THE UNIVERSITY of EDINBURGH

Centre for Intelligent Systems and their Applications
School of Informatics
The University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom
Tel: +44(0)131 651 4040

Personal comment on the Thesis and PhD defence of Daniel Garijo

Thesis title:
Mining abstractions in Scientific workflows

Commentary by Prof. Malcolm Atkinson *Malcolm.Atkinson@ed.ac.uk*

The following commentary was prepared and sent on 9th December 2015 as I was so impressed by the thesis and defence that I felt an explicit record should be made available immediately. I have left that commentary unchanged but add a few additional observations at the end. In summary, this was an exceptionally outstanding PhD and Dr Garijo, has continued to build on this work and to contribute substantially to advancing the semantics and usability of scientific workflows. A very high impact thesis launching a high-flying career which will deliver much needed advances in our ability to use data-driven methods to address global challenges.

This commentary is public and may be used for any purpose which those at UPM, including Daniel and his two supervisors consider appropriate. I am prepared to answer questions about any aspect of this document. It should **not** be construed as part of the formal PhD examination process, which was well conducted and came to a similar conclusion. I felt moved to write as I was so impressed by the very high quality of the thesis, the research behind the thesis and the discussions during the defence on 3 December 2015 at UPM.

The thesis has an unusually clear logical structure, which is well signposted and consistently carefully thought out and presented to advance the arguments behind the research. Each topic is carefully dissected and handled in one place. Each group of topics is introduced with the right level of detail, context setting and cross references. The writing throughout is crafted to be succinct, clear and sufficient. It is complemented by outstandingly well-drafted figures and clear, brief summary tables. Each group of topics is well summarised, with a careful lead into the next part. The arguments and observations are well supported by references and URLs. A balance between readability and precision is achieved by using a few mathematical definitions and brief statements of hypotheses, objectives and issues. The appendices are well thought out and continue the high standard. I have often encountered a thesis that achieved some of those merits, but I have not encountered another which was as outstanding throughout in every aspect in nearly 50 years of being involved in computing science research at universities and in industry.

The research covers a great deal of ground developing a systematic and rational path from identifying shortcomings in the existing support for understanding and re-using workflows, through a series of necessary steps to a demonstration that a class of motifs (recurring patterns) can be automatically mined and presented to help those creating workflows. The research maintains a careful balance between theoretical and conceptual advances and the demonstration

of their feasibility and utility. These depend on a breadth of abilities rarely seen in one PhD: going from a capacity to be creative in formal contexts, through the analysis of new technical opportunities triggered by the formal advances to practical tests of those technologies. The practical analysis a review of large repositories by hand, it included the measurement of the quality of results, where appropriate formulating new metrics, and it included interaction with practitioners to gather and analyse their opinions.

The coverage of the background was well structured under three headings and covered a considerable breadth of sources and workflow systems succinctly. The care taken to avoid the inclusion of extraneous material and the sharply defined conclusions were both commendable.

The third chapter of the thesis was unusual and extremely valuable. It established the hypothesis clearly, subdivided it into tractable issues and developed these into objectives neatly brought together in a roadmap. Rarely have I seen such clarity of thought about the research challenges and how they should be approached, even among colleagues much more experienced than PhD students. The full set of challenges and objectives involve a prodigious amount of research work that would have daunted most researchers and certainly well exceeds the normal scope of a PhD.

The fourth chapter tackles the research challenge of improving the representation of workflows, particularly when considering their publication and logical structure. The creation of OPMW extends P-Plan and builds nicely on previous work and standards and meets all three detailed requirements. It is a careful extension of the previous approaches that shows depth of understanding and ingenuity. As always, it is very clearly presented, with excellent summarising diagrams and careful writing. This is then used for publishing scientific workflows developed in Wings as linked open data and using the resulting representations for a range of purposes. The approach was validated against a demanding set of competency questions—a formalisation of assessment I hadn't seen used in computing science research before.

Improving the level of abstraction for workflows, the topic of the next chapter, is a very important challenge as workflows are the basis for developing, refining and defining scientific methods. Such improvement will underpin understanding and thereby enable effective sharing. I have long considered improving the level of abstraction used in this context to be a high priority. A key insight by Daniel, was to look for and to characterise *workflow* motifs, repeated patterns that occur both within and between workflows. The first step taken was to *manually* analyse 260 different workflows from active repositories covering Galaxy, Taverna, VisTrails and Wings workflow systems. Reading, comprehending, and characterising so many different workflows in different systems, with a range of ten application domains was a very brave and industrious undertaking. It yielded significant insights and proposals for a set of relevant abstractions: six for data-operation motifs and six for workflow-oriented motifs. Unsurprisingly there was a very high proportion of data-preparation motifs and these were subdivided into eight sub-categories. In my experience these include steps where the scientist is providing key insights on how to home in on relevant information and how to combine data from previously unrelated sources. Data preparation can also include mundane transformations to handle formats while preserving information and to even organise data movement in Taverna. That scientists should have to handle process steps that ought to be automated and hidden behind abstraction is reprehensible as it could impede innovation and discovery. On the other hand, scientists will always wish to selectively filter and transform each kind of data input, and to choose how they handle missing

observations. The advances made in identifying this category are therefore very welcome, but some aspects of data preparation will remain topics for which research scientists in many domains wish to carefully control. Perhaps this characterisation of the data preparation motifs is worthy of further clarification in due course to better articulate that boundary. The identification of visualisation and human-interaction motifs, which often go together, is an issue we have experienced in the VERCE project, and Alessandro Spinuso has facilitated through a provenance framework that enables us to circumvent the operational constraints inhibiting such actions when workflows run on HPC resources. This enabled seismologists to see whether their data-driven workflows were running as they hoped, and to save very significant resources and energy if they weren't.

The analysis of 851 different LONI pipeline workflows looks at two different patterns of use: single user dominated and multi-user, in a very well established collection of workflows enabling neurological image processing research. This is a further extension of the technologies and domains covered, and provides a path to a different user community with strong views and established working practices. The workflows were automatically transformed to a P-Plan representation, and then compared using generated SPARQL queries over the resulting RDF. As this is an order N^2 correlation, and N is effectively larger than 851 (though possibly correlations between the four corpora were not considered), as sub-graphs are taken into account, so just organising this analysis must have been a prodigious effort. The outcome is carefully reduced to an understandable set of tabulations that clearly show re-use in both patterns of workflow-corpus use. This observed re-use of workflows and workflow fragments was further investigated using a carefully designed web-based questionnaire. Given the prevalence of questionnaires in today's research, and the sharp focus on medical-imaging research in the LONI community, a result of 21 returned questionnaires was a commendable achievement. The outcome supported the case for improving the opportunities for workflow re-use and for workflow sharing.

Having established the motivation for workflow sharing the next stage of the research is a very ambitious step. Previously, sharing has relied on users, particularly developers adept at workflow creation, identifying sharable units or spotting fragments in other workflows they can re-use. As each corpus of available workflows grows, the ability of humans to identify potential sharing is being outstripped even within each repository. Reaching across repositories, taking advantage of a common notation for representing workflows, would be even more demanding. Yet the best scientific method for performing a particular task may well be encoded in another workflow language. Hence the investigation and establishment of an automated approach for discovering repeated fragments is a very significant research step, which will have far reaching benefits in the long run. The challenge is carefully formalised and identified as being an instance of the much-studied graph isomorphism problem. The combinatorial complexity of NP-complete, may be exacerbated by the potential range of graph fragments that could be considered from each workflow in all of the others, and from the added complexity of matching when approximation via a hierarchy of workflow step types is considered. The workflow graphs are pruned to their fundamentals, in preparation. The choice of candidate fragments is also carefully controlled to reduce the search space. Then one inexact and two exact matching algorithms are tried over the candidate corpora. Although these algorithms were "simply" imported, deciding how to use them, and organising their use over the target workflow graphs must have taken significant intellectual and technical effort. It looks simpler when reported in retrospect than I am sure it was when faced for the first time in practice. The careful filtering of

these results to eliminate fragments that were covered by larger fragments was a significant further step, significantly reducing the set of outputs. Once these candidate fragments had been identified, the use of automatically generated SPARQL graph queries to find their occurrences was a substantial further step, showing considerable mastery over the open linked data systems and methods. This has introduced and demonstrated the feasibility of the innovative formal and technical framework for finding, exposing and linking potentially re-usable fragments, so that they can be flagged to potential (re-)users, or so they can receive optimised support when they are used frequently. This will have a major impact on future scientific methods research. It will go beyond workflow re-use, as it has the potential to accelerate the recognition and re-use of the research methods and practices that the workflows encode.

Chapter 7 starts by defining evaluation metrics for assessing the quality and value of the results achieved by the automatic pattern matching in the previous chapter. These are well judged and inspired in part by traditional information retrieval metrics, but the usefulness measure is entirely new. It is appropriate to postpone performance issues to later research. The metrics were evaluated for all of the automated pattern-detection methods with and without generalisation, over a new corpus, a set of 17 text-mining workflows rendered in Wings. The evaluations and analyses are conducted with meticulous care, and the conclusions, even though they strongly support the hypothesis of the value of automated workflow-fragment mining, are expressed with appropriate caution. The final evaluation was performed by mining candidate re-usable workflow fragments from the four LONI pipeline corpora previously used. This provided further evidence of the potential utility of discovered workflow fragments based on assessing the introduced usability metric. A trial was then conducted with three domain experts, who were presented with samples from the mined patterns, and asked to evaluate their re-usability. This was a fairly small group of experts for an evaluation, and care had to be taken not to overload them, as they were presented with candidate re-usable fragments. As always, great care was shown in the design and interpretation of the experiment. The final analysis in this chapter composes a series of evidential steps to clearly demonstrate the power and value of the automated fragment mining approach in this setting. A *tour de force* argument for the hypothesis and its key components; made more remarkable when we consider that this was achieved in a very short elapsed period, and that very nearly all of the work was accomplished by the PhD student himself.

The concluding chapter draws all of this together in elegant and concise arguments without distracting ideas or wasted words. It is expressed with clarity. The initial impact and the further work make it very clear that the context of the work is well understood, and that Daniel Garijo is not only successful at leading, organising, doing and reporting his own research. He is also quite capable of planning and leading a campaign to take it much further.

The review of the detailed steps in developing the argument, creating so many experimental stages, and in delivering each analysis clearly, makes the power, agility and sustained commitment of the mind and individual very apparent. A remarkable number of research contexts, application domains, theoretical frameworks and technologies were mastered and effectively deployed. The patient attention to detail and precision while excluding extraneous side issues *throughout the whole thesis* is an extremely rare achievement. It is clear that Daniel is also adept at building effective research relationships very quickly, as evidenced by the number

of groups he worked with and the number of papers in which he participated–nine of which where he was first author.

This thesis provides unassailable evidence of his ability to organise his thoughts and communicate very effectively in written form. The viva defence showed that his oral communication is also exceptional. The presentation was beautifully crafted, with fine judgement about the audience and pace, with exemplary visual motifs and images. A first class set-piece performance that was clearly extremely thoughtfully and carefully prepared. His dialogue with examiners then showed he could think on his feet and make cogent responses without overstating his position. I asked about his initial goals and then probed his understanding of future research and application opportunities building on the work. For both questions he showed flexibility and sharp insights. He made clear that his actual research path had included adaptation and modification as it progressed. He also showed a sharp appreciation of some of the future issues that may be explored.

In summary, the opportunity to be an external examiner of this particular PhD has been a great privilege; not a claim I would make about many PhDs I have examined. It was an outstandingly good account of excellent research. Daniel not only showed his ability to do research well, for which he will be awarded his doctorate, but he also showed his enormous potential as a leading computing science researcher. His diverse skills and intellectual agility, his obvious capacity to work very effectively in diverse contexts on a wide range of topics, leads me to believe that he will achieve many significant successes during his career.

Evidence of developing impact

Since the *vive voce* examination reported above and the formal award of the doctorate, I have observed the anticipated impact beginning to be realised. I persuaded my colleague and Post-Doctoral Senior Research Associate, Dr Rosa Filgueira, to spend two months at ISI, USC, CA, where Dr Garijo now works. She came back overflowing with enthusiasm for the new methods and a bundle of publications and collaborative programs of research. Dr Garijo and his work, was a key reason for going, and he continued to contribute immense insights. As a result, Dr Filgueira, who is now a Senior Data Scientist at the British Geological Survey is making another two-month visit in September and October 2017.

A group of us, who have been engaged in data-intensive research since 2001, proposed a special edition of the Future Generation of Computing Systems to celebrate 10 years of the WORKS conference held in conjunction with Super Computing. We sought contributions from all leaders in the field of data-intensive and scientific workflows. I am delighted to say that Dr Garijo submitted an exceptionally strong paper reporting his thesis work, that sailed through the two-rounds of eliminations and requested revisions. The paper was highlighted in the editorial as providing evidence of substantial progress and a foundation for future methods of representing and reasoning about data-powered methods. The publication of this special edition is expected this summer. It is now in the hands of the printers (I am one of its four editors; the others are Drs Ian Tailor & Sandra Gesing, NDU, IN, Johan Montagnat, Universite Côte d'Azur, CNRS, FR).

I am currently developing plans for a research campaign improving our methods of pooling ideas, sharing methods and data as we address today's global and societal challenges. These require harnessing skills and intellectual effort from a wide diversity of disciplines and

viewpoints. They require representations of methods that last as technology changes. They require the federation of many autonomous organisations who must continue to address the agenda and priorities set by their funders and stakeholders while they contribute. This in turn requires new governance models and mechanisms for establishing rules governing data use and working practices across the distributed federation. The insights and methods Dr Garijo has pioneered will be vital here. I am delighted that he has agreed to help with this campaign.

Professor Malcolm Atkinson PhD, FRSE, FBCS, CITP, UKCRC

School of Informatics