Lab -- OpenRefine

OpenRefine Overview

Today, we'll navigate through our first hands-on lab. This allows for interactive learning, the slides are important here, so that you can follow along, click on links, and copy-paste during the demo.

I will start with a quick intro to OpenRefine. It's a fun and powerful tool for handling messy data. It's much more fun than Excel and I hope you will enjoy it. The slides are very important here, so that you can follow along, click on links, and copy-paste during the demo.

But first, we're just going to do a real quick kind of intro to OpenRefine so you can kind of understand use cases and why it's so awesome. And if you have questions, you know, just of course go ahead and let me know at any point.

Alright, so David Huynh, who's one of the original developers for OpenRefine, says it's a powerfuk tool for working with messy data and it's more powerful than a spreadsheet. It's **more interactive** and visual than scripting and it's better-suited for provisional exploratoration of your data, it's also more experimental and playful than a database.

Originally OpenRefine was developed by Google as GoogleRefine and it was supposed to be a kind of data tool that they that they developed and then in a certain point Google's funding for that project ran out and so it was transitioned to a open source project.

Now, it's called OpenRefine rather than Google refine but you're going to find a lot of tutorials who still call it Google Refine --good to know that they're the same thing.

So what does David Huynh mean by when he says it's a more powerful than a spreadsheet. Well, there's a lot more functionality in this than there is in say Excel or LibreOffice. You can do a lot of things with OpenRefine that you normally would have to use Python or R to do manipulation to transformations with data, so if a lot more powerful than using Excel.

It's also more interactive than scripting or cleaning your data with a programming language like Python or R. OpenRefine is especially useful because it is interactive, so you have all these abilities to actually explore your data without knowing what you want in advance. Openrefine it's really great at just being able to open something up and get a good sense of what's in there and explore it without really knowing anything about it.

It's going to look something like this when we're going to jump into it in just a few moments.

And it's going to say it's a **free open source extensible java app** that runs **offline** in your web browser. So what does that mean? Free and open source: *free* doesn't mean just that it doesn't cost anything, free also means that it is open source; the source code is there, you can modify it and build upon it -- this relates to OpenRefine's extensible quality as well. So people can add on plugins to this program and a lot of people have done so: they made different spinoffs, like a linked open data version or they made other modifications that make it work with certain types of data set.

It is a Java application that essentially runs a small server on your computer. This server serves the application into your web browser, allowing you to interact with it directly in the browser. However, it's important to note that it's not actually online; there's no connection to the outside world. It operates solely on your computer, making it a local application. This means you won't be putting your data at risk, and you don't need an internet connection to use it. Keep this in mind.

With OpenRefine you can handle all kinds of data. It represents data in a format that many are familiar with – tabular data. Once you import your data, it will appear in this tabular format. Open Refine uses specific terms for different aspects of this data. For example, in tabular data, what goes across is called a 'row', and what goes down a column is referred to as a 'column' in Open Refine. Each individual piece of data is called a 'cell'.

OpenRefine can import various formats, making it quite flexible. You can work with standard text-type data formats typically represented in tables, such as tab-delimited or comma-delimited files. It's also possible to define your own delimiter. Additionally, Open Refine can handle Excel, XML, JSON (which are not always tabular data but can be converted upon import), and RDF data.

One of the flexible aspects of OpenRefine is the variety of sources from which it can import data. You can upload files from your computer, a website URL, or even paste data directly into the application. It can connect with Google Drive to access data stored there. OpenRefine can also automatically open a zip archive, which is particularly handy if you have, for instance, an export from your data logger that's a zip archive of numerous text files. You can open these directly in OpenRefine. This feature is also useful when dealing with a directory of files, as you're not limited to opening one file at a time. Often, you might have a set of files with the same columns that you wish to use as a single file. Normally, this would require merging the files manually, but OpenRefine allows you to open them all at once in a batch, which is quite convenient.

It'll have really good performance for up to about 100,000 to hundreds of thousands of rows. You're going to experience good performance within this range. However, once you exceed that, it may start to slow down during certain operations. But, you can tweak your Java settings to enhance its capabilities, allowing it to handle millions of rows effectively. Additionally, there are some versions of it that have been developed to work in parallel processing environments, like Spark servers, for instance. These adaptations are particularly useful if you want to delve into more complex tasks and handle big data with it.

Use cases

- Clean your data: You get a dataset and you want to clean it up. You want to fix inconsistencies, use standard features, stats, and clustering transformation to fix your data.
- Transform your data: You have data in one form, and you need to reshape it to put it into some other form. For example, you might want to reshape it to put it into Tableau, or you want to reshape it and put it into Python just because it's easier to use. Sometimes you want to change the format from CSV into some kind of JSON with some changes on the way, and so it lets you do that in a real easy and visual way.
- Extend your data: It can actually reach out to your local system, collect data, scrape the web, reconcile with online databases, and you could geocode things like that. So, it's very good for enriching your data.

- Automate: Everything you're going to do with your data is going to be recorded as a routine. You can be able to grab the batch history and reapply it to new datasets as you go forward. So that's really a powerful thing if you're doing a lot of processing.

On top of data profiling and cleaning operations, OpenRefine extensions allow users to identify concepts in unstructured text, a process referred to as named-entity recognition (NER), and can also reconcile their own data with existing knowledge bases. By doing so, OpenRefine can be a practical tool to link data with concepts and authorities which have already been declared on the Web by parties such as Library of Congress or OCLC. Data cleaning is a prerequisite to these steps; the success rate of NER and a fruitful matching process between your data and external authorities depends on your ability to make your data as coherent as possible.

Resources

- OpenRefine Website
- Recommended book: Data Cleaning with OpenRefine

Messy Data

So what is messy data? This is kind of what we're talking about. And so, the first few rows here, this is kind of, you know, I work in text a lot, so my examples are going to be kind of text, but the same kind of problems apply to any bio-data or any other kind of data that you're working with. So, the top rows, all of the values are exactly the same essentially to a human, but to a computer and to your processing, these are all 100% different values, but they should be recognized by the computer as the same. So, we need to fix that inconsistency and format issues if we want to do analysis with this, otherwise, you're going to end up with bad results. So here, we have a date column with all different formats of the date, which I think was the last time I gave this workshop, which was October 2015.

Starting OpenRefine

- OpenRefine is a Java application, so you need to have Java installed on your computer. If you don't have Java installed, you can download it from java.com.
- To start OpenRefine, you can download it from the OpenRefine website. Once you have it installed, you can start it by clicking on the OpenRefine icon or by running the command `./refine` in the terminal.

1. NJ Shipwrecks

- We're going to start with a dataset of shipwrecks off the coast of New Jersey. This dataset is from the National Oceanic and Atmospheric Administration (NOAA) and is available on data.gov.

Creating a Project

You're going to create a project, open a project, import projects, and manage other settings. We'll start by creating a project. To do this, click on the 'Create Project' tab. One of the great features of OpenRefine is that it never alters the original data you provide. When you upload a file, OpenRefine creates a copy and then converts it into its unique format, which is stored in an archive image directory. This means all the

projects you've worked on in the past are saved and accessible, but your original file remains unchanged. This is a key difference from Excel, which often modifies the original file upon opening.

Next, we'll import data from this computer. OpenRefine is flexible and can import data from various sources. Choose the file you've recently downloaded and click 'Next'. Although it says 'uploading data', it's not actually uploading anywhere; it's just loading it into the OpenRefine app. OpenRefine will give you a preview of its first parsing option, which is usually correct, but you can change the character encoding if necessary. This is particularly important if you're working on Windows, as most text-based files will be in UTF-8, which Excel does not handle well.

In the parsing options, you can also change the file type if OpenRefine hasn't detected it correctly. An important option is 'Parse next _ lines as column headers', which you can turn on or off depending on whether your file has column headers. Once you've set these options, rename the project to something meaningful to help you remember what you're working on.

After creating the project, OpenRefine will display your data in a tabular format, similar to a spreadsheet. You have options to adjust how many rows to show, which is useful for large datasets. Remember, you'll primarily be performing operations on columns, so you're getting a high-level view of the dataset rather than working cell by cell.

Learning Objectives

- Understand the basic functionalities of OpenRefine
- Learn how to clean and transform data using OpenRefine
- Learn how to export data from OpenRefine
- Learn how to use OpenRefine to reconcile data with external sources
- Learn how to use OpenRefine to create a new project
- Learn how to use OpenRefine to import data from a URL

Agenda

When working with data, it is almost always necessary to spend a good portion of time cleaning and standardizing that data for analysis or presentation. This is especially true when re-purposing data created by others (always ensure you have permission and give due credit). Useful resources for this work include:

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data. Authority Lists: wherever possible you should use existing standards and commonly used authority files in order to create interoperable data. For example, VIAF: The Virtual International Authority File and Geonames for people and place identifiers.

XML to Excel

Things to do

1. Look at the Excel file, what do you see?

2. What are some of the issues with the data?

    - How can we find what issues there are?
    - Facets
    - How many divided backs are there?
    - How many undivided backs are there?
    - How many with a message?
    - How many without a message?
    - How many with a stamp?
    - How many in color, black and white,

3. Things to fix

- The column CopyrightDate is not in the right format (text, but it should be a date)
- The Publisher column has some issues -- various spellings of the same publisher
- Can we visualize what subjects are there? What format is this in column in?
- If statue is a column a
- Often lcsh geographic has double metnions...
- The Dinky now is not where the Dinky used to be... postcards before a specific time should have the old location
- Genre has two postcards with missing values!
- For some postcards AAT topics not filled in
- Add a column with VIAF for statues persons

Assign one column to a group of three students