Data Wrangling Tool Comparison

Abdullah AlMasaud[1*], Sandra Sampaio[1] and Pedro Sampaio[2]

[1*]Department of Computer Science, University of Manchester, Oxford Rd, Manchester, M13 9PL, UK.
[2]Alliance Business School, University of Manchester, Booth Street West, Manchester, M15 6PB, UK.

*Corresponding author(s). E-mail(s):
abdullah.almasaud@postgrad.manchester.ac.uk;
Contributing authors: s.sampaio@manchester.ac.uk;
p.sampaio@manchester.ac.uk;

Data Wrangling Exploration across tools

Self-service preparation has been brought forward to decrease reliance on IT Hellerstein, Heer, and Kandel (2018) and shorten the time to business insight Stodder and Matters (2016). However, the tools that perform these operations vary in terms of their capabilities and cost and the steepness of the learning curve associated with using these tools. TThis learning curve associated with tools, along with the cost, is why some organisations and users tend to employ spreadsheet tools such as Microsoft Excel to prepare data Stodder and Matters (2016) which in turn motivated solutions (e.g. OpenRefine) to adopt spreadsheet-like interfaces.

To identify the challenges presented by the various tools in the real world, we performed a brief study on a sample of the tools used in practice, with the focus being on the visual tools used in data preparation employing spreadsheet interface or visual workflow designers. The spreadsheet tools selected were Microsoft Excel, OpenRefine, Talend Data Preparation and Trifacta Wrangler. The dimensions considered in the comparison were divided into the following levels:

(i) Tool-level: This includes the tool's availability mode (e.g. open source), the supported platform, the structure of the imported data set, the primitive structure, how the data is viewed, file formats (importable and exportable file formats), the maintenance of the provenance trail, the ability to omit a performed step,

the ability to execute external code or web services and the ability of community development of components.

(ii) Control structure: This includes the facility to control the execution of steps using e.g. loops.

(iii) Dataset–level operations: These include operations (such as reshaping via transpose and graphing) executed on a complete dataset.

(iv) Record-level operations: These include inter-record operations such as removing and sorting records.

(v) Attribute-level operations: These include operations that are applied to a column of attribute values; e.g.: creating or rearranging attributes.

(vi) Value-based operations: These include value-specific operations such as string manipulation and the typecasting of values.

(vii) Specialist operations: These include specialist preparation operations such as outlier detection and masking data values to comply with privacy requirements.

Tables 1 and 2 summarise the result of the exploration of the tools used to prepare data. The tools in Table 1 were the spreadsheet-based tools most often found in the cited research. There are many similarities that can be identified in the table; however, in reality, differences exist, such as how an operation is named and how it performs the task. Excel contains many of the preparation tasks performed in this exploration; however, its primitive structure uses a cell as representative of the intersection between an attribute and a record address. For this reason, it requires manual and intensive work that can be very error-prone. For example, it is theoretically possible to join data sets; however, this can only be achieved via the look-up table functionality, and the written function must be copied into all required cells and replaced at run-time with the found values. In comparison, OpenRefine uses 'facets' to group similar values in a column of attributes together; these facets are the only mechanism required to filter data, detect duplicates and perform value transformations, while other operations can be applied to attributes of columns.

The tools selected in Table 2 were selected because they are either freely available GUI-based workflow design tools, which facilitates their review, or because they are widely adopted in data-intensive and scientific workflow orchestrations (i.e., Taverna Wolstencroft et al. (2013) and Pegasus Deelman et al. (2015)). While Taverna and Pegasus are the most commonly used tools, they are not the most user-friendly. Neither tool has a standard set of operations for its users; however, Taverna provides its user with a set of community-developed operations. Pegasus has no GUI but is commonly used for its ability to orchestrate workflows over distributed environments.

Data preparation tools are tailored to different users and have similarities and differences that make them particularly suitable for certain tasks. However, the usability of these tools varies, both in terms of their technical knowledge requirements and whether they are freely available to all users. Orange, RapidMiner Studio and KNIME each provide a set of built-in operations and a GUI, making them more user-friendly. Orange is the most limited of the three; it only provides a subset of applicable activities related to data preparation pipelines because it is targeted at data mining, which is reflected in the terminology it employs (e.g. domains, classes). RapidMiner Studio and KNIME are more complete tools that can be used in numerous data preparation activities.

Table 1: Comparison of different tools used in data preparation (spreadsheet-like applications). (Y = yes, N = no and P = partial.

Dim.	Item \ Tool	Excel	OpenRefine	Talend Data Preparation	Trifacta Wrangler
(i)	Open Source	N	Y	Y	N
	Platform	Multiple	Multiple	Multiple	Multiple
	Primitive structure	Cell	Tabular	Tabular	Tabular
	View	Full File	Page(s)	Full File	Page(s)
	Import format(s)	Multiple	Multiple	Multiple	Multiple
	Export format(s)	Multiple	Multiple	Multiple	Multiple
	Provenance trail	N	Y	Y	Y
	Omit steps	P	Y	Y	Y
(ii)	Condition	Y	N	N	N
	Loop	N	N	N	N
(iii)	Transpose/Pivot	Y	Y	Y	Y
	Graph	Y	P	Y	Y
	Join data sets	P	P	Y	Y
	Union data sets	Manual	N	Y	Y
	Aggregate	Y	N	Y	Y
(iv)	Remove	Manual	Y	Y	Y
	Filter	Y	Y	Y	Y
	De-duplicate	Y	Y	Y	Y
	Sort	Y	Y	Y	Y
(v)	Create	Y	Y	Y	Y
	Rename	Y	Y	Y	Y
	Remove	Y	Y	Y	Y
	Rearrange	Y	Y	Y	Y
	Split	Y	Y	Y	Y
(vi)	String manipulation	Y	Y	Y	Y
	Math functions	Y	Y	Y	Y
	Typecast	Y	N	Y	Y
(vii)	Outlier detection	N	N	N	Y
	Mask values	N	N	Y	N

Table 2: Comparison of the different tools used in data preparation (workflow tools). (Y = yes, N = no, M = multiple and P = partial.

Dim.	Tool / Item	Orange	RapidMiner Studio	KNIME	Taverna	Pegasus
(i)	Open Source	Y	N	Y	Y	Y
	Platform	M	M	M	M	Unix
	Primitive structure	Tabular	Tabular	Tabular	N/A	N/A
	Import format(s)	M	M	M	N/A	N/A
	Export format(s)	M	M	M	N/A	N/A
	Provenance trail	Y	Y	Y	Y	Y
	Omit steps	Y	Y	Y	Y	Y
	Call external code	Python	M	M	M	M
	Call web services	N	N	Y	Y	Y
	Community development	Y	N	Y	Y	Y
(ii)	Condition	N	Y	Y	P	P
	Loop	N	Y	Y	P	P
(iii)	Transpose/Pivot	Y	Y	Y	N/A	N/A
	Graph	Y	P	Y	N/A	N/A
	Join data sets	Y	Y	Y	N/A	N/A
	Union data sets	Y	Y	Y	N/A	N/A
	Aggregate	Y	Y	Y	N/A	N/A
(iv)	Remove	Y	Y	Y	N/A	N/A
	Filter	Y	Y	Y	N/A	N/A
	De-duplicate	N	Y	Y	N/A	N/A
	Sort	N	Y	Y	N/A	N/A
(v)	Create	Y	Y	Y	N/A	N/A
	Rename	Y	Y	Y	N/A	N/A
	Remove	Y	Y	Y	N/A	N/A
	Rearrange	N	Y	Y	N/A	N/A
	Split	N	Y	Y	N/A	N/A
(vi)	String manipulation	P	Y	Y	N/A	N/A
	Math functions	P	Y	Y	N/A	N/A
	Impute	Y	Y	Y	N/A	N/A
	Replace	P	Y	Y	N/A	N/A
	Typecast	N	Y	Y	N/A	N/A
(vii)	Outlier detection	Y	Y	Y	N/A	N/A
	Mask values	N	N	N	N/A	N/A

References

Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P.J., ... others (2015). Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, *46*, 17–35,

Hellerstein, J.M., Heer, J., Kandel, S. (2018). Self-service data preparation: Research to practice. *IEEE Data Eng. Bull.*, *41*(2), 23–34,

Stodder, D., & Matters, W.D.P. (2016). Improving data preparation for business analytics. *Transforming Data With Intelligence*, *1*(1), 41,

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., ... Goble, C. (2013, 05). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, *41*(W1), W557-W561, https://doi.org/10.1093/nar/gkt328 Retrieved from https://doi.org/10.1093/nar/gkt328 https://academic.oup.com/nar/article-pdf/41/W1/W557/3822473/gkt328.pdf