Common Motifs in Scientific Workflows: An Empirical Analysis

Daniel Garijo *, Pinar Alper †, Khalid Belhajjame †, Oscar Corcho *, Yolanda Gil ‡, Carole Goble †

* Universidad Politécnica de Madrid,
† University of Manchester,
‡ USC Information Sciences Institute

IEEE eScience 2012. Chicago, USA
• Empirical analysis on 177 workflow templates from Taverna and Wings

• Catalog of recurring patterns: scientific workflow motifs.
  
  • Data Oriented Motifs
  
  • Workflow Oriented Motifs

• Understandability and reuse
• Workflows as software artifacts that capture the scientific method
  • Addition to paper publication
  • Reuse

• Existing repositories of workflows (myExperiment)
  • Sharing workflows
  • Exploring existing workflows.

• **PROBLEMS to address:**
  • Sometimes workflows are difficult to understand
  • Workflow descriptions depend on tools/files
  • Decay of workflows
  • Identify good practices for workflow design

[myexperiment](http://www.myexperiment.org)
• Reverse-engineer the set of current practices in workflow development through an analysis of empirical evidence

• Identify workflow abstractions that would facilitate understandability and therefore effective re-use
Taverna and Wings

http://www.taverna.org.uk/

http://www.wings-workflows.org/
• Workflow motif: Domain independent conceptual abstraction on the workflow steps.

1. Data-oriented motifs: **What** kind of manipulations does the workflow have?
   - E.g.: 
     • Data retrieval
     • Data preparation
     • etc.

2. Workflow-oriented motifs: **How** does the workflow perform its operations?
   - E.g.: 
     • Stateful steps
     • Stateless steps
     • Human interactions
     • etc.
Data-Oriented Motifs

Data Retrieval

Data Preparation

Format Transformatic

Input Augmentation and Output Splitting

Data Organisation

Data Analysis

Data Curation/Cleaning

Data Moving

Data Visualisation
Data-Oriented Motifs

Data Retrieval

Data Preparation

Format Transformation

Input Augmentation and Output Splitting

Data Organisation

Data Analysis

Data Curation/Cleaning

Data Moving

Data Visualisation
Data Oriented Motifs

Data Retrieval

Data Preparation

Format Transformation

Input Augmentation and Output Splitting

Data Organisation

Data Analysis

Data Curation/Cleaning

Data Moving

Data Visualisation

IEEE eScience 2012. Chicago, USA
Data-Oriented Motifs

Data Retrieval

Data Preparation

Format Transformation

Input Augmentation and Output Splitting

Data Organisation

Data Analysis

Data Curation/Cleaning

Data Moving

Data Visualisation

IEEE eScience 2012. Chicago, USA
Data-Oriented Motifs

- Data Retrieval
- Data Preparation
- Format Transformation
- Input Augmentation and Output Splitting
- Data Organisation
- Data Analysis
- Data Curation/Cleaning
- Data Moving
- Data Visualisation
Data-Oriented Motifs

Data Retrieval

Data Preparation

Format Transformation

Input Augmentation and Output Splitting

Data Organisation

Data Analysis

Data Curation/Cleaning

Data Moving

Data Visualisation

IEEE eScience 2012. Chicago, USA
Workflow-Oriented Motifs

Intra-Workflow Motifs
- Stateful (Asynchronous) Invocations
- Stateless (Synchronous) Invocations
- Internal Macros
- Human Interactions

Inter-Workflow Motifs
- Atomic Workflows
- Composite Workflows
- Workflow Overloading
Workflow-Oriented Motifs

Intra-Workflow Motifs

Stateful (Asynchronous) Invocations

Stateless (Synchronous) Invocations

Internal Macros

Human Interactions

Inter-Workflow Motifs

Atomic Workflows

Composite Workflows

Workflow Overloading
Workflow-Oriented Motifs

Intra-Workflow Motifs

Stateful (Asynchronous) Invocations
Stateless (Synchronous) Invocations
Internal Macros
Human Interactions

Inter-Workflow Motifs

Atomic Workflows
Composite Workflows
Workflow Overloading
Workflow-Oriented Motifs

Intra-Workflow Motifs

Stateful (Asynchronous) Invocations
Stateless (Synchronous) Invocations
Internal Macros
Human Interactions

Inter-Workflow Motifs

Atomic Workflows
Composite Workflows
Workflow Overloading
Workflow-Oriented Motifs

Intra-Workflow Motifs
- Stateful (Asynchronous) Invocations
- Stateless (Synchronous) Invocations
- Internal Macros
- Human Interactions

Inter-Workflow Motifs
- Atomic Workflows
- Composite Workflows
- Workflow Overloading
• **177 Workflow templates**

- 111 from Taverna, sample from myExperiment
- 66 from Wings, available in public server (now as Linked Data)
- Diverse domains

![Chart showing workflow templates by domain](chart.png)
Over 60% of the motifs are **data preparation** motifs.
- Of the 4 subcategories, the most common across domains are **output splitting, input augmentation, and reformatting** steps.

- **Data retrieval** common in domains where **curated databases** exist

- **Data analysis** is often the main functionality of the workflow
Result Summary: Workflow Oriented Motifs

- Around **40% composite workflows and internal macros**
  - Workflow reuse is present even in some **atomic workflows**

- Human interactions steps increasingly used in some domains
• Data moving/retrieval, stateful interactions and human interaction steps are not present in Wings
  • Web services (Taverna) versus software components (Wings)
  • Wings has layered execution through Pegasus

• Data preparation steps are common in both systems

• Use of sub workflows is high
Our observations:

- **Obfuscation** of scientific workflows
  - The abundance of data preparation steps make the functionality of the workflow unclear.

- **Decay of scientific workflows**
  - Create an abstract description.

- **Good practices for workflow design**
  - Sub-workflows ✓

- **Workflow overloading** ✓/✗

http://www.sandensconsulting.com/images/DataObfuscation.jpg
• **Empirical analysis** of scientific workflows
  177 workflows
  • 2 different systems
  • A variety of heterogeneous domains

• **Workflow motif catalog**
  • Data oriented motifs
  • Workflow oriented motifs

• **Future work**: automatic abstractions on workflows
  ➔ Template analysis
  ➔ Trace analysis (**provenance**)  
  ➔ Include other workflow systems
Who are we?

• Pinar Alper
  School of Computer Science, University of Manchester

• Khalid Belhajjame
  School of Computer Science, University of Manchester

• Oscar Corcho
  Ontology Engineering Group, UPM

• Yolanda Gil
  Information Sciences Institute, USC

• Carole Goble
  School of Computer Science, University of Manchester

EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1).
(http://www.wf4ever-project.org)
Common Motifs in Scientific Workflows: An Empirical Analysis

Daniel Garijo *, Pinar Alper †, Khalid Belhajjame †, Oscar Corcho *, Yolanda Gil †, Carole Goble †

* Universidad Politécnica de Madrid,
† University of Manchester,
‡ USC Information Sciences Institute

IEEE eScience 2012. Chicago, USA