Research Paper

Access, claims and quality on the internet –Future challenges

Kim H. VELTMAN
European University of Culture

ABSTRACT

The vision of access to human knowledge has existed explicitly at least since the time of Aristotle In 1934, Otlet outlined a vision of comprehensive access to knowledge. Progress towards this vision entailed initial visions of hypertext, markup languages, the semantic web, Wikipedia and more recently a series of developments with respect to Open Source. A brief survey of these developments is provided.

The rhetoric of the Internet insists that everything should be accessible by everyone at anytime. This poses obvious technical challenges and serious philosophical problems of method. If everything is accessible then how do we separate the chaff from the grain and how do we identify quality? Following a survey of important developments, this essay suggests five dimensions that need to be included in a future web: 1) variants and multiple claims; 2) levels of certainty in making a claim; 3) levels of authority in defending a claim; 4) levels of significance in assessing a claim; 5) levels of thoroughness in dealing with a claim.

KEYWORDS

Internet access, quality, distributed repositories, networks

1 Introduction

The vision of access to the whole of human knowledge is probably as old as mankind. It inspired Aristotle, the *Natural History* of Pliny, the *Summas* of Thomas Aquinas, and the *Encyclopédie* of Diderot and d'Alembert. The 19th and early 20th centuries saw a quest to gain universal access to both primary and secondary literature. In 1934, this vision inspired Paul Otlet to outline the idea of comprehensive access to human knowledge

"a technology will be created acting at a distance and combining radio, X-rays, cinema and microscopic photography. Everything in the universe, and everything of man, would be registered at a distance as it was produced. In this way a moving image of the world will be established, a true mirror of his memory. From a distance, everyone will be able to read text, enlarged and limited to the desired subject, projected on an individual screen. In this way, everyone from his armchair will be able to contemplate creation, as a whole or in certain of its parts." [1]

By 1943, Otlet had sketched how such a machine to imagine the world might look (*machine à penser le monde*). In 1945, Vannevar Bush published a similar idea. In 1948, Claude Shannon, published a theory of information, which became one of the cornerstones of the Internet. In practical terms, the Internet began in the United Kingdom in 1968 and served as a basis for the U.S. Internet, which began in 1969. In the course of two decades the Internet became a tool for c. 100,000 academic users. The innovations of Tim Berners-Lee and Robert Cailliau (CERN) transformed the Internet into a World Wide Web (WWW). By March 2005, Google provided access to 8,058,044,651 web pages and to 1,187,630,000 images. [2] By the end of 2005, the Internet is predicted to reach 1 billion fixed line users.

Received July 6, 2005 ; Revised September 6, 2005 ; Accepted September 13, 2005 .
k.veltman@mmi.unimaas.nl
DOI : 10.2201./Niipi.2005.2.3

Problems remain qua finding meaningful hits, knowing whether they are reliable, creating new tools to make this possible and frameworks for searching and filtering that save us from a state of sheer chaos. The good news is that the past decades have brought many initiatives which point to new solutions. Hypertext, the semantic web, Wikipedia and Open Source have brought many positive steps forward. This paper surveys these developments and outlines some the challenges that lie ahead.

2 Hypertext developments

One impulse in this direction has come from the computer science community. The article by Vannevar Bush was a direct inspiration for Douglas Engelbart's ideas about collaborative and augmented knowledge. This inspired Ted Nelson to coin the term hypertext, which then made its way into the scholarly community in the 1980s. [3] Slowly the computer science community extended its interests beyond linking to the idea of annotating [4] and a vision of self-annotation. [5]

Much less publicized has been another impulse, which came from the heart of the scholarly community. In the years 1942-1946, Father Roberto Busa, while preparing his doctorate at the Pontificia Università Gregoriana (Rome), conceived the idea of linguistic analysis of Thomas Aquinas using computers. In 1949, the young Jesuit, approached the president of IBM about an electronic concordance to the collected works of Thomas Aquinas. In the next two decades he produced an *Index Thomisticus* of 10 million words, which contained every word and every expression of Aquinas, which amounted to 70,000 pages in 52 volumes in published form, and was made available on a single 12" disc. [6] In 1992, Father Busa went on to found the School of Lexicography and Hermeneutics (Scuola di Lessicografia ed Ermeneutica) at the Pontificia Università Gregoriana in Rome. [7]

In the 1960s and 1970s, computers began to have a serious impact in the study of the classics. [8] In 1972, Marianne McDonald set out to transform the *Thesaurus Linguae Graecae* (*TLG*) into an electronic data bank of ancient Greek Literature. The project was publicly released in 1982, was then made available in CD-ROM format (1985) and in April 2001, it "became available online to subscribing institutions and individuals. The web version currently provides access to 3,700 authors and 12,000 works, approximately 91 million words" (i.e. virtually every surviving ancient Greek text from 800 B.C. to 600 A.D). [9]

During the 1980s, Standard Generalized Markup Language (SGML) was applied to a number of other major projects such as the *Dictionary of Old English* (DOE), [10] and the *Records of Early English Drama* (REED). [11] Yuri Rubinsky, [12] one of the pioneers involved in these projects, founded SoftQuad, [13] partly to explore implications of these developments for publishing. By 1994, his ideas helped inspire new approaches to metadata that led to the Dublin Core initiative and subsequently to the vision of a semantic web. The work on SGML also inspired a group of scholars to found the Text Encoding Initiative (TEI). [14] A quest for simpler versions inspired the evolution of eXtensible Markup Language (XML) and TEI -Lite. As a result SGML, XML and variants are now used in a wide number of academic projects. [15] In Europe, one of the most famous of these is the *Thesaurus Linguae Latinae* [16] (*TLL*, founded in 1894 by Theodore Mommsen, the published version of which covers three feet of shelf space). [17] Other important projects are found around the world, such as a complete version of the Buddhist Pali Canon [18] in Korea, or the Emperor's library [19] and all the Classics (800 million characters in Unicode) [20] in China.

2.1 Freely available online

Some organisations, and individuals, often with funding from national and other bodies, have managed to make their resources available on-line without cost or by means of a minimal use fee. Significant examples are the Oxford Text Archive with 2,500 texts online; [21] the Marburg Archive with over 1.5 million photographs; [22] the Perseus Digital Library (Gregory Crane, Tufts); [23] 900,000 pages of the Max Planck-Institut für Europäische Rechtsgeschichte (Frankfurt, Manfred Thaller); [24] 130,000 very high resolution pages of *Codices Electronici Ecclesiae Coloniensis* (CEEC, Manfred Thaller); [25] the *New Media Encyclopedia* (Christine van Assche, Centre Pompidou) [26] and Netzspannung (Monika Fleischmann, Fraunhofer). [27] In Germany, the Prometheus [28] project, which provides online access to distributed slide collections for art history, entails a personal subscription fee of €20 annually.

Meanwhile thousands of reference works are becoming available online in a haphazard manner. There have been some attempts at aggregation. For instance, One Look Dictionary Search, [29] provides access to over 100 dictionaries including the *Merriam Webster English Dictionary*. There is a genuine need to make such resources more systematically available in the form of a virtual reference room and to link these resources to emerging digital libraries and virtual agoras for collaborative research and creativity. Eventually such virtual reference rooms will have various levels: namely, a section for reference materials

which are freely available, and other sections which are available via different subscription options.

2.2 Available online via intranets

In the sciences and particularly in physics, astronomy and chemistry, networks in the form of intranets assure a ready exchange of research (cf. § 2.3.2. below). In the realms of culture, humanities, and social sciences an enormous amount of research is still inaccessible beyond the intranets of the institution where they are being produced. The well known problem of the last mile, which has often been reduced to a challenge of the last 500 meters or even the last building, remains a stumbling block. A second obstacle remains the practical challenge of interoperability in a practical sense. A third, more insidious barrier is psychological, whereby institutions with terabytes of information in their databases are worried that their materials will be misused or stolen. A major challenge of the next decades lies in ensuring that these important results of research are shared more widely at least within the scholarly community.

2.3 Available online commercially

Meanwhile, there has been a trend for the results of scholarly research in terms of reference works, journals and books to come increasingly into the realm of commercial interests.

2.3.1 Reference

Traditionally scholarly research often led to reference works, which benefited the scholarly community as part of the public good. Such reference works are now increasingly being acquired by commercial companies and made available by subscriptions with an eye on profit. A five year individual subscription to the *Thesaurus Linguae Graecae* (*TLG*) costs $400. The *Oxford English Dictionary*, [30] is available online with an annual subscription for individuals at £195+VAT. [31] The *Allgemeine Künstler Lexikon* of Thieme Becker, a seminal reference work for art history, costs €298.90 annually. Hence, while the good news is that standard reference works are now available online, the bad news remains that they come at costs that are prohibitive for persons wishing to have the equivalent of reference rooms in an online context.

Increasingly such reference materials are being bundled by publishers and private companies. Here, one of the best known examples is Dialog (now owned by Thomson), which offers access to a wide range of databases including "coverage of scientific and technical research reports and publications from more than 150,000 journals. Abstracts to 1.2 million dissertations and more than 2 million conference papers. Pharmaceutical drug pipeline data from conception to launch." [32]

2.3.2 Journals

During the Renaissance, the world of learning was literally a world of letters whereby scholars shared ideas via scholarly correspondence. This world of letters gradually became transformed into scholarly journals. By the mid-20th century it was generally assumed that only a few great libraries would be able to collect the entire range of scholarly journals. The past 50 years have seen both a great diversification of titles and an enormous consolidation, whereby a very small number of major publishers now dominate the field. Rhetorically this was for reasons of efficiency. In practice, prices have continued to rise so dramatically that today not a single library can afford to collect all the journals that exist. The for profit attitude that was supposed to enable more effective publication is crippling journals as a medium for scholarly communication. As the Association for Research Libraries has noted, in the period 1986-2001, "The typical library spent 3 times as much but purchased 5% fewer titles." [33]

The scientific community has begun to take steps towards a new approach. In the 1970s and 1980s Douglas Engelbart and Bruce Schatz envisioned new approaches. By 1990, Stevan Harnad (Princeton University and Universite d'Aix Marseille II) had outlined a potential of making the preprint process accessible electronically. [34] In 1996, Paul Ginsparg, Los Alamos, gave a lecture on electronic publishing in science at UNESCO. In 2002, again at UNESCO, he outlined his vision of Creating a Global Knowledge Network. [35] Here he traced the history of an e-print arXiv (where "e-print" denotes self-archiving by the author), which began in 1991. As of June 2005, this has some 4,000 new submissions monthly and includes 323,889 preprints. [36] "arXiv is an e-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology." [37] Los Alamos has also been experimenting with distributed, open source, search and access methods. [38]

By 2003, the German Max Planck Gesellschaft, made a more dramatic announcement: that they would make freely available all the pre-prints of research results from their 83 institutes. [39] In 2003, the Max Planck group also organized an important conference on *Open Access to Knowledge in the Sciences and the Humanities* that led to a Berlin Declaration:

"to promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research pol-

icy makers, research institutions, funding agencies, libraries, archives and museums need to consider." [40]

Max Planck is also exploring how this online approach to science might be expanded into a European Cultural Heritage Online (ECHO), [41] which it sees as an Open Access Infrastructure for a Future Web of Culture and Science. At the 19th International Codata Conference, Adama Sammasekou gave a keynote on Open access for All. A Required Step towards a Society of Shared Knowledge. [42] Directly and indirectly such visions and efforts are inspiring other initiatives to make scholarly content readily accessible. The Scholarly Publishing and Academic Resources Coalition (SPARC) [43] and the Public Library of Science (PLOS) [44] are two examples. In the Netherlands, SURF is sponsoring the Digital Academic Repositories (DARE) project to make Dutch research available online. [45] This also entails six related projects including P-Web : a tool for online publication of proceedings' (at the Erasmus Universiteit, Rotterdam). [46] In the U. S., the National Institutes of Health (NIH) have made open access part of their policy. [47] In the United Kingdom, there are plans for open access to results from work supported by the Research Councils UK (RCUK). [48] The Association of Research Libraries (ARL) has a useful site on the issue of open access. [49]

2.3.3 Dissertations and books

Traditionally knowledge of dissertations was spread via dissertation abstracts and outstanding dissertations were published in a simple format by University Microfilms (1938), and subsequently University Microfilms International (UMI). In 1985, Bell and Howell acquired University Microfilms and the company was renamed ProQuest. In the United Kingdom, Chadwyck Healey, founded in 1973, amassed a number of standard reference works in electronic form including the 221 volumes of Migne's Patrologia Latinae. [50] In 1999, ProQuest acquired Chadwyck Healey. By 2000, ProQuest's "Dissertation Abstracts database archived over 1.6 million dissertations and master's theses. Some one million of them are available in full text in print, microform, and digital format." [51]

The trend towards commercialization, which began with reference works and journals, has by now spread to the whole of scholarly production. Small scholarly presses and even university presses are increasingly being incorporated into a handful of large multinational media companies such as Elsevier. In addition to its dissertations, ProQuest's collection now has 5.5 billion page images and adds "37 million images of contemporary information" [52] annually. The jewel in this collection is the Early English Books Online (EEBO), which "contains about 100,000 of over 125,000 titles listed in Pollard & Redgrave's *Short-Title Catalogue* (*1475–1640*) and Wing's *Short–Title Catalogue* (*1641–1700*) and their revised editions, as well as the *Thomason Tracts* (*1640–1661*) collection and the Early *English Books Tract Supplement*" [53] Early English Books Online is operated by ProQuest in conjunction with the Text Creation Partnership for which membership costs range from $15,000 for a small undergraduate institution to $60,000 for an Advanced Research Library (ARL). [54] This subscription merely assures participation. Access to copies of individual texts costs members $6 per text, Meanwhile, ProQuest continues to acquire other companies. On 1 March, 2005, ProQuest acquired Explore Learning, "producers of the world's largest online simulation library for math and science education." [55]

These developments are significant for a number of reasons. First, at a very simple level they mean that reference works, which are essential for research, have become a very profitable business for some. In 2004, ProQuest had a gross profit of $232.5 million. [56] An important corollary is that if one's institution is not a subscriber to the Text Creation Partnership and ProqQuest's various resources, a scholar is effectively deprived of the entire corpus of Early English printed books, and many of the key reference works, which the past two hundred years of scholarship have painstakingly created. Ultimately this convergence of reference tools, content and educational tools creates a digital divide throughout the developed world as well as so-called developing countries: between those with enough money for expensive subscriptions and those who fall outside this charmed circle.

In the past year, there has been a dramatic new player in this arena. Google's "mission is to organize the world's information, but much of that information isn't yet online. Google Print aims to get it there by putting book content where you can find it most easily – right in your Google search results." On 14 December 2004, Google announced that they were working with the "University of Michigan, Harvard University, Stanford University, The New York Public Library, and Oxford University to scan all or portions of their collections and make those texts searchable on Google." [57] These plans which entail over 16 million texts in full text will require at least ten years to be achieved. [58] A spectre of paid access to the Internet looms. [59] As will be noted below (§ 5) this has inspired dramatic reactions in the first half of 2005.

2.4 Commercial infrastructure

These trends towards convergence are the more disturbing because they are becoming ever more linked with infrastructure developments. In the United States, the plans of the Next Generation Internet (NGI) Initiative, [60] Internet 2 [61] and the National Light Rail [62] entail a relatively small number of institutions, which are increasingly intent on acquiring ownership or at least control over network infrastructures as well as the contents used for higher education. The quest for a Next Generation Internet, for Dublin Core Metadata, for Digital Object Identifiers (DOIs); the IEEE's quest to create a Learning Object Model (LOM); [63] the Army's efforts at a Sharable Courseware Object Reference Model (SCORM) [64] are all related and in the eyes of some are part of a single coherent vision. [65]

Critical observers such as John Perry Barlow (Electronic Frontier Foundation), [66] Lawrence Lessig (Creative Commons), [67] Clifford Lynch (Coalition of Networked Information) have warned that the spectre of trying to control the whole of education and all access of knowledge goes much deeper: e.g. plans to create books that self-destruct after a few readings; to forbid reviews and even to forbid reading books out loud. [68] The same technologies that could provide us with universal access could be used to limit our access more than ever before. Learning, which was once a challenge of ability, is increasingly becoming limited to those who are wealthy enough to afford it, in a world where an ever smaller number can afford more, and the overwhelming majority can afford ever less.

These dangers go far beyond inconveniences. Michael Giesecke (1994), in his standard book on the history of printing, noted that Gutenberg's real contribution lay not in the technology but rather in a decision to use printing for the common good. [69] Giesecke suggested that it was ultimately this attitude towards sharing knowledge that was a key to the modern world as we know it today. Independently, Jean Luc Guédon (2001) [70] made a related claim when he noted that the breakthroughs of early modern science resulted from a spirit of sharing knowledge through learned societies and academies and that trends towards commercialization of knowledge now threaten the advancement of learning. Others such as Philippe Quéau have gone further still to insist on knowledge as a Public Good and speak of a global common good (*le bien commun mondial*). [71] To neglect that public good endangers progress and indeed the very survival of civilization.

2.5 Government initiatives

Governments have begun to recognize these dangers and have begun to take action. The most conspicuous example thus far has been the United Kingdom, where the JISC (Joint Information Systems Committee), has been pioneering in arranging for collective licences for its member institutions to projects such as the Early English Books Online (EEBO) mentioned earlier. This means that at least those in universities and a number of Higher Education (HE) institutions again have normal access to basic reference works and content. It is important to recognize, however, that this important step does not solve the problem. The problem of access by the majority of citizens who are not connected to these university intranet networks remains.

In addition, the JISC is sponsoring some 135 current projects including: Building a Virtual Research Environment for the Humanities (BVREH); [72] Collaborative Stereoscopic Access Grid Environment (CSAGE); Digital Libraries for Global Distributed Innovative Design (DIDET) and a Virtual Research Environment for the History of Political Discourse 1500–1800. These are part of a larger vision in the direction of grids and an e-Science [73] strategy, whereby the UK hopes to provide a model for next generation information access for Europe and beyond.

By comparison, in France, the Maisons des Sciences de l'Homme (MSH) [74] are making contributions in bridging fields such as archaeology, anthropology, ethnography and social sciences, but on a much more limited scale. Meanwhile, former President Mitterand's plans for the new Bibliothèque Nationale de France (BNF, 1988–1994) [75] included a vision of access to full-text contents. As a result the BNF's Gallica project has made 76,000 books and 80,000 images available online. [76] Google's announcement in December 2004 radically altered the dimensions of this vision.

In early March 2005, the Director of the BNF urged "European governments to join forces and set up a digitization plan that would be a European response to Google Print." [77] By 17 March, 2005 President Chirac had given the go-ahead for a French project. [78] By 22 April, 2005, 19 National Libraries had signed an agreement that they were willing in principle to work together : [79]

"On 28 April 2005 6 EU countries sent an open letter to the European Commission and the Luxembourg Presidency of the Council asking for a European digital library. Inspired by the French president Jacques Chirac, the presidents or prime ministers of Poland, Germany, Italy, Spain and Hungary have signed the letter. On 3 May 2005 the European Commission responded with an announcement that it will boost its policy of preserving and exploiting Europe's written

and audiovisual heritage. The Commission plans to issue a communication by July outlining the stakes involved and identifying the obstacles to using written and audiovisual archives in the European Union. The communication will be accompanied by a proposal for a Recommendation aimed at enlisting all the public players concerned and facilitating public-private partnerships in the task of digitising the European heritage." [80]

As preliminary steps the efforts of the G7 pilot project, *Bibliotheca Universalis*, the Gateway to European National Libraries (GABRIEL) and the EC project The European Library (TEL) [81] are being co-ordinated within the BNF. EU Commissioner's Viviane Reding's i2010 vision is supporting these trends. [82] The vision of a European Digital Library has now become one of three flagship projects for the next five years. [83]

The plans thus far foresee a massive project that entails scanning four billion pages of text. The great open question remains whether these materials collected with the aid of public monies will be made available freely for the use of all citizens. If they are made readily accessible then the way will be open for a notebook for mankind on a scale that dwarfs previous efforts towards the semantic web, towards a Wikipedia and towards Open Source. Even so it will be fruitful to survey briefly these existing initiatives and outline their limitations before considering requirements for new authoring and search tools that overcome these limitations.

3 Semantic web

As noted earlier the WWW has made wonderful contributions in the domain of sharing knowledge. Thanks to their markup languages and protocols over 8 billion pages are now accessible online. If their quest for a semantic web were committed to semantics in the sense of "meaning," they could theoretically also address a) challenges of separating significant grains amidst the chaff of loose comments and b) elusive problems of quality. The W3 Consortium has, however, set itself a narrower goal. It is concerned primarily with machine-machine communication. As a result, it is focussed specifically on logical statements which can be verified within the binary logic of machines as either/or true/false. This is of enormous value in business, where the validity of orders, accounts and transactions needs to verified and certified. Hence, the goal of the W3 might more accurately be described as a quest for a transaction web. [84]

While both very useful and profitable, the present goals of the WWW focus only on meaning insomuch as it entails logical claims which are not open to ambiguity. Hence their quest does not address directly the needs of scholarship, where multiple meanings and ambiguities play a central role in interpretation and hermeneutics.

This is not ultimately a limitation of technology, but a conscious decision of the shapers of the technology to limit its applications. [85] Underlying this decision are deeper problems that face the computer science community as a whole. There is a fundamental assumption that the quest must be to reduce all meaning to operations which can be dealt with by machines in the absence of humans rather than a quest to use machines to record and communicate the complexities of meanings that have been developed and used by humans.

A handful of computer scientists have acknowledged this problem. Hence, Joseph Weizenbaum warned of the dangers of believing that machine-machine communication could replace human decision making, [86] as did Grant Fjermedal, [87] and Fred Brooks spoke of a need for Intelligence Augmentation rather than Artificial Intelligence (IA not AI), but for the most part this approach has been ignored. Machines and software which can be used to extend the range of man's meanings risk becoming limited systems caught in the tautologies of their own logic systems.

Indirectly, the quest of the W3 to create frameworks that verify and authenticate users, confirming that they are who they say they are, can be as useful in the world of scholarship as in the world of business. We need, at times, to be certain that the person who sends a claim, is indeed identical with the originator of the claim, or else be informed whether and how the message has remained intact in going via intermediaries. Ultimately, linking and hyper-linking can only be truly fruitful if they can bring us back, if necessary, to the original sources of content and claims.

Some members of the semantic web community, or more precisely, some communities concerned with the semantic web, are indeed concerned with meaning in a broader sense. For instance, those concerned with digital libraries are interested in creating standardized and interoperable thesauri. This is very important. Without clarity with respect to definitions of words and terms, there can be no certainty that we are even speaking about the same topic. To arrive at a full range of meanings, however, we need to have access to the equivalents of etymological dictionaries such as Oxford, and Grimm and these need to linked such that we can compare seamlessly changing meanings across languages. We have classification systems, dictionaries and encyclopaedias. We need to link these such that we can go from a term to its definition and explanation. We need

virtual reference rooms that provide much more than access to individual texts: we need to provide a new network of links between/among terms in reference works.

4 Wikipedia

The contributors to the Wikipedia [88] have also made great contributions to content on the web. Their commitment selflessly to add content for the greater good recalls a time-honoured mediaeval tradition that contributed greatly to the transmission of existing knowledge and considerably to the introduction and growth of new knowledge.

Traditionally, encyclopaedias attempted to summarize the state of knowledge at the time: Pliny, Vincent of Beauvais, Saint Thomas Aquinas and the *Encyclopédi*e of Diderot and D'Alembert are notable examples. [89] The *Encyclopaedia Britannica* continued this tradition until its 1911 edition. Thereafter it abandoned the quest for completeness. In 1992, *Encyclopaedia Britannica* introduced a distinction between a general Micropaedia and a Macropaedia for more detailed knowledge. These terms were patented and by 1995 the editors of a *Free Internet Encyclopedia* were advised that they could not use the term. [90] Accordingly they changed their terms to have been changed to "MicroReference" and "MacroReference" respectively.

As long as encyclopaedias were committed to recording the state of the art, it could be assumed that they covered the major literature or at least indicated the major surveys and reviews in a given field. Pauly Wissowa's *Realencyclopädie der classischen Altertums-wissenschaft* is a wonderful example of both the value and the dilemmas of such a quest. The two original authors died before the first edition was finished in 1852. A second edition began in 1861 and 1866 but remained unfinished. A third edition began in 1890 but took until 1980 to produce 84 volumes plus indexes and by then was too expensive for most individual scholars. [91]

One of the fundamental problems with the *Wikipedia* at present, is that there is no way of knowing how thoroughly a given article covers the topic in question. Some topics provide bibliographies, some do not. Some topics acknowledge using the 1911 edition of the *Encyclopaedia Britannica*, others do not. Critical tools concerning variants, certainty, authority, and significance are lacking.

5 Open source and open content

In Europe, there has been increasing attention to the possibility of open software, [92] which is made freely available without direct cost to individual users. This vision is linked with an open vision of Intellectual property. [93] By contrast, in the United States, Open Source has been linked with the notion of free software but free in the sense of "freedom" more than free in the sense of "without payment". The influence of Richard Stallman has been seminal in this context.

Table 1 Examples of emerging open source solutions.

Domain	Open Source Solution
Office	Open Office [94]
Photoshop	Gimp [95]
Illustrator	Inkscape [96]
Maya	Blender [97]
Premiere	Jahshaka [98]
Map software	Worldwind [99]
Bibliographies	Sourceforge [100]

His GNU project officially began in 1984 although he has traced its roots back to 1971. [101] With the introduction of Linus Thorvalds' Linux in 1992, [102] the Open Source Software movement began to take on new dimensions. The past five years have seen three fundamental developments: a) a dramatic increase in the range of open source tools (Table 1). [103] For instance, the Framasoft list now includes 905 examples of free software; [104] b) a related shift to include open source content through projects such as the Creative Commons; [105] the Open Content [106] and the Open Archives Initiative (OAI) [107] and c) discussions re: the future of scholarly communication [108] and the rise of new models for scholarly communication. [109]

Open Source, which was once seen as an interesting peripheral phenomenon is increasingly being adopted for crucial functions such as government administration [110] and even key commercial software. In June 2005, for instance, Nokia and Apple announced that they would use open source for their new mobile web browser. [111] The concept of open source software is being extended to include open content, open theory [112] and even open design. [113] The profound advantage of Open Source is that it offers new bridges across previously closed, proprietary solutions and thus potentially ushers in a new levels of interoperability across applications and systems.

The Wikipedia and Open Source are making enormous contributions and reflect the latest contributions of the World Wide Web, and a vision that goes back to the first half of the 20th century. While magnificent, since the early days of Artificial Intelligence in 1950s and throughout the emerging vision of a semantic web, there has been an underlying assumption in the computer science community that everything is either true or false; that the either/or approach of simple

logic offers a sufficient model; and that ultimately the quest is simply to document true statements.

Truth is the ideal and the quest to achieve it must remain paramount. Nonetheless, the realities of physical world and especially of the human world are more complex. Even in the world of engineering the limits of truth are recognized through the notion of tolerances which may be in terms of millimeters, sometimes much less and frequently much more.

In order to extend the potential usefulness of the semantic web to the semantic meanings of humans we need more than the logical propositions of either/or statements. Pioneers in computer science have rightly pointed to the need for sense-making tools through machines, but we also need to use machines to provide us with access to the senses of meanings that have already been provided by humans. To this end, we consider five new kinds of challenges, which have hitherto not been practical: 1) methods for integrating variants; 2) levels of certainty in making a claim; 3) levels of authority in defending a claim; 4) levels of significance in assessing a claim and 5) levels of thoroughness in supporting claims re: extant knowledge in a field. All five of these are important ingredients in a quest for discerning quality.

In some fields of science, where the emphasis is only on the latest results and discoveries, the methods here proposed may readily seem like an unnecessary amount of baggage. In the humanities, or more accurately, the human sciences (*scienze umane*, les *sciences humaines*), which include social sciences, ethnology, anthropology and archaeology, the situation is much more complex.

First and foremost, the crucial insights are about views which may not be universally accepted and yet remain fundamentally important. The writings of Dante or Milton, which are effectively commentaries on the Bible in the form of the *Divine Comedy* (La *Commedia divina*) and *Paradise Lost*, are expressions with an intrinsic value which is independent of whether this be the "right" or "true" interpretation of the *Bible*. They are an essential part of the literary tradition of Europe, for reasons similar to why the *Tale of Gengi* is central to the literary traditions of Japan. A significant part of the richness of these texts lies not in the texts themselves, but in the enormous amount of further editions, texts, commentaries and other expression that they have generated over the ages.

This has two fundamental consequences. First, we cannot understand Dante's or any author's importance simply through reading that author. We need access to the editions, commentaries and the rest. Second, it follows that the latest edition is not always the best in the way we assume that the latest finding is always the best in the scientific world. There is a cumulative dimension to knowledge, especially in the human sciences. This helps to explain why even the supremely ambitious head of Google who wants to gain access to all knowledge acknowledged in a recent interview that he estimated it would take 300 years.

Hence, this paper suggests adapting critical instruments which the scholarly and especially the library community has developed; to make the cumulative dimensions of knowledge part of our research programmes. We are concerned with vast new areas of human knowledge that need to be included within the semantic web, if we wish to have something that is truly useful for scholars as opposed to simply a tool for transactions which are necessary for business.

The representatives of memory institutions, especially librarians rightly insist that they have been working in this direction for millennia. It is true that they have created invaluable tools for bibliographic control. Yet the vision of a library has traditionally been to record the books, documents and materials it possesses, rather than indicating to what extent their collection reflects what is known about a person, a field or a discipline. To take a simple example: the Library of Congress catalogues tell us how many editions of Shakespeare or Goethe they have, but nothing about the extent to which their collection is a comprehensive one. Similarly Google and search engines tell us how many hits, but give us not the slightest hint as to what percentage this represents of what there is to be known on that topic.

So one challenge lies in using the tools of bibli ographic control from memory institutions and especially the library world and applying them to visions of the semantic web. A more subtle challenge lies in creating new frameworks that attempt to map not just isolated collections, but rather the extent to which any given collection or any given claim represents the state of knowledge about that person, field or discipline.

The five elements considered in this paper do not solve in a single stroke a problem that will take centuries to resolve. Even so they represent ingredients for going from a mentality of simply describing what we have in our collections to frameworks whereby we can see to what extent these collections represent what is known in the field. In the human sciences this means access to more than the standard versions of names, and standard versions of the facts, It necessarily means including those claims which are almost certainly true, probably true, or uncertain. At present these materials exist in secondary literature in our collections, but our bibliographic tools are about finding titles rather than contents. In the long term, we need new kinds of bibliographic instruments that will provide access not just

to contents, but provide us with knowledge about the claims made in those contents.

6 Variants and attributions

In an ideal world, scholarship is limited to eternal truths. In everyday life, many items are straightforward questions of true or false. Obviously in citing another work, the name of the author, the title and the date need to be precisely correct, otherwise they are wrong and misleading. In many cases, however, the situation is not so straightforward. We need to incorporate variant names, associations, attributions and claims. In many cases this knowledge/information is already being recorded in databases. The innovation that is being discussed here is how this knowledge/information is integrated into our search, retrieval and other tools.

6.1 Names

The most obvious of these entails different spellings of a given name. For much of the 20th century there was a conviction that if one could establish a standard version this could serve as an authority file and be adopted by or simply imposed on others. Library systems have complex systems for Machine Readable Cataloging (MARC and now MARC 21), which duly reflect standard and variant names. Ironically the potentials of this information are often not exploited fully even by the libraries themselves. In terms of everyday users such systems are not available and many experts would argue that even if they were available they would be much too complex to be used by "the man on the street," the non-expert.

Gradually there has been a recognition that these alternative names are effectively access points to earlier documents which were unaware of the current accepted spelling. So there is a new challenge to create online authority files with all possible variants built in. These lists can be online and freely available to users.

Here a first step lies in making alternatives visible to the user. Libraries have provided a partial solution through see also references, but although their internal catalogues and databases record all the variants used by an institution, this material is typically not available to users. In search engines the situation is worse. For instance, Google sometimes offers a guess but typically does not deal with variants. Hence a user who types in Martianus Cappella gets 100 hits with no clue that this is a variant spelling of Martianus Capella which gets 17,900 hits. In the Google approach as it is today these are two separate searches. By integrating lists of variants into search engines, entering a variant name effectively becomes part of the same search as a search using the standard name. Doing so would not require the user to do something more complex, but the results would be much richer. [114]

A prototype of the SUMS system illustrates the possibilities. [115] A user can type in a variant name such as Viator, arrive at the acknowledged modern name, Jean Pélerin and sees the other known variants.

A second step in this development would be provide users with tools to add further variant names as additional alternatives to the accepted authority names. Users could do so on a simple proviso: that they provide at least one historical document that uses the variant in question. This variant and its source would then become a regular part of the system. In using this method, non-expert users would be spared the deliberations of which variant to use. The system provides it for them. The variants remain accessible at the database level. Hence, even if the user forgets the official spelling next time round, the variants bring him/her back to the currently accepted version. Persons working in specific fields can work with subsets of the master lists to ensure that their tools match the complexity required, while saving them from unnecessary complexity.

Traditionally the quest for authority files was the domain of a very small group of librarians, and terminology experts. Much of their effort lay in trying to impose a given version and spelling of a term and trying to eradicate other forms, versions and variants. This was partly a necessity imposed by the limitations of the print medium. With electronic media, a new challenge looms: using one accepted version and linking this with all known versions. Variants now become tools for access rather than threats to the norm, and finding new variants can in fact become a task to which all users can contribute.

6.2 Classifications and associations

Connected with this theme of variant names are the variant classifications, thesauri and associations provided by previous attempts at systematic organization. Some links between thesauri, classification systems and titles exist already. The challenge is to deal with these much more systematically. There are a number of efforts which point in this direction such as multilingual classifications in the medical field [116]and classifications in the legal field. [117] Ideally one would be able to move systematically between subsumptive, determinative and ordinal relations. [118] One can imagine a system that allows users to choose whether they wish to deal only with physical instances (particulars) or also include various kinds of metaphysical (universals). In a simplest case this would entail an option between seeing a general universal version of, say an elephant and seeing particular exam-

ples of elephants. In more complex cases there would be a possibility to distinguish among different levels of metaphysical existence-: e.g. belief, phantasy, play, scenario and fiction at different levels of subsumption, some of which may not have direct correspondences in the physical world. Needed eventually is a classification of knowledge whereby we can see which subsumptive classes have a direct physical- metaphysical link and which do not. Ultimately we need new tools that allow us to go from subsumptive to determinative and ordinal relations. 3-D visualizations such as that provided by Spectasia can help in providing overviews of such classes.

These variants are often linked with simple numbers: e.g. the 4 seasons, 7 days, 8 winds, 9 muses, 12 months etc. A system which allows us to move seamlessly from any one term to other terms in a given set would be extremely useful by way of orientation. Providing visual overviews of these associations one can use their basic spatial positions as entry points into thought systems which have different names and associations as one moves from one culture to another. The 3 cardinal virtues and 4 points of compass are simplest examples. The 10-12 signs of zodiac are one step more complex. The 44 constellations of the Northern Hemisphere are a more complex example. One can imagine an interface that uses fundamental images such as the world tree, which can be viewed as the Milky way in astronomy, a physical tree of life in botany and the spine in anatomy. By choosing levels in the microcosm- macrocosm analogies one could move through these levels by a simple pointing metaphor. From this point of departure various expressions in different cultures could then be visualised in alternation.

Some of these variant associations are not spatial. To take an example from religion which has been the source of most great cultural expressions until the past century: the Virgin Mary is universally known in the West. Some call her Star of the Sea. There are over seventy such alternative names, [119] many of which are potentially useful in increasing the range of our search. Such an approach becomes essential when we are searching in other cultures. For instance, the Indian, Mother Goddess, Durga, has 108 names. [120] Such lists are effectively like mini-specialized thesauri applied to a given deity, person, idea or concept. The unfamiliar associations of a name may seem strange and eccentric to us, but if treated systematically these again offer new entry points into knowledge. Instead of seeing them as aberrations from what we know we need to discover their potentials in helping us discover what we do not know. Else the legitimate quest for standards in computer risks becoming a closed community of the familiar and the known rather than a tool towards discovery of the unfamiliar and the unknown.

Genealogical lists are another example of such contextualizing instruments. Hereby, a single name provides access to a range of related persons. Needed are frameworks, whereby these existing lists are made available to us as we embark on more serious searches. Such lists are again like classification systems and thesauri. They need to be linked with definitions (dictionaries), explanations (encyclopaedias), and titles (catalogues, bibliographies) elsewhere. All this belongs to the domain of virtual reference rooms (levels 1-5 in Table 2).

Librarians, especially cataloguers, indexers and classification are, of course, fully aware of the existence of such lists. Like their colleagues in the realms of computer science and artificial intelligence there have been enormous disagreements and wranglings about which system is ontologically true. Like all quests for truth, this quest has its uses and the search for a new and better system should continue.

In the meantime, our pragmatic suggestion is that, if we leave aside debates about which system is best, and focus rather on links between existing systems, there is enormous insight to be gained using the associations of the past as a tool for searches in the future. Instead of aiming at machines that will replace the rich ambiguities of human associations with unequivocal commands, perhaps we should aim to create machines that

Table 2 Virtual Reference Room, Distributed Digital Libraries and Virtual Agoras with different levels of reference and secondary literature.

Virtual Reference Room	
1. Terms	Classifications, Thesauri
2. Definitions	Dictionaries
3. Explanations	Encyclopaedias
4. Titles	Bibliographies, Catalogues
5. Partial Contents	Abstracts, Reviews
Primary Literature in Digital Library	
6. Full Contents	
Secondary Literature in Digital Library	
7. Texts, Objects	Analyses, Interpretation
8. Comparisons	Comparative Studies, Parallels,
9. Interventions in Extant	Object Conservation
10. Studies of Non-Extant	Object Reconstructions
Future Secondary Literature (Virtual Agora)	
11. Collaborative Discussions of Contents, Texts, Comparisons, Interventions, Studies	
12. E-Preprints of Primary and Secondary Literature in Collaborative Contexts	

use these rich ambiguities of historical documents as a source for new access to our past and our present.

Present systems such as Google assume that we know the detailed words necessary for the search, and indeed if we happen to know these then Google works surprisingly well. The problem with real research is that we usually do not know the important terms when we embark on our study. Being able to call on existing associations of earlier experts offers a way to go further. Implicit here is a notion that interfaces are something much more than physical screens. They should include the mental screens of earlier and existing experts. Their ways of organizing knowledge can serve as orientation tools in our own voyages of discovery.

6.3 Attributions

In the exact sciences only the latest version of an attribution is usually important. By contrast, in the humanities the cumulative history of attributions is potentially important. The latest claim is not always the best. Even if the latest is the best it is frequently still not definitive. For instance, in the case of a painting, one scholar may claim a) that the painting is by Leonardo, another may claim b) that it is by his pupil while a third claims c) that it belongs to his workshop.

An either/or mentality from computer science which creates a single category for creator/author in the Dublin Core Framework provides space for only one of these claims and in the process obscures the truth that in this case there exist debates on the question of precise attribution for this painting. This almost banal example illustrates how an overzealous quest for precision can be as misleading as it is meant to be helpful. Needed are tools a) to distinguish between these various claims re: attributions and b) to aggregate automatically the cumulative claims of the research literature such that we can see at a glance, for instance, that 3 scholars believe a given painting to be by the master himself, 15 claim that it was done by pupils; 7 claim that it belongs to his workshop and 2 feel that it merely belongs to his school.

One reviewer has reasonably suggested that such a criticism of Dublin Core is excessive: that this is an implementation issue and that it "can be handled by incorporating some mechanism in using RDF, for example." Of course there are ways that the matter could be handled using new tools. But everyday humans will never want to write all their essays in RDF, nor do they have time to explain how to convey the subtleties of 4,000 years of commentaries in a form that is made for machines. A real challenge for the computer science community is to create solutions that reflect what users do rather than what computer scientists assume they need to do.

6.4 Alternative claims

This principle extends also to alternative claims and interpretations. One scholar may claim that Galileo invented the telescope; or that Newton was the founder of early modern science. Hence, our records concerning an individual should provide systematic access both to an author's writings (primary literature) and to the (secondary) literature about those writings. We should be able to trace such publications alphabetically, chronologically and geographically. Ideally we would in addition be able to trace such publications in terms of both their attributions and claims.

As the 19th century made an increasing distinction between primary and secondary literature, especially in fields such as theology, philosophy and (English) literature, the initial emphasis was to focus on studies of texts and objects in isolation (*das Ding an sich*). This led to new levels of analysis in the form of interpretation, hermeneutics close reading, criticism. In the course of the 20th century so much attention was given to these problems by the de-, re- and post-constructivist schools that the cumulative, contextual dimensions of knowledge often faded into the background. Meanwhile, three further levels of analysis came into focus, which amounted to new layers within the notion of secondary literature. (levels 8-10 in Table 2). A first of these (level 8) related to comparisons (Comparative Studies, Parallels, Similarities); A second of these (level 9) entailed interventions in Extant Objects (Conservation, Restorations. A third (level 10) entailed studies of non-extant objects (Reconstructions).

Needed are systems that allow us to distinguish between these different kinds of reference and secondary literature. At present library catalogues provide us with titles of books and classification systems provide access to the subjects and concepts, but most of these systems still reflect an approach to knowledge via disciplines.

As a result studies of the Parthenon as a building are classed under history of architecture; studies of the Parthenon's location are classed under geography ; studies of restorations are usually classed under conservation; studies of reconstructions as to how it once looked are classed under art history, architecture or history. But aside from titles which happen to contain the word Parthenon, there is nothing to help find everything known about the Parthenon. The approach here suggested overcomes that limitation.

With the rise of new collaborative environments, virtual agoras can serve as a drafting ground for future secondary literature (levels 11-12 in Table 2).

An emerging challenge lies in integrating new per-

sonal and collaborative knowledge with the frameworks of enduring knowledge of memory institutions while acknowledging that they are not identical. Digital libraries will thus entail much more than scanning in printed texts: they will have at least three closely coupled features: virtual reference rooms; distributed digital libraries of primary and secondary literature and virtual agoras for collaborative research and creativity. The different levels (1–12 in Table 2) can also be seen as a knowledge life cycle: i.e. reference works point to primary literature, which inspires secondary literature, which prompts collaborative discussions (virtual agoras including discussion groups, blogs, Really Simple Syndication (RSS) [121] etc.), which in turn lead to new primary and secondary literature.

This model implies that a collective notebook for mankind can effectively be an extension of the systems from collective memory institutions [122] (libraries, museums and archives), a way of adding commentaries on a cumulative body of knowledge. This means that the critical apparatus of authority files (standard names and variants) and official terminology established by these institutions can be linked directly with personal variant names and personal terms of users at different levels of professional activity. By implication, the frameworks and search mechanisms already in place for enduring knowledge in memory institutions can be extended to the realms of new personal and collaborative knowledge in the Internet. To be effective this approach needs to include levels of certainty in making a claim (§ 7), levels of authority in defending a claim (§ 8) levels of significance in assessing a claim (§ 9) and levels of thoroughness in supporting a claim (§ 10).

7 Levels of certainty in making a claim

Ever since the advent of hypertext with Douglas Engelbart and Ted Nelson the emphasis has been on linking. Like all important ideas this built on earlier traditions. Footnotes and references were also concerned with linking. Electronic hypertext links introduced two fundamental steps forward: a) the link was only a click away; b) that click could potentially lead to a source outside the document being used at the moment. This is important because, traditionally a footnote in a scholarly book might conscientiously cite another article, book or manuscript in some remote library, to receive a copy of which took weeks or even months.

With electronic hypertext such a source is potentially only a click away. Google has filed patents in this domain and aims "to develop technologies that factor in the amount of important coverage produced by a source, the amount of traffic it attracts, circulation statistics, staff size, breadth of coverage and number of global operations" and searching for methods to determine the truth value of claims." [123] It is important to recall that many aspects of this quest are already reflected in our memory institutions. Instead of spending billions in creating entirely new models it is advisable to invest in linking the new instruments with existing frameworks. Some classification systems have some means of dealing with certainty of attribution in their categories. [124] Once again the challenge lies in making more of the enormous critical apparatus which memory institutions already possess visible to users. Of course, not every user will want to use the 800+ fields of the most complex systems; but the option to use them in various combinations should be there.

7.1 Direct and indirect links

Not all links are equally effective. A link from a reference concerning *Mona Lisa* in the Louvre to any of the dozens of sites containing a poor replica of the painting is less effective than a direct link to the Louvre website. One might distinguish between a) materials that are shown live, b) that come from the original location, c) via an agency, or d) via an official publication. In future, the extent to which scholarly books and articles link directly to original sources rather than to vague sites can become a new criterion for the quality of scholarship.

7.2 Degree of identity

Today, when we type in a word or term, search engines such as Google assume that we are looking for something that is identical to that word. It may also offer materials that are similar to that word but there are no tools in place to define the parameters of a match. Hence, typing in *Last Supper* (on 15 April 2005) produced 16,200 hits but there are no functions in place to search for cases that are identical in size, shape or colour. Over two millennia ago, Aristotle discussed the importance of attributes in defining objects. Adding attributes to our search parameters will mean that we can find things with the same name and then find subsets which are the same size, shape, colour etc. Eventually this could be extended to include attributes entailing all five senses and thus be able to discover surfaces, which look the same but literally feel different.

7.3 Levels of certainty

Needed also are new tools that allow authors to indicate the level of certainty behind their claims. In the sciences, such certainty is often continuous, or at least numerical, such that we can speak of parameters or

tolerances within which a technique or process will function. In the human sciences, these levels of certainty are often not continuous or quantitatively measurable. Even to attempt claims such as Shakespeare was x% (e.g. 98%) a genius as an author, would be a category mistake.

Such levels of certainty can be built in to cover claims about who, what, where and when? (Appendix 1). The precision with which one covers claims, including the detail with which one indicates the extent to which certainty is possible then becomes a further criterion for defining scholarship.

For the moment we shall focus on the problem of degree of certainty with respect to the question How? For instance, we are studying a painting of a woman's face. We encounter a related image on the web that suggests the painting is in fact a portrait of Madame X. On one occasion we may find additional evidence which is conclusive. On other occasions the link to be made might be very certain, quite certain, very probably, quite probably or only possibly. Ideally an editing tool makes available a small popup list of choices ranging from Authoritative to Possibly (Appendix 1).

Of course memory institutions and especially libraries have been struggling which such problems of bibliographic control for centuries and the results of their painstaking efforts are included in their internal catalogues and databases. Significantly, however, such results are often not accessible in usual queries in the Internet or even in libraries, museums and archives themselves. Hence if a person searches for Leonardo da Vinci, some systems give only titles definitely by Leonardo; others include items attributed to, students of, school of, followers of, copiers of. However, no system today provides us with a systematic overview of how these categories relate to each other, let alone how the numbers in these categories have changed over time.

Comprehensive lists of artists are a first step. However, search and retrieval systems that find a name such as Leonardo da Vinci will not solve this problem. Search engines must have as part of their system not just variant names but also the names of students, members of the school, names of followers etc. An expert on Leonardo knows that Bernardino Luini, and Andrea Solario were important students. But search engines and non-expert users do not know this. So this knowledge needs to be built into our search engines.

This matter is complicated by the reality that it is not just a question of scanning in some list of students and followers. There is debate on precisely which paintings were done by Leonardo, which by his students; who his students were etc. and this debate changes over time.

Indeed, scholars have traditionally used a whole range of vocabulary to indicate a spectrum of certainty ranging from simple assertions via conditionals to subjunctives. Hence phrases range from: "as we all know"; "it is generally agreed"; to "it is likely that"; "it would be possible to conclude"; "it is not to be excluded that."

A real challenge thus lies in incorporating such features into our authoring tools such that authors can record the state of their certainty as they are doing their work in order that we shall someday be able to trace systematically not only facts but also differing levels of certainty concerning these.

The system that we envisage, would allow us to choose the level of commitment. Levels 3–6 would require us to commit our name to the claim and invite documentation. Level 1, a claim that something is authoritative, requires documentation. Sceptics will rightly object that such a system will never be universally accepted. Many persons will prefer simply to dump their unsubstantiated claims on the web. In the interests of freedom of the spirit persons must be free to do so and free to state whatever they wish or not. Failure to permit these options takes one down a path where what a person writes, speaks or even what a person thinks could be seen as a threat to decision makers and the state. Science fiction movies such as *Minority Report* have warned us of the dilemmas of seeking mind and thought control.

Our approach rejects such basic censorship as a dead end. At the same time, by including rules and frameworks for levels of certainty, we have new possibilities of introducing search parameters which can sometimes choose to ignore unsubstantiated claims. Five centuries of experience with printing have led to similar solutions. We allow sensationalist newspapers such as the *Daily Mirror*, or the *Bild Zeitung* to publish many amazing, undocumented claims, but when we are writing a scholarly piece we usually ignore them as evidence. In future, learning how to use sources critically and being required to us sources with a given level of certainty can become new domains for learning in schools and universities.

Authorities, decision makers and sceptics generally will fear that all this assumes honesty in the system and will remain worried about dangers of the system being subverted by dishonest imposters. Fortunately, the simple rules of the game have some built in safety mechanisms. Anyone is free to state something, but anyone who claims to provide levels of certainty must also provide the supporting evidence. Hence, those who wish to use the cover of anonymity are free to do so, but thereby eliminate themselves from the certainty process. Those who add a source must provide a link

to that source. If the link is false or does not confirm the claim the system can reject it. If they refer to themselves they implicate their own reputation. If they include their organization, then their organization implicitly becomes liable for defending the claim. For this reason, authoring new tools for levels of certainty in making a claim need to become linked with levels of authority in defending a claim.

8 Levels of authority in defending a claim

This could at first sound like overkill. On reflection, this approach simply formalizes an approach that has been in place informally for centuries. Whenever we meet someone we expect a business card to tell us their affiliation. If they come from a world famous university or company we implicitly give them more respect and trust than if they come from an unknown organization. The purpose of a more systematic approach is not to check all the details of each source at every turn, but rather to have in place a framework that permits us to check these sources if necessary or desired. Hence, scholarly authors wishing to document their claim, might be prompted to indicate the source for this claim that something is authoritative in a further list, i.e. whether it originates in: 1) a memory institution; 2) an organization, usually a professional body, or 3) an individual.

Hereby, searchers will in future be able to use these parameters in their search criteria. For instance, within a library one might be searching for everything under a given name or subject or limit the search to specific forms of documentation (Table 3). The complexity of these lists will depend on the situation at hand. Sometimes, a simple distinction between scholarly and popular press might suffice. At other times a more detailed set of distinctions will be appropriate.

From such examples, we begin to see how the modules and lists for inputting knowledge and the lists to search for knowledge can gradually converge. Again we see that while anyone can make links, only those links which take us back to their sources are truly helpful. The need to cite sources was recognized by Renaissance humanists, who called for a return *ad fontes*. But whereas the Renaissance quest was limited to pointing to sources beyond the manuscript or book at hand, the new media allow a direct link with such sources. Hence, proper use of electronic equivalents of such sources can improve our success in accessing true and meaningful knowledge and at the same time provide new criteria for judging the quality of humanists and scholars in future. [125]

Table 3 Examples of different kinds of documentation that one might wish to access.

Source: Library
1. Book
2. Article
Peer Reviewed Journal Journal
3. Magazine
4. Newspaper
5. Television
Documentary Interview

9 Levels of significance in assessing a claim

History has taught us that significance is one of the most elusive characteristics to assess. Confucius had less than 30 followers when he died yet his ideas have profoundly affected more than two and a half millennia. Boethius spent the last year of life in jail before being beheaded on account of a false accusation and yet his *Consolation of Philosophy*, written in his prison cell became the most widely read book in the West, second only to the *Bible* for nearly a millennium. Milton wrote the most famous written defence of freedom in the English language, the *Aeropagitica*, while he was jailed for belonging to the wrong political party.

Some assure us that given the phrase "publish or perish," quantity of publications is the prime criterion for significance. Here caution is advised. Andrew Lang (1844−1912) was undoubtedly a significant writer. The *Wikipedia* records more than 140 books that he published. [126] Of Lao Tse only 81 paragraphs are extant. Yet many would rightly insist that the those 81 paragraphs in a slender book called the *Tao te Ching* that inspired Taoism had considerably greater significance than the writings of one of the most productive scholars and journalists of 19th century Britain. Meanwhile, peer review, citation indexes and the emerging field of automated citation indexes also termed dynamic contextualization offer further ways of assessing significance. These tools must be truly international and multilingual. Judging a European scholar in terms of how often they are cited in American publications can lead to distortions.

9.1 Peer review

In the 19th and early 20th centuries, when basic fields of study such as physics and chemistry had one standard journal the question of publication was fairly straightforward. Major scientists in the domain published their work in the standard journals, those at other levels published elsewhere. The enormous prolif-

eration of disciplines, and specialized applications means that no one continues to have a clear view of all that is written, even in fairly "narrow" disciplines. This has led to insistence on the importance of peer-reviewed journals and arguments that learned societies should have a greater role in the peer review process. [127]

Paul Ginsparg, (Los Alamos now Cornell University), has argued for a two-tiered approach whereby articles more articles are accepted almost automatically in the short term. The full peer review process is then applied to a considerably smaller subset in the longer term. [128] Here, once again the science community is suggesting new models that could potentially be used by the entire scholarly community. In terms of our model, the first tier would make personal and collaborative knowledge available at the level of e-preprints (level 12 in Table 2) and the second tier would act as filter in deciding what subset of this flux enters into the category of enduring knowledge (levels 1–10).

9.2 Citation indexes

In the 1970s, Derek de Solla Price developed the fields of bibliometrics and scientometrics, to address the problem of significance. [129] Over the past decades these fields have blossomed into a fashion for Citation Indexes. Such indexes are undoubtedly useful. Some would have us believe that they offer a chief criterion for judging scholarship, quietly overlooking that these indexes published in the United States focus mainly on Anglo Saxon publications. Others suggest that search engines such as Google are the new equivalents of, or even replacements for citation indexes. Like the citation indexes, the number of hits on Google is undoubtedly a useful indication. A simple example can quickly show why caution is needed with this approach. If we take ten leaders whose impact on the world (for better or worse) is universally recognized, we find that their ranking in Google is rather different than we might have expected (Table 4).

Table 4 Ten Political Leaders and their hits on Google (15.04.2005).

1. Charles V	29,600,000
2. George W. Bush	27,800,000
3. Alexander the Great	26,500,000
4. Hitler	8,560,000
5. Napoleon	8,410,000
6. Charlemagne	1,350,000
7. Mahatma Ghandi	1,120,000
8. Genghis Khan	374,000
9. Mao Tse Tung	361,000
10. Tamurlane	886

Taken literally this list would indicate that George W. Bush is 31,354 times more significant than Tamurlane, a dictator who once ruled over large parts of Asia, Russia and the Middle East and is said to be second only to Alexander the Great in terms of land ruled. Of course, to assume that the number of hits on Google alone constitutes serious proof in isolation would be simple-minded and ridiculous. But when we recall that careers of scholars are to some extent being determined by the number of times they are cited in citation indexes, the deceptive ways in which this kind of quantitative popularity contest is affecting our views of the world should give pause for concern.

Although it is not popular to say so, there are fashions in scholarship, just as there are fashions in clothes. Today, there is almost universal agreement that Rabelais was a great author of literature and that Leonardo da Vinci was a universal genius. It is sobering to note, however, that there were whole generations in the course of the past centuries when these figures were not at all appreciated. A generation after his death Leonardo's manuscripts were dispersed and many have never been found again. Almost than five centuries after his death we still have no complete works of Leonardo and books such as the *Da Vinci Code*, which are excellent novels that have virtually nothing to do with Leonardo, sell much better than the real thing. Such provocative examples serve simply to make a fundamental point that no simple criterion can solve the elusive question of significance.

9.2.1 Automatic citation indexes

A major breakthrough of the past few years is a trend whereby the process of citation indexes is becoming automated such that it can be integrated seamlessly into scholarly works and potentially reflect all citations rather than a sample as hitherto provided in American citation indexes. Michele Barbera and Nicolo D'Ercole (Pisa) and their team have developed *Hyperjournal*, [130] which includes [131] a Dynamic Contextualization, aP2P tool:

"which allows readers to visualize, while reading an article, all the articles quoted by and all those quoting the one they are reading. Dynamic Contextualization also enables you to easily carry out bibliometrical calculations such as: the number of quotations received by an article or by an author, citation source groupings by journal, by topic, by period."

If this approach were combined with our knowledge concerning kinds of journals (e.g. official journals in a field, journals published by key societies or Special Interest Groups (SIGs) of experts) and/or linked with standard collections of reviews, this could lead to new

insights concerning the influence of a given scholar. Meanwhile, this approach to dynamic contextualization is the more significant because Paolo d'Ilorio, the author of *Nietzsche Open Source* [132] and *Open Source Models in the Humanities. From Hyper Nietzsche to Hyper Learning* (April 2004), [133] has integrated this into his project on H*yper Learning* (*Hypermedia Platform for Electronic Research and Learning*):

"The overall objective of the Hyper-Learning project is to create an advanced e-learning system for the Humanities that will develop and enhance critical thinking skills. Hyper-Learning consists of four integrated components: 1) Research on functional programming for complex interactive web sites; 2) Development of a distributed web platform; 3) Establishment of Virtual Collaborative Learning Communities based around 13 representative European authors and 4) Creation of an appropriate pedagogical and legal framework." [134]

Ultimately we need some combination of quantity of output, quantity of citations and preferably also an indication of the extent to which authors are cited by experts in their own fields. Some authors establish fields, some authors contribute to accepted fields and some distinguish themselves by demonstrating the boundaries of strictly defined fields are too narrow to address the larger questions of scholarship. We need tools that will help us to recognize the contributions of all three of these types and not just the narrow experts for which German has a precise term (*Fach Idioten*). We need to maintain access to both generalist and specialized knowledge; to ability to provide surveys as well as capacity to focus on details (*minutiae* and *quisquilia*). Efforts in hierarchical classification might be useful in this context. [135]

10 Levels of thoroughness in supporting a claim

The above cautionary examples concerning significance may seem more evasive than incisive, but their combined thrust is that no single method offers a magic solution. Implicitly this suggests that thoroughness is the only way we can hope to achieve a balanced view. While attractive in theory this poses deep philosophical problems and challenges.

When a world expert gives a brilliant speech, attentive members of the audience are able to judge the points made in the speech. It would take another world expert of equal standing to have some sense of how much the brilliant speech omitted. The problems of brilliant speeches are also the problems of scholarship, which all too often is viewed as a series of brilliant books and articles or as a catalogue of those areas which are known and settled. Knowledge is presented as if it were a map of land conquered. All too often, however, we have no equivalent of a world map for knowledge, we have no clues as to how much has been covered so far. Roadmaps, a buzzword from the political arena, have become a fashionable term within the knowledge landscape. Alas they typically show us a few (possible information) highways and provide little indication of everyday roads, streets, paths and trails.

Even so, we know from history that such knowledge maps have frequently proved essential in the advancement of science and knowledge. In the early 19th century, once there was a periodic table, once one understood the scope and limits of chemical compounds one could start a process of looking for them systematically and filling in the missing gaps. It took a century, even then there a few bits to add, but it worked because there was a clear outline of what was not yet known (a map of ignorance in the true sense), which helped to guide explorers of new knowledge.

In spite of all the billions of printed and online pages today, we have remarkably little by way of serious tools to map our ignorance, to provide us some indication of level of thoroughness in dealing with a claim. Intuitively we recognize the problem perfectly well. If someone wanted to make great claims about Leonardo or any author, our first advice would be that they must study what Leonardo wrote and painted. Bibliographies exist but an updated list of all drawings, paintings of Leonardo and his school, a *catalogue raisonnée* in the traditional sense, does not yet exist.

Hence, while making knowledge accessible is obviously important, laudable and vital, it needs to be complemented by new kinds of cartography that map both our knowledge and our ignorance; the territory covered and the areas left uncharted. In some cases this is asking too much. There are always frontiers where we have no idea where to find even the next step. But if we make maps of accomplishments and dead ends there will be more hope of finding live ends and especially live non-ends.

Even in the absence of a clear programme, a number of intuitive steps have already been made in this direction. [136] We have international bibliographies, cumulative indexes of books, reviews, dissertations [137] and many other sources. We have reference rooms in the great libraries of the world with hundreds of thousands of reference works. Needed are virtual reference rooms where systematic connections between these resources can be created.

11 New criteria for scholarship

There was a time when scholars were a minority who could read and write amidst a majority who were

illiterate. In a world where theoretically everyone can read and where rhetorically everyone is an author, new criteria are needed to identify scholarship, and new criteria are needed to judge its quality. Everyday rhetoric points to the importance of multimedia and yet Gregory Crane, the author of the Perseus Project, was denied tenure at Harvard on the grounds that his work did not amount to publication. Even today some tenure committees overlook electronic contributions even if they entail peer review and major publishing houses such as Oxford University Press. Ultimately some combination of printed and electronic publication is likely to remain important in future. In both cases quantity alone is not a sufficient criterion. With respect to elusive questions of quality we have suggested five further ingredients that will prove useful.

12 Conclusions

The vision of access to the whole of knowledge goes back at least to Aristotle. The 19th century transformed this vision of individual thinkers into a more programmatic quest. By 1934, Paul Otlet (Mundanaeum, Brussels now Mons) had a vision of electronic access to the whole of knowledge. The past seventy years have been paradoxical. On the positive side technology has advanced greatly. Hundreds of thousands of books have already been scanned in. Soon there will be many millions of full-text books and other objects. At the same time there has been such an explosion of new knowledge and information that the possibility of systematic treatment seems more elusive today than a century ago. This is partly because scholars have focussed so much on studying texts and objects in isolation that the larger context and cumulative dimensions of knowledge have faded into the background. The growing commercialization of reference works, scholarly journals and even scholarly dissertations and books has cast a further shadow over the vision of free access to knowledge and information. This has led to a situation whereby even those with deep financial pockets cannot afford to see the big picture.

Meanwhile, the open source movement, impulses from science, and more recently initiatives from governments have re-introduced the feasibility of universal access to human knowledge. The quest for (distributed) digital libraries needs to be complemented by virtual reference rooms and virtual agoras. [138] Hereby, the ideal of a collective notebook can become an extension of existing systems for cataloguing and searching the cumulative knowledge of collective memory institutions.

The quest for full freedom of expression and open access in terms of quantity need not exclude the existence of criteria that highlight the central importance of quality. To this end, we have suggested that five new features that need to be added to such systems: 1) variants and multiple claims; 2) levels of certainty in making a claim; 3) levels of authority in defending a claim; 4) levels of significance in assessing a claim; 5) levels of thoroughness in supporting a claim. If these dimensions are integrated into an open source model there is reason for optimism about the potentials of the emerging technologies. The vision of open source knowledge on a fully semantic web may well take at least another century to achieve, but this only confirms that the goal is a noble one. In this context, if patience is a virtue, endurance and energy are a necessity.

Acknowledgements

I am very grateful to my colleague Professor Frederic Andres for kindly reading the text and providing suggestions for further references. I am grateful also to the reviewers whose positive comments and concrete suggestions have improved this paper.

An abridged 8 page version of this paper focussing on five new criteria was published in *Open Culture: accessing and sharing Knowledge. Scholarly production and education in the digital age*, Milan: Università Statale, June 2005.

The websites cited in this paper were checked anew on 28 August 2005.

References

[1] P. Otlet, *Monde: essaie d'universalisme -- connaissance du monde; sentiment du monde; action organis ée et plan du monde,* Brussels, du Mundaneum, Ed., 1935: http://www.laetusinpraesens.org/docs/otlethyp.php

[2] From 1990 − 1999 the WWW grew from 100 thousand to 200 million users. In spite of the dot.com bust, since 2000, the WWW has grown to 888,681,131 users of the fixed Internet (as of Mar. 24, 2005). Earlier concerns about finding enough hits have faded into the background.

[3] For an introduction see: W. McCarty, "A serious beginner's guide to hypertext research": http://www.kcl.ac.uk/humanities/cch/wlm/essays/diy/hyperbib.html; Willard McCarty, Home Page: http://www.kcl.ac.uk/humanities/cch/wlm/. For a survey of the state of the art in bibliographic control, see: A. G. Taylor, "The Organization of Information," Westport, CT, Libraries Unlimited, Libraries and Information Sciences Text Series, 2003.

[4] Cf. the WWW's Annotea project.

[5] P. Cimiano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," *Proc. of the Thirteenth World Wide Web Conf.* (WWW 2004), N.Y., pp. 462−471, May 17−22, 2004. This paper proposes

PANKOW Pattern-based Annotation through Knowledge on the Web), a method which employs an unsupervised, pattern-based approach to categorize instances with regard to an ontology.

[6] R. Busa, *Index Thomisticus: Sancti Thomae Aquinatis operum omnium indices et concordantiae*, Stuttgart-Bad Canstatt: Friedrich Fromann Verlag Günther Holzboog KG, 1974. *Thomae Aquinatis opera omnia, cum hypertextibus in CD-ROM*, Milano: Editoria Elettronica Editel, 1992. Cf. Thomism Today: Europe Video (transcript online below): http://www.innerexplorations.com/catchmeta/Thom.htm. J. Tomarchio, "Computer Linguistics and Philosophical Interpretation," Paideia: http://www.bu.edu/wcp/Papers/Meth/MethToma.htm.

[7] Corso Interuniversitario di Lessicologi aed Ermeneutica Tomistiche Computerizzate, Anno Accademico 2003–2004, Rome: http://www.unigre.it/pug/cael/cael.htm.

[8] Shane Houdek, *Classics and the Electronic Medium*, Department of English, University of Minnesota, English 3960, Junior-Senior Seminar: Electronic Text, Spring, 1996: http://mh.cla.umn.edu//houdek.html.

[9] Thesaurus Linguae Graecae: http://www.tlg.uci.edu/. Another of the pioneers in the TLG was Theodore Brunner (1913–1994).

[10] *The Dictionary of Old English*, Centre for Medieval Studies, University of Toronto : http://www.doe.utoronto.ca/.

[11] Records of Early English Drama: http://www.chass.utoronto.ca/~reed/reed.html.

[12] (Systémes) Grafnetix Systems Inc., "In Memory of Yuri Rubinsky": http://www.oasis-open.org/cover/yuriMemCSSC.html; J. Seybold, "Yuri Rubinsky" (1952–1996) : http://www.oasis-open.org/cover/seyboldYuri.html.

[13] SoftQuad (Toronto) was founded in 1984, sold to Corel (Ottawa) in 2001, which was bought by Vector Capital (San Francisco) in 2003.

[14] http://www.tei-c.org/. This now has the buzz-phrase "yesterday's information tomorrow." For an insight into early challenges see a report (1994) by the TEI's editor in chief, Michael Sperberg McQueen (1988–2000), "Trip Report Berkeley and Irvine, California, Mar.8–13 , 1994": http://www.tei-c.org/Vault/ED/edw44.txt.

[15] OASIS. Cover Pages, "Academic Applications, SGML/XML: Academic Applications. Contents": http://xml.coverpages.org/acadapps.html#tll. Pioneers in the field include Ian Lancashire, Homepage: http://www.chass.utoronto.ca/~ian/. L. Burnard, Homepage: http://users.ox.ac.uk/~lou/. Manfred Thaller, Homepage: http://www.hki.uni-koeln.de/people/thaller/mt.html.

[16] Consortium for Latin Lexicography, "The electronic version of the *Thesaurus Linguae Latinae*,"*1997*: http://web.archive.org/web/19981203074340/www.cs.usask.ca/faculty/devito/e-TLL/.

[17] A. DeVito, "Developing an Electronic *Thesaurus Linguae Latinae*," Consortium for Latin Lexicography, July, 1995: http://xml.coverpages.org/tll-ach.html.

[18] M. Madin, "Buddhist Studies Digital Library," Academic Info Inc., 2005: http://www.academicinfo.net/buddhismlibrary.html.

[19] C.-C. Chen and J. Z. Wang, "Large Scale Emperor Digital Library and Semantics- Sensitive region Based Material": http://www-db.stanford.edu/~wangz/project/imsearch/SIMPLIcity/DLOC/chen.pdf.

[20] EVA2002 Beijing Draft Outline Programme: http://www.tsinghua.edu.cn/docsn/cbx/5academic/conference/2002/eva1/schedule.html.

[21] The Oxford text Archive: http://ota.ahds.ac.uk/.

[22] F. Marburg, "Bildindex der Kunst und Architektur": http://www.bildindex.de/rx/apsisa.dll/init?sid={63c57939-0373-49da-a0cb-84cc87466e7b}&cnt=84020&%3Asysprotocol=http%3A&%3Asysbrowser=ie6&%3Alang=de&.

[23] G. Crane, "The Perseus Digital Library", Tufts University: http://www.perseus.tufts.edu/.

[24] Max Planck Institut für europäische Rechtsgeschichte, "Bibliothek": http://www.mpier.uni-frankfurt.de/dlib. Cf. M. Thaller, "From the Digitized to the Digital Library", *D-Lib Magazine*, vol. 7, no. 2, Feb., 2001. http://www.dlib.org/dlib/february01/thaller/02thaller.html.

[25] CEEC (Codicis Electronici Ecclesaie Coloniensis): http://www.ceec.uni-koeln.de/.

[26] C. van Assche, "New Media Encyclopaedia ," Centre Georges Pompidou, Paris, 2005: http://www.newmedia-art.org/.

[27] Netzspannung.org: http://netzspannung.org/index_flash.html.

[28] Prometheus: http://www.prometheus-bildarchiv.de/.

[29] One Look Dictionary Search: http://www.onelook.com/.

[30] The electronic Oxford English Dictionary was developed at Waterloo University and led to the development of the Open Text Corporation. Which as since developed "the Livelink ECM Platform, an integrated framework that combines a shared content repository with user, content, and process services in a Service-Oriented Architecture (SOA)." Open Text Corporation: http://www.opentext.com/.

[31] Oxford English Dictionary: http://www.oed.com/subscribe/individuals-rw.html.

[32] Thomson Dialog, "Sources": http://www.dialog.com/sources/.

[33] Association of Research Libraries (ARL), "Framing the Issue: Open Access": http://www.arl.org/scomm/

open_access/framing.html.
[34] S. Harnad, "Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry," *Psychological Science*, 1, pp. 342–343 (reprinted in *Current Contents*, 45, pp. 91–13, Nov. 11, 1991). http://cogprints.org/1581/00/harnad90.skywriting.html.
[35] P. Ginsparg, "Creating a global knowledge network," *Invited contribution for Conf. held at UNESCO HQ, Paris,* Feb. 19–23, 2001, *Second Joint ICSU Press - UNESCO Expert Conf. on Electronic Publishing in Science, during session Responses from the scientific community*, Tue. 20 Feb., 2001. http://arxiv.org/blurb/pg01unesco.html.
[36] arXiv monthly submission rate statistics: http://arxiv.org/show_monthly_submissions.
[37] arXiv.org e-Print archive: http://arxiv.org/. Front for the Mathematics ArXiv , University of California (UC) Davis: http://front.math.ucdavis.edu/.
[38] A. Vance, "Los Alamos lends open source hand to life sciences," *The Register*, June 23, 2003: http://www.theregister.com/2003/06/29/los_alamos_lends_open_source/.
[39] MPI, "Preprints of the MPI": http://www.mpim-bonn.mpg.de/html/preprints/preprints.html
[40] Max Planck Gesellschaft, "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," *Conf. on Open Access to Knowledge in the Sciences and Humanities*, Berlin, Oct. 20–22, 2003: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html. Max Planck Gesellschaft, "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," *Conf. on Open Access to Knowledge in the Sciences and Humanities*, Berlin, Oct. 20–22, 2003:, Conference Synopsis: http://www.zim.mpg.de/openaccess-berlin/.
[41] European Cultural Heritage Online (ECHO). Open Access Infrastructure for a Future Web of Culture and Science: http://echo.mpiwg-berlin.mpg.de/home.
[42] CODATA XIX, *Int. Conf., The Information society: New Horizons for Science*, Berlin, Nov. 7–10, 2004: http://www.wsis-si.org/CODATA/codata-samassekou-en.pdf.
[43] SPARC, Europe: http://www.sparceurope.org/.
[44] Public Library of Science: www.plos.org.
[45] DARE (Digital Academic Repositories): http://www.darenet.nl/nl/page/language.view/home. L. Waaijers, "By them going, pathways are growing, DARE; a work in progress," *SURF*, Mar. 7, 2005: http:///64.233.183.104/search?q= cache:Hz1ECridCaQJ:www.lib.helsinki.fi/finnoa/esitykset/Waaijers_Helsinki2004.ppt+leo+waaijers+surf+dare&hl=en.
[46] L. Waaijers, "By them going, pathways are growing, DARE; a work in progress," *SURF*, Mar. 7, 2005: 6 projects. CoMa: Copyright Management', Universiteit van Tilburg. 'DARC: Distributed Africana Repositories Community,' Universiteit Leiden. 'P-Web: een tool voor het online publiceren van proceedings,' Erasmus Universiteit Rotterdam. 'Scripties Online,' Universiteit Twente; Erasmus Universiteit Rotterdam; Rijksuniversiteit Groningen. 'Stroomlijning en digitalisering van het review proces,' Wageningen Universiteit 'Universitair Wetenschappelijk Archief (UWA),' Universiteit van Amsterdam.
[47] National Institutes of Health, Public Access: http://www.nih.gov/about/publicaccess/.
[48] S. Pincock, "RCUK draft mandates open access," *The Scientist*, June 23, 2005: http://www.the-scientist.com/news/20050623/01.
[49] Association of Research Libraries (ARL), "Framing the Issue: Open Access": http://www.arl.org/scomm/open_access/framing.html.
[50] ProQuest. Information and Learning Company:http://www.proquest.co.uk/products/ch_titlelists.html.
[51] ProQuest. Information and Learning Company, About UMI: http://www.il.proquest.com/umi/ab-about.shtml.
[52] ProQuest. Information and Learning Company: http://www.il.proquest.com/umi/ab-about.shtml.
[53] Early English Books Online, "About EEBO": http://eebo.chadwyck.com/marketing/about.htm.
[54] Early English Books Online, "Text Creation Partnership": http://www.lib.umich.edu/tcp/eebo/partner/partner_pricing.html. Pricing for the Text Creation partnership.

INSTITUTION TYPE	PARTNERSHIP FEE
ARL or equivalent institution	$60,000
Non-ARL graduate degree granting institution with more than 15,000 FTE	$45,500
Non-ARL graduate degree granting institution with fewer than 15,000 FTE	$30,000
Undergraduate institution only with more than 2,500 FTE	$21,000
Undergraduate institution only with fewer than 2,500 FTE	$15,000

[55] PR Newswire, "ProQuest Acquires ExploreLearning," *Cold Fusion Developer's Journal*, Mar. 1, 2005: http://cfdj.sys-con.com/read/62085.htm.
[56] ProQuest, Annual Report, 2004: http://www.proquestcompany.com/investor2/ar2004.shtml.
[57] Google, "Library Project": http://print.google.com/googleprint/library.html. Google, "Google Checks Out Library Books": http://www.google.com/press/pressrel/print_library.html. Even within the US there

is also opposition to the Google vision. Kimberley A. Kicheniuk, "Google Begins Digitalization": http://www.thecrimson.com/article.aspx?ref=507937.

[58] D. Welle, European Libraries Fight Google-ization, Apr. 27, 2005: http://www.dw-world.de/dw/article/0,1564,1566717,00.html.

[59] Yahoo's announcement of paid subscriptions for deep search on June 17 2005 is seen by some as a sign of things to come. Editor, "Yahoo 'deep search', " The Bosh, June 17, 2005: http://thebosh.com/archives/2005/06/yahoo_deep_sear.php.

[60] "About the NGI": http://www.ngi.gov/.

[61] Internet2: http://www.internet2.edu/.

[62] Jacqueline Brown, "Pacific Wave, Pacific Light Rail and National Light Rail," *CANS2002*, Shanghai, Aug. 22, 2002: www.oit.umd.edu/projects/cans/2002/Presentations/Brown.ppt.

[63] IEEE, "WG12: Learning Object Metadata": http://ltsc.ieee.org/wg12/.

[64] SCORM: http://www.rhassociates.com/scorm.htm.

[65] See for instance the author's, "American Visions of the Internet": http://www.sumscorp.com/articles/html/visions_22_dec.htm.

[66] http://www.eff.org/.

[67] http://www.lessig.org/blog/.

[68] CNI, "Clifford A. Lynch, CNI's Executive Director, Publications": http://www.cni.org/staff/clifford_publications.html. See especially: C. A. Lynch, "The Battle to Define the Future of the Book in the Digital World," *First Monday* vol. 6 no. 6, June, 2001; C. A. Lynch, "The New Context for Bibliographic Control in the New Millennium," *Bicentennial Conf. for the New Millennium: Confronting the Challenge of Networked resources and the Web,* Washington, D.C, Nov. 15–17, 2000.

[69] M. Giesecke, *Der Buchdruck in der frühen Neuzeit - Eine historische Fallstudie über die Durchsetzung neuer Informations- undKommunikationstechnologien*. Frankfurt/ Main (Suhrkamp) 1991, 2. Aufl., 1994; durchgesehene und mit einem Nachwort versehene Ausgabe 1998. Cf. Michale Giesecke, Homepage: http://www.michael-giesecke.de/giesecke/menue/index_h.html.

[70] J.-C. Guédon (Université de Montréal), "In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing:" http://www.arl.org/arl/proceedings/138/guedon.html.

[71] "Le savoir, un bien public mondial": http://www.freescape.eu.org/biblio/rubrique.php3?id_rubrique=11;http://2100.org/conf_queau1a.html. This idea has also been explored in the present author's Augmented Knowledge and Culture, University of Calgary Press, 2005 (in press).

[72] JISC, "Building a Virtual Research Environment for the Humanities (BVREH)": http://www.jisc.ac.uk/index.cfm?name=vre_bvreh&src=alpha.

[73] G. Fox and D. Walker, "e-Science Gap Analysis," *UKeS*,-2003-01: http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/.

[74] Maisons Science de l'Homme, "Centres, réseaux, associations de recherche hébergés ou domiciliés": http://www.msh-paris.fr/la_recherche/centres_recherche/cr_autres.htm.

[75] BNF, "A Major Project. 1988–1994: From a major project to the new Bibliothéque nationale de France": http://www.bnf.fr/site_bnf_eng/connaitrgb/projetgb.htm.

[76] BNF, "Gallica: about the Project": http://gallica.bnf.fr/FranceAmerique/en/D1/T1-1-Intro.htm.

[77] V. Khanna "French cry havoc over Google's library plans," Library Staff Blog, Mar. 28, 2005: http://www.library.upenn.edu/blos/staffweb/Current_Readings/french_cry_havoc_over_googles_library_plans.html.

[78] "French To Provide Alternative To Google Library Project," *Web Rank Info*., Mar. 17, 2005: http://www.webrankinfo.com/english/seo-news/topic-2267.htm.

[79] D. Welle, European Libraries Fight Google-ization, Apr. 27, 2005: http://www.dw-world.de/dw/article/0,1564,1566717,00.html.

[80] EDRI-gram, Initiative European Libraries to Digitise Books, no. 3.9, 4 May 2005: http://www.edri.org/edrigram/number3.9/digilibrary.

[81] "Over 11m Digital Records Now Available At European Library," *Managing Information News*, May 31, 2005: http://www.managinginformation.com/news/content_show_full.php?id=3893.

[82] Europa, "Viviane Reding Member of the European Commission responsible for Information Society and Media i2010," Europe Must Seize the Opportunities of the Digital Economy. *Press Conf. on the occasion of the launch of the initiative European Information Society 2010*, Brussels, June, 1 2005": http://europa.eu.int/rapid/pressReleasesAction.do?reference=SPEECH/05/312&format=HTML&aged=0& language=EN&guiLanguage=en; "New 'i2010' programme to unleash digital services in the EU," Euractiv.com: http://www.euractiv.com/Article?tcmuri=tcm:29-134976-16&type=News.

[83] To close the gap between the information society "haves and have nots," the Commission will propose: an Action Plan on e-Government for citizen-centred services, 2006; three "quality of life" ICT flagship initiatives (technologies for an ageing society, intelligent vehicles that are smarter, safer and cleaner, and digital libraries making multimedia and multilingual European culture available to all, 2007; and actions to overcome the geographic and social "digital divide," culminating in a European Initiative on e-Inclusion, 2008, UNI Global, EU: Launches new European Communications strategy "i2010": http://www.union-

network.org/unitelecom.nsf/0/5e33c14432197c30c125701400269b61?OpenDocument.

[84] For a discussion of the problems involved see the author's: Towards a Semantic Web for Culture, *JoDI* (*J. of Digital Information*), Special issue on New Applications of Knowledge Organization Systems, vol. 4 no. 4, Article no. 255, Mar. 15, 2004: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Veltman/.

[85] Recently Sir Tim Berners-Lee it in an interview with Andrew Updegrove explicitly stated that this direction was not the goal of the W3C: http://www.consortiuminfo.org/bulletins/semanticweb.php.

[86] J. Weizenbaum, *Computer Power and Human Reason*: *From Judgment To Computation,* San Francisco, W. H. Freeman, 1976.

[87] See: G. Fjermedal, *The Tomorrow Makers, A Brave New World of Living Brain Machines*, Redmond, Tempus Books, p. 188, 1986.

[88] For other work in the direction of free encyclopaedias see the work of Torsten Wöllert. "Offene Enzyklopädie": http://www.opentheory.org/enzyklopaedie/text.phtml.

[89] Encyclopaedias are but one expression of a much more complex tradition. For an introduction see: F. Yates, *The Art of Memory*, Chicago, Chicago University Press, 1966. For details see: G. Tonelli, *A Short-Title List of Subject Dictionaries of the Sixteenth, Seventeenth and Eighteenth Centuries as Aids to the History of Ideas*, London, 1971.

[90] Free Internet Encyclopedia, "Your Comments," Last updated Aug. 22, 1995: http://www2.cs.uh.edu/~clifton/com.html: "Please be advised that Encyclopaedia Britannica is the owner of federal trademark registrations 1,672,590 for the mark 'Macropaedia' and 1,672,591 for the mark 'Micropaedia'."

[91] The abridged Kleine Pauly in 12 volumes initially cost DM 268/volume. See: "H. Cancik and H. Schneider (Ed.), Der Neue Pauly. Enzyklopaedie der Antike. Altertum, Band 1 (A-Ari). Stuttgart: J.B. Metzler, 1996. Pp. liii, 577. DM 268/volume (subscription price)." *Bryn Mawr Classical Review*, Mar. 15, 1997: http://ccat.sas.upenn.edu/bmcr/1997/97.03.15.html.

[92] P. Noisette and T. Noisette, *La bataille du logiciel libre, dix clés pour comprendre*, Paris: éditions La Découverte, collection Sur le Vif, Oct., 2004. http://www.freescape.eu.org/biblio/article.php3?id_article=202.

[93] J. Boyle, "A manifesto on Wipo and the Future of Intellectual Property," Mis en ligne le mercredi 29 Sep. 2004; http://www.freescape.eu.org/biblio/article.php3?id_article=194.

[94] Free office suite, Open office.org: http://www.openoffice.org.

[95] The GIMP: http://www.gimp.org,

[96] Inkscape: http://www.inkscape.org.

[97] Blender: http://www.blender3d.org.

[98] JAHSHAKA: http://www.jahshaka.com.

[99] NASA, WorldWind 1.03: http://worldwind.arc.nasa.gov.

[100] SourceForgeNet, "Open standards and software for bibliographies and cataloging": http://wwwsearch.sourceforge.net/bib/openbib.html.

[101] R. Stallman, "The GNU Project": http://www.gnu.org/gnu/thegnuproject.html. cf. Richard Stallman's personal site: http://www.stallman.org/.

[102] Linux International, "Linux History": http://www.li.org/linuxhistory.php ; Linux Online: http://www.linux.org/.

[103] For a serious list of such alternatives see: J. Poulsen, DebianLinux.Net, "Freedom": http://debianlinux.net/freedom.html.

[104] Framasoft, "939 logiciels libres dans l'annuaire," http://www.framasoft.net/.

[105] Creative commons: http://creativecommons.org/.

[106] Open Content: http://www.opencontent.org/.

[107] Cover Pages, "Open Archives Initiative Releases Specification for Conveying Rights Expressions": http://xml.coverpages.org/ni2005-05-06-a.html.

[108] *Proc. of the New Ways, New Technologies Conference*, University of Calgary, Calgary Alberta Canada, University of Calgary Press. Oct. 15, 2004.

[109] Towards an Integrated Knowledge Ecosystem, A Canadian Research Strategy: http://www.kdstudy.ca/references.htm.

[110] R. Bloor, "The government open source dynamic," The Register, Jan. 7, 2005: http://www.theregister.co.uk/2005/01/07/gov_open_source_dynamic/.

[111] K. Regan, "Nokia, Apple Develop Open-Source, Mobile Web Browser," Linux Insider, June 13, 2005: http://www.linuxinsider.com/rsstory/43768.html.

[112] Open theory: http://www.opentheory.org/.

[113] CODES:Collaborative Open Design System for Integration of Information Webs with Design and Manufacturing Tools: http://codes.edrc.cmu.edu/CODES/contents.html. Open has many meanings. The Open design Alliance is open only to a small group of industry partners. Cf. Open Design Alliance: http://www.opendwg.org/.

[114] In a refinement of this approach the user can be offered a choice of seeing only a subset of the complete list which uses one of the variant names.

[115] SUMS: http://sumscorp.com/develop/.

[116] There is for instance the Galen Classification Workbench (ClaW): http://www.openclinical.org/dm_galenClaw.html. There are also the efforts of the Dutch WHO Collaborating Centre for the ICIDH to act as an intermediary for international classifications between WHO and the Netherlands: http://www.rivm.nl/who-fic/Annuals/Bethesda.08.doc.

[117] O. Streiter and L. Voltmer, "Document Classification

for Corpus-based Legal Terminology," European Academy, Viale Druso 1, Bolzano, Italy: http://dev.eurac.edu:8080/autoren/publs/iasi/.

[118] For a more detailed discussion see the author's "Towards a Semantic Web for Culture," *JoDI* (*J. of Digital Information*), vol. 4, no. 4, Article no. 255, Special issue on New Applications of Knowledge Organization Systems, Mar. 15, 2004: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Veltman/.

[119] Dan Corner, "Some Revealing Catholic Names, Titles and Prayers To Mary": http://www.evangelicaloutreach.org/marytitles.htm. Holy Mary Holy Mother of God; Most honored of virgins; Chosen daughter of the Father Mother of Christ; Glory of the Holy Spirit Virgin daughter of Zion, Virgin poor and humble, Virgin gentle and obedient, Handmaid of the Lord, Mother of the Lord, Helper of the Redeemed, Full of grace, Fountain of beauty, Model of virtue, Finest fruit of the redemption, Perfect disciple of Christ, Untarnished image of the Church, Woman transformed, Woman clothed with the sun, Woman crowned with stars, Gentile Lady, Gracious Lady, Our Lady, Joy of Israel, Splendor of the Church, Pride of the human race, Advocate of grace, Minister of holiness, Champion of God's people, Queen of love, Queen of mercy, Queen of peace, Queen of angels, Queen of patriarchs and prophets, Queen of apostles and martyrs, Queen of confessors and virgins, Queen of all saints, Queen conceived without original sin, Queen assumed into heaven, Queen of all earth, Queen of heaven, Queen of the universe (pp. 190,191) From the "Litany of Loreto" Mother of the Church, Mother of Divine grace, Mother most pure; Mother of chaste love; Mother and virgin, Sinless Mother, Dearest of Mothers, Model of motherhood, Mother of good counsel; Mother of our Creator; Mother of our Savior; Virgin most wise; Virgin rightly praised; Virgin rightly renowned; Virgin most powerful; Virgin gentle in mercy; Faithful Virgin; Mirror of justice; Throne of wisdom; Cause of our joy; Shrine of the Spirit; Glory of Israel, Vessel of selfless devotion; Mystical rose; Tower of David; Tower of ivory; House of gold; Ark of the covenant; Gate of heaven; Morning star; Health of the sick; Refuge of sinners; Comfort of the troubled; Help of Christians; Queen of the rosary; Queen of peace (pp. 191,192)

[120] M. Pilgrim, "What is RSS?": http://www.mantraonnet.com/108durganames.htm.

[121] http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html; Webref, "Introduction to RSS": http://www.webreference.com/authoring/languages/xml/rss/intro/.

[122] The term memory institutions to describe the combination of museums, libraries and archives was introduced within the European Commission in the early 1990s and has only gradually moved towards international recognition. Some use the term cultural institutions, others speak of the ALM (archives, libraries, museums) sector. Cf. Lorcan Dempsey, Scientific, Industrial, and Cultural Heritage: a shared approach, Ariadne, no. 22, Dec. 21, 1999: http://www.ariadne.ac.uk/issue22/dempsey/.

[123] O. Gibson, "Coming soon: Googling the truth," The Guardian, June 18, 2005: http://www.guardian.co.uk/international/story/0,3604,1509281,00.html.

[124] Art Libraries Society of North America, Cataloging Advisory, "Anonymous Artist Relationships in the MARC 21 Bibliographic Format," Discussion Paper 115, May 14, 1999. Cf: http://www.loc.gov/marc/marbi/dp/dp115.html. For an example from geology, see: POSC Specifications Version 2.2.2, "Classification System": http://www.posc.org/Epicentre.2_2/DataModel/LogicalDictionary/StandardValues/classification_system.html.

[125] For an example of current approaches see: T. DiLauro, "Choosing the components of a digital infrastructure", *First Monday*, no. 9: http://www.firstmonday.org/issues/issue9_5/dilauro/.

[126] http://en.wikipedia.org/wiki/Andrew_Lang; http://www.users.globalnet.co.uk/~crumey/andrew_lang.html.

[127] E.g. Charles Phelps (Provost, University of Rochester): "The central idea would have, the learned societies expand their role to undertake a certification process for articles, independently of whether they are submitted for, or are eventually published in the standard paper journal system. Under such a system, scholars could submit articles for review (with an agreed-upon submission fee), and the normal refereeing process of the learned society would determine whether the article qualified for their "seal of approval," which, if received, could be affixed to any electronic version of the article as retrieved by others. With such a certification, if appropriately "honored" in processes that rely upon such certifications, including tenure and promotion in colleges and universities and grant applications from governments and foundations, the necessity to carry on with paper publication (which serves only the certification and editing processes in addition to the distribution, indexing and archiving that the computer file-server system can serve) could diminish or vanish at least in some settings. But until such refereeing, and in some settings, editorial functions are provided, the paper journal system will persist in parallel with whatever electronic system emerges." C. E. Phelps, "The Future of Scholarly Publication. A Proposal for Change:" http://www.econ.rochester.edu/Faculty/PhelpsPapers/Phelps_paper.html.

[128] arXiv monthly submission rate statistics: http://arxiv.org/show_monthly_submissions.

[129] D. DeSolla, 1922–1983: http://www.asis.org/Features/Pioneers/price.htm.
[130] HyperJournal Website, Core Team: http://www.hjournal.org/team.
[131] HyperJournal Website, Features: http://www.hjournal.org/features. It also includes 1) Open Archive OAI-PMH protocol compliance; 2) unlimited number of scientific and editorial committees; 3) online anonymous peer-review.
[132] P. D'Orio, "Nietzsche Open Source": http://www.freescape.eu.org/biblio/article.php3?id_article=64.
[133] P.D'Iorio, "Sharing Knowledge in Web Communities," *Open Source Models in the Humanities*, University of Chicago, Apr. 26, 2004: http://cmig.uchicago.edu/diorio.htmlPogrettare. Cf. Scuola Normale, Conference: Progettare su web.Digital Libraries di parole e immagini (centri di ricerca e grandi biblioteche), Pisa, Apr., 2005: http://www.sns.it/it/strumenti/ufficiostampa/formazione/annoincorso/geografia/. For other views on the present state of scholarly communication see: R. King and E. Callahan, "Electronic journals, the Internet and scholarly publishing," *Annu. rev. of Information Science*, vol. 37, pp. 127–177, 2003; R. Kling, "The Internet and unrefereed scholarly publishing," A*nnu. Rev. of Information Science and Technology*, vol. 38, pp. 591–631, 2004.
[134] Hyperlearning, HyperMedia Platform for Electronic Research and Learning: http://www.hypernietzsche.org/doc/hyper-learning/.
[135] Ashwin K Pulijala Susan Gauch , "Hierarchical Text Classification": http://computing.breinestorm.net/classification+classes+hierarchical+concept+assigning/.
[136] The challenges of linking the two National central libraries of Italy in Rome and Florence offer a case in point. Maria Patrizia Calabresi, "Two national central libraries in Italy: bibliographic co-operation or competition?," *66th IFLA Council and General Conf.*, Jerusalem, Israel, Aug. 13–18, 2000: http://www.ifla.org/IV/ifla66/papers/066-123e.htm.
[137] E.g. Texas Woman's University, "TWU Libraries": http://www.twu.edu/library/res/res_thesesgen.htm.
[138] Cf. S. Keene and F. Monti, "The DEER: Distributed European Electronic Resource." Final Report, 2003. In *E-Culture Net: Work Package 6*, *Deliverable 11a. IST-2001-37491*: http://www.eculturenet.org.

Appendix 1 Different levels of certainty in making a claim in terms of basic questions.

Claim (How)
1. Authoritative
2. Very Certain
3. Quite Certain
4. Very Probably
5. Quite Probably
6. Possibly
Claim (Who)
1. Author
2. Student
3. Workshop
4. Follower of
5. Copier of
6. After
Claim (What)
1. Object
2. Class
3. Species
4. Genus
Claim (Where)
1. House
2. Street
3. City
4. Province
5. Country
6. Continent
Claim (When)
1. Date
2. c.
3. c.- c-
4. c.?-c.?
5. fl.
6. Century

Kim H. VELTMAN

Kim H. VELTMAN is an author and consultant re: implications of new media for scholarship, culture and society. He has taught at the universities of G– tingen, Rome, Carleton; was Director of the Perspective Unit, McLuhan Program, Toronto (1990−1996), and Director of the Maastricht McLuhan Institute (1998−2005). He has worked as a consultant to the CEO of Bell Media Linx (1996−1998); is a permanent consultant to the Scuola Normale Superiore, Pisa.; on the Board of the Special Interest Group for the Semantic Web and Information Systems (SIGSEMIS), and a member of the International Who's Who of Professionals.