Work at ISI 2014 Collaboration with LONI Group

Daniel Garijo Verdejo

Supervisors: Oscar Corcho, Yolanda Gil

Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

Date: 25/06/2014

Outline

Index

1. Background
2. What do I do?
3. Motivation
4. Work done in the internship
    1. Detecting common fragments in LONI pipeline
    2. Survey on workflow reuse

About me

- 3rd year PhD student at Universidad Politécnica de Madrid
  - Supervisors: Oscar Corcho, Yolanda Gil.
- Knowledge representation and Semantic Web.
  - Ontologies, Linked Data, RDF, etc.
- eScience, reproducibility, reuse.
- Third time at ISI with Yolanda

What am I working on?

- Workflow representation
    - Plan/template representation
    - Provenance trace representation
    - Link between templates and traces

CH1: Can we export an abstract template of the method being represented?
CH2: How do we interoperate with other workflow results?
CH3: How do we access the workflow results?
CH4: How do we link an abstract method with several implementations?

- Creation of abstractions/motifs in scientific workflows
    - Abstraction catalog
    - Find how different workflows are related

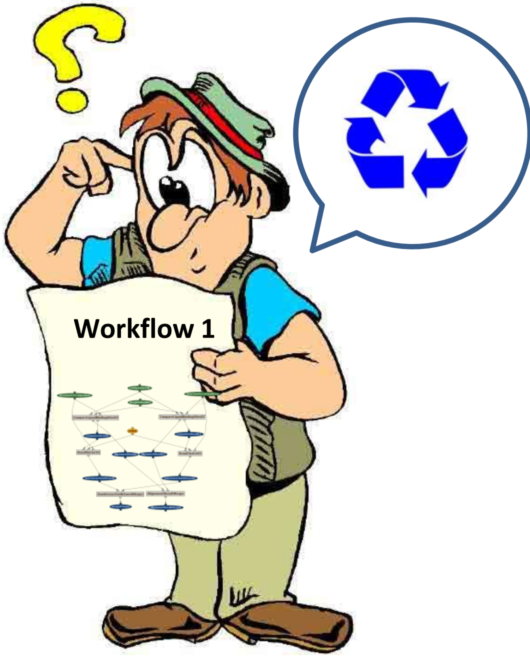
CH5: How can we detect what are the typical operations in scientific workflows?
CH6: How can we detect them automatically?

- Understandability and reuse of scientific workflows
    - Relation between the workflows involved in the same experiment (Research Objects)

CH7: Which workflow parts are related to other workflows?
CH8: How do workflows depend on the other parts of the experiments?

Motivation

- As a designer: Discovery
  - Workflows with similar functionality fragments/methods
  - Design based in previous templates.
- As user/reuser: Understandability, Exploration
  - Search workflows by functionality
  - Commonalities between execution runs
  - Component categorization
  - Workflow summarization



Work done in the internship

1. Workflow fragment detection in the LONI Pipeline
   - Integration of previous work with the LONI Pipeline.
   - Evaluation against 3 different corpora
   - User-based preliminary evaluation

2. Study on workflow reuse
   - Series of interviews with LONI pipeline users
   - Survey on workflow reuse
   - Discussion of the results

Workflow fragment detection in the LONI Pipeline

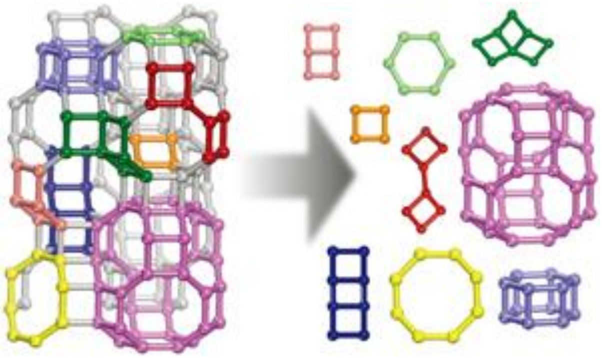
Fragment detection The problem

Problem statement:

*Given a **repository of workflows, what are the workflow fragments I can deduce from it?***

Useful for:

- Finding relationships between workflows and sub-workflows.
    - Most used fragments, most executed, etc.
    - Workflow cloud of reused fragments.
- Proposing new templates with the popular fragments.
- Summarizing existing workflows using popular fragments.



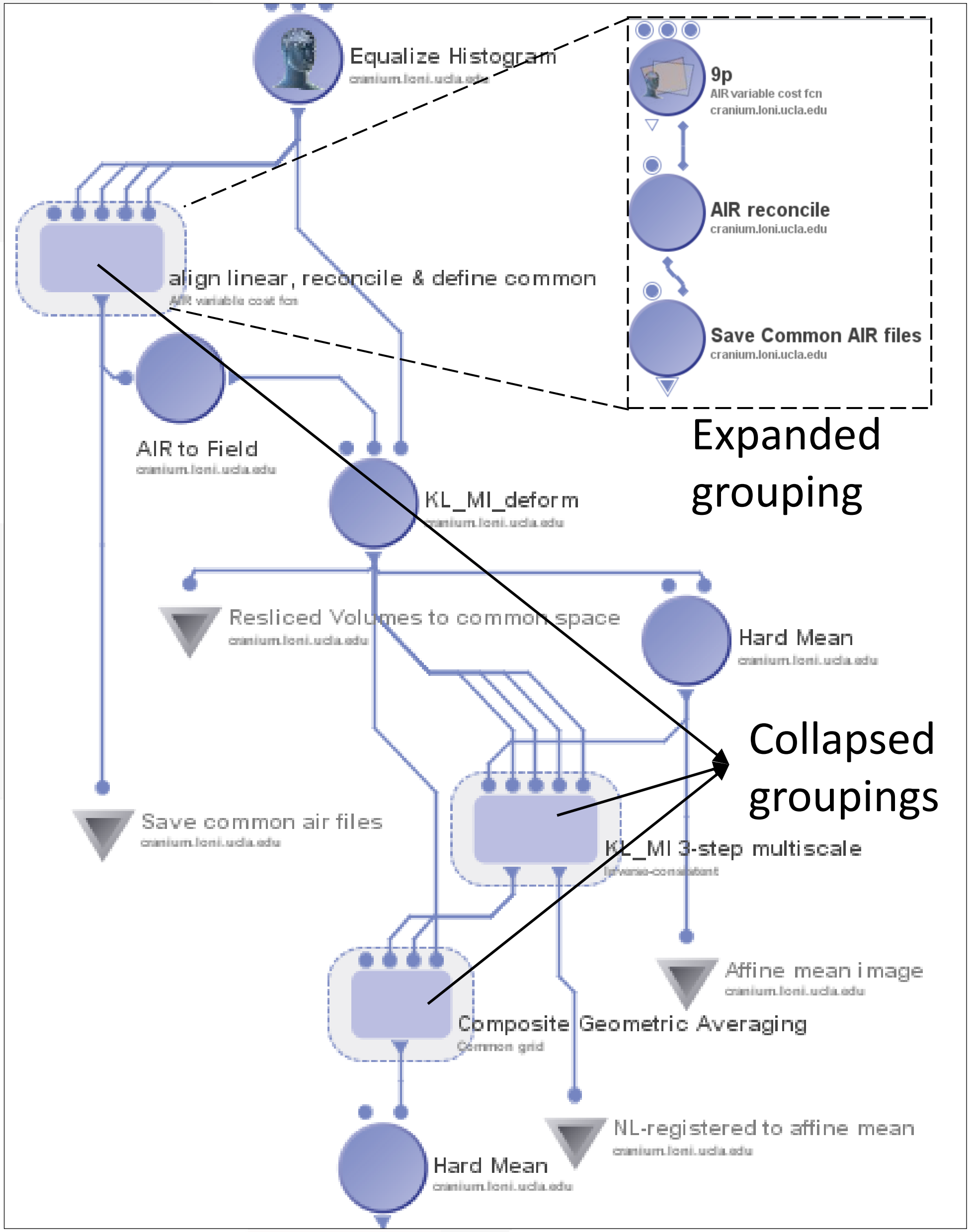
Challenges: Common workflow fragment detection

- Given a collection of workflows, which are the most common fragments?
  - Common sub-graphs among the collection
    - Sub-graph isomorphism (NP-complete)

- We use 2 different algorithms to find the fragments
  - The SUBDUE algorithm **[Holder et al 1994]** (hierachical clustering
    - Inexact Frequent Graph Matching (some fragment candidates might be not returned)
    - Graph based hierarchical clustering
      - Each cluster corresponds to a workflow fragment
    - Iterative algorithm with two measures for compressing the graph:
      - Minimum Description Length (MDL)
      - Size
  - gSpan (http://www.cs.ucsb.edu/~xyan/software/gSpan.htm)
    - Exact Frequent Graph Matching (all possible fragments are returned)
  - More algorithms to come!

[Holder et al 1994]: **Substructure Discovery in the SUBDUE System** L. B. Holder, D. J. Cook, and S. Djoko. AAAI Workshop on Knowledge Discovery, pages 169-180, 1994.

The LONI pipeline



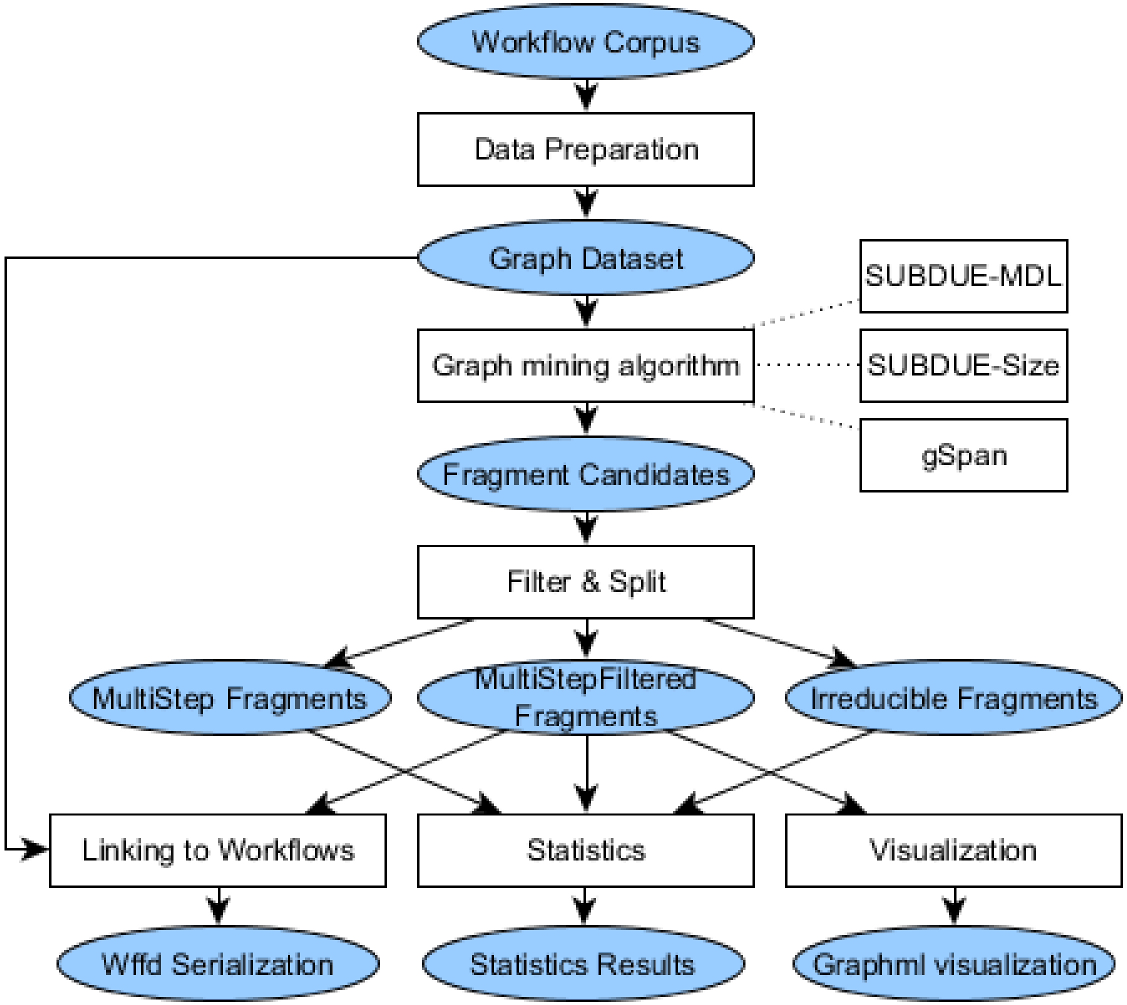
Expanded grouping

Collapsed groupings

LONI Pipeline:

- Workflow system for neuroimaging analysis
- Many workflows recorded and published
- Workflows are likely to reuse parts of other workflows (common library of components)
- Users can group sets of steps to organize, simplify and reuse workflows (groupings).

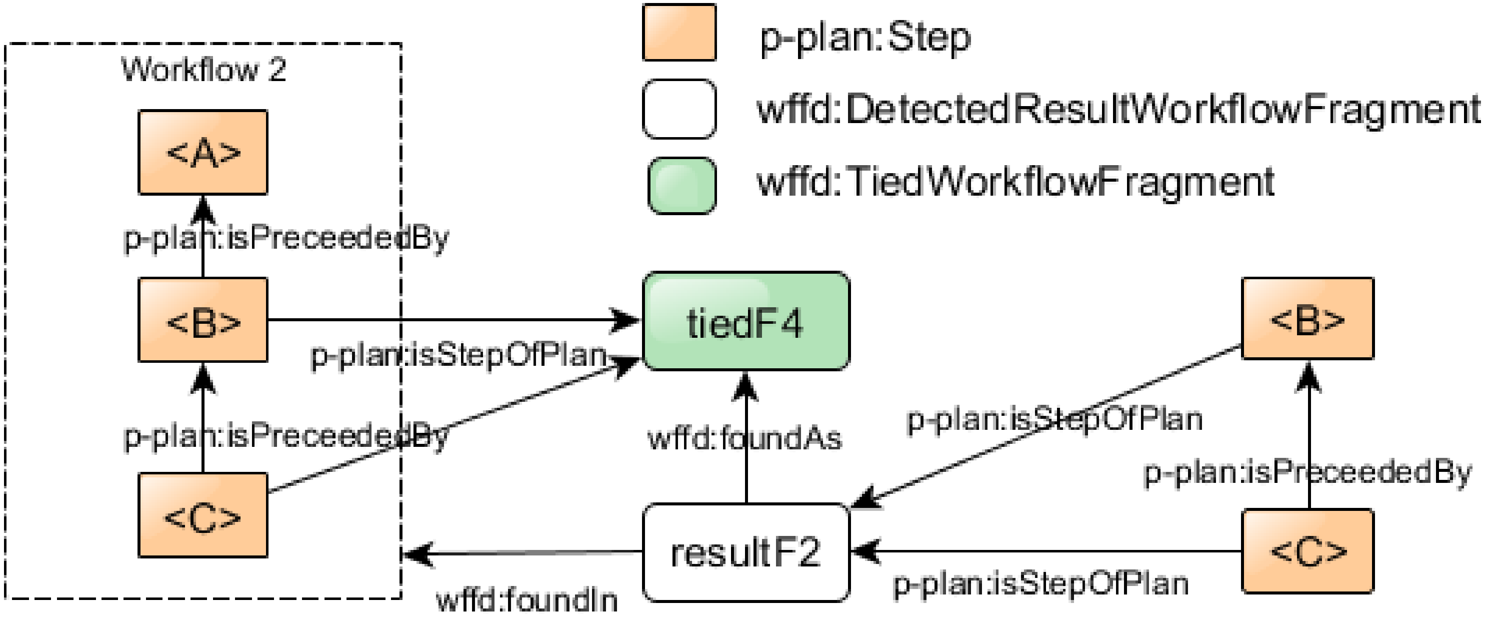
Workflow for fragment detection



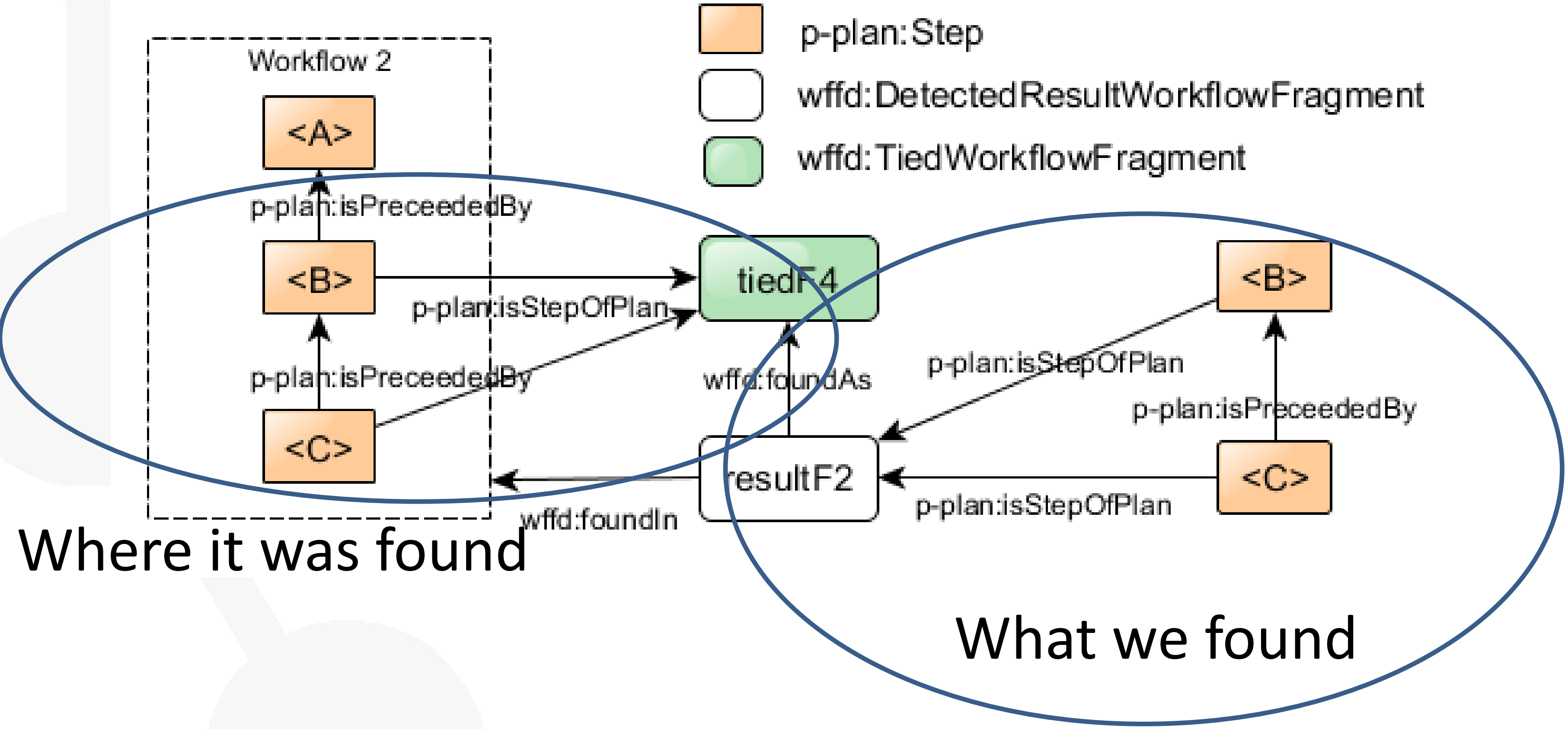
The data has to be prepared, and the results filtered.

1. A data preparation step filters duplicate workflows and remove single step workflows
2. The graph mining algorithm calculates fragment candidates
3. Fragment candidates are filtered to simplify the final results.
4. Results are linked back to workflows and visualized.

Linking workflows: an example



Linking workflows: an example



Where it was found

What we found

Fragment detection: Corpora

Evaluation:

- 3 different corpora:
  - WC1: single user corpus, with 475 different unique workflows
  - WC2: single user corpus (plus collaborations with others), with 96 different unique workflows
  - WC3: multi user corpus, with 357 different workflows from 62 users. Submited to the LONI Pipeline during January, 2014

Fragment detection : Evaluation

We measure our results by comparing them against the groupings defined by users

Metrics defined:

$$P(\text{Exact}) = \frac{|\text{FragFlow Frag} \cap (\text{LONI Gr.} \cup \text{LONI wfs})|}{|\text{FragFlow Frag}|}$$

$$R(\text{Exact}) = \frac{|\text{FragFlow Frag} \cap (\text{LONI Gr.} \cup \text{LONI wfs})|}{|(\text{LONI Gr.} \cup \text{LONI wfs})|}$$

We also relaxed the previous metrics to measure if there was an overlap from the common detected fragments and the user defined groupings.

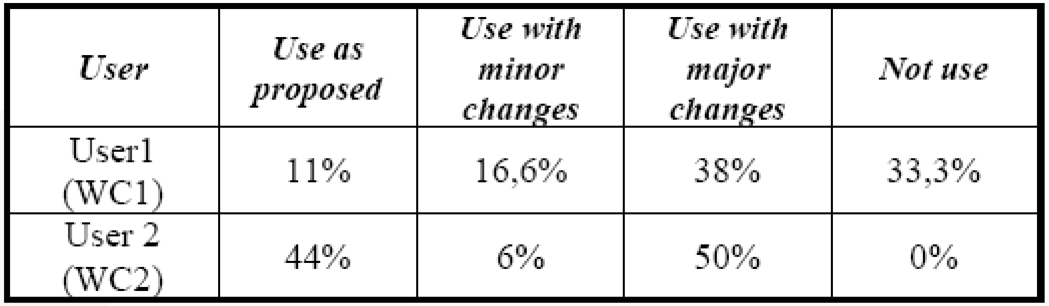
Fragment detection : Results

Results

•30% to 75% of the total fragments found correspond directly to user defined groupings in the single user corpora.

•In the multi user corpus, the best results are 50% to 56% with minimum frequency. If we consider the overlap of 80% of the steps, the precision is 40% to 80%.

•Users find our proposed fragments as useful candidates for groupings, and therefore useful for reuse in their workflows



User	*Use as proposed*	*Use with minor changes*	*Use with major changes*	*Not use*
User1 (WC1)	11%	16,6%	38%	33,3%
User 2 (WC2)	44%	6%	50%	0%

Survey on workflow and grouping reuse

Interview with users

- What are the main benefits of using workflows and groupings in the LONI Pipeline?
  - Sharing workflows with collaborators
  - Time savings
  - Teaching
  - Visualization
  - Design for modularity
  - Design for understandability
  - Design for standarization
  - Design for debugging
  - Paper writing
  - Reproducibility and inspectability

Types of users

Interview with users

- Types of users in the LONI Pipeline
  - Developers: usually write components, develop programs and create workflows. Usually bioinformaticians and engineers.
  - Beginner programmers: can write small scripts and program spreadsheets for statistical analysis. They mainly reuse workflows from others. Typically, they are neuroscientists.
  - Non-programmers: cannot write code. They reuse workflows from others. Typically students.

Results

Survey with 30 questions

- 25 responses recorded (from the LONI pipeline group at USC).

Main findings:

- Writing code is considered very important for this area of research. Sharing code is not considered as important.

- The majority of responders found the workflow system useful.

- Creating workflows is very useful, but the reuse of workflows was not seen as useful.

- Workflows are useful for both nonprogrammers and for teaching new students.

- Reusing groupings from one's own work is more useful than reusing groupings from others. Groupings help simplify workflows. Groupings save time and also make workflows more understandable by others.

- Workflows are not systematically linked to publications. Most responders believe that the link between a workflow and a publication is kept in private laboratory notes

1. Workflow fragment detection in the LONI Pipeline
   - Integration of previous work with the LONI Pipeline.
   - Evaluation against 3 different corpora
   - User-based preliminary evaluation

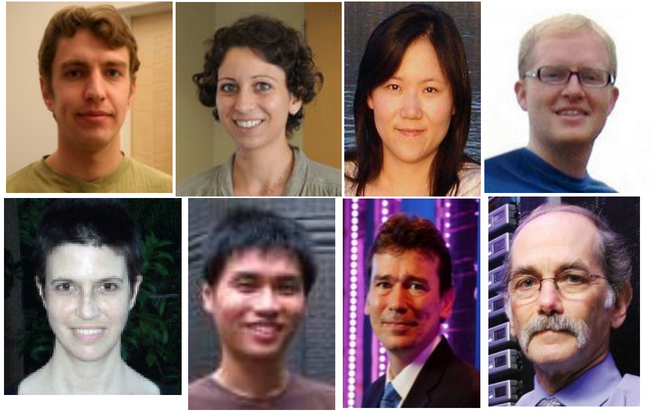
2. Study on workflow reuse
   - Series of interviews with LONI pipeline users
   - Survey on workflow reuse
   - Discussion of the results

Involved People

- Ontology Engineering Group (UPM)
  - Daniel Garijo, Oscar Corcho

- Information Sciences Institute (USC)
  - Yolanda Gil, Varun Ratnakar

- Laboratory of Neuro Imaging
  - Boris A. Gutman, Neda Jahanshad, Xue Hua, Derrek Hibar, Meredith Braskie, Zhizhong Liu, Paul Thompson and Arthur W. Toga



- University of Michigan School of Nursing
  - Ivo Dinov

Work at ISI 2014 Collaboration with LONI Group

Daniel Garijo Verdejo

Supervisors: Oscar Corcho, Yolanda Gil

Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

Date: 25/06/2014