SCIENTIFIC WORKFLOW

METHODS AND CONCEPTS

CONTENTS

Definitions

What is scientific workflow system?

A scientific workflow system is a specialized form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or workflow, in a scientific application [1].

What is workflow?

Workflows are defined as a set of interrelated computational and data handling tasks designed to achieve a specific goal [1].

APPLICATIONS

Some projects from various scientific domains that are dealing with large-scale distributed data[1]:

- Optical Astronomy
- Radio astronomy
- Seismology
- Experimental Biology
- Environmental Science

These projects, like many others, involve the challenges of data creation, exploration, exploitation, and preservation in many scientific communities. The rapidly growing and diverse data opens many new opportunities in business, research, design, policy formulation, and decision making, but these opportunities can only be exploited if we improve our knowledge discovery apparatus as we enter the data-intensive era [1].

Related topics to workflow

- **Workflow Characteristics.**
- **Workflow Architectures.**
- Notable systems
- Sharing workflows
- Analysis

WORKFLOW CHARACTERISTICS

Workflow Characteristics:

- —These phases are from the scientists' perspective as they create and run workflows.
- —Scientists compose, operate, analyze, and refine workflows.
- —Scientific workflows are exploratory; that is, it is common to reuse workflows and refine them using trial and error.
- —Scientific methods are often repeated; that is, scientists rerun workflows with different parameters and datasets.
- —Runtime monitoring and diagnostics are important; that is, scientists monitor progress and may steer or decide to abort or suspend an execution.

WORKFLOW ARCHITECTURES

Workflow Management System (See [1] and references there in):

- Pegasus [Deelman et al. 2015],
- Kepler [Lud¨ascher et al. 2006], Taverna [Wolstencroft et al. 2013], Triana [Taylor et al. 2007a],
- Swift [Zhao et al. 2007], Trident [Barga et al. 2008], Galaxy [Blankenberg et al. 2010],
- ASKALON [Fahringer et al. 2007], WS-PGRADE/gUSE [Kacsuk et al. 2012],
- Meandre [Llor`a et al. 2008], and Apache Airavata [Marru et al. 2011].

WORKFLOW ARCHITECTURES

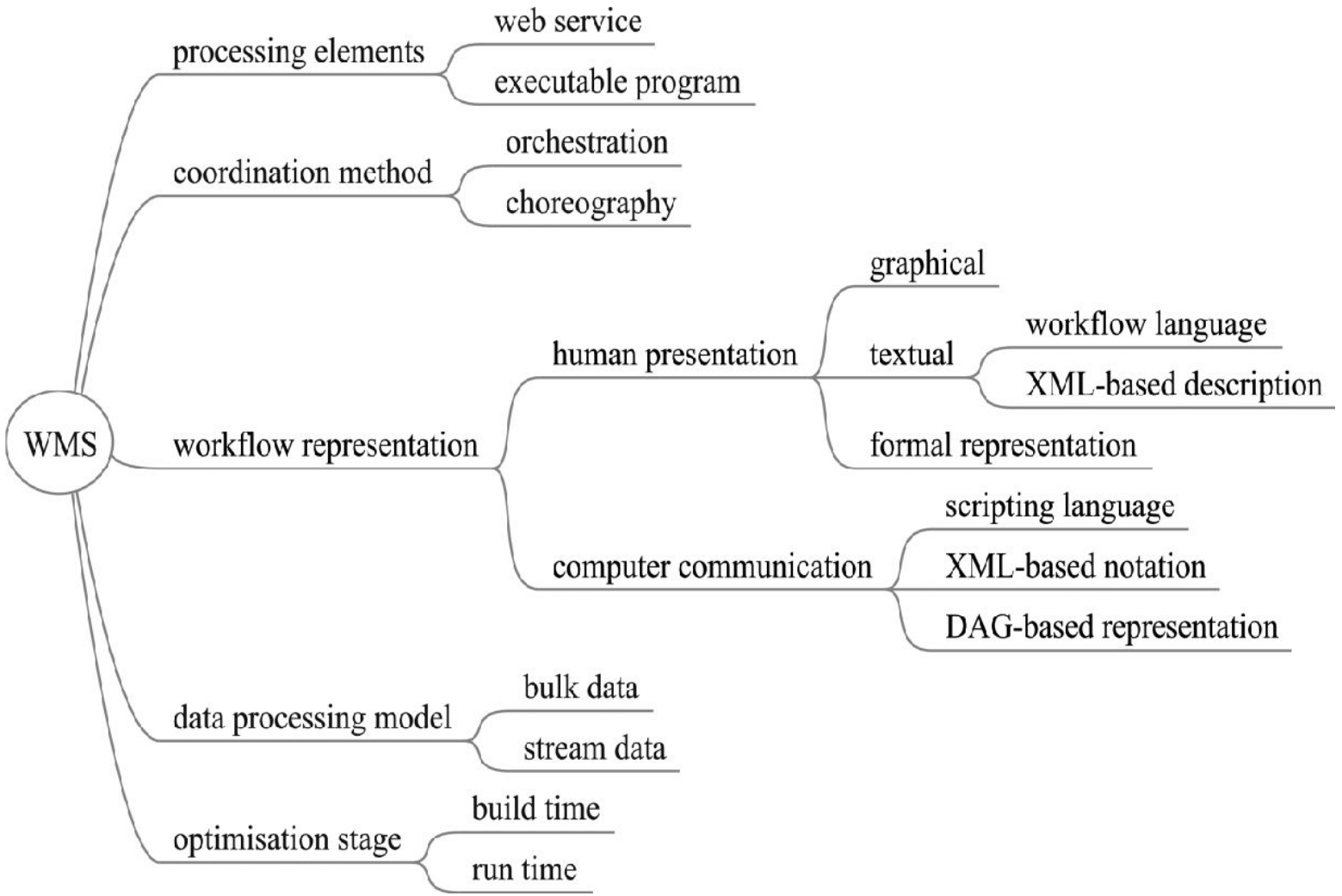

Fig. 1. Architectural characterizations of WMSs.

NOTABLE SYSTEMS

Workflow composition tool (colored in green), the resource mapping mechanism (colored in orange), and the workflow execution engine (colored in yellow).

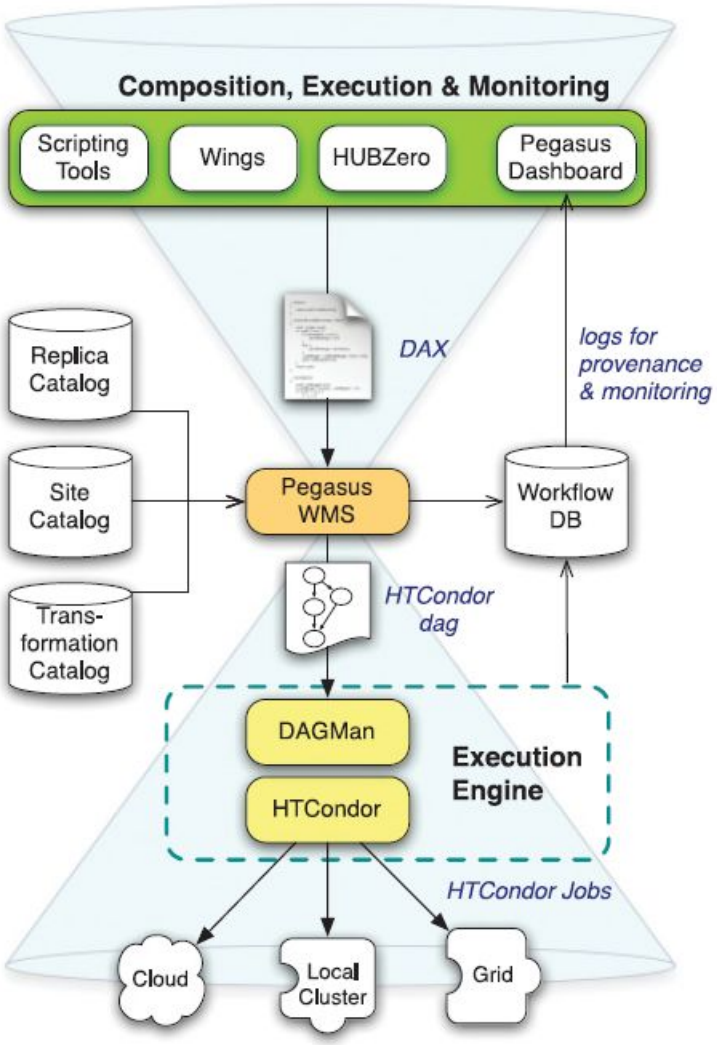

Fig. 1. Pegasus architectural diagram.

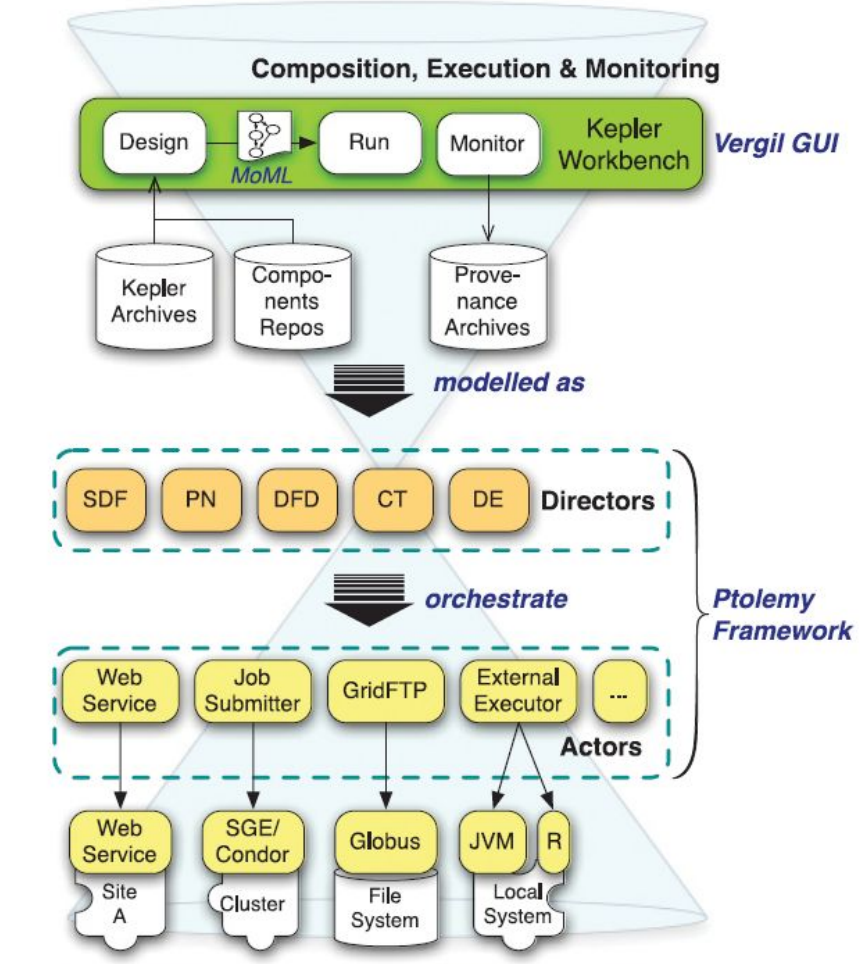

Fig. 2. Kepler architectural diagram.

NOTABLE SYSTEMS

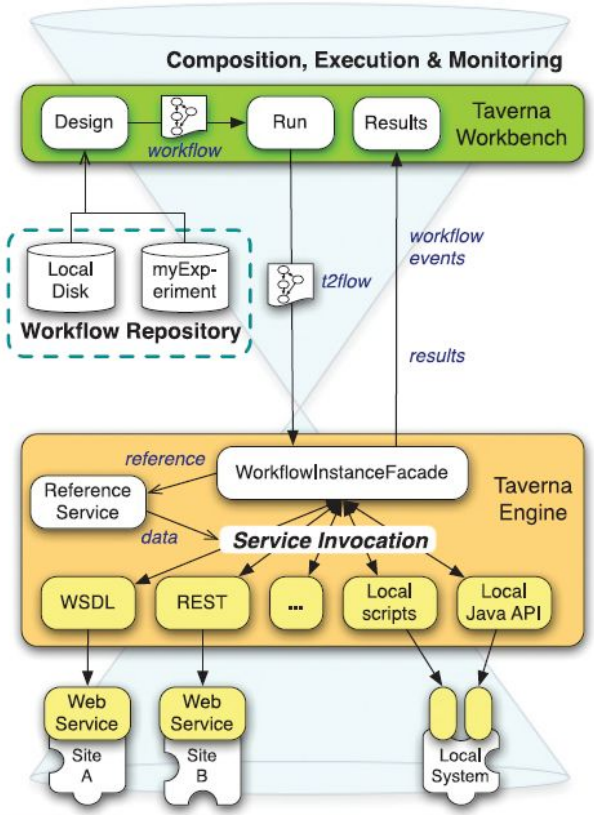

Fig. 3. Taverna architectural diagram.

NOTABLE SYSTEMS

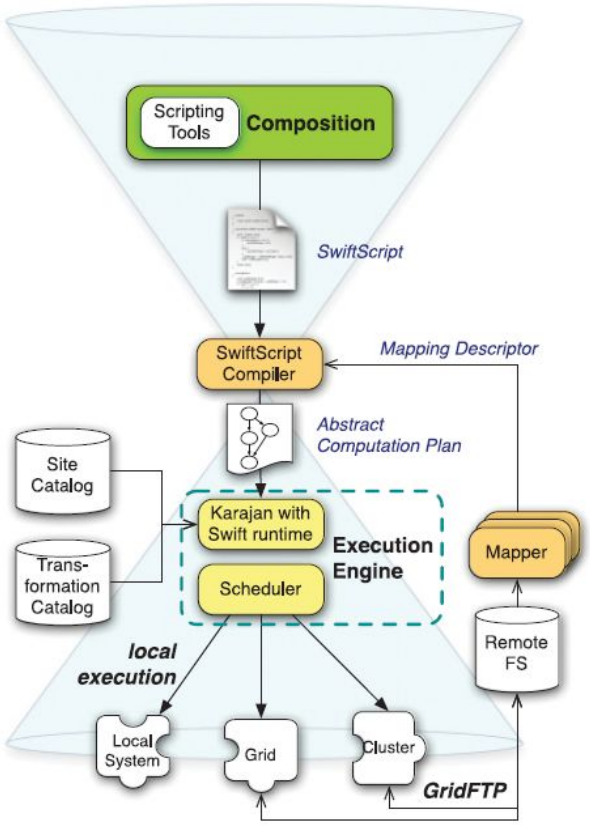

Fig. 3. Swift architectural diagram.

NOTABLE SYSTEMS

NOTABLE SYSTEMS

Table I. Characterizing the Workflow Management Systems

	Pegasus	Kepler	Taverna	Swift	KNIME	Airavata	Meandre
processing element	executable program	executable program & web service	executable program & web service	executable program	executable program	executable program & web service	executable programs
system architecture	orchestrate	orchestrate	orchestrate	orchestrate	orchestrate	orchestrate	orchestrate
optimization stage	build time	none	none	runtime	runtime	none	runtime
user interface	textual	graphical	both	textual	graphical	both	both
data processing model	bulk data	bulk data & stream data	bulk data & stream data	bulk data & stream data	bulk data	bulk data & stream data	stream data

SHARING WORKFLOWS

- The quantity and diversity of data are growing rapidly because the capacity of storage is increasing [Walter 2005], digital communication is pervasive and increases in capacity [Zhao et al. 2011], and the sensitivity, speed, diversity, and deployed numbers of digital data collection devices exhibit a compound growth. [1]
- This is combined with a growing drive to share data Interagency Working Group on Digital Data 2009; EU Parliament 2007], enabled by many organizations' standardization efforts (e.g., W3C, OGC, FDSN68 IVOA, and RDA) and a growing need to combine data across discipline boundaries to address today's societal challenges. [1]

ANALYSIS

The benefits include

- (1) increased productivity and lower error rates as tedious chores are automated,
- (2) improved scientific methods as many different specialists pool advances to their parts of a method, and
- (3) achievement of new goals by combining computational power with the increased wealth of data.

ANALYSIS

We review two questions [1]:

- (1) Why are scientific workflow systems unable to support this increased use?
- (2) What research will be needed to make them ready?

Each community develops its own culture—a body of knowledge, established methods, practices, and ethics—shaped for its own research goals and professional practices.

WORKS CITED

1. Chee Sun Liew, Malcolm P. Atkinson, Michelle Galea, Tan Fong Ang, Paul Martin, and Jano I. Van Hemert. 2016. Scientific workflows: Moving across paradigms. ACM Comput. Surv. 49, 4, Article 66 (December 2016), 39 pages. DOI: http://dx.doi.org/10.1145/3012429

2. mmmm