(Meta-)Datamanagement with KNIME

SWIB 2017 Workshop

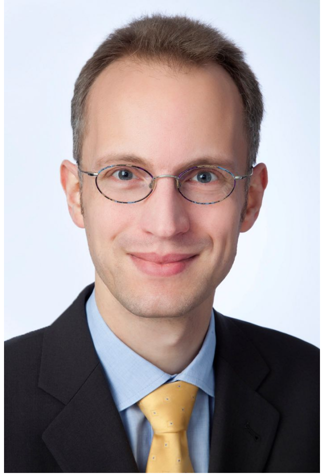
Your mentors

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

JudaicaLink

Funding by

DFG

Consortium

Universitätsbibliothek
J.C. Senckenberg UB

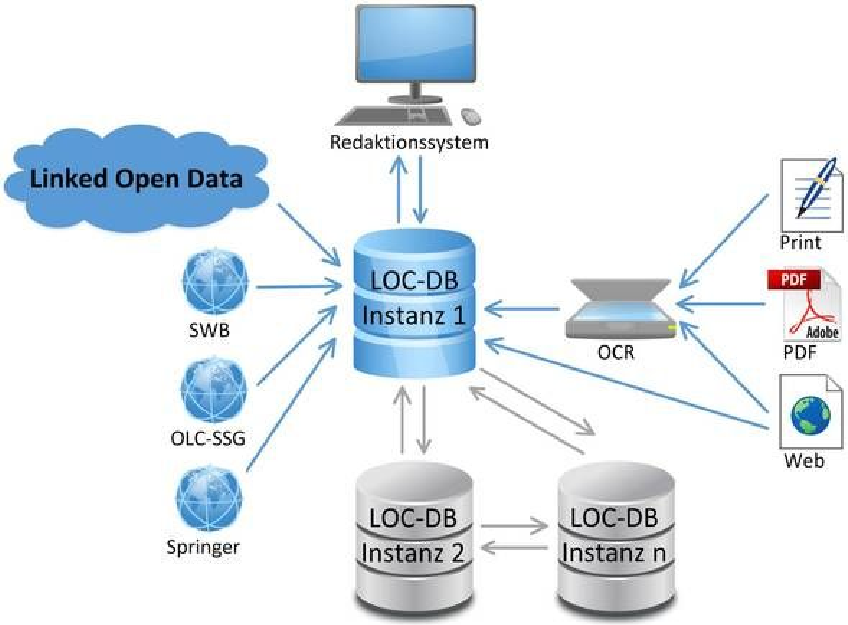
HOCHSCHULE
DER MEDIEN

Current projects with data focus

Linked Open Citation Database

Challenges:

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



Funding by DFG

Consortium

UB MANNHEIM

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

HOCHSCHULE DER MEDIEN

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

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

HOCHSCHULE DER MEDIEN

UNIVERSITÄTS BIBLIOTHEK ubl LEIPZIG

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

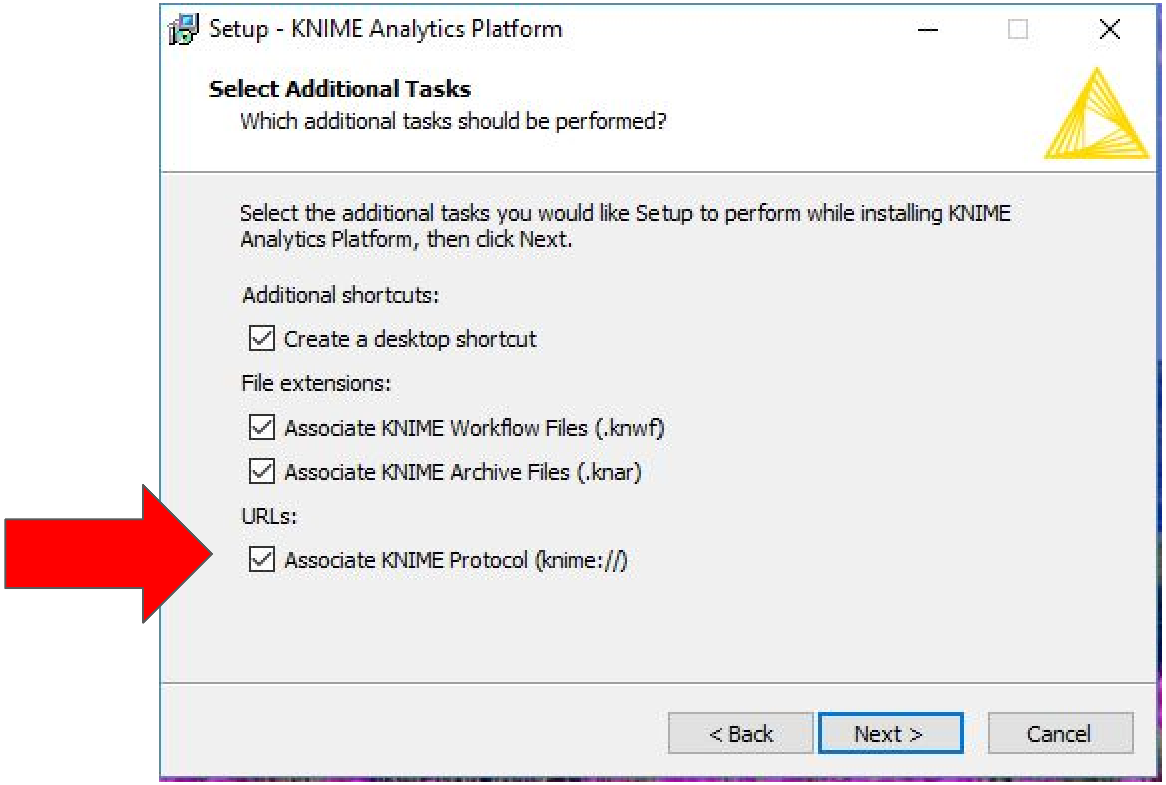
- Part 2: Real-world uses (~ 1 hr)
  - Using the RDF nodes to read and output linked data
  - Creating an enriched bibliographic dataset
    - Fixing errors in the input dataset
    - Downloading bibliographic data as XML from the web
    - Enriching with classification data from a different source
    - Data output

- Part 3: 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



Setup - KNIME Analytics Platform

Select Additional Tasks

Which additional tasks should be performed?

Select the additional tasks you would like Setup to perform while installing KNIME Analytics Platform, then click Next.

Additional shortcuts:

☑ Create a desktop shortcut

File extensions:

☑ Associate KNIME Workflow Files (.knwf)

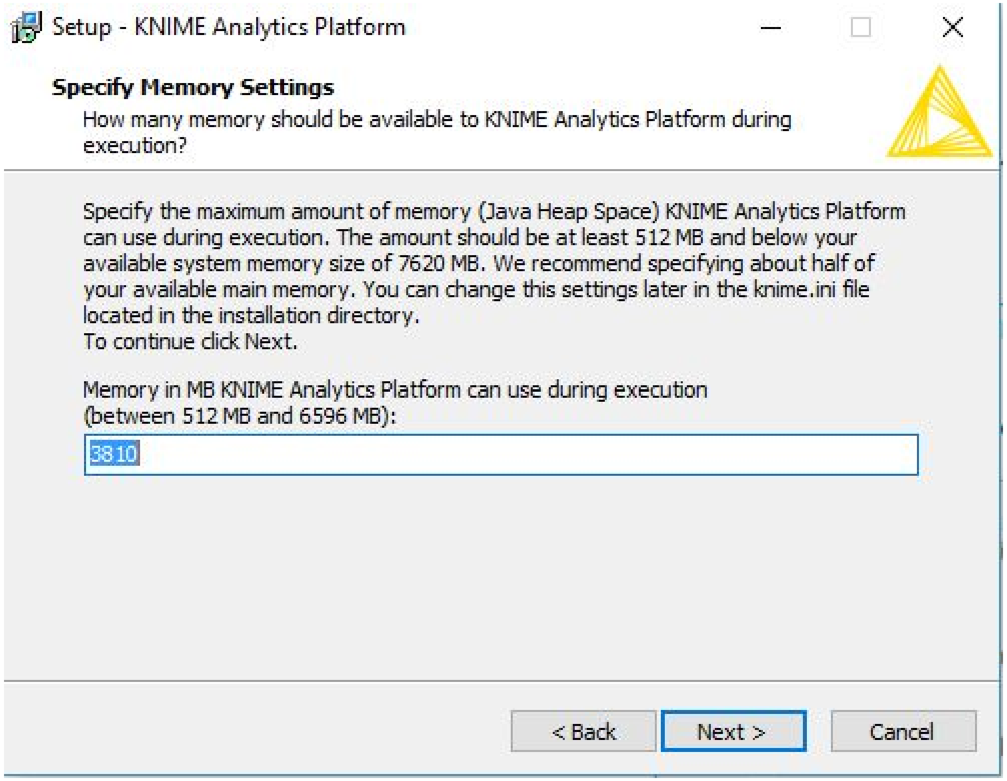
☑ Associate KNIME Archive Files (.knar)

URLs:

☑ Associate KNIME Protocol (knime://)

Installation

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



Setup - KNIME Analytics Platform

Specify Memory Settings

How many memory should be available to KNIME Analytics Platform during execution?

Specify the maximum amount of memory (Java Heap Space) KNIME Analytics Platform can use during execution. The amount should be at least 512 MB and below your available system memory size of 7620 MB. We recommend specifying about half of your available main memory. You can change this settings later in the knime.ini file located in the installation directory.
To continue click Next.

Memory in MB KNIME Analytics Platform can use during execution (between 512 MB and 6596 MB):

3810

Why KNIME?

Possible alternative: Develop own software tools?

Upside: Maximum flexibility

Downsides:

- Very complex, coding knowledge a necessity
- Own code cat 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

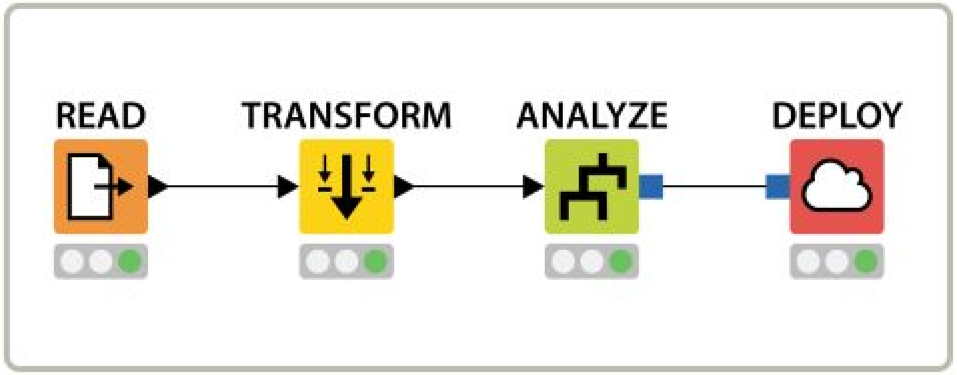
Classic data workflow: Extract, Transform, Load (ETL)

KNIME adds:

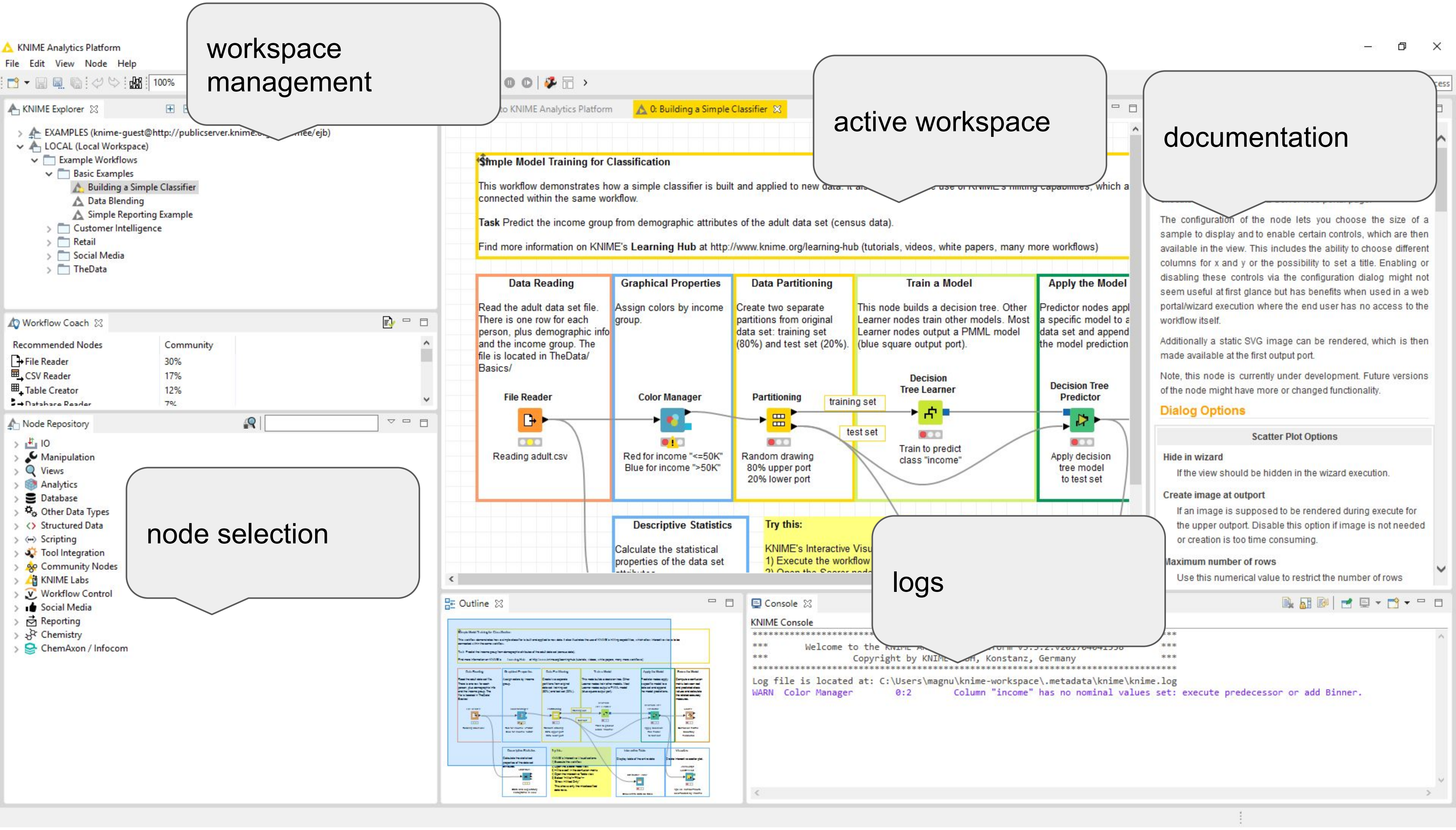
Extensions for analysis and visualization

Extensions for machine learning

...and much more



KNIME GUI



workspace management

active workspace

documentation

node selection

logs

Recommended Nodes	Community
File Reader	30%
CSV Reader	17%
Table Creator	12%
Database Reader	7%

Simple Model Training for Classification

This workflow demonstrates how a simple classifier is built and applied to new data. It als... ...e use of KNIME's ... capabilities, which a... 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/

File Reader: Reading adult.csv

Graphical Properties

Assign colors by income group.

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

Data Partitioning

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

Partitioning: Random drawing 80% upper port 20% lower port

training set

test set

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).

Decision Tree Learner: Train to predict class "income"

Apply the Model

Predictor nodes appl... a specific model to a... data set and append the model prediction

Decision Tree Predictor: Apply decision tree model to test set

Descriptive Statistics

Calculate the statistical properties of the data set attributes

Try this:

KNIME's Interactive Visu...
1) Execute the workflow...
2) Open the Scorer node...

The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself.

Additionally a static SVG image can be rendered, which is then made available at the first output port.

Note, this node is currently under development. Future versions of the node might have more or changed functionality.

Dialog Options

Scatter Plot Options

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

Create image at outport
If an image is supposed to be rendered during execute for the upper outport. 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

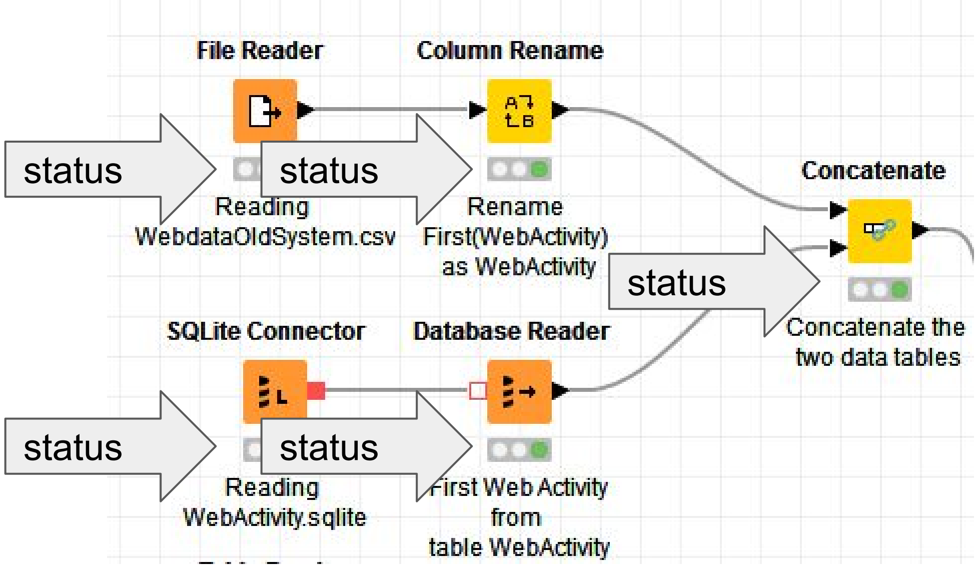
KNIME Console

```
***********************                                          ***
***       Welcome to the KNIME A...       ...form v3.3.2.v201704041338     ***
***                  Copyright by KNIME ...bH, 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


File Reader
Column Rename
status
status
Reading WebdataOldSystem.csv
Rename First(WebActivity) as WebActivity
Concatenate
status
Concatenate the two data tables
SQLite Connector
Database Reader
status
status
Reading WebActivity.sqlite
First Web Activity from table WebActivity

Example: “Data Blending”

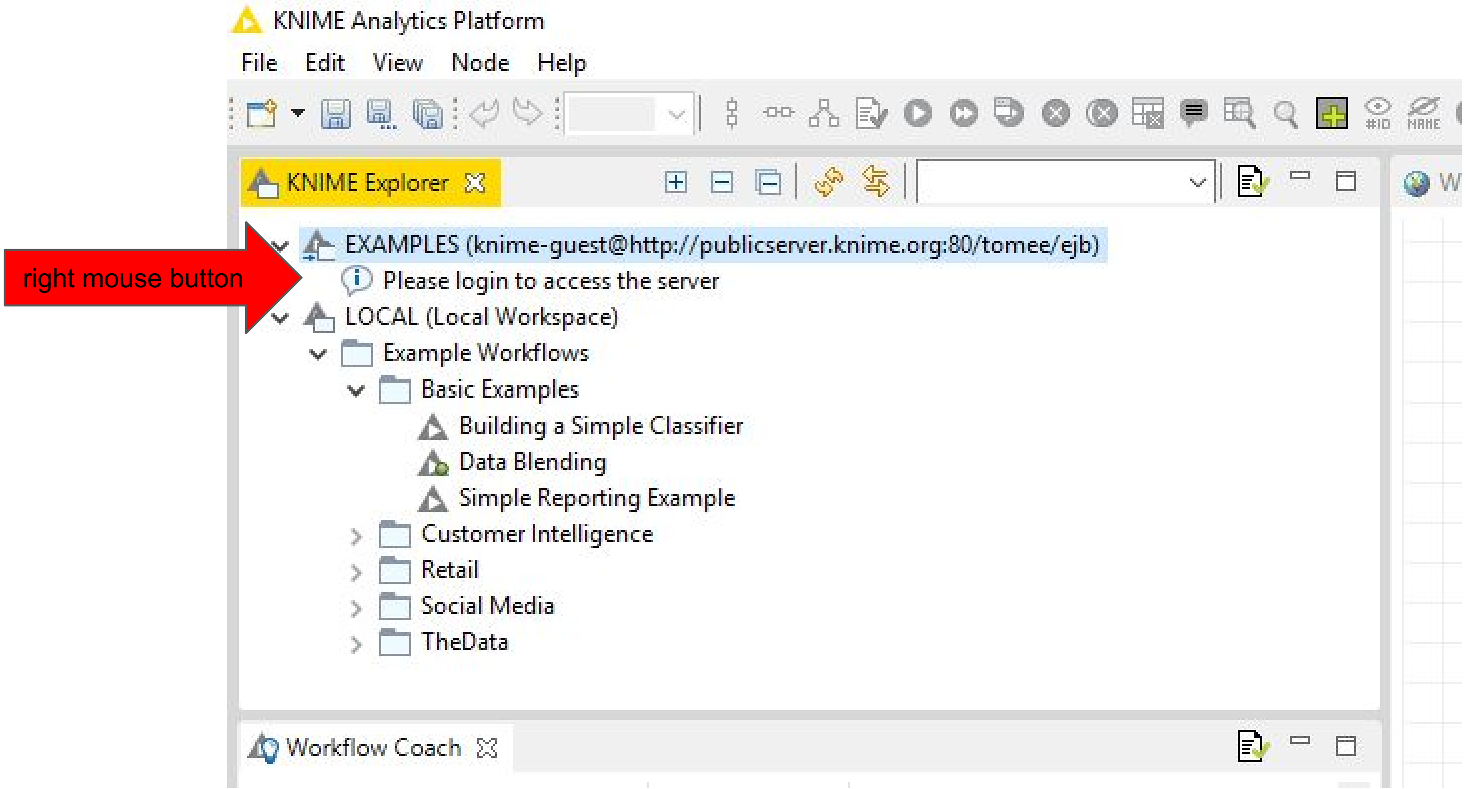
Local example workflow included in the KNIME distribution

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

(Demo)

Example: a simple ETL workflow

Login to the EXAMPLES server of KNIME



EXAMPLES (knime-guest@http://publicserver.knime.org:80/tomee/ejb)
Please login to access the server
LOCAL (Local Workspace)
Example Workflows
Basic Examples
Building a Simple Classifier
Data Blending
Simple Reporting Example
Customer Intelligence
Retail
Social Media
TheData

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)

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

(Demo)

knime://EXAMPLES/03_Visualization/02_JavaScript/04_Example_for_JS_Bar_Chart

(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

(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

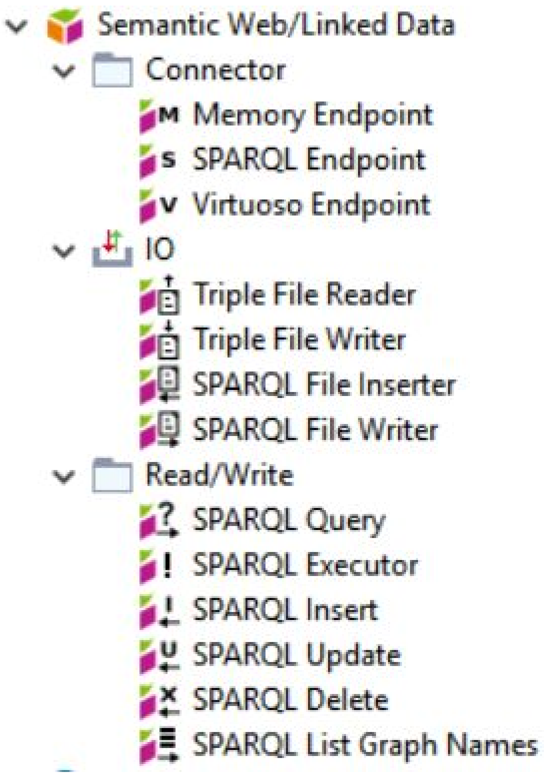
RDF in KNIME

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 graps to/from file
- Important table structure: subj, pred, obj

- Free SPARQL queries can be used to query for additional data.
- RDF data manipulation



Semantic Web/Linked Data
- Connector
  - Memory Endpoint
  - SPARQL Endpoint
  - Virtuoso Endpoint
- IO
  - Triple File Reader
  - Triple File Writer
  - SPARQL File Inserter
  - SPARQL File Writer
- Read/Write
  - SPARQL Query
  - SPARQL Executor
  - SPARQL Insert
  - SPARQL Update
  - SPARQL Delete
  - SPARQL List Graph Names

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

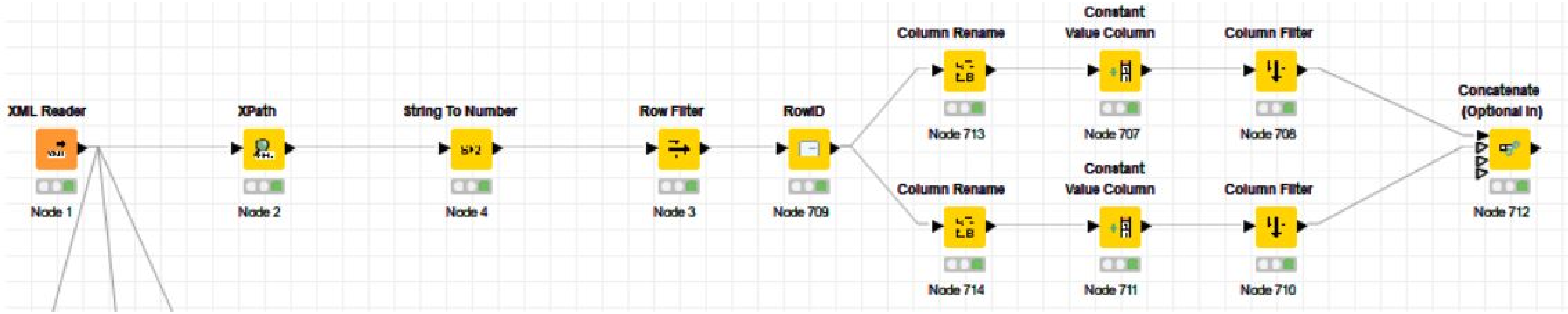
(DEMO)

Fixed version:
10_Using_Semantic_Web_to_generate_Simpsons_TagCloud_FIXED.knwf

- 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

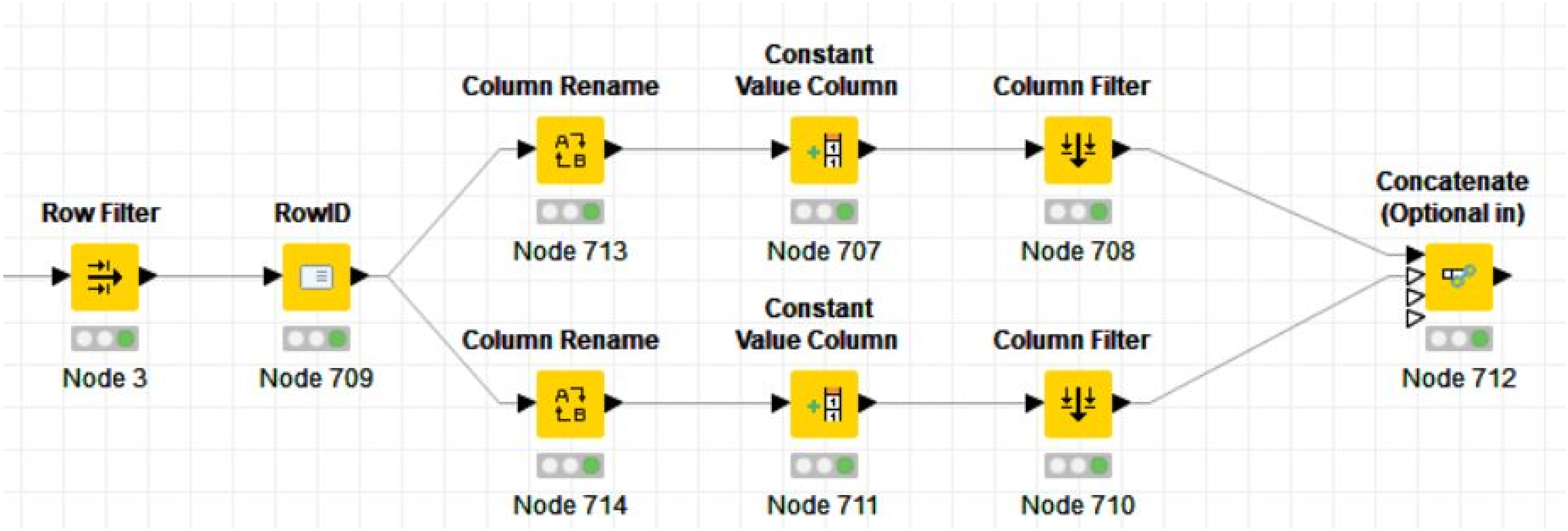
Use your movie workflow to produce triples for title and year of a movie.



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)

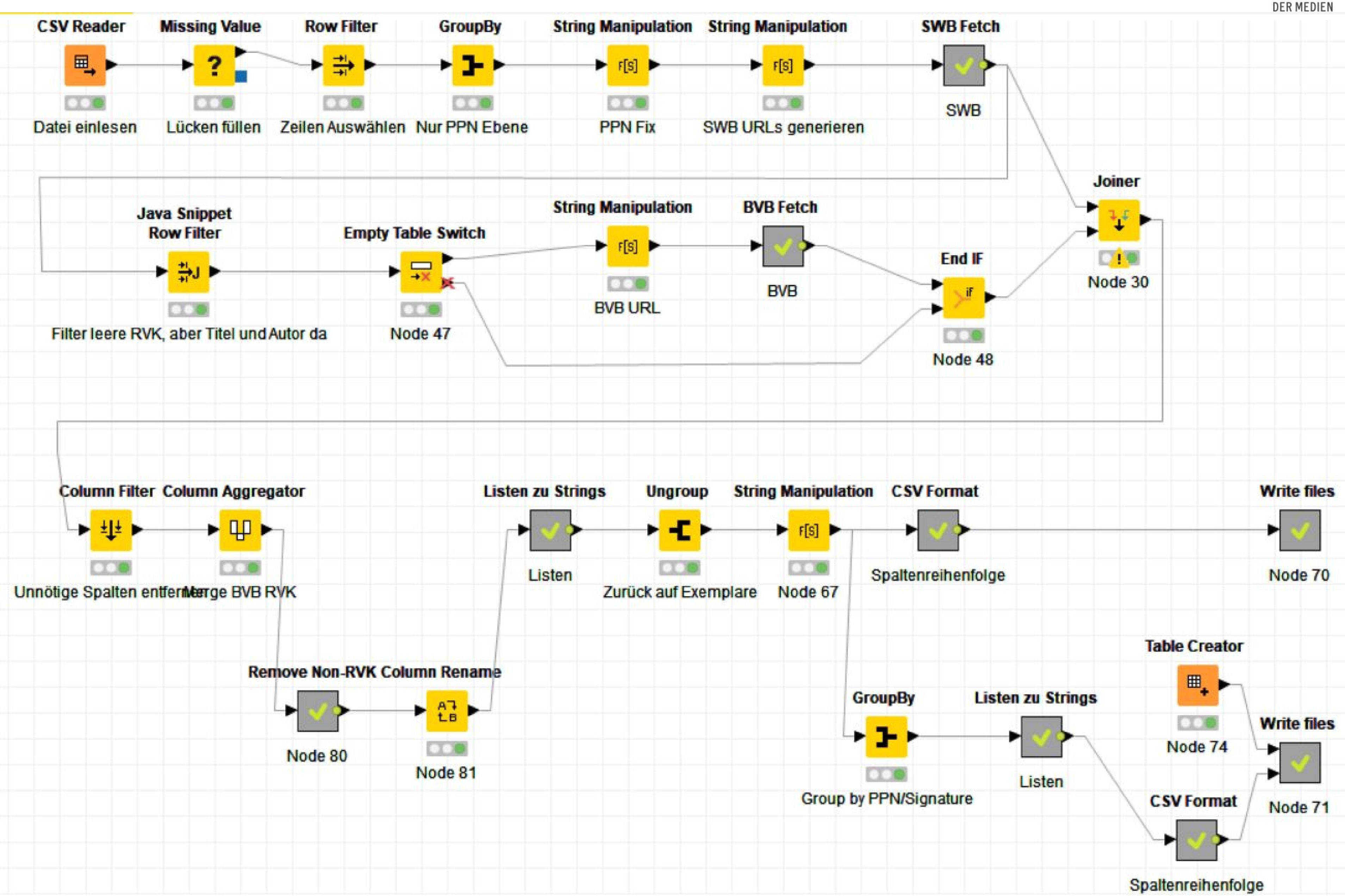
Producing RDF in KNIME



(DEMO) Kurs_Movies_Filter_With_RDF.knwf

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

Case Study: Metadata enrichment

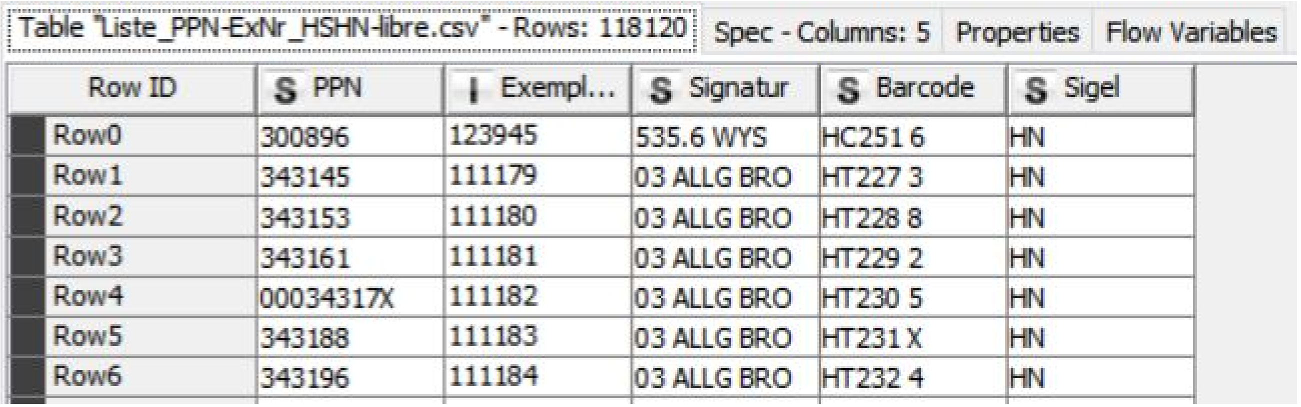

CSV Reader
Datei einlesen
Missing Value
Lücken füllen
Row Filter
Zeilen Auswählen
GroupBy
Nur PPN Ebene
String Manipulation
PPN Fix
String Manipulation
SWB URLs generieren
SWB Fetch
SWB
Joiner
Node 30
Java Snippet
Row Filter
Filter leere RVK, aber Titel und Autor da
Empty Table Switch
Node 47
String Manipulation
BVB URL
BVB Fetch
BVB
End IF
Node 48
Column Filter
Unnötige Spalten entfernen
Column Aggregator
Merge BVB RVK
Remove Non-RVK
Node 80
Column Rename
Node 81
Listen zu Strings
Listen
Ungroup
Zurück auf Exemplare
String Manipulation
Node 67
CSV Format
Spaltenreihenfolge
Write files
Node 70
GroupBy
Group by PPN/Signature
Listen zu Strings
Listen
Table Creator
Node 74
CSV Format
Spaltenreihenfolge
Write files
Node 71

Input

- A table of library holdings:
  - Item number and barcode to 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.



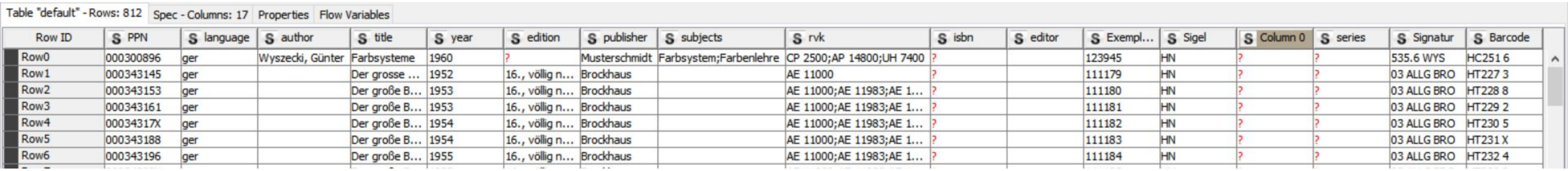
Row ID	PPN	Exempl...	Signatur	Barcode	Sigel
Row0	300896	123945	535.6 WYS	HC251 6	HN
Row1	343145	111179	03 ALLG BRO	HT227 3	HN
Row2	343153	111180	03 ALLG BRO	HT228 8	HN
Row3	343161	111181	03 ALLG BRO	HT229 2	HN
Row4	00034317X	111182	03 ALLG BRO	HT230 5	HN
Row5	343188	111183	03 ALLG BRO	HT231 X	HN
Row6	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!)

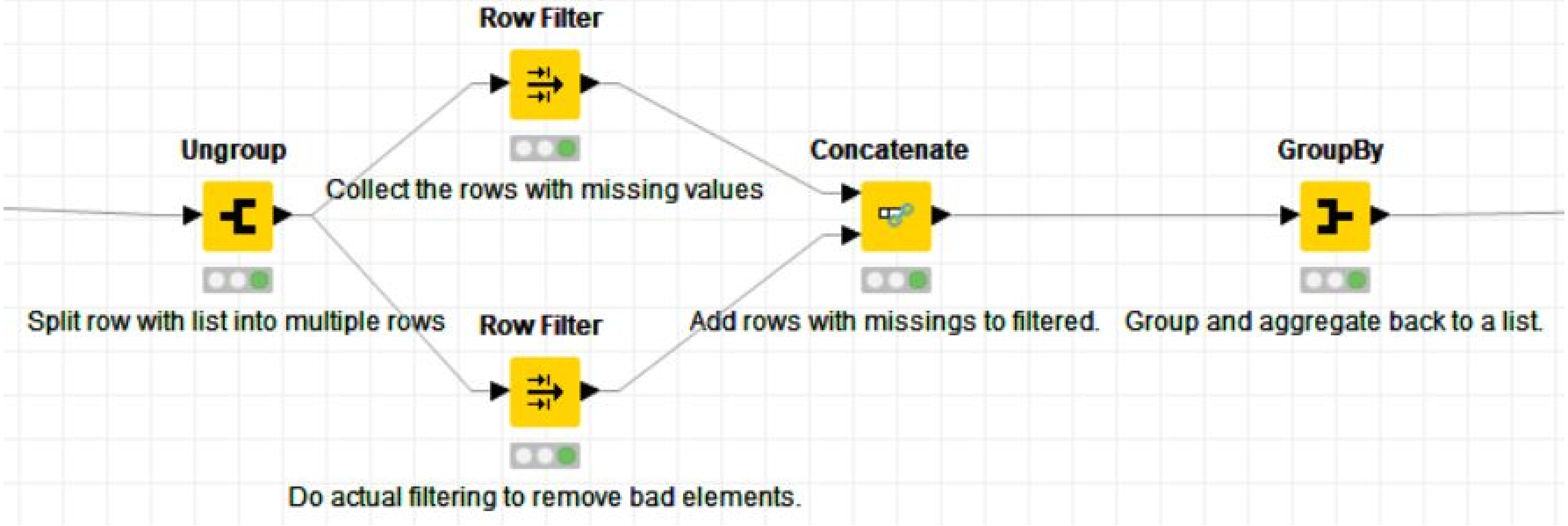
5. Data is then processed in another application to do manual quality checks and add additional RVK.
6. Afterwards, there is another workflow to ungroup back to item level.



Row ID	PPN	language	author	title	year	edition	publisher	subjects	rvk	isbn	editor	Exempl...	Sigel	Column 0	series	Signatur	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):


Row Filter
Ungroup
Collect the rows with missing values
Concatenate
GroupBy
Split row with list into multiple rows
Row Filter
Add rows with missings to filtered.
Group and aggregate back to a list.
Do actual filtering to remove bad elements.

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.

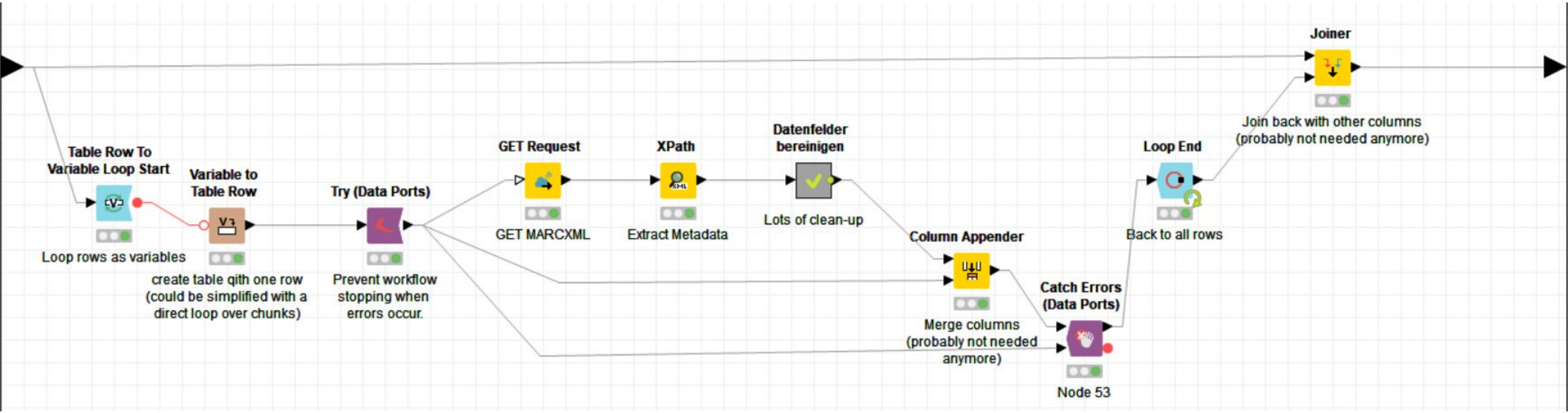
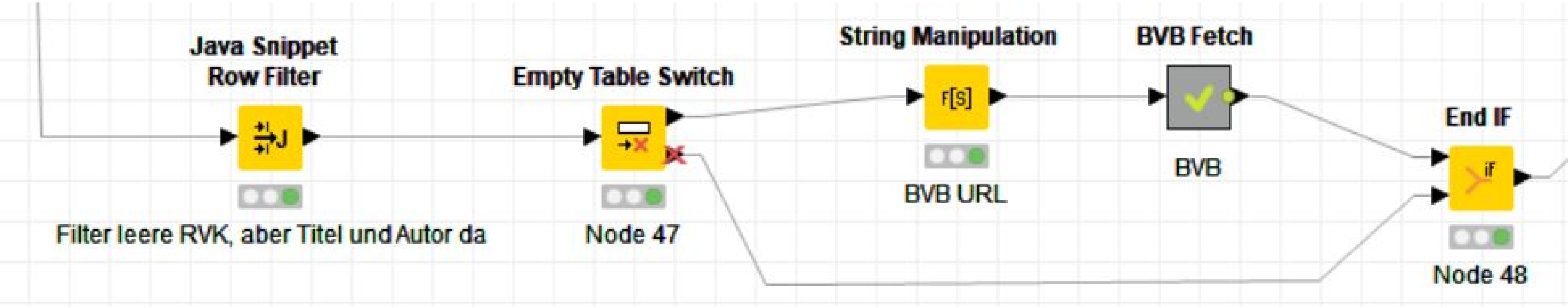


Table Row To Variable Loop Start
Loop rows as variables
Variable to Table Row
create table qith one row (could be simplified with a direct loop over chunks)
Try (Data Ports)
Prevent workflow stopping when errors occur.
GET Request
GET MARCXML
XPath
Extract Metadata
Datenfelder bereinigen
Lots of clean-up
Column Appender
Merge columns (probably not needed anymore)
Catch Errors (Data Ports)
Node 53
Loop End
Back to all rows
Joiner
Join back with other columns (probably not needed anymore)

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.


Java Snippet Row Filter
Filter leere RVK, aber Titel und Autor da
Empty Table Switch
Node 47
String Manipulation
BVB URL
BVB Fetch
BVB
End IF
Node 48