Data and Metadata Profile

The data in question are the salaries of every employee of the University of Vermont from 2009 to 2021.[1] The dataset contains information about year, employee name, base pay in a given year, and job title, and one of the files in the dataset also adds data about department and college. The data are by default arranged alphabetically by the last name of each employee and then chronologically, so each employee's change in salary over time is visible. These data come from the website of the University of Vermont, which publishes annual lists of faculty and staff base pay in PDF form. The information about which department and college within the university employees report to comes from the University of Vermont's department websites and list of full-time undergraduate faculty, also freely available on the university's website.

Stakeholders for this data include its creator, Tyson Pond, a University of Vermont student who compiled the data as part of his Master's of Science in Mathematics program. The University of Vermont itself, as an institution, is also a stakeholder, as it initially published the salary information included in this dataset and is also the producer of the data (because it is the institution paying these salaries to these people). The other main stakeholders are the UVM employees themselves, whose names, positions within the university, and base pay over time make up the dataset. Anyone using the dataset to study university salaries would also be considered a stakeholder, but it is not clear from the dataset's page on Kaggle that anyone is, in fact, using the dataset; there are no comments or discussion on the dataset, and it does not link to or list any publications that have been written based on the dataset. The data have been downloaded from Kaggle 780 times in total, but that does not necessarily mean the dataset's

[1] https://www.kaggle.com/datasets/tysonpo/university-salaries.

downloaders have used it for further analysis or research. At the very minimum, I could not locate any publications that cite this particular dataset.

The dataset comprises three files: a spreadsheet of all faculty and staff salaries without department or college information, a spreadsheet of the salaries of only faculty and staff who are affiliated with an academic department and a college within UVM (i.e. not maintenance, facilities, or dining staff, but only academic and clerical staff), and a data dictionary explaining the various meanings of the abbreviations in the College column of the latter spreadsheet. All three of these are .csv files that do not require any particular program or special information to download and open. While the data dictionary is very helpful and comprehensive, including notes on any unusual cases within the College values, the rest of the metadata is sparse, and other than the column headings in the spreadsheets themselves, there is no other metadata included within the dataset. Kaggle does provide a dedicated metadata section associated with each dataset on the site, and the metadata included in this section is also sparse: only the owner (Tyson Pond), the temporal and geospatial coverage of the dataset (August 8, 2009 to May 30, 2021, in Burlington, VT), the provenance (the UVM website), and the license (CC0) are present. No other authors or collaborators are listed, and the expected update frequency of the dataset is set to "Never." None of the metadata appears to be structured according to a specific existing metadata schema, and since the data dictionary covers only one aspect of the data, it also does not fall in line with unofficial best practices for data dictionaries from organizations such as the Smithsonian[2] and the U.S. Geological Survey[3], which recommend describing all of the data and

[2] Smithsonian Data Management Best Practices - Smithsonian libraries. *Smithsonian Libraries*. (n.d.). https://library.si.edu/sites/default/files/tutorial/pdf/datadictionaries20180226.pdf.

[3] Data dictionaries. Data Dictionaries | *U.S. Geological Survey*. (n.d.). https://www.usgs.gov/data-management/data-dictionaries.

including system-level descriptions or diagrams of the relationships between entities in the dataset. Granted, this is a relatively simple dataset, but the existing inclusion of any data dictionary suggests that including a full dictionary in line with best practices recommendations would have been possible.

In addition to a more comprehensive data dictionary, which would allow for reuse by more inexperienced users, the dataset could be enriched with a variety of additional data. The dataset does not include, for example, information about which positions at UVM are represented by unions and what those collective bargaining agreements specify about compensation over time, nor does it include job description information about the positions whose job titles and salaries are included in the dataset. All of this information should also be available publicly through UVM's website, so no additional data sources would be necessary, and all of it would make the dataset more comprehensive, allow for more points of comparison and study, and thus make it more flexible for reuse or reinterpretation by other users. Salary information from, for example, other state universities in New England could also enrich this dataset. None of this would improve a user's ability to discover the dataset in the repository environment, necessarily, since the dataset is already featured with other education-related datasets on Kaggle and is simply entitled "University Salaries," but including salary information from other universities would certainly make the dataset live up to its title more fully.

Repository Profile

For the University of Vermont faculty and staff salaries dataset, I selected the Dryad data repository.[4] I explored several more niche repositories before settling on Dryad. Other options included Harvard's Henry A. Murray Research Archive and the Research on Gender Relations in Academia repository, which both store datasets with similar themes (e.g. academia and sociology) to the UVM salaries dataset. The Research on Gender Relations in Academia repository in particular curates comparative data about employment in higher education, but its holdings are exclusively German datasets pertaining to German academia, which would have made a Vermont-based dataset a poor fit. Similarly, the Henry A. Murray Research Archive stores only data from Harvard-based researchers. I ultimately opted for Dryad because, as a generalist repository, it is designed for datasets that do not fit into more niche or institution-specific repositories, and because Dryad exclusively publishes data under a CC0 license. The UVM salary dataset is already published under a CC0 license, so there shouldn't be any licensing issues upon ingest. Furthermore, Dryad allows for versioning, but many of the datasets I examined have not been versioned, and the UVM salary dataset is not expected to be updated by its creator.

Dryad is open to submissions from anybody with an account and an associated ORCID, and it places no limits or restrictions on the types of files that can be uploaded as part of a dataset, although their submission walkthrough does recommend submitting data in open file formats whenever possible so that no proprietary software is required to view or use the data. Dryad has no subject- or domain-based restrictions on datasets. Its browser-based online submission form only accepts datasets up to 300 GB, but Dryad data curators can assist with

[4] https://datadryad.org/stash.

dataset submissions up to 1 TB.[5] The main data submission form breaks out its submissions into three categories: data (examples include .csv and .xsl files), software (code packages and scripts used to collect, process, or view the data), and supplemental information (like figures or tables supporting the data). Submitted datasets can contain one or more of these categories of information.

The submission form guides the submitter through the creation of a SIP. The expectation is that a dataset will be linked to its related works (such as journal articles or preprints), and metadata includes the dataset's title, author, research domain, granting or funding organization, an abstract for the dataset, subject keywords for discoverability, an optional description of the data collection methods, and a README file conforming to Dryad's README standards. All README files must contain the dataset's title, a description of the data and file structure, sharing and access information (including other ways to access the data and sources from which the data was derived), and, optionally, a description of the code and software in the dataset, if any. There is no built-in human assistance or consulting through the online submission form, but Dryad does have an email address that submitters can contact for help or with questions about the dataset submission and publication processes.

Because Dryad takes submissions through an online form, submitters do not have to format their metadata themselves; metadata is automatically formatted to conform with Dryad's standards. Dryad makes metadata recommendations (for example, that all files be named according to an internally consistent schema), but it does not mandate the use of a particular schema. Dryad's metadata standards seem to be based on a mix of the Darwin Core and Dublin Core schemas, but because it is a generalist repository, its metadata are not drawn from any

[5] https://zenodo.org/records/7946938.

specific or single-use metadata schema. The authors are listed at the top of each dataset in last name, first name alphabetical order along with their institutions and the publication date and DOI of the dataset, and there are headings for the abstract, methods, and usage notes associated with each dataset.

No login is required to download data, although the process of creating an ORCID and an associated account is relatively straightforward; users need to supply their name, email, and institutional affiliation, if applicable, and confirm their account from an email link. My own account creation took fewer than five minutes in total. Users have the option to download all of the data files in a dataset as a .zip file or to download version files like READMEs and individual project files, separately. Other than the 'download full dataset' button and the individual version file downloads, I couldn't find any other data access mechanisms; if Dryad provides anything like an automated script for downloading datasets or a multi-dataset bulk download option, they do not advertise it well enough for a novice user to find and take advantage of.

There is not, as best as I can tell, any non-user-generated information in Dryad's Dissemination Information Package. Downloaded dataset .zip files include the user-created README file and the files that make up the dataset itself. Unless the metadata mentioned above that is displayed on the webpage for each dataset (from which the dataset can be downloaded) counts as part of the DIP (which it might based on the OAIS definition of a DIP but I am not entirely certain), downloaded datasets do not include metadata about authorship or institution other than what is presented in the README.

Recommended Data Citation

Pond, Tyson (2021) University Salaries. Burlington, VT: Kaggle. https://www.kaggle.com/datasets/tysonpo/university-salaries.

Preservation Considerations

I do not anticipate any major challenges to the long-term preservation of this dataset. It is extremely small, with just three files totaling 525 kB, so there should be no file size or dataset size issues with long-term storage or migration from platform to platform. All files in the dataset are CSV files, which require no special software to open or use, and because CSV files are highly interoperable and do not require maintenance or updates by a specific organization, it is unlikely that the data would ever have to be switched to a different file format for accessibility reasons or that they would become inaccessible in their current format. The biggest issue of preservation is that the dataset's original creator, Tyson Pond, last updated the dataset in 2021 and has no intention of updating or maintaining it again, so if Kaggle (where the dataset was originally uploaded) or this GitHub repository were to experience issues or shut down, the dataset might be rendered inaccessible and have no one invested in preserving it and restoring access.

Copyright

Since the data were pulled from freely-available documents on the University of Vermont website and uploaded to Kaggle under a CC0 license, any fully open copyright license would be appropriate for this dataset. I selected the MIT License for the dataset on GitHub because, like a CC0 license, the MIT License provides access free of charge without restrictions or limitations.

Human Subject Considerations

The data contain publicly available and extremely personally identifiable information about University of Vermont employees, including full names and job titles. The data have not been de-identified and could not, by nature, be anonymized. The dataset's original creator seems to have been working under the assumption that, because the data are publicly available on the UVM website, the people who are the subjects of the data by-default consented to having their names, positions, job titles, and salaries included in this dataset. The data are arguably most valuable in their current form, since if they were de-identified by removing subjects' names, researchers would lose information about employees' gender and ethnicity that could be used to do valuable comparative research into pay disparities at UVM. Regardless, it is clear that because of the publicly-available nature of the data, potential privacy concerns were not taken into account in the creation of this dataset.