

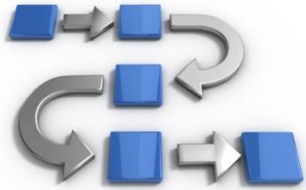
Impact of Software Environment on Replicability of Biomedical Workflows

Tomasz Miksa

Vienna University of Technology, Austria

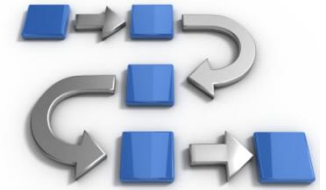
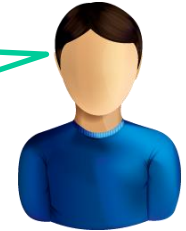
miksa@ifs.tuwien.ac.at

Replicability



Original
experiment

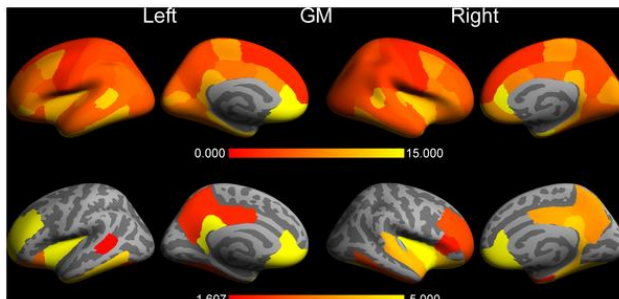
- Can I reuse her workflow?
- Do the results match?
- Have I done it correctly?



Re-executed
experiment



- Current studies show very low reproducibility in
 - medicine
 - economy
 - computer science
- Reproducibility requires
 - well documented research workflows
 - precise information on the experiment's environment



OPEN ACCESS Freely available online PLOS ONE

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild^{1,2*}, Petra Habets^{1,2}, Heidi L. L. Jacobs^{1,2,3}, Ron Mengelers^{1,2}, Nico Rozenaad^{1,2}, Jim van Os^{1,2,4}, Machteld Marcellis^{1,2}

1 Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, Alzheimer Center Limburg, The Netherlands, 2 European Graduate School of Neuroscience (EURON), Maastricht University, Maastricht, The Netherlands, 3 Cognitive Neurology Section, Institute of Neuroscience and Medicine-3, Research Centre Jülich, Jülich, Germany, 4 King's College London, King's Health Partners, Department of Psychosis Studies, Institute of Psychiatry, London, United Kingdom

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies. The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Citation: Gronenschild EHM, Habets P, Jacobs HLL, Mengelers R, Rozenaad N, et al. (2012) The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. PLoS ONE 7(6): e38234. doi:10.1371/journal.pone.0038234

Editor: Satoshi Hayasaka, Wake Forest School of Medicine, United States of America

Received: January 12, 2012; Accepted: May 1, 2012; Published: June 1, 2012

Copyright: © 2012 Gronenschild et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Geesick program of the Dutch Health Research Council (CON-MW, grant number 10-000-1002), and the European Community's Seventh Framework Program under grant agreement No. HEALTH-F2-2009-241909 (Project EU-GIG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ed.gronenschild@maastrichtuniversity.nl

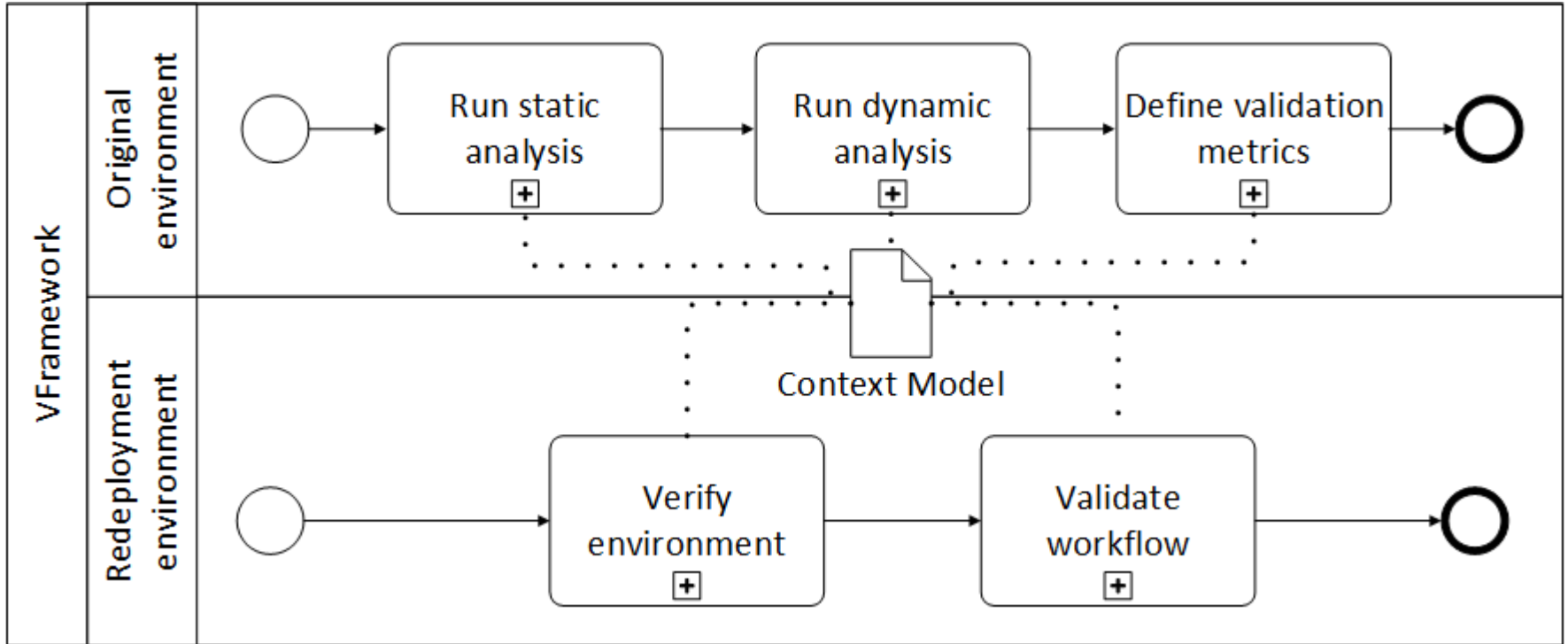
Introduction

FreeSurfer (Athinoula A. Martinos Center for Biomedical Imaging, Harvard/MIT, Boston) comprises a popular and freely available set of tools for deriving neuroanatomical volume and cortical thickness measurements from automated brain segmentation (<http://surfer.nmr.mgh.harvard.edu>), recently summarised by Fischl [1]. A number of reported studies discussed the accuracy of the technique by comparing the volume of specific brain structures, such as the hippocampus or amygdala, with manually derived volumes [2–5]. The measurement of cortical thickness was validated against histological analysis [6] and manual measurements [7,8]. Also the reliability of the measurements was subject of a number of investigations. Some of these studies addressed the effect of scanner-specific parameters, including field strength, pulse sequence, scanner upgrade, and vendor (cortical thickness) [9,10].

Since FreeSurfer is CPU-intensive (20–30 hours per brain for a full segmentation is not exceptional), it is common practice to distribute the computational load among the available central processor units (CPUs) on a single workstation and/or among several workstations. Given this context, a number of questions suggest themselves: (1) does every CPU produce the same results; (2) is there any interaction between the processes running simultaneously on the same workstation; (3) does every workstation produce the same results?

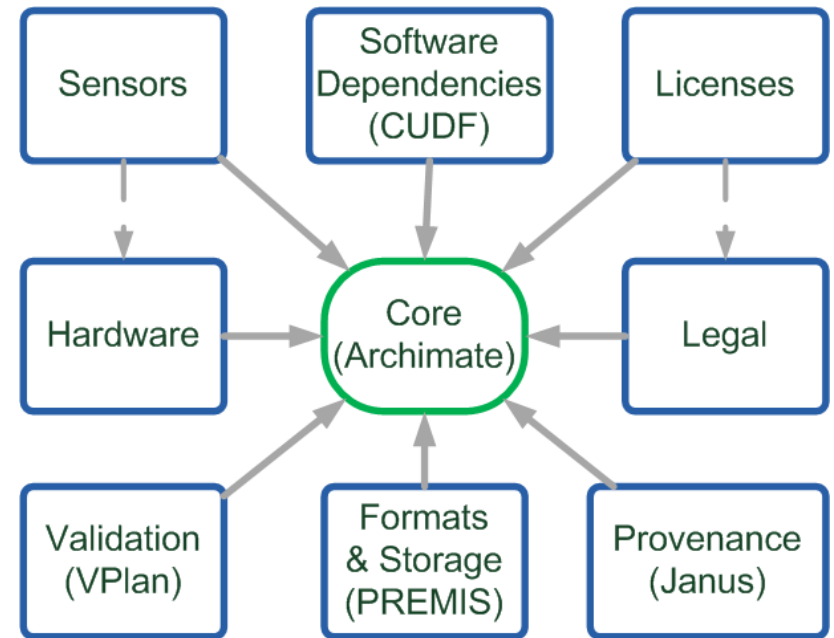
Just like similar neuroimaging packages, new releases of FreeSurfer are issued regularly, fixing known bugs and improving existing tools and/or adding new ones. Each release is accompanied with documentation describing the changes relative to the previous release (<http://surfer.nmr.mgh.harvard.edu/swiki/ReleaseNotes>). However, transition to a new release during the course of a study may affect the results and is therefore

VFramework



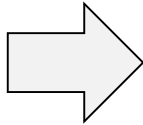
Context Model

- OWL ontology
- Modular architecture
 - Domain Independent Ontology
 - Domain Specific Ontologies
- Process preservation

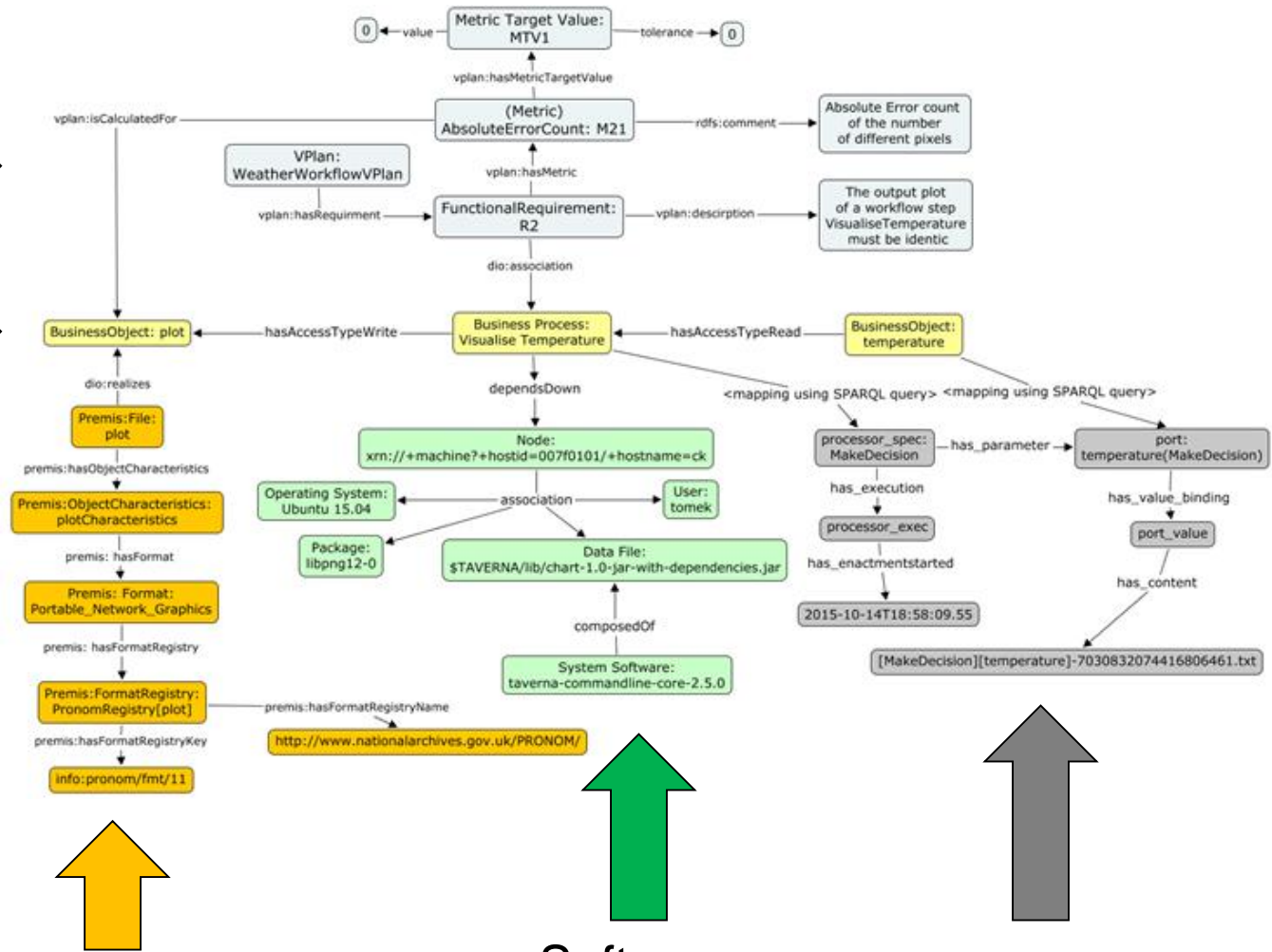
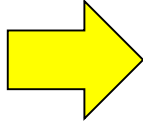


Context Model

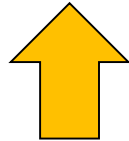
Validation metrics



Workflow definition



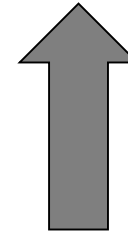
File formats



Software dependencies



Provenance



Validation report for the WeatherExample

Evaluation result: There are 2 not fulfilled metrics. Please see tables below for details.

Comparison performed using following workflow execution traces

Original Workflow

ID: 37b4d2fb-e71c-4b67-b7b3-17888ee82977

Timestamp: 2015-10-14 18:58:06.475

Compared Workflow

ID: ede04b87-5f58-4a89-b0c3-e179957cbad0

Timestamp: 2015-11-13 13:59:56.443

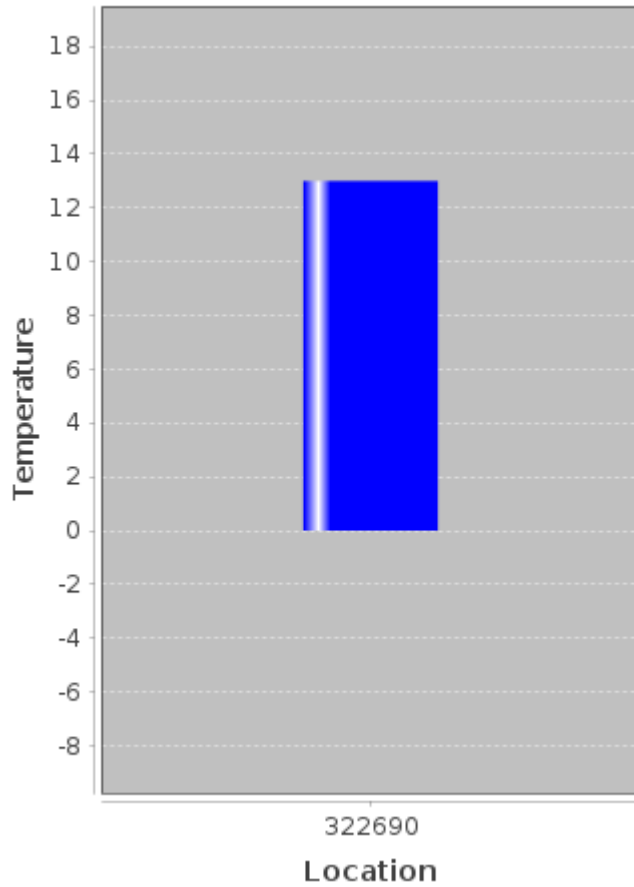
Table 1: Overview of requirements

Requirement	Description	Is Fulfilled
R1	The inputs to the workflow are the same	true
R2	The outputs of the workflow are the same	false
R3	The workflow step ExtractTemperature must have identical outputs	true
R4	The workflow step GetWeatherData must have identical outputs	true
R5	The workflow step MakeDecision must have identical outputs	true
R6	The workflow step ExtractWeatherType must have identical outputs	true
R7	The workflow step VisualiseTemperature must have identical outputs	false
R8	Execution duration of each of the workflow steps shall be similar	true

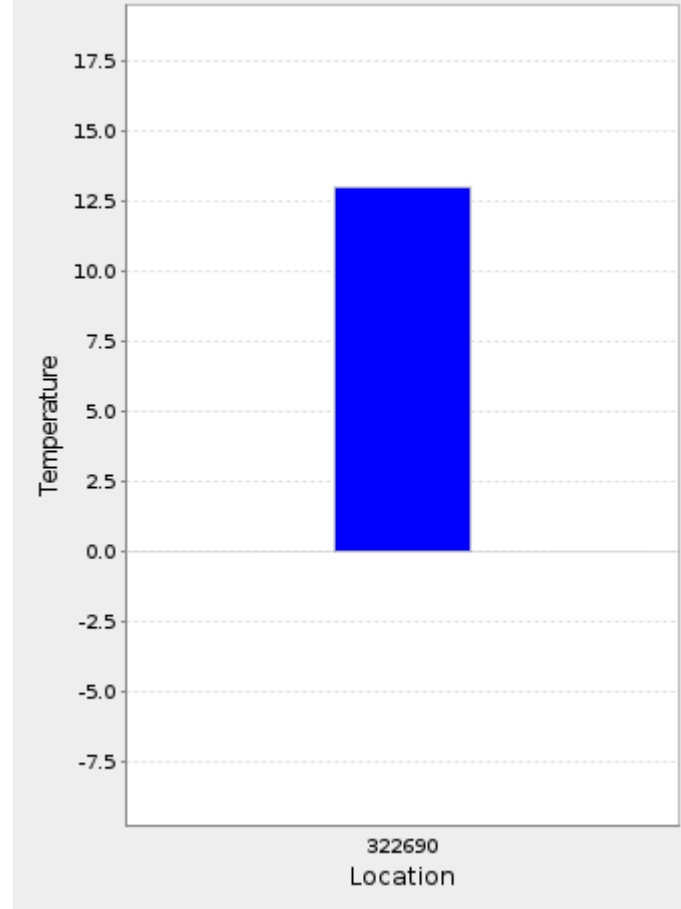
Table 2: List of requirements and metrics that failed.

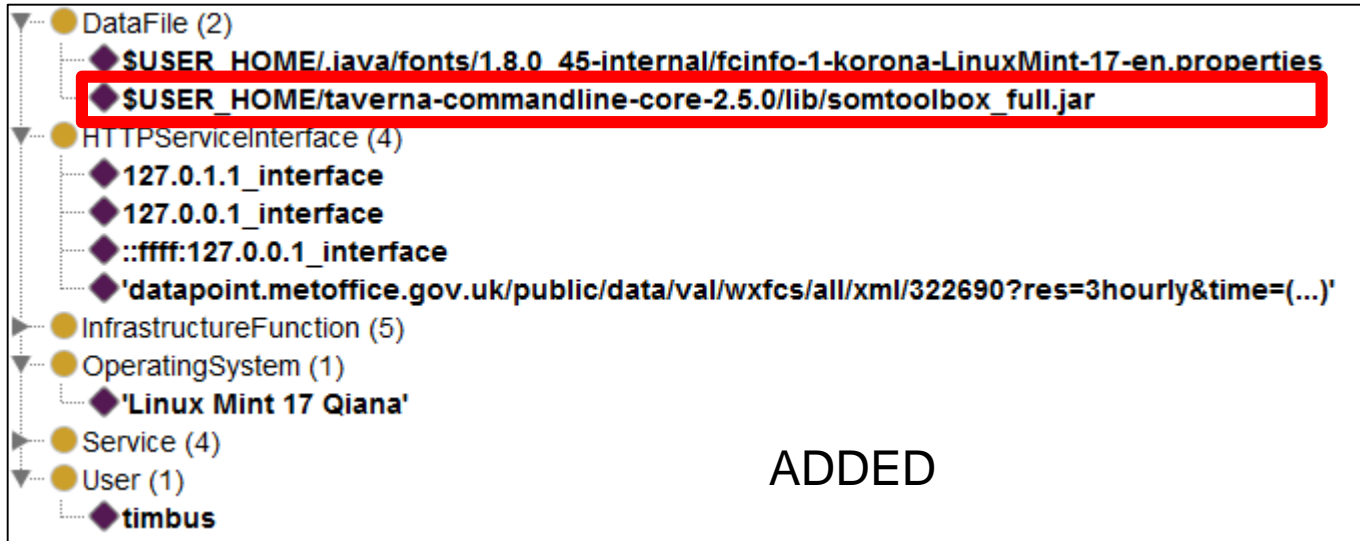
Req	Sub-req	Sub-requirement description	Measurement point	Metric	Validity
R2	R7.1	The output plot of workflow step VisualiseTemperature must be identical	plot	ImageFingerPrintEquality	false
				ImageResolutionEquality	true
				AbsoluteErrorCount	false
R7	R7.1	The output plot of workflow step VisualiseTemperature must be identical	plot	ImageFingerPrintEquality	false
				ImageResolutionEquality	true
				AbsoluteErrorCount	false

Temperature in 322690



Temperature in 322690

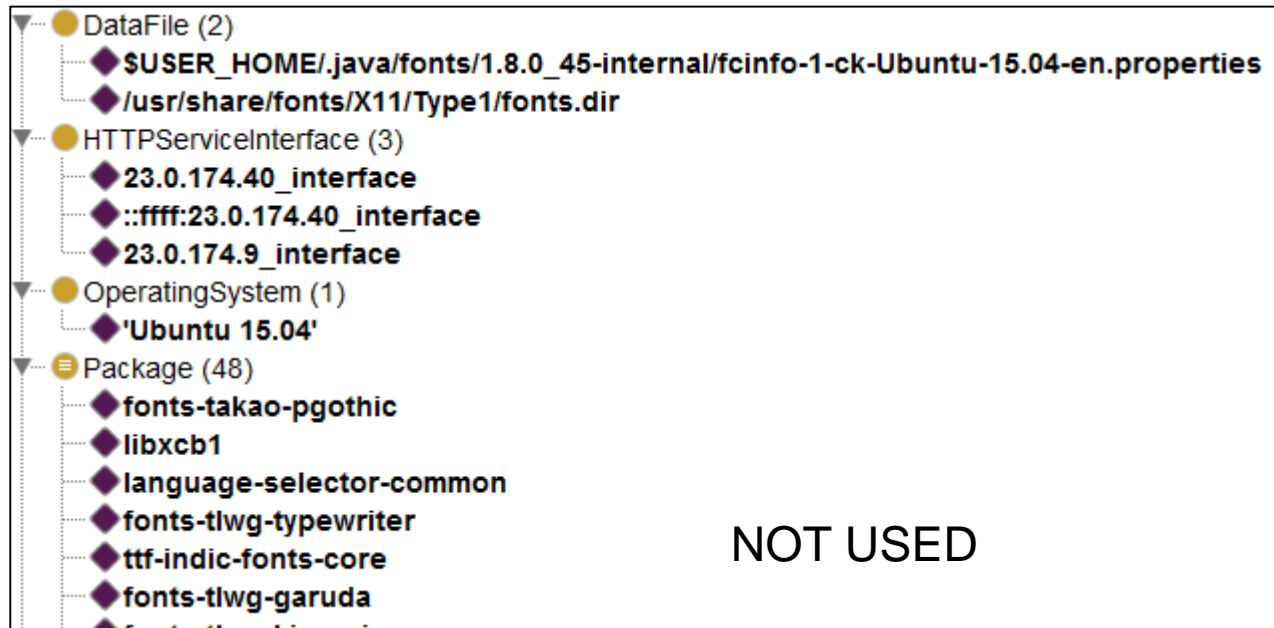




File system tree showing the following structure:

- DataFile (2)
 - \$USER_HOME/.java/fonts/1.8.0_45-internal/fcinfo-1-korona-LinuxMint-17-en.properties**
 - \$USER_HOME/taverna-commandline-core-2.5.0/lib/somtoolbox_full.jar**
- HTTPServiceInterface (4)
 - 127.0.1.1_interface
 - 127.0.0.1_interface
 - ::ffff:127.0.0.1_interface
 - 'datapoint.metoffice.gov.uk/public/data/val/wxfcs/all/xml/322690?res=3hourly&time=(...)'
- InfrastructureFunction (5)
- OperatingSystem (1)
 - 'Linux Mint 17 Qiana'
- Service (4)
- User (1)
 - timbus

ADDED



File system tree showing the following structure:

- DataFile (2)
 - \$USER_HOME/.java/fonts/1.8.0_45-internal/fcinfo-1-ck-Ubuntu-15.04-en.properties
 - /usr/share/fonts/X11/Type1/fonts.dir
- HTTPServiceInterface (3)
 - 23.0.174.40_interface
 - ::ffff:23.0.174.40_interface
 - 23.0.174.9_interface
- OperatingSystem (1)
 - 'Ubuntu 15.04'
- Package (48)
 - fonts-takao-pgothic
 - libxcb1
 - language-selector-common
 - fonts-tlwg-typewriter
 - ttf-indic-fonts-core
 - fonts-tlwg-garuda

NOT USED

- Characteristics of Taverna workflows on myExperiment

- 50% use Beanshells
- 30% use WSDL web services
- 15% local tool invocations



- Access to the original environment was necessary

- 5 Taverna workflows from 3 domains

- 3 *biomedical* workflows (Leiden University, NL)
- 1 *music classification* workflow (TU Wien)
- 1 *sensor data analysis* workflow (LNEC, Portugal)



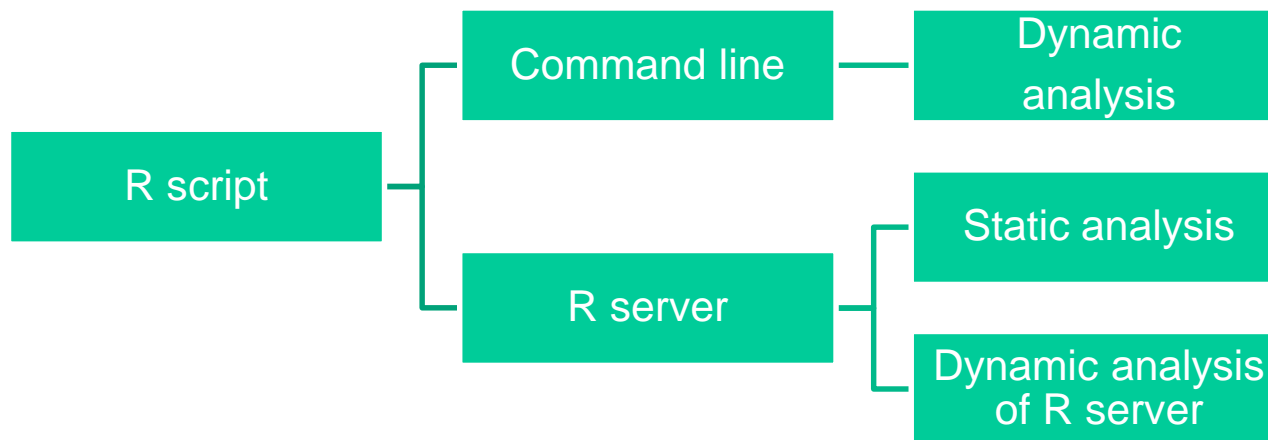
Use Cases

- Workflows required
 - specific dependencies in the environment
 - external services
- Re-executions on different
 - version of operating system
 - distribution
 - system architecture



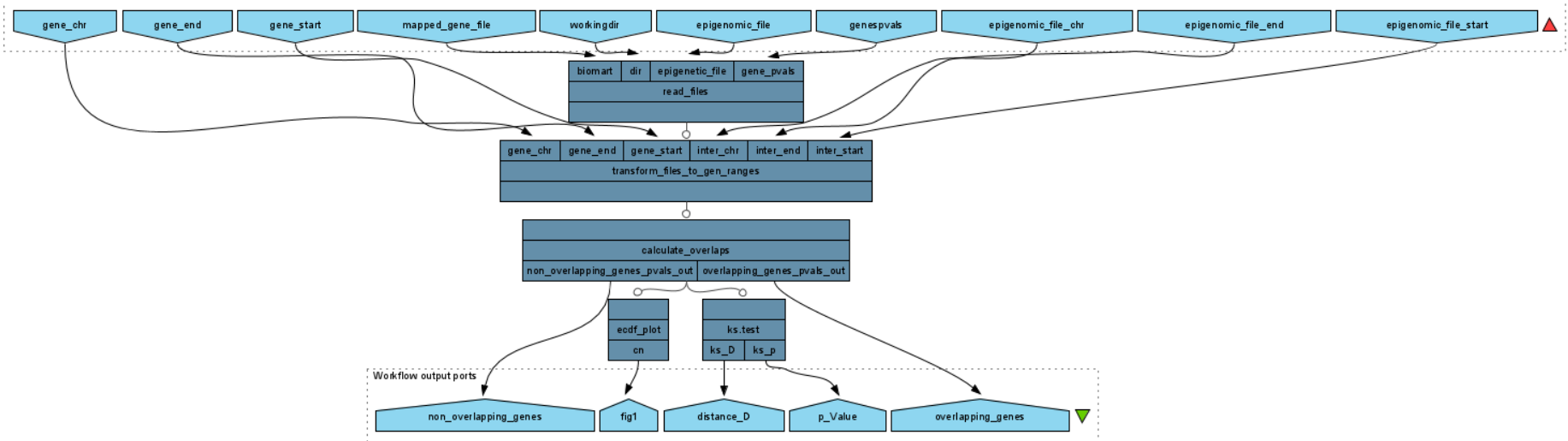
```
#!/bin/bash
```

- All dependencies were identified
 - R dependencies
 - Ruby dependencies
 - Java libraries
 - local tools (e.g. pdflatex)



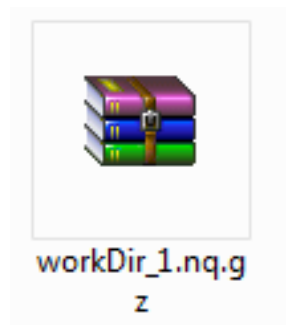
Results– external services

- All external communications detected (web services)
- Limited validation for external services (Rserve scripts)
 - global variables
 - steps with no outputs
 - only final result of workflow computation is validated
 - black box testing



- Dynamic analysis detected
 - data created through shell calls
 - such files are not a part of provenance traces in Taverna
 - ‘real’ workflow outputs
 - Taverna traces can contain paths only, but not the content
 - Taverna workflows can create files not linked to any output

```
I, [2015-11-30T15:44:14.448859 #8787] INFO -- : row 1
skipped. I, [2015-11-30T15:44:15.159204 #8787] INFO -- :
===== running time total: 0.710367914 =====
No of statements in a file 29590 I,
[2015-11-30T15:44:15.159824 #8787] INFO -- : 0.708272803
```



- detected changes on different stages of workflow processing
 - metadata included in data
 - generation timestamp
 - file format comparison improved the results
 - ZIP archives

```
% latex table generated in R 3.0.2 by xtable 1.7-1 package
% Tue Apr 21 13:37:05 2015
\begin{table}[ht]\centering
\begin{tabular}{rrrrrr}
\hline & Min & 1Q & Mediana & Media & 3Q & Max \\
\hline Residuos (mm) & -2.42 & -1.23 & -0.52 & -0.00 & 0.45 & 5.52 \\
\hline
\end{tabular}
\caption{Residuos do modelo}
\end{table}
```

Tabela 1: Dados gerais do modelo de IQ

Utilizador admin	
Data da execu \tilde{c} o	21-04-2015 12:37
Barragem	275
Fim da constru \tilde{c} o	1943-04-01
In \tilde{c} io do primeiro enchimento	1960-01-01
Tipo de instrumento	125
Grandezas	Desl. radial (mm) Desl. tangencial (mm)
N \tilde{u} instrumentos	4
Designa \tilde{c} o	FP1 FP1 FP2 FP2
Elementos da matriz X	CTE H, H4 COSD, SEND T
Per \tilde{c} odo an \tilde{a} lise	[2005-01-01,2013-01-01]
Fim da IQ	2008-01-01

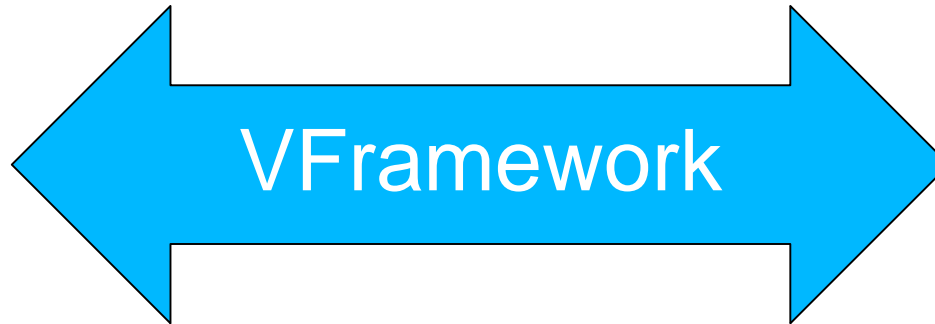
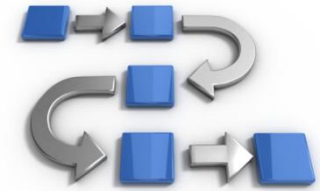
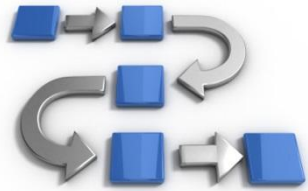
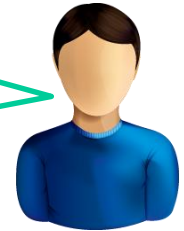
Recommendations

- Analyse dependencies and evade shell calls
 - e.g. use scripting mechanisms provided by the workflow engine
- Write code that runs on all platforms
 - e.g. do not encode specific paths
- Publish experiment setup and context
 - e.g. exact versions of tools used
- Publish validation data
 - e.g. provenance but also other files created during execution
- Test the replicability on your own
 - e.g. try rerunning your experiment in a clean virtual machine

Conclusions



I have repeated her experiment
in the same way!
I got the same results!
I can reuse any part of it!



Original
experiment



Re-executed
experiment