Scufl2 – because a workflow is more than its definition

Document Version
Other version

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Stian Soiland-Reyes
myGrid, School of Computer Science
University of Manchester, UK

SCUFL2: BECAUSE A WORKFLOW IS MORE THAN ITS DEFINITION
Taverna workflows

- A set of (local and remote) services to analyze or manage data
- Nested workflows are also services
- Data-links connects services
  - i.e. output from service A is input to service B and C
  - Describes the desired dataflow instead of process coordination
- Automatic iterations
- Can customize list handling and control links
Taverna workflow features

- Nested workflows
  - Reuse existing components
- Implicit iterations
  - With customizable list handling
- Pipelining
  - Process partial iteration results early
- Parallelisation
  - Run as soon as data is available
- Retries, failover, looping
  - For stability and conditional testing
- Plugin-extensible execution control
  - Ideas: caching, error detection, dynamic service lookup
Extensible workflow engine

- New service types/protocols
- Execution control like looping/branching/start/stop
- Service discovery at runtime
Taverna’s existing wf formats

- **Taverna 1: SCUFL 2004-2007**
  - Lightweight XML format
  - Extensions in arbitrary formats ("*Put your XML element here*")
  -Parsed and written by many 3rd parties

- **Taverna 2: t2flow 2007-2011**
  - Supports control and service extensions of T2
  - Annotation system
  - ‘Heavy’ XML serialisation of Java beans
  - Tricky to use by 3rd parties
SCUFL2 motivations

- Easy to use for **third-parties**
- **Semantic web**/linked open data compatible
  \[ \rightarrow \text{identifiers and ontology} \]
- **Extensible** in a controlled manner
  - Plugins should provide a minimal ‘schema’
  - But parsers should not need to know about plugin
- Bundle relevant **resources**
  - Input/output data, provenance, scripts, binaries
- **Annotations** on anything
- **Reproducible** workflow results
Research objects

- **Reusable** – used as part of new study;
- **Repurposeable** – reuse the pieces in a new (and different) study. Substitute alternative data sets, methods;
- **Repeatable** – repeat the study, possibly years later;
- **Reproducible** – a special case of repeatability with a complete set of information/results to work towards;
- **Replayable** – go back and see what happened;
- **Referenceable** – cite in publications;
- **Revealable** – provenance and audit;
- **Re-interpretable** – crossing boundaries;
- **Respectful** – appropriate credit and attribution;
- **Retrievable** – discover and acquire.
Research Object vision

- ROs: **Aggregations** to support **sharing/publication**
- Incorporating *methods, data, people*
- Research Objects will allow us to conduct research in ways that are
  - **Efficient**: cheaper to borrow than recreate
  - **Effective**: larger scale through reuse
  - **Ethical**: Benefiting wider communities, not just individuals
- **Could I have a copy of your Research Object please?**
WfEver project

- Architecture/implementation for workflow preservation, sharing and reuse
- Research Object models
- Workflow Decay, Integrity and Authenticity
- Workflow Evolution and Recommendation
- Provenance
- Driven by Use Cases: Biology & Astronomy

...technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows in a range of disciplines.
What is SCUFL2?

- File **format** for *(Taverna)* workflows and data
- **Specification** *(wiki)*
- **Ontology** *(OWL)*
- **Schema** *(XSD)*
- **API** *(Java +Ruby?)*
- **Conversion Tool** *(Java)*
SCUFL2 bundles

ZIP archive

- mimetype
- META-INF/manifest.xml
- META-INF/container.xml
- workflowBundle.rdf
- workflow/HelloWorld.rdf
- workflow/someNestedWf.rdf
- profile/tavernaServer.rdf
- profile/tavernaCloud.rdf
- resources/script.pl
- inputs.rdf
- annotations/myExperiment.rdf
- inputs/data.txt
- outputs.rdf
- outputs/result.txt
- run/provenance.rdf
- http://example.com/largeData
- http://example.net/tavernaServer
- http://example.org/binary
- http://myexperiment.org/workflow/1337
- http://example.net/tavernaServer
Bundle manifest

the-workflow-bundle.scufl2

"Structured" zip-file (can be unpacked to be exposed on the web)

Root file: Primary document of the bundle (ePub OCF/Adobe UCF).

Alternative representations of same workflow bundle allowed (Turtle, JSON, HTML, etc)

Self-documenting media type (OpenOffice ODF, ePub OCF, Adobe UCF) – for tools like file(1) and mime magic

META-INF/container.xml

<container version="1.0"
xmlns="urn:oasis:names:tc:opendocument:xmlns:xmlns:container">
<rootfiles>
  <rootfile full-path="workflowBundle.rdf" media-type="application/rdf+xml" />
  <!-- <rootfile full-path="workflowBundle.ttl" media-type="text/turtle" /> Alternative repr. -->
</rootfiles>
</container>
Workflow bundle

**Unique identifier** which can be used as *prefix* for all relative references in bundle.

Suggests **main workflow** and **main profile**, but executor could (e.g. by parameter) run a different workflow or profile.

**Additional (non-executable) annotations and metadata**
Technologies and standards

- OpenOffice ODF *manifest*
- ePub OCF/Adobe Universal Container Format (UCF) *container rootfile, mimetype*
- RDF/XML and OWL *structure and annotations*
- XML Schema for ‘structured’ RDF/XML
- Open *Annotation Collaboration* (OAC)?
- ORI-ORE? *Aggregation*
- Open *Provenance Model* (OPM)?
Want people to try it out
Open for suggestions and feedback
Beta to be released in Autumn 2011
  - Native format for upcoming Taverna 3 in 2012
  - Load/Save plugin for Taverna 2.3 2011 Q4
Future plans
  - Explore further research object & provenance
  - Converting to/from wf systems like Galaxy?
Acknowledgements

Funded by: EPSRC EP/G026238/1; European Commission’s 7th FWP FP7-ICT-2007-6 270192; FP7-ICT-2007-6 270137
More information

- http://www.taverna.org.uk/
- http://www wf4ever-project.org/
- http://www.mygrid.org.uk.uk/
- http://github.com/mygrid/scufl2