RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

197

Toward an Epistemic Web

Malcolm D. Hyman and Jürgen Renn

April 2012

Federal Ministry
of Education
and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Toward an Epistemic Web[1]

Malcolm D. Hyman[†] and Jürgen Renn

Introduction

In the beginning knowledge was local. With the development of more complex forms of economic organization knowledge began to travel. The Library of Alexandria was the fulfillment—however partial and transitory—of a vision to bring together all the knowledge of the world. But to obtain the knowledge one had to go to Alexandria. Today the World Wide Web promises to make universally accessible the knowledge of a world grown larger. To be sure, much work remains to be done: many documents need to be made available (i.e. digitized if they are not already, and freed from restrictive access controls); and various biases (economic, legal, linguistic, social, technological) need to be overcome. But what do we do with this knowledge? Is it enough to create a digital library of Alexandria, with (perhaps) improved finding aids? We propose that the crucial question is how to structure knowledge on the Web to facilitate the construction of new knowledge, knowledge that will be critical in addressing the challenges of the emerging global society. We begin by asking three questions about the Web and its future. In the remainder of the paper we explore the possibility of an "Epistemic Web" in the context of a more general discussion of "knowledge representation technologies," technologies used for storing, manipulating and spreading knowledge.

---

[1] This paper will be published in the forthcoming volume *The Globalization of Knowledge in History*, Berlin: Edition Open Access, http://www.edition-open-access.de.

What is Fundamentally New About the Web as a Knowledge Representation Technology?

The World Wide Web is a recent phenomenon, but it belongs in a long chain of knowledge representation technologies. In fewer than twenty years the Web has developed from a small tool used by a specialized research community to a technology with more than a billion users, and a volume of data added each year that exceeds the content held, for example, in the Library of Congress by a factor of hundreds of millions. Apart from its rapid growth, what makes the Web different from other knowledge representation technologies?

1. The Web offers a high *impact potential* to an unprecedented number of people. Personal weblogs can receive hundreds of thousands of visitors daily.
2. The Web offers high *collaborative scalability*. Thousands of people (or more) can collaborate in the creation of such products as an open-source operating system (GNU/Linux) or an encyclopedia (Wikipedia).
3. The Web promises nearly universal *interconnectivity*. Discrete documents participate in a vast network of relations to other documents.
4. The Web exhibits exceptional *plasticity*. It can readily accommodate new ways of organizing content as well as new types of content. Content can be changed rapidly and frequently.
5. The Web allows *ambient findability*. Amidst the vast stockpiles of information, desired knowledge can be located almost instantaneously from anywhere in the world (Morville 2005, 6).
6. The Web provides extremely *low latency*. News spreads worldwide within minutes after an event; photographs and telemetry within seconds. Data with radically disparate lifetimes converge: today's news story already finds its place in the encyclopedia.

What Are the Shortcomings of the Present-Day Web?

None of the Web's distinctive potentials have yet been systematically realized. The present Web remains a prototype of what the Web might become, and of what its founders envisioned (Gillies and

Cailliau 2000). The democratizing impact potential is hindered by a "digital divide"—inequality in access to digital sources and services—that results not only from economic disparity but also from technocratic culture, linguistic bias (Paolillo 2005) and the absence of key enabling technologies. Collaborative scalability is limited by the lack of tools for shared annotation of heterogeneous data. Universal interconnectivity cannot be achieved without tools for visualizing and manipulating the complex structures of relations between documents. Plasticity is impeded by the lack of standards for linking non-textual media at a fine granularity. Findability fails without some formal means of disambiguating natural language. And despite the potential of low latency, the time-to-publication of scholarship is scarcely lower on the Web than in traditional print culture, since social practices have not evolved at the same rate as technology.

More generally, however, there is a "central" problem, namely, how to represent human knowledge adequately on the Web. Any solutions that fail to address this problem must fail radically. It is not enough to look to semantics, or social networks, or increased interactivity, or more sophisticated computation—although all these things are indeed useful and necessary.

What Are the Options for Future Developments of the Web?

Proposals for how to transform the present-day Web abound. The explosion of technology opens up a maze of possible directions for creating a new civic and scholarly infrastructure, an *embarras de richesse*. Three paths have most notably captured the attention of technological visionaries:

1. The idea of the *Semantic Web* was first publicly aired by Berners-Lee and colleagues in 2001; they proposed "an extension of the current [Web], in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[2] In the Semantic Web, documents are enriched with structured metadata to allow for intelligent information retrieval and automated inferences about document content. Ontologies capture the relations between terms within a specific knowledge domain. Semantic Web research has led to the development of potentially fruitful technologies such as RDF

[2] See (Berners-Lee et al. 2001; Halpin 2004).

(Resource Description Framework) and OWL (Web Ontology Language) (Yu 2007). Yet few compelling applications have emerged so far. Moreover, it is not clear in which context relationships are established or what happens when fundamental disagreement occurs (as it inevitably will). A centralized approach cannot be the solution! To state matters provocatively, the approach to meaning in the Semantic Web resembles the claims of universal validity once offered by the Catholic Church and the Soviet Union. Although the Semantic Web can increase the ability of *computers* to assist in managing the complexity of the Web, it does not solve the problem of how *humans* can integrate the Web into a coherent body of knowledge.

2. *Web 2.0* is a term first used in 2004 not to describe a vision of what the Web *might* become but rather to name a set of *actual* developments that seemed to point to the future (O'Reilly 2005). This is the *Social Web*. Instead of the formal ontologies of the Semantic Web, Web 2.0 evangelists embraced *folksonomies*—a neologism for informal, bottom-up, overlapping classifications created in an egalitarian fashion by users (Morville 2005, 136). Web 2.0 sites allow anyone to add *tags*—short, simple metadata labels—to resources such as photographs and blog entries; other people can then use these tags in searching for resources. Social Web sites such Technorati, flickr and del.icio.us have become contagiously popular. By allowing for the sharing of sets of tags, these websites connect not just *documents* but also *people*. Web 2.0, in which meaning is assigned not by central authorities but by ordinary citizens, is the Protestant version of the Semantic Web. Yet while this reformation has undoubtedly created a new type of networked community, and although serious scientific applications have emerged (Schröder 2007), it is not clear that such communities can develop into serious scholarly or civic communities organized around a meaningful body of shared knowledge.

3. Futurists envision a *Web of Things* in which *physical objects* become manipulable in many of the same ways that we now manipulate hyperlinked documents. This Web of Things will be enabled by such technologies as low-cost RFID chips, GPS and (in general) the decreasing cost and size of electronic

> components. Bruce Sterling conceives of web-enabled things as spimes, objects whose changes in space and time are recorded, which can be searched, and around which user communities will form (Morville 2005, 84). Others imagine ubiquitous computing in which computers are embedded in, or can communicate with, everyday objects. These scenarios are derided by critics that imagine a series of (often laughable) interactive appliances—and feared by those that imagine a surveillance society of unprecedented reach. The Web of Things offers the potential of expanding the concept of *document* to include all kinds of physical things that indeed constitute *objects* of human knowledge (Morville 2005, 148). But it too ignores the central problem of how systematically to represent human knowledge itself.

All these paths lead somewhere interesting and we by no means view them as misguided. But we insist that a new way is needed: an Epistemic Web, that is, a universe of knowledge on the Web that parallels human knowledge.

We need a deeper understanding of the relationship between knowledge and representation and how that relation has evolved over human history. Such an understanding will allow us to formulate the challenges for the future and to make a proposal for the development of a new Web that is a plausible continuation of the previous evolution of knowledge representation technologies.

The remainder of this paper consists of two parts, each of which begins with a theoretical discussion and concludes with a practical analysis. In the first part we articulate the approach to knowledge taken by historical epistemology and provide a brief history of knowledge representation technologies. In the second part we use three fundamental premises about knowledge to explore the challenges for the future development of the Web and conclude with concrete proposals for the Epistemic Web.

Knowledge: The Perspective of Historical Epistemology

Historical epistemology, as explained in the introduction, is the study of the historical development and transmission of knowledge in light of social, cultural and cognitive factors and with attention to the interaction between individual thinking and institutionalized systems of knowledge.

And as is also explained in the introduction, knowledge is not representation-independent, and the media of knowledge representation affect the structure of knowledge.

Once knowledge is represented externally, it is subject to transfer in a knowledge economy. Particular knowledge representation technologies shape this economy in different ways, since these technologies vary along a set of economic dimensions:

1. *Portability*: Can a representation travel? How fast? Radio and television broadcasts propagate very quickly, whereas inscribed monoliths generally don't move at all.
2. *Durability*: How lasting is a representation? Cuneiform tablets have endured for thousands of years; spoken language has vanished without a trace.
3. *Ownership*: Who has access to the means of production? How easily can this access be controlled? It is considerably easier to regulate printing presses than pen and ink.
4. *Rivalness*: Does an individual's use of a representation decrease the value of that representation for others? Only one person can read a manuscript at a time, but many people can listen to a story teller or watch a television program.
5. *Reproducibility*: At what cost can a representation be copied? Books were more expensive before the invention of printing with movable type; now they can be photocopied inexpensively, and the cost of a digital copy approaches zero.
6. *Interactivity*: How flexibly can a representation be accessed? A monologue can only be listened to from beginning to end; parts of a book can be skipped or re-read; an electronic text can be searched in more powerful ways.
7. *Recursiveness*: Can higher-order knowledge about a representation be externalized and integrated with the representation? Books can be annotated in the margins, but electronic texts can be annotated more extensively and easily; a spoken monologue, on the other hand, can't be annotated at all.
8. *Connectivity*: To what degree, and how explicitly, is a representation connected to other knowledge? An epic poem may contain allusions to other literature, but these are less direct connections than the footnotes in a scholarly article or (a fortiori) hyperlinks in a Web document.

People strive to maintain an equilibrium between their own cognitive

structures and the environment (Piaget 1985). Knowledge from the environment must be assimilated in the context of what an individual already knows, and internal knowledge representations must be accommodated to knowledge acquired from the environment (for instance from external representations). This process is called *equilibration*. The high degree of interaction between internal and external knowledge representations entails that knowledge representation technologies play a key role in equilibration. Equilibration occurs not only with respect to individual knowledge, but also with respect to shared knowledge. Thus equilibration results from an encounter between local and global knowledge (e.g. prior notions of healing and the body are adjusted when global biomedicine is imported into a culture of traditional medicine), or between expert and egalitarian[3] knowledge (e.g. specialist consensus and non-mainstream conceptions are integrated in the collaborative construction of an online encyclopedia article).

Just as certain factors facilitate or hinder cognitive maturation, certain factors facilitate or hinder knowledge growth in a social context. The growth of shared knowledge depends on equilibration and on a knowledge economy in which knowledge circulates widely, is not lost, is not excessively regulated, can be enjoyed by many, is interactive, is open to recursive processes of knowledge formation, and is highly connected. Thus the growth of shared knowledge is shaped by available knowledge representation technologies. We arrive at our vision of the Epistemic Web by reasoning deductively from the factors that facilitate knowledge growth and the technological capabilities of networked computer systems. Before we come to our discussion of the Epistemic Web, we will examine the history of knowledge representation technologies, stressing historical dynamics and the impact of particular technologies for the knowledge economy and the structure of knowledge.

A Short History of Knowledge Representation Technologies

Much animal and human communication is *context-dependent*; elements of the communicative repertoire are exploited only in response to a specific context. The ancestors of Homo sapiens sapiens developed sophisticated language based on the gestural modality; this language

[3] Cf. (Sanger 2007).

contained *context-independent* elements and was characterized by complex syntax (Armstrong et al. 1995). With the evolution of laryngeal descent, humans became capable of articulating the full range of speech sounds used in modern languages, and syntax was co-opted for the organization of spoken language (although its original function remains for sign language users). Spoken language constitutes the baseline for the knowledge representation technologies that we discuss below. It is portable, if not at all durable, it is difficult to control, and it is not very rival. Dialogic speech has rich potential in terms of interactivity, recursiveness, and connectivity, while monologic speech is highly restricted in these respects.

What follows is a summary of the development of important knowledge representation technologies in human history. Such technologies have their ultimate origin in the first use of symbols, which are known from the Upper Paleolithic. These technologies developed not in direct succession but in overlap, and all persist today. We do not see a simple story of more highly developed technologies replacing more primitive ones. Nor do we find useful the often told story of a few technological "revolutions" that punctuate periods of relative stagnation: the invention of writing, printing with movable type, the Web. The history of knowledge representation technologies rather exhibits complex historical interrelationships between technologies, changing social attitudes toward the technologies, and a dynamic tension between conservatism and innovation.

1. *Mnemotechnics* is unique among the technologies described here in that it involves primarily internal representations. Yet these internal representations are structured in the context of a shared symbol-based technology that is learned, and they involve *loci* that are characteristically dependent upon external representations. Mnemotechnics has its origin in traditions of oral-formulaic poetry that are known in many parts of the world. Verseform functions as a set of constraints that structure content so that it can be recalled for oral performance multiple times with good accuracy (Rubin 1995). These techniques of formal mnemotechnics (traditionally ascribed to the Greek poet Simonides in the early fifth century BCE) involve establishing a mental chain of *loci*—typically envisaged as wax tablets or papyri—in a fixed order; the *loci* are internalized and serve as the background against which concepts, arguments, physical objects and words are memorized (Lewis 2006, 7–8). Mnemotechnics was practiced especially widely and with unique

sophistication among Roman rhetoricians and in medieval monasteries. In the early modern period, mnemotechnics led to the development of such phenomena as commonplace books and tables of knowledge: “forms of technology that exteriorize the means of recollection used in mnemotechnique” (Lewis 2006, 23).

2. *Writing* arose around the end of the fourth millennium BCE (ca. 3300) in southern Babylon (modern Iraq). The earliest written documents are clay tablets impressed with numerical notations and sealings that likely indicated institutional contexts. Although these documents led eventually to the development of cuneiform writing used for the representation of texts in Sumerian, Akkadian and other languages, the earliest writing constituted a symbol system independent of spoken language and used as an instrument of administration for the construction and control of centralized economic systems. On a parallel track, early writing led to calculating techniques and mathematical concepts. Early documents are very closely tied to their particular administrative context and do not represent background knowledge shared by the social actors in this context; in this respect early writing exhibits much of the context-dependence of face-to-face communication. At the same time, writing, in presenting a system of manipulable symbols, allowed for the emergence of new kinds of reflexiveness (Damerow 1996, 46–54).
3. *Glottography* is writing that represents spoken language—although written texts differ in a number of structural ways from speech (Hyman 2006). The potential of writing as a tool for permanently documenting spoken language was discovered only slowly and with increasing usage. When glottographic writing first emerged in the Fara period (ca. 2500 BCE) it served as a mnemonic aid to recording oral genres (proverbs, incantations, hymns, etc.). Glottography led to an increased awareness of language (Krebernik 2007). Subsequently written and spoken language developed as partly independent, partly interpenetrating systems. Glottographic writing eventually spread widely and diverged greatly in form, in response to differences of language typology, social usage and physical media.
4. *Paper* was made from rags as early as the third century BCE in China, but the technique of papermaking using fresh plant

materials is supposed to have been the invention of the Chinese court official Cai Lun in 105 CE (Tsien 1987, 2). In the following centuries, paper improved in quality and popularity, becoming the standard writing material by the third or fourth century. Paper technology spread westward, reaching the Arab world before the seventh century and Europe in the tenth; European manufacture began in the twelfth century (Tsien 1987, 293–303). Paper was a necessary enabling technology for printing (and thus a key advance to increasing the portability and reproducibility of knowledge), which began in China around 700, with movable type introduced by the mid-eleventh century.

5. Although *movable type* had been used for four centuries in China, the printing press, a fifteenth-century German invention, came to have a profound and worldwide effect on the dissemination and production of documents (Eisenstein 1980). It is as a result of this technology that mass literacy was achieved in Europe and other parts of the world in the nineteenth and twentieth centuries. Yet the printing press, for all its potential of empowering the masses with literature, was a technology carefully controlled by the Church or by other authorities. Witness the following report of the attitudes of British colonial officials in India:

   > During the administration of Lord Minto this dread of the free diffusion of knowledge became a chronic disease, which was continually afflicting the members of Government with all sorts of hypochondriacal day-fears and night-mares, in which visions of the Printing Press and the Bible were ever making their flesh to creep, and their hair to stand erect with horror. (Kaye 1854, 247–248)

6. With the Industrial Revolution, new technologies extended printing along several vectors. *Hot metal typesetting*, exemplified by the Mergenthaler Linotype (1886) and Lanston Monotype (1889), increased *automation* by replacing the process of manual composition (in which types were picked one by one from a typecase) with the keyboarding of text (Steinberg 1961, 286). The typewriter, first commercially manufactured in the United States in the 1870s, eliminated the centralized ownership of the means of mechanical production of texts and allowed mechanical

technology to be used for the creation of even ephemeral documents. *Teletype* machines, which originated around 1907, allowed for the remote transmission and printing of text.

7. Jacquard's *punchcard-controlled* loom (1804) and Hollerith's *tabulating machines*, developed to deal with the massive data that needed to be processed for the 1890 United States Census, first exemplified modern techniques of information processing (Austrian 1982).
8. The *mass media* of radio and television in the twentieth century allowed for extremely quick dissemination of knowledge to unprecedented numbers of people, but the ease with which they could be controlled and their low interactivity made them ideal tools of propaganda.
9. *Mimeographic* and *photocopy* technologies, by lowering the barriers of cost, skill, and time associated with the reproduction of printed documents, allowed for the flourishing of popular self-published literatures (*samizdat*).
10. The first *digital computers* greatly augmented human capabilities in managing knowledge in political and economic administration, engineering and the natural sciences. Computers led first to advances in the culture of calculation. Their application to text and language processing followed at first only slowly, but led eventually to a revolution in which the computer came to augment through external technology human mnemonic and linguisticcapacities.[4]

One aspect apparent in this history is a frequent conservatism, in which features of previous knowledge representation technologies and economies are uncritically imported into new ones. Gutenberg's 42–line Bible of the mid 1450s employed a font of almost 300 characters, including a large number of ligatures, alternate letter forms, accented letters and abbreviations: elements that had in the past arisen to *speed up* the copying of manuscripts but that now *slowed down* reading (Steinberg 1961, 20, 30). In much the same way, scholarly articles on the Web make use of features taken over from the book—such as numbered footnotes—although the hypertext medium offers much better alternatives. In general, this history has been shaped by technology, rather than by the purposeful project of

[4] For a recent historical overview, see (Dyson 2012). For the role and meaning of knowledge representation in Artificial Intelligence, see (Brachman and Levesque 2004).

creating a new architecture for knowledge. Knowledge representation technologies hold implications for the forms of knowledge. In Greco-Roman antiquity, for instance, precise citations in texts were extremely rare, as scrolls of papyrus made the checking of sources laborious and time-consuming. Today standardization of publication formats in academia fosters a culture that takes quantity of publications or impact factor (how often and where one is cited) as measures of achievement, although these at best are weak proxies of intellectual merit, and at worst constitute an economy that rewards a high output of low-quality work. By studying how knowledge representation technologies have historically fostered or impeded the growth of shared knowledge, we are afforded a better perspective for redesigning such technologies in the future.

Challenges for the Future of the Web

We organize our exploration of the challenges for the future development of the Web around three general theses about knowledge. We use these theses to draw conclusions about the design the Epistemic Web should take and discuss present obstacles to this design.

Knowledge Is Collectively Produced and Changes in Quantity and Structure

Traditional media such as print, TV and radio are shaped by and reinforce a sender-receiver model of knowledge production and consumption. In contrast, the actual production and appropriation of knowledge typically occurs in a co-operative manner without such a clear distinction between sender and receiver. In a scientific context, the results of knowledge production quickly become tools for the production of further knowledge. Media favoring efficient knowledge production must therefore support these interactive and recursive features and rely on open accessibility to knowledge. They must also be characterized by an equally open availability and adaptability of tools serving to process and network this knowledge. A co-development of knowledge and knowledge infrastructure is required that allows knowledge producers to participate in the development and adaptation of tools appropriate to their purposes.

The large-scale production of knowledge over history is not simply the accumulation of the expertise of a few outstanding individuals. Rather knowledge is produced under complex and dynamic social conditions, in

which external representations play a crucial role in the transmission, appropriation, reorganization and equilibration of shared knowledge. Ideally, therefore, external representations should be dynamic. But traditionally most existing knowledge has been locked into static representations. Thus the processes of the accumulation of knowledge and its restructuring in the aftermath of major conceptual advances remain largely hidden. The integration of old and new knowledge is hindered by the fact that knowledge is fragmented across various media and protected by access control measures that restrict its availability. The complex and dynamic structures of links between documents on the Web represent the relations between different areas of knowledge and in themselves constitute an important kind of knowledge. Yet the present Web lacks means for annotating these structures and creating new knowledge about them; indeed the structures themselves remain largely invisible to both humans and computer agents. Only by increasing connectivity between knowledge and by making the relations between discrete elements of knowledge explicit can the Web overcome the limitations of traditional static knowledge representation technologies.

Knowledge is Produced Recursively

An external representation is internalized, and higher-order knowledge can be formed about this internal representation; this higher-order knowledge can then be converted into a new external representation. The traditional boundary between the production and dissemination of knowledge results from the limitations of prior technologies and now hinders the recursive production of knowledge. New tools are needed to integrate access to existing knowledge with facilities for the production of new knowledge both within and outside science. Existing popular and scholarly publications tend to be superficial, and indeed the traditional media of publication are structured (by limitations of length and established generic conventions) in such a way that such superficiality is almost guaranteed. Publications in computer science don’t include executable code. Historians and political commentators rarely reproduce their primary sources, which remain in public—or, worse, private!—archives and collections. Articles in scientific journals don’t provide sufficient details to allow for the reproduction of experiments.

Experimental data and historical sources are often reproduced only in a piecemeal fashion that does not allow for verification of the authors’

conclusions without extensive research on one's own part. In social and behavioural sciences publications don't allow the production of statistical results due to the fact most of the analyzed micro data is not available because "data protection laws" apply. Moreover, the traditional media of dissemination are not well integrated. Print media contain both images and text, but techniques for linking these are only rudimentary. In recent years, books are sometimes accompanied by other media such as DVDs that allow for the distribution of audio and video, but here the relation between media is even looser. Media are somewhat more tightly integrated in Web publications, but even there they are not linked at a consistent level of granularity or presented with a seamless interface.

Today's social networking sites (in particular Facebook, but also Google+, Flickr and others) function as data silos into which contributions can be pumped, but only extracted—if at all—with extreme difficulties. Shared collections of sources from various platforms are not very easy to realize so that recursiveness is impeded. Moreover, the providers in their "terms-of-use" for these applications authorize themselves to reuse contributions as they see fit. Neither is the problem of sustainability solved. Should Facebook decide to delete contributions, then these simply disappear.

This makes it all the more necessary for knowledge producers to retain possession of their data and to ensure open access to them. Tools such as editorial servers or even the desktop should make it possible for the user to choose through which frameworks their contributions should be made accessible. Contributions should be kept in an accessible standard format, such as XML or Markdown, on an editorial server, remain the property of the owner and be moved whenever necessary to another platform or into another collection at any time. Strategies must be developed to ensure the archiving and longterm availability of contributions on diverse editorial servers. Nevertheless, science platforms have much to learn from Facebook and Co. The possibility of forming groups, having real-time discussions, but also of asynchronous communication can be powerful tools for the production of knowledge.

The quest for open access is not a matter of content communism. Without open access, the Web is bound to replicate the insular structure of information in the print world. Lack of open access constitutes one of the main obstacles to the full exploitation of the potential of the Web to support the recursive character of research and scholarship. But while the actual content in form of digital objects is moving more and more into the public

domain, fired by the open-access movement, semantics is becoming increasingly privatized. Google, Facebook and others are monopolizing the relations between documents and the users interacting with them. Google Books, for instance, makes documents openly available as far as possible, but not background structures such as search algorithms and full texts. The requirements of open access hence needs to be coupled with those of open source.

Knowledge Includes both Data and Models

The evolution of large bodies of shared knowledge is organized around conceptual models that frame data, but the accumulation of data results necessarily in the periodic revision and substitution of these conceptual models. The contemporary knowledge explosion—not only in the sciences but also in the ever-increasing complexity of social and political life in a global culture—results in an acceleration in the change of conceptual models. To prevent the potential ruptures caused by these changes, it is necessary to integrate conceptual models and data within single representations. Only such an integration will allow for research and thinking that address overarching theoretical concerns in the context of concrete, empirical data—so that we can escape the Scylla of empty speculation and the Charybdis of aimless accumulation of detail. If conceptual models were universally shared, they could safely be left unstated; but models differ between communities and change over time even within a single community. Traditional modes of exposition, both academic and popular, are highly conservative and often assume a shared understanding that does not correspond to reality. The problem of the contemporary fragmentation of knowledge necessitates a plastic knowledge representation technology that accommodates both data and models.

The Epistemic Web

In this last part of the paper we begin by articulating the fundamental principles underlying our vision of a Web that can represent human knowledge adequately. We next discuss the architectural cornerstones upon which the Epistemic Web can be built. Finally, we are ready to paint a scenario of how the Epistemic Web should function and to indicate the gains we expect it to yield.

Fundamental Principles

The Web will become a universe of knowledge that parallels human knowledge. After a lifetime of laborious memorization, study and intellectual activity, some individuals manage to obtain a set of rich internal representations of knowledge that provide good overall coverage of a single domain. Experts can summon up numerous items of knowledge quickly. But it takes a lifetime to reach this point, and few manage. Moreover, this store of knowledge perishes with its owner; there is no way of imparting the whole to students or readers. The Web of the future offers hope: powerful search tools will allow immediate access to a wealth of knowledge (primary and secondary sources; echoes and commentary; critiques and response) in a random-access fashion that parallels, but supersedes the limitations of, human memory. And the Web will be able to represent not only the complete store of structured knowledge accumulated in a single lifetime by a single expert, but the collective knowledge of humanity, structured with equal care and richness.

Private reading (and browsing) will be replaced by the public creation of information. The present economy of knowledge on the Web is strikingly atavistic, incorporating anachronistic features of print culture that stretch back to Gutenberg and indeed to the medieval scriptorium. A traditional publication—and most Web publications are precisely this—is a freeze-frame of active, dynamic research and thought. The process of publication involves technical and social infrastructure that typically lies beyond the range of a single author. And what is published on the Web is *browsed*—a term that signifies a casual association of documents. In the Epistemic Web, *browsing* will be replaced by the purposeful *federation* of documents. Users will (in accord with their interests and needs) choose which documents to view together; which documents they wish to select as entryways into the universe of knowledge; and which documents should serve as *master documents*, controlling the views of secondary documents. These decisions do not remain *private* (like annotations in books kept at home); rather, they may result in the creation of *public*, shareable knowledge. One person's views will be made available to, and serve as potential starting points for the explorations of, others. Of course, the publishing of federations will be voluntary. On the current Web, user behavior is subject to surreptitious methods of information capture (by advertisers etc.); the Epistemic Web, by making federation an explicit

activity, will give users control over the information they produce.

All data will be metadata, and all documents will be perspectives into the universe of knowledge. Librarians ordinarily conceive of metadata as a canonical structured vocabulary that describes the contents and form of certain knowledge representations. By allowing for greatly enriched links between documents (incoming as well as outbound links; multi-directional links; transitive and intransitive links; links with attached semantic labels; links with specified behaviors), the Epistemic Web will allow documents to describe one another. Since any document can refer to any other set of documents, a document may be understood as a *projection* of the universe of knowledge that is instantiated in the Web. Each document serves as a *perspective* into the entire universe of available knowledge, and the extent of the view from this perspective is a function of the document's degree of connectivity. Thus documents resemble Leibniz's monads, which "are nothing but aspects [*perspectives*] of a single universe" (Leibniz 1898, §57). Any document that is connected to other documents is in one or another sense *about* those other documents, and it can be construed as metadata.

Architectural Cornerstones

To increase interactivity and reflexiveness a new paradigm is needed to replace the browser/server paradigm. The knowledge consumer and knowledge producer will merge in the knowledge *prosumer*, a term that describes an individual who "co-innovates and coproduces the products they consume" (Tapscott and Williams 2006, 126). We use the term *interagent* to refer to the key piece of software that will enable the Epistemic Web. The interagent will allow the Epistemic Web prosumer to annotate existing documents and create new documents as easily as the current Web user can browse documents. The interagent, like the Roman god Janus, looks in more than one direction: it is the software that mediates interactivity; it allows information production as well as consumption; and it breaks down the division between browser and server. We envision the interagent as a thin client that runs on a user's computer, but that is radically extensible through Web services. Not only does the interagent provide access to the universe of knowledge; it brings a *world of services* to the prosumer's desktop. The interagent can extend its repertoire of behaviors by discovering and utilizing services available on the Web—for instance, when it encounters a new document type, or a new natural language, or a new set of technologies for working with data of a particular type.

A key way of extending knowledge on the Epistemic Web is federation of documents. A group of *federated documents* is brought together by means of a *federating document*. For example, a collection of geographical data sets may be federated into a *mappa mundi*. Or several editions, translations, and commentaries on a literary work may be federated into a *synoptic edition*. In general, federation is a way of bringing together knowledge from existing documents to represent new knowledge. Whereas in the traditional Web the structures of links between documents are mostly hidden and do not allow for annotation, in the Epistemic Web these structures will be exposed as federating documents containing enriched links. In turn such federating documents may be annotated or recursively federated. The interagent will offer facilities for federation, which will be assisted by content analysis technologies that can automatically create provisional federating documents; these documents will then be available for extension and modifications by humans.

Scenario

The Epistemic Web will not be built all at once. Innovation demands the narrowing of the gap between developers and users. The architects of the next-generation Web can promote a technically informed public by creating powerful, flexible and modular tools that are easy to learn, easy to use, and guaranteed not simply to disappear one day. The creation of such tools is an ideal task for the flourishing open-source software community. New technologies will arise from a *virtuous circle* in which technical developments support knowledge production, which in turn leads to new technical developments. Compelling applications will attract users, leading to positive network externalities, more contributors and further gains.

The Epistemic Web depends, of course, on content. Digitization of current knowledge stores is essential but is not enough: knowledge must be accessible, findable, and available for the recursive production of new knowledge. Here there are technical challenges as well as the legal and social challenges of evolving property rights and data protection measures to fit the new knowledge economy. Open-access content is crucial for the growth of knowledge.

The development of knowledge in new areas will necessitate new models for federating documents. Current models such as the encyclopedia model (exemplified by Wikipedia) and the geospatial model (exemplified by Google Earth) are powerful structures for organizing a large amount of

knowledge. But they ultimately are only incremental improvements on content models that have been in use for more than a millennium. As we begin systematically to explore new large-scale topics, such as the comparative study of globalization processes in history and social sciences, we will need new knowledge representation forms to accommodate such phenomena as layered time developments within a geospatial context. One research area of considerable importance is visualization methods, that is "systematic graphic formats, that can be used to create, share, or codify knowledge" (Lengler and Eppler 2007).

The Epistemic Web will have to be a sustainable ecology of knowledge, affording a place for established knowledge and creating space for new knowledge. There will be niches for grassroots innovation as well as for conservative institutions. The Web will grow in an *innovation-stabilization* cycle. Some innovations will showcase powerful new ideas that need to be reimplemented with greater generality. Some innovations will serve the purpose for which they were constructed, and all that will be needed is an infrastructure to ensure their longevity. Some innovations will be dead ends; they can be forgotten, or remembered only as negative examples. Stabilization will ensure that the Web is not cobbled together from prototypes and experiments. Successful innovations will become infrastructure that allows for the next wave of innovation.

The accumulation of knowledge is only possible when mechanisms exist to ensure reliability. Knowledge must be grounded at a low level. In established genres of writing, baseless statements can be couched in the language of authority, allowing them to masquerade as reliable knowledge. Ultimately, higher-level knowledge must be grounded in low-level, concrete, foundational knowledge. A knowledge representation technology based on the principle of high connectivity will help ensure that there is a chain of explicit links that allows knowledge to be verified.

Current discourse about the Web centers around *information*, a word that suggests an undifferentiated, interchangeable commodity, and which is often used in an imprecise way that reflects a "conceptual creolization" (Nunberg 1996). Knowledge, by contrast, is highly structured and is tied to agents: it is what individuals, or social groups, or all people know. Knowledge arises dynamically through equilibration processes. The Epistemic Web constitutes a novel technology that accommodates both local and global, both egalitarian and expert knowledge. By allowing for the equilibration of such disparate kinds of knowledge on an unparalleled scale, the Epistemic Web will make possible the next stage in the globalization of

knowledge.

We have presented a scenario for the Epistemic Web that poses considerable technical and social challenges. We believe, however, that new thinking is needed to transform the Web into a technology that facilitates the production of knowledge in a complex global society. Left to develop in a haphazard fashion, the Web will not spontaneously evolve in an utopian direction. Indeed, the alternative to an Epistemic Web may be a Web in which there is a growing digital divide of competence, a commercial monopoly on content, de-facto monopolies of content due to unsolved data protection problems, a lack of open standards and infrastructure, restrictions on innovation, and ultimately a forking into two Webs: a Web of slick, mainstream content for the many; and an underground, alternative Web for the few.

Acknowledgments

This work originated in the context of a workshop *Infrastructure for Cyberscholarship* held in Phoenix, Arizona, April 17–19, 2007, and sponsored by the US National Science Foundation (NSF) and the Joint Information Systems Committee (JISC) of the UK. We gratefully acknowledge the role that our long-standing collaborators Peter Damerow and Mark Schiefsky have played in shaping our ideas.

Bibliography

Armstrong, David F., William C. Stokoe, and Sherman E. Wilcox. 1995. *Gesture and the Nature of Language*. Cambridge: Cambridge University Press.

Austrian, Geoffrey D. 1982. *Hermann Hollerith, Forgotten Giant of Information Processing*. New York: Columbia University Press.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American* 284 (5): 34–43.

Brachman, Ronald, and Hector Levesque. 2004. *Knowledge Representation and Reasoning, The Morgan Kaufmann Series in Artificial Intelligence*. Amsterdam: Morgan Kaufman Publ. Inc.

Damerow, Peter. 1996. *Abstraction and Representation: Essays on the Cultural Revolution of Thinking*. Translated by R. Hanauer. Vol. 175, *Boston Studies in the Philosophy of Science*. Dordrecht: Kluwer.

Dyson, George. 2012. *Turing's Cathedral: The Origins of the Digital Universe*. New York: Pantheon.

Eisenstein, Elizabeth L. 1980. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe; Vols. I and II*. Cambridge: Cambridge University Press.

Gillies, James, and Robert Cailliau. 2000. *How the Web Was Born: The Story of the World Wide Web*. Oxford: Oxford University Press.

Halpin, Harry. 2004. The Semantic Web: The Origins of Artificial Intelligence Redux. Edinburgh: University of Edinburgh. http://www.ibiblio.org/hhalpin/homepage/publications/html/airedux/.

Hyman, Malcom D. 2006. Of Glyphs and Glottography. *Language and Communication* 26 (3–4): 231–249.

Kaye, John William. 1854. *The Life and Correspondence of Charles, Lord Metcalfe: Late Governor-General of India, Governor of Jamaica, and Governor-General of Canada; From Unpublished Letters and Journals Pressed by Himself, His Family, and His Friends*. 2 vols. Vol. 2, *The Life and Correspondence of Charles, Lord Metcalfe*. London: Bentley.

Krebernik, Manfred. 2007. Zur Entwicklung des Sprachbewusstseins im Alten Orient. In *Das geistige Erfassen der Welt im Alten Orient*, edited by C. Wilcke. Wiesbaden: Harrasowitz.

Leibniz, Gottfried Wilhelm. 1898. *The Monadology and Other Philosophical Writings*. Translated by L. Robert. Oxford: Clarendon Press.

Lengler, Ralph, and Martin Eppler. 2007. Towards a Periodic Table of Visualization Methods for Management. In *Proceedings of the IASTED International Conference on Graphics and Visualization in Engineering: January 3–5, 2007, Clearwater, Florida, USA*, edited by M. S. Alam. Anaheim, CA: ACTA Press.

Lewis, Rhodri. 2006. *From Athens to Elsinore: The Early Modern Art of Memory, Reconsidered*. Preprint 319. Berlin: Max Planck Institute for the History of Science.

Morville, Peter. 2005. *Ambient Findability*. Sebastopol, CA: O'Reilly.

Paolillo, John C. 2005. Language Diversity on the Internet. In *Measuring Linguistic Diversity on the Internet*, edited by J. C. Paolillo. Paris: UNESCO.

Piaget, Jean. 1985. *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development*. Translated by B. Terrance and T. Kishore Julian. Chicago, Ill.: University of Chicago Press.

O'Reilly, Tim. 2005. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.

Rubin, David C. 1995. *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-Out Rhymes*. New York: Oxford University Press.

Sanger, Larry. 2007. Who Says We Know: On the New Politics of Knowledge. www.edge.org/3rd_culture/sanger07/sanger07_index.html. Edge.
Schröder, Sebastian. 2007. Web 2.0 und der Einsatz in der Wissenschaft. Ph.D. Thesis. Department Informatics and Media: University of Applied Sciences Brandenburg.
Steinberg, Sigfrid Henry. 1961. *Five Hundred Years of Printing*. 2. ed. Harmondsworth, UK: Penguin Books.
Tapscott, Don, and Anthony D. Williams. 2006. *Wikinomics: How Mass Collaboration Canges Everything*. New York: Portfolio.
Tsien, Tsuen-Hsuin. 1987. Chemistry and Chemical Technology. Part 1: Paper and Printing, edited by J. Needham. Cambridge: Cambridge University Press.
Yu, Liyang. 2007. *Introduction to the Semantic Web and Semantic Web Services*. Boca Raton, Fla.: Chapman and Hall/CRC.