|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  | **Humanities Data Praxis**  **10 March 2017** |
| Normalizing Metadata Using Open Refine |  | Instructor: Andrea Thomer Email: [thomer2@illinois.edu](mailto:thomer2@illinois.edu)  Thanks to Megan Senseney for originally creating this handout |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Creating a New Project  1. Choose ‘Create Project’ from the menu of the right. 2. Make sure the NYPL Menus Data file has been unzipped into a folder somewhere on your machine. 3. It’s best to upload files one at a time. We recommend starting with ‘Menu.csv’ 4. OpenRefine will parse the file and ask you to choose some options.     The most important of these options are:   * Set character encoding as ‘utf-8’ * Make sure that ‘commas (CSV)’ are selected as a column separator * Make sure that 1 line from the file is being parsed as the column header   Click ‘Update Preview’ and take some time to review. Speak up if something doesn’t look right!   1. Give your project a name and click the ‘Create Project’ button at the top right. |  | Installations  |  | | --- | | OpenRefine v. 2.6 (formerly Google Refine) **http://openrefine.org**  RDF Refine Extension  **http://refine.deri.ie/** Datasets Download from  [**https://github.com/elliewix/Humanities-Data-Praxis**](https://github.com/elliewix/Humanities-Data-Praxis)  **We'll be using "menu.csv"** | |

|  |  |  |
| --- | --- | --- |
| Basic Normalization  1. Select the downward pointing arrow next to the name of a column 2. Go to Edit cell > Common transforms 3. Try a few of the following options, and see how they affect your data  * Trim leading and trailing whitespace * Collapse consecutive whitespace * Pick a capitalization scheme  Facets and Clustering  1. Starting with the drop-down menu for the ‘sponsor’ column, go to Facet > Text facet 2. Inspect the values that appear in the box on the right (try sorting by ‘count’) 3. Once you’ve reviewed the values, click the ‘Cluster’ button – a menu will appear 4. You can select or supply the preferred value for a cluster of similar items in the text box, then select the check box under ‘Merge?’ 5. Do a few clusters then click the ‘Merge Selected & Re-Cluster’ button   *This exercise was initially created at the Maryland Institute for Technology in the Humanities and developed through the Digital Humanities Data Curation Institute series, generously funded by the National Endowment of the Humanities. Special thanks to Trevor Muñoz and Lydia Zvyaginsteva!* |  | Further Readings**OpenRefine Manual** Verborgh, R. & De Wilde, M. (2013). Using OpenRefine. Birmingham: Packt Publishing. **OpenRefine for DH Projects** Curating Menus’ three-part post “More work with NYPL’s open data”: **http://www.curatingmenus.org**  Programming Historian’s lesson “Cleaning data with OpenRefine”: **http://programminghistorian.org/lessons/cleaning-data-with-openrefine** **OpenRefine Community** GitHub User Documentation: **https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users**  OpenRefine Google Group: **https://groups.google.com/forum/?fromgroups#!forum/openrefine**  OpenRefine on Twitter: **https://twitter.com/OpenRefine**    Untitled:private:var:folders:p1:yvr7s6cx2bq3d58x_n3kf1vm0000gp:T:TemporaryItems:wordmark_black.png |