Developing Career Word Clouds for the O*NET System

Justin A. Lewis, Jeremiah Morris, and Phil Lewis
National Center for O*NET Development

Prepared for

U.S. Department of Labor
Employment and Training Administration
Office of Workforce Investment
Division of National Programs, Tools, & Technical Assistance
Washington, DC

Submitted by

The National Center for O*NET Development
April 4, 2023

Table of Contents

List of Figures

Introduction

The Occupational Information Network (O*NET) is a comprehensive system developed by the U.S. Department of Labor that provides information on over 900 occupations within the U.S. economy. This information is maintained in a comprehensive database (at the time of this report, the most current version is the O*NET 27.2 Database; National Center for O*NET Development, 2023). To keep the database current, the National Center for O*NET Development (hereafter referred to as "the Center") is involved in a continual data collection process aimed at identifying and maintaining current information on the characteristics of workers and occupations. The Center is also responsible for disseminating data in both original and summarized form via databases, websites, and web services (i.e., APIs) that it develops and maintains. The purpose of this project was to develop a process for creating customized "career word clouds" for each of the 923 data-level occupations included within the O*NET-SOC 2019 taxonomy (Gregory et al., 2019).

The Center has a long-standing history of developing new ways to summarize, simplify, synthesize, and/or display the original, raw data that it collects (e.g., see Easy Read Content; O*NET OnLine Occupation Summary Report; and Developing Related Occupations for the O*NET Program). This project extended this effort by developing customized word clouds for inclusion within the career overview reports of the My Next Move and My Next Move for Veterans websites.[1] Along with adding visual interest to the career reports, the clouds provide customers with a quick overview or snapshot of key attributes for each career.

Background

Data visualization, "…presenting data in graphical or pictorial form which makes the information easy to understand…and helps to explain facts and determine courses of action" (Sadiku et al., pg. 11) has been around for a long time (Aparicio & Costa, 2014). The availability and ability to process large amounts of data, new technology and computer applications, and the emergence of the data science and business intelligence fields all have led to the widespread use of data visualization to help make sense of complex problems such as the COVID-19 pandemic (Lim et al., 2022; Sawicki, 2020). Application-specific tools have been developed for a variety of settings including public health, renewable energy, environmental science, library-decision making, mainstream media, education, academia, and personal usage (Korab, 2021; Sadiku et al., 2016; Viegas et al., 2009).

The goal of this project was to select a visualization technique that would work best in helping to simplify and make engaging career exploration information for students and young adults. A variety of options related to career pathways were discovered (e.g., Li et al., 2018). However, recommended visualizations specifically related to communicating career and occupation

---

[1] Spanish-language career word clouds may be developed in the future for inclusion within the Mi Próximo Paso website.

descriptive information were not found. The search was then expanded to the related fields of education and personality assessment.

The literature points to support and frequent use of "word clouds" within classroom and education settings to simplify and make information more engaging (Burns, 2021; Peachey, 2020; deNoyelles & Reyes-Foster, 2015; Bromley, 2013; Bandeen & Sawin, 2012; Nickell, 2012; Perry, 2012; Baralt, 2011; Viegas, 2009). Examples of word clouds being used to help with the interpretation of personality and knowledge, skills, ability, and vocational interest information were also found (Putka et al., 2022; Park et al., 2014). A word cloud is defined as:

> *an image composed of words used in a particular text, in which the size of each word indicates its frequency or importance* (Korab, 2021 pg. 2).

Word clouds are "eye-catching, engaging, and easy to produce using online tools" (Hearst, 2019 pg. 1). Much of their popularity in various settings is based on their "playful visual appeal" (Hearst et al., 2020 pg. 1). Like many types of data visualization, there is both a technical and "artistic" component to creating effective clouds (Sawicki, 2020). For examples of word clouds, see Figures 1a, 1b, and 1c.

The described characteristics and application within related fields led us to move forward with investigating how to create customized word clouds for each of the careers included within the My Next Move websites.

Figure 1a Word Cloud Example "Wordle"

Cloud example from Viegas et al., 2009 pg. 1

Figure 1b Word Cloud Example "Covid 19 Post Outbreak"

Cloud example from Karab, 2021 pg. 8

Figure 1c Word Cloud Example "Social Interests"

Cloud example from Putka et al., 2022 Supplemental C Materials

Career Word Cloud Development and Implementation

The development and implementation of the career word clouds involved four steps:

1. Input selection (i.e., which words and text to include from each occupation's O*NET data profile);

2. Development of word cloud statistics (i.e., a weighted word listing) for each career using a word cloud generator;

3. Review and finalize the career word clouds;

4. Incorporate the career word clouds within the My Next Move websites.

The sections below describe each of the four steps.

Input Selection

Word clouds are created based on the analysis of a specified set of words or text. For each O*NET-SOC occupation (also referred to as "career"), the O*NET System has a wide variety of information available to potentially include in a cloud analysis. Within the O*NET System, the occupational information collected is guided by the O*NET Content Model, the conceptual foundation of O*NET (Mumford & Peterson, 1999). See Figure 2 for a graphical depiction of the O*NET Content Model. The model provides a framework that identifies the most important types of information about work and integrates them into a theoretically and empirically sound system. The model includes "worker-oriented" and "job-oriented" categories of information, which have both "cross-occupation" and "occupation-specific" information types.

When reviewing the types of input to include within the cloud analysis, the goal was to involve a variety of categories and types of information, while also focusing on information typically used in career exploration. Emphasis was placed on more specific types of information (e.g., reported job titles, task statements, and detailed work activity statements) because they are more text-based and they also would help lead to more differentiated clouds across careers. Lastly, all types of information selected were also included in the career overview reports in the My Next Move websites.

The selected input was drawn from the following types of information:

- Occupation Title
- Occupation Description
- Sample of Reported Job Titles
- Task Statements
- Generalized Work Activities
- Detailed Work Activities
- Abilities
- Interests
- Knowledge
- Skills

For the implementation rules used to generate the career cloud input files, see Appendix A.

Figure 2 O*NET Content Model
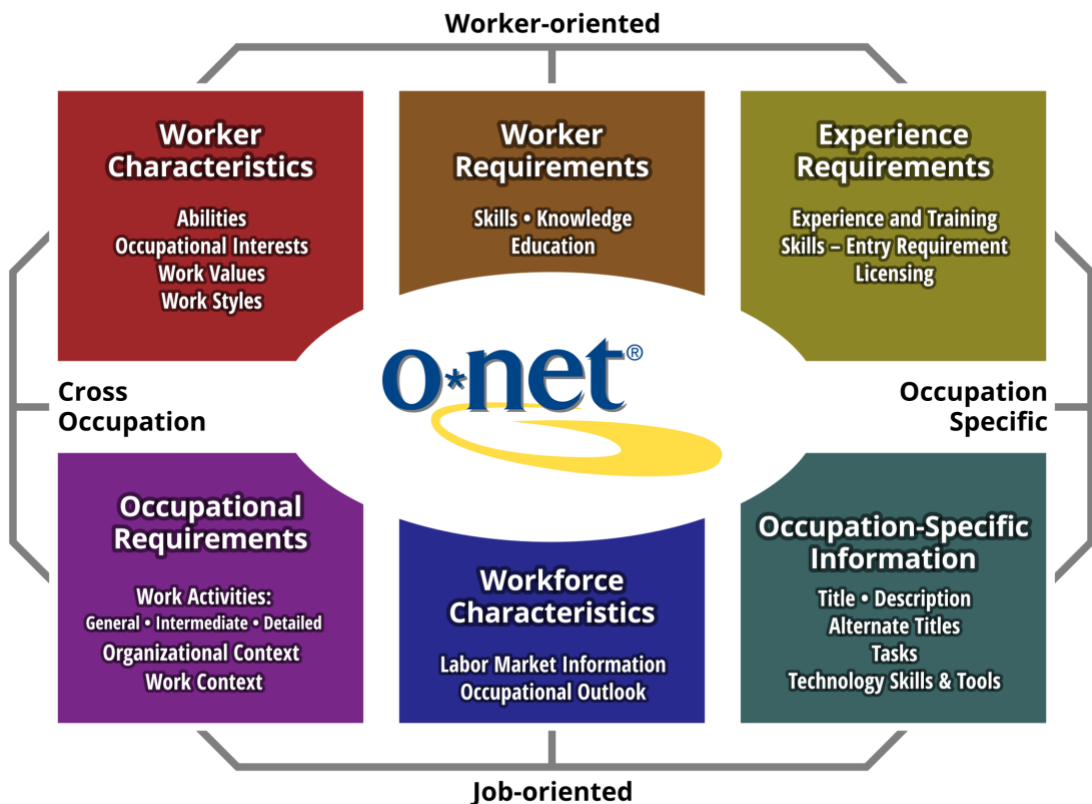
Graphic from National Center for O*NET Development (retrieved February 4, 2023).

Development of Word Cloud Statistics

A "word cloud generator" is a tool that computes statistics for creating clouds, assigning weights to words by breaking down input text into component words and counting how frequently they appear in the body of the input text. Word weights are then used to select which words are displayed, their size, font/boldness, and relative placement to other selected words.

There are a wide variety of available free and paid subscription-based word cloud generators, each having pros and cons depending on a user's need and purpose (Bouchrika, 2022). The applications range from online interfacing tools (cut and paste input text or upload a single file) to batch processors via APIs. Each use their own algorithms for creation, leading to clouds with different looks and feel. Tools offer varying options for output, exporting, and/or direct incorporation into websites and documents (e.g., interactive clouds). They range in the ability of the user to customize or manipulate the mathematics involved in the generation algorithm. They also provide varying control over aesthetic features such font, color palette, or customized cloud shape (e.g., heart-shaped versus traditional cloud shape).

This project reviewed and tested several free word cloud generators. "MonkeyLearn WordCloud Generator" (MonkeyLearn Inc, 2023b) was the selected. Their "AI-Powered" application led to

intuitive, visually pleasing, easy to interpret career clouds. Along with the visual cloud generation, the application also generates lists of the most relevant or frequently used terms, which provides insight into how many times a word was used and helps explain its relevance score. This term-generation capability was separately available through the MonkeyLearn API (MonkeyLearn Inc., 2023a), allowing batch processing and additional control over the text content. The online WordCloud Generator interface allowed for easy manipulation of cloud parameters (e.g., max number of words displayed) and control over a variety of aesthetic features (e.g., font, color palette, image background). Lastly, a wide range of file format exports (e.g., SVG, PNG, and CSV) provided flexibility in processing and implementation options for the project.

The implementation rules for career cloud input (see Appendix A) were applied to extract a datafile from the O*NET 27.1 Database (National Center for O*NET Development, 2022) containing the cloud input data for each of the 923 data-level O*NET-SOC occupations. The MonkeyLearn WordCloud Generator customizable parameters were set (i.e., font, color palette, maximum number of words, and file export type) and word-cloud SVG images were generated for an initial subset of occupations. In addition, through the MonkeyLearn API, 923 weighted term lists were generated covering each data-level occupation.

Review and Finalize Career Word Clouds

While the testing of the MonkeyLearn WordCloud Generator led to positive results across several tests of differing careers, it was determined that the initial clouds generated by the tool would need to be reviewed and potentially manipulated prior to finalization.

A review of applications designed to allow the further customization and manipulation of created word clouds led to the use of the free application "WordClouds.com" (Zygomatic, 2023). The application allows for the import of created and formatted word clouds (i.e., statistics and formatting rules) via word-cloud format files (wcld2). Its user-friendly online interface then enables easy modification and manipulation of the word cloud. After completion, output files can be saved in a variety of formats (e.g., SVG, PNG, PDF). Modified input files can also be easily saved.

Using the term lists generated earlier, along with default image parameters, a batch of 923 word-cloud format files (wcld2) were produced for use as input data. A trained business intelligence analyst used WordClouds.com to review and finalize each career word cloud. The following action steps were implemented:

- Remove non-descript or non-differentiating words (e.g., "confer" or "work");

- Correct words or phrases with poor grammar, unneeded auxiliary words, or misspellings;

- Ensure that at least one "context" specific word is included to help differentiate clouds across careers (e.g., add teaching subject area for the different types of university professor careers by reviewing language included in the occupation reports of O*NET OnLine);

- Ensure the high-point interest area is included (when available);

- Manipulate the final shape of image into as near of a traditional cloud shape as possible.

Once finalized, word cloud images were saved as an SVG file. Also, the modified word-cloud format input files were saved.

Incorporate Career Word Clouds within the My Next Move Websites

With the content and general visual look of the word clouds finalized, they were integrated into a redesign of the My Next Move and My Next Move for Veterans career overview reports. The clouds provide visually-interesting context for each career. See Figures 3a and 3b for examples of the new career overview report headings.

Figure 3a My Next Move Career Overview Report Header-Special Effects Artists & Animators

Figure 3b My Next Move for Veterans Career Overview Report Header-Aerospace Engineering & Operations Technologists & Technicians

Visitors can click on a "zoom in" button to see the career word clouds in a larger pop-up window (see Figures 4a and 4b for examples). The pop-up includes a text list of the image's terms, with the largest (i.e., most frequent in the input material) listed first, so researchers or job seekers can easily copy them into a resume or summary document.

During the integration process, some visual details of the word clouds did not mesh well with existing design elements in the My Next Move sites. Fortunately, the SVG image format allows for customization, as components of the image are stored as a series of drawing commands (for example, "draw the text 'Engineer' in red at the top left corner"). These drawing commands were adjusted to fit the specific needs of the My Next Move sites. Adjustment steps included:

- The font used for word cloud text was changed from "Oswald" to "Open Sans Condensed," which is used for headings throughout the websites. The two fonts have similar metrics (i.e., letter widths and heights are roughly the same) so the substitution did not introduce significant gaps or overlapping areas which might change the look or readability of the clouds.

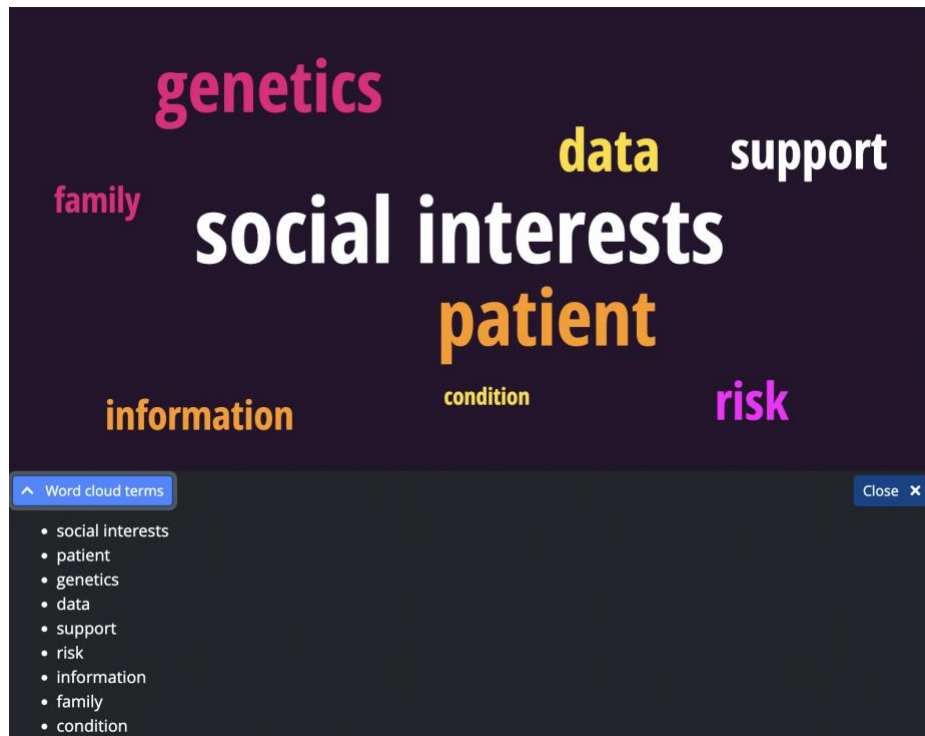**Figure 4a Zoomed-in Word Cloud in My Next Move-Genetic Counselors**

Figure 4b Zoomed-in Word Cloud in My Next Move for Veterans-Data Scientists

- Some images were not visually centered within their canvas area, or had excess empty space due to a tighter grouping of words. Each image was resized and repositioned so that the words were centered, and filled the image area without extending beyond its borders. The Chrome-based tool Puppeteer (Chrome Developers, 2023) was used to measure the word boundaries and calculate the necessary adjustments in a batch process.

- Text colors were lightened and desaturated, and the background color was changed, so that the word clouds did not overpower or clash with adjacent colors in the My Next Move site header or report headings. A separate color palette was developed for use in My Next Move for Veterans, where the site header features a different set of strong colors. In the zoomed-in view, the word cloud's originally-chosen colors are used.

A total of 478 career overview reports were enhanced with word clouds when the new design debuted within My Next Move and My Next Move for Veterans. Career word clouds for the remaining occupations will be published when analyst reviews and processing have been completed.

Conclusion and Future Directions

In its ongoing effort to provide easy-to-understand and visually-appealing summaries of occupational content, the Center developed a process for creating data visualization "career word clouds" encapsulating occupations in the O*NET-SOC 2019 taxonomy. Leveraging a combination of automation and analyst insight, this process was successfully applied to create nearly 500 unique word clouds, which are prominently displayed within two websites in the O*NET site family.

In the future, this process will be used to not only complete career word clouds for all 923 data-level occupations in the O*NET-SOC taxonomy, but to update them as new data is available. Between 2003 and 2022, on average 721 occupations were updated yearly (National Center for O*NET Development, n.d.). When these updates affect the items included as word cloud input data, the image generation and review process can be repeated to pull in new or changed terms. This will allow the career word clouds to reflect the latest available information, such as new task statements suggested by job incumbents, or revised analysis of occupational interests.

In addition to maintaining a comprehensive and current set of career word clouds, another potential direction is to make the word clouds more widely available. O*NET OnLine has a set of occupation reports which may benefit from the addition of these word clouds. O*NET Web Services may provide these career word clouds, and/or the ranked term information depicted in them, to the 3,500+ users of its real-time API.

Engaging and appealing data visualization techniques, including "career word clouds," are an effective way to make the O*NET System's unique occupational information accessible to a wider audience. Word clouds and related visualizations are likely to evolve and grow as important parts of the Center's mission to disseminate and summarize key characteristics of workers and occupations.

References

Aparicio, M., & Costa, C.J. (2014) Data Visualization. *Communication Design Quarterly* (Volume 3, Issue 1,November 2014 pp 7–11), DOI: 10.1145/2721882.2721883

Baralt, M. & Pennestri, S. & Selvandin, M. (2011). Using Wordles to teach foreign language writing. *Language Learning & Technology*, 15(2), 12–22, DOI: 10125/44245

Bandeen, H.M., & Sawin, J.E., (2012) Encourage Students to Read Through the Use of Data Visualization, *College Teaching*, 60:1, 38-39, DOI: 10.1080/87567555.2011.580638

Bouchrika, I. (2022). 7 Best Work Cloud Generator Tools for School and Work. *Research.com* (Software Oct 14, 2022). https://research.com/software/best-word-cloud-generators

Burns, M. (2021). 5 Ways to Use Word Clouds in the Classroom in Grades 3-12. *George Lucas Educational Foundation. Edutopia.* https://www.edutopia.org/article/5-ways-use-word-clouds-classroom/

Bromley, K. (2013). Using word clouds in the classroom. *The Utah Journal of Literacy. 16*(1), 39-41.

Chrome Developers (2023). Puppeteer. https://developer.chrome.com/docs/puppeteer/

Dahlke, J. A., Putka, D. J., Shewach, O., & Lewis, P. (2022). *Developing Related Occupations for the O*NET Program (Report 2022 No. 036).* Alexandria, VA: Human Resources Research Organization. https://www.onetcenter.org/reports/Revised_Deviance.html

deNoyelles, A., & Reyes-Foster, B. (2015). Using Word Clouds in Online Discussions to Support Critical Thinking and Engagement. *Online Learning.* (v19 n4 Sep 2015). https://olj.onlinelearningconsortium.org/index.php/olj/article/view/528

Gregory, C., Lewis, P., Frugoli, P., & Nallin, A. (2019). *Updating the O*NET®-SOC taxonomy: Incorporating the 2018 SOC structure*. Raleigh, NC: National Center for O*NET Development. https://www.onetcenter.org/reports/Taxonomy2019.html

F. B. Viegas, M. Wattenberg and J. Feinberg, "Participatory Visualization with Wordle," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137-1144, Nov.-Dec. 2009, DOI: 10.1109/TVCG.2009.171

Korab, P. (2021). Guide to Using Word Clouds for Applied Research Design. *Towards Data Science*. https://towardsdatascience.com/guide-to-using-word-clouds-for-applied-research-design-2e07a6a1a513

Lim VY, Peralta LMM, Rubel LH, et al. Keeping pace with innovations in data visualizations: A commentary for mathematics education in times of crisis. *ZDM : the International*

Journal on Mathematics Education. 2022 Dec:1-10, DOI: 10.1007/s11858-022-01449-0. PMID: 36532825.

Hearst, M. (2019) Word Clouds: We Can't Make Them Go Away, So Let's Improve Them. *Multiple Views: Visualization Research Explained.* https://medium.com/multiple-views-visualization-research-explained/improving-word-clouds-9d4a04b0722b

M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski and S. Franconeri, "An Evaluation of Semantically Grouped Word Cloud Designs," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 9, pp. 2748-2761, 1 Sept. 2020, DOI: 10.1109/TVCG.2019.2904683

M. Li, *et al*., "CareerVis: Hierarchical Visualization of Career Pathway Data," in *IEEE Computer Graphics and Applications*, vol. 38, no. 6, pp. 96-105, 1 Nov.-Dec. 2018, DOI: 10.1109/MCG.2018.2874514

MonkeyLearn Inc. (2023a). MonkeyLearn API v3. https://monkeylearn.com/monkeylearn-api/

MonkeyLearn Inc. (2023b). MonkeyLearn WordCloud Generator. https://monkeylearn.com/wordcloud/

Mumford, M. D., & Peterson, N. G. (1999). The O*NET content model: Structural considerations in describing jobs. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 21–30). American Psychological Association. DOI: 10.1037/10313-002

National Center for O*NET Development. Easy Read Content. O*NET Resource Center. Retrieved February 3, 2023, from https://www.onetcenter.org/easyread.html

National Center for O*NET Development. O*NET OnLine Summary Report. O*NET OnLine. Retrieved February 3, 2023, from https://www.onetonline.org/help/online/summary

National Center for O*NET Development. O*NET Graphics. O*NET Resource Center. Retrieved February 4, 2023, from https://www.onetcenter.org/graphics.html#content-model

National Center for O*NET Development. (2022). O*NET 27.1 database. https://www.onetcenter.org/dictionary/27.1/excel/

National Center for O*NET Development. (2023). O*NET 27.2 database. https://www.onetcenter.org/dictionary/27.2/excel/

National Center for O*NET Development. (n.d.). O*NET Occupation Update Summary. O*NET Resource Center. Retrieved March 21, 2023, from https://www.onetcenter.org/dataUpdates.html

Nickell, J. (2012). Word Clouds in Math Classrooms. *Mathematics Teaching in the Middle School*, *17*(9), 564–566, DOI: 10.5951/mathteacmiddscho.17.9.0564

Peachey, N. (2020, March 19). Tech tools for teachers: Word clouds. *Macmillan Education*. https://www.macmillanenglish.com/us/blog-resources/article/tech-tools-for-teachers-word-clouds

Perry, L. (2012) Using Word Clouds to Teach About Speaking Style, *Communication Teacher*, 26:4, 220-223, DOI: 10.1080/17404622.2012.702918

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2022). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, 1-26.

Sadiku, M. N. O., Shadare, A. E., Musa, S. M., & Akujuobi, C. M. (2016). DATA VISUALIZATION. *International Journal of Engineering Research and Advanced Technology (ijerat) (E-ISSN 2454-6135) DOI: 10.31695/IJERAT*, *2*(12), 11–16. Retrieved from https://ijerat.com/index.php/ijerat/article/view/191

Sawicki, J. (2020). Why is this Chart Bad? The Ultimate Guide to Data Visualization Evaluation using GoDVE (Grammar of Data Visualization Evaluation). *Towards Data Science.* https://towardsdatascience.com/why-is-this-chart-bad-5f16da298afa

Zygomatic (2023). WordClouds.com. https://www.wordclouds.com/

Appendix A: Implementation Rules for Career Cloud Input

Data Type	Implementation Rule
Occupation Title	Full title.
Occupation Description	Full description.
Sample of Reported Job Titles	Up to 10 reported alternate titles.
Task Statements	Include all "core" task statements. Include all "emerging" task statements.
Generalized Work Activities	Include the element name and description of the highest rated variable on the importance scale.
Detailed Work Activity (DWA) Statements	Include all linked DWA statements.
Abilities	Include the element name and description of the highest rated variable on the importance scale.
Interests	Include the element name listed below for the identified high-point interest code (repeat the element name three times[2]). "Realistic Interests" "Investigative" "Artistic" "Enterprising" "Social Interests" "Conventional Interests"
Knowledge	Include the element name and description of the highest rated variable on the importance scale.
Skills	Include the element name and description of the highest rated variable on the importance scale.

---

[2] Interest element names are included three times to increase the likelihood of the high-point interest being displayed in the career word cloud.