Using Kepler Workflows In Ecology

Claas-Thido Pfaff Karin Nadrowski Christian Wirth
Universität Leipzig, Spezielle Botanik und Funktionelle Biodiversität, Leipzig, DE

UNIVERSITÄT LEIPZIG

Questions

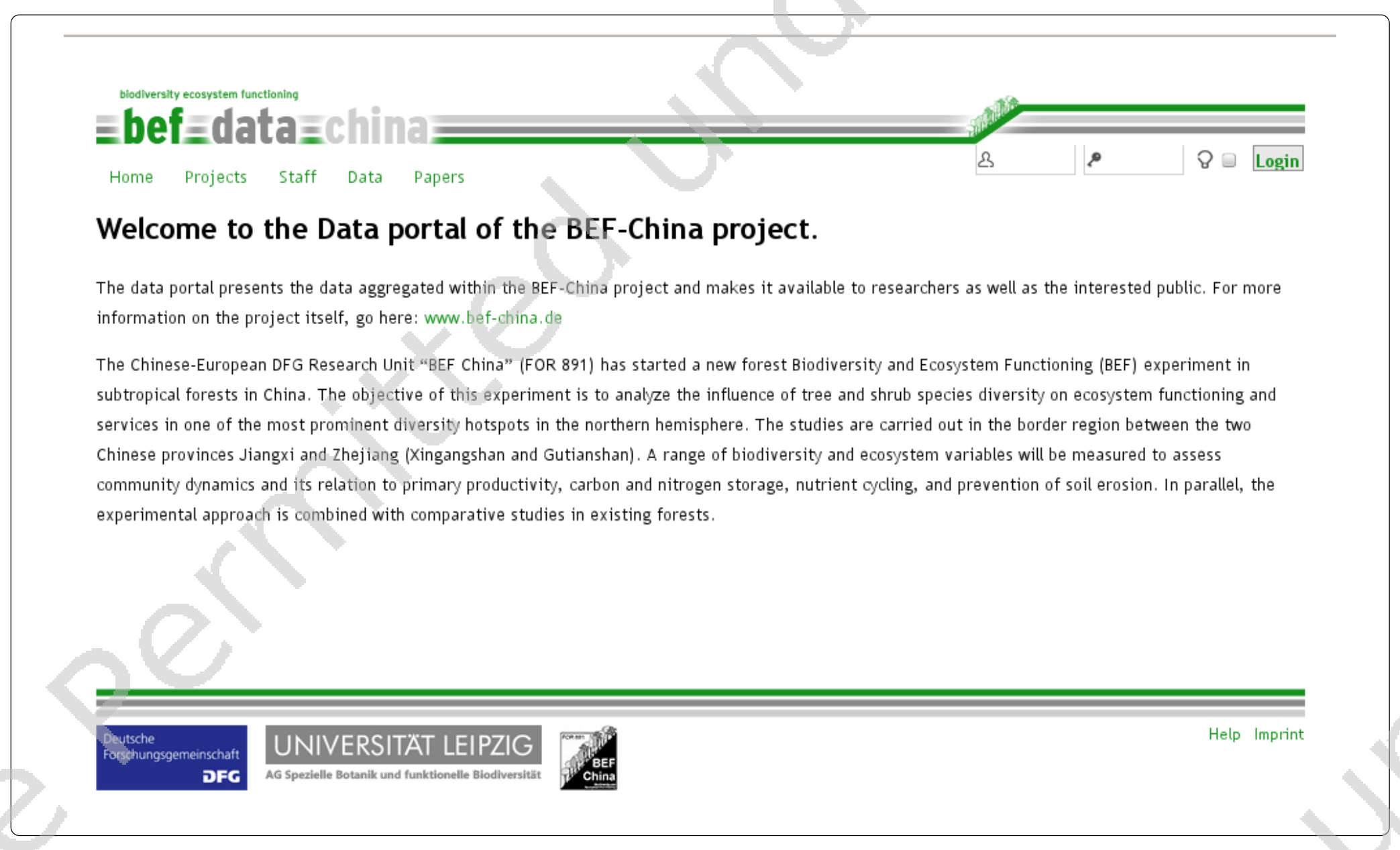
1. Which analysis tasks occur in a workflow and how diverse are they?
2. Is the level of complexity stable throughout a scientific workflow?
3. Which data manipulation steps show highest complexity?
4. Which data manipulation steps are potential candidates for automatization?

Introduction

Ecology has become a more collaborative, interdisciplinary, and data intensive science. We have access to a growing global data pool filled mainly with data from extensive long term research projects using sensor or satellite networks, but also with a growing amount of data from smaller projects spread all over the world. Especially the potential of small idiosyncratic datasets is still underutilized. They can be integrated in meta analyses to address questions to a deeper level of detail and on a greater temporal and spatial scale. These integration processes are very time and labour intensive which is related to the high heterogeneity and missing additional information (meta data) about the small datasets. We need effective mechanisms which assist researchers to maintain and store their metadata and help them to deal with the high complexity of the integration of ecological data.

We show how the BEFdata (github, `http://tinyurl.com/cf58j2s`) portal (Fig. 1) can be used to integrate small datasets in a workflow constructed by the Kepler software (Fig. 2) (Kepler, `http://tinyurl.com/9zg8u35`). We use an ongoing carbon meta analysis that estimates the carbon stocks of the comparative study sites of the BEF-China Experiment (DFG, FOR 891, Biodiversity - Ecosystem Functioning, `http://tinyurl.com/8kzy4y5`).

We created means to measure workflows and analysed the created Kepler workflow about the carbon stock analysis. This workflow analysis is a first attempt to measure scientific workflow processes with the help of a workflow software for a better understanding of how data is manipulated during the preparation of data and during an ongoing scientific analysis. We conclude with an outlook in how knowledge organization systems (controlled vocabularies, thesauri, ontologies) may be helpful in automating common tasks in ecological analyses.

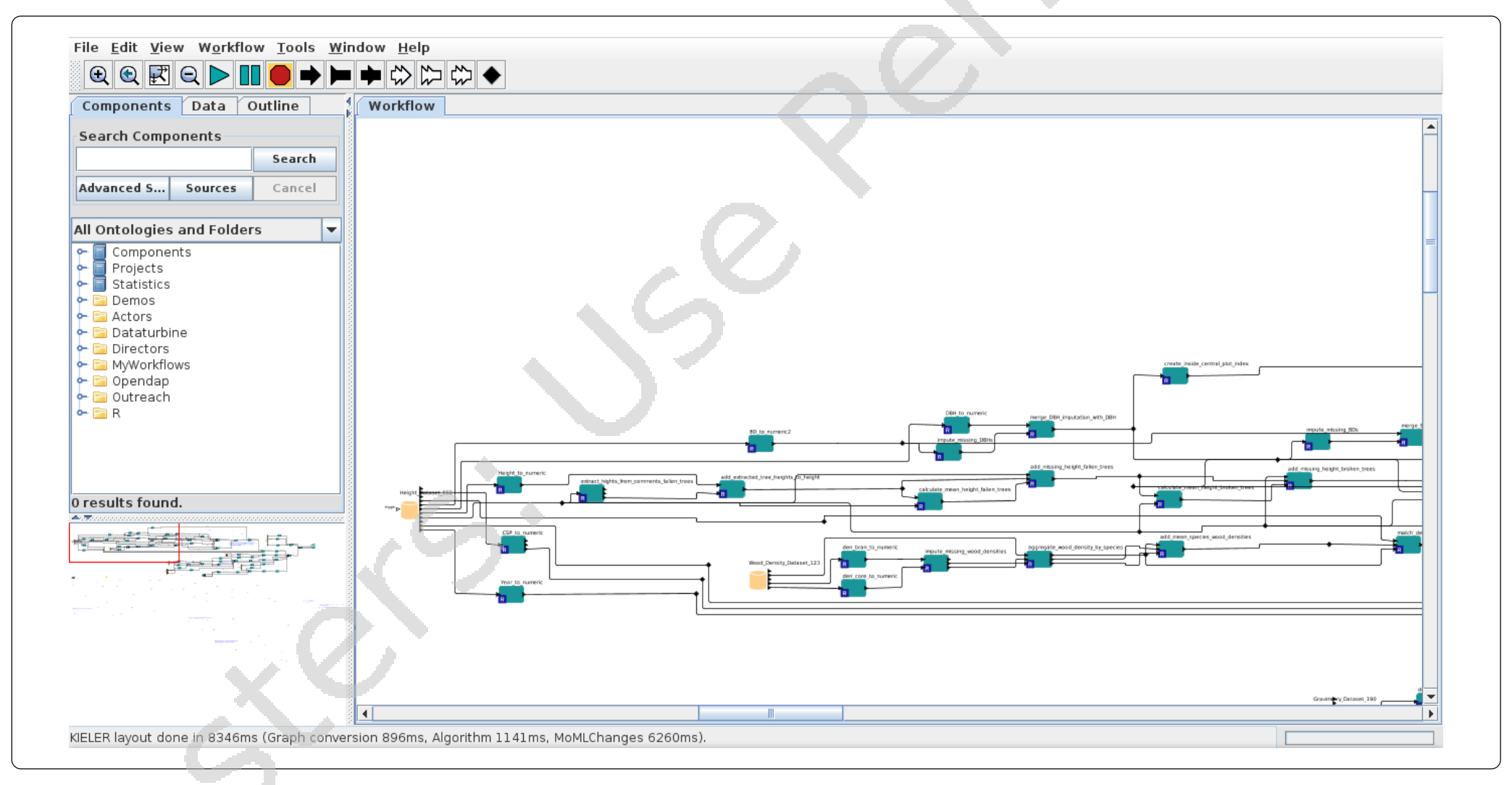

biodiversity ecosystem functioning
bef data china
Home
Projects
Staff
Data
Papers
Login
Help
Imprint
Deutsche Forschungsgemeinschaft DFG
UNIVERSITÄT LEIPZIG
AG Spezielle Botanik und funktionelle Biodiversität
BEF China

Welcome to the Data portal of the BEF-China project.

The data portal presents the data aggregated within the BEF-China project and makes it available to researchers as well as the interested public. For more information on the project itself, go here: www.bef-china.de

The Chinese-European DFG Research Unit "BEF China" (FOR 891) has started a new forest Biodiversity and Ecosystem Functioning (BEF) experiment in subtropical forests in China. The objective of this experiment is to analyze the influence of tree and shrub species diversity on ecosystem functioning and services in one of the most prominent diversity hotspots in the northern hemisphere. The studies are carried out in the border region between the two Chinese provinces Jiangxi and Zhejiang (Xingangshan and Gutianshan). A range of biodiversity and ecosystem variables will be measured to assess community dynamics and its relation to primary productivity, carbon and nitrogen storage, nutrient cycling, and prevention of soil erosion. In parallel, the experimental approach is combined with comparative studies in existing forests.

Figure 1: *The BEFdata portal was used to extract data for the creation of the workflow about the carbon stocks in the BEF-China experiment.*



0 results found.

KIELER layout done in 8346ms (Graph conversion 896ms, Algorithm 1141ms, MoMLChanges 6260ms).

Figure 2: *The Kepler Workflow software shown with a part of the created workflow of the carbon stock analysis.*

Methods

All primary data and the corresponding Ecological Metadata Language (EML) files were downloaded form the BEFdata portal. The primary data was imported into Kepler by the use of the `EML 2 Dataset` actor which reads all information about the data from the EML, creates output ports for each column in the primary data and sets the data type for the column automatically (Fig. 3).

All data manipulation in the workflow was performed via the `RExpression` actor which is an interface to the R programming language. Each actor was measured by their position in the workflow the count of input and output ports, the count of R code lines as well as the used R commands and packages (Fig. 3).

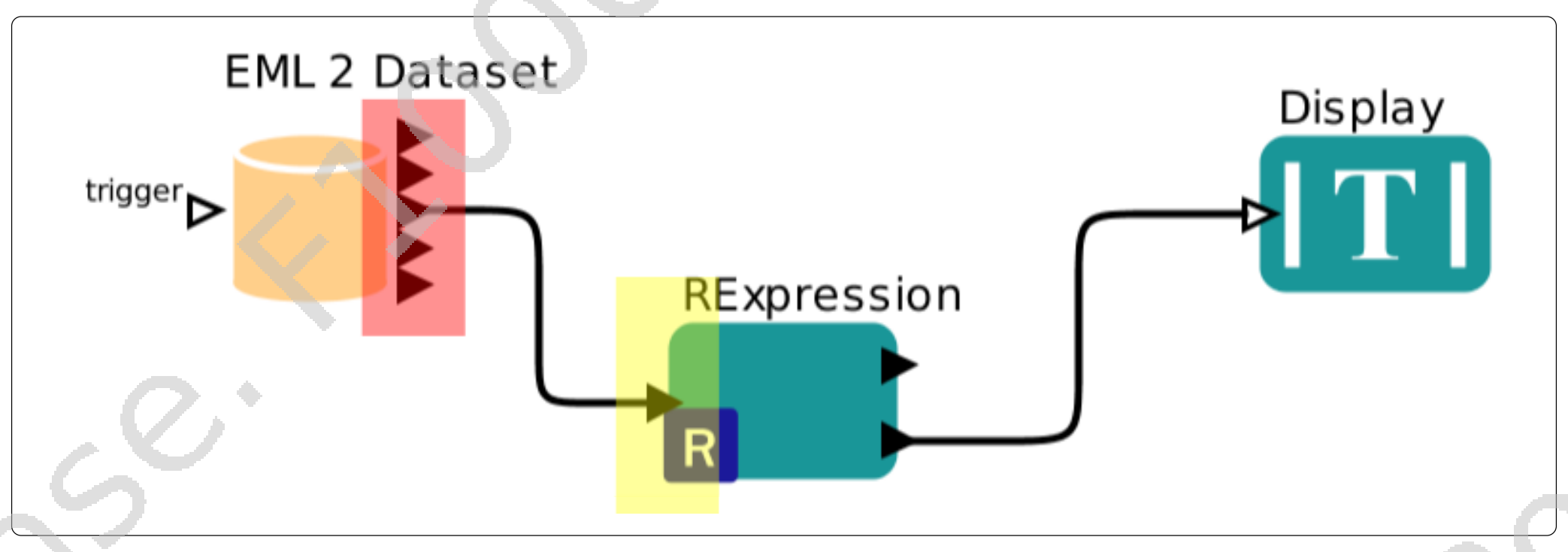


Figure 3: *The `EML 2 Dataset` actor on the left side serves as a data source. It provides data via its output ports (color=red) for further processing in an R script in the `RExpression` actor on the right side, which consumes the data through its input ports (color=yellow). Finally the data gets handed over to a display to show the results of the analysis of the R script*

The taken values were used to calculate a measure of complexity for each actor. The absolute complexity of an actor (ac = actor complexity) is calculated as a sum of the components: p = count of ports (in + out), loc = count of R code lines, fc = count of R functions, pa = count of R packages used (Equ. 1). Additionally each actor was assigned to a data manipulation purpose (Tab. 1).

$$ac = \Sigma p + \Sigma loc + \Sigma fc + \Sigma pa \qquad (1)$$

Table 1: *Analysis tasks (purposes) that occurred in workflow about the carbon stock analysis. The purposes were assigned and defined ad-hoc to the workflow components.*

Purpose	Description
data source	Emits data and does not consume it
data type transformation	Transform a variable type (mainly text to numeric)
merge data	The actor matches and merges data (vertical merge)
data aggregation	Horizontal aggregation of data
create new vector	Create a vector filled with new data
data imputation	Impute data, mostly linear regressions on data subsets
modify a vector	Modify a complete vector by a factor or basic arithmetic operation
create new factor	Create a new factor
data extraction	Extract data values (e.g from comment strings)
sort data	Sort data
data modeling	All kinds of model comparison related operations (ANOVA, AIC)

Results

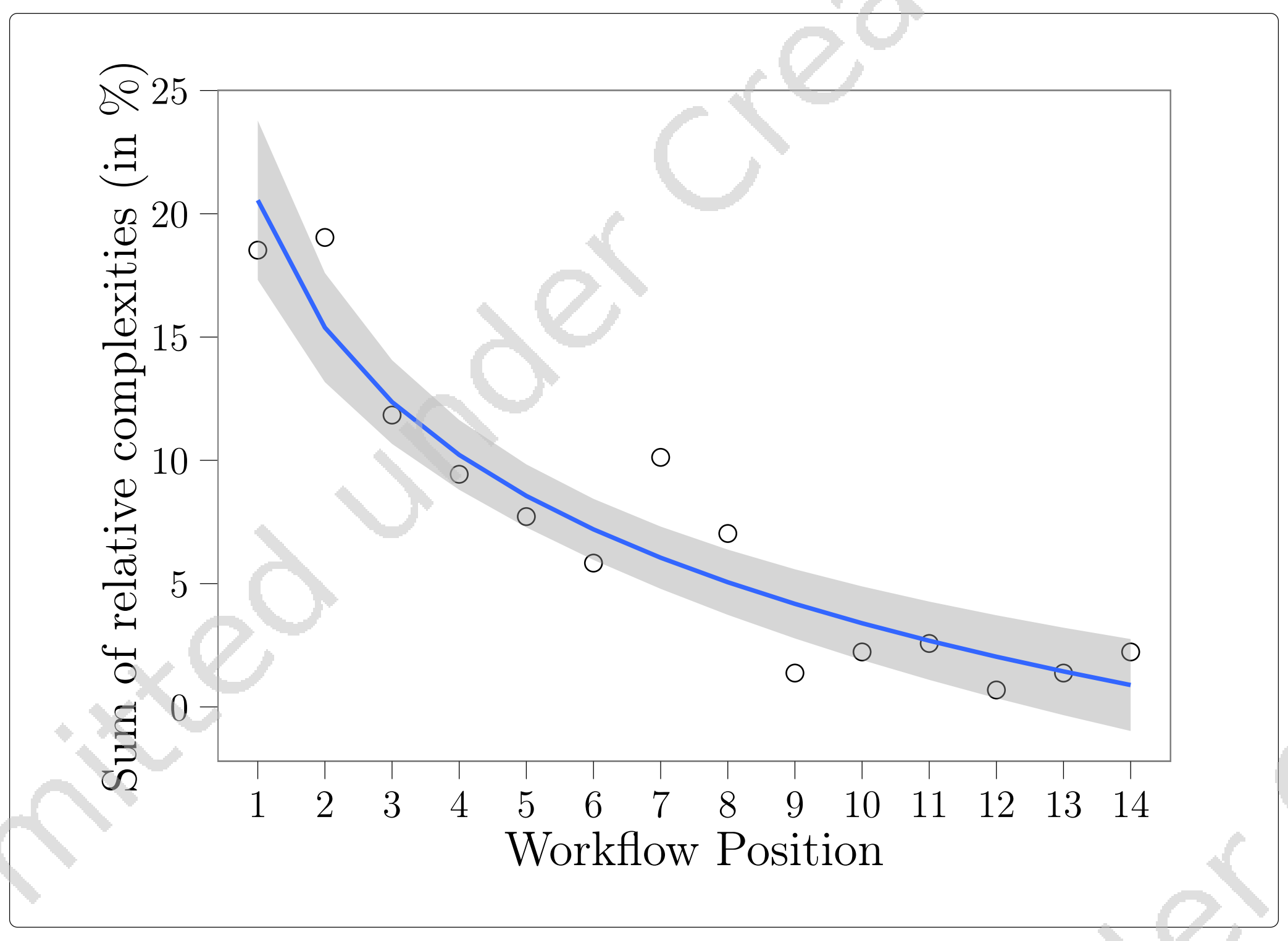


Figure 4: *The workflow position on the x-axis plotted against the sum of relative complexities per workflow position on the y-axis. The gray shading displays the standard error. Model: $lm(sum\ rel\ complexity \sim log(position))$ R-squared=0.9, F-statistic: 29.16 on 3 and 10 DF, p-value: 1.943e-05*

Answers

1. The analysis tasks are shown in table 1 with their description. Their diversity is displayed in the boxplot in Figure 6!
2. The level of complexity is not stable, it decreases throughout the workflow!
3. The data manipulation step "data imputation" shows the highest complexity and also the highest variation!
4. Data manipulation steps with a combination of low complexity and variability are potential candidates for automatization processes!

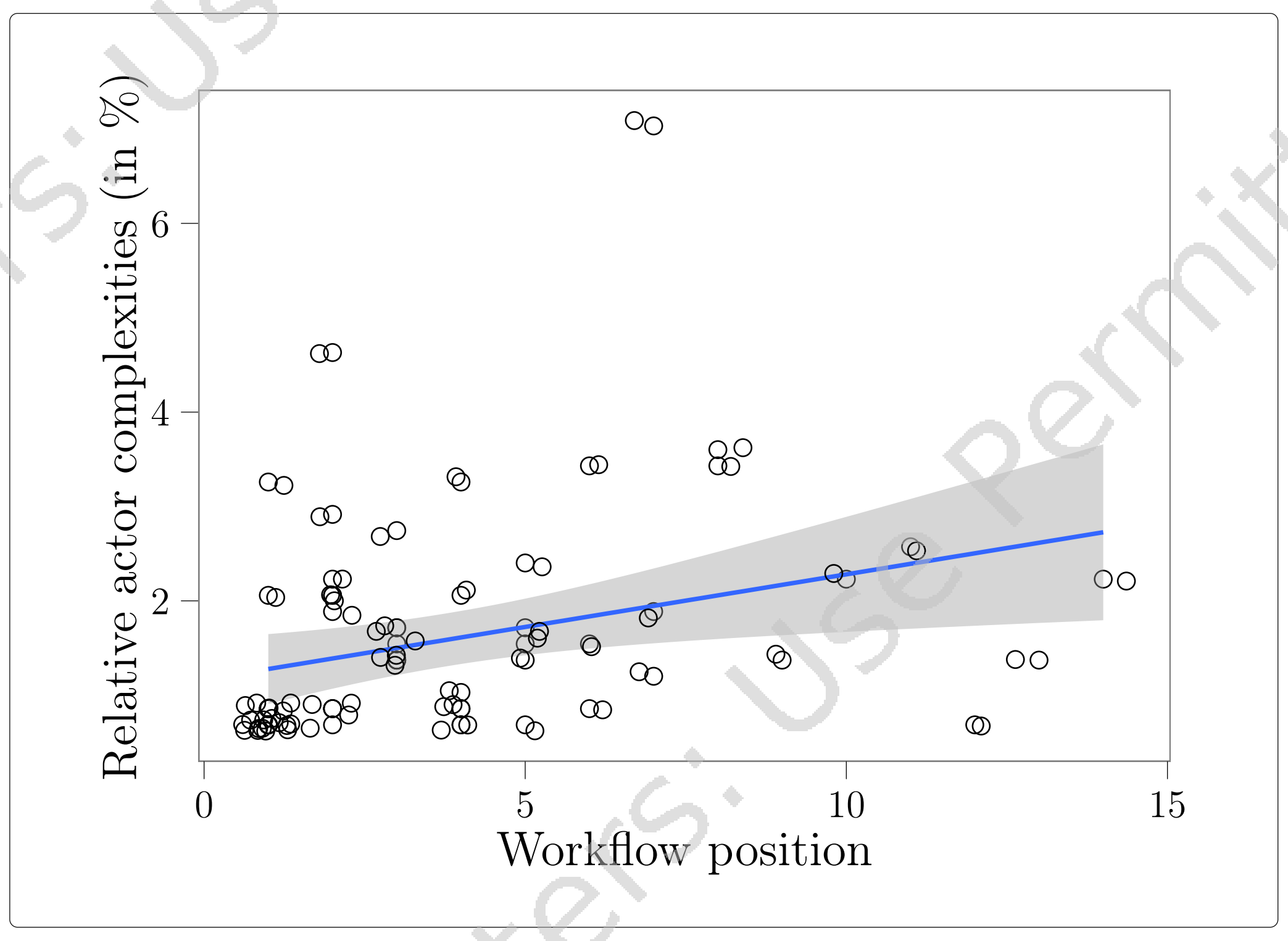


Figure 5: *Relative actor complexities along the workflow of the carbon analysis. The points are slightly jittered to handle overplotting. At each position in the workflow there are actors of different type and complexity. Model: $lm(rel\ complexity \sim position)$ R-squared: 0.09, F-statistic: 6.57 on 1 and 61 DF, p-value: 0.01285*

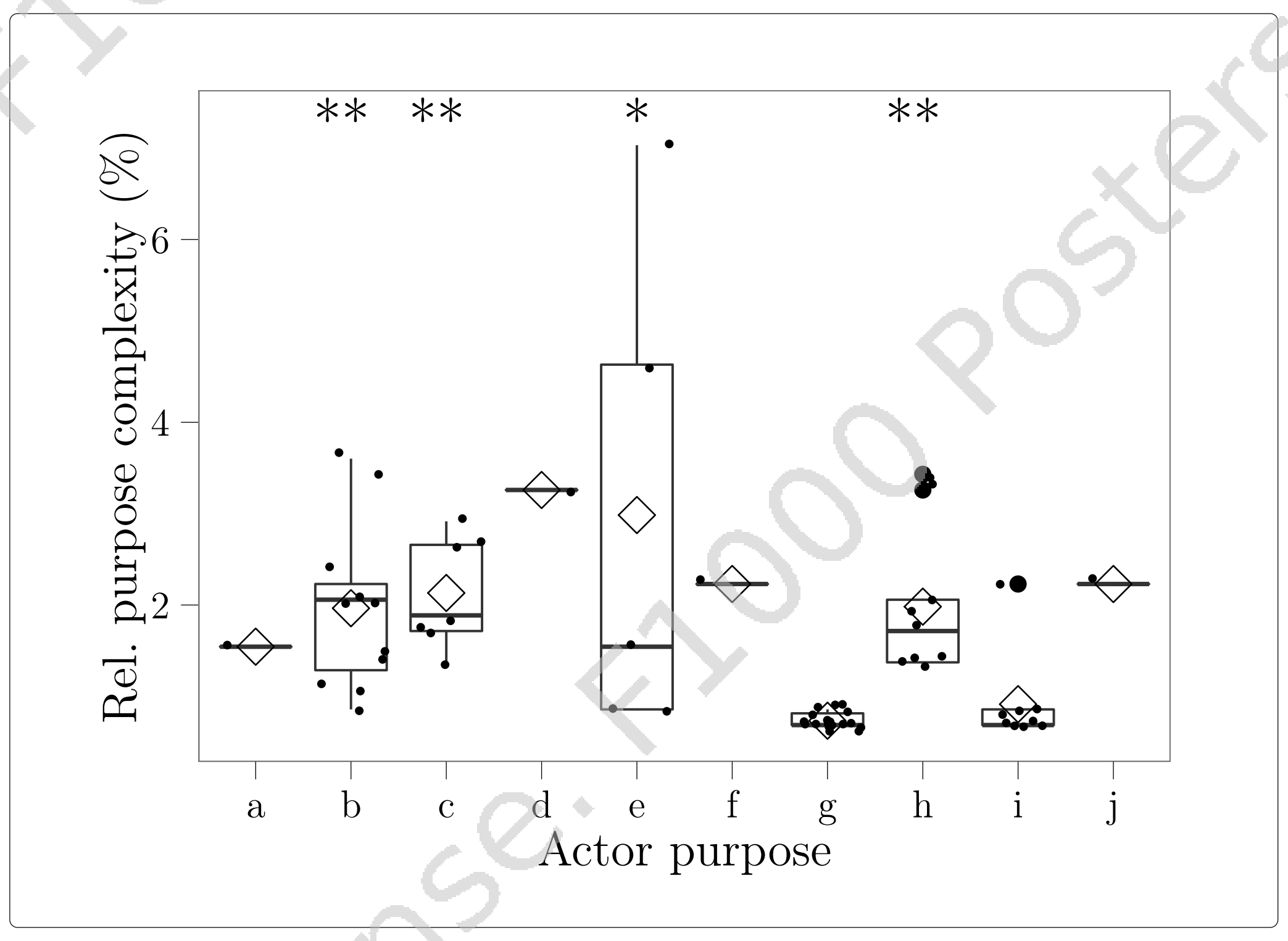


Figure 6: *The median, 25% and 75% quantiles of the relative actor complexities for the actor purposes. Letters refer to: a=create new factor, b=create new vector, c=data aggregation, d=data extraction, e=data imputation, f=data modeling, g=data type transformation, h=merge data, i=modify a vector, j=sort data. The small dots are the relative complexities, the diamonds the means. The whiskers are 25% quantile - 1.5 * IQR and 75% quantile + 1.5 * IQR, big black circles are outliers. Signific.: * = 0.05, ** = 0.001.*

Discussion

The sum of complexities per workflow position decreases throughout the whole workflow (Fig. 4). Although the sum of introduced complexity is high in the first working steps the relative complexities for each of the actors are low. This could be an indicator for places in a workflow where automated processes could come in handy to reduce the amount of work needed to prepare or handle the data. But also the variability of the purposes plays an important role for computer assisted processing. The lower the variability the more conformity we have inside a certain data manipulation step and the easier it is to handle by automated processes. Especially the purposes data type transformation and modify a vector had a very low variability of complexities. In fact they perform tasks already discussed as application for ontology frameworks like the Extensible Observation Ontology (OBOE). The developed measures can be used to investigate scientific workflows to find suitable tasks for automatization and the use of knowledge organization systems.