete verification view thus allowing users to individually inspect each workflow.

When the user clicks a link between a workflow (see Figure 6.17) NeuroProv provides users with useful metadata information required for further experimentation. This is displayed alongside the Sankey Diagram in the window. This enhances the user experience by allowing users to view detailed information regarding the changes over subsequent versions of a workflow or when new workflows are generated using previous versions of workflows. For instance in Figure 6.17 a basic user may want to find out the changes that caused ‘BrainSuitev1.0’ to evolve into ‘BLASTv1.0’ workflow. Upon clicking on the link between the two workflows NeuroProv provides annotations associated with the cause(s) for transforming the former workflow into the latter workflow. Annotations provided by NeuroProv reveal that the addition of ‘CopyUV’ module led to the generation of ‘BLASTv1.0’ from ‘BrainSuitev1.0’. Also note that upon clicking on the link it is highlighted to bring the user’s focus to the workflows under inspection.

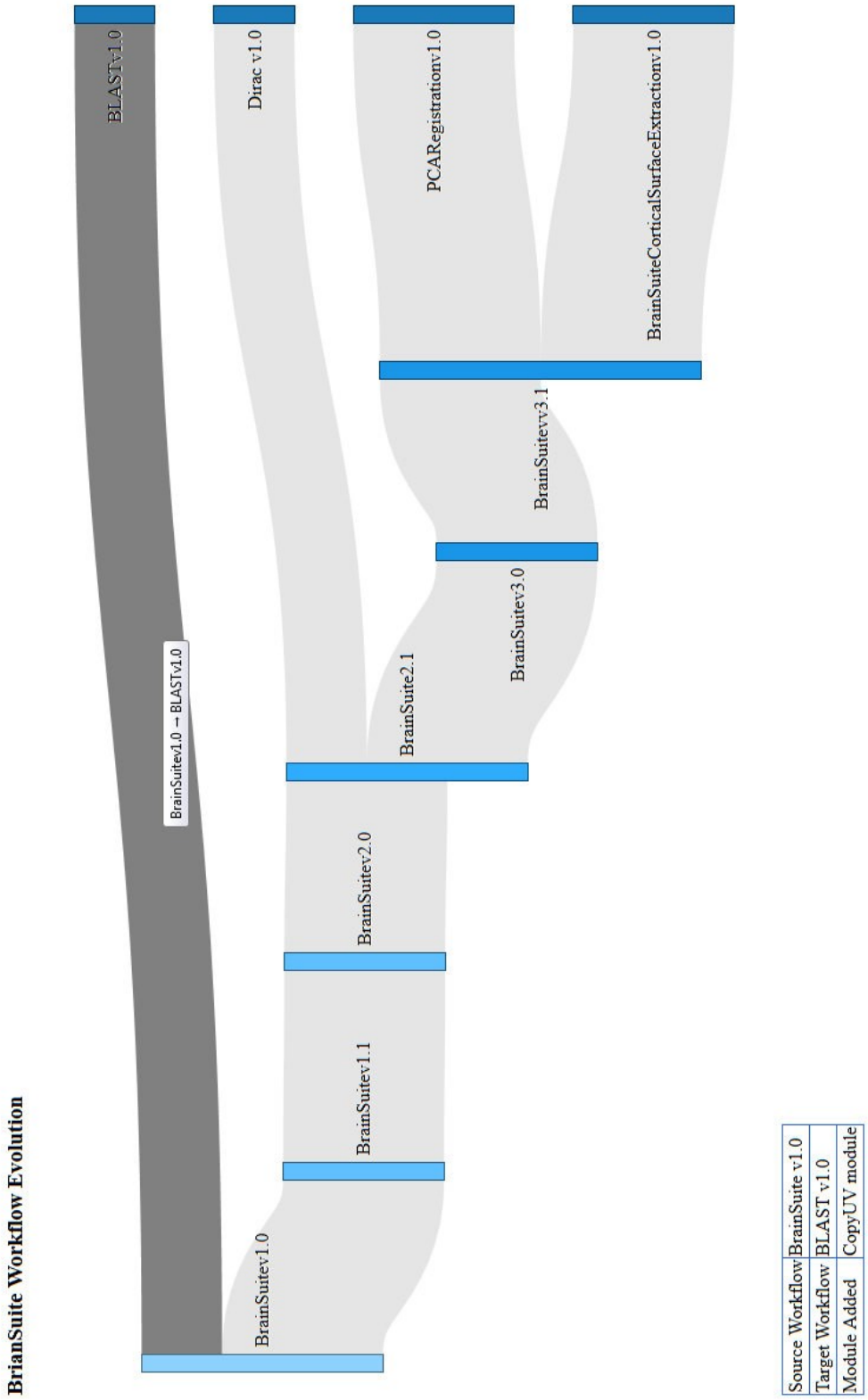


Figure 6.17 BrainSuite Workflow Evolution Link Clicked

Taking another example from the same workflow's evolution figure 6.18, the user may want to bring 'PCARegistrationv1.0' under inspection. The link between the source i.e. 'BrainSuitev3.1' and the target i.e. 'PCARegistrationv1.0' is highlighted, and once the user clicks the link between the two workflows any annotations provided with how the workflows have evolved are presented on the screen. In this scenario the input image to workflow 'BrainSuitev3.0' is 'T1 Image' while the output file generated is 'SVPASEG'. This leads to the transformation of 'BrainSuitev3.0' to 'PCARegistrationv1.0'. The user can click either of the workflow nodes to inspect an individual workflow's provenance and further study the workflows in detail.

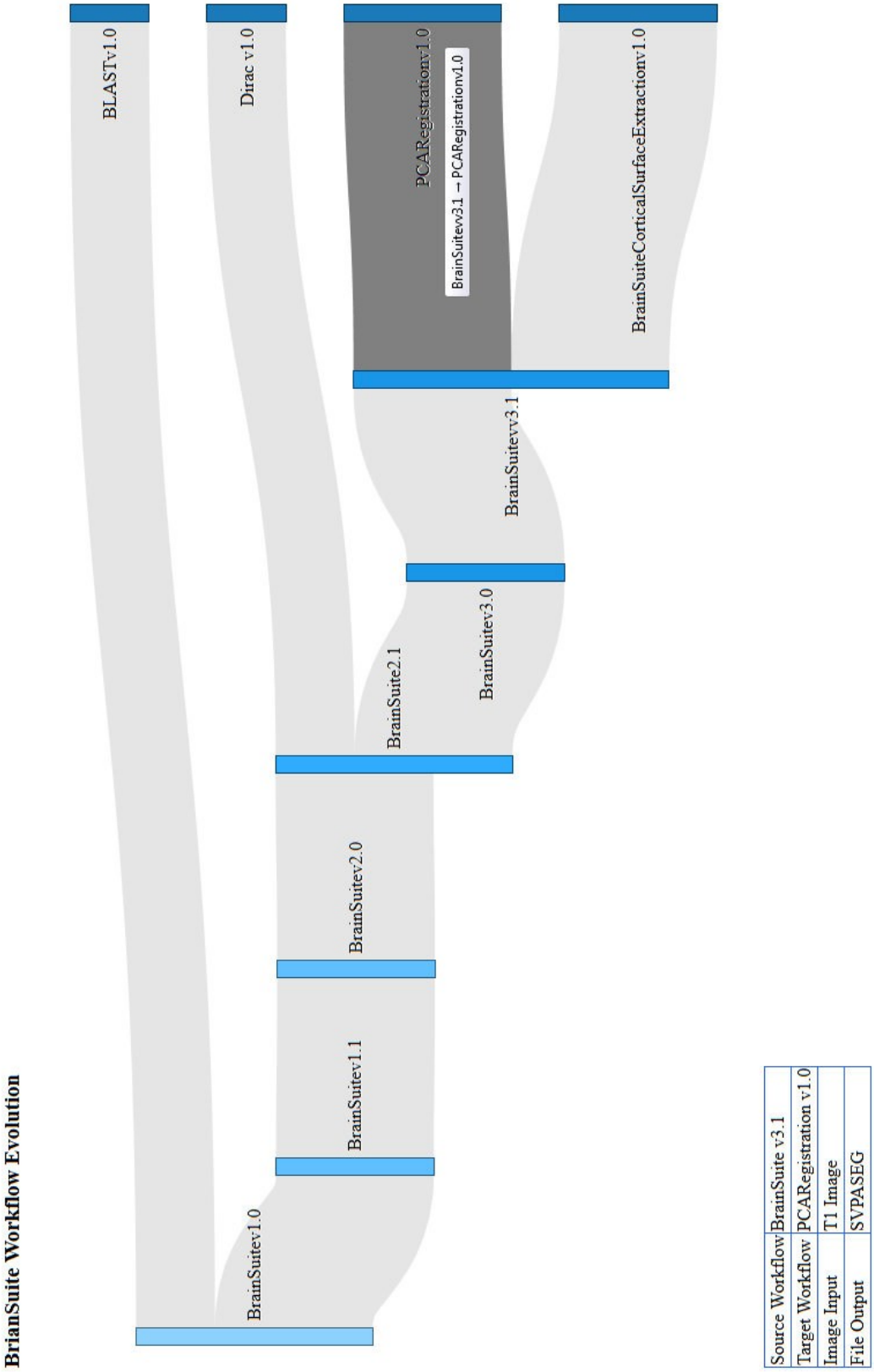


Figure 6.18 BrainSuite Workflow link highlighted

Table 4 highlights the metrics that hold true for NeuroProv's Workflow(s) Evolution Use-Case. NeuroProv's ability to track the evolution of workflows allows researchers to inspect how the workflow has evolved over the passage of time, what caused the evolution, by whom and for what purposes. This saves time to verify which of the previous versions of a workflow are correct or not, thus allowing users to save useful time to perform experimentation on the current workflow versions and to conduct research in an organised manner.

Table 4 –Metrics for Workflow(s) Evolution

Sr No	Metrics	Workflow(s) Evolution
1	Display complete provenance for workflows evolution	✓
2	Display workflows versions as nodes and changes as edges	✓
3	Highlight trace for a particular workflow version	✓
4	Ability for users to drill down	✓
5	Provide annotations with workflow versions	✓
6	Provide users with the ability to see how WF's evolve over time	✓
7	Provide tracking of expanded stages	✓
8	Search feature e.g. workflow versions	✓
9	View workflow attributes	✓
10	When a user clicks a particular workflow version display in a new window	✓

6.2.4 Progression – Use-Case 4

Figure 6.19 WordCount Workflow Progression

Figure 6.19 displays visualisation of provenance data for the ‘WordCount’ workflow and its subsequent versions. NeuroProv allows researchers to visually view how a workflow has progressed into subsequent versions over the course of time. NeuroProv also provides users with metadata information and annotations that will help users to visually monitor the progression of workflows and later if required to use a particular version to perform future experimentation. The user can click on any particular version of a workflow that will allow the workflow to be opened in another window and the user can then perform further analysis. The user can view workflows in any view required and perform future experimentation.

Figure 6.20 WordCount Workflow Progression link clicked

Figure 6.20 visually represents the progression of ‘WordCountWF v1.0’. The user can click the link between the nodes ‘WordCountWF v1.0’ and ‘WordCountWF v2.0’ to see what caused the progression of the workflow version on the left to the workflow version on the right. The Sankey Diagram highlights the trace to bring the focus of the user to the workflow versions under inspection. Upon clicking the node the user is provided with the annotation stating that the change in the version was due to addition of an activity to ‘WordCount Workflow v1.0’ i.e. ‘stage_in_condor_pool_1_0’. This helps the user to identify the cause of the change and to verify if the progression is accurate as per the user’s understanding of the workflow. The user can click on either of the workflow version to inspect complete provenance for that version of the workflow thus facilitating future experimentation on that version to verify results or to make modifications to use for some other purpose of the user’s requirements.

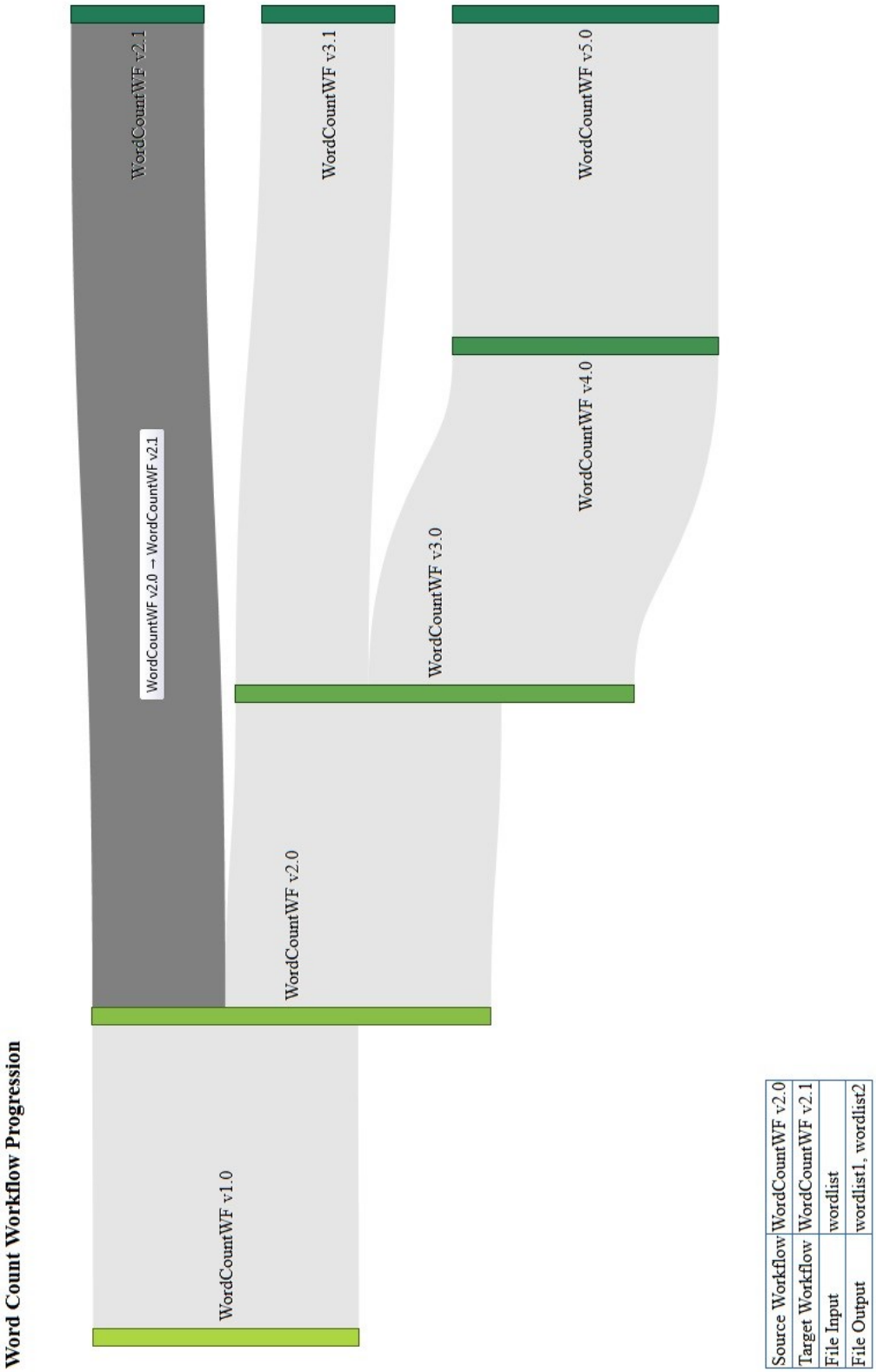


Figure 6.21 Annotations provided when link clicked between WordCountWFv2.0 & WordCountWF v2.1

Figure 6.21 shows the progression of ‘WordCountWF v2.0’ to ‘WordCountWF v2.1’. The user can hover over the link to highlight the workflow versions under inspection. Upon clicking the node the user is provided with annotations related to the progression of the workflow from the former version to the latter. It can be deduced from the annotation that the input and output of certain files led to the progression of ‘WordCountWf v2.0’ as shown in the diagram.

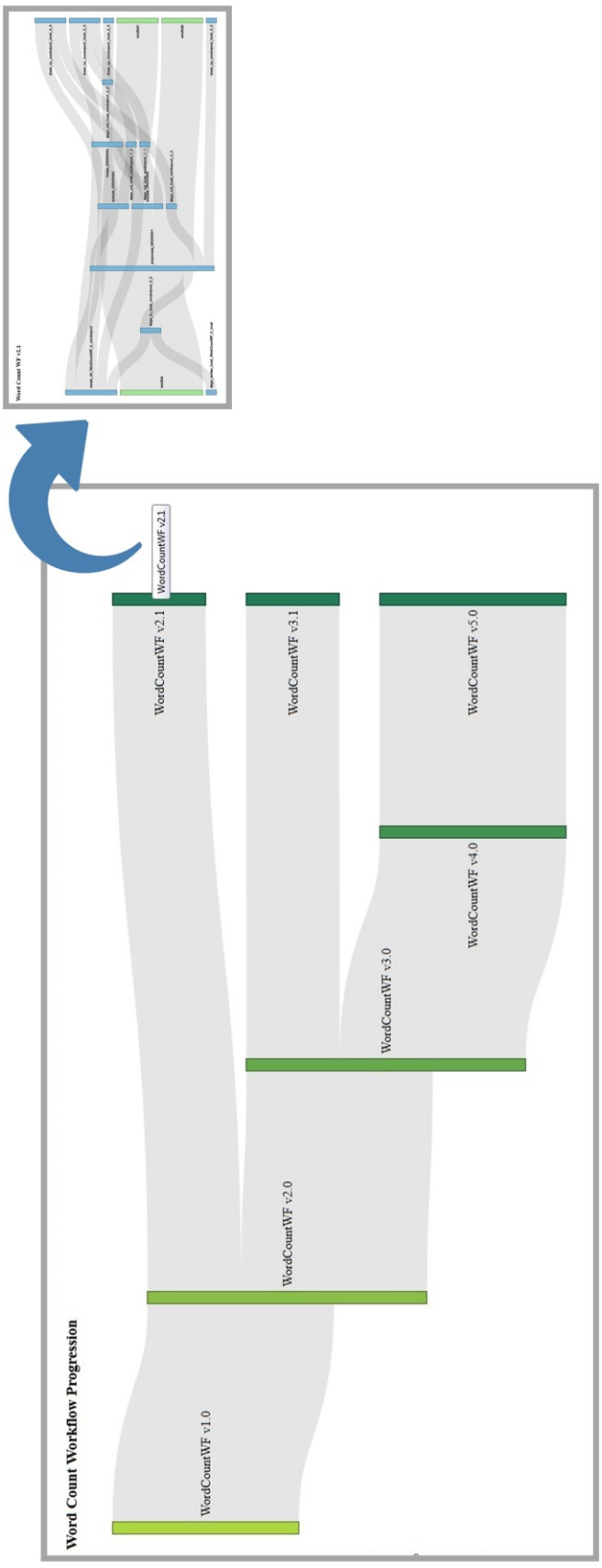


Figure 6.22 Workflow version clicked - opens in a new window

Figure 6.22 shows that the user can inspect the individual provenance for 'WordCountWF v2.1'. NeuroProv provides the capability of inspecting provenance in a separate window so that the current window does not get cluttered making it difficult to identify the essential features of the workflow required for inspection by the user. Opening the workflow in a separate window provides all the basic capabilities of NeuroProv such as annotations provided with that workflow version, details such as to who executed the workflow, when and for what purposes. All this provenance data is essential to verify the authenticity of the results and the workflow itself. Table 5 (below) provides a brief analysis of all the metrics that hold true for Progression of a Workflow in NeuroProv Use-Case.

Table 5 –Metrics for a Workflow's Progression

Sr No	Metrics	Workflow Progression
1	Display complete provenance for a workflow's progression	✓
2	Display a workflow's versions as nodes and changes as edges	✓
3	Highlight trace for a particular workflow version	✓
4	Ability for users to drill down	✓
5	Provide annotations with workflow's versions	✓
6	Provide users with the ability to see how a workflow changes over time	✓
7	Provide tracking of expanded stages	✓
8	View workflow's attributes	✓
9	When a user clicks a particular workflow version display in a new window	✓

6.3 Conclusion

We have provided results and their analysis in this chapter based on the experimentation performed for the evaluation of our hypothesis 'Visualisation techniques can enhance the utility of provenance data for neuroimaging analysis'. The results allowed us to examine the workings on NeuroProv based on the use-cases defined in Chapter 3. According to the results of the experimentation, visualisation techniques do enhance the utility of provenance data for neuroimaging analysis. The results clearly suggest that a system such as NeuroProv encompasses aspects required for visualisation of provenance data that were either completely or partially missing from such visualisation systems. The metrics defined in chapter 4 have been addressed separately for each use-case to evaluate the efficacy of NeuroProv. In the following chapter our expected contribution towards the neuroimaging community is discussed to give users a better perspective of how our research work fits into the wider neuroimaging community.

Chapter 7

Conclusions, Contribution to Knowledge and Future Directions

7.1 Conclusions

Tremendous development has been observed in the domain of E-Science platforms; one of the major barriers to their widespread use is the lack of provenance support and usage. Provenance means the history, ownership and its processing in some domain of interest. For neuroscientific systems such as NeuGRID and N4U provenance helps clinical researchers in the study of MRI scans to determine biomarkers for the onset of Alzheimer's disease. In such a collaborative environment it is essential to keep track of who did what, when and for what purpose. In order to understand the lineage of both the data products and processes, researchers require visualising provenance data to extract meaningful information. Current visualisation systems partially or completely lack support for provenance visualisation thus our research endeavour undertook to develop a system that could address the needs of visualising provenance for neuroimaging analyses. The research was conducted in the context of the NeuGRID and N4U projects, which aimed to provide traceability to support research analysis processes in the study of biomarkers for Alzheimer's disease, but are generically applicable across medical systems.

In our research work we identified general requirements for provenance visualisation particularly for neuroimaging analysis (in Chapter 2) and further detailed functional requirements based on our proposed system, NeuroProv's use-cases (in Chapter 3). This provided us with a set of requirements that were categorised as essential, desirable and optional. These identified requirements showed that we have been successfully able to satisfy our first research question i.e. 'What are the functional requirements for provenance visualisation in neuroimaging analysis?'. Eliciting requirements gave a fair idea of which needs are to be met essentially in order to demonstrate the utility of provenance data visualisation for neuroimaging analysis. Other identified requirements would have been taken into account had time permitted.

Based on the requirements identified for research question number 1 the next research question was about which technique would be suitable for visualising provenance data for neuroimaging analysis. We have fulfilled this requirement by using a visualisation technique known

as ‘Sankey Diagrams’ to generate visualisation. This technique has been carefully selected based on the previous research question regarding requirements.

The third research question i.e. ‘Which metrics can be used to evaluate the utility of provenance data for neuroimaging analysis?’ was answered through visualisation metrics elicited in Chapter 4 that helped in fulfilling requirements for the provenance visualisation identified in research question number 1 and the use-cases investigated in chapter 3. The research concluded that visualisation techniques do indeed enhance the utility of provenance data for neuroimaging analysis thus proving our research hypothesis holds true. The results presented in chapter 6 clearly show that visualisation techniques can increase the utility of provenance data for neuroimaging analysis. NeuroProv consequently enhances the user’s experience by providing tools and techniques to visualise provenance data for neuroimaging analysis.

7.2 Contribution to Knowledge

NeuroProv allows neuroimaging researchers to visually represent provenance data in an intuitive manner thus allowing clinicians and users to exploit the true potential of provenance data. Clinical researchers can benefit from the use of NeuroProv to understand provenance data for neuroimaging analysis, can verify a result or intermediate result, can compare two or more workflow visualisations and can visually investigate how a workflow has evolved over the period of time and see the progression of workflows for future use. The use of the Sankey Diagram 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 the neuroimaging community with a system to visualise provenance data for neuroimaging based on the following use-cases: the verification of results; the validation of workflows; the comparison of workflows/datasets; the evolution of workflows/datasets and the progression of workflows/datasets. This provides the user community and, in particular, the research clinicians a means to authenticate experimental results and to reproduce an experiment essential for the domain of neuroimaging analysis. Other visualisation systems as discussed in detail in Chapter 2 either completely or partially lack the ability to visualise provenance for neuroimaging analysis. Provenance data can be large, sometimes as much as an order of magnitude greater than the data for which the provenance is recorded. This is also true for the domain of neuroimaging analysis where researchers tend to record provenance data potentially of greater magnitude than the data recorded for initial experimentation.

NeuroProv provides researchers 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. 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 User' may wish to inspect detailed visualisation of provenance to determine errors/anomalies.

Over the course of the study we have managed to identify requirements for provenance visualisation (as elicited in Chapter 3), derived from the use-cases designed for neuroimaging analysis. The use-cases encompass all aspects of provenance visualisation for the domain of neuroimaging analysis. Particular focus has been paid on the usage of provenance data in order to determine the requirements. The detailed study of state-of-the-art reveals that most of the current visualisation systems 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. The first research question is being taken into account and requirements for provenance visualisation for neuroimaging analysis have been drafted. Thus 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. Our other contribution is the results of our qualitative study, devising metrics to evaluate visualisation for provenance data.

The following table provides a summary of how NeuroProv compares to existing visualisation systems based on identified by inspecting the requirements for provenance visualisation in the previous chapter.

Table 6 - Comparison of Visualisation Systems vs NeuroProv

Metrics	Visualisation Systems										
	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 (more than two)			•								•
Display workflows as nodes, edges and relationships		•	•	•		•	•	•	•	•	•
Highlight trace for a particular data or process											•
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 etc.			•	•							•
Visually view attributes (visually encoded)	•	•	•		•		•	•		•	•
Workflow execution timeline									•		•

Based on Table 6 (above) NeuroProv addresses all the metrics when compared to other visualisation systems which partially address the metrics elicited for visualisation of provenance data for neuroimaging. This provides a basis for stating that NeuroProv's capability to effectively and efficiently visualise provenance data for neuroimaging analysis holds true and that it can be further developed to encompass provenance for other domains in future.

NeuroProv provides different classes of users in N4U namely Basic, Intermediate and Advanced clinical researchers' the means to verify and authenticate their results. NeuroProv helps N4U users to generate visualisations of provenance data to validate results, compare two or more visualisations to identify anomalies, visually to see how workflow(s) evolve over time and to inspect the progression of a workflow. NeuroProv's workings rely on the querying the NeuroProv Store (shown in Figure 5.2, Chapter 5) and provides NeuroProv with results to generate visualisation. As a consequence of our research work, NeuroProv improves the workflow definition by ensuring that the appropriate pipelines, algorithms and images have been selected for execution. In the verification module, researchers can verify if any anomaly occurred due to the changes in workflow specifications. Within the context of N4U, NeuroProv is a means to generate

visualisations depending on user's requirements. NeuroProv minimises an individual researcher's burden for providing details such as provenance and metadata to dramatically improve compliance. This frees the user to focus on performing neuroimaging research rather than exhaustively documenting provenance.

One of the limitations of the current system is that it can only support expansion down to two levels of detail. Ideally the user would be able to incrementally drill down to multiple levels of details. For example, one link can be expanded to two links, each of which can be further expanded. Drilling down to multiple levels of details may prove quite complex to implement because it involves multiple levels of inferencing rules. Another limitation of the system is that it does not provide users with the ability to export provenance visualisations, to be used in other systems. Furthermore NeuroProv does not have the ability to live update in view of the currently visualised workflow provenance. Additionally the users cannot save visualisations as bookmarks so they can be revisited later in a session. Unlike [41] NeuroProv lacks the ability to generate visualisation focusing on a particular activity, this might be helpful in a scenario where the user only wants to focus on a particular activity without the context of the entire workflow.

7.3 Future Direction

One major direction for the continuation of this work is generalising the system to perform in other domains. NeuroProv currently facilitates the visualisation of provenance data for neuroimaging analysis. One future direction in this regard would be that the results from this research could be validated externally to assess if 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 can be modified in future in order to accommodate other provenance systems in order to visualise data. Many of the systems in other domains have the ability to keep track and store provenance data to some extent but systems do not holistically capture provenance for (re-)analysis. Systems in Geo-Spatial domains tend to capture provenance of data but not of processes which is essential to reproduce and verify an experiment. Recently the medical domain has been at the forefront of adopting standards that ensure that provenance is completely captured for both data and processes. Thus it would seem plausible that other domains will follow a similar path in the not too distant future.

Another future direction in an aim to extend the existing NeuroProv system will be to add further functionality that will aid the users in the research process. Adding a statistical bent to NeuroProv is a logical direction. For example for pipelines and datasets the users might want to know the number of attempts to access, or the number of times a system has crashed. Other

measures include the ability to display usage metrics to identify which datasets or pipelines are commonly used. This provides users with an advanced idea of what other researchers commonly used to conduct similar kind of workflows. Another aspect could be the use of heat maps or clusters to represent commonly used images for analysis.

In the future we intend to research and integrate within the NeuroProv a module that will enable applications to learn from their past executions and improve and optimise new studies and processes based on the previous experiences. Models that will inform researchers of missing processing stages, suggest available and verified processing modules and warn users of incompatible data types. Another possible addition to the system can 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 analysis with limited modification. Models can be formulated that can derive the best possible optimisation strategies by learning from the past execution of experiments and processes. These models will gradually evolve over time and will facilitate decision support in generating visualisation of future processes and workflows in a domain.

One essential future direction in this regard is provenance interoperability. Currently NeuroProv uses the emerging PROV [19] interoperability standard. This will allow N4U users to share their provenance data in other PROV-compliant systems. Currently NeuroProv uses the PROV-XML standard for generating visualisations. In order to give the readers a better overview of what provenance data for neuroimaging looks like when represented in a PROV-XML format refer to Appendix B. Workflow id 132 is represented in PROV-XML format for the readers in Appendix B. Another future direction would be to display only relevant provenance related to the role of the user in order to reduce data overloading on the screen. This will enable basic researchers to view a high level summary of the visualisation while a more experienced user with further details for inspection relevant to the user.

Our primary future effort will be to develop useful applications based on the work presented in this thesis, gain access to real world infrastructure for system deployment and engage users in using provenance to perform neuroimaging research, encouraging subsequent data and provenance sharing, enhanced peer-reviewed publications and support multi-centre collaboration.

References

[1]. Friedman, V. (2008) *Data Visualisation and Infographics* in Graphics, Monday Inspiration.

[2] CMS Collaboration. (2005) *The Computing Project Technical Design Report* in CERN/LHCC-2055-023.

[3] Foster, I., Zhao, Y., Raicu, I. and Lu, S. (2008) *Cloud Computing and Grid Computing 360-Degree Compared* in IEEE Grid Computing Environments (GCE08), co-located with IEEE-ACM Supercomputing.

[4] Mehmood, Y., Habib, I., Bloodsworth, P., Anjum, A., Lansdale, T., McClatchey, R., and The neuGRID Consortium. (2009, pp.1-4) *A Middleware Agnostic Infrastructure for Neuro-imaging Analysis* in 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS).

[5] Redolfi, A., McClatchey, R. et al., (2009, pp. 703-722) *Grid Infrastructures for Computational Neuroscience: the neuGRID Example* in Future Neurology vol4, no. 6, DOI 10.2217/fnl.09.53

[6] Deelman, E., Gannon, D., Shields, M. and Taylor, I. (2008) *Workflows and eScience: An Overview of Workflow System Features and Capabilities* in Future Generation Computer Systems.

[7] Simmhan, Y, L., Plale, B., Gannon, D. (2005, pp.31-36) *A survey of Data Provenance in e-science* in Sigmod Record.

[8] Simmhan, Y, L., Plale, B. and Gannon, D. (2006, pp.72) *Towards a quality model for effective selection of data in collaboratories* in Workshop on Workflow and Data Flow for Scientific Applications (SciFlow06), held in conjunction with ICDE, IEEE.

[9] https://neuGRID4you.eu/en_GB [Accessed: April 2014]

[10] Experimental Research: http://www.experiment-resources.com/experimental-research.html [Accessed: April 2014]

[11] *Workflow on the Web*, the Applied Technologies Group, www.globotron.com/htm/whitepapers/workweb.pdf, [Accessed: May 2014]

[12] Workflow definition: www.e-workflow.org, [Accessed: May 2014]

[13] Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Linvy, M., Moreau, L. and Myres, J. (2007, pp.24-32) *Examining the Challenges of Scientific Workflows* in Computer, vol 40.

[14] Miles, S., Deelman, E., Groth, P., Vahi, K., Mehta, G. and Moreau, L. (2007, pp.179-186) *Connecting Scientific Data to Scientific Experiments with Provenance* in IEEE International Conference on e-Science and Grid Computing.

[15] MacKenzie-Graham, A., VanHor, J., Woods, R., Crawford, K. and Toga, A. (2008, pp.178-195) *Provenance in Neuroimaging* in NeuroImage Vol 42. 178–19510.1016/j.neuroimage.

[16] Horn, J, D., Grafton, S., Rockmore, D. and Gazzaniga, M. (2004, pp.473-481) *Sharing Neuroimaging Studies of Human Cognition* in Nature Neuroscience 7.

[17] Lliinsky, N. and Steele, J. (2011) *Designing Data Visualisations; Intentional Communication from Data to Display*. : O'Reilly Media.

[18] Kunde, M., Bergmeyer, H. and Schreiber, A. (2008, pp.241-252) *Requirements for a Provenance Visualization Component* in Provenance and Annotation of Data and Processes.

[19] Groth, P. and Moreau, L. (2013) *An Overview of the PROV Family of Documents*. Details available for the W3C Consortium at: www.w3.org/TR/2013/NOTE-prov-overview-20130430/ [Accessed: January 2015]

[20] Bose, R. and Frew, J. (2004) *Composing Lineage Metadata with XML for Custom Satellite-Derived Data Products* in Scientific and Statistical Database Management, Proceedings in the 16th International Conference.

[21] Stevens, R., Robinson, A. and Goble, C. (2003, pp.302-304) *MyGrid: Personalised Bioinformatics on the Information Grid* in Bioinformatics, vol 19.

[22] Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D. and Greenwood, M. (2004, pp92-106) *Using Semantic Web Technologies for Representing e-Science Provenance* in The Semantic Web (ISWC).

[23] Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. and Goble, C. (2010) *Taverna reloaded* in SSDBM, Heidelberg, Germany.

[24] Quan, D., Huynh, D. and Kargar, D. (2003, pp.738-753) *Haystack: A Platform for Authoring End User Semantic Web Applications* in International Semantic Web Conference (ISWC).

[25] Haystack BioDASH Demonstration (2005) http://www.w3.org/2005/04/swls/BioDash/Demo/Haystack%20Demo%20home.html [Accessed: March 2014]

[26] Freire, J., Silvia, C., Callahan, S., Santos, E., Scheidegger, E., and Vo, H. (2006) *Managing Rapidly Evolving Scientific Workflows* in International provenance and Annotation Workshop (IPAW).

[27] Rio, N. and Silva, P. (2007, pp.732-741) *Probe-It! Visualisation Support for Provenance* in Advances in Visual Computing (ISVC), vol 4842 of LNCS, Springer.

[28] B. Ludaščher and et al. (2005) *Scientific Workflow Management and the Kepler System* in Concurrency and Computation: Practice & Experience, 2005. Special Issue on Scientific Workflows.

[29] Silva, P., mcGuinness, D. and Fikes, R. (2006, pp.381-395) *A Proof Markup Language for Semantic Web Services* in Information Systems 31(4-5): ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-04-01.pdf [Accessed: February 2014]

[30] Myers, J., Pancerella, C., Lansing, C., Schuchardt, K. and Didier, B. (2003) *Multi-Scale Science: Supporting Emerging Practice with Semantically Derived Provenance* in ISWC 2003 Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data.

[31] Cheung, K. and Hunter, J. (2006) *Provenance Explorer – Customized Provenance Views Using Semantic Inferencing* in 5th International Semantic Web Conference (ISWC 2006), vol 4279 of Lecture Notes in Computer Science. Springer-Verlag, (doi: http://dx.doi.org/10.1007/11926078_16 [Accessed: January 2014]).

[32] Frew, J. and Bose, R. (2001, pp.180-189) *Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products* in 13th International Conference on Scientific and Statistical Database Management, Fairfax, VA.

[33] AT&T. *Graphviz* (2006) AT&T Labs – Research. http://www.graphviz.org [Accessed: March 2014]

[34] Simmhan, Y., Plale, B. and Gannon, D. (2006) *A Framework for Collecting Provenance in Data-Centric Scientific Workflows* in International Conference on Web Services, Chicago, Illinois.

[35] Rio, N. and Silva, P. (2006) *Towards Scientific Provenance Visualisation on the Web* Technical Report, Department of Computer Science, University of Texas at El Paso, El Paso, TX.

[36] Federal Geographic Data Committee, Content standard for digital geospatial metadata: Content Standard for Digital Geospatial Metadata (CSDGM), FGDC-STD-001-1998, Federal

Geographic Data Committee, Washington, DC, (revised June 1998). http://www.fgdc.gov/metadata/csdgm/ [Accessed: December 2013].

[37] Federal Geographic Data Committee, Content standard for digital geospatial metadata: Extensions for Remote Sensing Metadata (ERSM), FGDC-STD-012-2002, Federal Geographic Data Committee, Washington, DC, http://www.fgdc.gov/metadata/ersm/ [Accessed: December 2013].

[38] Macko, P., Setlzer, M. (2011) *Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs* in TaPP

[39] http://businessanalystlearnings.com/ba-techniques/2013/3/5/moscow-technique-requirements-prioritization [Accessed: May 2014]

[40] http://bost.ocks.org/mike/sankey/ [Accessed: June 2014]

[41] Hoekstra, R. and Groth, P. (2014) *PROV-O-Viz Understanding the Role of Activities in Provenance* in the Proceedings of the International Provenance and Annotation Workshop (IPAW), Cologne, Germany.

[42] Suriarachchi, I., Zhou,Q. and Plale, B. (2014) *Komadu: A Provenance Collection and Visualization System* in Journal of Open Research Software.

[43] McClatchey, R., Branson, A., Shamdasani, J. and Kovacs, Z. (2015) *Designing Traceability into Big Data Systems*. In: 5th Annual International Conference on ICT: Big Data, Cloud and Security (ICT: BDCS 2015), Singapore, 27-28 July 2015. [In Press]

[44] Tory, M. and French, S.S. (2008) *Qualitative Analysis of Visualization: A Building Design Field Study* in BELIV '08 Proceedings of the 2008 Workshop on Beyond time and errors: novel evaluation methods for Information Visualization, Article No. 7, ISBN: 978-1-60558-016-6.

[45] Borkin, M.A., Yeh, C. S., Boyd, M., Macko, P., Gajos, K. Z., Seltzer, M. and Pfister, H. (2013, pp 2476-2485) *Evaluation of Filesystem Provenance Visualization Tools* in IEEE Transactions on Visualization and Computer Graphics, 19(12).

[46] Bjorn Meyer, B., Prohaska, S. and Hege, H. C. (2009) *Provenance Visualization and Usage*. Technical report.

[47] ProvViz: https://provenance.ecs.soton.ac.uk/vis/ [Accessed: December 2014].

[48] Charles Minard: http://www.edwardtufte.com/tufte/minard [Accessed: December 2014].

[49] LHC: http://home.web.cern.ch/topics/large-hadron-collider [Accessed: April 2014]

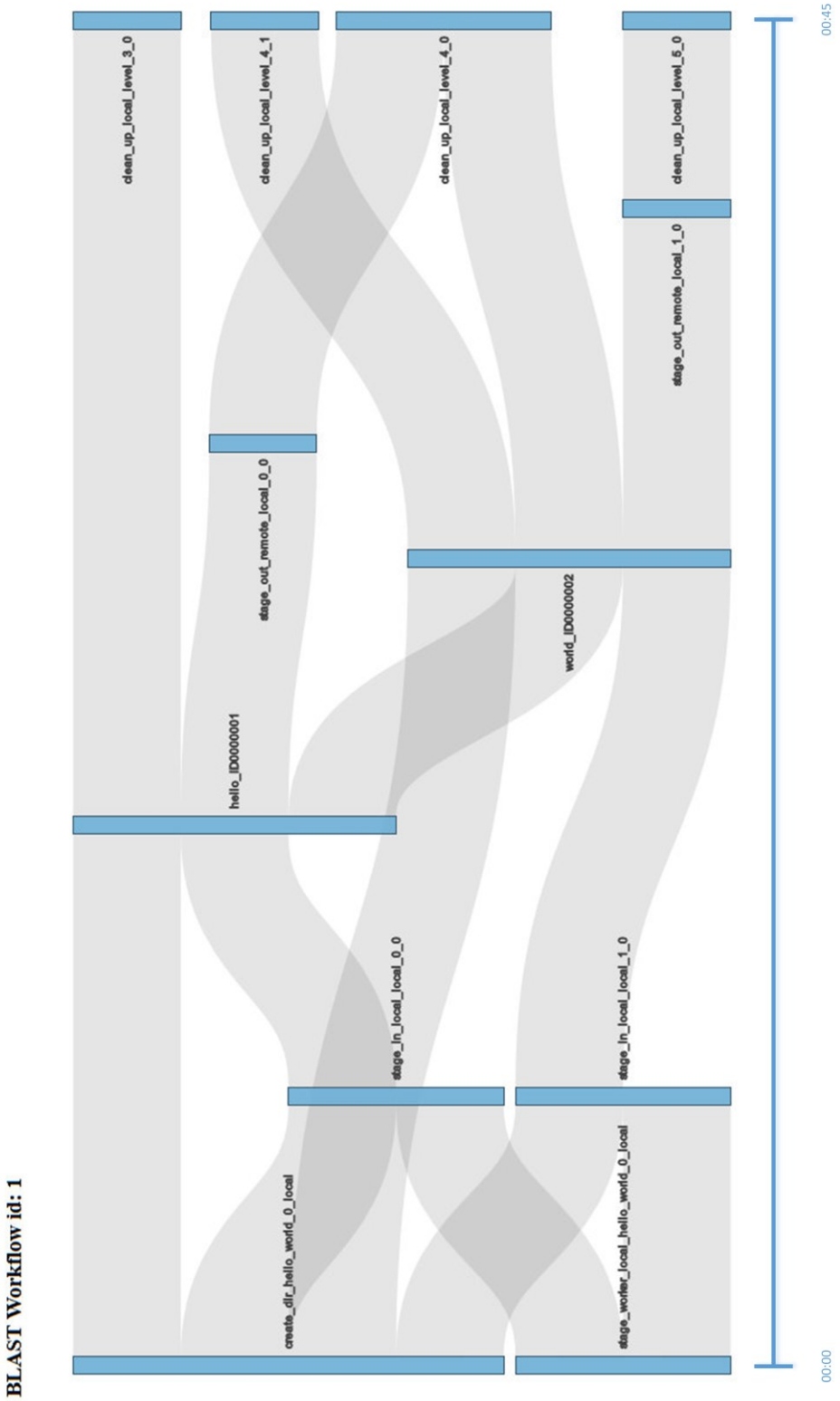
[50] Pegasus: http://pegasus.isi.edu/ [Accessed: February 2014]

[51] Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction*. John Wiley & Sons.

[52] Non-Functional Requirements: http://www.sqa.org.uk/e-learning/SDM03CD/page_02.htm [Accessed: September 2015]

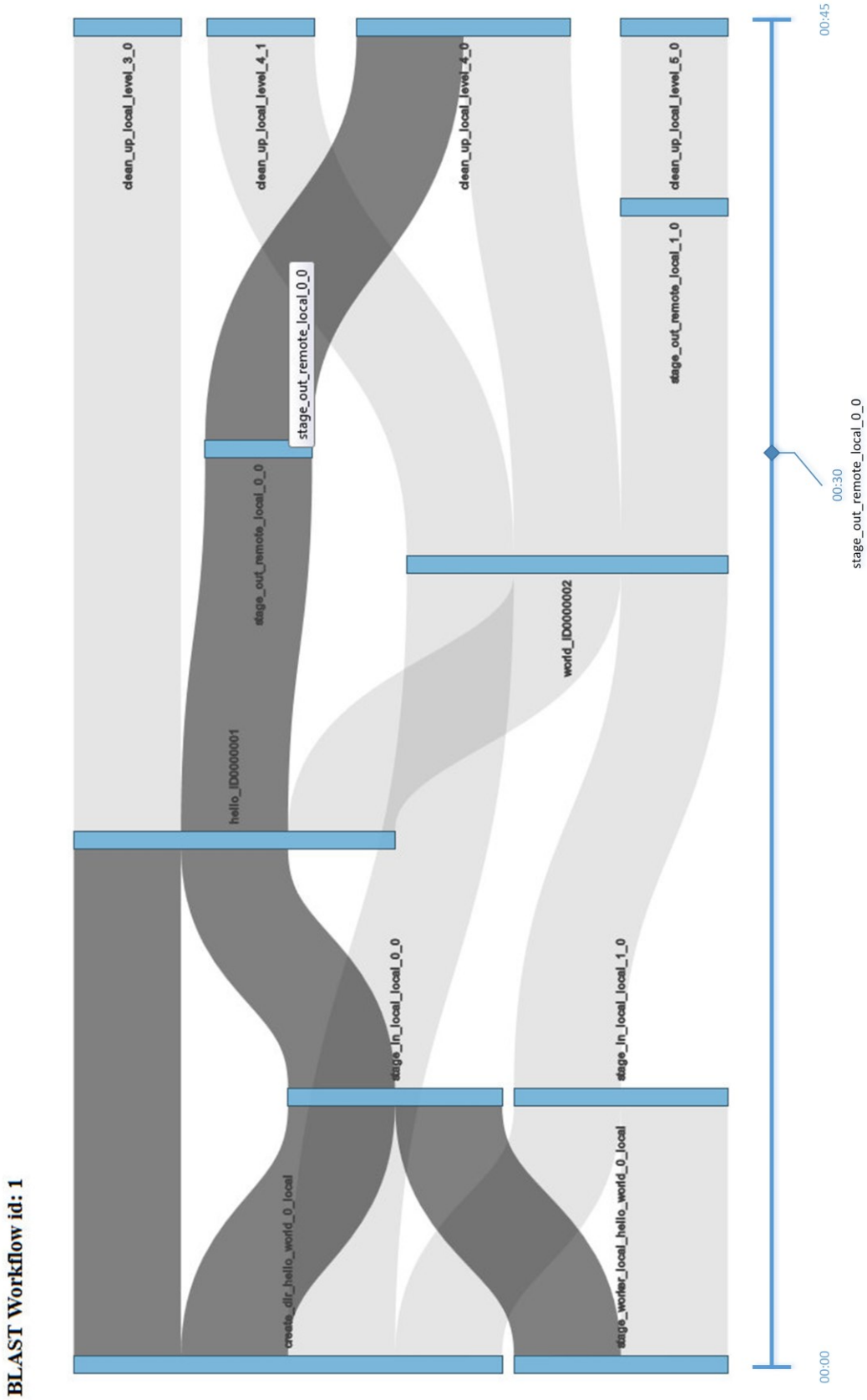
Appendix A

Use-Case 1, Case-Study 2:



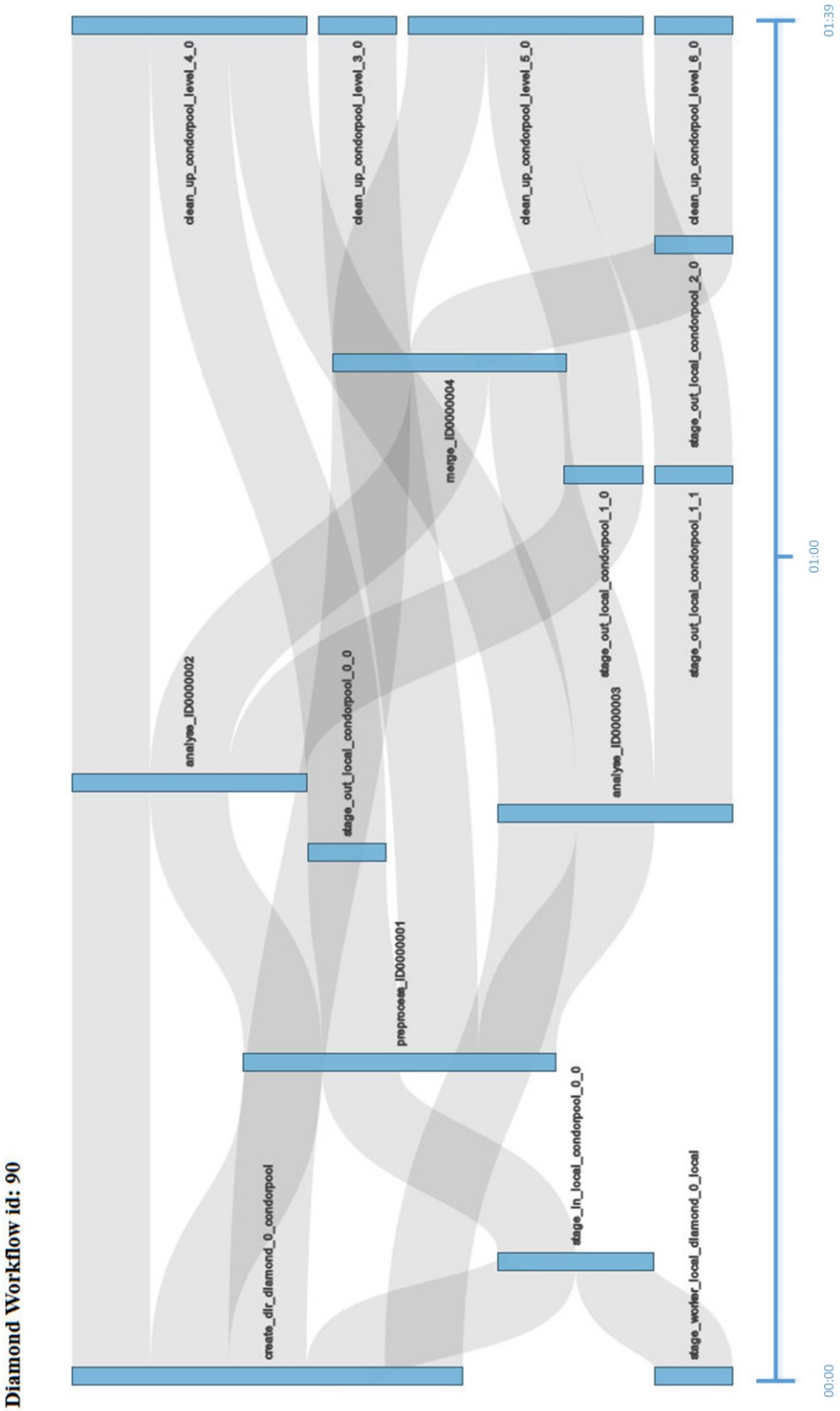
Appendix A Figure 1 - 'BLAST' Workflow

Appendix A Figure 1 shows the visualisation of provenance data for Workflow id 1. This workflow contains 12 different activities that are either consumed or generated over the course of the workflow. Appendix A Figure 2 shows that a user can click on one of the nodes for instance to inspect how an activity is produced and later consumed by different activities. The entire path is highlighted providing users with the ability to identify which activities took part in the generation or consumption of that activity and all the intermediary activities that are generated as a result of an activity under consideration.



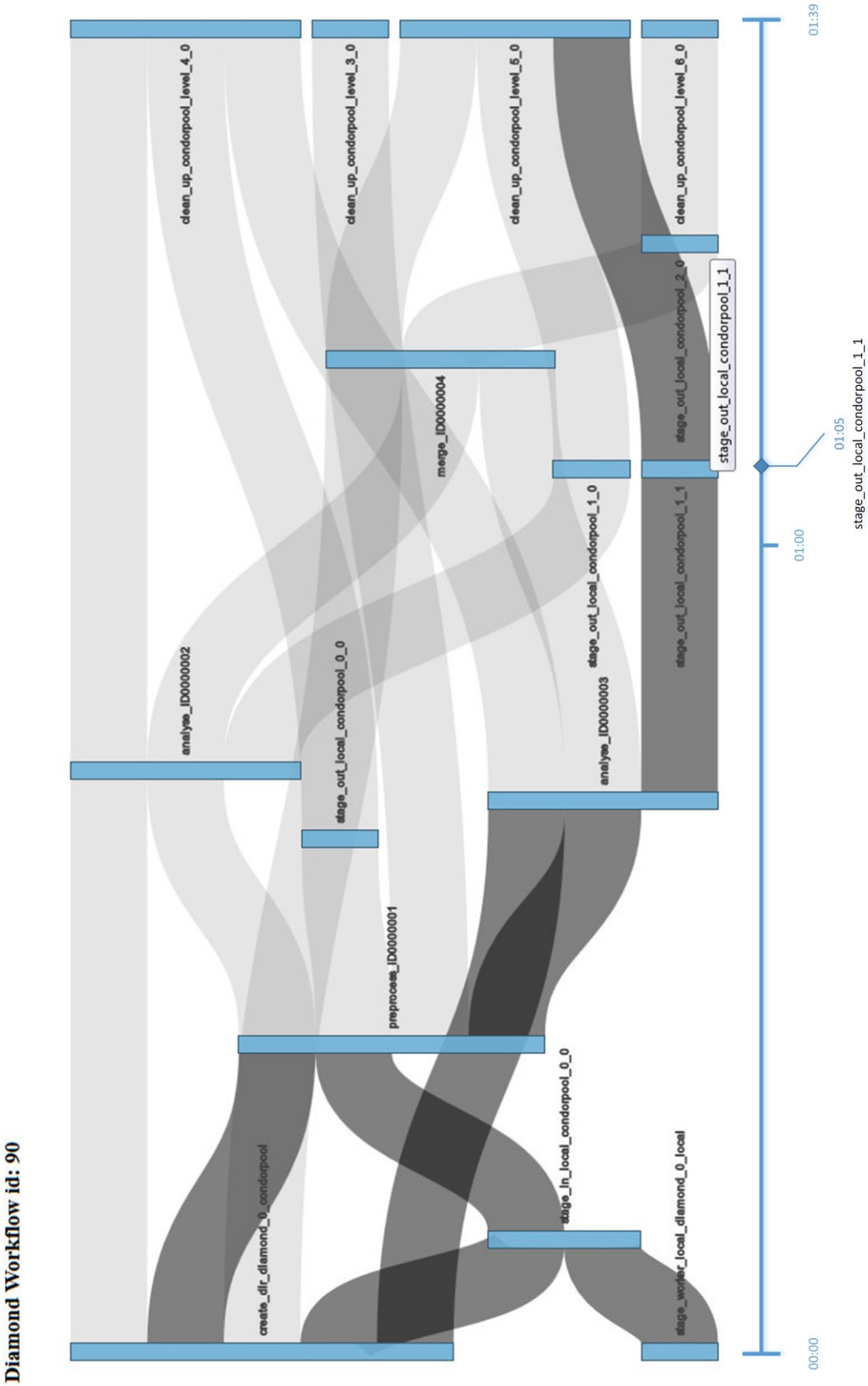
Appendix A Figure 2 'BLAST' Workflow with selected entity

In the ‘BLAST’ Workflow id number 1, the user has clicked on the ‘stage_out_remote_local_0_0’ activity and NeuroProv highlights the trace for that activity. Appendix A Figure 2 shows the trace highlighted for the activity under inspection. The activity is generated by ‘hello_ID0000001’ and led to the participation of generation of another activity i.e. ‘clean_up_local_level_4_0’. The following activities led to the generation of the activity under inspection namely ‘create_dir_hello_world_0_local’; ‘stage_worker_local_hello_world_0_local’; ‘stage_in_local_local_0_0’ and ‘hello_ID0000001’. All the activities that are highlighted in the trace are in blue. Furthermore NeuroProv also provides the user with the ability to view when an entity or an activity was generated in the context of the workflow execution timeline. Our activity ‘stage_out_remote_local_0_0’ was generated at the 30th minute of the execution time whilst the entire workflow took 45 minutes to successfully execute.



Appendix A Figure 3 'Diamond' Workflow

Appendix A Figure 3 shows the visualisation of provenance data for Workflow id 90. This workflow contains 15 different activities that are either consumed or generated over the course of the workflow. Appendix A Figure 4 shows that a user can click on one of the nodes for instance to inspect how an activity is produced and later consumed by different activities. The entire path is highlighted providing users with the ability to identify which activities took part in the generation or consumption of that activity and all the intermediary activities that are generated as a result of an activity under consideration.



Appendix A Figure 4 'Diamond' Workflow with selected activity

In the ‘Diamond’ Workflow id number 90, the user has clicked on the ‘stage_out_ local_condorpool_1_1’ activity and NeuroProv highlights the trace for that activity. Appendix A Figure 4 shows the trace highlighted for the activity under inspection. The activity is generated by ‘analyse_ID0000003’ and led to the participation of generation of another activity i.e. ‘clean_up_condorpool_local_level_5_0’. The following activities led to the generation of the activity under inspection namely ‘create_dir_diamond_0_condorpool’; ‘stage_worker_local_diamond_0_local’; ‘stage_in_local_condorpool_0_0’; ‘preprocess_ID0000001’ and ‘analyse_ID0000003’. All the activities that are highlighted in the trace are in blue. Furthermore NeuroProv also provides the user with the ability to view when an entity or an activity was generated in the context of the workflow execution timeline. Our activity ‘stage_out_local_condorpool_1_1’ was generated at the 1 Hour and the 5th minute of the execution time whilst the entire workflow took 1 hour and 39 minutes to successfully execute.

Appendix B

Workflow id 132 represented in PROV-XML to provide users with the neuroimaging workflow provenance in the PROV interoperability standard:

```
<prov:document xmlns:prov="http://www.w3.org/ns/prov#" xmlns:ex="http://example.org/"
xmlns:dct="http://purl.org/dc/terms/" xmlns:foaf="http://xmlns.com/foaf/0.1/">
<!-- Entities -->
<prov:entity prov:id="exg:wordlist"/>
<prov:entity prov:id="exg:wordlist1"/>
<prov:entity prov:id="exg:wordlist2"/>
<prov:entity prov:id="exg:analysis1"/>
<prov:entity prov:id="exg:analysis2"/>
<prov:entity prov:id="exg:merge_output"/>
<!-- Activities -->
<prov:activity prov:id="exc:analyse_ID0000002"/>
<prov:activity prov:id="exc:analyse_ID0000003"/>
<prov:activity prov:id="exc:create_dir_WordCountWF_0_condorpool"/>
<prov:activity prov:id="exc:merge_ID0000004"/>
<prov:activity prov:id="exc:preprocess_ID0000001"/>
<prov:activity prov:id="exc:stage_in_local_condorpool_0_0"/>
<prov:activity prov:id="exc:stage_in_local_condorpool_1_0"/>
<prov:activity prov:id="exc:stage_out_local_condorpool_0_0"/>
<prov:activity prov:id="exc:stage_out_local_condorpool_1_0"/>
<prov:activity prov:id="exc:stage_out_local_condorpool_1_1"/>
<prov:activity prov:id="exc:stage_out_local_condorpool_2_0"/>
<prov:activity prov:id="exc:stage_worker_local_WordCountWF_0_local"/>
<prov:activity prov:id="exc:clean_up_condorpool_level_3_0"/>
<prov:activity prov:id="exc:clean_up_condorpool_level_4_0"/>
<prov:activity prov:id="exc:clean_up_condorpool_level_5_0"/>
<prov:activity prov:id="exc:clean_up_condorpool_level_6_0"/>
<!-- Usage and Generation -->
<prov:used>
<prov:activity prov:ref="exc:preprocess_ID0000001"/>
<prov:entity prov:ref="exc:wordlist"/>
</prov:used>
<prov:used>
<prov:activity prov:ref="exc:analyse_ID0000002"/>
<prov:entity prov:ref="exc:wordlist1"/>
</prov:used>
<prov:used>
<prov:activity prov:ref="exc:analyse_ID0000003"/>
<prov:entity prov:ref="exc:wordlist2"/>
</prov:used>
<prov:used>
<prov:activity prov:ref="exc:merge_ID0000004"/>
<prov:entity prov:ref="exc:analysis1"/>
</prov:used>
<prov:used>
<prov:activity prov:ref="exc:merge_ID0000004"/>
```

```
<prov:entity prov:ref=“ex:analysis2”/>
</prov:used>
<prov:wasGeneratedBy>
<prov:entity prov:ref=“exc:preprocess_ID0000001”/>
<prov:activity prov:ref="exc:wordlist1"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:entity prov:ref=“exc:preprocess_ID0000001”/>
<prov:activity prov:ref="exc:wordlist2"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:entity prov:ref=“exc:analyse_ID0000002”/>
<prov:activity prov:ref="exc:analysis1"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:entity prov:ref=“exc:analyse_ID0000003”/>
<prov:activity prov:ref="exc:analysis2"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:entity prov:ref=“exc:merge_ID0000004”/>
<prov:activity prov:ref="exc:merge_output"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref=“exc:analyse_ID0000002”/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref=“exc:analyse_ID0000003”/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref=“exc:merge_ID0000004”/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref=“exc:preprocess_ID0000001”/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref="exc:stage_in_local_condorpool_0_0"/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
<prov:activity prov:ref="exc:stage_in_local_condorpool_1_0"/>
<prov:activity prov:ref="exc:create_dir_WordCountWF_0_condorpool"/>
</prov:wasGeneratedBy>
<!-- Communication -->
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_4_0"/>
<prov:informant prov:ref="ex:analyse_ID0000002"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
```

```
<prov:informed prov:ref="ex:merge_ID0000004"/>
<prov:informant prov:ref="ex:analyse_ID0000002"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_out_local_condorpool_1_0"/>
<prov:informant prov:ref="ex:analyse_ID0000002"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_4_0"/>
<prov:informant prov:ref="ex:analyse_ID0000003"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:merge_ID0000004"/>
<prov:informant prov:ref="ex:analyse_ID0000003"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_out_local_condorpool_1_1"/>
<prov:informant prov:ref="ex:analyse_ID0000003"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:analyse_ID0000002"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:analyse_ID0000003"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:merge_ID0000004"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:preprocess_ID0000001"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_in_local_condorpool_0_0"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_in_local_condorpool_1_0"/>
<prov:informant prov:ref="ex:create_dir_WordCountWF_0_condorpool"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_5_0"/>
<prov:informant prov:ref="ex:merge_ID0000004"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_out_local_condorpool_2_0"/>
<prov:informant prov:ref="ex:merge_ID0000004"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
```

```
<prov:informed prov:ref="ex:analyse_ID0000002"/>
<prov:informant prov:ref="ex:preprocess_ID0000001"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:analyse_ID0000003"/>
<prov:informant prov:ref="ex:preprocess_ID0000001"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_3_0"/>
<prov:informant prov:ref="ex:preprocess_ID0000001"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_out_local_condorpool_0_0"/>
<prov:informant prov:ref="ex:preprocess_ID0000001"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_4_0"/>
<prov:informant prov:ref="ex:stage_out_local_condorpool_0_0"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_5_0"/>
<prov:informant prov:ref="ex:stage_out_local_condorpool_1_0"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_5_0"/>
<prov:informant prov:ref="ex:stage_out_local_condorpool_1_1"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:clean_up_condorpool_level_6_0"/>
<prov:informant prov:ref="ex:stage_out_local_condorpool_2_0"/>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_in_local_condorpool_0_0"/>
<prov:informant prov:ref="ex:stage_worker_local_WordCountWF_0_local"/>
</prov:wasInformedBy>
<prov:wasInformedBy>
<prov:informed prov:ref="ex:stage_in_local_condorpool_1_0"/>
<prov:informant prov:ref="ex:stage_worker_local_WordCountWF_0_local"/>
</prov:wasInformedBy>
</prov:document>
```