The dataset chosen for this analysis was the Phoneme Dataset, whose aim is to acoustically distinguish between vowels produced completely using the oral cavity and vowels having a nasal color. These vowels (named oral and nasal respectively) were sampled at three instances each for the amplitudes of their five component frequencies. This dataset was published on datahub.io and was easily findable with a simple Google search. Once found, the dataset was accessible with robust metadata and a ReadMe file. The dataset also clearly displayed an Open Data Commons Public Domain Dedication and License, which places the data in public domain with all rights waived by the author. The dataset is interoperable, as it was published in multiple file formats including csv, arff, and json. The metadata also includes code snippets explaining how to integrate the dataset into multiple different programming language workflows, including data-cli, cURL, R, Pandas, Python, and JavaScript.

The one FAIR principle which this dataset could improve upon is reusability. While the ReadMe file includes background information explaining the purpose of the dataset and describing the phenomenon it seeks to solve, it doesn’t define the fundamental concepts well enough for someone outside the field to follow coherently. For instance, the average person will not know what harmonics or spectra are, even though these terms are highly common in the field of speech recognition and phonetics. Additionally, there is an error in the metadata with respect to variable definitions. While the metadata describes vowel classes nasal and oral with assigned variable names “Class 0” and “Class 1”, the dataset class column instead uses variables “Class 1” and “Class 2”. This creates a discrepancy in how the vowel classes are defined in the data, and a user must infer which variable refers to which class. After conducting an analysis, it appears that “Class 1” refers to nasal vowels and “Class 2” refers to oral vowels, however this is not guaranteed and must be corrected in either the metadata or the actual dataset for optimal reusability.