Dark Data

One of the most intriguing new concepts of this past century is dark matter. A number of observational experiments, based mainly on astronomical images that appear to show gravitational lensing, are the major evidence for the existence of dark matter, although its makeup is unknown. So, while its existence provides interpretations that help to explain the current experimental facts, there is the possibility that dark matter may be a latter-day luminiferous aether, which would have be discarded when the nature of the physical universe is better understood.

In our day there is another set of experimental facts that are, for the most part, unseen and unseeable. Although they are accessible to some, providing evidence of new scientific discoveries, for everyone else they might just as well be located across the universe. These are the data that reside in our notebooks and hard drives, and perhaps on ½-in. reel-to-reel tapes. You might call them dark data.

Each day a great deal of data is generated in labs across the earth. But the number of persons available to evaluate the data is usually limited to a researcher, his or her assistants, and grad students. Many times the data reveal nothing of great importance or they ratify work that has already been published. So, these data, for all intents and purposes, disappear. Even the interesting data vanish, once the good stuff has been extracted, analyzed, graphed, and written up. If the discarded data were interpreted erroneously or important features contained in the published experiment were overlooked, those too are, in effect, gone.

In ancient times (when I started doing research), the data were in the form of strip chart records, IBM computer printouts on green-and-white 14-in.-wide paper, or punched tapes from an ASR33 Teletype. Today, most data are recorded and then assembled using standard formats in spreadsheets, databases, graphs, or digital images−all of which could be stored and accessed by others. Now that memory is cheap (a half-terabyte hard drive for under $200 these days!), the idea of storing the results of all our work is not far-fetched.

In some instances, this is already taking place. For example, observatories around the world are turning out astronomical amounts of data. For researchers, it's like trying to drink from a fire hose. But, these days, the data are stored and can be accessed on-line by others. For example, the National Virtual Observatory (http://us-vo.org/) in collaboration with the International Virtual Observatory Alliance, allows astronomical researchers to find, retrieve, and analyze astronomical data generated by ground- and space-based telescopes worldwide.

For unique research facilities such as observatories and accelerators, this approach makes sense. But most research is done in modest size labs by individual researchers or small groups. Even here, duplicating most experiments costs a lot of time and money. If institutions provided raw research data with appropriate metadata to describe the experiment, others could, with proper credit, use the data to make additional discoveries. In some cases, the ability to examine such data could prevent other researchers from going down an unfruitful path or cause them to rethink their experiment along more fruitful lines.

A number of institutions are beginning to move beyond cataloging and archiving the work of their faculty to establishing repositories for the raw materials of research. In a way, the library collections of famous authors represent a form of this raw storage. But it is almost always at the end of a career or after the author's death. In contrast, once the researcher has finished his or her analysis, dark data would have to be posted, since most scientific data soon get stale.

Who knows where the publication and sharing of data will go? Some of the efforts will come from societies such as SPIE, others from individual universities or academic consortia. But individual researchers will have to contend with institutional requirements and their own individual inclinations to publish their data. With new trends in social organization on the Web, there may be totally new reasons to distribute and analyze the contents of these repositories, thus eliminating dark data.

Donald C. O'Shea
Editor