Lumping and splitting

Next month, the National Postdoctoral Association will convene its annual meeting in Cleveland, Ohio, where, among many topics, the matter of how data can inform policies to improve post-Ph.D. career pathways will be discussed. Many such data are indeed “out there,” but a major question is whether they are accessible and usable—which brings me to the value of taxonomies.

In 1758, Carl Linnaeus provided a systematic framework for the classification of animals and plants based on detailed visual observations of many organisms. This taxonomy captured important relationships between different groups of organisms and provided a systematic method for naming organisms, including those yet to be discovered. Moreover, the taxonomy was refined over time as more information became available.

The organization of information in this way has turned out to be enormously useful in data analysis across many different enterprises. Taxonomies allow classification of items based on their characteristics so that closely related items are grouped together and given the same name. They are hierarchical so that items that share many but not all characteristics are associated. For characteristics that vary continuously rather than in a discrete fashion, taxonomies can include cut-off points to allow binning of items for subsequent analysis. A key factor in developing taxonomies involves the preference for “lumping” versus “splitting,” terms that evolved early in discussions of biological taxonomies. Lumpers prefer broader categories that include items that share important features despite some differences; splitters prefer narrower categories, emphasizing variations rather than common features.

What happens when there is too much lumping or too much splitting? Consider the classification of cancer cells. Too much lumping, such as defining cancer cells only by their tissue of origin, will obscure biologically and clinically important differences, whereas too much splitting will yield too many categories with insufficient information available about any given one. We are rapidly accumulating data about cancer cells—genomic, proteomic, metabolomic—which may be combined with other data, such as sensitivity to therapeutic agents, to enhance the richness and value of such taxonomies.

During my time as a director at the U.S. National Institutes of Health, I sought to answer a seemingly simple question regarding a pool of individuals who had recently received their first major grant funding: How many postdoctoral fellowships did these early-career scientists complete? Examining individual biographical sketches, I found that the job titles for many scientists changed multiple times during the period between completion of his/her doctoral training and initiation of his/her independent position. A 2014 analysis by the National Postdoctoral Association revealed that 37 different job titles had been used to describe positions that appear, to a reasonable observer, to be postdoctoral fellowships. The excessive splitting here sometimes is related to differences in sources of funding or other factors but, in many cases, is simply due to lack of an agreed taxonomy. This has inhibited many analyses of this important component of many scientific training paths. The development of a proposed taxonomy for career paths for those with Ph.D. degrees in biomedical sciences (see peerj.com/preprints/3370/) represents an important step in facilitating the compilation of career outcomes data in a manner suitable for detailed analysis, particularly when coupled with recent commitments for transparency from leading institutions (see science.sciencemag.org/content/358/6369/1388.full). This taxonomy includes classification by sector, career type, and job function and can help illustrate the many post-Ph.D. career alternatives and the frequencies of their use.

Creating useful taxonomies requires careful analysis about underlying data and the intended uses, as well as engagement with the communities that may be affected by their use. The development of detailed taxonomies for post-Ph.D. career alternatives and commitments to collect and share the underlying data lays the foundation for much important analysis in the near future.

–Jeremy Berg


10.1126/science.aat5956
Lumping and splitting
Jeremy Berg

Science 359 (6382), 1309.
DOI: 10.1126/science.aat5956

http://science.sciencemag.org/content/359/6382/1309

http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service