Cognitive Dimensions of Between-Table Context Support in Wrangling Applications

Steve Kasica
Dec. 03, 2019
Why am I doing the thing that I’m doing?

• More journalists are using the tools and techniques of data science under the term data journalism.
• Little is understood about the unique issues of this group as they related to cleaning, transforming, and otherwise wrangling their data.
• But there’s a lot of open-source and commercial wrangling applications available to journalists.
• However, do these interfaces support the kind of tasks and data that journalists actually do in the wild?
• This is an analysis project in the course
What is the thing that I’m doing?

• Replicate the wrangling workflows done by real journalists working in a programming environment with these GUI-based tools
• Discuss the trade-offs that exist between dimensions in the wrangling activity.
• Compare and contrast the strengths are weaknesses of these two tools in the wrangling of journalistic data.
Prior Work

• Over the summer, I conducted an artifact-mediated indirect observational study of data wrangling in journalism.
• Identified high-level wrangling actions done by journalists
• Also identified exemplar data and wrangling sequences.
Workflows and Tools

• Reproduce each wrangling workflow with both tools
• Workflows are abstracted to not a sequence of steps because that would be trivial,
• Workflows are a sequence of intermediate table forms to reproduce
• One workflow-tool combination may include may branches for getting to different table states as there are different means to the same end.
Workflows to reproduce

• Longterm managed care records in New York
  • Sarah Cohen's CAR 2016 tutorial on data cleaning with OpenRefine
  • Performs the following tasks: extract data from column, remove non-data rows, remove rows that contain notes, remove bad-data rows, remove rows with missing values, aggregate join, resolve entity names

• Water usage over time in California
  • Wrangling performed by Ben Welsh at *Los Angeles Times*
  • Performs the following tasks: configure analysis tools, subset raw data to relevant, string-ify date, filter data, remove rows
Applications Considered

- There are many applications for wrangling: OpenRefine, Cloud Dataprep, Tableau Prep, Trifacta Wrangler, Workbench
- Focus on OpenRefine and Cloud Dataprep
  - Were recommended by a in the MOOC Data Journalism and Visualization with Free Tools offered from Knight Center for Journalism in the Americas
OpenRefine
Google Cloud Dataprep
Cognitively Dimensions

• There are 13 different dimensions to create a common, interface-independent vocabulary to discuss usability in user interfaces
• Each interface occupies 13-dimensional space, thus improving an interface in one aspect impacts the others
• One goal of this project is to identify these tradeoffs in data wrangling interfaces in general, in addition to compare and contrasting the two tools.
Viscosity

• “Resistance to change” [Blackwell et al, 2003]
Visibility

• “Ability to view components easily” [Blackwell et al., 2003]
  • Can we see all components in VPL? [Blackwell et al., 2003; Green, 1996]
• In data wrangling, visibility because an issue as datasets become large
  • Is every part of the relevant data simultaneous visible?
  • In high-dimensional data you have to scroll to view all columns
  • In data with many observations, you have to scroll to view rows.
• **Focal point:** Would increasing visibility may decrease error-proneness?
• Visualization may help here. Charts are great at representing data compactly, a.k.a data-ink ratio [Tufte, 1983]
Visibility in Dataprep
Visibility in OpenRefine

![OpenRefine interface](image)
Premature Commitment

• “Constraints on the order of doing things” [Blackwell et al., 2003]
Hidden dependencies

• “Important links between entities are not visable” [Blackwell et al., 2003]
• The output of each transformation step in a wrangling process serves as the input for the next. So in wrangling dependencies are highly sequential
• But often the sequences doesn’t matter unless it’s a transformation that restructures the dataset.
Role-Expressiveness

• “The purpose of an entity is readily inferred” [Blackwell et al., 2003]
• In data wrangling, it is already difficult to verbally express table transformations.
• Different tools use different vocabulary to describe the same thing.
  • Entity resolution: “cluster and edit” and “mass edit” in OpenRefine and “standardize” in DataPrep
  • DataPrep does include little icons, which are more helpful than no icons.
Error-Proneness

• “The notation invites mistakes and the system gives little protection.” [Blackwell et al, 2003]

• In data wrangling, errors often creep in when filtering as Type I vs Type II errors in the gulf of execution and evaluation [Hutchins et al., 1985]
  • Type I / false positive: A row was removed, but it should have been kept.
  • Type II / false negative: A row was kept, but it should have been removed.

• You often have to approve operations on rows that you don’t know the values of.
Abstraction

- “Types and availability of abstraction mechanisms” [Blackwell et al, 2003]
- Wrangling actions may encapsulate many small, low-level actions.
Secondary notation

• “Extra information in means other than formal syntax” [Blackwell et al, 2003]

• Secondary notation is often used in specifying column extraction methods
  • Python and “index slicing” such as foo[0:5]
  • Regular Expressions
Closeness of mapping

• “Closeness of representation to domain” [Blackwell et al, 2003]
• As examples of direct-manipulation interfaces, both interfaces enjoy a very close mapping between notation and results it’s describing.
Consistency

• “Similar semantics are expressed in similar syntactic forms” [Blackwell et al, 2003]
Diffuseness

• “Verbosity of language” [Blackwell et al, 2003]
• In some tasks, the notation can be too concise, when you have to specify a sequence of three transformations that might be encapsulated in one transformation.
• Perhaps diffuseness and abstraction are two interrelated dimensions.
Provisionality

• “Degree of commitment to actions or marks” [Blackwell et al, 2003]
• Both tools support a preview function that addresses provisionality.
• It makes sense that these interfaces may not suffer from pre-mature commitments because they both use this idea.
Addressing provisionality with previews
Progressive evaluation

• “Work-to-date can be checked at any time” [Blackwell et al, 2003]
• This may be a barrier to collaboration in OpenRefine as it doesn’t support concurrent modifications per project.