A Methodology For Measuring Information Accessibility In Science

Puneet Kishor, University of Wisconsin-Madison

Submitted for the 2010 Workshop on The Role of VGI in Advancing Science

It is axiomatic that information is the very basis of science, its raw material, and its final product. The more easily we can get to more information, the more expeditious will be our scientific process. Hence, accessibility is an important quality of information. But, while we talk about information accessibility, what do we really mean when we tag some information as inaccessible — is it not visible at all? Is it difficult to understand? Is it not documented? Does it have to be paid for, and is its price too high? Is it encrypted? Is its reuse constrained by onerous legal clauses and expectations of reciprocal obligations?

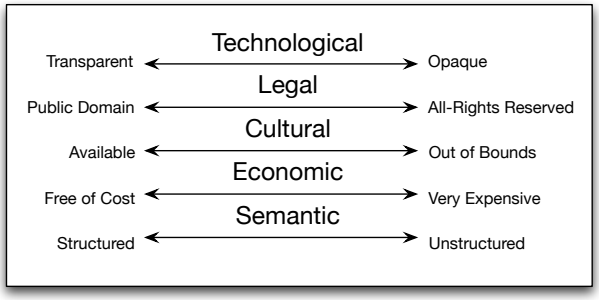
As modes of information production, distribution and use blur, and producers become users and users become producers, questions of openness, accessibility and interoperability become more important. As the internet is allowing us to create and share vast amounts of information very rapidly (Gantz 2008; Lyman and Varian 2003), new tools and techniques are being developed to restrict sharing of information with the aim of protecting the rights of its creators (Mccurry et al. 2006; NRC 2003). As a result, there is a growing tension between the creators, users, managers and beneficiaries of information, further exacerbated when information crosses national and jurisdictional boundaries (Special 301 Report 2007).

Given the above, the following seems inevitable — information *will* be produced and shared with increasing ease, while access to and use of information *will* be restricted with increased vigilance. Whether one is advocating freedom or restriction of information, being able to quantify or otherwise indicate the accessibility of information will be useful but intractably difficult (Peterson Bishop 1998).

Measuring Information Accessibility. Accessibility is not a universally uniform quality. What may be easy to access for one person might be highly restrictive to another. For example, a $100 subscription to a source of information might not be cheap but could be possible for someone in the U.S., but it could be prohibitively expensive for someone from a poor country. Similarly, a source of information that is free of monetary cost, but has non-commercial use restrictions may just as well not exist for someone who may want to found a for-profit business on that information. The concept of accessibility is made up of many different component ideas that are rooted in culture, law, technology, usability and economics. I propose a new methodology for measuring information accessibility in science.

Scientific inquiry both produces and uses various types of information. Digital scientific information can be divided into three major classes: data (further divided into raw and interpreted data), software and source code, and scientific literature ranging from informal output such as magazine articles, presentations, web broadcasts, and books to formally published and peer-reviewed journal articles.

In science, data, software and journal articles all constitute information. Access to information refers to how easily information can be acquired and used, and is represented by `Qinf = {q | qualities of information accessibility}`. Access to scientific information can be restricted technologically, legally, culturally, monetarily and semantically. It should be noted that these are primarily ordinal criteria for measuring access. Other than measuring relative order, it is hard to say that a particular source of information is n-times as restrictive than some other information. These five means for restricting information or freeing it are explored below.



Technology: From "transparent" to "opaque," when applied to data, technological barriers exclude anyone without the proper software or hardware device capable of using that data. When applied to software, such barriers prevent its modification to accommodate variations in the problem domains.

Law: From "public domain" to "all-rights reserved," legal restrictions prevent the uptake of information by attaching restrictive use conditions as either licenses or contracts that impose burdensome requirements to legally using the data.

Culture: From "available" to "out-of-bounds," information access and use can be governed by subtle and not-so-subtle cultural norms. Research in public universities is generally accepted to be an open activity, its products and by-products available for peer reuse, review and verification.

Economics: From making information "free of cost" to "prohibitively expensive," the monetary cost of procuring it has been around for as long as information has been around. Digital information has made it easy to monetize information, but the ability to non-destructively copy information *ad infinitum* has also made it easy to circumvent paying the costs.

Semantics: From "structured" to "unstructured," semantic organization of data affects its accessibility, especially programmatically. Even if the data format is published, making meaning and context of the data difficult to access and process will raise the barrier to its reuse. This becomes particularly important when volume of data grows beyond the ability of humans to parse and make sense of it manually without the assistance of computers.

Indicator of Information Accessibility. At their simplest, the above five indicators are binary — on or off, transparent or opaque, free or for money, open or restricted, etc. If different levels of intensity are employed, even then, the indicators would be ordinal rather than provide and sense of useful cardinality. It would be useful to have a composite indicator that is meaningful across different information sources. To do this, I turn to Qualitative Comparative Analysis (QCA), a technique particularly suited for analysis when the set of entities being studied is small, and the variation in the set qualities is very narrow, or even binary (Ragin 1989). My hope is that a formal methodology for assessing the accessibility of information will decrease uncertainty associated with that information, and assist in information policy as well as information use. This will, in turn, increase the interoperability of information, making the scientific process more efficient and equitable.

REFERENCES

Gantz, John F. 2008. The Diverse and Exploding Digital Universe.

Lyman, Peter, and Hal R. Varian. 2003. How Much Information? Berkeley, CA.

Mccurry, D, James Campbell, M Lutz, Harlan Onsrud, and K Williams. 2006. Managing intellectual property issues in a commons of geographic data. *Digital Libraries* Proceeding:347.

NRC. 2003. *The Role Of Scientific And Technical Data And Information In The Public Domain*: The National Academies Press.

Peterson Bishop, Ann. 1998. Logins and Bailouts: Measuring Access, Use, and Success in Digital Libraries. *Journal of Electronic Publishing* 4:27-38.

Ragin, Charles. 1989. *Comparative method : moving beyond qualitative and quantitative strategies*. 1st ed. Berkeley, CA: University of California Press.

Special 301 Report. 2007. Office of the United States Trade Representative.