

# Turning Numbers into Knowledge

**MASTERING THE ART OF PROBLEM SOLVING**

Third Edition

**JONATHAN G. KOOMEY, PH.D.**



Analytics Press  
PO Box 4933

El Dorado Hills, CA 95762

<http://www.analyticspress.com>

<http://www.numbersintoknowledge.com>

## SHARE AND SHARE ALIKE

One of the biggest problems with data is that they often aren't linked to the critical information needed to make the data useful, and thus are "disembodied". For example, a downloaded spreadsheet will usually not contain a complete citation for the data itself or a description of the methods and sources used to compile it.

Disembodied data are also commonplace on the World Wide Web. Consider a fictional address: 54321 Broadway Avenue, Sometown, NY 12345. If it appeared on a normal web page, search engines or software agents could only rely on the pattern of characters to match against a search query, using "natural language processing". They could not know for sure that 12345 is a zip code, or that 54321 Broadway Avenue is a street address.

In both cases, the data lack *metadata*, described by Wikipedia as "data about data".<sup>102</sup> The American Library Association more precisely defines it as "structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities". Most computer users are familiar with the vCard format for address book data, which allows software to import and export contact information. vCards associate all parts of addresses with correct metadata (e.g., all zip codes are correctly labeled as such).

Lack of metadata prevents software from making more intelligent use of data on the web. To solve this problem, Tim Berners-Lee and other web pioneers developed the "Semantic Web" in the early 2000s. This innovation allows web page creators to associate metadata with important data types, and to make the characteristics of and the relationships between such metadata universally accessible to users of the web. In practice this would allow the web to continue to grow rapidly at the same time enhancing the abilities of software tools to access and use relevant data. Berners-Lee summarizes his big-picture view of the potential as follows:

Human endeavor is caught in an eternal tension between the effectiveness of small groups acting independently and the need to mesh with the wider community. A small group can innovate rapidly and efficiently, but this produces a subculture whose concepts are not understood by others. Coordinating actions across a large group, however, is painfully slow and takes an enormous amount of communication. The world works across the spectrum between these extremes, with a tendency to start small—from the personal idea—and move toward a wider understanding over time.

An essential process is the joining together of subcultures when a wider common language is needed. Often two groups independently develop very similar concepts, and describing the relation between them brings great benefits. Like a Finnish-English dictionary, or a weights-and-measures conversion table, the relations allow communication and collaboration even when the commonality of concept has not (yet) led to a commonality of terms.

The Semantic Web, in naming every concept simply by a Uniform Resource Identifier (URI), lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. This structure will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together.

Inadequate metadata also presents a challenge to good analytical practice that is in some ways more complicated than that for the web itself. Data shared among analysts is even more diverse in its structure and format than the bulk of those posted on the web. Full documentation requires deep conceptual understanding and a level of attention to detail that is rare. And gleaning meaning from data often requires expert contextual knowledge that is time consuming and difficult to summarize for those who are not familiar with the practices and terminology of the relevant expert community.

Even with those challenges, the potential benefits of more efficient, effective, and accurate data sharing could be substantial, and that has led historically to a profusion of web sites devoted to sharing data and helping users graph it. Unfortunately, those sites have a terrible record of capturing these benefits, in part because they underestimated the importance of structured

data. This chapter reviews the promise and pitfalls of recent attempts to enhance data sharing, and describes how they might be improved in future.

In this chapter, I'm focused specifically on what I call *public data*, which are freely available to anyone. These data include those available from government sources, but are not limited to those sources. For example, the income statement for any publicly traded company for 2017 contains public data—anyone who wants to know about profits and losses for that company in that year can request it. Society has determined, for reasons of transparency and accountability, that those data should be calculated in a standardized way and made available for public scrutiny (the accounting fraud associated with the Enron and Worldcom debacles notwithstanding). I am not specifically focused in this chapter on proprietary data internal to companies, although much of the discussion here will be relevant to those data as well.

## THE PROMISE OF DATA SHARING

In creating data for use by others, there are two basic approaches: the “wiki” model (where anyone can contribute information or change documents, with the document becoming the summation of the contributions of many members of the community) and the “expert” model (where a recognized expert on a topic creates a document or compiles data and the community trusts the compilation because of that person's expert knowledge). The struggle between these two philosophies has existed since the dawn of the World Wide Web, and it is unlikely to be resolved anytime soon. Wikipedia, of course, champions the wiki model, while journal reviews written by scientists are the most successful example of the expert model.<sup>103</sup>

For some kinds of information the pure wiki model works well, but for data sharing, the expert model has advantages. Numeracy skills matter, and competent experts are better able to compile data in a way that is meaningful and useful. People who are not experienced in numerical issues can make mistakes that a more numerate person generally will avoid, and those mistakes will propagate quickly in an online data sharing community. Of course, some parts of the wiki model related to online communities, like user comments on data (which represent a type of informal peer review),

will still be important when using the expert model, but the central role of experts in creating and posting the data would not be diminished by using collaborative web tools.

Assume for a moment the existence of data sharing web sites that enforce rigorous documentation of sources and methods, track the change history for the data, attach units to all numbers, and track multifaceted rankings of the reliability of the data, as well as the reputation of its creators and the institutions where they work and publish. If sites of that type existed, what benefits would they yield for users? Three major areas immediately come to mind: quality control, ease of finding and using information, and visual discovery.

### ***Quality control***

Data quality is a multidimensional concept encompassing completeness, accuracy, and contextual assessments of the relevance and credibility of data. Web sites sharing structured data could easily enforce basic measures of data quality. Incomplete sources, inadequate explanatory documentation, missing values, and unlabeled units would negatively affect the confidence users have in its underlying quality.

Improved quality control would make analysis more rapid and accurate and minimize some of the pitfalls described earlier in this book. Cleaning the data, finding complete references, and understanding how the data were derived would be made easier. All analysts would benefit.

### ***Ease of finding and using information***

Web sites that facilitate data sharing would allow searches to be much more efficient. Full references and descriptions of how the data were created would provide quicker and more accurate searches. Key-word categories developed specifically for certain types of data would also improve matters. And once users found information, they would have the confidence and quality rankings to help them assess it, plus the full documentation to help them reference it correctly in their work.

### ***Visual discovery***

Sharing of structured data would allow software to more easily and accurately graph it. For example, having the units attached to each data

series would allow software to identify conflicts and prevent data series with incompatible units from being plotted on the same graph. While having the ability to graph data online is convenient, making good graphs is already possible using existing desktop software—it is *sharing* of structured data that is most important at this stage (a point that is unfortunately often neglected by current data sharing sites).

## DATA SHARING SITES HAVE (MOSTLY) FAILED

For the second edition of this book, I listed the following web sites devoted to data sharing (in alphabetical order): Data 360, Freebase, Google Base, Graphwise, Many Eyes, Numberpedia, and Swivel. All are now defunct except for *Data 360*,<sup>104</sup> but even that web site has limited recent activity.

Data 360 allows easy posting of data as well as graphing of those data. It explicitly focuses on sharing of public data, but also allows companies to post their own data for a fee. It seems to tilt towards the expert model—I couldn't find any way to comment on or change data sets posted by others, although the site hints that people other than the creator can update data.

Data sharing sites attempted to apply social networking and online collaboration tools to enable data sharing. They have historically shared a few common attributes:

- All allow(ed) posting/sharing of data.
- All force(d) some structure on the data, but not much.
- Most allow(ed) users to comment on the data.
- Some allow(ed) users to rank or rate data (although what criteria they are using to rank is usually unclear).
- Some allow(ed) easy download of data to Excel or other easily readable formats.
- Some allow(ed) graphing of data.
- None enforce(d) adequate documentation.

What is clear from recent history is that the business models and justifications for these sharing sites have not been successful. What might work instead?

## A BETTER APPROACH?

The designers of the defunct or dormant data sharing sites described above have applied the social networking and collaborative models that have worked well for YouTube and Wikipedia without giving careful thought to what would work best for data. The scientific community has taken a different approach, in which the technical, legal, and logistical issues of data sharing are front and center.

The Nature Publishing Group has attempted to address problems of data sharing by starting an open-access journal called *Scientific Data* to document scientific data sets,<sup>105</sup> specifying how<sup>106</sup> and where<sup>107</sup> data can be documented and stored. Time will tell if this approach will be successful, but the scientific community's need for data sharing is well established and *Nature* is one of the most prominent scientific journals, so its chances are as good as any.

Creative Commons has addressed ambiguities in current law that impede sharing of databases by creating open access data protocols.<sup>108</sup> Another group, Open Knowledge International,<sup>109</sup> has addressed similar issues as part of their "Open Data Commons" project.<sup>110</sup> These two groups correctly (in my view) understand that core issues of data sharing (including legal issues and technical/practical issues about structured data) need to be resolved before data sharing web sites can reach their full potential. This realization is as true for all data sharing sites, not just for their scientific counterparts.

Enabling data sharing will only be as valuable as the software tools that sharing enables, and if data sharing sites can't deliver measurable customer benefits, they will fail. This realization points to the "chicken and egg" nature of the implementation problem for data sharing. Identifying a killer app to use the data may be the key to fostering the development of a vibrant data sharing community, but developing the killer app depends on the existence of a functional data sharing system. The people who crack this code first will reap vast benefits and finally deliver on the promise of data sharing to the broader community.

## CONCLUSIONS

Ideal data sharing sites would allow the community to rank the credibility and quality of data. They would make it easier to find data and to design software tools to graph those data in a way that would yield insights, accelerating both technological and intellectual innovation. There are clear advantages to computer-aided sharing of structured data, but at this writing (mid 2017), data sharing web sites have fallen far short of reaching this promise.

---

*Whenever I found out anything remarkable, I have thought it my duty to put down my discovery on paper, so that all ingenious people might be informed thereof.*

—**ANTONIE VAN LEEUWENHOEK (1632–1723)**





# Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

## You are free to:

**Share** — copy and redistribute the material in any medium or format

The licensor cannot revoke these freedoms as long as you follow the license terms.

---

## Under the following terms:



**Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



**NonCommercial** — You may not use the material for [commercial purposes](#).



**NoDerivatives** — If you [remix, transform, or build upon](#) the material, you may not distribute the modified material.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

---

## Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.