

On the Reusability of Data Cleaning Workflows

Lan Li and Bertram Ludäscher


lanl2@illinois.edu
ludaesch@illinois.edu



**School of
Information Sciences**

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Outline

- Data Cleaning with OpenRefine 
- What do we mean by “reusing a data cleaning workflow”?
- Reusing Data Cleaning Workflows: Challenges & Opportunities
- A Simple Conceptual Model for Recipe Reuse
- Improving the Reusability of Data Cleaning Workflows
- Conclusions and an Invitation (quick survey)

Data Cleaning with OpenRefine

7594 rows													
Show as: rows records Show: 5 10 25 50 rows													
All	id	name	host_id	host_name	neighbourhood	neighbourhood	latitude	longitude	room_type	minimum_nights	number_of_reviews	last_review	reviews_per_month
41.	207218	(Historic Pullman Artist Flat)	1019125	Rebecca	hyde park	OHARE	41.7888649	-87.58670891	Private room	2	137	11/12/18	2.92
42.	221109	(1 Bedroom Apartment -Noble Square)	1146738	Reem			41.92926222	-87.66009125	Entire home/apt	4	93	8/12/18	0.81
43.	223777	(Bohemiam Bedroom in H...Hood)	1163561	Sarah	West Town		41.90289494	-87.6818216	Entire home/apt	2	321	10/29/18	2.81
44.	225314	(Cozy & comfy with AC close to California Blue Line)	1173654	At Home Inn	Lincoln Park		41.91768924	-87.63787944	Entire home/apt	3	33	10/14/18	0.59
45.	228273	(Amazing Lincoln Park Location!!)	1190842	Lois And Ed	hyde park		41.79708495	-87.59194894	Private room	2	31	7/29/18	0.63
46.	230836	(Huge Bedroom with pvt marble bath)	1185573	At Home Inn	Lincoln Park		41.91182685	-87.63999816	Entire home/apt	3	9	11/5/18	0.16
47.	233933	(BROWN ROOM W/PRIVATE BATH-WALK TO TRAIN)	1224828	Dominic	O?Hare		41.9312564	-87.65227338	Private room	165	8	10/24/17	0.18
48.	234442	(Beautifully Furnished Studio SW2R)	87231	Sharon And Robert	Uptown		41.96331169	-87.66088494	Private room	2	4	10/17/18	0.1
49.	236792	(Modern Luxury Condo in Lakeview)	1241447	Craig	Irving Park		41.95602263	-87.72780856	Private room	60	9	10/12/14	0.15
50.	241514	(Gay friendly apartment in Chicago)	1267472	Jeff And JoAnne	Lincoln Park		41.91336569	-87.63918044	Entire home/apt	1	174	10/21/18	1.69

Table 1. Chicago Airbnb Dataset

OpenRefine CHI_Airbnb

Facet / Filter Undo / Redo 1 / 4

Extract... Apply...

Filter:

1. Create project

2. Text transform on 1409 cells in column name: `greel:value.replace(" ");replace(" ");replace(" ");`

3. Text transform on 6 cells in column host_name: `value.trim()`

4. Text transform on 898 cells in column neighbourhood: `value.toTitleCase()`

5. Text transform on 6699 cells in column last_review: `greel:toStringToDate(value); "yyyy-MM-dd"`

Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

☒ Text transform on cells in column name using expression `greel:value.replace(" ");replace(" ");replace(" ");`

☒ Text transform on cells in column host_name using expression `value.trim()`

☒ Text transform on cells in column neighbourhood using expression `value.toTitleCase()`

☒ Text transform on cells in column last_review using expression `greel:toStringToDate(value); "yyyy-MM-dd"`

```
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "columnName": "name",
  "expression": "greel:value.replace(' ');replace(' ');replace(' ');",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on cells in column name"
},
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "columnName": "host_name",
  "expression": "value.trim()",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on cells in column host_name"
},
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "columnName": "neighbourhood",
  "expression": "value.toTitleCase()",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on cells in column neighbourhood"
},
{
  "op": "core/text-transform",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "columnName": "last_review",
  "expression": "greel:toStringToDate(value); 'yyyy-MM-dd'",
  "onError": "keep-original",
  "repeat": false,
  "repeatCount": 10,
  "description": "Text transform on cells in column last_review"
}
```

Select All Unselect All

Close

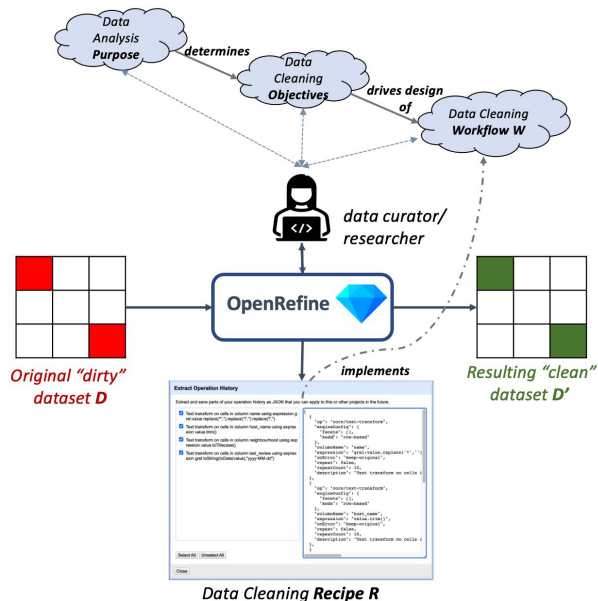
Data quality issues (from left to right):

1. **Special characters** in column name;
2. Leading **whitespaces** in column `host_name`
3. **Spelling variations and typos** in column `neighbourhood`, e.g., “OHARE” and “O?Hare”
4. **Non-standard date format** in column `last_review`

Example OpenRefine recipe with four steps

3

What do we mean it mean by “Reusing a Data Cleaning Workflow”?



First attempt at a definition:

Let R be the recipe for the data cleaning workflow W that was used on dataset D .

We say that recipe $R (= R_{D,W})$ is being **reused** whenever we apply it on a **different** dataset E .

Simple! What could possibly go wrong?

A lot, as it turns out ... ;-)

*Executing a data cleaning workflow on D yields D' and an **operation history** (“recipe”) R :*

$$D' = R(D)$$

*R can be **reused** later! (hopefully ..)*

Reusing Data Cleaning Workflows is **Desirable** but **Challenging**

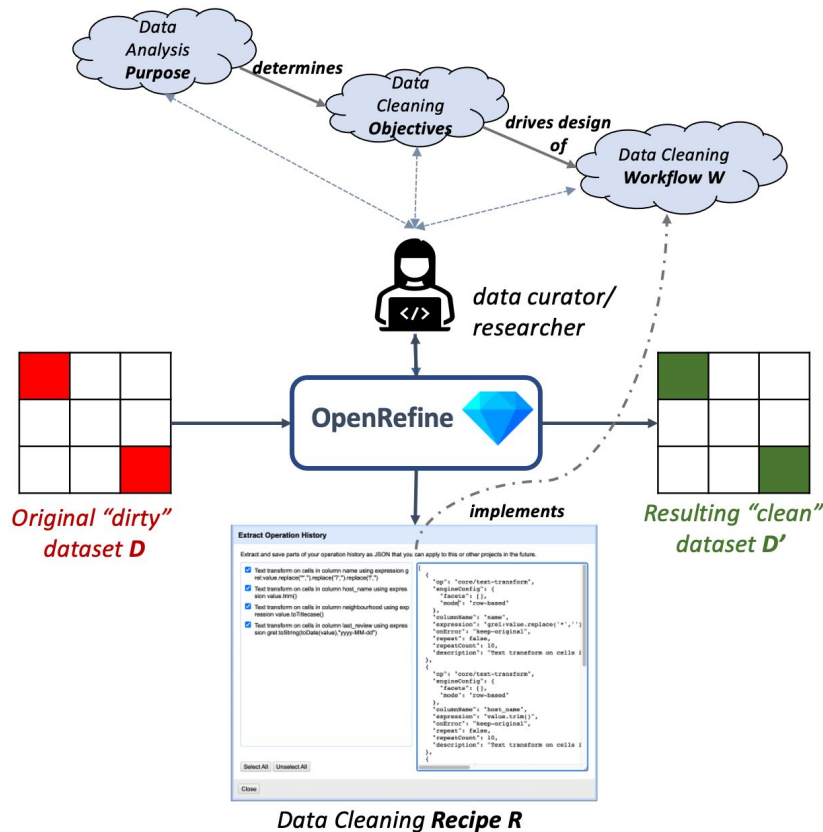
Recipe reuse: Applying the old recipe R (recorded for D) on a new dataset E :

$$E' = R(E)$$

When does this **work/not work**? For example:

- R might **not be safe** for E , e.g.,
 - if D has numeric type in column C , but E has string type in column $C \Rightarrow$ applying arithmetic operations on C is OK in D but a **type error** in E .
- R might be “useless” (**not meaningful**) on E , e.g.,
 - if the **schema** of E is disjoint (or very different) from D
 - Even if $\text{schema}(E)=\text{schema}(D)$, there might be further problems:
 - If the **semantics** of data in E is different from one in D . E.g, apply the recipe that is used to clean menu dataset on biological dataset.
 - if the **purpose** for cleaning D is different from the purpose for cleaning E

A Simple Conceptual Model for Recipe Reuse



Recipe Creation:

1. A researcher or data curator has a **data analysis purpose P** in mind.
2. The purpose P can be identified by one or more **questions/queries Q** on the given dataset D .
3. **Data cleaning objectives O** are determined by the analysis purpose P (and the associated questions/queries Q).
4. The objectives drive the design of the **data cleaning workflow W** .
5. Executing W yields a **recipe R** (if **provenance** is captured).

Recipe Reuse (later – not shown on the left):

6. Ensure R is (type) **safe** and (semantically) **meaningful** before applying it on new data E .
7. Sometimes R can be reused **"as is"** on E (rarely). Usually: decompose R into smaller **modules** and **adapt** (e.g., to schema changes).

Reusability of Recipes (second attempt)

We say that $R (=R_{D,W})$ is **directly reusable** for a new dataset E ,

- If $\text{schema}(E) = \text{schema}(D)$ and $\text{purpose}(E) = \text{purpose}(D)$.

Otherwise, we say that R is **possibly reusable with modifications**, i.e. if there are schema changes and/or changes in the purpose of E relative to the given D .

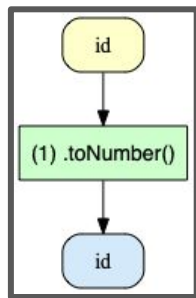
Research Question:

- How to modify R to R' to make it more reusable?

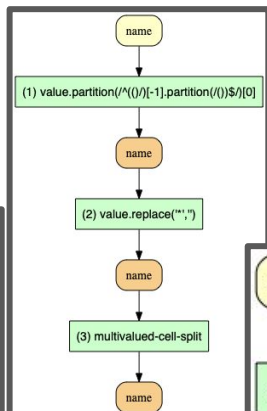
Improving the Reusability of Data Cleaning Workflows

... by exploiting the **modular structure** of recipes

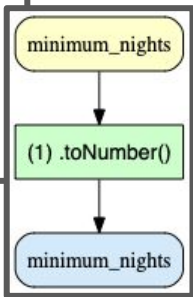
0. Create project
1. Text transform on 7594 cells in column id: value.toNumber()
2. Text transform on 7594 cells in column name: gregl:value.partition(/^(\[-1].partition(/(\[-1]\$)/[0]
3. Remove column neighbourhood_group
4. Text transform on 7594 cells in column minimum_nights: value.toNumber()
5. Mass edit 321 cells in column neighbourhood
6. Text transform on 77 cells in column name: gregl:value.replace("","")
7. Text transform on 7594 cells in column number_of_reviews: value.toNumber()
8. Create new column reviews_representation based on column number_of_reviews by filling 7594 rows with gregl:if(value>100,'high','low')
9. Mass edit 47 cells in column host_name
10. Text transform on 0 cells in column host_name: value.trim()
11. Split 7591 cell(s) in column host_name into several columns by separator
12. Rename column host_name 1 to First_host
13. Rename column host_name 2 to Second_host
14. Text transform on 7594 cells in column availability_365: value.toNumber()
15. Create new column comment_hotel based on column reviews_representation by filling 7594 rows with gregl:if(and(value=='high', cells.availability_365.value <100),'popular','other')
16. Split multi-valued cells in column name
17. Text transform on 6699 cells in column last_review: value.toDate()
18. Text transform on 6699 cells in column last_review: gregl:value.split('T')[0]



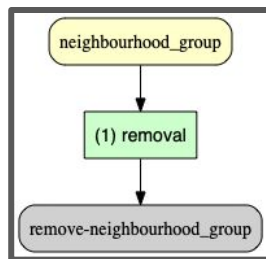
Module 1



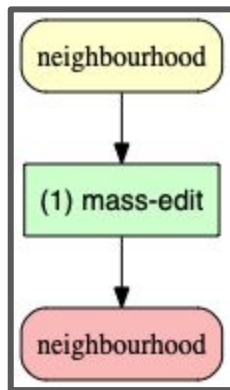
Module 2



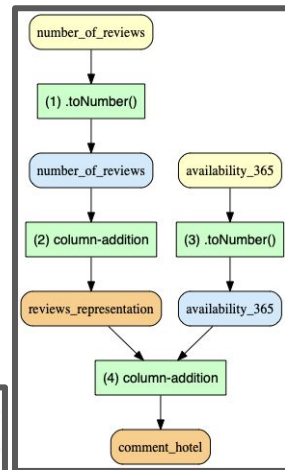
Module 4



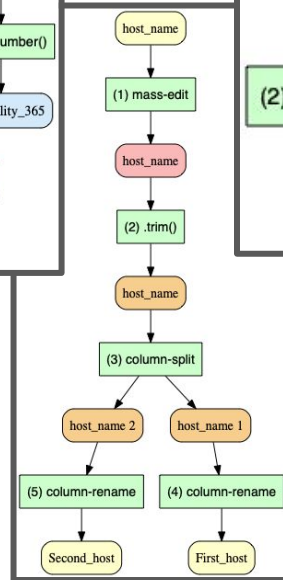
Module 3



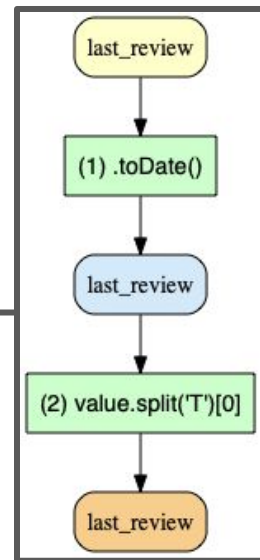
Module 5



Module 6



Module 7



Module 8

Li, L., Parulian, N., & Ludäscher, B. (2021). **Automatic Module Detection in Data Cleaning Workflows: Enabling Transparency and Recipe Reuse**. 16th Intl. Digital Curation Conference (IDCC), 2021. <https://doi.org/10.2218/ijdc.v16i1.771>.

Improving the Reusability of Data Cleaning Workflows

... by **generalizing** data cleaning operations

Example:

Consider two separate operations:

- o_1 : US_date \rightarrow ISO_date
- o_2 : EU_date \rightarrow ISO_date

.. for mapping date formats (MM/DD/YYYY and DD.MM.YYYY) to the standard format (YYYY-MM-DD)

Combining o_1 and o_2 into a single operation ..

- o_3 : (US_date U EU_date) \rightarrow ISO_date

yields a **more reusable** operation o_3 since it can handle a larger set of inputs than o_1 and o_2 individually.

Summary

1. Reusing Data Cleaning workflows/recipes is **desirable** but **challenging**.
 - a. e.g. before reusing R recipe on new dataset E , we need to make sure that it is **safe** and **meaningful** to do so
2. We have developed research prototypes (companion tools) for **OpenRefine** to **capture recipes** and analyze and enrich their **provenance** information...
 - a. e.g. to decompose R it into smaller **modules** (subworkflows) that can be customized and reused in other recipes for new datasets.
 - b. **generalizing** operations by extending their domain is another way some operations can be made more reusable (future work)
3. We'd like to hear from you!
 - a. Quick survey to share your data cleaning experiences, desiderata, etc.

Acknowledgments. We thank **Nikolaus Parulian** and **Timothy McPhillips** for fruitful discussions and joint work on data cleaning research and tool development!

Thank You ... and an Invitation!

*Please share your data cleaning **experiences**, **use cases**, **desiderata** etc.*

Quick Survey:

bit.ly/data-cleaning-survey

We are always looking for new real-word use cases and potential collaborators!