Review of OpenRefine Software

By Dom Bortruex

OpenRefine is a free, open-source desktop application for analyzing, cleaning, and transforming messy data. It is similar to spreadsheet applications but behaves like a database. Users of all experience levels can analyze and manipulate data without complicated scripting. Due to its ease of use, OpenRefine is an ideal tool for working with bibliographic data, including tabular data, XML, and even MARC records.

OpenRefine operates on rows and columns of data. Data can be viewed as rows or records. With a few clicks from the column tools, users can split or join columns. Rows and records can be filtered using facets. Multiple facets can be viewed at a time. Clusters of facets group similar content to provide an in-depth view of the data. This is useful for identifying variations in data values including misspellings and unexpected characters. For example, a user could facet the MARC fields 490 and 830 to compare the fields, then use clusters to identify discrepancies.

Once data is explored through facets and clusters, users can query and clean the data through built-in transformations and simple expressions using the Google Refine Expression Language (GREL). Built-in transformations include changing data values to text, numbers, or dates, stripping white space, and changing the text case. OpenRefine provides extensive documentation on GREL, empowering users to easily learn the language. Through GREL transformations, users can split or join columns, look up values from other projects, or edit values. Users can also write complex transformations using Python.

OpenRefine's reconciliation services retrieve data from external sources such as Library of Congress Subject Headings (LCSH), Virtual International Authority File (VIAF), Open Researcher and Contributor ID (ORCID), Wikidata, and others. Through the reconciliation tool, users can retrieve and input controlled names and terms. For example, the reconciliation service could be applied to the MARC field 100 to control creator names. This service, which includes "best candidate scores" to help identify accurate matches, allows users to review and select reconciled information. It also provides a link to the authority page for easy review of controlled terms.

OpenRefine allows exports of projects and data in multiple tabular formats, including Excel and many others. Additionally, you can export steps performed in a project, including facets, clusters, mass edits, and transformations; however, single-cell edits are not reflected in these exported steps. Users can import the project steps and apply them to other sets of data.

OpenRefine's power and ease of use makes it an ideal tool for working with messy bibliographic data. The ability to easily drill down into data fields, query other projects and reconciliation services, and perform complex transformations through built-in features and expressions gives users control over their data with a minor learning curve.

Dom Bortruex is the Manager of the Monographic FastCat Unit at Yale University.