Created by Steven Braun
Last updated 2015-12-10 (Steven Braun)

Mapping Trends in Data Curation in the CLIR Postdoctoral Fellowship Program

A general framework for creating a guided interactive visualization interface delivered over the web

Documentation

The content of this documentation provides information about the data sources, tools, and processes/methodologies used to analyze trends in data curation fellowships offered through the Council on Library and Information Resources (CLIR). The code for this project is being released as a general framework for creating a guided interactive visualization interface that can be delivered via the web.

CONTENTS

PROJECT OVERVIEW

The objective of this project is to produce an interactive visualization illustrating the emerging importance of data curation activities in CLIR postdoctoral fellowship programs as well as the breadth of external projects in which CLIR fellows have been involved over time. The final product was initially presented to the CLIR board of directors November 19-20, 2015, but later audiences may include funders and prospective fellows and institutional hosts. This project is a collaborative effort between **Alice Bishop** (abishop@clir.org, *CLIR*), **Steven Braun** (braunsg@gmail.com OR s.braun@northeastern.edu, *Northeastern University Libraries*), **Justin Schell** (jmschell@umich.edu, *University of Michigan Libraries*), and **Rita Van Duinen** (RVanDuinen@clir.org, *CLIR*).

GITHUB REPOSITORY

The full open source code for this project is distributed under a GNU General Public License (version 2) and is accessible on GitHub at the following repository address:

https://github.com/braunsg/clir-data-curation-viz

DATA SOURCES

Two primary textual sources are considered in this project. The first consists of the full text for all fellowship position postings since 2012. These data were scraped from online postings, cleaned, and organized for database storage (see Database Structure). The second source of data consists of projects, microgrants, and publications to which CLIR data curation fellows have contributed. These data were also cleaned and structured for database storage. All data sources were provided by Alice Bishop and Rita Van Duinen with additional cleaning performed by Justin Schell.

In addition to these textual sources, biographical data about CLIR postdoctoral fellows since 2012, including names, Ph.D. disciplinary foci, and host institutions, were also provided by Alice Bishop and Rita Van Duinen.

For position postings, Justin transformed the source documents into plain text (.txt) format using TextWrangler, stripping out any formatting that could interfere with analyses. The resulting text files were combined together via command line.

PACKAGE CONTENTS

This package includes a collection of scripts, sample data, and other files that are used to produce the final visualization product. These contents are described below.

Data analysis scripts and files

The following scripts, found in the subfolder /scripts, were used for all data analysis processes:

Script or file name	**Script description**
mysql-token-analysis.php	Performs token frequency analyses, *i.e.*, finding all unique tokens across all data (fellowship position descriptions), computing their frequency, capturing their semantic context, and injecting the results into the database to be transformed as node and link data for network representation
mysql-token-links.php	Calculates co-occurrence distances between highest-frequency tokens determined by msyql-token-analysis.php and transforms those distances into link records for network analyses
mysql-token-count-over-time.php	Calculates token frequencies per year (since 2012), thus giving a temporal description of how token counts have changed over time through position descriptions from each cohort
mysql-reduce-nodes-links-data.php	Collapses token node and link data to a reduced form to facilitate faster processing with the semantic network visualization, based on specifying a maximum token frequency, maximum token node count, and inter-token distance threshold
mysql-inject-fellow-data.php	Injects biographical data about fellows into the database, based on data retrieved from a file; this is a custom script particular to the form of biographical data provided but is included for legacy purposes
mysql-inject-lat-long.php	Injects latitude and longitude data for all academic institutions represented by fellows, including both their Ph.D.-granting institutions and fellowship host institutions
mysql-update-pubs-table-structure.php	Updates the table structure for fellow publications and projects; not necessary for data analyses, but included for legacy purposes
inc/default-config.php	Stores global configuration parameters for the scripts above, including parameters for MySQL database connections
inc/stopwords.txt	A list of English stop words excluded from token analyses, taken from http://www.ranks.nl/stopwords/

Sample data

The subfolder /sample-data includes example data used for token analyses, specifically plain text versions of a handful of fellowship position postings.

Table definitions

MySQL table definitions (queries), along with dummy data, are included in the subfolder /table-definitions. These scripts may be run directly in a MySQL environment to generate the tables necessary for the analysis processes to work.

Interface files

The subfolder /interface includes all of the web components (HTML/PHP files, CSS stylesheets, Javascript libraries, and visualization scripts) required to run the visualization interface via a web server. More information about the interface can be found in the Interface section below.

ANALYSIS OVERVIEW

In this section, a summary of the main analytics processes carried out in this project is provided.

Token Analyses

The visualizations produced in this project are the result of a standard process that 1) determines the natural frequency of unique terms and keywords (henceforth “tokens”) in the source texts, and 2) calculates the extent to which those tokens appear together in any given text as a measure of the word distance between them. Here, “text” refers to the text derived from position description postings for CLIR fellowships (See Data). The steps taken in this process are described below.

Step	Process description	File name	Notes
1	The text to be analyzed is split by the parser (script, mysql-token-analysis.php) into individual sentences/lines, the ends of which are demarcated by a period (.) or a line break (carriage return)	mysql-token-analysis.php	Before this splitting occurs, any and all instances of “Ph.D.” (and any derivatives such as “Ph.D”) in the text are converted to “PhD” (without periods) to reduce duplicative line breaks resulting from non-stopping punctuation
2	Each resulting line is stripped of excess punctuation, *e.g.*, commas, semicolons, and question marks; hyphens are handled specially and transformed into spaces	mysql-token-analysis.php	
3	Each line (sentence) is split into individual word tokens by breaking at each space between words	mysql-token-analysis.php	
4	The parser loops through each individual token word, transforms it to lowercase, and compares it to a standard list of English token exceptions (*i.e.*, stop words), such as is, are, and be, as well as a special list of exceptions specific to the data set (such as university, phd). If the token is found in the list of exceptions, it is encoded as a stop word and recorded in a pool of “cleaned tokens”; if the token is not found in the list of exceptions, the token is encoded without flagging to ensure it receives further analysis	mysql-token-analysis.php	See the Appendix for a full list of stop words used
5	Once all tokens in the selected text have been compared to the list of exceptions, the parser loops through each non-exception token and records the number of instances within the text of each unique token; this procedure is continued for token window lengths of 1, 2, and 3 words in sequence and across all texts, producing a single list of all terms/keywords ranked	mysql-token-analysis.php	The list of non-exception tokens and their frequencies is compiled across all position descriptions or fellow biographies

	by frequency	
6	For each token recorded in the resulting list above, a new parser (script, mysql-token-count-over-time.php) calculates their frequency of appearance within fellowship/position postings for each consecutive year since 2012	mysql-token-count-over-time.php
7	For these same tokens, a new parser (script, mysql-token-links.php) determines the number of times the term appears in the same text (*i.e.*, position description or fellow biography) with all other terms in the list within a specified link distance threshold (here, a maximum distance of 3 intervening words between two tokens)	mysql-token-links.php
8	For each full dataset, the parser uses the results of the previous step to produce a list of nodes (terms/keywords) and links (two terms/keywords appearing within the same text within a specified distance threshold) that are stored in the database for visualization and analysis; these lists are reduced to a specified maximum number of nodes to enhance data access and reduce memory overhead for the web application/interface	mysql-token-links.php, mysql-reduce-nodes-links-data.php

A concrete example of the procedure described above will now be provided. Consider the following text, which is the biography for Inna Kouper (CLIR fellow, 2012-2014; note that while biographies were not used as a final data source for analysis, the following example serves the same illustrative purpose):

```
Inna Kouper (Indiana University) received her Ph.D. in
Information Science from Indiana University. Her fellowship was
based within the Data to Insight Center (D2I) at IU and focused
on collaborative initiatives between D2I and the IU Libraries.
These included cornerstone research projects of the center such
as the National Science Foundation-funded DataNet SEAD
(Sustainable Environment-Actionable Data) Virtual Archive, the
HathiTrust Research Center, and work on non-consumptive research
methodologies funded by the Alfred P Sloan Foundation.
```

In the first step, beginning with the parser script *mysql-token-analysis.php*, the original text is split into separate lines based on breaks indicated by periods (.). Note that before this happens, the instance of the term "Ph.D." is transformed to "PhD" (without periods) to prevent the parser from splitting the text in the middle of the term. For the example provided, the result is an array of three lines (sentences):

```
Array [
        "Inna Kouper (Indiana University) received her PhD in
Information Science from Indiana University",
      "Her fellowship was based within the Data to Insight Center
(D2I) at IU and focused on collaborative initiatives between D2I
and the IU Libraries",
      "These included cornerstone research projects of the center
such as the National Science Foundation-funded DataNet SEAD
(Sustainable Environment-Actionable Data) Virtual Archive, the
HathiTrust Research Center, and work on non-consumptive research
methodologies funded by the Alfred P Sloan Foundation"
]
```

In the next step, the parser loops through each line and removes any extra punctuation that may be accidentally included as belonging to individual words. Note that hyphens (-) are specially handled and replaced with spaces instead of being deleted, as is illustrated here (using the third line from the original biography text):

> These included cornerstone research projects of the center such
> as the National Science Foundation-funded DataNet SEAD
> ~~(~~Sustainable Environment-Actionable Data~~)~~ Virtual Archive~~,~~ the
> HathiTrust Research Center~~,~~ and work on non-consumptive research
> methodologies funded by the Alfred P Sloan Foundation

The result of this deletion/replacement is the string

```
These included cornerstone research projects of the center such
as the National Science Foundation funded DataNet SEAD
Sustainable Environment Actionable Data Virtual Archive the Hathi
Trust Research Center and work on non consumptive research
methodologies funded by the Alfred P Sloan Foundation
```

Note that the decision to split the text by line breaks is not arbitrary but rather critical to disambiguating false term sequences that might be detected as the result of combining the final word of one sentence with the beginning word of the next. For example, a sentence ending with the word "data" and followed by a sentence beginning with "management" would erroneously encode an instance of "data management" for the parser when the individual components are not actually semantically related in the original text structure.

The parser now loops through each line, splitting them into individual word tokens by breaking at each space. The result is an array of the form

```
Array ["These", "included", "cornerstone", "research", "projects",
"of", "the", "center", "such", "as", "the", "National", "Science",
"Foundation", "funded", "DataNet", "SEAD", "Sustainable",
"Environment", "Actionable", "Data", "Virtual", "Archive", "the",
"HathiTrust", "Research", "Center", "and", "work", "on", "non",
"consumptive", "research", "methodologies", "funded", "by", "the",
"Alfred", "P", "Sloan", "Foundation"]
```

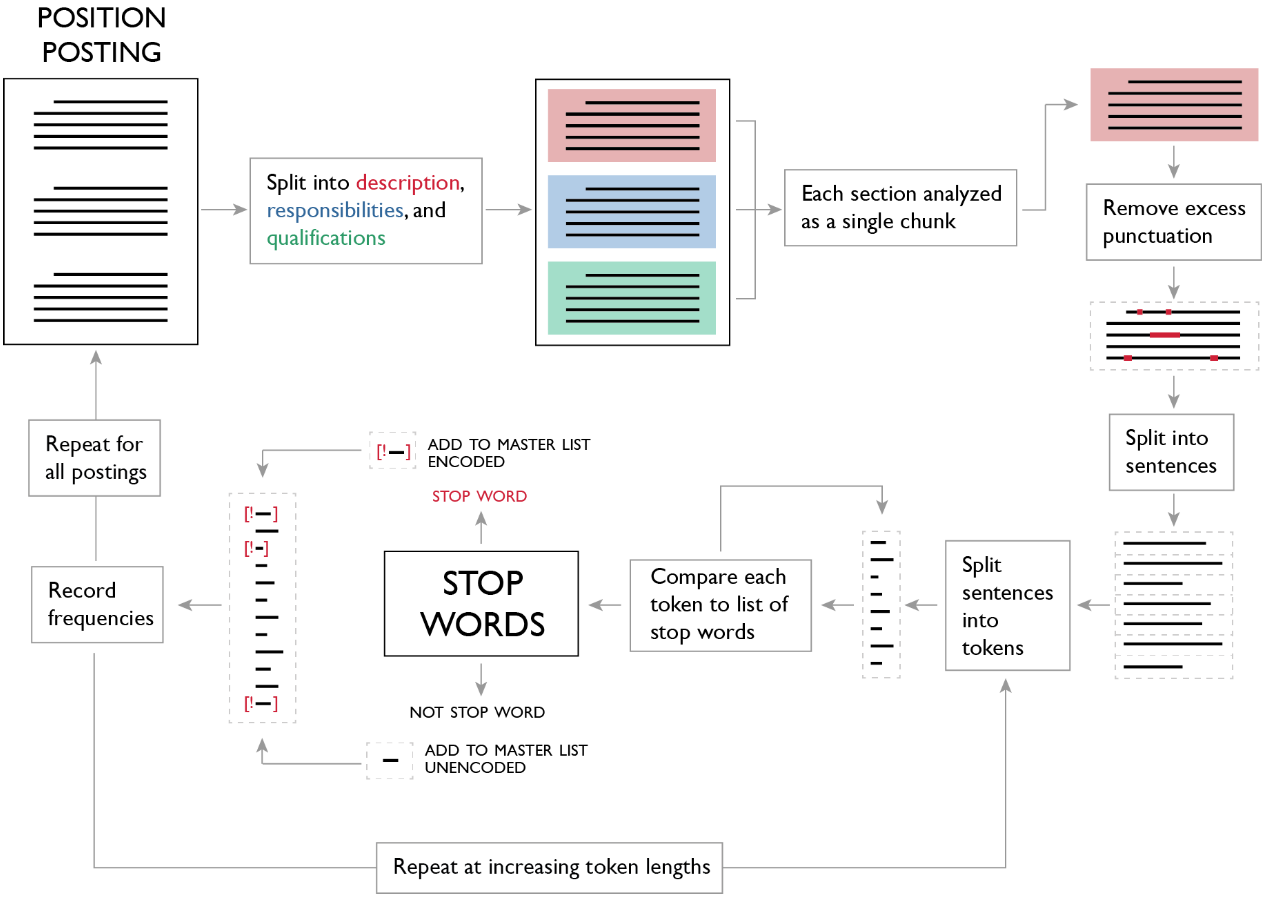
This array is then looped through one by one. For each token, the parser converts the word to lowercase and compares it with a list of special token exceptions that includes both standard English stop words (such as of, the, as) as well as special tokens that appear with high frequency but are semantically irrelevant (such as university, phd). (See the Appendix for a full list of stop words used.) If the word is found in the list of token exceptions, the parser encodes the word with an exclamation point within square brackets (for example, “the” becomes [!the]) and adds the encoded token to a new array of “cleaned” tokens. If the word is not found in the list of exceptions, it is added to the cleaned array without any such encoding. In our example, this array of cleaned tokens would look like the following:

```
Array [  [0] => “[!these]”
         [1] => “included”,
         [2] => “cornerstone”,
         [3] => “research”,
         [4] => “projects”,
         [5] => “[!of]”,
         [6] => “[!the]”,
         [7] => “center”,
         [8] => “[!such]”,
         [9] => “[!as]”,
         [10] => “[!the]”,
         [11] => “national”,
         [12] => “science”,
         [13] => “foundation”,
         [14] => “funded”,
         [15] => “datanet”,
         [16] => “sead”,
         [17] => “sustainable”,
         [18] => “environment”,
         [19] => “actionable”,
         [20] => “data”,
         [21] => “virtual”,
         [22] => “archive”,
         [23] => “[!the]”,
         [24] => “hathitrust”,
         [25] => “research”,
         [26] => “center”,
         [27] => “[!and]”,
         [28] => “work”,
         [29] => “[!on]”,
         [30] => “non”,
         [31] => “consumptive”,
         [32] => “research”,
         [33] => “methodologies”,
         [34] => “funded”,
         [35] => “[!by]”,
         [36] => “[!the]”,
         [37] => “alfred”,
         [38] => “p”,
         [39] => “sloan”,
         [40] => “foundation” ]
```

Note that this array of cleaned tokens retains the original index position (starting with 0) of each token from the source text.

Next, the parser moves through each non-stopping (*i.e.*, unflagged) token in the array and compiles the number of instances of each such unique term in the full data set. This process is continued across all texts in the data set (*i.e.*, all fellowship position descriptions, responsibilities, and qualifications sections; see Sample Data) and at progressively longer window lengths, up to a maximum length of 3. This means the parser detects the number of times singular words such as data occur as well as the number of times strings of singular words occur, such as is the case with expressions like data curation, data management, and research data. While singular stop words are not encoded as unique tokens in this analysis step, multistring tokens may include stop words (for example, "methodologies funded [!by]").

The following figure provides a schematic description of the process described above.


POSITION POSTING
Split into description, responsibilities, and qualifications
Each section analyzed as a single chunk
Remove excess punctuation
Split into sentences
Split sentences into tokens
Compare each token to list of stop words
STOP WORDS
STOP WORD
[!—] ADD TO MASTER LIST ENCODED
NOT STOP WORD
— ADD TO MASTER LIST UNENCODED
[!—]
[!-]
[!—]
Record frequencies
Repeat for all postings
Repeat at increasing token lengths

A unique list of extracted tokens is created separately for position descriptions, responsibilities, and qualifications. Once this process is completed for all tokens, a new parser (*mysql-token-links.php*) determines the frequency with which pairs of terms/keywords appear together in the same categorized text source within a single position posting within a specified distance threshold, using as a basis the top-ranked keywords across the entire dataset (*e.g.*, the 100 most-frequent terms across all position qualifications sections). Consider the following dummy example sentence:

```
John Doe has experience working with data management as well as
extensive expertise with biomedical data and guiding researchers
through metadata and curation practices
```

For this text, let us assume that the terms curation, collaboration, experience, data, and metadata are on the list of most-frequent keywords for the total dataset. In this step, the parser loops through each of these terms and calculates word distances between them and all other tokens in the list. For example, the token list and example sentence above would yield the following token distances:

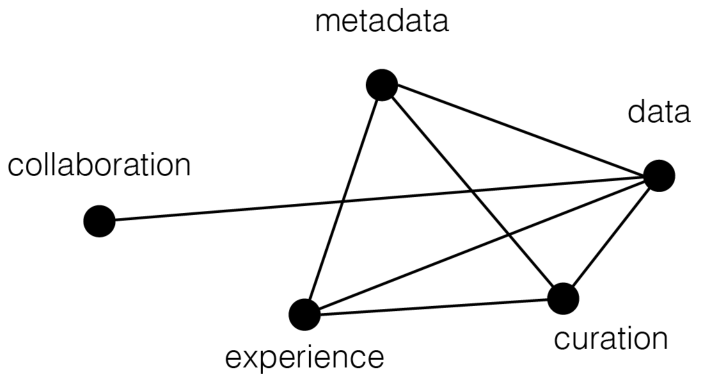
Token 1	Token 2	Distance
curation	collaboration	None
curation	experience	19
curation	data	16
curation	data	7
curation	metadata	2
collaboration	experience	None
collaboration	data	None
collaboration	metadata	None
experience	data	3
experience	data	12
experience	metadata	17
data	metadata	14
data	metadata	5

Token distances are measured as the number of singular tokens (non-compound words) between two tokens. When two tokens are adjacent in a sentence (such as "data" and "management" together as "data management"), token distance defaults to 1.

In the table above, note that there are no token distances for any calculation involving the word "collaboration" since this term does not appear in the given sentence. Also note that some word pairs have multiple distances, such as is the case with any pair including the word "data," since this particular token occurs more than once in the sentence. Given these two observations, extending this token distance calculation process across all sentences and across all text sources results in a long list of distances between all possible token pairs.

We can subsequently graph connections between these tokens as a measure of their token distance. Specifically, in constructing a node and link diagram, individual tokens can act as nodes while links between nodes may be derived from the token distances between them. For simplicity, we can limit these links to only pairs of tokens that have token distances less than or equal to a specified distance threshold (here, this threshold is a maximum token distance of 3). For any pair of tokens with a distance within this threshold, we can count the number of instances in which the pair appears with the given distance across all texts and use these counts as a measure of relative link strength or link distance. In such a diagram, pairs of tokens that frequently appear together within the specified token distance threshold will be closer in proximity to one another than other pairs. The resulting diagram is a network of tokens and links between them indicating highly correlated co-occurrences.

With a sufficiently large set of data, a graph of the following form may be produced:



Again, nodes represent individual tokens, and links represent pairings of tokens with token distances that fall within the specified threshold. Pairs of nodes that are closer together (*e.g.*, "data" and "curation") represent pairs of tokens that appear together at short distances with high frequency. Note that this network is undirected and thus any non-unique pairings (*e.g.*, experience – curation vs. curation – experience) are not included as duplications/redundancies in the final analyses.

For the purposes of the final visualization product, the resulting network diagram was optimized by reducing the number of nodes (tokens) to a set number (50) for each data source type (position descriptions, responsibilities, and qualifications) and limiting connections between tokens to the specified token distance threshold (3).

Geocoding

Latitude and longitude coordinates for Ph.D.-granting and host institutions were retrieved by Justin Schell using an online tool available at [http://www.findlatitudeandlongitude.com/batch-geocode/]. Institutions were searched for by name. Justin subsequently confirmed the returned geocoordinates by mapping them and corrected any anomalies that were found.

DATABASE STRUCTURE

CLIR position description and fellow biographical data have been structured and stored in a MySQL database for ease of access and analysis. This database, named clir, has the following table structure:

Table name	**Table description**
fellows_data	Data about current and past CLIR postdoctoral fellows
institution_data	Data about Ph.D.-granting and fellowship host institutions
links_distance	Calculated distances between pairs of tokens
links_distance_reduced	A reduced form of links_distance, holding distances between pairs of tokens within the specified distance threshold (threshold = 3)
position_descriptions	Content from position descriptions used to recruit CLIR postdoctoral fellows

publications_projects	Data about external projects and publications with which CLIR fellows have been involved
tokens	Data for all unique tokens (singular and multi-length) captured in the data analysis processes
tokens_reduced	A reduced form of tokens, holding data for only the top 50 most frequent tokens across each data/text source
token_counts_over_time	Data about token frequencies across position postings within single years since 2012, limited to the top 200 most frequent tokens across each data/text source
token_instances	Data about high-frequency token instances in the original texts; *i.e.*, original textual context for individual tokens across all position postings

TABLE DEFINITIONS

The following section provides more detailed notes about fields defined for each table listed above.

fellows_data

Field name	**Field description**
fellow_row_index	(int) An auto-incremented counter for the given fellow/record
fellow_id	(varchar) A unique semantic + numeric identifier for the given fellow
f_name	(varchar) The fellow's full first and last name
f_startyear	(year) The start year of the fellow's fellowship
f_endyear	(year) The end year of the fellow's fellowship
f_phd	(varchar) The fellow's doctoral disciplinary focus (Ph.D. area)
f_phd_class	(varchar) A flag indicating whether the fellow's disciplinary focus is in the arts and humanities (AH), social sciences (SS), or natural sciences (NS)
f_institution	(varchar) The institution from which the fellow received their Ph.D.
i_id	(varchar) The unique identifier for the fellow's Ph.D.-granting institution
host_institution	(varchar) The fellow's host institution
host_i_id	(varchar) The unique identifier for the fellowship host institution
f_location_notes	(varchar) Special notes about the Ph.D.-granting institution location
fellow_bio	(varchar) A short biographical text of the fellow, where applicable and given
fellow_dc	(tinyint) A flag indicating whether or not the individual is a data

	curation fellow (0 = no, 1 = yes)
fellow_dc_track	(varchar) The data curation track name, if applicable

institution_data

Field name	**Field description**
i_row_index	(int) An auto-incremented counter for the given institution
i_id	(varchar) A unique semantic + numeric identifier for the institution
i_name	(varchar) The full name of the institution
i_dc_host_status	(varchar) A flag indicating whether or not the institution is the host site for a data curation fellowship ("Y" = yes, "N" = no)
i_lat	(varchar) The institution's latitudinal coordinate
i_lon	(varchar) The institution's longitudinal coordinate

links_distance

Field name	**Field description**
token_link_row_index	(int) An auto-incremented counter for the given token pair
source_id	(varchar) The unique identifier for the source token, taken from the table *tokens*
target_id	(varchar) The unique identifier for the target token, taken from the table *tokens*
link_distance	(int) The calculated distance between the two tokens
weight	(int) The weight of the linked pair as a function of the pair's distance, *i.e.*, the number of times the given pair of tokens occurs at the specified distance across all text sources
link_source_type	(varchar) A flag indicating the text source to which the link is applicable ("pds" = position descriptions, "responsibilities" = position responsibilities, "qualifications" = position minimum and preferred qualifications)

links_distance_reduced

Field name	**Field description**
token_link_rd_row_index	(int) An auto-incremented counter for the given (reduced) token pair
source_id	(varchar) The unique identifier for the source token, taken from the table *tokens*
target_id	(varchar) The unique identifier for the target token, taken from the

	table *tokens*
link_distance_threshold	(int) The maximum token distance threshold at which the given pair is being analyzed and stored (only token pairs that occur within this distance threshold are counted and stored in this table)
weight	(int) The weight of the linked pair as a function of the pair's distance, *i.e.*, the number of times the given pair of tokens occurs at the specified distance across all text sources
link_source_type	(varchar) A flag indicating the text source to which the link is applicable ("pds" = position descriptions, "responsibilities" = position responsibilities, "qualifications" = position minimum and preferred qualifications)

position_descriptions

Field name	**Field description**
pd_row_index	(int) An auto-incremented counter for the given position posting record
pd_id	(int) A unique semantic + numeric identifier for the given fellowship position description
pd_dc	(tinyint) A flag indicating whether or not the fellowship position is on the data curation track (0 = no, 1 = yes)
pd_institution	(varchar) The name of the institutional host of the fellowship
pd_institution_description	(varchar) Descriptive information about the institutional host
pd_startyear	(year) The start year of the fellowship
pd_endyear	(year) The end year of the fellowship
pd_title	(varchar) The formal title of the fellowship
pd_description	(varchar) The central descriptive text about the fellowship
pd_responsibilities	(varchar) Fellowship responsibilities listed for the position
pd_qualifications_min	(varchar) Minimum qualifications listed for the position
pd_qualifications_preferred	(varchar) Preferred qualifications listed for the position, if applicable
pd_discipline	(varchar) The intended disciplinary focus for the position, if applicable
pd_addl_info	(varchar) Any additional information about the fellowship that does not fit in other fields
pd_url	(varchar) The URL for the position description

publications_projects

Field name	**Field description**
pub_row_index	(int) An auto-incremented counter for the given project/publication record
pub_id	(varchar) A unique semantic + numeric identifier for the given project/publication
pub_citation	(varchar) The full citation for the project/publication
pub_title	(varchar) The title of the project/publication
pub_type	(varchar) A short semantic descriptor of the project/publication type (*e.g.*, publication, presentation)
pub_link	(varchar) The URL for the project/publication, if applicable
pub_year	(year) The year of publication or start year of the project
cohort_year	(year) The fellowship cohort year of the author(s) of the project/publication; in the case of fellows from multiple years, defaults to the start year of the individual in the earliest cohort represented
contributor_fids	(varchar) A list of fellow IDs indicating fellows who contributed to the given project/publication
pub_display	(tinyint) A flag indicating whether or not the project/publication should be displayed in the visualization interface (0 = no, 1 = yes)

tokens

Field name	**Field description**
token_row_index	(int) An auto-incremented counter for the given token record
token_id	(varchar) A unique semantic + numeric identifier for the given token record
token_label	(varchar) The token string
token_display	(tinyint) A flag indicating whether or not the token string should be displayed in the product visualization (0 = no, 1 = yes)
token_length	(int) The length (word count) of the token string
token_stop_words_ct	(int) The number of stop words included in the token string
token_instance_ct	(int) The total number of times the token string appears across all specified source texts
token_entity_ct	(int) The total number of source texts in which the token string appears at least once
token_coverage_pct	(float) The percentage of source texts in which the token string appears at least once, represented as a float (1.0 = 100% of source

	texts)
token_source_type	(varchar) A descriptive flag indicating the source text class in which the token string appears ("pds" = position descriptions, "qualifications" = position qualifications, "responsibilities" = position responsibilities)

tokens_reduced

Field name	**Field description**
token_id	(varchar) The unique identifier for the given token string
token_label	(varchar) The token string
token_length	(int) The length (word count) of the token string
token_instance_ct	(int) The total number of times the token string appears across all specified source texts
token_coverage_pct	(float) The percentage of source texts in which the token string appears at least once, represented as a float (1.0 = 100% of source texts)
token_source_type	(varchar) A descriptive flag indicating the source text class in which the token string appears ("pds" = position descriptions, "qualifications" = position qualifications, "responsibilities" = position responsibilities)

token_counts_over_time

Field name	**Field description**
token_count_row_index	(int) An auto-incremented counter for the given token count record
token_id	(varchar) The unique identifier for the given token string, taken from the table *tokens* (or *tokens_reduced*)
count_2012	(int) The total number of times the token string appears in all source texts from the year 2012
count_2013	(int) The total number of times the token string appears in all source texts from the year 2013
count_2014	(int) The total number of times the token string appears in all source texts from the year 2014
count_2015	(int) The total number of times the token string appears in all source texts from the year 2015
pct_2012	(float) The total percentage of source texts from the year 2012 in which the token string appears at least once
pct_2013	(float) The total percentage of source texts from the year 2013 in

	which the token string appears at least once
pct_2014	(float) The total percentage of source texts from the year 2014 in which the token string appears at least once
pct_2015	(float) The total percentage of source texts from the year 2015 in which the token string appears at least once

token_instances

Field name	**Field description**
token_instance_row_index	(int) An auto-incremented counter for the given token instance record
token_id	(varchar) The unique identifier for the given token string, taken from the table *tokens*
token_pretext	(varchar) The pretext string that immediately precedes the token string in this instance
token_posttext	(varchar) The string that immediately follows the token string in this instance
entity_id	(varchar) The unique identifier for the entity from which this instance is recorded (*i.e.*, the position posting), taken from the table *position_descriptions*
entity_type	(varchar) A descriptive flag indicating the source text class in which the token string appears ("pds" = position descriptions, "qualifications" = position qualifications, "responsibilities" = position responsibilities)

VISUALIZATION INTERFACE

The results of the data analysis processes described above provide the foundation for the visualization interface, the structure and implementation of which is described in this section. The interface can currently be accessed at http://www.stevengbraun.com/dev/clir/index.php.

Overview

The final product is designed as a guided interactive visualization and is composed of a series of "frames," each of which holds a single visualization. Using guided navigation, the interface leads the user through these frames and their animations one by one and allows the user to pause at various steps along the way to explore the data in more detail.

The title of the visualization series is "Ecologies of Innovation and Discovery: A Snapshot of CLIR Postdoctoral Fellowships in Data Curation." The following table lists and describes these frames:

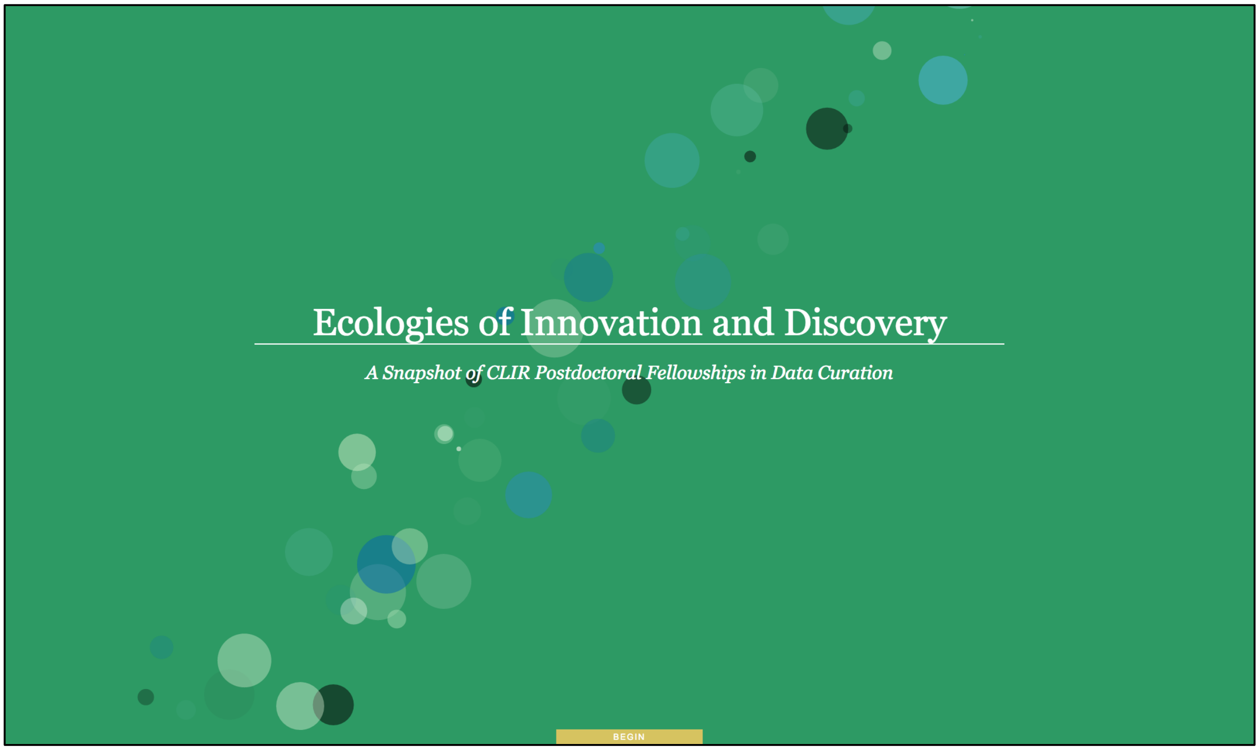
Frame name	**Frame description**
Introduction *frames/full_title.php*	The title frame and the entry point for the user
Fellows Map *frames/fellows_map.php*	An interactive and dynamic visualization that shows fellows' Ph.D.-granting institutions, their disciplinary foci, fellowship host institutions, and collaborative connections between them on publications and projects
Language of Data *frames/token_count_descriptions.php*	A series of dynamic charts showing token string patterns (e.g., high-frequency tokens) across the source texts, including an option that allows users to view token strings in their original textual context
Semantic Network *frames/force_static_div.php*	A force-directed node and link network diagram showing high-frequency tokens across position descriptions, qualifications, and responsibilities sections as well as connections between them as measured by proximity within the source texts
Acknowledgments *frames/closing.php*	A concluding slide with acknowledgments

Architecture and Libraries/Dependencies

The visualization interface is built on a standard LAMP architecture, with the output view and internal data processes described with PHP, data stored in a MySQL database, an Apache server handler, and a Linux environment. The visualizations themselves are constructed using D3.js version 3.5.10 (http://d3js.org/) and TopoJSON version 1.0 (https://github.com/mbostock/topojson) with additional functionality achieved through the use of jQuery version 1.11.2 (http://jquery.com/).

APPENDIX

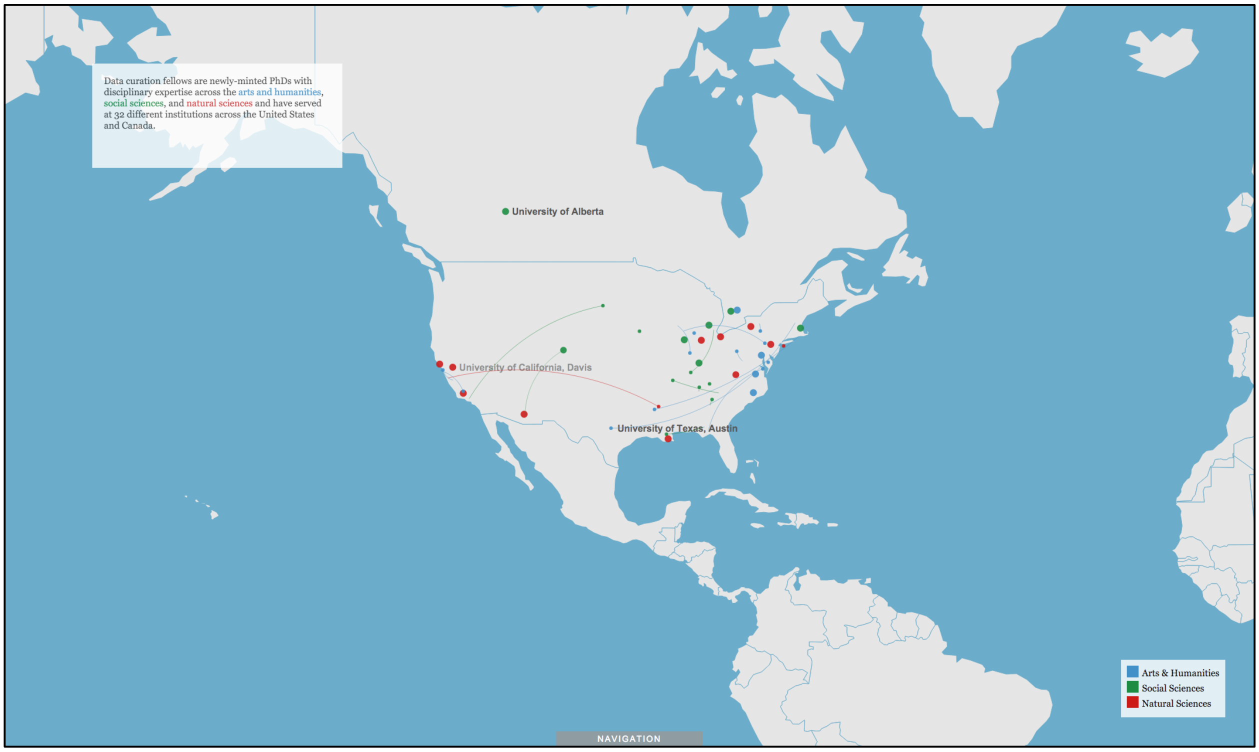
Screenshots



Ecologies of Innovation and Discovery

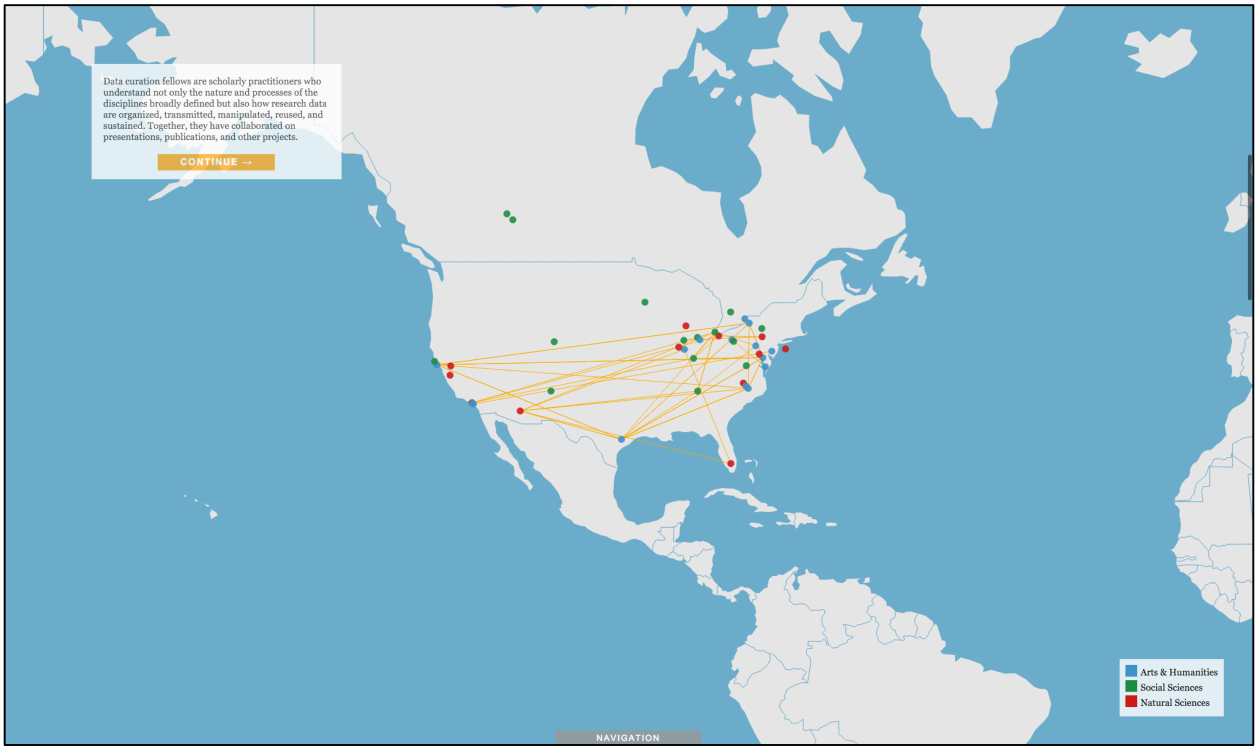
A Snapshot of CLIR Postdoctoral Fellowships in Data Curation

Title slide; landing page for the user



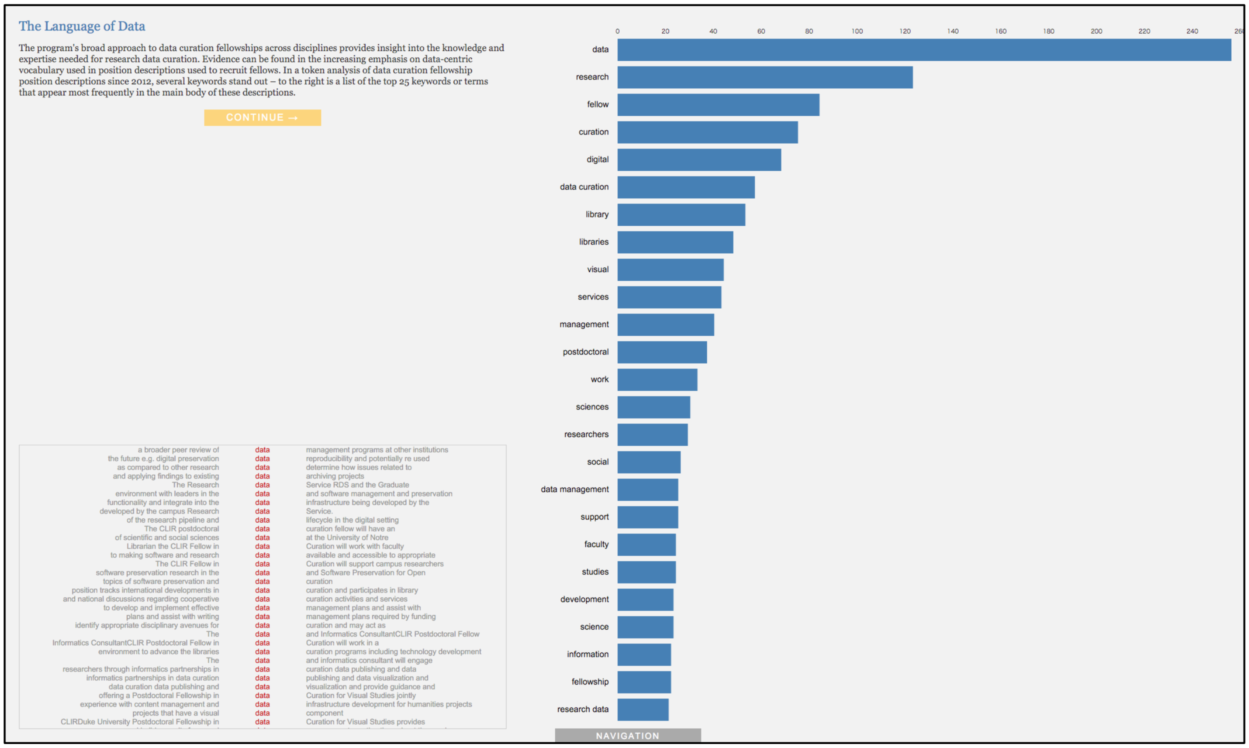
Data curation fellows are newly-minted PhDs with disciplinary expertise across the arts and humanities, social sciences, and natural sciences and have served at 32 different institutions across the United States and Canada.

Map of fellows showing fellowship host institutions



Data curation fellows are scholarly practitioners who understand not only the nature and processes of the disciplines broadly defined but also how research data are organized, transmitted, manipulated, reused, and sustained. Together, they have collaborated on presentations, publications, and other projects.

Map of fellows showing collaborative connections on projects and publications

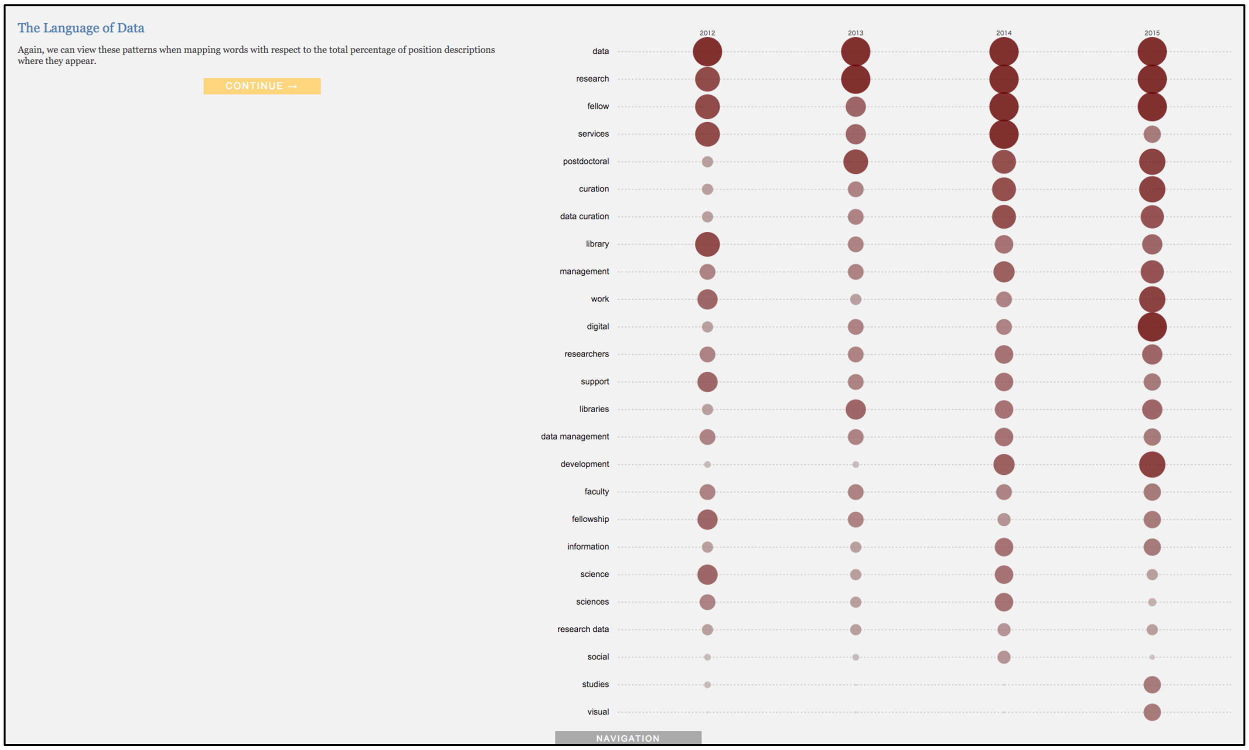


The Language of Data

The program's broad approach to data curation fellowships across disciplines provides insight into the knowledge and expertise needed for research data curation. Evidence can be found in the increasing emphasis on data-centric vocabulary used in position descriptions used to recruit fellows. In a token analysis of data curation fellowship position descriptions since 2012, several keywords stand out – to the right is a list of the top 25 keywords or terms that appear most frequently in the main body of these descriptions.

a broader peer review of	data	management programs at other institutions
the future e.g. digital preservation	data	reproducibility and potentially re used
as compared to other research	data	determine how issues related to
and applying findings to existing	data	archiving projects
The Research	data	Service RDS and the Graduate
environment with leaders in the	data	and software management and preservation
functionality and integrate into the	data	infrastructure being developed by the
developed by the campus Research	data	Service.
of the research pipeline and	data	lifecycle in the digital setting
The CLIR postdoctoral	data	curation fellow will have an
of scientific and social sciences	data	at the University of Notre
Librarian the CLIR Fellow in	data	Curation will work with faculty
to making software and research	data	available and accessible to appropriate
The CLIR Fellow in	data	Curation will support campus researchers
software preservation research in the	data	and Software Preservation for Open
topics of software preservation and	data	curation
position tracks international developments in	data	curation and participates in library
and national discussions regarding cooperative	data	curation activities and services
to develop and implement effective	data	management plans and assist with
plans and assist with writing	data	management plans required by funding
identify appropriate disciplinary avenues for	data	curation and may act as
The	data	and Informatics ConsultantCLIR Postdoctoral Fellow
Informatics ConsultantCLIR Postdoctoral Fellow in	data	Curation will work in a
environment to advance the libraries	data	curation programs including technology development
The	data	and informatics consultant will engage
researchers through informatics partnerships in	data	curation data publishing and data
informatics partnerships in data curation	data	publishing and data visualization and
data curation data publishing and	data	visualization and provide guidance and
offering a Postdoctoral Fellowship in	data	Curation for Visual Studies jointly
experience with content management and	data	infrastructure development for humanities projects
projects that have a visual	data	component
CLIRDuke University Postdoctoral Fellowship in	data	Curation for Visual Studies provides

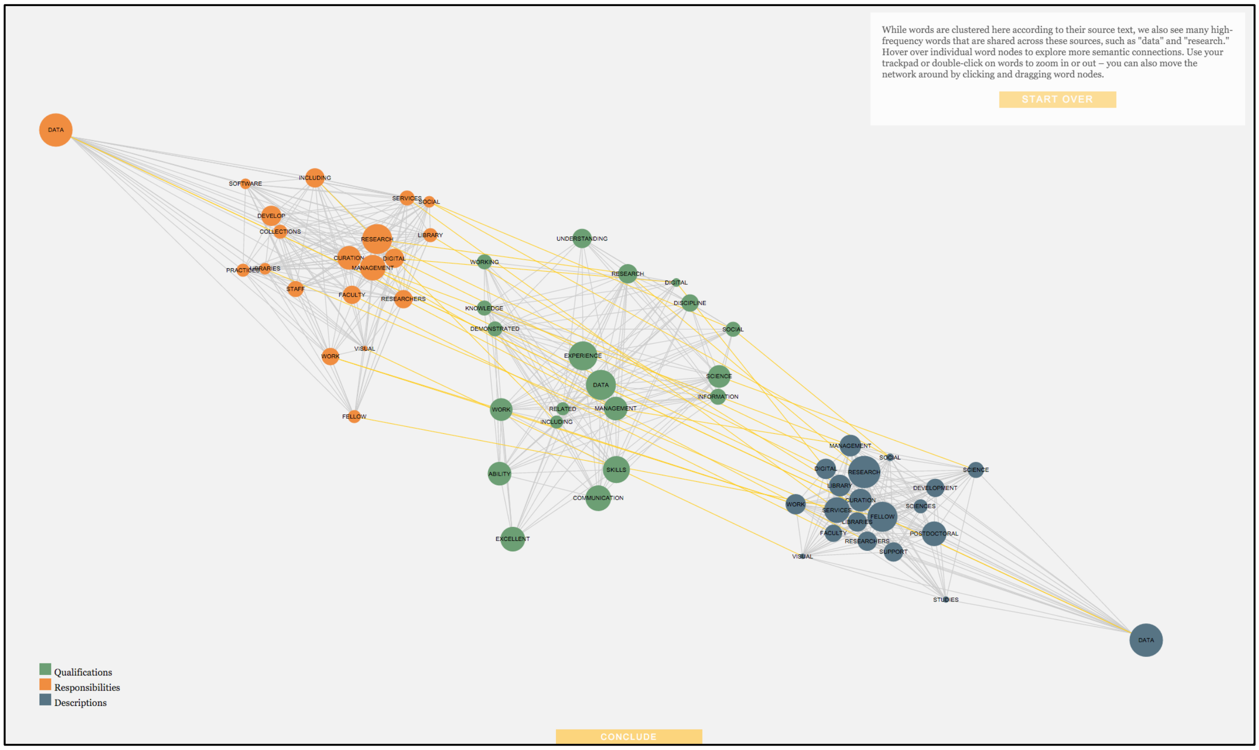
Example list of high-frequency tokens



The Language of Data

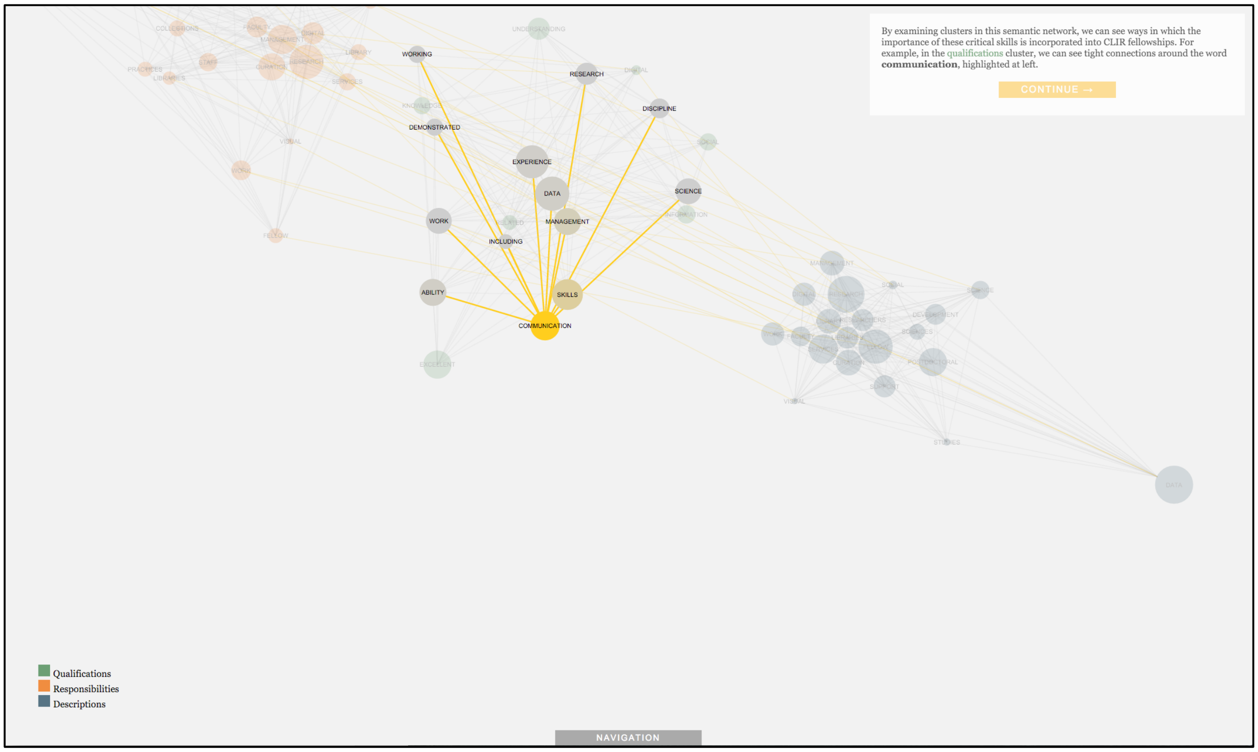
Again, we can view these patterns when mapping words with respect to the total percentage of position descriptions where they appear.

Changes in token frequencies over time



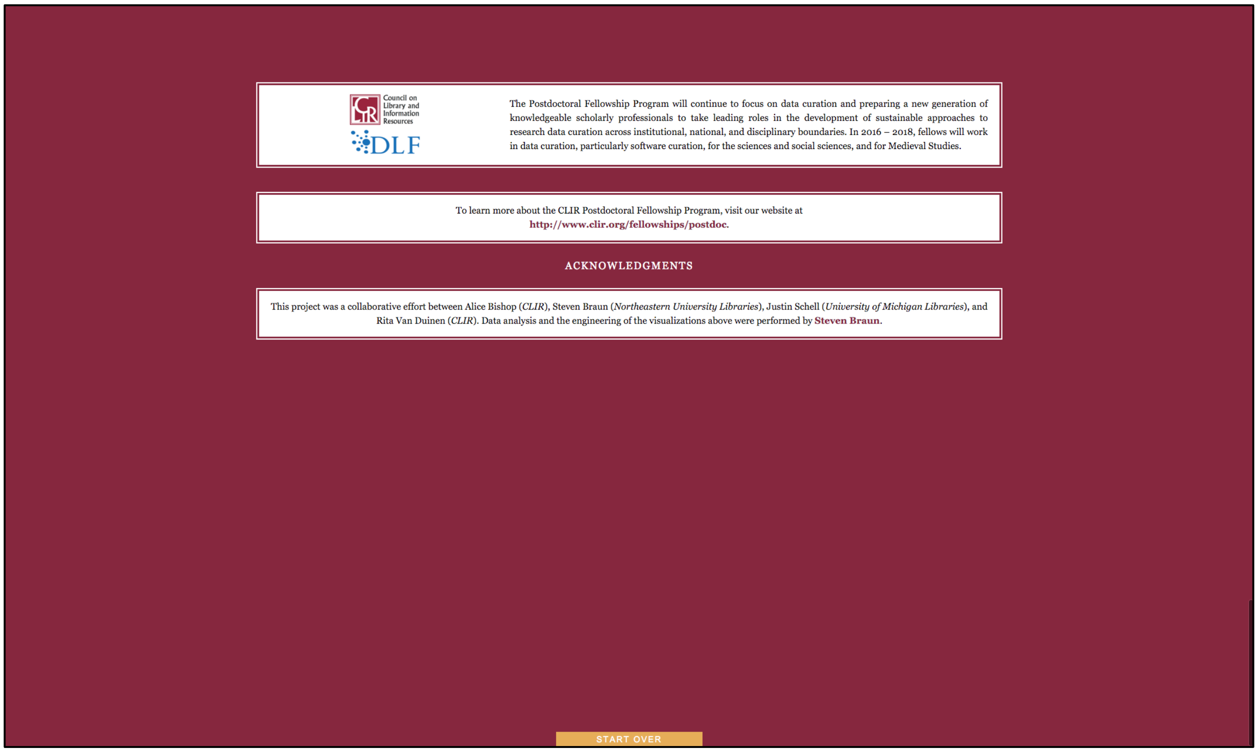
While words are clustered here according to their source text, we also see many high-frequency words that are shared across these sources, such as "data" and "research." Hover over individual word nodes to explore more semantic connections. Use your trackpad or double-click on words to zoom in or out – you can also move the network around by clicking and dragging word nodes.

Network diagram showing high-frequency tokens and connections between them based on co-occurrence patterns



By examining clusters in this semantic network, we can see ways in which the importance of these critical skills is incorporated into CLIR fellowships. For example, in the qualifications cluster, we can see tight connections around the word **communication**, highlighted at left.

Network diagram highlighting an example cluster of tokens



The Postdoctoral Fellowship Program will continue to focus on data curation and preparing a new generation of knowledgeable scholarly professionals to take leading roles in the development of sustainable approaches to research data curation across institutional, national, and disciplinary boundaries. In 2016 – 2018, fellows will work in data curation, particularly software curation, for the sciences and social sciences, and for Medieval Studies.

To learn more about the CLIR Postdoctoral Fellowship Program, visit our website at **http://www.clir.org/fellowships/postdoc**.

ACKNOWLEDGMENTS

This project was a collaborative effort between Alice Bishop (*CLIR*), Steven Braun (*Northeastern University Libraries*), Justin Schell (*University of Michigan Libraries*), and Rita Van Duinen (*CLIR*). Data analysis and the engineering of the visualizations above were performed by **Steven Braun**.

Concluding slide with acknowledgments

Stop Words

The following is a list of token exceptions consisting of common English words that were excluded from final analyses. This list is obtained from http://www.ranks.nl/stopwords/.

a
about
above
after
again
against
all
am
an
and
any
are
aren't
as
at
be
because
been
before
being
below
between
both
but
by
can't
cannot
could
couldn't
did
didn't
do
does
doesn't
doing
don't
down
during
each
few
for
from
further
had
hadn't
has
hasn't
have
haven't
having
he
he'd
he'll
he's
her
here
here's
hers
herself
him
himself
his
how
how's
i
i'd
i'll
i'm
i've
if
in
into
is
isn't
it
it's
its
itself
let's
me
more
most
mustn't
my
myself
no
nor
not
of
off
on
once
only
or
other
ought
our
ours
ourselves
out
over
own
same
shan't
she
she'd
she'll
she's
should
shouldn't
so
some
such
than
that
that's
the
their
theirs
them
themselves
then
there
there's
these
they
they'd
they'll
they're
they've
this
those
through
to
too
under
until
up
very
was
wasn't
we
we'd
we'll
we're
we've
were
weren't
what
what's
when
when's
where
where's
which
while
who
who's
whom
why
why's
with
won't
would
wouldn't
you
you'd
you'll
you're
you've
your
yours
yourself
yourselves

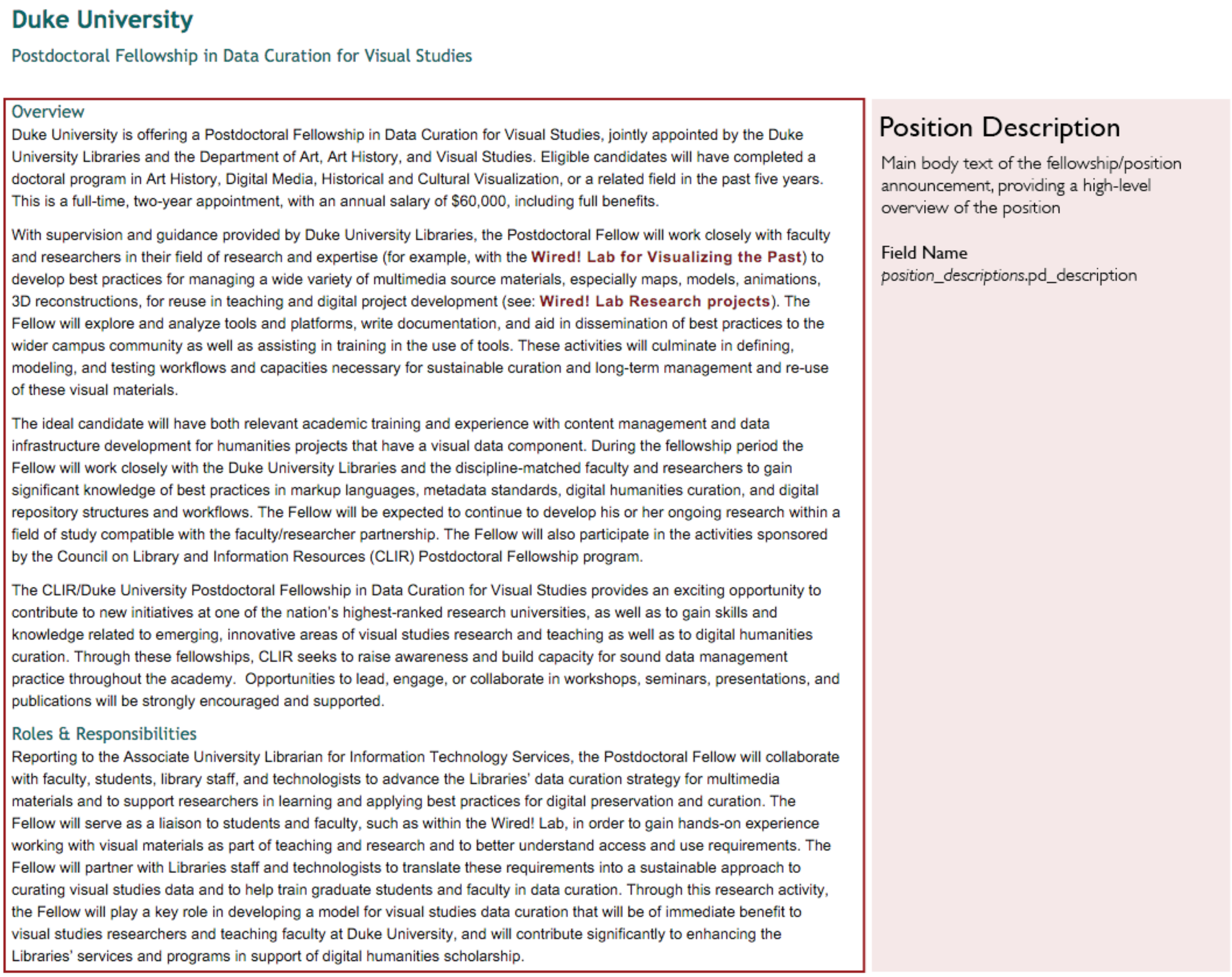
Additional token exceptions were selected based on early analyses that revealed high-frequency terms with low semantic relevance. In addition to those listed above, the following words were excluded from final analyses:

university
phd
will

Sample data, position description text

A sample fellowship position description is included in this package (see /sample-data/duke-2015-position-description-plaintext.txt). The way these postings are broken up into their constituent sections for storage in the database is described in the coded illustration below.

Note: the original position posting can be found online at http://www.clir.org/fellowships/postdoc/applicants/duke2015



Duke University

Postdoctoral Fellowship in Data Curation for Visual Studies

Overview

Duke University is offering a Postdoctoral Fellowship in Data Curation for Visual Studies, jointly appointed by the Duke University Libraries and the Department of Art, Art History, and Visual Studies. Eligible candidates will have completed a doctoral program in Art History, Digital Media, Historical and Cultural Visualization, or a related field in the past five years. This is a full-time, two-year appointment, with an annual salary of $60,000, including full benefits.

With supervision and guidance provided by Duke University Libraries, the Postdoctoral Fellow will work closely with faculty and researchers in their field of research and expertise (for example, with the **Wired! Lab for Visualizing the Past**) to develop best practices for managing a wide variety of multimedia source materials, especially maps, models, animations, 3D reconstructions, for reuse in teaching and digital project development (see: **Wired! Lab Research projects**). The Fellow will explore and analyze tools and platforms, write documentation, and aid in dissemination of best practices to the wider campus community as well as assisting in training in the use of tools. These activities will culminate in defining, modeling, and testing workflows and capacities necessary for sustainable curation and long-term management and re-use of these visual materials.

The ideal candidate will have both relevant academic training and experience with content management and data infrastructure development for humanities projects that have a visual data component. During the fellowship period the Fellow will work closely with the Duke University Libraries and the discipline-matched faculty and researchers to gain significant knowledge of best practices in markup languages, metadata standards, digital humanities curation, and digital repository structures and workflows. The Fellow will be expected to continue to develop his or her ongoing research within a field of study compatible with the faculty/researcher partnership. The Fellow will also participate in the activities sponsored by the Council on Library and Information Resources (CLIR) Postdoctoral Fellowship program.

The CLIR/Duke University Postdoctoral Fellowship in Data Curation for Visual Studies provides an exciting opportunity to contribute to new initiatives at one of the nation's highest-ranked research universities, as well as to gain skills and knowledge related to emerging, innovative areas of visual studies research and teaching as well as to digital humanities curation. Through these fellowships, CLIR seeks to raise awareness and build capacity for sound data management practice throughout the academy. Opportunities to lead, engage, or collaborate in workshops, seminars, presentations, and publications will be strongly encouraged and supported.

Roles & Responsibilities

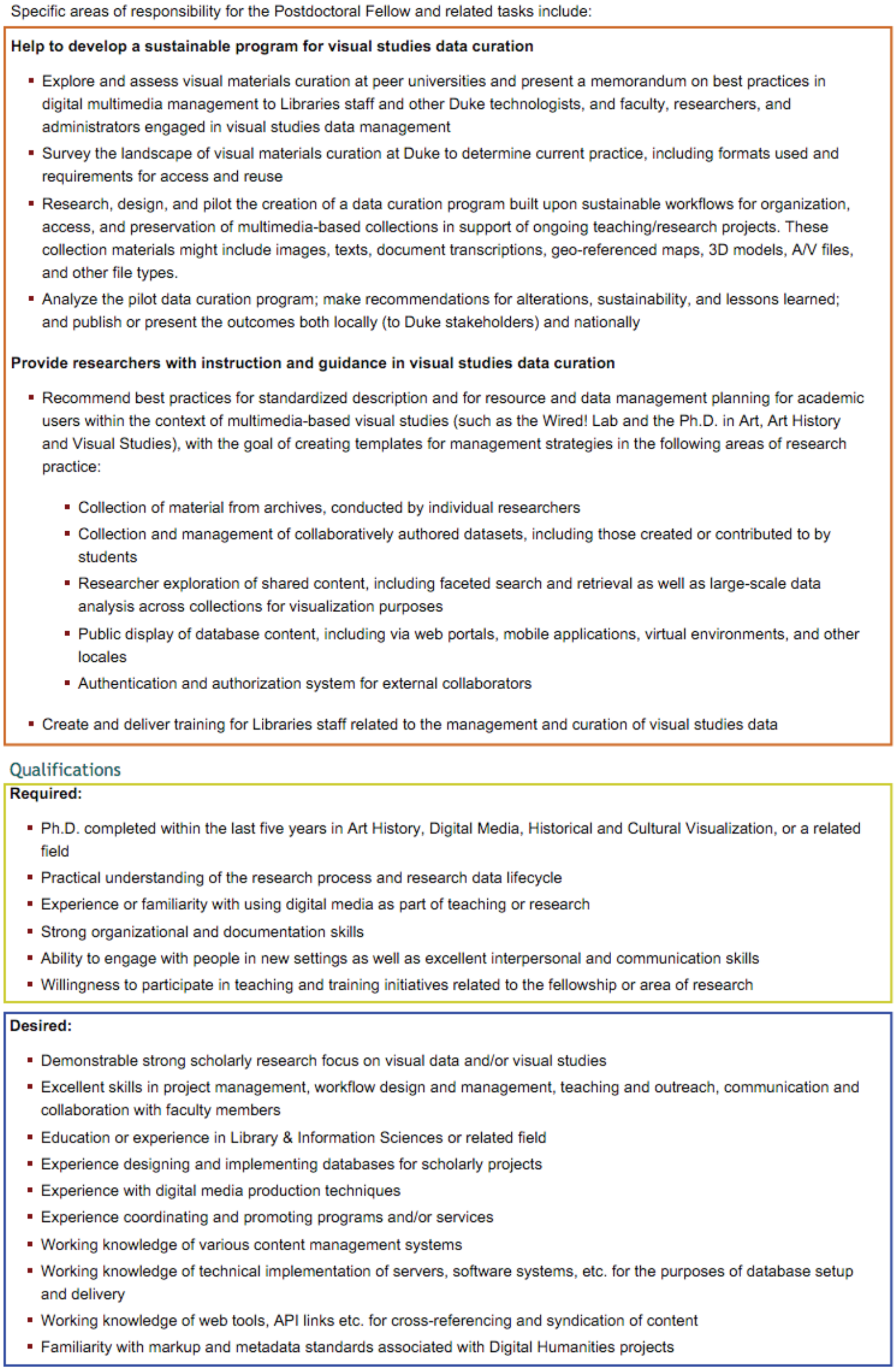
Reporting to the Associate University Librarian for Information Technology Services, the Postdoctoral Fellow will collaborate with faculty, students, library staff, and technologists to advance the Libraries' data curation strategy for multimedia materials and to support researchers in learning and applying best practices for digital preservation and curation. The Fellow will serve as a liaison to students and faculty, such as within the Wired! Lab, in order to gain hands-on experience working with visual materials as part of teaching and research and to better understand access and use requirements. The Fellow will partner with Libraries staff and technologists to translate these requirements into a sustainable approach to curating visual studies data and to help train graduate students and faculty in data curation. Through this research activity, the Fellow will play a key role in developing a model for visual studies data curation that will be of immediate benefit to visual studies researchers and teaching faculty at Duke University, and will contribute significantly to enhancing the Libraries' services and programs in support of digital humanities scholarship.

Position Description

Main body text of the fellowship/position announcement, providing a high-level overview of the position

Field Name

position_descriptions.pd_description



Specific areas of responsibility for the Postdoctoral Fellow and related tasks include:

Help to develop a sustainable program for visual studies data curation

- Explore and assess visual materials curation at peer universities and present a memorandum on best practices in digital multimedia management to Libraries staff and other Duke technologists, and faculty, researchers, and administrators engaged in visual studies data management
- Survey the landscape of visual materials curation at Duke to determine current practice, including formats used and requirements for access and reuse
- Research, design, and pilot the creation of a data curation program built upon sustainable workflows for organization, access, and preservation of multimedia-based collections in support of ongoing teaching/research projects. These collection materials might include images, texts, document transcriptions, geo-referenced maps, 3D models, A/V files, and other file types.
- Analyze the pilot data curation program; make recommendations for alterations, sustainability, and lessons learned; and publish or present the outcomes both locally (to Duke stakeholders) and nationally

Provide researchers with instruction and guidance in visual studies data curation

- Recommend best practices for standardized description and for resource and data management planning for academic users within the context of multimedia-based visual studies (such as the Wired! Lab and the Ph.D. in Art, Art History and Visual Studies), with the goal of creating templates for management strategies in the following areas of research practice:
    - Collection of material from archives, conducted by individual researchers
    - Collection and management of collaboratively authored datasets, including those created or contributed to by students
    - Researcher exploration of shared content, including faceted search and retrieval as well as large-scale data analysis across collections for visualization purposes
    - Public display of database content, including via web portals, mobile applications, virtual environments, and other locales
    - Authentication and authorization system for external collaborators
- Create and deliver training for Libraries staff related to the management and curation of visual studies data

Qualifications

Required:

- Ph.D. completed within the last five years in Art History, Digital Media, Historical and Cultural Visualization, or a related field
- Practical understanding of the research process and research data lifecycle
- Experience or familiarity with using digital media as part of teaching or research
- Strong organizational and documentation skills
- Ability to engage with people in new settings as well as excellent interpersonal and communication skills
- Willingness to participate in teaching and training initiatives related to the fellowship or area of research

Desired:

- Demonstrable strong scholarly research focus on visual data and/or visual studies
- Excellent skills in project management, workflow design and management, teaching and outreach, communication and collaboration with faculty members
- Education or experience in Library & Information Sciences or related field
- Experience designing and implementing databases for scholarly projects
- Experience with digital media production techniques
- Experience coordinating and promoting programs and/or services
- Working knowledge of various content management systems
- Working knowledge of technical implementation of servers, software systems, etc. for the purposes of database setup and delivery
- Working knowledge of web tools, API links etc. for cross-referencing and syndication of content
- Familiarity with markup and metadata standards associated with Digital Humanities projects

Responsibilities

Section outlining the responsibilities or objectives of the fellowship/position

Field Name
position_descriptions.pd_responsibilities

Minimum Qualifications

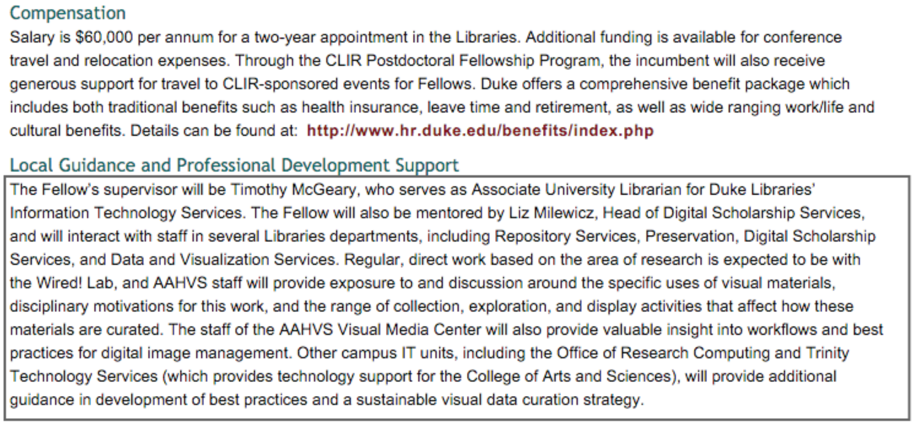
Minimum/required qualifications for the fellowship/position

Field Name
position_descriptions.pd_qualifications_min

Preferred Qualifications

Preferred qualifications for the fellowship/position

Field Name
position_descriptions.pd_qualifications_preferred



Compensation

Salary is $60,000 per annum for a two-year appointment in the Libraries. Additional funding is available for conference travel and relocation expenses. Through the CLIR Postdoctoral Fellowship Program, the incumbent will also receive generous support for travel to CLIR-sponsored events for Fellows. Duke offers a comprehensive benefit package which includes both traditional benefits such as health insurance, leave time and retirement, as well as wide ranging work/life and cultural benefits. Details can be found at: **http://www.hr.duke.edu/benefits/index.php**

Local Guidance and Professional Development Support

The Fellow's supervisor will be Timothy McGeary, who serves as Associate University Librarian for Duke Libraries' Information Technology Services. The Fellow will also be mentored by Liz Milewicz, Head of Digital Scholarship Services, and will interact with staff in several Libraries departments, including Repository Services, Preservation, Digital Scholarship Services, and Data and Visualization Services. Regular, direct work based on the area of research is expected to be with the Wired! Lab, and AAHVS staff will provide exposure to and discussion around the specific uses of visual materials, disciplinary motivations for this work, and the range of collection, exploration, and display activities that affect how these materials are curated. The staff of the AAHVS Visual Media Center will also provide valuable insight into workflows and best practices for digital image management. Other campus IT units, including the Office of Research Computing and Trinity Technology Services (which provides technology support for the College of Arts and Sciences), will provide additional guidance in development of best practices and a sustainable visual data curation strategy.

Additional Information

Any additional or miscellaneous information pertinent to the position but not captured elsewhere

Field Name
position_descriptions.pd_addl_info