Documentation and visualisation of workflows for effective curation@source

Cerys Willoughby & Jeremy Frey
I’m recovering from a brain injury

Sometimes my words don’t work!
What is a scientific workflow?

Computational methods to process, analyse, and model scientific data

INPUT → TRANSFORMATION → OUTPUT

'in silico' experiments
Can be automated
Can be reused
Workflow platforms
Transformations
Sharing and reuse

Workflows are ideal for sharing and building upon in theory, but lots of studies have shown that workflows need additional data and context to make them reusable.

Even with additional materials, still can’t understand it without having the original software. Issues with access to O/S, correct version, underlying software stack, licenses, hardware, lack of skills, and so on.
An example of the problem..

An email arrived with a workflow file attached, an excel file, and this blurry picture of the workflow. The email provided a sentence explaining where the input data was derived, but no explanation about what the workflow was doing or how it worked!

Chemical compounds, identifiers, names, and lots and lots of numbers!
What can we capture that might help?

<table>
<thead>
<tr>
<th>Composition</th>
<th>Construction</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYPOTHESIS</td>
<td>INTERMEDIATE DATA</td>
<td>CONCLUSIONS</td>
</tr>
<tr>
<td>DESIGN</td>
<td>INTERMEDIATE DATA</td>
<td>SHARING</td>
</tr>
<tr>
<td>TECHNOLOGY CHOICE</td>
<td>SCRIPTS</td>
<td>PUBLICATION</td>
</tr>
<tr>
<td>PROTOCOL</td>
<td>MODELS</td>
<td>REUSE</td>
</tr>
<tr>
<td>PARAMETERS</td>
<td>RESULTS DATA</td>
<td>CURATION &amp; ARCHIVE</td>
</tr>
</tbody>
</table>

Research Objects and Workflow Packages
Still the problem of relying on access to the execution environment and a disconnection between the underlying configuration and the workflow
An Interactive Workflow Visualisation

• Using KNIME
An example workflow in KNIME

This Example Workflow uses a File Reader node to import the Iris dataset (included). It then assigns visual properties with a Color Manager node and computes some basic statistics with a Statistics node. The data is split into training and testing fractions with a Partitioning node. The Decision Tree Learner generates a predictive model in PMML from the training fraction which is then applied to the test fraction using the Decision Tree Predictor. Model performance is evaluated with a Scorer node, which is applied after the Decision Tree Predictor. Finally, errors can be explored interactively, by using an Interactive Table node to highlight certain classes of errors which can then be visualized using a Scatter Plot node.
Visualise Workflow (Subflow) Node
Constructing an interactive visualisation
Example Workflow

Author: gabriel Last edited by: cerys
Flow variable name: Project value: WF Class: STRING
Flow variable name: Run Name value: 1.1 Class: DOUBLE

Example workflow design

Author: - Microsoft Office User Date: Mon Oct 10 01:00:00 BST 2016

In a real-world scenario our researcher would want to include some background information and provide the story of the development of their workflow including some rationale and details of decisions made. For example, how was the code for scripts developed, where did the data come from, has there been any preprocessing, what are the planned uses for the results? This could also contain the methods information and observations made whilst developing the workflow. Potentially also conclusion information.Example workflow design

This Example Workflow uses a File Reader node to import the iris dataset (included). It then assigns visual properties with a Color Manager node and computes some basic statistics with a Statistics node. The data is split into training and testing sets with a Partitioning node. The Decision Tree Learner generates a predictive model in PMML from the training fraction which is then applied to the test fraction using the Decision Tree Predictor. Model performance is evaluated with a Scorer node, which is applied after the Decision Tree Predictor. Finally, errors can be explored interactively, by using an Interactive Table node to highlight certain classes of errors which can then be visualized using a Scatter Plot node.

![Workflow Diagram]

Decision Tree Predictor: Apply model (Node 4)

No custom description has been provided. This node is EXECUTED.

This node is part of the KNIME Base Nodes from KNIME GmbH, Konstanz, Germany. This node is version 3.2.1.v20160816059.

Model properties:
- UseGainRatio 10000
- ShowDistribution false
- prediction column name Prediction()
- change prediction false
- class probability suffix

Variable properties:
Workflow Provenance

Example Workflow

Author: gabriel Last edited by: cerys
Flow variable name: Project value: WF Class: STRING
Flow variable name: Run Number value: 1.1 Class: DOUBLE

Workflow metadata

File name

User-defined flow variables
Text and metadata extracted from a Word document containing background information

Example workflow design

Author: Microsoft Office User Date: Mon Oct 10 01:00:00 BST 2016

In a real-world scenario our researcher would want to include some background information and provide the story of the development of their workflow including some rationale and details of decisions made. For example, how was the code for scripts developed, where did the data come from, has there been any preprocessing, what are the planned uses for the results? This could also contain the methods information and observations made whilst developing the workflow. Potentially also conclusion information.
This Example Workflow uses a File Reader node to import the Iris dataset (included). It then assigns visual properties with a Color Manager node and computes some basic statistics with a Statistics node. The data is split into training and testing fractions with a Partitioning node. The Decision Tree Learner generates a predictive model in PMML from the training fraction which is then applied to the test fraction using the Decision Tree Predictor. Model performance is evaluated with a Scorer node, which is applied after the Decision Tree Predictor. Finally, errors can be explored interactively, by using an Interactive Table node to highlight certain classes of errors which can then be visualized using a Scatter Plot node.
The Flow of Data (PMML doc)
The Flow of Data (PNG Img)
The Flow of Data (Table data)
Node Provenance

Example Workflow

In a real-world scenario, our researcher would want to include some background information and provide the story of the development of their workflow. Including some context and details of each step can be useful. For example, what was the role of each developer? What did the team discuss, and how was the workflow designed to improve their understanding and performance?

Decision Tree Predictor: Apply model (Node 4)

This node is part of the KNIME Base Nodes from KNIME GmbH, Konstanz, Germany. This node is version 3.2.1.v201608161059.
Node configuration

Example Workflow

Example workflow design

Node settings, metadata, & links to intermediate or results data
‘Get Data’ node and setting the directory for storing data
Links to data

Scatter Plot (JFreeChart): View test data (Node 464)

No custom description has been provided. This node is EXECUTED.

This node is part of the KNIME JFreeChart nodes from KNIME GmbH, Konstanz, Germany. This node is version 3.2.0.v201605301230.

Link to data file from Port 1 for Scatter Plot (JFreeChart): View test data (Node 464).

**Model properties:**

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>graph_title</td>
<td>Title</td>
</tr>
<tr>
<td>graph_title_BOOL</td>
<td>false</td>
</tr>
<tr>
<td>graph_height</td>
<td>600</td>
</tr>
<tr>
<td>graph_width</td>
<td>400</td>
</tr>
<tr>
<td>CFG_TYPE_IMAGE</td>
<td>PNG</td>
</tr>
<tr>
<td>CFG_PlotAlpha</td>
<td>0.2</td>
</tr>
<tr>
<td>CFG_FSScale</td>
<td>1.0</td>
</tr>
<tr>
<td>col_x</td>
<td>sepal length</td>
</tr>
<tr>
<td>col_y</td>
<td>sepal width</td>
</tr>
<tr>
<td>Item_Size</td>
<td>10.0</td>
</tr>
<tr>
<td>CGG_NR_ROW</td>
<td>2500</td>
</tr>
<tr>
<td>cfg_autoRange</td>
<td>true</td>
</tr>
<tr>
<td>CFG_SHOW_HISTORY</td>
<td>false</td>
</tr>
</tbody>
</table>

**Variable properties:**

![Scatter plot with data points illustrating sepal length vs. sepal width.]
Statistics: Calculates statistic measures: mean, max, min, variance, median, etc. (Node 9)

No custom description has been provided. This node is EXECUTED.

This node is part of the KNIME Statistic Nodes from KNIME GmbH, Konstanz, Germany. This node is version 3.2.0.v201606221332.

Link to data file from Port 1 for Statistics: Calculates statistic measures: mean, max, min, variance, median, etc. (Node 9).

<table>
<thead>
<tr>
<th>Model properties:</th>
</tr>
</thead>
<tbody>
<tr>
<td>compute_median</td>
</tr>
<tr>
<td>num_nominal-values</td>
</tr>
<tr>
<td>num_nominal-values_output</td>
</tr>
<tr>
<td>image format</td>
</tr>
<tr>
<td>histogram width</td>
</tr>
<tr>
<td>histogram height</td>
</tr>
<tr>
<td>enable HiLite</td>
</tr>
<tr>
<td>show min max</td>
</tr>
</tbody>
</table>

| Variable properties:               |

<table>
<thead>
<tr>
<th><strong>Column</strong></th>
<th><strong>Min</strong></th>
<th><strong>Max</strong></th>
<th><strong>Mean</strong></th>
<th><strong>Std. Deviation</strong></th>
<th><strong>Skewness</strong></th>
<th><strong>Kurtosis</strong></th>
<th><strong>Overall sum</strong></th>
<th><strong>No. missing</strong></th>
<th><strong>No. NaN</strong></th>
<th><strong>No. s=0</strong></th>
<th><strong>No. s=0</strong></th>
<th><strong>Median</strong></th>
<th><strong>Row count</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>sepal length</td>
<td>4.3</td>
<td>7.9</td>
<td>5.843</td>
<td>0.828</td>
<td>0.315</td>
<td>0.552</td>
<td>576.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td>sepal width</td>
<td>2</td>
<td>4.4</td>
<td>3.057</td>
<td>0.436</td>
<td>0.19</td>
<td>0.225</td>
<td>148.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td>petal length</td>
<td>1</td>
<td>6.9</td>
<td>3.758</td>
<td>1.766</td>
<td>3.116</td>
<td>-0.275</td>
<td>1.402</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td>petal width</td>
<td>0.1</td>
<td>2.5</td>
<td>1.199</td>
<td>0.762</td>
<td>0.581</td>
<td>-0.103</td>
<td>1.341</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>150</td>
</tr>
</tbody>
</table>
Limitations for KNIME
Futures for KNIME
Automatic packaging of Research Objects
Other platforms: VisTrails

Current Workflow Example

Provenance trails for previous workflow runs

Trail for user changes to workflow
Other platforms: Kepler

Workflow with annotation

This workflow demonstrates the use of the data transformation actors to process a genetic sequence. The sequence is displayed in three different ways, first in its native format (XML), second as a sequence element that has been extracted from the XML format, and third as an HTML document that might be used for display on a web site. Both of the latter two operations are performed using a composite actor that hides some of the complexity of the underlying operation. These composites can be thought of as 'sub-workflows' that execute a potentially complex set of tasks when called.

Author: Ilkay Altintas, May 2006, SDSC
For reusability and understandability, the workflow is only as good as the documentation and context provided!
Thanks for listening!
I’m recovering from a brain injury

Sometimes my words don’t work!