USC
INFORMATION
SCIENCES
INSTITUTE

Information Sciences Institute
Viterbi School of Engineering
University of Southern California
4676 Admiralty Way
Marina del Rey CA, 90292

June 25, 2017

To whom it may concern,

I am writing to fully support Daniel Garijo's application for the 2017 SWSA Dissertation Award. I believe his thesis work is exceptional and deserves recognition.

The main goal that Daniel tackled was improving the reuse of scientific workflows. Scientific workflows capture complex multi-step computational experiments for data analysis as dataflow graphs. Workflow systems manage the distributed execution of workflows, validate the correctness of the computations, and facilitate reuse. Workflow reuse is an important topic, as reusing workflows improves the efficiency of scientists since they would not have to re-implement every software step of the workflow and instead would simply re-run the workflow. Workflow reuse is also important because it improves the quality of scientific data analysis by ensuring that the computations are executed as intended. Finally, workflow reuse would facilitate reproducibility by making it easy for others to re-run original experiments. Although workflows have become popular with scientists and workflow reuse holds great promise for all these reasons, it is unfortunately not common practice.

In order to understand the challenges of reuse, he analyzed workflows from many domains and from workflow systems, both from shared repositories such as myExperiment, and the workflows in the WINGS workflow system developed by my group. This analysis revealed that current workflows are hard to understand by scientists who did not develop them for several reasons: 1) the workflows are often described too close to the implementation of the software steps; 2) the workflows are generally poorly documented by authors, to others cannot easily understand how to reuse them; and 3) workflows are often only partially reused but it is hard to find which workflows contain the portions needed.

The key idea in Daniel's thesis is that capturing abstractions is key to workflow reuse. In the thesis, Daniel puts forward two kinds of abstractions. One kind of abstraction is the decomposition of a step into substeps. This is typically a conceptual abstraction that is in the mind of the workflow designers and is not typically captured in a workflow. A second type of abstraction concerns the function and criticality of the different computational steps of the workflow. For example, a reformatting step is not very critical to a computational experiment, it is only included to convert the data to the format needed by another step that is necessary for the experiment. Typically, scientists search for workflows to reuse based on these abstractions, so it is important to capture this kind of abstraction.

While traditional semantic web research is concerned with representing and sharing information about objects, very little work is devoted to representing and sharing procedures. This is a major differentiator in Daniel's thesis. His work has resulted in an open shared repository of hundreds of scientific workflows (and growing!), which can be searched by type of step, by combinations of steps, and by the characteristics of the data that they ingest or generate. Daniel's specific contributions to the semantic web are threefold:

1. Representations of workflow abstractions as labeled directed acyclic graphs (LDAGs). In LDAGs the labels indicate the type of abstraction or characteristics of data and computational steps. In addition, Daniel had to distinguish between high-level workflow descriptions, their specializations into executable workflows, and finally their executions. This required defining LDAGs that could capture these different kinds of workflows. He was one of the co-authors of the W3C PROV provenance standard, where he incorporated requirements for publishing workflow executions. In addition, he developed the OPMW representation that extends PROV to include workflow abstractions.
2. Workflow publication as complex web objects shared as linked open data. This enables scientists to search and reuse workflows. A workflow can be published individually, but it should be linked to related workflows, so that a scientist can search for all the workflow executions that correspond to a given abstract workflow. His approach is able to link these different kinds of workflows, to indicate that a single abstract workflow could be run with different datasets and algorithms and each of those in turn could be executed many times with different parameter values. For workflow publication, he

extended an existing approach to ontology publication. Today, hundreds of workflows have been published with this approach and are openly available to the community. Daniel showed that different workflow systems and tools could exchange information about workflows through this publication, demonstrating how semantic web technologies can be used in mediators of procedural knowledge.

3. Mining of labeled directed acyclic graphs that represent complex workflows and their abstractions. Daniel developed in his thesis a novel approach to mine those graphs in order to automatically learn workflow abstractions. He developed the FragFlow workflow mining system to extract workflows and common workflow fragments from a corpus of workflow examples. FragFlow uses off-the-shelf process mining techniques including both inexact and exact graph matching algorithms. Inexact techniques use heuristic measures, such as minimum description length (MDL) and size reduction, to calculate the similarity between graphs in an efficient but not complete manner. Exact techniques return all the possible frequent subgraphs included in a dataset, and we use deep first search and breadth first search strategies. FragFlow can be configured with different algorithms and settings, varying the frequency of the mined workflow fragments, their size, and the type of graph mining algorithm. Therefore, FragFlow can be easily customized to particular users or contexts, and can be easily extended with new graph mining algorithms.

Daniel did an impressive evaluation based on hundreds of workflows created by dozens of different users of the LONI Pipeline workflow system for neuroimaging genomics. To do this work, he had to create a mapping of the LONI Pipeline workflows into his representation, run his workflow mining algorithms with them, and evaluate with users whether the abstractions found were useful. He found that the majority of the abstract fragments extracted would be selected by users for new problems either directly or with minor changes. It is very unusual to find evaluations of this magnitude and practical implications.

Daniel was unequivocally the major force and driver of the work in this thesis, but the issues involved were significant and therefore he involved others in collaborations along the way. This includes both workflow researchers as well as scientists who are users of workflows. The many papers that have resulted from this thesis are co-authored with others, but Daniel was the main force behind all of them. This is demonstrated by the fact that he is first author in all these papers. The co-authors are different groups in different sets of papers. Daniel is a great collaborator in developing new semantic web technologies, and an avid ambassador to encourage their adoption in science.

In summary, Daniel's thesis has major qualities that set it above others in the area. First, it addresses the representation and publication of procedures, rather than descriptive information about objects, and contributed to the development of the W3C PROV standard. Second, it focuses on learning to extract abstractions from published linked data in order to enable users to find and reuse relevant pieces. Third, the work was evaluated with a very large collection of actual scientific workflows created by scientists, and through interviews with scientists who found his approach useful for workflow reuse.

Please let me know if I can help with any further questions.

Sincerely,

Yolanda Gil
Research Professor of Computer Science
Associate Director for Research, Intelligent Systems
Associate Director of Informatics for Joint Degrees
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 (USA)
Phone: +1-310-448-8794
Email: gil@isi.edu
http://www.isi.edu/~gil