Astronomy in Focus, Volume 1, Focus Meeting 3
XXIXth IAU General Assembly, August 2015
Piero Benvenuti, ed.

doi:10.1017/S1743921316002702

That over-used and much abused 4-letter word: DATA

Elizabeth Griffin

Dominion Astrophysical Observatory, 5072 West Saanich Road,
Victoria, BC, V9E 2E7, Canada
email: Elizabeth.Griffin@nrc-cnrc.gc.a

Abstract. In its prime state, DATA is a Latin word meaning "[things] given", a plural noun derived from the verb "To Give". Its singular form is DATUM. Modern conversation equates DATA with "Information", while modern philosophies on information management are getting entwined with parallel philosophies on knowledge management. In some ways that is a positive development, and is greatly assisted by Open Access and Internet policies, but in others it is more detrimental, by threatening to blur the essential distinction between objectivity and subjectivity in our science. We examine that essential distinction from the view-points of observers, authors (and publishers), and database managers, and suggest where, when and how the distinctiveness of their fundamental contributions to the communication and validation of research results should be respected and upheld.

Keywords. Astronomical data bases:miscellaneous; catalogues

1. The Facts

The word "Data" is a plural Latin noun meaning "[things] given". It derives from the verb "to give"; its singular form is "datum". However, the way that modern conversation tends to equate "data" with "information" is actually a central contributor to the *apparent* problems surrounding the variety and complexity of 'data in science' or 'data in publishing'.

To start with, "data" is emphatically *not* the same as "information". Science needs a word to describe raw, or just preliminarily processed, observations or records – the unmodified signals which have been captured as an image or a spectrum – and "data" is the traditional word for that. Raw observations are objective: untouched, and distinct from any interpretation of what they show. They are the base reference; regardless of what laws we want to apply, or of what evidence for some theory we hope to extract, those original data constitute the same pristine observation or record for each and every researcher to study.

Admittedly we need to qualify the description "raw", since an observation will unavoidably bear the signature of the detector itself, or (in the case of a spectrum) of the spectrograph – its designed characteristics of limiting resolution and point profile, and its selected ones of wavelength region and focus. It will also have been pre-processed by (for example) a CCD readout, a telemetric-bit conversion, or a photographic development. But the observation stands by itself, and does not rely upon any introduced concept or interpretation like classification type or temperature to explain it, nor is any justification required for what it shows.

2. The Problem

Some of the confusion between "data" and "information" has sneaked in through somewhat sloppy journal editing, or lack of it, and the consequences of what has been happening are still not widely appreciated. Many of today's authors who do not have English as their first language tend to copy the way their English-born peers use their own language, and so one sees a number of new terms, new spellings and even new words gradually creeping into our papers, including this critical lack of appreciation of when, and when not, to use the word "data". Information embraces what is known about some object, process, or whatever; it includes what can be deduced from the basic observation by applying laws, theories and measurements, and represents an end-product of what an analysis of the original observation has produced *on this occasion*. The analysis may be different when carried out by different people, and is therefore fully subjective; the thinking behind each analyses is likely to evolve with time, so the information is neither static nor conclusive. Enough information gleaned from a suitable range of appropriate sources ultimately contributes to *knowledge*, which sounds as though it should be more stable and more comprehensive than the separate strands of information which fed into it. But these stages are a long way removed from the original Data from which they stemmed.

It is therefore essential to respect the differences between Data and Information, and to ensure that that respect prevails throughout our publications, our libraries, our archives and our databases. Unless we do, researchers will be presented with materials that already confuse issues and render it unclear just what has been measured, and what has been deduced and therefore has a temporal quality. One good example of this is the *Bright Star Catalogue*, which includes not only the fundamental positions of the stars and recorded *measurements* of radial velocities, photometry, proper motions and parallaxes, but mixes in the classification type and an opinion as to whether the object has a variable velocity or not, and (from that) whether it is likely to be single or multiple. Many of the velocity observations in the literature of hot stars, in particular, were made manually on photographic spectra, and when the object was rotating so that its (already rather few) lines were broad, it was difficult to measure line-positions (from which the velocity was then derived) with very high precision. As a consequence, the velocities tabulated in the literature show scatter, and rather than suspect that the scatter was caused by low measuring precision the *Catalogue* suggests that all the objects thus affected have variable velocity. It takes a very long time and resources to prove that something once labelled as 'variable' does not in fact vary. A recent study of 12 such systems, for which the *BSC* gave the verdict of variable velocity for 11 of them, finally showed – after numerous observations spanning 5 years – that *all but one* have constant velocity; the one exception proved to be a previously-unrecognized spotted star, whose rapid rotation gave rise to line-profile changes.

3. A Solution?

Despite the scientist's need to restrict the term "data" to a very special aspect of the discipline, the way that the same word is also used very loosely in everyday conversation as a synonym for "facts", "characteristics" or "parameters" is spilling over into science. The development of Open Access and the concomitant involvement of an increasing variety of relevant expertise in information management is now another factor that is adding to the confusion of our descriptive language. That development cannot and should not be checked, but our science is suffering in important ways. How often do young

https://doi.org/10.1017/S1743921316002702 Published online by Cambridge University Press

authors commence a paper with "These stars are known to be ...", or "It is well known that ...", when the truth of the matter is that someone once proposed a hypothesis which then got printed in a paper, and once it was printed rather than just discussed orally it immediately gained an undue credibility: the simple act of publishing the hypothesis conferred upon it a level of proof and acceptability that it did not [yet] deserve. Insisting on a more clear and rigorously maintained distinction between Data and Information will teach the need to honour the fundamental difference between the objective and the subjective. But Open Access has its own momentum, and any attempt to clean up our conversation so as to respect that basic scientific distinction is as futile as trying to stem a breached dyke-wall with one finger. Perhaps science should invent a new word of its own?