A new Approach for Publishing Workflows: Abstractions, Standards and Linked Data

Daniel Garijo

Ontology Engineering Group, Departamento de Inteligencia Artificial. Universidad Politécnica de Madrid

Yolanda Gil

Information Sciences and Institute
University of Southern California, Marina del Rey

Date: 14/11/2011

Index:

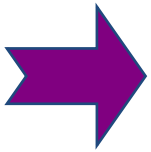
Typical Published Article

Text:
Narrative of method,
software packages used

Data:
Key datasets and figures/plots

NOT published,
loosely recorded:

Software:
scripted codes + manual steps +
notes/emails



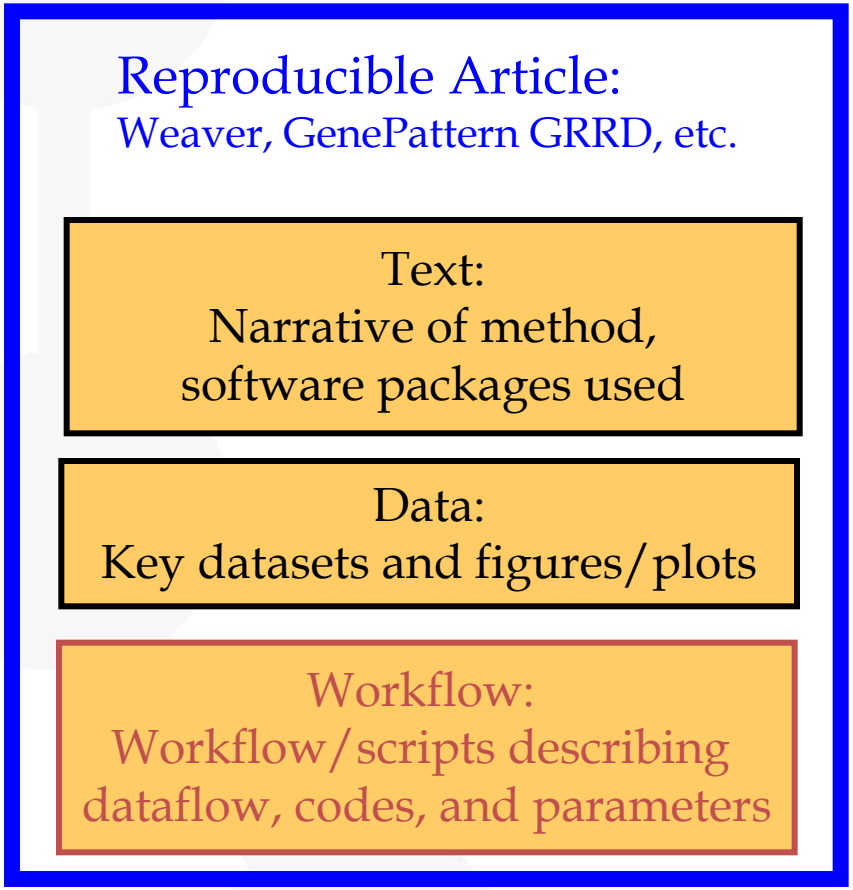
Reproducible Article:
Weaver, GenePattern GRRD, etc.

Text:
Narrative of method,
software packages used

Data:
Key datasets and figures/plots

Workflow:
Workflow/scripts describing
dataflow, codes, and parameters

Current issues with existing publication approaches



Only executable workflow is published:

1. Must have the **same codes** to re-execute the workflow, but:
   - Codes become unavailable
     - Eg: eHits was proprietary and replaced by AutodockVina
   - Different labs prefer different codes
     - Eg: R vs Matlab
     - Eg: viz in Citoscape vs yEd
2. Must have the **same workflow framework** to re-execute the workflow
   - Must have R for Weaver
3. Must **import files** to local file system and workflow framework
   - Must import bundle of workflow/data/code files to reproduce

Key Features of our approach

- Publish an **abstract workflow** in addition to executable workflow
  - Description of workflow that is independent of the codes executed
  - Maps to the codes executed (the “executable workflow”)
- Publish both abstract and executable workflow **using the OPM standard**
  - OPM (Open Provenance Model) is independent of workflow framework and is widely implemented
  - Other groups can import to their own workflow framework
- Publish data and workflows as **Linked Data** on the Web
  - All workflows and related files are web-accessible
  - Simple mechanism to share across local file systems

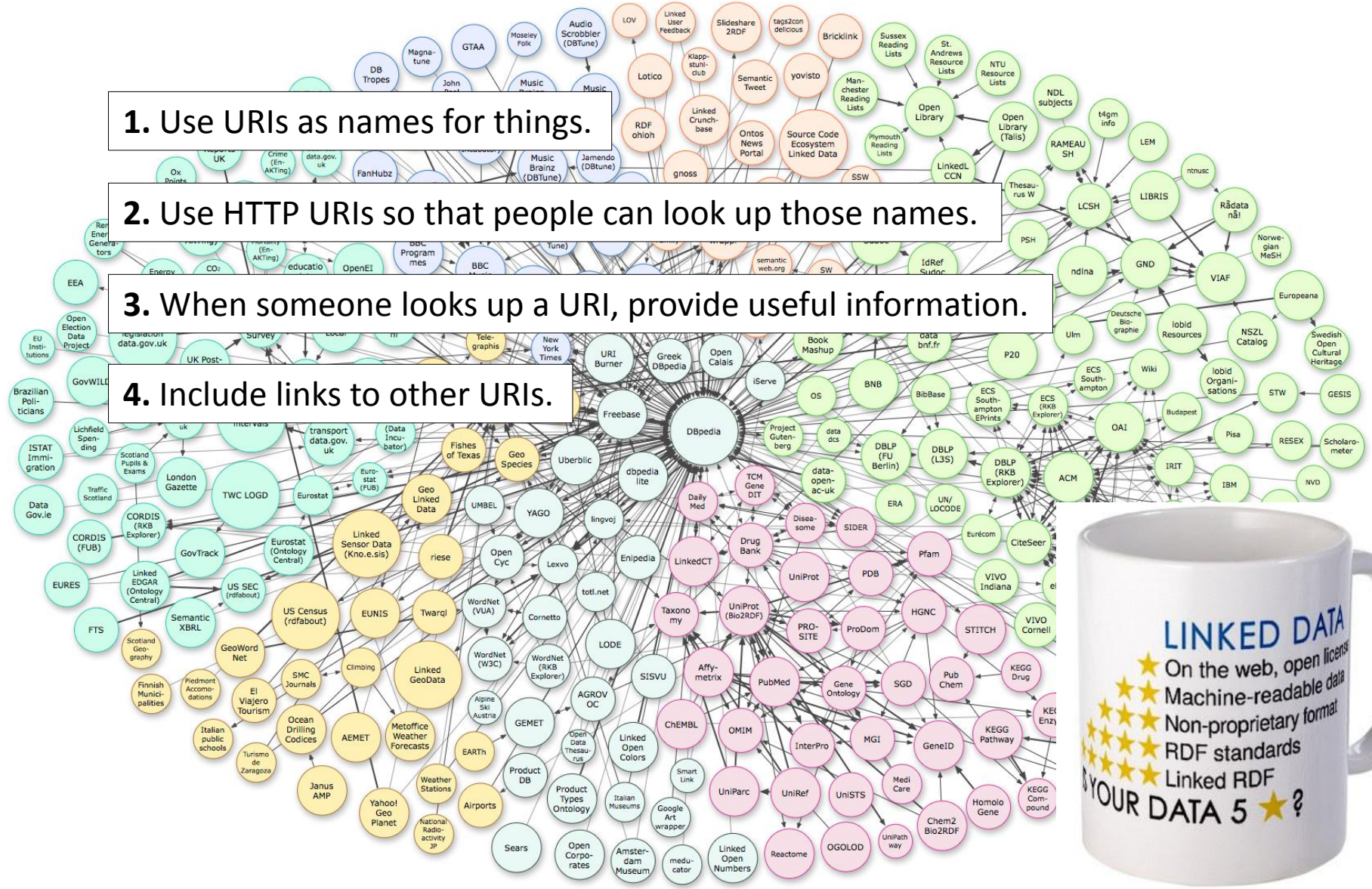
What is Linked Data

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

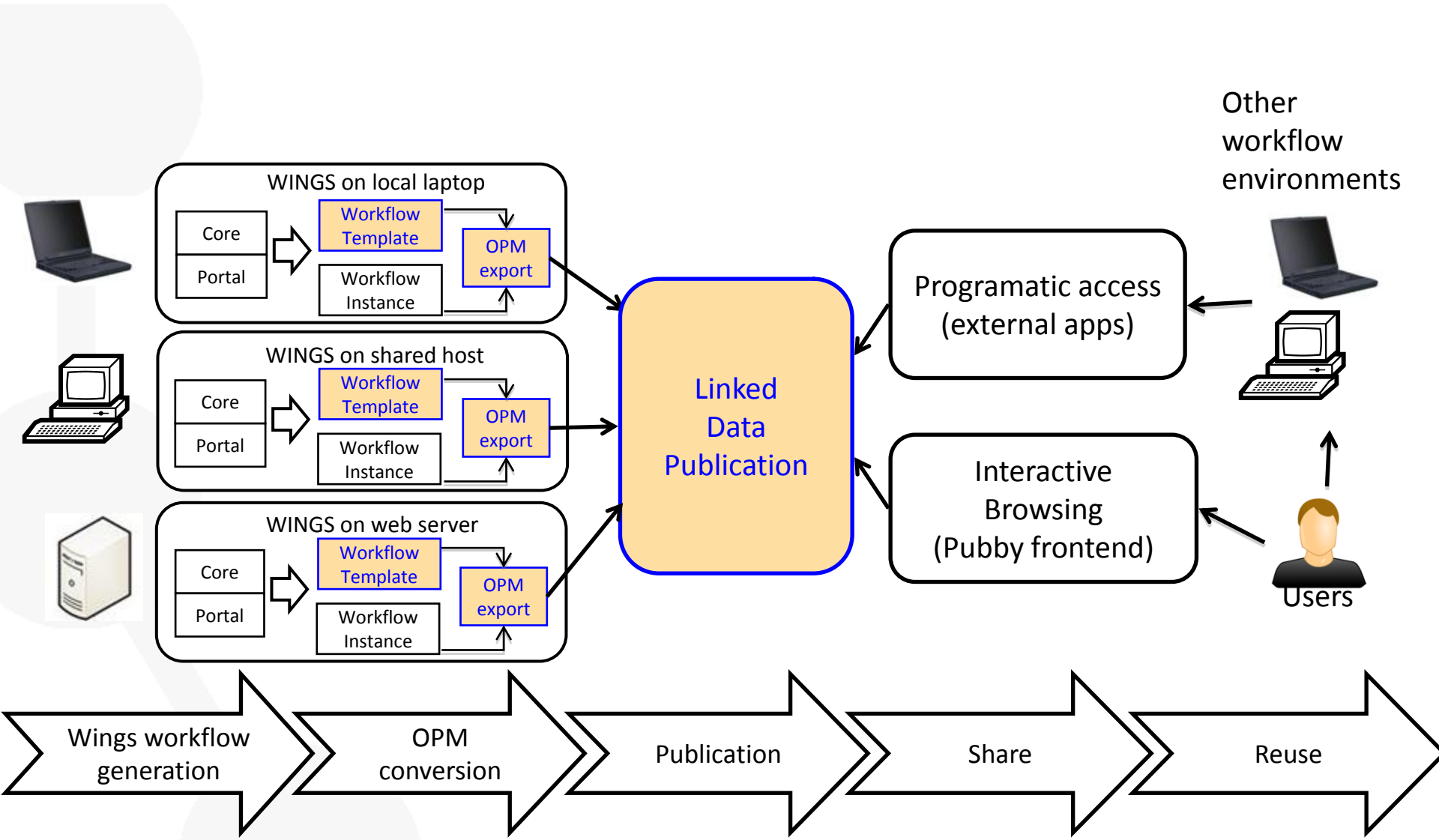
3. When someone looks up a URI, provide useful information.

4. Include links to other URIs.

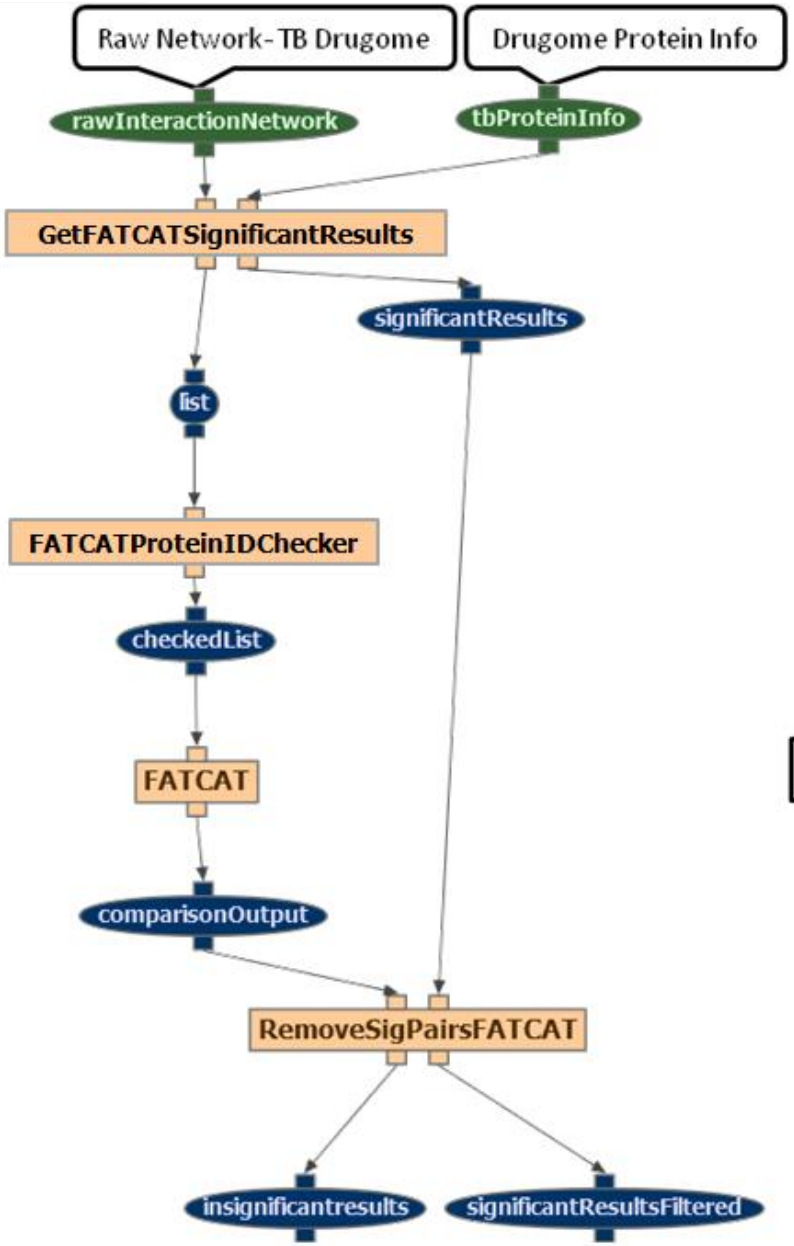


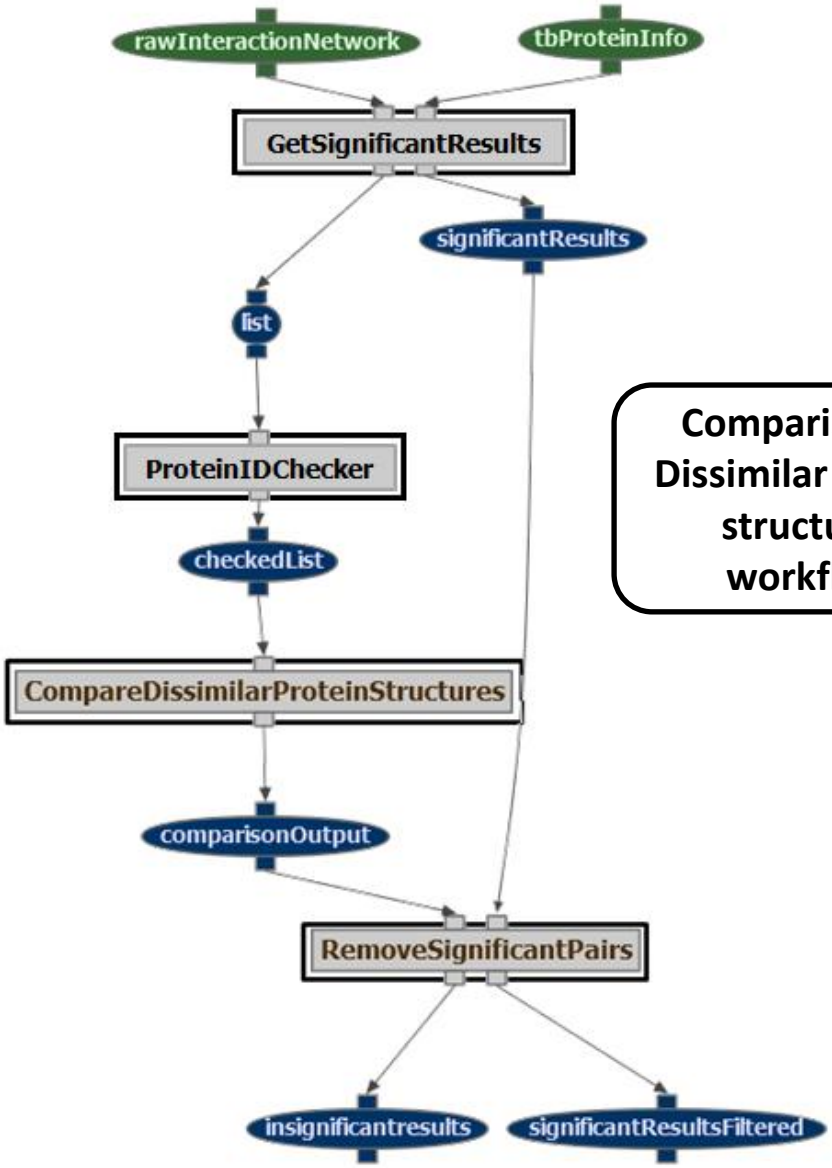
"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

High level architecture



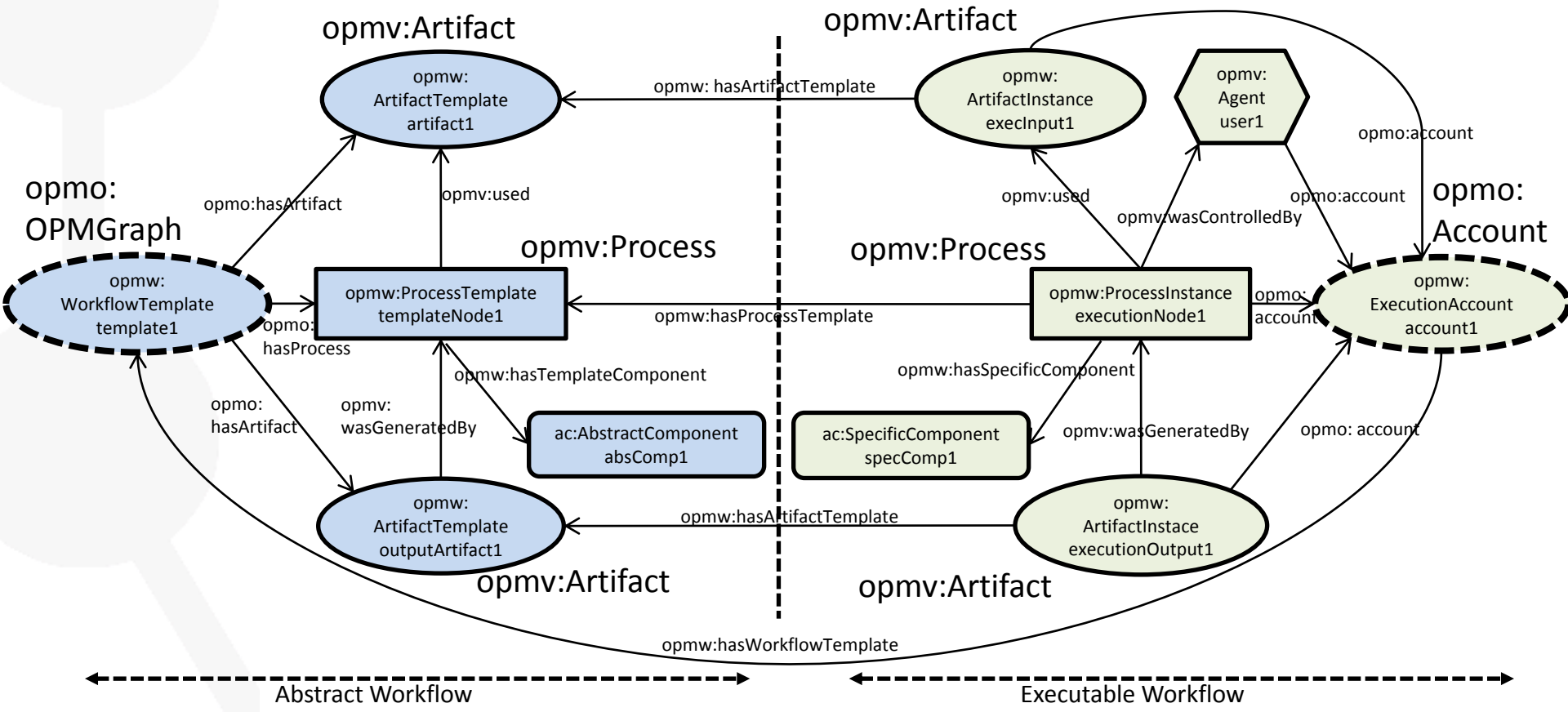
Publishing the abstract workflow



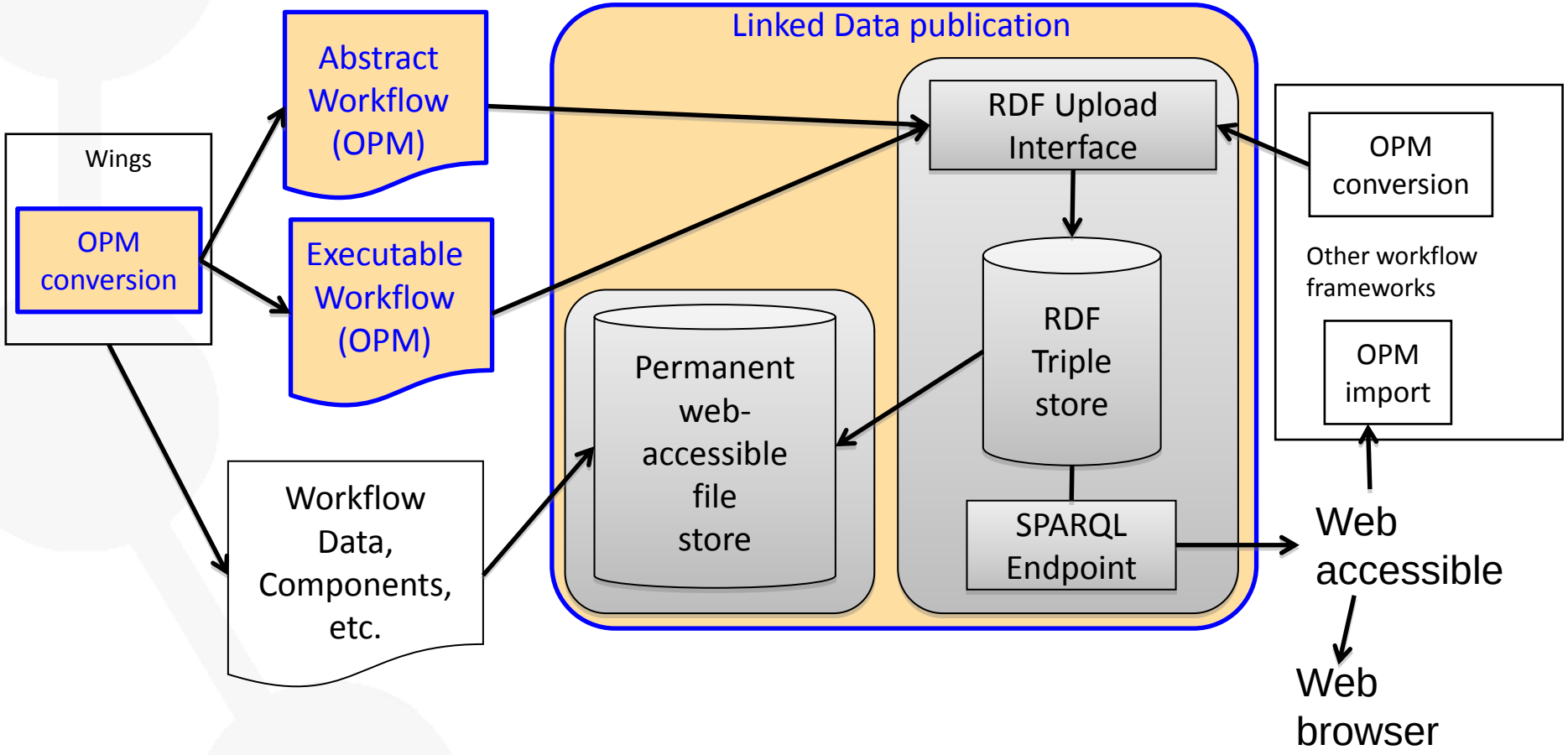


Comparison of Dissimilar protein structures workflow

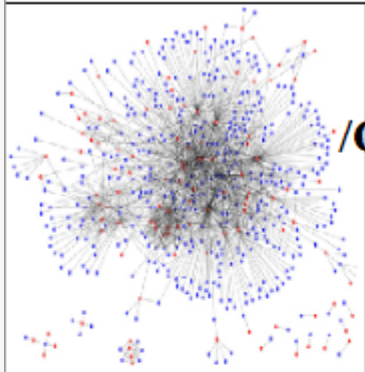
OPMW Ontology



Publication of Workflows as Linked Data



Searching/Browsing Workflows as Linked Data



Types of search

Workflow search | Author and contributor search | Global search

http://wings.isi.edu/opmexport/resource/ProcessInstance/COMPAREDISSIMILARPROTEINSTRUCTURES1314662017568

Global search (all resources)

Resource URI (Process instance)

Insert the name of a resource to chech its related information

compareDissimilarProteinStructures

Autocomplete search bar

http://wings.isi.edu/opmexport/resource/ProcessInstance/FORMATSTEPTOCOMPAREDISSIMIL

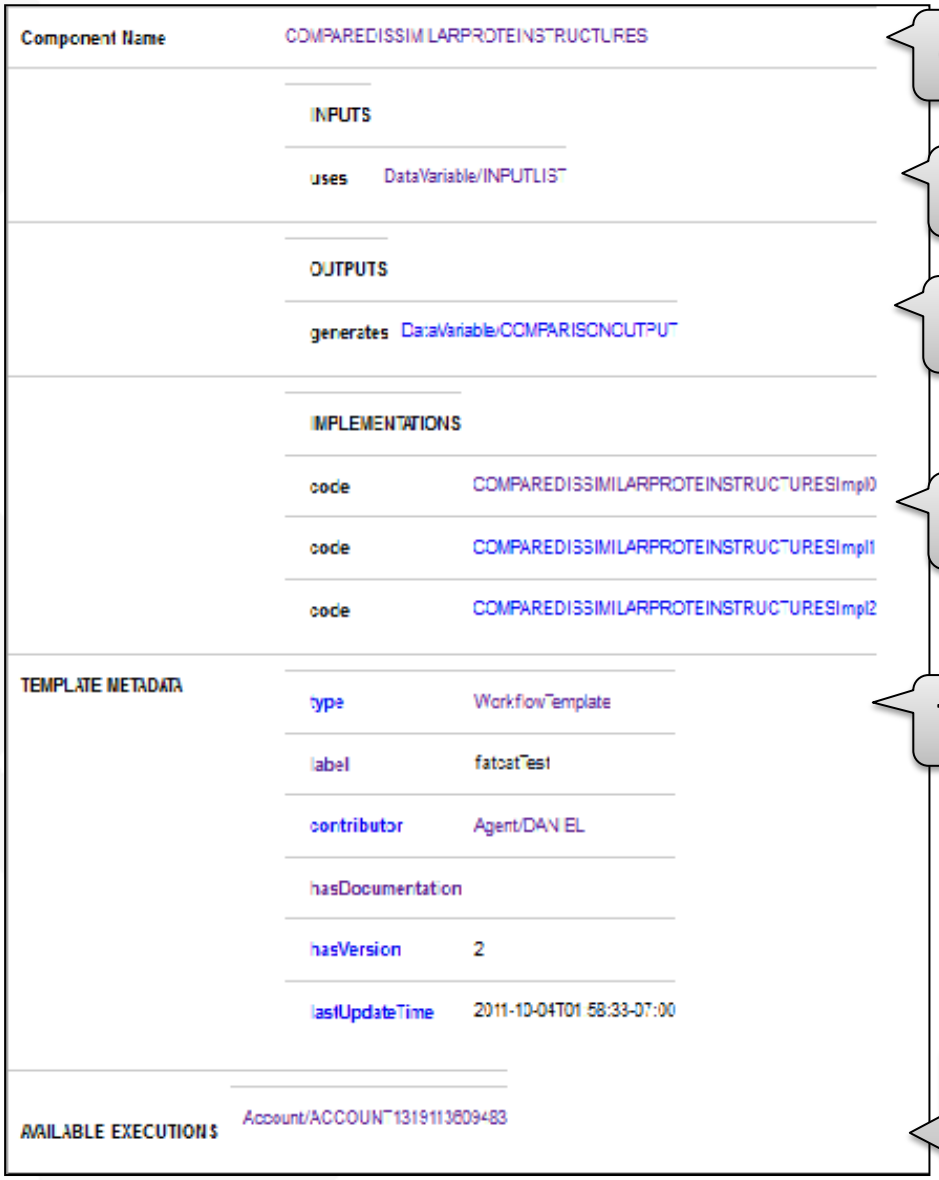
http://wings.isi.edu/opmexport/resource/ProcessInstance/COMPAREDISSIMILARPROTEINSTRU

Properties

Specific component for this process instance

hasSpecificComponent	FATCAT/COMPONENTFATCAT
hasProcessTemplate	ProcessTemplate/ABSTRACCTSUBWFFATCAT_COMPAREDISSIMILARPROTEINSTRUCTURES
wasControlledBy	Agent/DANIEL
used	ArtifactInstance/31A41E8CDA244334B692F13A2CB7C4C1
account	Account/ACCOUNT1314662017568
type	ProcessInstance

Searching/Browsing Workflows as Linked Data



Component Name

Component Inputs

Component Outputs

Code Implementations

Template additional metadata

Record of the different executions of this workflow

Conclusions

1. Publication of an **abstract workflow** that represents the computational method in an **execution-independent manner**.

2. Publication of **the abstract workflow** and the **executed workflow** using the **OPM standard** that is independent of the execution environment used.

3. Publication of the workflows, components, codes and datasets **as Linked Data** on the web.

Future work

- Extensions to abstract workflow publication
  - Be able to provide abstractions **on several steps**.
  - Incomplete provenance.

- Create an OPMV/W3C PROV-O profile for common workflow representation.
  - **Increase interoperability** with other workflow representation systems.

- Workflow reuse in **different workflow systems**.
  - Import and execute workflows in other workflow frameworks.

References

- WINGS workflow system: http://seagull.isi.edu/marbles/
- The Open Provenance Model Specification: http://openprovenance.org/
- OPMO: http://openprovenance.org/model/opmo
- OPMV: http://open-biomed.sourceforge.net/opmv/ns.html
- TB Drugome Wiki (Evolution of this work): http://seagull.isi.edu/wings-drugome/index.php/Main_Page
- W3C PROV-O current ontology (draft): http://www.w3.org/2011/prov/wiki/PIL_OWL_Ontology
- Principles of Linked Data: http://www.w3.org/DesignIssues/LinkedData.html

Acknowledgements

- UCSD people:
  - **Li Xie**
  - **Lei Xie**
  - **Sarah Kinnings**
  - **Phil Bourne**
- ISI people:
  - **Varun Ratnakaar**
- OEG people:
  - **Oscar Corcho**

A new Approach for Publishing Workflows: Abstractions, Standards and Linked Data

Daniel Garijo

Ontology Engineering Group, Departamento de Inteligencia Artificial. Universidad Politécnica de Madrid

Yolanda Gil

Information Sciences and Institute
University of Southern California, Marina del Rey

Date: 14/11/2011