



Netherlands
Bioinformatics
Centre

Opening new gateways to workflows for life scientists

Kostas Karasavvas

Marco Roos

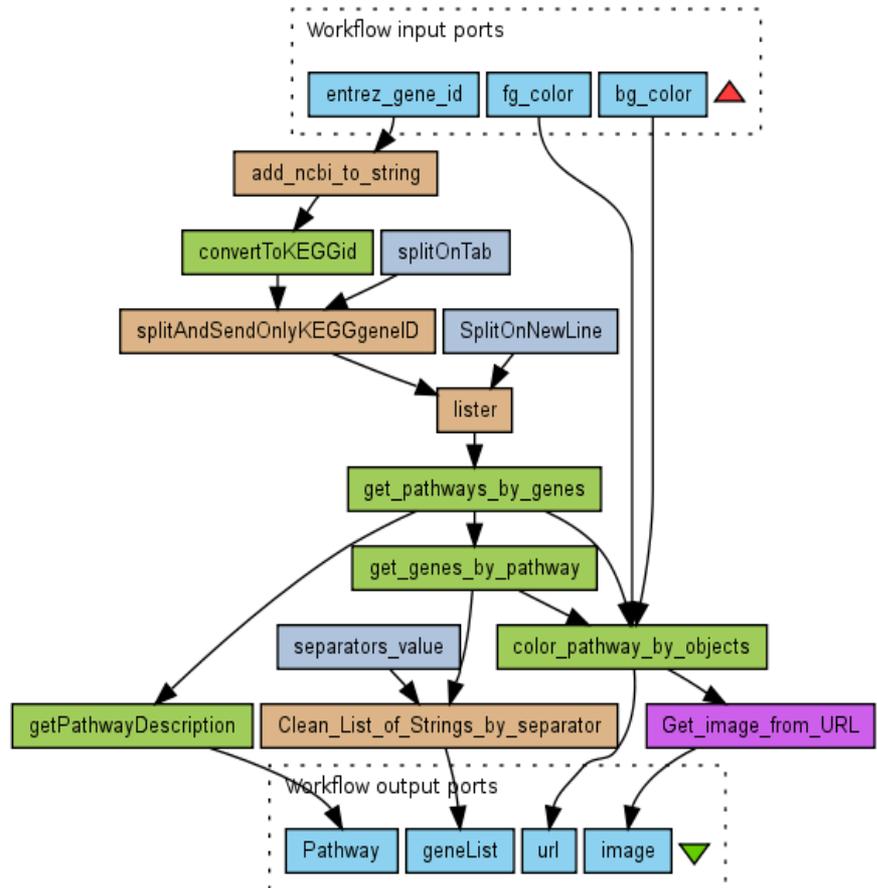


Leiden University Medical Center



Workflow approach

- Structured approach to performing bioinformatics experiments
- Steps are exposed
- Helps evaluation by supervisors and peers
- Compares to 'Materials and Methods', but you can run them
- Easier to reuse & extend (manage complexity)



Observations

- Two widely-used systems: Galaxy & Taverna
- Typical usage scenario's
 - Galaxy:
 - 'Scripting bioinformatician' adds command line tools to local Galaxy server for their biological colleagues
 - Biologists interactively 'play with data'
 - Save interactive analysis as a workflow
 - Strength: simplicity
 - Taverna
 - Bioinformatician designs a workflow to perform complicated data analysis / data integration
 - Strength: sophisticated workflows

Why interoperate?

- Use each other's tools, use each other's workflows
 - Example:
 - calculate genome region overlaps in Galaxy
 - annotate genes by text mining Web Services in Taverna
- Exploit strengths
 - Galaxy: simplicity
 - Taverna: sophistication
- Allow user communities to benefit from each other's work
 - greater accessibility

Taverna Workbench

The screenshot displays the Taverna Workbench interface, which is divided into several panels:

- Service panel:** Located on the left, it contains a search filter, a "Clear" button, and an "Import new services" button. Below these are "Available services" categorized into "Service templates", "Local services", and several "WSDL" services from various sources like Moby, Soaplab, and EBI.
- Workflow diagram:** The central area shows a flowchart of a workflow. It starts with three "Workflow input ports": `entrez_gene_id`, `fg_color`, and `bg_color`. The process flow includes:
 - `add_ncbi_to_string` (orange box)
 - `convertToKEGGid` (green box) and `splitOnTab` (blue box)
 - `splitAndSendOnlyKEGGgeneID` (orange box) and `SplitOnNewLine` (blue box)
 - `lister` (orange box)
 - `get_pathways_by_genes` (green box)
 - `get_genes_by_pathway` (green box)
 - `color_pathway_by_objects` (green box)
 - `separators_value` (blue box)
 - `getPathwayDescription` (green box)
 - `Clean_List_of_Strings_by_separator` (orange box)
 - `Get_image_from_URL` (purple box)The workflow concludes with four "Workflow output ports": `Pathway`, `geneList`, `url`, and `image`.
- Workflow explorer:** Located at the bottom left, it shows a tree view of the current workflow, including "Workflow input ports", "Workflow output ports", and "Services".

Galaxy

The screenshot shows the Galaxy web interface. At the top, there is a browser window with the URL 'localhost:8080' and a search bar. Below the browser window is the Galaxy navigation bar with tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. The main content area is divided into three panes:

- Tools Pane (Left):** Contains a list of tools under various categories like 'Taverna Workflows', 'Get Data', 'Send Data', etc. The tool 'Get Pathway-Genes by Entrez gene id' is selected and highlighted.
- Tool Configuration Pane (Center):** Displays the configuration for the selected tool. It includes:
 - Select source for entrez_gene_id:** A dropdown menu set to 'Type manually'.
 - Enter entrez_gene_id:** A text input field containing '3064'.
 - Would you also like the raw results as a zip file:** A dropdown menu set to 'No'.
 - Execute** button.
- History Pane (Right):** Shows a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Below the tool configuration pane, there is a section titled 'What it does' with a description: 'Given a specific entrez gene id, returns the pathways that this gene participates in and for each of those pathways which genes are associated with.' This is followed by 'Inputs' (entrez_gene_id) and 'Outputs' (Pathway, geneList). A warning message states: 'Please note that some workflows are not up-to-date or have dependencies that cannot be met by the specific Taverna server that you specified during generation of this tool. You can make sure that the workflow is valid by running it in the Taverna Workbench first to confirm that it works before running it via Galaxy.' Another warning message states: 'Please note that there might be some repetitions in the workflow description in some of the generated workflows. This is due to a backwards compatibility issue on the myExperiment repository which keeps the old descriptions to make sure that no information is lost.' A final information message says: 'For more information on that workflow please visit <http://www.myexperiment.org/workflows/2805>'.

At the bottom of the interface, there is a search bar with the text 'time' and navigation buttons for 'Previous', 'Next', 'Highlight all', and 'Match case'.

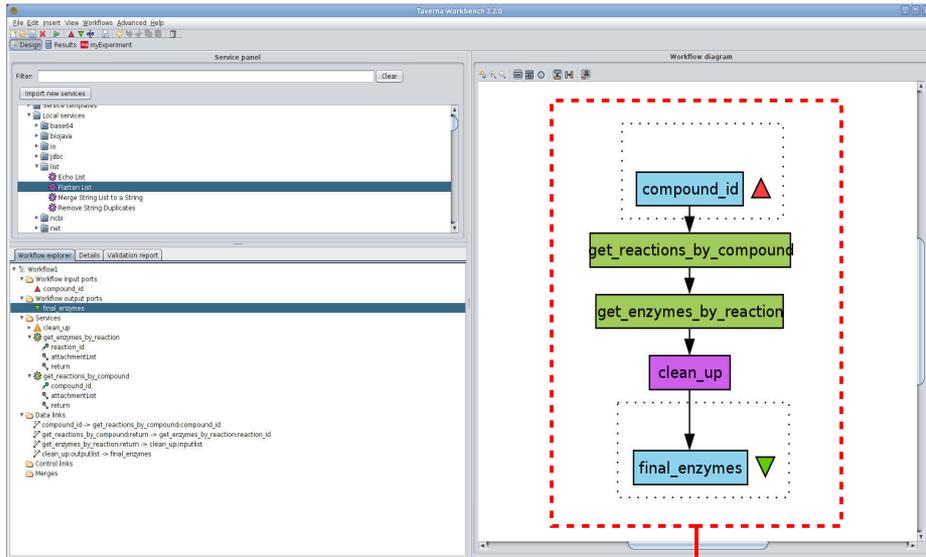
myExperiment

The screenshot shows the myExperiment website in a Mozilla Firefox browser window. The browser's address bar displays the URL <http://www.myexperiment.org/>. The website features the myExperiment logo on the left and a yellow banner with the text: "myExperiment makes it easy to find, use and share scientific workflows and other Research Objects, and to build communities." Below the banner is a search bar with a dropdown menu set to "All" and a "Search" button. The main content area is divided into three columns. The left column, titled "First time visitor? Try these videos:", lists "Project Introduction" and "Bioinformatics Case Study", followed by "Use myExperiment to..." with links for "Find Workflows", "Share Your Workflows and Files", "Create and Find Packs of Items", "Find People and Make Friends", and "Create and Join Groups". The middle column features a large "Explore" button, a workflow diagram, and a "Find Workflows" button. The workflow diagram shows "Workflow inputs" (query, program, database) leading to "species_file", "chromosome_file", and "blast_query", which then lead to "blastRecompiler" and "Workflow outputs" (compared_output, blast_output). The right column has a "Register" button, "or Login:", and a login form with fields for "Username or Email:", "Password:", and "Remember me:" (checkbox), along with an "Or use OpenID:" section.

Our work

- Taverna workflows available in Galaxy
- Taverna workflows available on a web browser
- Demonstration server with Galaxy & Taverna servers, T2Web
 - Galaxy tools/workflows available in Taverna
 - Virtual Image (for production as well)

Taverna Workflows in Galaxy (1)



Galaxy

Analyze Data Workflow Shared Data Help User

Tools manipulation NGS: Mapping NGS: Indel Analysis NGS: RNA Analysis NGS: SAM Tools NGS: Peak Calling NGS: Simulation SNP/WGA: Data; Filters SNP/WGA: QC; LD: Plots SNP/WGA: Statistical Models Human Genome Variation VCF Tools Taverna Workflows

- FBI InterProScan for Taverna 2
- Workflow1
- Workflow2
- Get enzyme classifications of a compound

Get enzyme classifications of a compound

Select source for compound_id:
Type manually ▾

Enter compound_id:
C15973

Would you also like the raw results as a zip file:
No ▾

Execute

What it does
Given a compound we want to know all the reactions that it participates in so that we get all the enzymes that drive those reactions. It uses KEGG services.

Inputs

- compound_id The compound id (from KEGG). Examples include:
 - C15973

Outputs

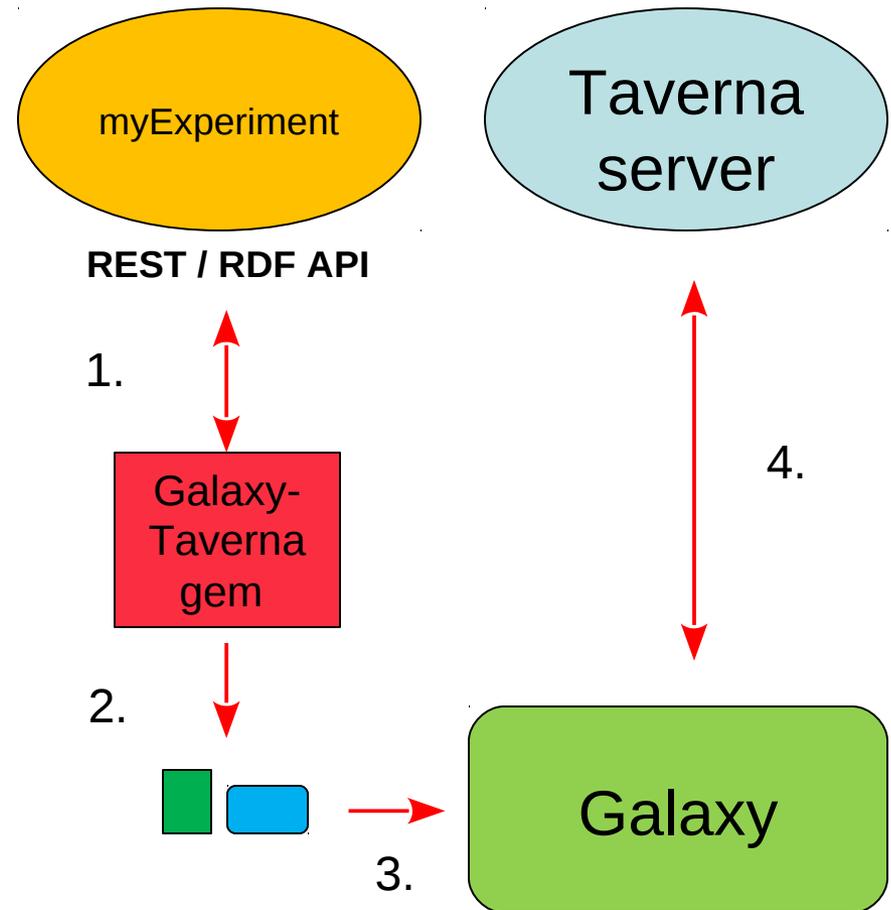
- enzymeClassificationList Enzyme classifications Examples include:
 - ec:2.3.1.168

History Options ▾

Your history is empty. Click 'Get Data' on the left pane to start

Taverna Workflows in Galaxy (2)

- Galaxy-Taverna component
 - ruby gem
 - behind the scenes
 - generates a Galaxy tool
 - requires a workflow description
- Workflow description
 - myExperiment
 - workflow file
- Galaxy
 - tool needs to be manually installed



Taverna Workflows in Galaxy (3)

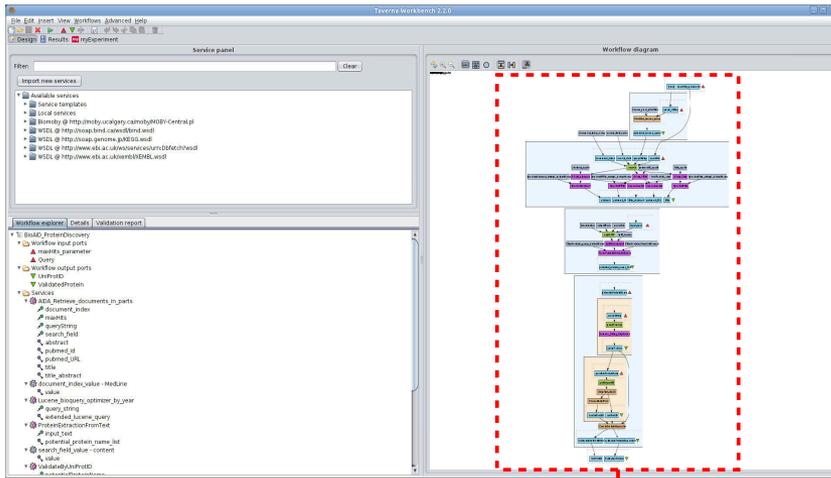
The screenshot displays the myExperiment website interface. At the top, the browser address bar shows the URL <http://www.myexperiment.org/workflows/820.html>. The website header includes the myExperiment logo, navigation links (Home, Users, Groups, Workflows, Files, Packs, Services, Topics), and user options (Logout, Give us Feedback, Invite). The main content area features a search bar and a navigation breadcrumb: Home » Workflows » EBI_InterProScan for Taverna 2. The central focus is the 'Workflow Entry: EBI_InterProScan for Taverna 2', which includes metadata such as creation and update dates, license information, and a list of credits. A 'Preview' section shows a complex workflow diagram with various tools and data flows. To the right, a sidebar provides user information for 'Kostas', including profile links, membership details, and a list of friends and groups.

Taverna Workflows in Galaxy (4)

The screenshot shows a web browser window displaying the Taverna workflow page for EBI InterProScan. The browser's address bar shows the URL <http://www.myexperiment.org/workflows/820.html>. The page content is organized into several sections:

- Diagram:** A workflow diagram is shown at the top, with a button labeled "Download Scalable Diagram (SVG)".
- Description:** A section titled "Description" provides details about the InterProScan analysis, including the input sequence and the use of the EBI's WSInterProScan service.
- Download:** A section titled "Download" contains two buttons: "Download Workflow File/Package (T2FLOW)" and "Download Workflow as a Galaxy tool". A red arrow points to the latter button.
- Run:** A section titled "Run" provides instructions on how to execute the workflow in the Taverna Workbench, including a link to the workflow's download page.
- Sidebar:** A sidebar on the right contains several panels: "People/Groups" (listing users like Stian Solland-Reyes), "Attributions (5)" (listing related workflows), "Tags (5)" (listing tags like interpro, interproscan, and sequence), "Shared with Groups (1)" (listing the myGrid group), "Featured In Packs (1)" (listing the Taverna 2.1 beta 2 pack), and "Ratings (0)".
- Right Panel:** A panel on the far right contains "My Favourites", "My Tags", and "Popular Tags" sections.

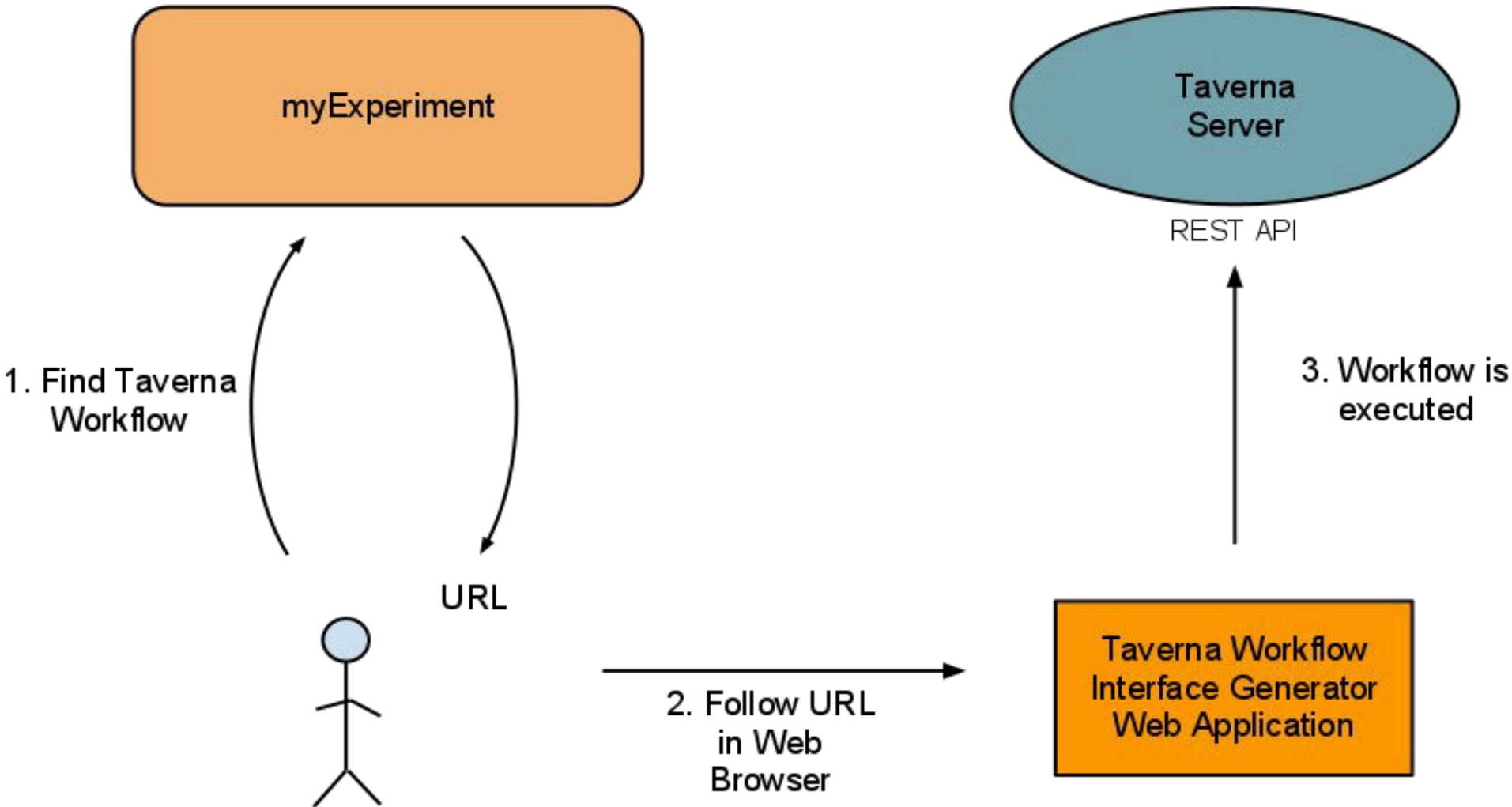
Taverna Workflows on a web browser (1)



The screenshot shows a web browser window displaying the Taverna workflow 'BioAID_ProteinDiscovery' configuration page. The page is titled 'Workflow: BioAID_ProteinDiscovery' and is attributed to 'workflow by Marco Roos'. It features a 'Configure Workflow Inputs' section with two input fields: 'Enter Query:' (containing 'scarsabacans proteins" AND anyload') and 'Enter maxHits_parameter:' (containing '3'). There are 'Upload file?' checkboxes and an 'Execute' button. Below this is a 'Workflow Description' section with the text: 'The workflow extracts protein names from documents retrieved from MedLine based on a user Query (cf Apache Lucene syntax). The protein names are filtered by checking if there exists a valid UniProt ID for the given protein name.' At the bottom, there is a table with columns 'Output', 'Description', and 'Examples', showing 'ValidatedProtein' and 'UniProtID'. A 'Please Note' section is also present.



Taverna Workflows on a web browser (2)



Taverna Workflows on a web browser (3)

The screenshot shows a Mozilla Firefox browser window with the URL <http://workflow.mybiobank.org/t2web/workflow/74> in the address bar. The page title is "Workflow: BioAID_ProteinDiscovery" and it is attributed to "workflow by Marco Roos". The interface includes the nbic logo (netherlands bioinformatics centre) and the Leiden University Medical Center logo. The main content area is titled "Configure Workflow Inputs" and contains two input sections. The first section, "Enter Query:", has a text box containing the query `"transmembrane proteins" AND amyloid` and an "Upload file?" checkbox. The second section, "Enter maxHits parameter:", has a text box containing the value `3` and an "Upload file?" checkbox. A red arrow points to the "Execute" button below the second section. Below the input section is a "Workflow Description" section with the text: "The workflow extracts protein names from documents retrieved from MedLine based on a user Query (cf Apache Lucene syntax). The protein names are filtered by checking if there exists a valid UniProt ID for the given protein name." Below the description is a table with three columns: "Output", "Description", and "Examples". The "Output" column lists "ValidatedProtein" and "UniProtID". Below the table is a "Please Note" section.

Taverna Workflows on a web browser (4)

Results: - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://workflow.mybiobank.org/t2web/enact

ValidatedProtein

UniProtID

Workflow: BioAID_ProteinDiscovery

workflow by Marco Roos

netherlands bioinformatics centre

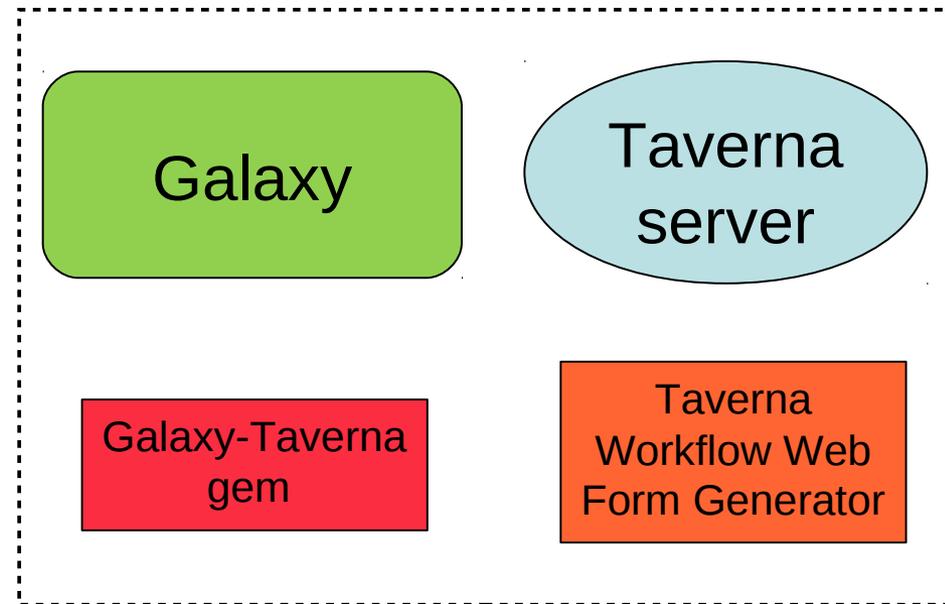
Leiden University Medical Center

P70386
Q02527
Q09327
Q10470
Q14CK5
Q6IC49
Q9UH32
A8K7C2
O73815
P02571
P02579
P12714
P14104
P53478
P60010
P63259
P63260
P63261

Transferring data from workflow.mybiobank.org...

Demonstration Server (1)

- Galaxy+Taverna Server 'in one box' and more
 - demonstration
 - preconfigured
 - 'playground'
 - Taverna → Galaxy



Demonstration Server (2)

- Galaxy+Taverna Server 'in one box' and more
- Galaxy Server
 - <http://galaxy.nbic.nl/galaxy>
 - with some example taverna workflows
- Taverna Server
 - <http://galaxy.nbic.nl/demo/taverna-server>
- Taverna workflows to Galaxy tools generator
- Taverna workflows web interface generator web application
 - <http://galaxy.nbic.nl/t2web/workflow/74>

Conclusions

- More interoperable Galaxy – Taverna workflows
- Taverna workflows can be accessed in Galaxy
 - ... and thus take part in a Galaxy workflow
- Taverna workflows can be accessed via the web
 - Bioinformatician creates the workflow
 - ... sends the URL to biologist
- Demonstration Server
 - Virtual machine

Acknowledgements

- NBIC BioAssist developers
- myGrid team (developers of Taverna)
- Galaxy developers
- Users 'in the loop'
 - Eleni Mina
 - Harish Dharuri
 - Pieter Neerincx
 - Jelle Scholtalbers
- Our colleagues at the Human Genetics Department, LUMC, NL
 - BioSemantics group LUMC-Leiden/EMC-Rotterdam
- NBIC-BioAssist/Wf4ever (EU-FP7)



Questions?

- More information
 - **Taverna-Galaxy** → <https://trac.nbic.nl/elabfactory/wiki/eGalaxy>
 - **Taverna-Web** → <https://trac.nbic.nl/elabfactory/wiki/t2web>
 - **Virtual Machine** → https://wiki.nbic.nl/index.php/Galaxy_VM
 - **Demonstration Server** → <http://galaxy.nbic.nl/galaxy>
 - **Galaxy** → <http://galaxy.psu.edu/>
 - **Taverna** → <http://www.taverna.org.uk/>
 - **myExperiment** → <http://www.myexperiment.org/>
- Contact
 - elaboratory-users@trac.nbic.nl (preferred)
 - kostas.karasavvas@nbic.nl
 - m.roos@lumc.nl