ELECTRONIC RESOURCES REVIEWS

OpenRefine (version 2.5). http://openrefine.org. Free, open-source tool for cleaning and transforming data.

Overview

In recent years, the proliferation of open source data has been a boon to librarians who serve multidisciplinary research communities in academic and other settings. However, this type of data is often not readily usable because of errors, inconsistencies, problems with data entry methods, or file format incompatibilities. While technically possible, it is usually impractical to clean up the data line by line or with Excel tools. Enter OpenRefine (previously, Google Refine), a free, open source tool for turning messy data into usable, clean data [1, 2]. OpenRefine is a standalone tool that runs like a web server on an individual's computer. It uses a web browser as its interface, but the data stay local. It is a powerful utility that is capable of transforming and cleaning large datasets. However, it is also a complex tool that has a steep learning curve for the average user who does not have some experience with programming and expression language strings. The good news is that recent trends in data management and curation have led to new courses and programs in academia and library schools, so more recent graduates are likely to have the requisite skills for using these kinds of tools.

Intended audience

OpenRefine will interest librarians, scientists, data curators, researchers, business analysts, data journalists, and digital repository managers in a variety of disciplines who need clean, usable data. While the librarian may or may not be an OpenRefine expert user, knowledge of its existence and functionality will be beneficial for users in need of such a tool. Just as librarians hold classes on using library resources and EndNote as complementary tools for research, they might educate and train users about finding datasets and cleaning the data with OpenRefine. Once familiar with this tool, librarians themselves may find that OpenRefine serves a useful purpose in their own work environments.

Major features

OpenRefine is very powerful. Users can explore data to see the big picture, clean and transform data, and reconcile data with various web services. Examples of OpenRefine abilities are:

- OpenRefine works with local files or data from web addresses in a number of file formats, including CSV, TSV, XLS, XML, and other formats.
- It has the ability to filter or search for certain elements that need to be changed in some way, which restricts the view to just the relevant cells, rows, or columns that contain the elements. Then the user can perform the desired action on just those entries.
- It can find duplicate entries, empty cells, entry variations, inconsistencies, and patterns of errors for bulk fixing and cleaning.
- It provides a quick analysis of the data contained in the file; for instance, the word facet tool can analyze the words in a column and return a count of each of the unique words, and the results sort alphabetically by default, but when sorted by count, any trends can be seen at a glance.
- It provides an Undo/Redo function for all actions performed on the data, which saves time and effort by extracting and reusing commands; for example, an address field may include city, state, and zip code, and an operation can be performed to separate the data into three separate columns. When the same issue is encountered in another project, it is a simple task to copy (extract) the operation from one project and apply it to the data field.
- It uses the Google Refine Expression Language (GREL) as its native language to transform existing data or to create new data and supports Jython and other programming languages; according to the documentation, GREL is designed to resemble Javascript. For instance, a GREL expression can be used to find all instances of a particular text string and replace it with another string of text.
- Users might find that the data need to be reconciled, linked, or extended with authoritative sources, of which several are available in OpenRefine. The VIVO Scientific Collaboration Platform allows reconciling data corresponding to VIVO entities such as faculty members, journal titles, and other data entries. Several reconciliation services are mentioned in the user documentation for other uses.

Accessibility and usability

OpenRefine is easy to download and install. Words and phrases used to describe this tool by actual users include ''essential,'' ''wonderful,'' a ''time saver,'' and ''can't live without.'' However, usability depends in large part on user background and skills. For nonprogrammers, OpenRefine comes with a steep learning curve, to the point that it may be unusable for any but the most basic actions. For users with more technical skills (or access to technical support staff), the tool has great potential for its utility and time-saving functions.

Documentation and tutorials

The OpenRefine wiki includes links to overview videos and detailed documentation for users [3]. The documentation is the result of the efforts of many developers working together in an open source environment. As such, it is geared toward programmers and those familiar with coding language and terms, so the instructions and explanations might be difficult to grasp for nonprogrammers. An Internet search for articles and other tutorials retrieved user blog postings and articles about user experiences and real-world examples, which may be helpful.

Deficiencies and disadvantages

One potential disadvantage is that future development will depend on the commitment and efforts of interested users in the open source community. No formal technical support is available; rather, support will come from user forums and the development community. Computer processing power is essential for working with large amounts of data.

Similar product

The Stanford Visualization Group offers DataWrangler, a free, web-based tool for data cleaning and transformation [4]. At a glance, DataWrangler appears to have a more user-friendly interface than OpenRefine. At the present time, only the web application is available with no support or documentation. According to the official site blog, DataWrangler is a client-side JavaScript application, which limits the amount of data that can be ''wrangled''at a time.

Conclusion

When dealing with data, the ability to modify and transform many records at once allows users to save tremendous amounts of time and create usable data. This snapshot of OpenRefine just barely scratches the surface in the ways the data can be viewed, filtered, and modified. For users with some programming experience, the experience is likely to be straightforward and rewarding; however, its full utility will be out of reach for the non-programmer. The only way to know if OpenRefine is right in a particular setting is to try it. At the very least, librarians will benefit by understanding its capabilities and knowing when to recommend it to their users.

Kelli Ham, MLIS, kkham@library.ucla.edu, National Network of Libraries of Medicine, Pacific Southwest Region, Louise M. Darling Biomedical Library, University of California, Los Angeles, Los Angeles, CA

References

1. OpenRefine. A free, open source, power tool for working with messy data [Internet]. [cited 1 Feb 2013]. <http://www.openrefine.org>.
2. Magdinier M. From Freebase Gridworks to Google Refine and now OpenRefine [Internet]. GoogleRefine blog [10 Mar 2012; cited 1 Feb 2013]. <http://www.googlerefine.blogspot.com/2012/10/from-freebase-gridworks-to-google.html>.
3. OpenRefine wiki [Internet]. [cited 1 Feb 2013]. <https://github.com/OpenRefine/OpenRefine/wiki>.
4. Stanford Visualization Group. DataWrangler [internet]. The Group [cited 1 Feb 2013]. <http://vis.stanford.edu/wrangler/>.

DOI: http://dx.doi.org/10.3163/1536-5050.101.3.020

PsycTESTS. 750 First Street Northeast, Washington, DC 20002-4242; http://www.apa.org/pubs/index.aspx. Institutional subscriptions from assorted vendors, with pricing available on request; individually with a subscription to American Psychological Association's PsycNET.

Purpose and description

PsycTESTS is a relatively new addition to the American Psychological Association (APA) stable of subscription database resources and is a research database focused on psychological assessments. It is intended to supplement PsycINFO, their flagship bibliographic database of psychological literature, but it can be used independently as well. It includes a mix of full-text and abstracted content. If a library has other APA databases, journal abstracts may cross-link to the full-text articles. This is a niche database that has no direct electronic peer or competitor and is intended to fill a gap in resources for available test references.

PsycTESTS has a wide range of included dates, subjects, and sources. Its earliest tests were first described in the literature in 1910, but more than half date since 2000, and it is updated monthly with new records and data. Its measures address development, identity, physical health, personality, neuropsychological issues, aptitude, competency, intelligence, resilience, educational levels, and so on. Tests and test information are sourced directly from authors, peer-reviewed journals, dissertations, books and handbooks, commercial test publishers, websites, and the Archives of the History of American Psychology Test Collection at the University of Akron in Ohio.

Content and audience

This database provides ''scholarly, published accounts of test development or assessment'' [1]. While it includes many records for commercially published tests, its focus is on unpublished tests, defined as those only referred to in journal articles, books, or gray literature. This information is of great interest to researchers and students in such areas as psychology, psychiatry, counseling, social work, occupational therapy, rehabilitation, education, and other related fields, but is frequently difficult to find.

Each test record includes APA-added descriptive metadata, a summary description and history, and all available reliability and validity data. Most records for noncommercial tests also include the actual test instrument and an assortment of linked records, such as relevant dissertations, technical reports, published reviews, related peer-reviewed literature, and so on. Most content is available in portable document format (PDF), with the exceptions of some tests that include multimedia or software content. APA has also included a record field for ''permissions,'' which explicitly marks each test as being freely available for reuse (most) or requiring author permission. As it is common for graduate students and other researchers to wish to use noncommercial tests in their research, this is likely to be well appreciated by users.

Usability and platforms

PsycTESTS is available from APA, EBSCO, ProQuest, and Ovid. As it is highly likely than any new subscriber would already have a subscription to PyscINFO and other APA databases, the database will be most usable for a particular library on whichever platform provides those databases. Based on the assumption that other APA databases are already subscribed to, most users new to it will find it familiar and simple to use. The reviewer used it on both EBSCO and Ovid's platforms and found that both provide clear, obvious search limiters that are specific to PsycTESTS, as well as their usual