

From Research Challenges of the Humanities to the Epistemic Web (Web 3.0)

*Malcolm Hyman and Jürgen Renn
(Max Planck Institute for the History of Science, Berlin)*

This paper is an outcome of a longstanding cooperation with Peter Damerow and Mark Schiefsky dealing with the implications of the new information technologies for scholarly research in the humanities. On the background of our joint experience we feel that time has come for directing work towards the creation of a new implementation of the World Wide Web. This should not be entirely surprising. The original invention of the Web was the result not of technical improvements nor was it the work of a committee. It rather emerged from reactions to challenging problems posed by the enterprise of scientific work, that is, from an innovative usage scenario. The Web of the future will similarly be the response to challenging intellectual and practical problems and not just grow out of accelerating technical developments. In view of the obvious role that understanding the representations, structures, contexts, and developmental processes of knowledge will play in this process, the humanities and, in particular fields like the history of science, cognitive science and linguistics, must bring their competence to this endeavor. The future Web will be an Epistemic Web, connecting not just texts but knowledge and people. It will grow out of the research challenges of the humanities which are ever more closely related to the problems posed by transforming the present Web of garbage information into a Web of science and culture.

For this reason, we shall begin by formulating some of the challenges, needs, and ideals of humanistic research, in order to then discuss the obstacles facing such research. Next, we shall confront our vision of a future Web with those that are currently discussed. The architecture of a future Epistemic Web will then be discussed, as well as some of the obstacles impeding its construction. We close by indicating some developmental pathways that may lead to the realization of our vision.

Research Challenges

A discussion of research challenges confronting the sciences and, in particular, the humanities is naturally shaped by the experience with traditional media. In order to discuss the potential of the new media for future research perspectives, it seems helpful to formulate these challenges first with a view to intellectual goals in order to then assess their attainability in a particular environment.

Creating dynamic representations of knowledge

Knowledge in the humanities as well as in science generally is collectively produced and steadily changes both in quantity and structure. External knowledge representations should therefore ideally have themselves a dynamic character, serving, at the same time, as a reliable collective storage and a convenient medium for the communication of knowledge. Even more importantly, any external representation of knowledge typically becomes itself the object of higher-order thinking (as when numbers representing arithmetical knowledge become the objects of higher-order mathematical thinking), thus engendering an iterative process of representation and reflection. The crucial role of representations for shaping the historical development of knowledge derives from this iterative process. Creating dynamic representations of knowledge could optimally support this process.

Integrating research and dissemination

Research and dissemination are but two aspects of the process of generating shared knowledge, involving an equilibration process between private and public dimensions at each stage. Their neat separation is largely an artifact of traditional media with high latency. Producing new knowledge in individual research involves actively accessing the available knowledge of mankind in ways that change both, the individual and the shared knowledge. The reflectivity generated in such interchanges is ultimately the only source of the reliability and “objectivity” of scientific knowledge. Integrating research and dissemination could thus enhance both knowledge production and quality control.

Facilitating the recursiveness of scholarship

Scholarship is intrinsically recursive in that new knowledge can only be generated by preserving and changing prior knowledge. Knowledge is always also self-referential, a characteristic that is only to a very limited extent embodied in traditional citation networks, fulfilling diverse functions at the same time, from linking chunks of knowledge via ensuring reliability of information to attributing or pretending merit. Facilitating the recursiveness of scholarship beyond such ambiguous networks could therefore support the self-organization of knowledge without imposing artificial constraints on its structure such as those inherent in the traditional publication format.

Integrating conceptual models and data

Scholarship is typically directed at understanding epistemic objects such as the medieval city, the globalization of knowledge or climate change. For this purpose ever new heterogeneous data are related to ever changing conceptual models of these epistemic objects which, in turn, may themselves be modified during the process. Limited access and linking capabilities together with a traditional division of labor have, however, tended to keep the relation between conceptual models and data rather opaque, resulting in a dominance of idiosyncratic expertise and a lack of synthetic research. Integrating conceptual models and data could instead foster the study of overarching questions, avoiding the unfortunate alternative of speculation without empirical support and nitpicking without theoretical perspective.

Incorporating intelligence into working environments

Scholarly research makes continuously use of earlier results transformed into instruments of further investigation – another way of looking at the recursiveness of scholarship. The transformation of the consequences of research into presuppositions of further work materially embodying the intelligence earlier invested becomes most clear when research results enter the construction of ever more sophisticated laboratory equipment. But the same process takes place also in the humanities. Publications may become sources of information or evidence in ongoing investigations; some scholarly results enter handbooks, dictionaries, lexicons, or other reference works specifically serving as research tools. The quality of scholarly working environments crucially depends on both, the effectiveness with which earlier results are transformed into instruments of further work and the seamless availability of these instruments. Incorporating intelligence into working environments can thus be seen as a crucial factor enabling good scholarship.

Research Obstacles

Considering the challenges, needs, and ideals associated with scholarly research helps to distinguish unavoidable limitations imposed by a historically specific state of development from unnecessary obstacles due to inadequate use of existing potentials. Often these obstacles are also related to external interests overlaying and interfering with those of scholarly research, such as greed for power and money.

Data locked into static representations

As long as data are locked into static representations, they can only indirectly reflect the process of the accumulation of knowledge, let alone its restructuring in the course of major conceptual advances. Under this condition, connecting new to old data is a process that resists automatization because it involves individual activities outside the primary medium of knowledge representation such as looking up references in a library or paying for an information service on the Web. The Web offers, on the other hand, the potential for overcoming the static and insular character of traditional knowledge representations in favor of a seamless and visible connectivity. But the present Web can realize this potential only to a limited degree. While the hyperlinks of the Web represent structures of meaning that transcend the meaning represented by individual documents, these structures of meaning can so far hardly be made themselves the object of interventions by the Web community. Due to the static distinction between servers and clients, there is at present no way to construct complex networks of meaningful relations between Web contents. In fact, the providers have no influence on the links to the contents provided by them and the users have no impact on the available access structures to the content, except by becoming content providers themselves.

Superficial representations of research results

Just as much of the epistemic connectivity of research remains invisible in the traditional medium so does the information hinterland of research results, from observational data, computer models, visual material, to archaeological records and manuscript material. The established publication system amounts to a very limited, superficial representation of the underlying knowledge which remains largely concealed or implicit in the social interactions and material techniques of research. The Web has altered the preconditions for communicating research, not only with regard to speed and connectivity but also with regard to depth. Exploiting this potential could contribute to the reliability and transparency of scientific knowledge but also to a more effective sharing of resources. Instead the lack of openness of the present Web – in the humanities e.g. the lack of open access to cultural heritage – as well as some aspects of the career system of science favor the privatization of knowledge and its superficial representation in public communication.

Fragmentary character of knowledge representations

The hallmark of the traditional publication system is the fragmentary character of knowledge representation. Indeed, the traditional publication system and the disciplinary organization of the sciences are sustaining a knowledge explosion which is at the same time a knowledge fragmentation. In periods of normal science, the integration of these fragments into a meaningful whole is largely taken for granted rather than being made itself the objective of research. Since the mechanisms of knowledge integration tend to remain rather implicit, problems that might engender a reorganization of knowledge may thus remain inconspicuous. The competitiveness of the academic system further contributes to the fragmentation of knowledge and the neglect of

integrative ventures. The Web has so far hardly offered new perspectives of integration but has rather made the puzzle even more complex by dramatically increasing the availability of non-descript fragments of knowledge.

Lack of structural correspondence between content and representation

The standardization of publication formats is part of the well established traditional information circuit, from research, via refereeing and publication, to archiving and retrieval as a presupposition of further research. Standardization refers both to external formats (monographic and edited books, research and review papers, abstracts, etc.) as well as to internal formats (table of contents, footnotes, index, bibliography, etc.). The fact that these formats are largely content-independent makes the traditional publication system robust and seemingly contributes to its “objectivity” as counting publications and citations can be taken as a measure of scientific productivity and academic advancement. The abstraction from content has, however, also serious drawbacks. Quantitative success in the publication system (as measured by the impact factor) is ever more becoming a value in itself, leading to an impact fetishism. What is more important, the implicit network of publications referring to each other can hardly be considered an adequate representation of the epistemic network of the contents they are dealing with, while such a structural correspondence is quite conceivable in the electronic medium, as innovative knowledge representations such as Google Earth demonstrate: While the relation between a text and a footnote is devoid of any content, the link between an electronic map and information about a particular place is part of a structure that itself carries a meaning. Even in the present Web, the embedding of a publication within a network of links is much more telling than the traditional counting of its citations, as this network embodies (often immediate) qualified reactions to the contribution of a given piece of research to the developing shared knowledge. But this is hardly more than a glimpse at the potentially far-reaching consequences of the self-reflecting character of the new information system.

Missing links between media

In the traditional system the media of research and the media of dissemination are separated from each other, as are individual sites of research and individual publications. The Web has facilitated the transfer of content between the worlds of research and dissemination but its potential for overcoming the insular character of research sites and individual publications has largely remained unused. Different types of software are being used to pursue research, to write a paper and to make it available on the Web. The seamless integration of texts, images, films and sound in research publications seems an obvious request but is still not accomplished. Powerful technologies for analyzing and processing images or natural and formal language content such as mathematical expressions have long been available offline but are still not available on the Web. For mathematical content even elementary facilities for the transfer of content between the Web and offline resources are missing.

Conventional Vision

The conventional vision actually corresponds to a sequence of attempts at understanding the new medium according to expectations rooted in experiences with traditional knowledge representations. They are evolving because of their emerging incapacity to cope with the radical nature of the break constituted by establishing the Internet as a potentially all-encompassing representation of human knowledge. Yet they are not evolving adequately because they focus on what seem to be deficits of the new medium with regard to the old that need compensation, a framework for a well-structured division of intellectual labor in the sense of the traditional

information circuit, an orderly universe of meaning allowing to relate the pieces of the information puzzle to each other, and niches like speaker's corner for free-wheeling discussions. The innovative power with which the compensatory measures are being undertaken can hardly distract from their impotence to accomplish the profound transition to a new information age the Web is piloting.

Replicating the traditional knowledge economy in the new medium

The easiest way to establish a framework for a well-structured division of intellectual labor on the Web is certainly to emulate in the new medium the traditional information circuit from research, via refereeing and publication, to archiving and retrieval as a presupposition of further research. In this way, scholars, publishers, and librarians could maintain their well-defined roles as well as their respective sources of recognition, revenue, and respect. The stability of this traditional system is, however, endangered: due to the increase in the number of publications and their rising costs, adequate access to research results is becoming ever more difficult; due to the novel economics of the Web, the traditional distribution of labor is collapsing. Nevertheless, there is a natural tendency by all stake holders to map the traditional system into the new medium, ignoring rather than exploring its novel possibilities or to rethink the implications of research challenges when pursued in a new context. The prevalence of such a conservative vision is characteristic of all technological revolutions; it is like building cars according to the model of horse carriages. In the case of the Web, however, the pursuit of this vision is expensive, the investments having to be subtracted from those necessary for building up a more adequate infrastructure optimizing the new technological potentials for addressing research challenges. Creating toll-access structures to scientific knowledge thus hampers research twice, impeding connectivity by blocking access and by preventing the creation of connective working environments.

Solving the problem of semantics on the Web by central planning

If the Web only knew what it is talking about, it would understand itself much better – but who is going to teach it? Evidently, the potential of the Web as a universal representation of human knowledge and communication would be greatly enhanced if its sites could “speak” to each other in the sense of recognizing if two figures refer to the same date whatever the format or if two texts refer, say, to soccer whether the word occurs or not. The Web's semiotic connectivity would thus be transformed into a semantic connectivity. One of the strategies of adding meaning to data is using metadata establishing the data's significance, for instance by referring to ontologies offering common frames of semantic reference. Historically, attempts of creating such a second world of meaning with a claim to universal validity are familiar from the Catholic Church and the Soviet Union. They typically involve a great deal of central planning and are characterized by the rule of technocrats as well as the incapacity to cope with developmental dynamics. Providing meaning to the Web with the help of metadata created by expert groups and committees ultimately amounts to an Orwellian vision of the Web in which adopting Newspeak is obligatory for being part of the accepted community. Natural language works differently in that meaning emerges rather than being predefined – as dangerous as that may be for any classificatory order. It would thus be better to learn from natural language: One of the reasons for the power of natural language in representing and furthering human thinking is its inbuilt reflexivity: natural language is its own metalanguage. The separation of language and metalanguage makes it to a certain degree possible to fix the semantics clarifying communication and avoiding paradoxes but severely limits the potential of creating new meaning.

Moving from a textual network to a social network; but not yet a knowledge network

Web 2.0 is the protestant version of the Semantic Web: where central authorities have failed in mediating between the real world of data and the transcendental world of meaning, smaller, self-organized groups feel that they are called upon to open up direct access to this transcendental world in terms of their own interpretations of the Great Hypertext. The traditional separation between providers/priests and clients/laymen is thus modified in favor of a new social network in which meaning is actually created bottom up. The unrealistic idea of taxonomies inaugurated by top-down measures is being replaced by the more feasible enterprise of “folksonomies” spread by special interest groups. As their scope remains, however, rather limited and the separation between data and metadata essentially unchallenged, the chances for developing such a social network into a knowledge network fulling coping with the real world of data are slim.

Alternative Vision

Creating a universe of knowledge on the Web that parallels human knowledge

After a lifetime of laborious memorization and study, some individuals manage to obtain a rich internal (mental) representation that provides good overall coverage of knowledge in a single domain. A relevant example is a traditional Sanskrit *pandit* — an orally-trained scholar who is able to quote Pāṇinian sutras, provide instances of their application, and refer to many centuries of commentary and meta-commentary on them. We might consider also a Latinist who has internalized the universe of Latin letters; for instance, a scholar of Vergil, who can instantaneously identify allusions to Ennius and Lucretius, who can note the places in which Vergil is echoed by Silver Latin epicists, and who can discourse seemingly ad infinitum on any word of the Vergilian text. The scholar who has internalized such knowledge is happy indeed — able immediately to summon from memory a daunting quantity of information. The scholar’s memory is *random access* — providing quick and flexible recall of virtually *anything* at *anytime*, ad libitum. It takes a lifetime to gain such fluency; and, even in a lifetime, few manage. The Renaissance ideal of the *memory palace* is a persisting dream for scholars; yet for all but a handful of aged and wise savants, it remains beyond reach.

Today’s digital technology offers new hope. Scholars don’t need to scroll through long papyri, or flip endlessly through codices. Powerful search tools will — we emphasize *will*, because the Web as a place for research still lacks much — allow virtually instantaneous *random access* to a wealth of information: primary sources, secondary sources; echoes and commentary; critiques and response. Ideally, a gifted young scholar of tomorrow will possess tools that in the past were the exclusive property of senior *magistri*.

The mental resources of the traditional scholar are formidable: but how much more formidable if transferred from internal *mens* to external *machina* — and combined with resources of countless other scholars? The private domain of the sage then becomes part of a public domain that represents in external form, accessible to all, the collective knowledge of humanity.

Turning (private) reading into the (public) creation of information

Today’s digital economy of knowledge is strikingly atavistic — incorporating anachronistic survivals from the age of print publication, stretching back to Gutenberg, and indeed to the medieval *scriptorium*. An author decides to publish, and in so doing presents a freeze-frame of active, dynamic research to the world; to *host* this publication on a *server* requires complex

technical and social infrastructure, often beyond the means of individual scholars. Once the work is published, the *public* or *audience* (these terms are not rigorously defined) have the “luxury” of *browsing* the publication — an activity that in some ways is even more passive than *reading*.

In the future Web that we envision, users will not *read*, and a fortiori will not *browse*. Instead, they will *federate documents*: that is, choose which documents to view together, choose (according to their interests) which document will provide an access point into the *universe* of digital knowledge, and which documents will be displayed as subsidiary documents of the *master document*. These decisions will not remain private. Future scholars will not follow the model of the erudite gentleman who sits in front of his fireplace and annotates personal volumes. Rather, private *reading* becomes public *federation* — in choosing which documents to view, and in establishing the relationship between these documents, the scholar of tomorrow will be creating *public, shareable* information. Such views will be made available to, and will serve as a starting point for the work of, other scholars.

Allowing all data to be metadata and all documents to be windows into the universe of knowledge

(Digital) librarians are enamored of metadata — they see metadata as a canonical set of structured vocabulary that can be used to describe the content held by a (digital) library. This view is short-sighted. On our model, documents can contain enriched links — links along the lines of the XLink standard of the W3C. That is: incoming as well as outbound links, bidirectional links, links that can be followed transitively, links that are semantically labeled, and links whose behavior is explicitly specified. Since any document can refer to any other set of documents, a document may be understood as a *projection* of the universe of data (or content, or knowledge) that is instantiated in the Web. That is to say, it systematically maps a subset of the total knowledge that is represented on the Web. Each document has the potential of serving as a window into the total universe of knowledge; and the richness of the view through this window is a simple function of the richness of the controlling document. Thus the Epistemic Web offers two (compatible) models that define the relation of a single document to the sum of all documents. First, the process of *federating* documents allows for the novel bringing-together of a set of arbitrary documents. From such federated documents, we may expect emergent properties. Second, the process of *projecting* documents allows for the specification of how any single document provides a window into the entire universe of knowledge. Obviously, the richer the set of links in a document, the richer this document is (potentially) as a projection of the entire *docuverse*. Since any document allows outgoing as well as incoming links, any document is potentially *about* one or more other documents. This *aboutness* is the quintessence of *intentionality*; thus any document in the universe is open to being construed as metadata about one or more foreign documents.

Architecture

The Epistemic Web will demand a new architecture, which builds on the architecture of the current Web, and incorporates elements of certain “alternative Webs” (both those that exist only in theory, and those that are *in statu nascendi*): the *semantic Web*, the *social Web*, and the *geospatial Web*.

Moving from servers and browsers to interagents that allow people to interact with information

The current paradigm of the Web — in which the user *browses*, leaving behind a click-trail that is of interest primarily to marketers — falls far short of the needs of scientists and scholars.

Browsing the Web is scarcely more interactive than surfing television channels. The Web server — and the server administrator —, in turn, are survivals from the traditional print economy. True interactivity — which will allow the Web finally to achieve its potential as a medium for scholarly, political, and social dialogue — demands something other than the current browser/server paradigm. A new approach will be needed, where developers recognize that information consumers are also information producers. In fact, the salient individual here is the *prosumer*, a term invented by Don Tapscott to describe the individual who “co-innovates and coproduces the products they consume.”¹

Although the democratizing of information production (and not just consumption) is emerging in the “Web 2.0” paradigm, true interactivity demands a new tool: not a browser, but an *interagent*. Such a tool will allow the user of the future Web to annotate existing publications and create new publications with the same facility that she or he currently enjoys in *browsing* or *reading* existing publications. The interagent, like the Roman god Janus, looks in several directions: it is the software that *mediates* interactivity; it allows information *production* as well as information *consumption*; and it blurs the role of the Web browser and Web server. We see the interagent as a thin client that runs on a user’s computer, but that is radically extensible through *Web services*. In other words, it brings not just a *world of data*, but also a *world of services* to the user’s desktop, leveraging the increasingly service-oriented architecture (SOA) of the new Web. In practical terms, 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.

Replacing browsing and searching with projecting and federating information

Browsing, as we have argued, is *passive* and *private*. In the Epistemic Web, the user does not aimlessly surf. Rather she or he actively chooses which documents to view. But in this model documents are not information islands, but rather something akin to Leibnizian monads. Each document is potentially a projection of some part of the entire universe of knowledge. Any document supports both outgoing links (familiar from HTML) and incoming links (an XLink technology, the potential of which is only beginning to be realized.) We might propose a new motto: *information wants to connect*. A further key concept in this new model is the *federation of documents*. A group of documents (we call these *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*. Powerful linking facilities will allow one document to serve as a viewpoint for multiple other documents.

Enabling automated federation through an extensible service architecture

Current Web browsers are constructed according to an essentially monolithic architecture. To be sure, they allow for “plugins,” but the very term suggests an *afterthought*, an *appendage*. The interagent, by contrast, is a thin software application (one might call it a “kernel”) that runs locally and discovers richer behaviors through a Web-services architecture. The interagent doesn’t know how to segment Thai words? It finds a Web service for that. It doesn’t know how to transliterate Tibetan? It finds a Web service for that too. It can’t handle “Ruby” annotation of Japanese? It can learn. It can’t lemmatize Sanskrit inflected word forms? It finds a Web service for

1 Don Tapscott and Anthony D. Williams, *Wikinomics: How Mass Collaboration Changes Everything*, Portfolio (2006), p. 126.

doing that too.

But such extensibility is only one aspect of the problem. Future Web techniques will depend on new software — and also on old software that is ported to the Web. Why not allow documents to be indexed by Latent Semantic Analysis (LSA)? Why not provide a powerful Web-accessible statistics toolkit, which can be applied to any data on the Web? Surely the infrastructure can be accommodated to allow automatic linking to lexica, thesauri, specialized references, and so on. Data should be effortlessly pluggable into specialized tools. For example, if I have quantitative data, I should be able to drag-and-drop it into Mathematica, or Maple, or R (or SPSS). If I have visual data, I should be able to apply domain-specific image enhancement. Why shouldn't OCR work over the Web? Where are the tools that allow us to perform flexible searches on multimedia data, such as sound, video, and still images?

Extending current hypertext architecture with granular addressing and enriched links

Key to the architecture that we envision are two extensions to the current architecture of the Web. First, we need granular sub-addressing of documents, which is independent of the explicit structure of those documents. For XML documents, we can obtain arbitrary document fragments through standards such as XPath, XPointer, and XQuery. But to do so, we must know how the document is marked up. We need a layer of indirection that allows us to abstract from document structure. In particular, four means of sub-addressing XML documents are needed: (1) specification of a structural division of an XML document, without knowing the specific grammar of the XML document; such abstraction depends on a translation layer that maps the markup vocabulary of the XML document to a *lingua franca*; (2) reference to a particular word in an XML document, where the word may either be explicitly tagged, or where NLP technology segments words automatically; (3) specification of an arbitrary range from *word₁* to *word₂* (inclusive), where the range is specified in terms of words referenced according to the scheme described above; (4) a set of structural sub-addresses, drawn from the four enumerated items, which may include an ordered combination of {*structural divisions, words, ranges, sets*}. These, we emphasize, are the basic requirements for core humanistic disciplines such as history and literary studies. Certain disciplines, such as philology and linguistics, will need more powerful and granular techniques for subaddressing documents.

The second requirement is *enriched links*. The XLink recommendation of the W3C is a considerable move in this direction. Links as implemented in (X)HTML are notably impoverished. For all practical purposes, there exists only a single kind of link (the *anchor* element with the *href* attribute, and this link has a single specified behavior: when the user activates the link (in a implementation-specific way), the current document view is replaced by the view of the link target. Enriched linking standards such as XLink suggest a great variety of alternatives: links that are displayed inline, rather than as a replacement of the current view; incoming as well as outgoing links; specification of whether links are to be automatically followed or only followed by user interaction (as in the (X)HTML model); specification of the semantics and behaviors of links; amalgamation of a set of links; links that may be followed transitively. These new linking behaviors, especially combined with a mechanism for granular sub-addressing of document structure, promise the basic building blocks for the Epistemic Web.

The foregoing discussion has concentrated on links within textual resources, primarily those represented in XML. But the principles are readily extensible to other sorts of resources: image data, which may be sub-addressed by Cartesian coordinates; geospatial data which may be sub-addressed by geospatial coordinates; and audio and video data which may be sub-addressed by temporal ranges. Other types of media propose new challenges, and we call for basic research in

these areas.

Development Obstacles

The ideas and technologies for transforming the Web from a data whisk to a knowledge representation clearly exist. The realization of this potential suffers, however, from a number of development obstacles, ranging from the growing gap between scholars and developers, closely connected with the lack of modular working environments and authoring tools, via the confusion due to the maze of possible directions opened by the explosion of technology, to the tyranny of information technocrats and the unremitting lack of openness due to the persistence of proprietary interests.

The growing gap between scholars and developers

While an emerging tradition of computational humanities has achieved remarkable breakthroughs, it has largely remained at the level of pilot ventures that have failed to bring along the mainstream. The gap between research practice and technological potentials characteristic of the humanities is actually growing. The bulk of cultural heritage representing the object of humanistic studies is still not openly available online, nor do electronic tools for its analysis belong to the standard methodology of the humanities. Most scholars in the humanities are not involved in bringing their experience to bear on forging the cultural techniques of the future appropriate to the Web, nor are they even familiar with the existing standard methods for extracting and encoding text structures in the electronic medium. Developers, on the other hand, are usually not involved in innovative usage scenarios, focusing instead – driven by uninformed “needs analyses” – on emulating traditional practices in the new medium. Scholars and developers thus form an unholy alliance of ignorance: scholars fail to see the technological potentials and developers are blind to the intellectual visions these potentials may give rise to.

The lack of modular working environments and authoring tools

The developmental potential inherent in the combination of new technologies with intellectual visions might be freed if scholars were able to quickly learn about and flexibly use the rapidly developing new technologies. Making new tools broadly available requires, however, to decouple the knowledge necessary to invent and construct such tools from that necessary to use them, just as knowing how to use a refrigerator should not depend on mastering thermodynamics. The tools should, on the other hand, be suitable for being customized and combined in innovative ways so as to avoid just replicating standard working environments. What are needed are thus modular working environments functioning according to the box of blocks principle as the optimal synthesis of the competencies of the block builders and the block users. Such working environments should, in particular, also include authoring tools allowing for the seamless integration of individual research and “knowledge weaving” on the Web.

The maze of possible directions opened up by the explosion of technology

While the spreading of the new information technologies within scholarly communities requires a certain standardization and black-boxing of tools, any such channeling is risking, on the other hand, to go astray and to miss new opportunities brought about by the rapid technological development. The explosion of technology opens up a maze of possible directions for creating a new scholarly infrastructure, an *embarras de richesse* that itself has become an obstacle of development. Therefore, while scholars do not have to be familiar with all the technical details,

they nevertheless better keep a critical eye on the general development, accumulating, if possible, a reflective knowledge keeping them from missing opportunities but also from falling into the traps of false technological promises.

The tyranny of information infrastructure technocrats

The challenges associated with the transition to a new medium of knowledge representation have begun to trigger a new distribution of labor. Nevertheless, the old structure shapes the new one according to its own image, as each group – scholars, librarians, publishers, developers – tends to somehow preserve its role by looking for “new tasks” rather than altogether dropping its original identity. Another ugly imprint left by the old structure on the incipient new one is the focus on the deficits rather than on the potentials of the new medium when compared to the print world, such as the lack of reliability, stability, ranking and order of the available information. The spell of the old under which the new information world is born is embodied in the emergence of a new class of specialists who represent technological competence to the scholarly world and scholarly competence to the world of developers but does actually not possess either one. To everybody they offer both, the exciting impression of being part of an unavoidable revolution and the tranquillizing feeling that nothing really will change. No doubt, the arrival of the infrastructure technocrats follows a historical logic. As it became ever more clear that a new continent of information could not be settled by scout solutions and pilot ventures, colonial officers were called upon to create an infrastructure. Its promise are reliable repositories, stable references, transparent data structures, seamless interfaces, and maybe even a self-sustaining dynamics of transition to the new medium. But building infrastructures quickly became an end in itself, also as a *raison d'être* of the infrastructure architects. They have actually come to dominate the discourse about the transition to the new medium due to the resources they administer, the committees they rule, and the time they can invest into the enterprise. The infrastructures they conceive are accordingly administration oriented rather than research driven, focusing on issues of authentication, standards, and metadata rather than on innovative usage scenarios. They are typically developer rather than user oriented, they tend to decouple technical developments from research and to be over-engineered, risking to quickly become obsolete. A dynamic synthesis of scholarship and technology will only become possible once the tyranny of the infrastructure architects is broken.

The persistence of proprietary interests and the unremitting lack of openness

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 innovative potential of the Web for research and scholarship. In the sciences open access refers to publications as well as their hinterland of data, simulations, software etc.. In the humanities open access should similarly refer not only to publications but also to testimonies of cultural heritage, to historical works of art, literature, and science, to image, film and sound collections, to statistical data, etc.. There is, however, a major difference between the humanities and the sciences: while in science the raw data constituting the hinterland of research are typically produced and kept by the same people who write the publications, authors in the humanities are as a rule not those who collect and preserve cultural heritage or provide access to it. Research institutions and cultural heritage institutions tend to perceive their interests in different ways. While most research institutions see their mission only half accomplished if they are not employing the optimal tools for granting access to their output, holders of cultural heritage tend to conceive electronic reproductions not as a new way of preserving and sharing the memory of mankind but merely as a new source of

revenue they can use in compensation of dwindling public funds to fulfill their traditional function. The duplication of the world of cultural heritage in the electronic medium has actually triggered a gold rush. It motivates museums, libraries, archives as well as private companies to stake out proprietary claims in this new territory. They tend to speculate on the quick commercial exploitation of resources rather than fostering their integration into a global representation of human knowledge. Unfortunately, they are assisted in this exclusive policy by those humanists who are all too willing to compromise open access in the interest of their own exclusive, academic niches. The arduous goal of open access in the humanities can only be achieved when public institutions no longer invest into endeavors with proprietary output.

Pathways

The alternative vision that we have developed is ambitious. We believe, however, that it is realistic; to stress its feasibility, we close by discussing some concrete pathways to the Epistemic Web.

Establishing an ecology for innovation by fostering research-driven developments, opening creative niches, and creating positive network externalities

Much technology is *developer-centric*; developers code what they want, and they employ the concepts that they know. Sometimes developers engage in a “needs analysis”: they go around to (potential) users and ask open-ended questions about what users would like. Users often have no understanding of the realities of large-scale technology development, and so what comes out of the “needs analysis” procedure is all too often a laundry list of desiderata, some of which may well be mutually incompatible. We propose a model that is based on a *virtuous circle* in which researchers and technical developers provide each other with feedback. Research in the humanities and sciences opens up challenges for technical developers, while technical developers can offer researchers new perspectives on, and tools for, their work. The cycle works iteratively to improve the quality of both technical tools and scholarly research.

Ideally, however, we should work to close the gap between technical developers and “users.” Individuals engaged in many areas of research would benefit from basic yet *extensible* tools, as well as a modicum of technical know-how. They need not be — in fact, no one expects them to be — experts in software development. Yet insights that arise in the course of their research projects may allow them to build innovative tools. We might call such people *technically informed scholars*. In the past we have seen powerful and useful creations arise from such *technically informed scholars* — often developed with systems such as HyperCard, or FileMaker, or scripting languages. The power of these systems is considerable; yet we have no real equivalents for the *technically informed scholar* who wishes to create innovative tools on the Web. Somebody who felt comfortable writing a small program ten or twenty years ago is likely to feel overwhelmed by today’s technologies. Yet *technically informed scholars* are crucial to innovation in scholarly computing. The architects of the next-generation Web must provide them with powerful, flexible, modular tools that are easy to learn, easy to use, and guaranteed simply not to disappear one day. The creation of such tools is an ideal task for the ever-growing community of open source software developers that are strongly committed to contributing their expertise to the public good.

Moreover, if innovative developments are to flourish, they must be adopted. The value of software systems increases with the number of users — the relevant economic term is *positive network externalities*. As more people use Linux and Firefox, these tools become increasingly valuable. A *fortiori* the value of the Internet increases as more people use it. The free software and open source movements are replete with evangelists. The scholarly computing community is less well-

served — in reality, it is often a dangerously isolated niche in academic and research institutions. If we build an Epistemic Web, we must make sure that scholars understand its benefits, and that the barriers to entry are low. Furthermore, we must put our money where our mouth is; we cannot blithely continue to publish our research primarily in traditional paper journals and books.

Ensuring sustainability through an innovation-stabilization cycle, focusing on open-access content as mediator between research and technology

Open-access content is key to the viability of the Epistemic Web. If knowledge is locked in the box of outmoded intellectual property rights, it cannot be accessed, augmented, or modified in the ways we envision here. It is incumbent upon public institutions that possess a wealth of intellectual property to make it available under an open-access scheme. Universities and research centers are *public institutions*; as such, they should eschew *private property rights*. Just as the value of the Epistemic Web increases with the number of users, so it increases with the amount of content that it includes. Without content, researchers cannot do research. At the same time, content is critical to the technological developments that will lead to the Epistemic Web. Content is not just *data* — things that are given, what is already there. Content is not streams of bytes: we will not measure the value of the Epistemic Web by counting gigabytes, terabytes, or petabytes. Nor is content merely such things as digitized books, digitized paintings and photographs, or numerical climate data acquired over the past century. Content includes scholarly annotations; relationships between places, persons, and events; semantic maps; mathematical models; and conceptual structures of all kinds. We have less experience in dealing with content of these kinds, because so much effort has been invested in digitizing the Gutenberg universe. As a result, work is needed to develop schemas for the representation of additional sorts of documents, and technical developers will have to work hand-in-hand with researchers.

The Epistemic Web will not be built all at once. There will be no launch date. Rather its development will depend on an *innovation-stabilization* cycle. We have already argued that an ecology for innovation is needed, and that researchers themselves will play a crucial role in innovation. Some innovations will showcase powerful new ideas; these will need to be generalized and reimplemented by professional developers. Some innovations will serve the purpose for which they were constructed well; all that is needed in this case is an infrastructure that ensures their longevity. Some innovations will be dead ends; they can be forgotten, or remembered only as negative examples. But innovation — *the revolution* — must continue. The Epistemic Web must not develop into a static system with a frozen function set. Besides innovation, *stabilization* is the other key phase of the development cycle. We don't want a Web that is cobbled together from prototypes, experimental software, and unfinished projects. What is learned from innovation must be reused and reimplemented with maximum generality and robustness. Yesterday's innovation become tomorrow's infrastructure. And it is this stable (but evolving!) infrastructure that will provide the platform for further innovation.

Launching Web 3.0 techniques for core resource types

Digitized texts, images, audio, and video, as well as formal language content and numeric datasets together account for much of the material that scholars will draw on for their research. Early work on the Epistemic Web should give priority to addressing improved ways of working with this content.

First, more computation needs to be brought onto the Web. A wide variety of tools that enhance the usability of content are either presently available or actively being researched. But most of

these tools are only available offline, and it is not easy to apply them to resources on the Web. For textual content, we urgently need more NLP technologies brought to the Web: especially technologies for lemmatization, morphological analysis, transliteration, term extraction, and cross-language information retrieval (CLIR). Language technology on the Web should provide extensive coverage of both modern and ancient languages and demands standard interfaces and middleware, as well as mechanisms for resource discovery. For formal language content, manipulable interfaces and visualization technologies are needed, which will allow such actions as: plugging numeric values into a formula; translating between variant notational systems; graphing a function; or displaying a three-dimensional, rotatable model of an organic molecule. For image data, enhancement tools are needed (not just the general purpose filters available in software such as Photoshop or the GIMP, but algorithms tuned to specific purposes, such as enhancing illegible or barely legible manuscripts) as well as powerful search engines for images (e.g. word spotting or Supervised Multiclass Labeling (SML)).

New and more sophisticated means for representing quantitative information need to be fostered, and online visualization toolkits would provide an ideal way of disseminating better techniques. Scholarly articles are often full of what Edwin Tufte calls “chartjunk.” Sometimes scholars present data in tabular format, simply because they are unfamiliar with the software to present it in a more “readable” way. Few scholars are familiar with such techniques as *sparklines* (“intense, simple, word-sized graphics”).² Few use mapped pictures, despite the fact that they constitute a powerful way of combining images with textual and/or quantitative data. As researchers begin systematically to explore new large-scale topics, such as the comparative study of globalization processes in history, they will need new representations for such phenomena as layered time developments within a geospatial context.

Finally, we believe it is time to create 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 quantity of information. But really they are only incremental improvements on content models that have been used for more than a millennium. This is not to disparage the models; and certainly not to deny the impressiveness and usefulness of their Web-based reincarnations. But it is symptomatic of stale thinking that one can find such a limited number of federating models on the Web. New ways of organizing knowledge are needed, and the Epistemic Web is where they will flourish.

2 Edwin Tufte, *Beautiful Evidence*, Graphics Press (2006), pp. 7–19.