Creating a Science of the Web

Tim Berners-Lee[1], Wendy Hall[2], James Hendler[3], Nigel Shadbolt[2], Daniel J. Weitzner[1]

[1]Computer Science and Artificial Intelligence Laboratory, MIT, USA.
[2]School of Electronics and Computer Science, University of Southampton, UK.
[3]Computer Science Department, University of Maryland, USA.

Since its inception, the World Wide Web has changed the ways scientists communicate, collaborate, and educate. There is, however, a growing realization among many researchers that a clear research agenda aimed at understanding the current, evolving, and potential Web is needed. If we want to model the Web; if we want to understand the architectural principles that have provided for its growth; and if we want to be sure that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, then we must chart out a research agenda that targets the Web as a primary focus of attention.

When we discuss an agenda for a science of the Web, we use the term "science" in two ways. Physical and biological science analyzes the natural world, and tries to find microscopic laws that, extrapolated to the macroscopic realm, would generate the behavior observed. Computer science, by contrast, though partly analytic, is principally synthetic: It is concerned with the construction of new languages and algorithms in order to produce novel desired computer behaviors. Web science is a combination of these two features. The Web is an engineered space created through formally specified languages and protocols. However, because humans are the creators of Web pages and links between them, their interactions form emergent patterns in the Web at a macroscopic scale. These human interactions are, in turn, governed by social conventions and laws. Web science, therefore, must be inherently interdisciplinary; its goal is to both understand the growth of the Web and to create approaches that allow new powerful and more beneficial patterns to occur.

Unfortunately, such a research area does not yet exist in a coherent form. Within computer science, Web-related research has largely focused on information-retrieval algorithms and on algorithms for the routing of information through the underlying Internet. Outside of computing, researchers grow ever more dependent on the Web; but they have no coherent agenda for exploring the emerging trends on the Web, nor are they fully engaged with the emerging Web research community to more specifically focus on providing for scientists' needs.

Leading Web researchers discussed the scientific and engineering problems that form the core of Web science at a workshop of the British Computer Society in London in September 2005 (1). The participants considered emerging trends on the Web and debated the specific types of research

needed to exploit the opportunities as new media types, data sources, and knowledge bases become "Webized," as Web access becomes increasingly mobile and ubiquitous, and as the need increases for privacy guarantees and control of information on the Web.

The workshop covered a wide range of technical and legal topics. For example, there has been research done on the structure and topology of the Web (2, 3) and the laws of connectivity and scaling to which it appears to conform (4–6). This work leads some to argue that the development of the Web has followed an evolutionary path, suggesting a view of the Web in ecological terms. These analyses also showed the Web to have scale-free and small-world networking structures, areas that have largely been studied by physicists and mathematicians using the tools of complex dynamical systems analysis.

The Web yesterday and today. (**Left**) The World Wide Web circa 1990 consisted primarily of text content expressed in the Hypertext Markup Language (HTML), exchanged via the hypertext transfer protocol (HTTP), and viewed with a simple browser pointing to a Universal Resource Locator (URL). (**Right**) Users of the Web now have a variety of top-level tools to access richer content including scalable vector graphics, the Semantic Web, multimodal devices (e.g., voice browsers), and service descriptions. These are expressed in extended markup language (XML), exchanged by newer protocols [e.g., HTTP 1.1 and SOAP (simple object access protocol)] and are addressed by uniform resource identifier (URI) schemes.
CREDIT: ADAPTED FROM THE WORLD WIDE WEB CONSORTIUM

The need for better mathematical modeling of the Web is clear. Take the simple problem of finding an authoritative page on a given topic. Conventional information-retrieval techniques are insufficient at the scale

of the Web. However, it turns out that human topics of conversation on the Web can be analyzed by looking at a matrix of links (7, 8). The mathematics of information retrieval and structure-based search will certainly continue to be a fertile area of research as the Web itself grows. However, approaches to developing a mathematical framework for modeling the Web vary widely, and any substantive impact will, again, require a new approach. The process-oriented methodologies of the formal systems community, the symbolic modeling methodologies of the artificial intelligence and semantics researchers, and the mathematical methods used in network analyses are all relevant, but no current mathematical model can unify all of these.

One particular ongoing extension of the Web is in the direction of moving from text documents to data resources (see the figure). In the Web of human-readable documents, natural-language processing techniques can extract some meaning from the human-readable text of the pages. These approaches are based on "latent" semantics, that is, on the computer using heuristic techniques to recapitulate the intended meanings used in human communication. By contrast, in the "Semantic Web" of relational data and logical assertions, computer logic is in its element, and can do much more.

Researchers are exploring the use of new, logically based languages for question answering, hypothesis checking, and data modeling. Imagine being able to query the Web for a chemical in a specific cell biology pathway that has a certain regulatory status as a drug and is available at a certain price. The engineering challenge is to allow independently developed data systems to be connected together without requiring global agreement as to terms and concepts. The statistical methods that serve for the scaling of language resources in search tasks and the data calculi that are used in scaling database queries are largely based on incompatible assumptions, and unifying these will be a major challenge.

Despite excitement about the Semantic Web, most of the world's data are locked in large data stores and are not published as an open Web of inter-referring resources. As a result, the reuse of information has been limited. Substantial research challenges arise in changing this situation: how to effectively query an unbounded Web of linked information repositories, how to align and map between different data models, and how to visualize and navigate the huge connected graph of information that results. In addition, a policy question arises as to how to control the access to data resources being shared on the Web. This latter question has implications both with respect to underlying technologies that could provide greater protections, and to the issues of ownership in, for example, scientific data-sharing and grid computing.

The scale, topology, and power of decentralized information systems such as the Web also pose a unique set of social and public-policy challenges. Although computer and information science have generally concentrated on the representation and analysis of information, attention also needs to be given to the social and legal relationships behind this information (9). Transparency and control over these complex social and legal relationships are vital, but require a much better-developed set of models and tools that can represent these relationships. Early efforts at modeling in the area of privacy and intellectual property have begun to establish the scientific and legal challenges associated with representing and providing users with control over their own information. Our aim is to be able to design "policy aware" systems that provide reasoning over these policies, enable agents to act on a user's behalf, make compliance easier, and provide accountability where rules are broken.

Web science is about more than modeling the current Web. It is about engineering new infrastructure protocols and understanding the society that uses them, and it is about the creation of beneficial new systems. It has its own ethos: decentralization to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and fairness. It uses powerful scientific and mathematical techniques from many disciplines to consider at once microscopic Web properties, macroscopic Web phenomena, and the relationships between them. Web science is about making powerful new tools for humanity, and doing it with our eyes open.

References

1. Workshop on The Emerging Science of the Web, British Computer Society, London, 12 to 13 September 2005. See **www.cs.umd.edu/~hendler/2005/WebScienceWorkshop.html**
2. R. Milo et al., Science **298**, **[824]** (2002).
3. R. Milo et al., Science **303**, **[1538]** (2004).
4. A.-L. Barabási, R. Albert, Science **286**, **[509]** (1999).
5. J. M. Kleinberg, Nature **406**, 845 (2000). **[CrossRef]**
6. S. H. Strogatz, Nature **410**, 268 (2001). **[CrossRef]** **[Full text]**
7. S. Brin, L. Page, in Proceedings of the 7th International World Wide Web Conference (Elsevier Science, Amsterdam, 1998), pp. 107–117. **[Full text]**
8. Z. N. Oltvai, A.-L. Barabási, Science **298**, **[763]** (2002).
9. L. Lessig, Code and Other Laws of Cyberspace (Basic Books, New York, 1999). **[publisher's information]**