

(Meta-)Datamanagement with KNIME

SWIB 2017 Workshop

Slides and Datasets: goo.gl/S6y3ER

Your mentors

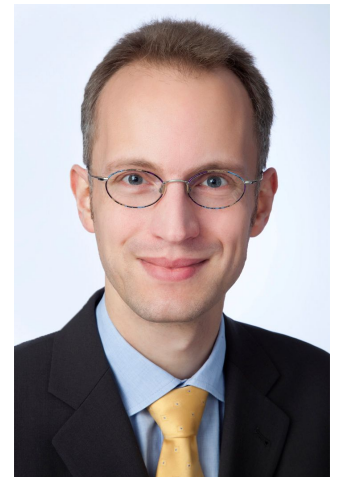
Prof. Dr. Kai Eckert

- Stuttgart Media University
- Focus: web-based informations systems



Prof. Magnus Pfeffer

- Stuttgart Media University
- Focus: information management



Current projects with data focus

Specialised information service for Jewish studies

Challenges:

- Integration of heterogenous datasets
- Contextualization using external sources
- Merging data across language and script barriers



Funding by



Consortium

Universitätsbibliothek
J.C. Senckenberg UB

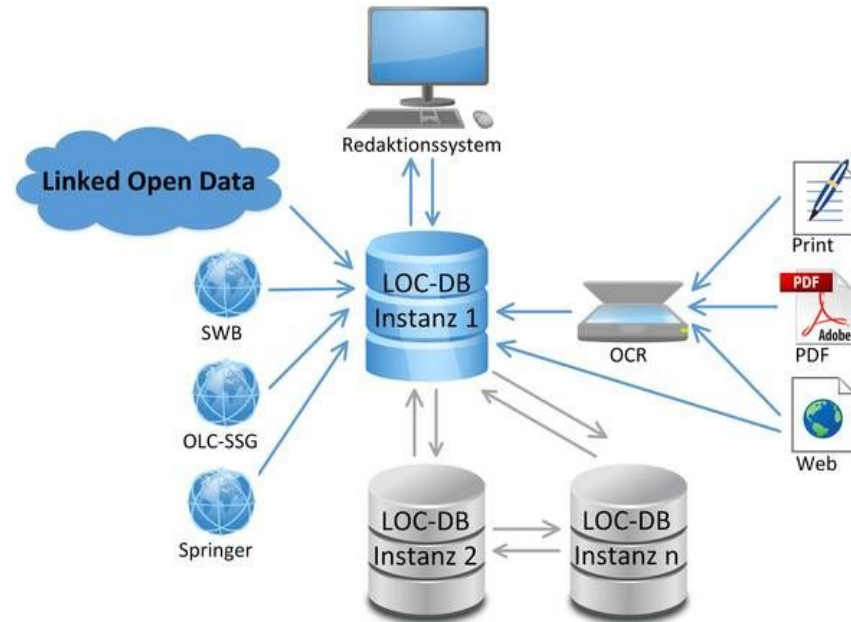


Current projects with data focus

Linked Open Citation Database

Challenges:

- Bad data
 - ... OCR'd references...
 - ... created by the authors...
- Identity resolution
- Complex data model
- Natural Language Processing



Funding by
DFG

Consortium



Current projects with data focus

Japanese visual media graph (funding pending...)

Challenges:

- Multitude of entities and relations
 - Work, release, adaption, continuation
 - Creators, producers, staff, actors
 - Characters
- No traditional data sources (libraries, etc.)
- Fan-produced data is the best available source



Consortium



JAPANOLOGIE LEIPZIG

Today's Workshop

- Part 1: Introduction (~ 2 hrs)
 - Installation and preparation
 - Basic concepts
 - Basic data workflow
 - Loading
 - Filtering
 - Aggregation
 - Analysis and visualization
 - Advanced workflow
 - Dealing with errors and missing values
 - Enriching data
 - Using maps for visualization

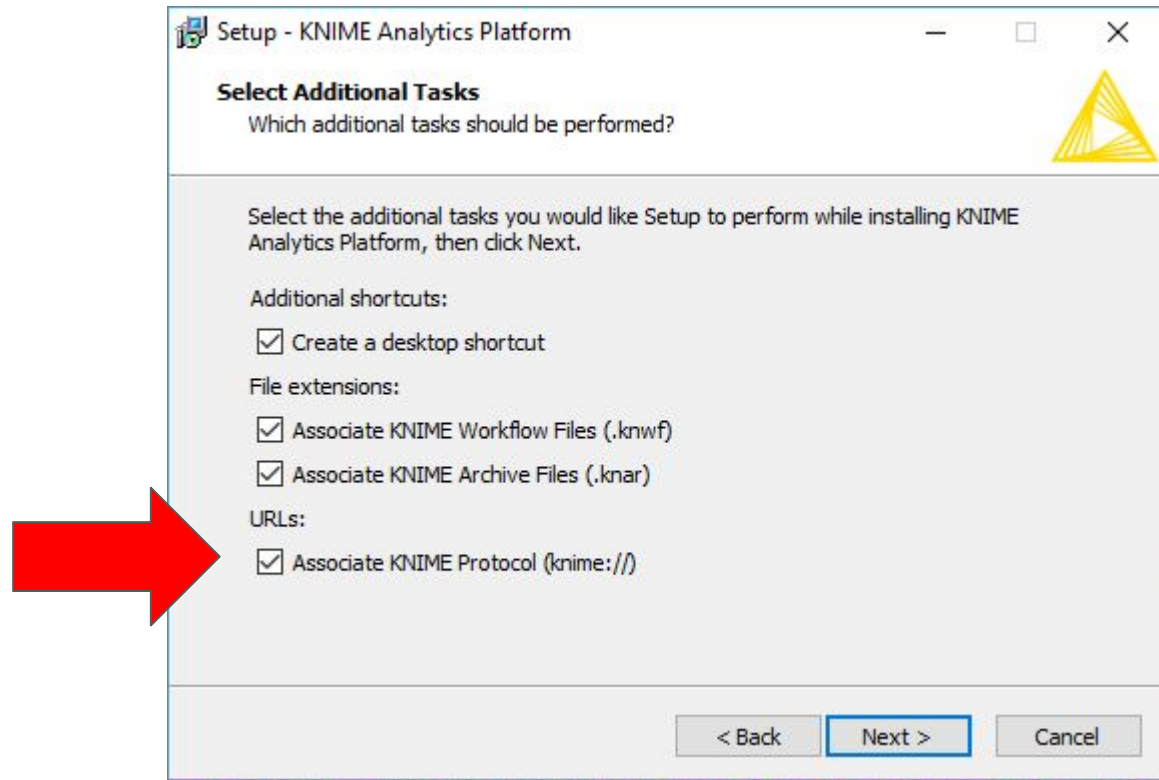
Today's Workshop

- Part 2: Real-world uses (~ 1 hr)
 - Using the RDF nodes to read and output linked data
 - Creating an enriched bibliographic dataset
 - Fixing errors in the input dataset
 - Downloading bibliographic data as XML from the web
 - Enriching with classification data from a different source
 - Data output
- Part 3: Machine Learning (~ 45 mins)
 - Split data into training and test data
 - Learn model
 - Apply model
 - Scoring
- Data challenge
 - Did you bring interesting data? Do you have any specific needs?

Part 1: Introduction

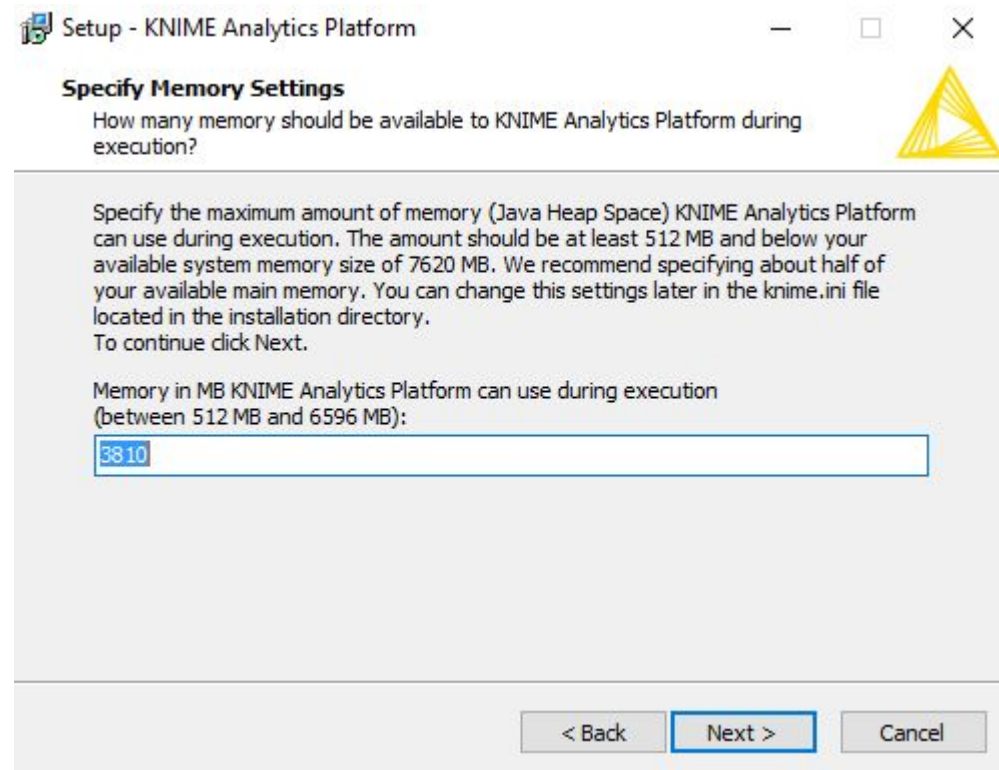
Installation

- Please chose the 64bit version whenever possible
- KNIME:// protocol support must be activated
- Use the full package, so there is no need to download modules later



Installation

- Watch out for the memory settings, allot enough memory to KNIME
- Can be changed by editing the config file KNIME.ini



Why KNIME?

Possible alternative: Develop own software tools?

Upside: Maximum flexibility

Downsides:

- Very complex, coding knowledge a necessity
- Own code can get messy, hard to maintain and document
- Shared development can lead to friction and overhead
- Modules and standard libraries often do not cover all aspects

→ Maybe it is better to use an existing toolset for metadata management

Why KNIME?

Alternative: Toolsets?

Some exist:

- Simple command-line tools and tool collections
- Catmandu
- Metafacture

→ Single tools are very inflexible

→ Toolsets are still quite complex, need coding proficiency and still are very challenging for new users

→ So maybe an application-type software would be better?

Why KNIME?

Alternative: Application software for data management?

Examples:

- OpenRefine
- d:swarm

→ Easy access, but limited functionality

→ Fixed workflow (OpenRefine) or fixed management domain (d:swarm)

→ Extensions are hard to do

That is why KNIME

Open source version available (extra functionality requires licensing)

GUI-driven data management application

Supports multiple types of different workflows

Very good documentation, self-learning support for newcomers

Many extensions exist, and creating your own is well supported

Development in a team or using other people's data workflows is integral to the software

Workflows

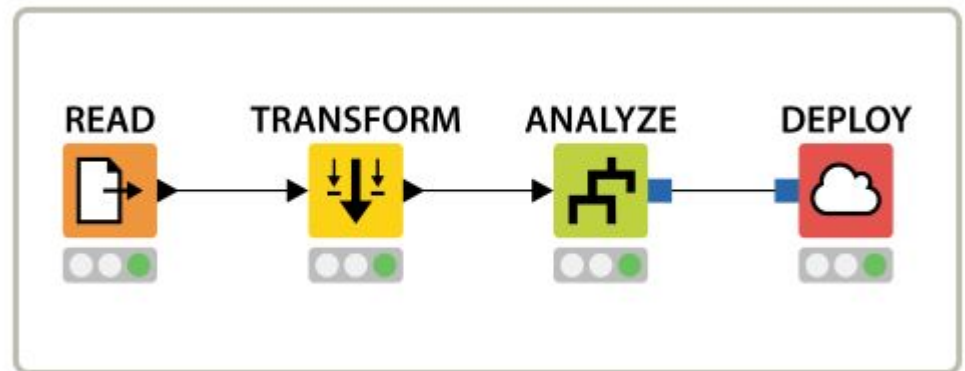
Classic data workflow: Extract, Transform, Load (ETL)

KNIME adds:

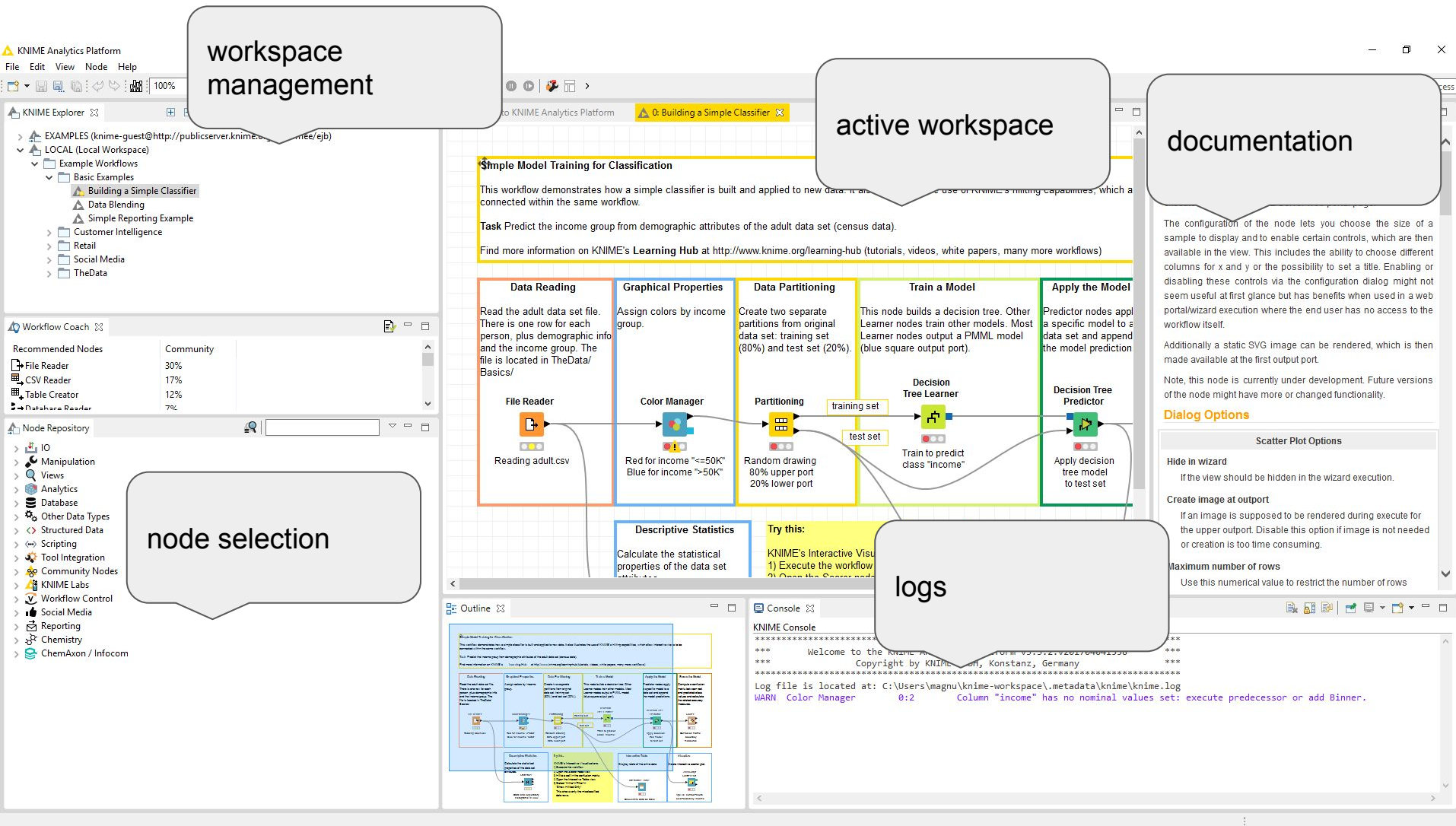
Extensions for analysis and visualization

Extensions for machine learning

...and much more



KNIME GUI



workspace management

active workspace

documentation

node selection

logs

Simple Model Training for Classification

This workflow demonstrates how a simple classifier is built and applied to new data. It uses KNIME's mining capabilities, which are connected within the same workflow.

Task Predict the income group from demographic attributes of the adult data set (census data).

Find more information on KNIME's **Learning Hub** at <http://www.knime.org/learning-hub> (tutorials, videos, white papers, many more workflows)

Data Reading
Read the adult data set file. There is one row for each person, plus demographic info and the income group. The file is located in TheData/ Basics/

Graphical Properties
Assign colors by income group.

Data Partitioning
Create two separate partitions from original data set: training set (80%) and test set (20%).

Train a Model
This node builds a decision tree. Other Learner nodes train other models. Most Learner nodes output a PMML model (blue square output port).

Apply the Model
Predictor nodes apply a specific model to a data set and append the model prediction

File Reader
Reading adult.csv

Color Manager
Red for income "<=50K"
Blue for income ">50K"

Partitioning
Random drawing
80% upper port
20% lower port

training set

test set

Decision Tree Learner
Train to predict class "income"

Decision Tree Predictor
Apply decision tree model to test set

Descriptive Statistics
Calculate the statistical properties of the data set

Try this:
KNIME's Interactive Visual Editor
1) Execute the workflow
2) Open the Scatter plot

Dialog Options

Scatter Plot Options

Hide in wizard
If the view should be hidden in the wizard execution.

Create image at output
If an image is supposed to be rendered during execute for the upper output. Disable this option if image is not needed or creation is too time consuming.

Maximum number of rows
Use this numerical value to restrict the number of rows

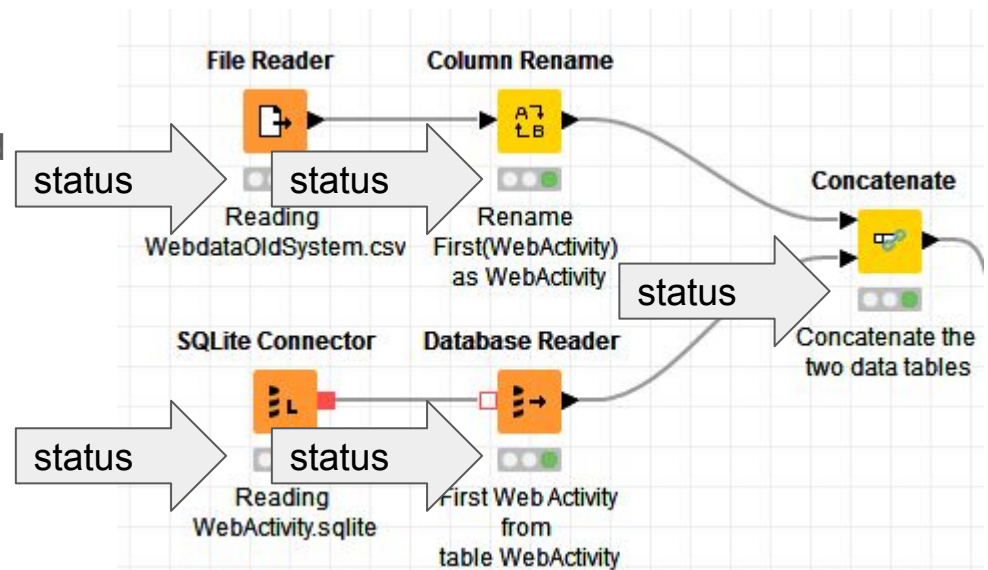
KNIME Console

```
*****
*** Welcome to the KNIME Analytics Platform ***
*** Copyright by KNIME AG, Konstanz, Germany ***
*****
Log file is located at: C:\Users\magnu\knime-workspace\metadata\knime\knime.log
WARN Color Manager 0:2 Column "income" has no nominal values set: execute predecessor or add Binner.
```


Nodes

Basic KNIME idea: nodes in a graph form a “data pipeline”

- Nodes for all kinds of functions
- Configuration is done using the GUI
- Directed links connect nodes to each other
- Processing follows the links
- Transparent processing status
 - Red: inactive and not configured
 - Yellow: configured, but not executed
 - Green: executed successfully



Example: “Data Blending”

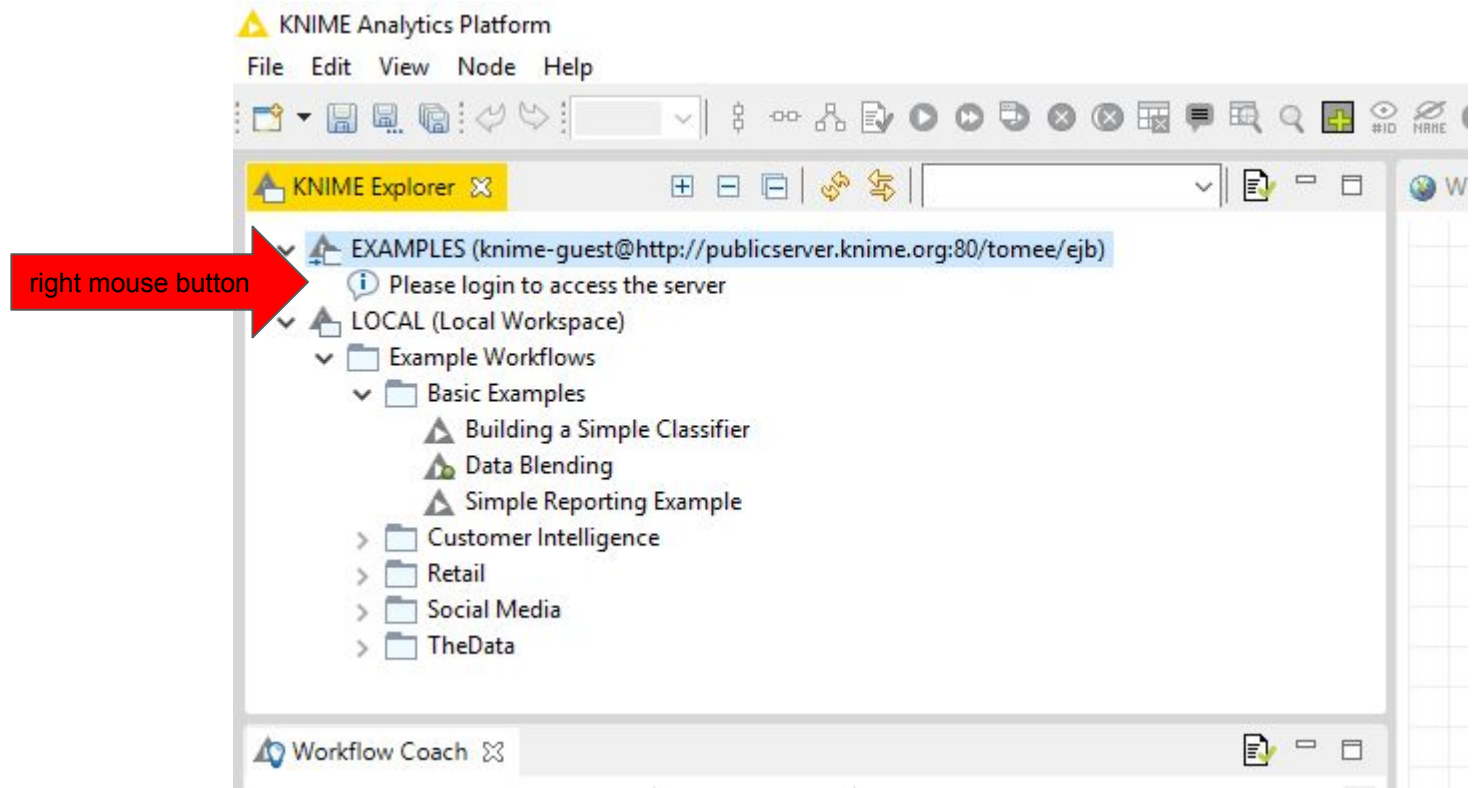
Local example workflow included in the KNIME distribution

KNIME://LOCAL/Example%20Workflows/Basic%20Examples/Data%20Blending

(Demo)

Example: a simple ETL workflow

Login to the EXAMPLES server of KNIME



Example: ETL Basics

KNIME://EXAMPLES/02_ETL_Data_Manipulation/00_Basic_Examples/02_ETL_Basics

(Demo)

My first workflows

Generate some data (Excel or LibreOffice)

- Columns author, title, year, publisher
- 3-4 sample datasets
- Save as both CSV file and Excel spreadsheet

In KNIME:

- Use a file node to open the CSV file
- Use a filter node to limit columns to title and year
- Use a filter node to select only those rows where year > 2000
- Use a file node to save the result as a CSV file

My first workflows

We prepared an XML file with data on the TOP 250 entries of IMDB.com (movies.xml)

goo.gl/S6y3ER

In KNIME:

- Preparation: Open the file, create a table from XML data
- Filter 1: Only title and year information
- Filter 2: All information on films from 2012
- Filter 3: What are the titles of the films from the years 2000-2010?
- Analysis 1: What genres are contained in the file?
- Analysis 2: Which director appears most often?

Example: Data visualization

Example data visualization.knwf in dropbox

(Demo)

knime://EXAMPLES/03_Visualization/02_JavaScript/04_Example_for_JS_Bar_Ch
art

(Demo)

My first visualization

Using movies.xml

In KNIME:

- Determine the countries, in which the movies take place and count their occurrence
- Use a pie chart to show the numbers
- Use a bar chart to show the numbers

Advanced exercise: What information is missing to visualize the countries as discs on a world map, with the size of the disc corresponding to the number?

Using external sources to enrich data

json demo.knwf in dropbox

(Demo)

Using external sources to enrich data

Using web APIs

KNIME://EXAMPLES/01_Data_Access/05_REST_Web_Services/01_Data_API_Using_REST_Nodes

(Demo)

My first enrichment

Have address, want geo-coordinates? Geocoding!

<https://developers.google.com/maps/documentation/geocoding/start>

In KNIME:

- Extend the list of countries to contain an URL for the google API
- Use the GET-node and query google
 - Warning: there is a rate control on the google APIs!
 - Use the node configuration to slow down the queries

Did we get correct coordinates for all countries? How did you check?

Example geo-visualization

KNIME://EXAMPLES/03_Visualization/04_Geolocation/04_Visualization_of_the_World_Cities_using_Open_Street_Map_(OSM)

(Demo)

Using geo-visualization

Again using movies.xml

In KNIME:

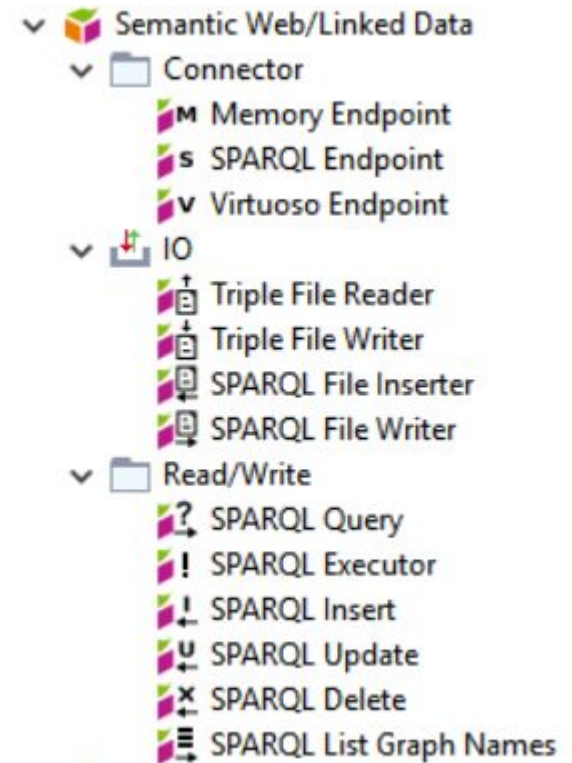
- visualize the countries that the movies are taking place in as discs on a world map, with the size of the disc corresponding to the number

Part 2: RDF and a real-world example

RDF in KNIME

Node group: Semantic Web/Linked Data

- Memory Endpoint as internal storage
- SPARQL Endpoint to read/write data
- IO is very basic:
 - Triples from tables to/from file
 - Triples from graphs to/from file
- Important table structure: subj, pred, obj
- Free SPARQL queries can be used to query for additional data.
- RDF data manipulation



Consuming RDF in KNIME

knime://EXAMPLES/08_Other_Analytics_Types/06_Semantic_Web/11_Semantic_Web_Analysis_Accessing_DBpedia

(DEMO)

Use the right tools!

knime://EXAMPLES/08_Other_Analytics_Types/06_Semantic_Web/10_Using_Semantic_Web_to_generate_Simpsons_TagCloud

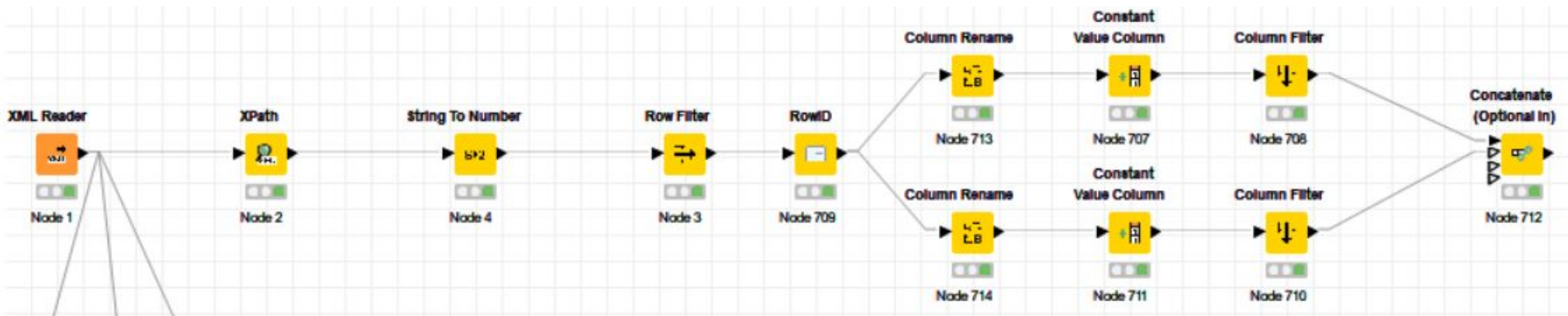
(DEMO)

TODO: Add link to file (fixed version)

- The demo needs some fixes to actually get the word cloud.
- Most part of the workflow is about trimming and filtering RDF strings (e.g., get rid of the xsd types).
- It is great that it is possible to do this in KNIME, but the creation of a proper CSV file outside KNIME might be easier.

Producing RDF in KNIME

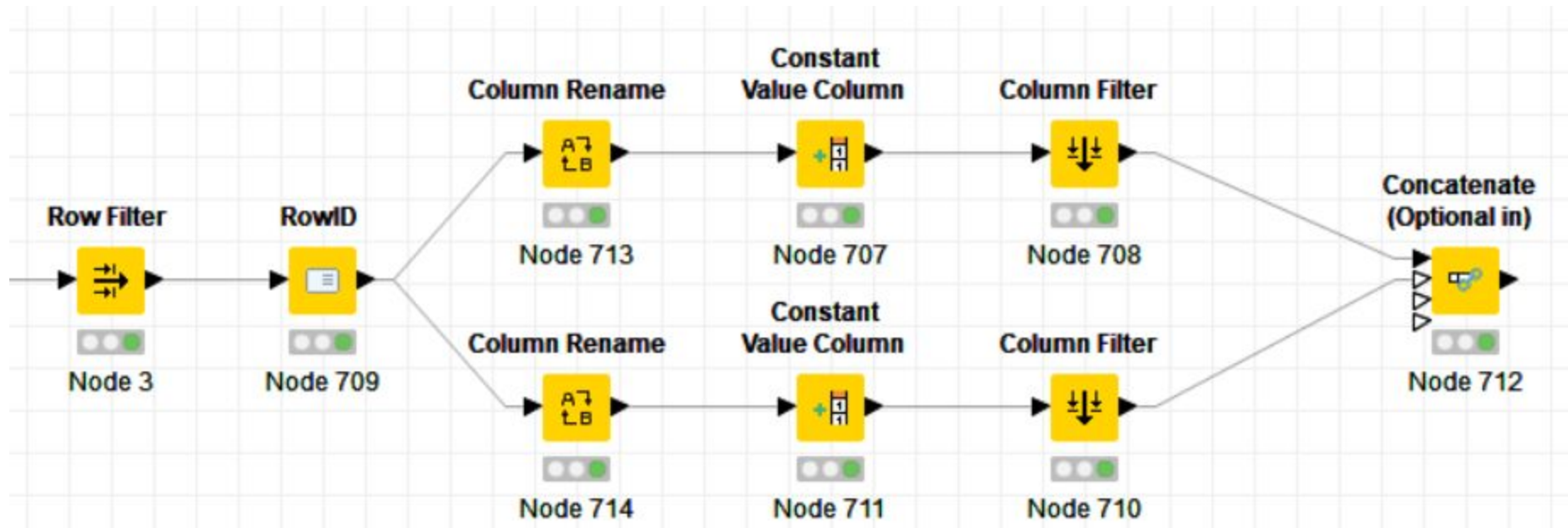
Use your movie workflow to produce triples for title and year of a movie.



Approach:

1. Create a column **subj** containing the subject of each row
2. For each predicate to be written:
 - a. rename the column containing the value to **obj**.
 - b. add a column **pred** containing the desired property.
 - c. filter to keep only the columns **sub**, **pred**, **obj**.
3. Concatenate the resulting tables (or write them to a triple store)

Producing RDF in KNIME



(DEMO) TODO: Add link to file

Again the question: Is creating triples from CSV outside KNIME easier?

Case Study: Metadata enrichment

All files: goo.gl/S6y3ER

Input

- A table of library holdings:
 - Item number and barcode to identify an item.
 - PPN to identify the manifestation of each item.
 - call number (Signatur) and location (Sigel) for each item.
- No metadata!
- Goal: Get classification data (RVK) for each item.

Table "Liste_PPN-ExNr_HSHN-libre.csv" - Rows: 118120						Spec - Columns: 5	Properties	Flow Variables
Row ID	\$ PPN	Exempl...	\$ Signatur	\$ Barcode	\$ Sigel			
Row0	300896	123945	535.6 WYS	HC251 6	HN			
Row1	343145	111179	03 ALLG BRO	HT227 3	HN			
Row2	343153	111180	03 ALLG BRO	HT228 8	HN			
Row3	343161	111181	03 ALLG BRO	HT229 2	HN			
Row4	00034317X	111182	03 ALLG BRO	HT230 5	HN			
Row5	343188	111183	03 ALLG BRO	HT231 X	HN			
Row6	343196	111184	03 ALLG BRO	HT232 4	HN			

Output

1. Group per PPN
2. Add Metadata from SWB union catalog.
3. For entries without RVK: Add RVKs from BVB.
4. Modify result table to match required CSV format.

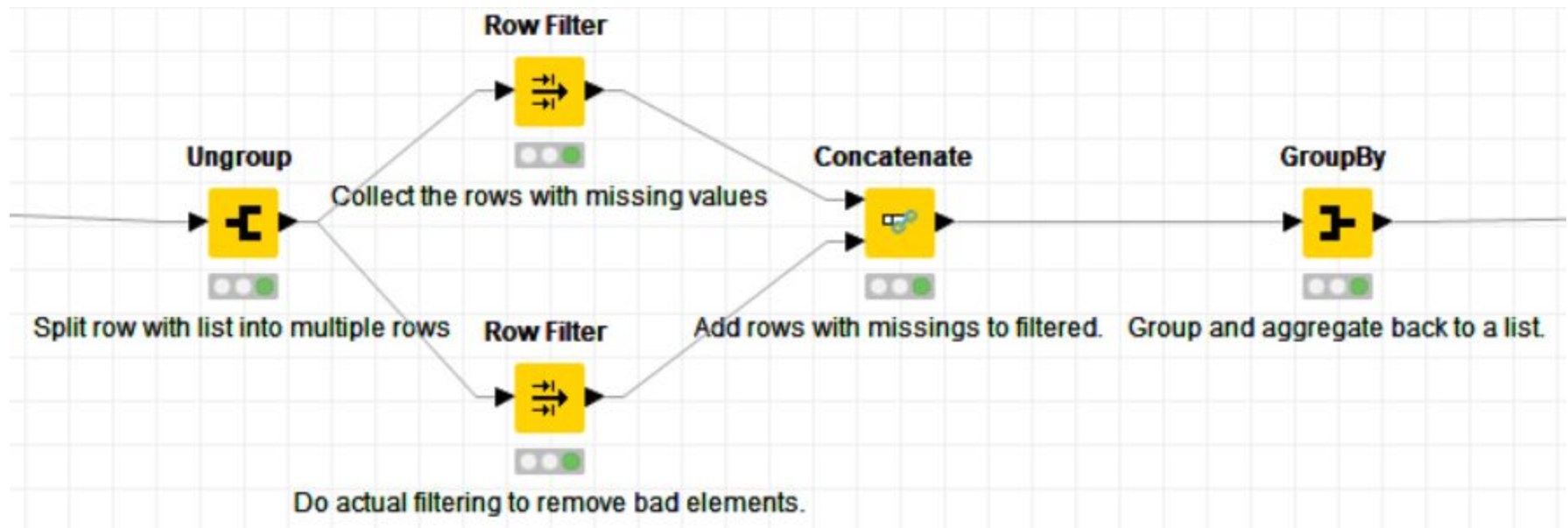
(This workflow ends here!)

5. Data is then processed in another application to do manual quality checks and add additional RVK.
6. Afterwards, there is another workflow to ungroup back to item level.

Table "default" - Rows: 812																	Spec - Columns: 17	Properties	Flow Variables
Row ID	S PPN	S language	S author	S title	S year	S edition	S publisher	S subjects	S rvk	S isbn	S editor	S Exempl...	S Sigel	S Column 0	S series	S Signatur	S Barcode		
Row0	000300896	ger	Wyszecki, Günter	Farbsysteme	1960	?	Musterschmidt	Farbsystem;Farbenlehre	CP 2500;AP 14800;UH 7400	?		123945	HN	?	?	535.6 WYS	HC251 6		
Row1	000343145	ger		Der grosse ...	1952	16., völlig n...	Brockhaus		AE 11000	?		111179	HN	?	?	03 ALLG BRO	HT227 3		
Row2	000343153	ger		Der große B...	1953	16., völlig n...	Brockhaus		AE 11000;AE 11983;AE 1...	?		111180	HN	?	?	03 ALLG BRO	HT228 8		
Row3	000343161	ger		Der große B...	1953	16., völlig n...	Brockhaus		AE 11000;AE 11983;AE 1...	?		111181	HN	?	?	03 ALLG BRO	HT229 2		
Row4	00034317X	ger		Der große B...	1954	16., völlig n...	Brockhaus		AE 11000;AE 11983;AE 1...	?		111182	HN	?	?	03 ALLG BRO	HT230 5		
Row5	000343188	ger		Der große B...	1954	16., völlig n...	Brockhaus		AE 11000;AE 11983;AE 1...	?		111183	HN	?	?	03 ALLG BRO	HT231 X		
Row6	000343196	ger		Der große B...	1955	16., völlig n...	Brockhaus		AE 11000;AE 11983;AE 1...	?		111184	HN	?	?	03 ALLG BRO	HT232 4		

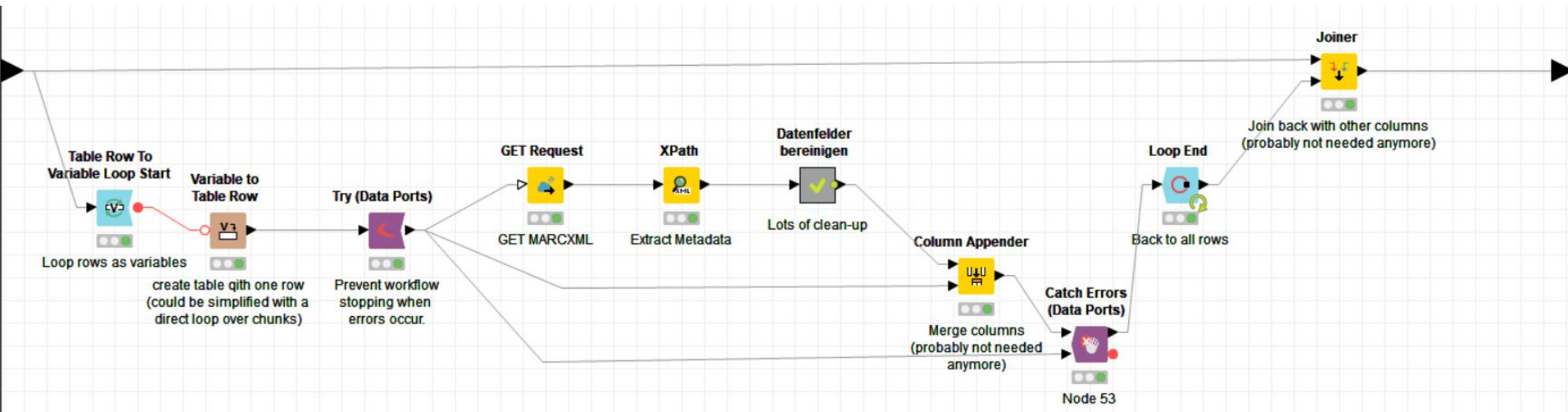
Group/Ungroup

- A typical step is to switch the levels of aggregation to make use of KNIME operators.
- Here is an example where a row filter is used to actually filter elements of a list element (“Remove Non-RVK” in the workflow):



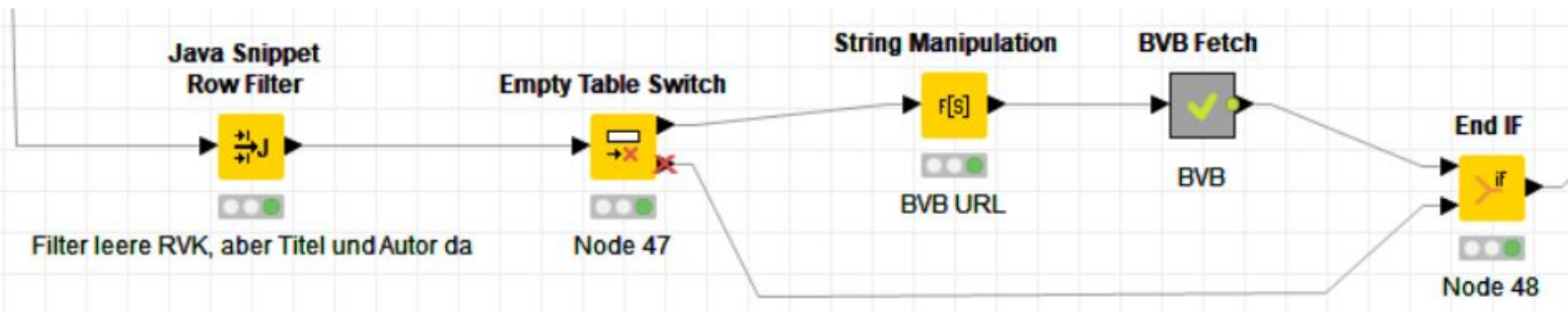
Looping over rows

- When the workflow was created, the GET operator could retrieve data for a whole table, but if one request failed, the whole operator failed and the workflow stopped.
- Moreover, the GET operator did not pass through other columns than the URL columns.
- Both problems are dealt with in the SWB fetch part:
 - A loop is created over all rows.
 - The resulting table with (additional) columns is joined with the original table.



Deal with empty results

- Sometimes whole parts of the workflow can be skipped.
- Example: We filter for all rows who have no RVK but have author and title information available (as we need this to search for matching records).
- Depending on the (sampled) input data, there might be no rows who qualify. Then we just bypass the whole RVK enrichment part of the workflow.



Part 3: Machine Learning

Finding patterns in data: scatterplot analysis

Example data visualization.knwf in dropbox

(Demo)

Machine learning: classification

Basic process

- Get and arrange data
- Split data into training and test data
- Learn model using training data
- Test model using test data
- Score results

→ Apply learned and tested model to new data

Learning and testing a model

machine learning.knwf in Dropbox

(Demo)