What is an Analogue for the Semantic Web and Why is Having One Important?[1]

mc schraefel
IAM Group, Electronics and Computer Science
University of Southampton, UK
mc+ht07 at ecs.soton.ac.uk

ABSTRACT
This paper postulates that for the Semantic Web to grow and gain input from fields that will surely benefit it, it needs to develop an analogue that will help people not only understand what it is, but what the potential opportunities are that are enabled by these new protocols. The model proposed in the paper takes the way that Web interaction has been framed as a baseline to inform a similar analogue for the Semantic Web. While the Web has been represented as a Page + Links, the paper presents the argument that the Semantic Web can be conceptualized as a Notebook + Memex. The argument considers how this model also presents new challenges for fundamental human interaction with computing, and that hypertext models have much to contribute to this new understanding for distributed information systems.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext and Hypermedia; H.5.2 [**User Interfaces**]: Human Information Processing.

General Terms
Design, Human Factors, Documentation.

Keywords
Memex, notebooks, hypertext argumentation, interaction design, Semantic Web, Jourknow, mSpace, Tabulator

1. INTRODUCTION
In order to design either a system or an interface to support a technology, it helps to know what it is - or failing that - to have a model around which we can conceptualize what it is, what it does, and somewhat how it works. It is not unusual for a new technology to be introduced via an analogue of a previous, familiar technology "it's like this thing - but for this new bit." Word processors for instance used to be described as “like typewriters except for copy and paste.” The familiar along with the New Idea. The Web has likewise frequently been explained along these lines: the Web = a Page + Links. The concept of the printed page is one with which we are all familiar. It's clear, easy to grasp. The link offers only one new concept to understand, and it is largely communicable in practice: click on the link; go to a new Page, with Links. The rapidity with which people started creating and using new Pages for the Web demonstrates the success of the model: one creates some text (with images if desired); adds links to other similar types of Pages, and voila, one has a Web Page. Based on the success of the Web, a new suite of Web technologies and protocols have been developed, collectively called the Semantic Web. This grouping of technologies promises new and more powerful ways to interact with information on the Web and to build new knowledge from those interactions. While this all sounds very good, there has been no analogue proposed for the Semantic Web that is similar in communicative power to the Web as a Page plus Links.

What is the equivalent analogue for the Semantic Web to help make it tractable? It is not obvious. It may be argued that the lack of such an analogue for communicating the Semantic Web to communities outside Semantic Web research is a contributor to the relatively slow or resistant take up of the Semantic Web within communities whose work could greatly inform its development: human computer interaction, information retrieval, information architecture, and what should be its proper home, Hypertext. It is important to note that the motivation for this question of analogue is not a marketing/packaging question to help sell the Semantic Web, but is simply a matter of fundamental importance in any research space: it is critical to have both a shared and sharable understanding of a (potentially new) paradigm. If we do not have such a shared understanding, we cannot interrogate the paradigm for either its technical or, perhaps especially, its social goals.

In the following sections, how technology models based on older familiar models actively assist development of new technologies is considered. Then by looking at how this modeling approach has informed the Web, we propose a possible way to construe the Semantic Web via a model steeped in Hypertext tradition. The paper closes with a consideration of how this model may open new design paradigms beyond the Semantic Web and for computing interaction, as well as for the new field of Web Science. These seem like bold claims. They are not meant to be proclamations, but more a contemplation of a possible research agenda to include *other* ways we might think about computing if we start with a blank page in a fresh notebook, and let hypertext ideas be, literally, re-presented in a call to renew perhaps, rather than just to the new.

This is a pre-print of the paper to be available for.
HT’07, September 10–12, 2007, Manchester, United Kingdom.

[1] Parts of this paper have appeared in a blog post at http://dig.csail.mit.edu/breadcrumbs/node/184.

2. The Web, the Page, and History

There is no argument that the Web is a success story. It has changed not only the way we access information, but it has also changed our expectations for information: if it is not on the Web, it does not exist. For example, as bibliometric studies have shown, citation rates are significantly higher for material that is accessible on the Web, compared with material only available in print [7]. There have been many things that have contributed to the success of the Web, from powerful search engines that make content discoverable, to commercial take up of the Web as a core medium for communication. Significantly, it has brought people into contact with computers and global network who otherwise would have had no contact with such systems. We might argue that this success of the Web is largely because the paradigm of the Web is powerfully familiar. That is, despite the newness (to most people) of this complex of networks and protocols known as "the Web," its paradigm is based on prior, well-established, well-used technology from the past millennia at least. The Web *page* is in many ways, a simulacrum of both a technology and form of communication with which we have tremendous familiarity: the read-only text of the printed page.

Figure 1: One of Banksy's anti-ads in London, UK, 2005 [1]

We have a long history with read-only text, whether as official public communication, such as obelisks that communicated history and cultural imperatives, to government posters, such as the famous 1917 "I Want YOU" [19]. With the growth of the printing press, unofficial counter-commentary from 17th Century political handbills glued to lamp posts to more contemporary anti-ads like the artist Banksy's political commentary (shown in Figure 1) it has become easier to make alternative views publicly available. We also have a long experience (400+ years) of a particular technology's deployment of words and images in a page – taking us from the relative exclusivity of hand copied illuminated manuscripts to early printed texts with woodcut illustrations (Figure 2).

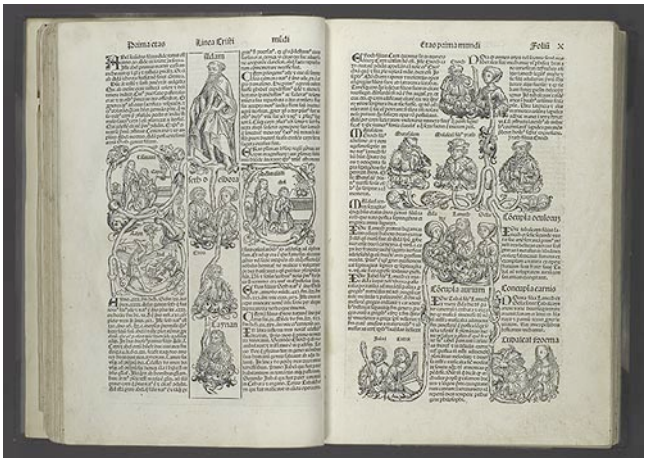


Figure 2. *Liber Chronicarum*, 1493. [17]

The Web draws on this familiarity: it does not look like some strange new technology that requires strange new devices; it does not remind us of its stateless, network accessing, server dependent vastness. Rather, the Web looks very familiar. The Web as it was introduced to us, and largely how it has evolved draws on this highly familiar mode of the printed page for communicating content. The one new thing added in the Web to the notion of the page - the thing that makes it a Web page - is the hypertext link. The link is the core new concept introduced to the page, and more times than not, that link's job is to link the current page to another page. The mental model for understanding the Web, with its unary links, can be well supported by the page. Indeed, the Web's fundamental "ease of use" is often attested to by the uptake of the technology by largely self-taught Senior Citizens [26]: if elders can do it, goes the argument, it must be easy; if they are doing it, it must be ubiquitous.

This is not to say that there are not a myriad of design and usability challenges for making that page+link approach useful, usable and accessible. We have developed whole suites of conventions on how to deliver pages effectively and have gone through now what are referred to as "generations" of web design to ensure that text, image and link work [29]. Yet despite over a decade of technological evolutions informing the Web, how it can access content, how browsers can present that content dynamically and programmatically, the paradigm for describing what we create with the Web is the same: it's a page. With Links. That paradigm informs how we design Web content: not as a spreadsheet; not as a network diagram, but as a page.

Even with Web 2.0, with RSS feeds, blogs, mashups, we still have pages. The only slight page model variant in Web 2.0 may be with location based mashups. In these pages, the main content rather than text is now a map. And again, maps are also highly familiar technologies that have been around for millennia, and accessed in posters and printed books. Maps are a technology most of us have even had some formal training on how to use at various points in our education.

In terms of communicating functionality to people – how to use the thing – the model of the Web page as page is clear, familiar, highly expressive, and rapidly communicates what the Web is largely about: enabling people to communicate ideas, and with that one special tool, the link, to hook their ideas into the myriad of other ideas available on other Web pages. The great new concept of the Tag[2] to mark and aggregate content like blog entries or photos for rapid representation, for example, still outputs its results in catalogue-like page indices of "Tag clouds" where size of tag represents its popularity in a given system. It is in large part because there is such a clear model of how to access Web content and make use of Web technology that there has been such rapid adoption of that technology across sectors.

The Semantic Web changes all that.

3. HYPERTEXT IN THE MACHINE

While the Web may be over ten years old and can claim world domination, even at five years old it had become a tour de force. The Semantic Web has effectively just turned five: it has been five years since the original Scientific American article on the Semantic Web was published [4]. A five years on article has recently been published [27]. While the community of Semantic Web researchers can claim increasing traction within some parts of the computing industry, there is still considerable skepticism on two sides of the computing space: back end technologists and front end researchers, designers and lest we forget, users. There is far less understanding, even within the computing space, about what the Semantic Web is, five years on, compared with the Web at five. At meetings with leaders in Information Retrieval over a year ago, misconceptions about the Semantic Web abounded: "isn't that just that old [*i.e.* failed] AI stuff?" was a common theme. At a Human Factors conference recently, the response from people who should know better was "I don't care what the back end is; I'm platform agnostic." And yet, it is the capabilities enabled by the back end that often inform how we imagine the possible of what can be delivered at the front end.

One might suggest that the technology deserves what it gets: if it is not being picked up by researchers or the commercial sector in large measure, then perhaps there is a reason: it is fundamentally flawed, or damaged goods. After all, that kind of argument has been made of hypertext – until the Web made (a version of) it "real" to a far greater population than the limited set of hypermedia researchers. Today, indeed, the annual Web conference attendance surpasses numbers at either Hypertext or at the International Semantic Web Conference itself. Indeed, comparisons between the Semantic Web and Hypertext are not unknown. Leading lights in the Semantic Web community have been quoted as saying "we don't want what happened to Hypertext to happen to the Semantic Web." Of course such statements are informed by ignorance of the actual hypertext community, but such comments also make clear how critical it is to communicate not only what the technology and research agenda is about, but what the potential benefits of that work are. That is, what problems is this new technology going to solve that makes the cost of adoption worth the supposed benefits? And by the way, what is being adopted? What is the Semantic Web?

How best to answer this question perhaps needs to take into account the people the Semantic Web community wish to attract to be involved as practitioners, innovators, creators, and discoverers in this space. If that population is to include the same range of passions and expertise that have brought so much to the Web from the arts, humanities and sciences, among others, then how this question is answered becomes critical.

Consider for a moment how the Semantic Web has been described in the new First Stop Shop for What Something Is, Wikipedia. The Wikipedia entry for the semantic web begins:

> The Semantic Web is an evolution of the World Wide Web in which information is machine processable (rather than being only human oriented), thus permitting browsers or other software agents to find, share and combine information more easily. It is a manifestation of W3C director Tim Berners-Lee's vision of the Web as a universal medium for data, information, and knowledge exchange.
>
> At its core the Semantic Web consists of a data model called Resource Description Framework (RDF), a variety of data interchange formats (e.g RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL) that facilitate formal description of concepts, terms, and relationships within a given domain. The burgeoning Semantic Web comprises newly created and/or transformed web data sources endowed with computer-processable meaning (semantics). [2]

All that description tells anyone about the semantic Web is that it is for Machines. As a Semantic Web researcher, who works with a community of Semantic Web researchers, one would be hard pressed to find a majority opinion that believes that the end game imagined for the Semantic Web is to make data easier for machines to process. Machine-processable data is truly a gnarly problem, but it is a means to an end, not the end itself. The end, as with the Web, is still about people, and *people* being able to build knowledge by moving through linked information. Consider the following statement from the founders of the Web Science Research Initiative, who are leaders in Hypertext, the Web and the Semantic Web.[3] In the *Science* article "Creating a Science of the Web" they state the following rationale for starting a Web Science discipline:

> Since its inception, the World Wide Web has changed the ways scientists communicate, collaborate, and educate. There is, however, a growing realization among many researchers that a clear research agenda aimed at understanding the current, evolving, and potential Web is needed. If we want to model the Web; if we want to understand the architectural principles that have provided for its growth; and if we want to be sure that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, then we must chart out a research agenda that targets the Web as a primary focus of attention [3].

The emphasis here is on human engagement with this Web technology. Indeed, the article describes the exemplar motivation

[2] Tag (metadata) http://en.wikipedia.org/wiki/Tag_(metadata).

[2] Wikipedia is a fluid source. The quotation reflects the state of the entry as of Jan. 31, 2007. http://en.wikipedia.org/wiki/Semantic_web

[3] http://www.webscience.org/about/people/

for the Semantic Web as how it will aid a scientist in drug discovery: "Researchers are exploring the use of new, logically based languages for question answering, hypothesis checking, and data modeling. Imagine being able to query the Web for a chemical in a specific cell biology pathway that has a certain regulatory status as a drug and is available at a certain price [3]."

We might ask, then, if the Semantic Web has effectively the same human-oriented goals as the Web, why not use the same model for describing it: pages with links. While that was in large part the approach proposed in the foundational 2001 article, there is a growing awareness that the page is not necessarily robust enough to support what more we get from the Semantic Web's linking capacity to connect information across domain axes. The above drug example would seamlessly connect information about a particular cell to a variety of possible relevant domains: regulatory status, dispensers, related research, use in parallel investigations. We can imagine more radically diffuse but still logically associatable shifts from domain to domain that the Semantic Web can support. Consider someone exploring a music space (however that may be represented) who has heard something they like that turns out to be by Wagner. In a Works domain they can see all his compositions. The Semantic Web data model promises to make it possible to link in data on say performances to compositions and then project the data through a Timeline visualization. With such a representation, it becomes possible to see that there have been key periods as well as geographical locations where Wagner has been performed, contrasted with periods where his work has been seemingly ignored. Now connect in information on Historical Events and locations, and it becomes possible to correlate an influx of performances in Germany during WWII and a decrease internationally post WWII; indeed performances of his work in Israel more recently have become points of strong social and ethical controversy. The above interaction with data to explore associations across all these domains takes us outside the page.

One might say that the whole rationale of Information Visualization and Information Seeking is to provide means to support identification of moments of interest in data spaces, hence what is new with the above Semantic Web scenario? IBM's new Many Eyes tool[4] to enable researchers to upload data to the web and share representations of a spreadsheet worth of data with others is a compelling example of where a little bit of Web 2 can get one. The Semantic Web, however, provides the technologies to make explorations *across* domains dynamically in a kind of 6 degrees of separation approach technically tractable. These resources are also not fixed single data files but cut cross dynamic, multiple, heterogeneous sources and data providers. A critical challenge then becomes just *how* to represent these new affordances to enable and take advantage of this rich interlinking of (meta)data for exploration.

Some of us in the Semantic Web & User Interaction community[5] have been considering these problems: mSpace, Exhibit, Haystack, Topia are exemplars of efforts to take advantage of not only the metadata, but the cross-domain linking that the Semantic Web might enable. Tabulator is a more recent and even wilder approach as it attempts to leap from RDF source to RDF source across unknown schemas and enable these diverse sources to be queried (and thus integrated) dynamically.

[4] http://services.alphaworks.ibm.com/manyeyes/home

[5] http://semanticweb.org/swui

This kind of emphasis on rich interlinking of data sources, focusing on representing not only the data but the metadata of an object explodes representation parameters beyond the page into other kinds of exploratory models for discovery and knowledge building. Indeed, these models reach back to fundamental hypertext and hypertext systems and forward to new kinds of representations and interaction challenges when applied at Web Scale. But how do we describe this potential? For a community steeped in rich link models, Hypertext is an obvious conceptualization. But beyond this community, Hypertext equals "a page with links" – it equals the current Web, not the rich possibility of what we might call Real Hypertext, which was modeled in Note Cards and Microcosm. We may ask then is Hypertext as imagined in the late 20th Century a better framing for the Web Scale possibilities enabled by the Semantic Web? These early hypertext models were imagined largely as local systems. Do we then need to go further back? The original coin of hypertext with Nelson's transpointing and transclusions [20] was certainly not restricted imaginatively to local-only systems. But it was largely constrained by traditional notions of documents and pages in particular. Long pages, but pages in documents nonetheless: components of other people's work could readily be used either to support argument in a new document (transclusion) or to provide commentary on another document (transpointing). This is the view of the world as ongoing narratives, of interactive prose. Of *literary* machines.

The Semantic Web promotes thinking of information as, if not more then at least as also other than, and also often prior to, a page or a document. In this respect, the metadata is as valuable as the data as is the provenance of that data. By extension, the meanings, the semantics, the ways of interpreting and hence the ability to link/associate these sources with related sources automatically becomes an alternative way of thinking about the hyperlink *as* meaning. That is, the *way* meaning is communicated that is not via the explicit prose page or catalogue page, but is via the exposure of the ways in which data is associated, and can be discovered, by direct semantic association, for the reader/interactor/explorer to make meaning.

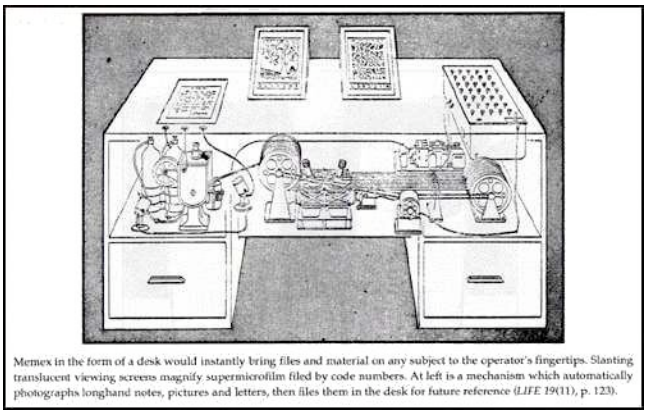
Thus, we see that beyond the Wikipedia definition for the Semantic Web, the Semantic Web's promise is to *enable* people to explore, associate, and connect information to build new knowledge. Thus if Nelson's model of hypertext does not capture these metadata or subdata strata of information, perhaps we need to go back further, prior to the coining of the hypertext term, and return to an early source, Bush's Memex, and see how it may help communicate the possible to be enabled by the Semantic Web.

4. MEMEX AS PARTIAL SW MODEL

Most people in the Hypertext community (and much of the Computing community beyond it [10]) can immediately site the source article for the system described as the Memex, V. Bush's As We May Think [8] (imagined a few years after publication as shown in Figure 3). One of the key parts of the Memex is the making and sharing of associations crafted among diverse sources by the person using the Memex. Bush imagined professions of "trail blazers" (section 8 of As We May Think) enabled by the Memex who would go about creating these connexions and publish them in new kinds of encyclopedias. His goal was to enable people to move across information "associatively" – modeled on how, he said, the brain builds knowledge. These associative connections have been translated into hypertext links,

and the ubiquitous unary Web link. It may be worth arguing that tagging is evolving into a very rapid lightweight way of making at least new connections, if not the richer notion of Bush's trails.

There is nothing either explicitly semantic or automatic about the description of trail-making in the Memex. Even rediscovery of resources is based on remembering and retyping the name of the label the operator gives to a work they have added to their personal Memex store. The Semantic web on the other hand promises that associations can be made inferentially and automatically by taking advantage of the use of both explicit semantic structures and the use of logic to reason over those structures.



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Figure 3. Drawing of Bush's theoretical Memex machine (Life Magazine, November 19, 1945) [1]

Interestingly, the earlier part of Bush's article, prior to describing the Memex, explicitly focuses on calculations machines should be able to carry out through the application of logical processes. Bush makes the distinction between "repetitive thought" and "creative thought" and that there ought to be "powerful mechanical aids" for the former. He goes on, "Whenever logical processes of thought are employed—that is, whenever thought for a time runs along an accepted groove—there is an opportunity for the machine" (section 5). We have seen just this kind of automation of patterns throughout computing, but when combined with trail making, Bush's description has in part been realized in Semantic Web practice. For instance, the myGrid project developed workflows for bioinformaticians to explore gene databases, running variations of the same processes to generate results to interrogate genetic patterns. Work that took days or weeks or more could be reduced to hours [30]. Likewise, the Haystack project used similar kinds of patterns with a direct manipulation interface to pull together resources in an integrated scheduling scenario for trip planning [21]. The Haystack scenario in particular draws in one's own data to mix with external information: personal calendar data and travel/flight information, for example.

The imagined automatic, logical processing of "repetitive thought tasks," and the ability to make (or infer where appropriate) links associatively across heterogeneous resources in new and unexpected ways related to either these kinds of tasks, or to the "creative thought" processes, gives us a strong model that captures at least part of the Semantic Web, and as shown, has already been explored in research from the scientific to the personal. The Memex offers us a model of the "what's new" part of our analogue approach to describing technology. Where the Web is the Page + Links (the familiar + the new), the Memex is the second part of the sum, the Semantic Web = Blank + Memex. We are left still to define critical familiar part of the equation. The description of interactions with the Memex points to a potential model.

5. WORK IN PROGRESS & NOTEBOOKS

The end game of the Memex is to enable the scientist to "extend the record." As Bush puts it,

> Presumably man's [sic] spirit should be elevated if he can better review his shady past and analyze more completely and objectively his present problems. He has built a civilization so complex that he needs to mechanize his records more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory. His excursions may be more enjoyable if he can reacquire the privilege of forgetting the manifold things he does not need to have immediately at hand, with some assurance that he can find them again if they prove important.

The above describes processes of building new thought based on connecting new ideas with previous personal and public data. It foregrounds the need to be able to forget about data management and focus on the present "creative thought" *with some assurance* that the material forgotten can be retrieved. What Bush describes here could in large measure be the mandate for research in personal information management [14]: to address the challenges of information capture and the problems of later retrieval. But what Bush adds to the description that takes it beyond a data management problem, is that the data management is in the service of a particular goal: to support work in progress. Bush wants a tool that will support creative thought.

We have a mechanism at least as successful and pervasive as the page which has for centuries served the function of personal information management for work in progress: the notebook. In the following discussion, we will consider how a model of Notebook + Memex can be used as an analogue to express the rich potential of the Semantic Web not just as a read-only mechanism like the Web, but as a mechanism for the ongoing work of our own review of our shady past and analyze more completely and objectively our present problems, which include both local personal and public informing sources.

5.1 Affordances[6] of the Notebook

Most of us have some experience of the notebook as support tool for our own work in progress, whether to capture short thoughts, experimental observations or ideas gleaned during a meeting. It is a highly flexible tool. It supports a variety of input types (pencil, pen) and data types (sketches, photos, samples, text). It also has attributes to support multiple retrieval processes: the ordered sequence of pages can be used to support temporal progress; physical width can be used for random access to relocate a note (it was around the middle of the book). In particular, the notebook also affords easy capture of this rich variety of idiosyncratic notes, what we have been calling "information scraps" [6] that information which may have no other formalized home, like an

[6] See Mads Soegaard, Affordances, Norman's Use of the Term, Section: Encylopedia, Interactions-design.org, http://www.interaction-design.org/encyclopedia/affordances.html

address book or calendar (for a more complete catalogue of a paper notebook's affordances, see "Breaking the Book"[24]).

The notebook while using *pages* as media, breaks the printed-page paradigm prevalent in the Web as well structured, well presented, largely read-only information space. In the notebook, we are really looking at *a blank surface* bound into a single, portable container. As such these books are fundamentally unlike what we usually think of as the Web in at least one particular way: the web is public; we use its protocols to publish work. Lab/note books are usually personal, idiosyncratic, again emphasizing work in progress (Figure 4). Even the complete capture of an experiment in a formal lab context is not the finished work, but is the raw observations and in-progress annotations to be available for the analysis of that work towards some understanding of an hypothesis [24]. Only under certain circumstances are notebooks called into a more public use as evidence for tracking the genesis of an idea or discovery. More casually if they are shared it is to offer a glimpse of an idea to a colleague– usually with close supervision, and for the purposes of interacting with the data directly, synchronously with the collaborator.

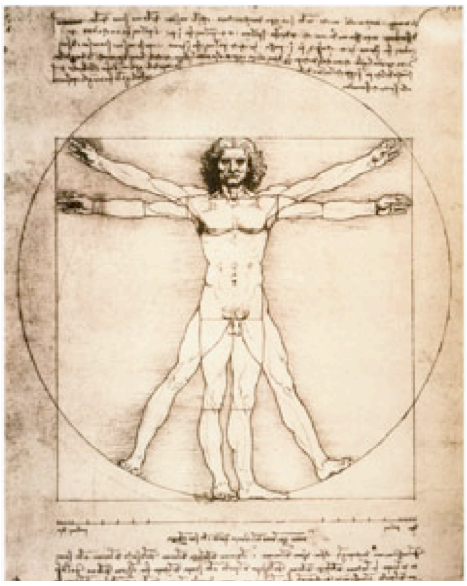


Figure 4. Quintessential version of a scientist's notebook capturing ideas/work in progress: a page from Da Vinci's notebook working out a sketch to accompany a translation of Viturvius's work on Architecture. The text is a translated quotation from Viturvius's work, which the figure illustrates [22].

This is not to say that we do not see traces of, if not work in progress, then what we might call the persona in progress on the current Web: there is a growing trend of "social stalking" on social networking sites, and "self-stalking" web-services. Blog spaces like Facebook[7] publish rapid updates of information added by one member as immediate alerts to associated members/"friends" of the person. Likewise Twitter[8] enables people to post from their phone fast updates of what they're doing, where. Pithy posts such as "getting on the bus" are not infrequent. Such collections might be construed as valuable contextual material for work/thoughts in progress, if not as the primary material of notes on work itself: they may act in the same way a phone number or meeting reminder might be scribbled beside the first few bars of a new sonata. One item can act as a way of refinding the other: "I put that by the notes for the sonata; the new sketch is by Peter's phone number." But in the Web context, even in these brief bursts of personal, we see that they are produced for publication, at Web scale levels of access rather than for direct support of personal reflection, idea generation or work progress. This is not to say that the Web is not trying to support these more private branches of endeavor. Various Web 2.0 services like Web-based stickies and note keepers do exist, including of course one by the increasingly ubiquitous Google with Google Notebook, a clipping service where links can be annotated and grouped into collections. In related work surveying knowledge workers, none of the 27 people we worked with used these Web based tools for note taking or information management [6]. Applications for collaborative writing, from Sub Etha Edit to Google Docs have far greater take up. It is not clear the degree to which these online word processors are being used as notebooks rather than task specific tools for completing a specific writing project.

5.2 Non-affordances of Digital Capture

One would be hard pressed to say that right now using a computer is as easy for data capture in particular as using a paper notebook. Research in personal information management [14] suggests that one of the key problems of taking the kinds of information we readily capture on paper over to the digital is an issue of both data capture and data retrieval. That is (a) there is a high cost to get the data into the computer and (b) it is not always easy to get it back out [16]. Consider the problem of digitally capturing a phone number of someone met just once. If using a paper source, one might use a scrap of paper, note the number and stick the note in a book or on the corner of a desk; indeed the note may be moved to a variety of locations, and reinforce awareness of its location. On the computer, one may feel very clever and have the person beam their contact information, including phone number, from their phone to their laptop, thus avoiding the multi-step process of opening an address book application, creating a new form, and entering data into the form's fields – a timely process at best. In either case, one month later, how will one find the phone number if all the person remembers is where they were when the data was captured, but not the person's name? The only option is brute force search through the address book. With the paper notebook, one can say "ah, that number is next to the notes for that meeting that happened just before I left X." In other words, the notebook provides both excellent rapid input as well as usable multiple context cues for rediscovery of data. Our digital tools tend to denature the information we capture from any context. In the case of the phone number, while there may be rich data about the person, their job title and their address captured in addition to the phone number within that beamed transfer, the context of capture, that incidental data critical to its recovery, is lost. Bush's goal of a tool that will enable temporary forgetting of data in the confidence that it will be rediscoverable when needed is not met in such a circumstance.

Bush imagined the Memex to have an easy interaction for data capture that did not denature it.

[7] For an overview of Facebook features, see http://www.facebook.com/sitetour/

[8] Twitter.com home page: "A global community of friends and strangers answering one simple question: What are you doing? Answer on your phone, IM, or right here on the web!"

> One can now picture a future investigator in his [sic] laboratory. His hands are free, and he is not anchored. As he moves about and observes, he photographs and comments. Time is automatically recorded to tie the two records together. If he goes into the field, he may be connected by radio to his recorder. As he ponders over his notes in the evening, he again talks his comments into the record. His typed record, as well as his photographs, may both be in miniature, so that he projects them for examination (Section 3).

This scenario implicitly foregrounds two critical facets: interaction with the system is transparent; some metadata is added to preserve some context to be able to associate related data . With just one automatically added metadata tag, time, two records, notes and images, are linked. How the evening's spoken notes are associated with the field notes is less clear, but it is obvious that semantics are being used to maintain connections among related types of information. Indeed, in the SmartTea project [13] we used a similar type of lightweight semantics to tie parts of synthetic chemistry experiments together with the goal of enabling groups of them to be interrogated in various ways.

In Bush's example, there is not a form in sight; no one is required to put a first name into a first name field and a last name into a last name field and so on. Likewise, the data captured is not hived off into discrete applications for each data type. The information is available as captured. Bush does not explicitly speculate on the value, however, of being able to get at the structured properties of the data captured, such as kingdom or class of a photographed organism or the fact that 27-6-45 is a combination number not a date. But again, implicitly, Bush's quest for automation of repetitive thought practices and retrieval of assets when needed both beg the question, well then, why not do so via the metadata of a captured artefact? It is in the structure of the data, identifying one string as type meeting and another as type person or type phone number or type musical inspiration that lets us carry out queries like "what were all the phone numbers I recorded when I was last in the office at X?" Such retrieval would potentially improve upon what is possible to do with even the best notebook: it would make it possible to query the captured information from a multiplicity of associative contexts. The challenge for such a system becomes how might we combine the easy interaction of notebooks or even Bush's more advanced voice and image field recorders with the rich capabilities afforded by structured data capture? To capture data structure currently, we must use separate forms in usually separate applications that share data and data structure often grudgingly. The rapid input of the notebook is lost.

Enter the Semantic Web as both personal and Web Scale data mechanism. By using Semantic Web technologies like RDF for data representation and triple stores as knowledge bases, data can be shared in a single "data soup" as the Apple Newton used to refer to it[9], where the data in the soup is accessible to all applications on the platform. By using either lightweight grammars (what natural language experts refer to as "pidgins") it becomes feasible to capture data structure from idiosyncratic data entry of text strings. A string like "meet w Ch. @ 6 re jourknow" can readily be translated into a calendar event to be associated with notes on the project jourknow and referenced to Chiang as the person involved in the meeting. We have described this process elsewhere [31].

The advantage of automatic structure extraction to a shared data source means that data can be explored in its native context, such as the note it was when entered, or from a variety of other contexts, such as activities that took place at the time it was created or locations used or as it relates to a particular activity or project, or as a marker to what other documents were being worked on when that note was created. Time and location are easy details to capture from wireless devices; document state is also tractable. Using the same protocols for association, external services can be developed to support these local contexts: in an academic context, for example, relevant conferences may be found that relate to areas of work for particular projects, and deadlines scheduled automatically. Awareness of others working on similar projects can also be discovered, and their related work captured. These kinds of automatic or semi-automatic associations with external data sources enable the notebook space to retain the easy affordances of the physical model while going beyond the physical limitations into the benefits of a networked computer with access to Web scale data. In this respect, we do not slavishly copy the page model of the notebook, but rather as Dix suggests [9] endeavor to capture its affordances, its experiential qualities. We then enhance them with these Semantic Web technologies.

5.3 Note Cards Redux: Even More Hypertext

A compelling affordance of going digital indeed is that we can deploy a variety of representations for the same data, and take advantages of the affordances they offer. While the notebook is a well used, well trusted mechanism for keeping notes together, it does have limitations: page binding enforces linearity; it is difficult to see page 6 next to page 36. A well-studied model for idea capture that breaks that linearity is the notecard stack. Indeed, one of the earliest hypertext systems, NoteCards [13] used the notecard stack as a model for idea capture and reordering. This work was to be followed by the commercial and pre-Web Hypercard and Supercard applications**.** The cards not only contained data, but links *and* functions. There were also specific data types assigned to card types. Hypercard defined these cards very explicitly: the Home card, address cards and so on. Cards within card stacks could be visited either sequentially or arbitrarily. Spatial hypertext systems from VIKI [18] to Tinderbox [5] have also capitalized on the the affordances of card stacks, but added another affordance from the physical realm of cards: the ability to spread out and reorganize virtual card stacks, where space in their organization communicates a kind of meaning – at least to the author of the structures. Tinderbox also adds AI processing and data mining to extract new kinds of information and associations from the local data in the cards.

The history of card use and structuring of cards comes from a well-designed practice of card use in pre-digital scholarship and journalism. In this research model, there were three kinds of cards: idea cards, quotation/paraphrase cards, and bibliography cards. These cards are interlinked: quotations to citations; ideas to either. These cards could be created in any order as material was discovered or ideas occurred: "only one idea to a card; only one quotation per card; only one reference per card" were the only constraints on card use. The idea being of course that individual cards could be organized and reorganized spatially for getting a picture of the developing paper. Not all cards would be used, but gaps could also be detected. The organized cards could then be

[9] "Data Soup", Apple Newton entry, Wikipedia. http://en.wikipedia.org/wiki/Apple_Newton

put into one pile, and the paper written effectively from iterating through the cards one at a time. Indeed, an outline for the paper or chapter could be generated from the organization of the cards before proceeding to the paper writing.

The relevance of the note card model to the concept of the Semantic Web as personal work space with associated public data is in the integration of personal ideas with external sources: the idea cards are backed up with/informed by the quotations from external sources. In the case of note cards, these associations are either manually created by the researcher/author, or are presented by (and thus attributed to) another author. The goals are the same: building new knowledge by capturing one's own ideas, and working with those of others - whether these are ideas that come up in a conversation with others and are hastily jotted down, or are captured from a published source. There is interplay here, a making of meaning. Mark Bernstein's Tinderbox software very much follows the note card paradigm to support just this kind of intermix activity between the card stack, the card layout, and capture of ideas and other sources. It enables links to be copied from the web into cards, and of course enables other kinds of data to be written into the cards. It blends capture of the external with capture of the personal. Digital notebook software, like Circus Ponies's Notebook, supports live capture of web content into a notebook page, and provides a single, knowable source for keeping track of digital ideas, whether as short bursts or longer thoughts. However that tool is currently locked to the paper page concept of the Notebook page metaphor. Based on the benefits of these various types of representations for our information, our tools need to provide multiple representations of the information – from pages, to cards, to timelines, to maps to facetted browsers, to whatever mode – to best support this *work in progress* paradigm.

6. NOTEBOOK+MEMEX=HUMAN FOCUS

Setting issues of particular embodiment aside, whether of discrete cards or sequential pages, it is the affordances of the analogue notebook/note card stack for developing and progressing ideas and for interleaving idea content with Memex-like associations across newly discovered, richly associated work that can stand as a tractable analogue for the Semantic Web. The Semantic Web = Notebook+Memex.

One may argue that the Memex is still to unfamiliar a concept to be useful, but this is the "something new" part of the "Page+Links" "Familiar+New" equation for introducing a new technology. There was a time when Links were Something Very New to the general population as well, and that the demonstration of how they worked quickly clarified their role. In this case, the Memex is the means to help make, discover or recover contexts and connexions among work in progress at any point in the "creative thought process" from quiet self-reflection and engagement with related work and making associations among and between therein, to more broadly sharing material for in progress feedback. While the "+ Memex" reflects this movement between the local and the network/web, the "Notebook" component reflects the very active, yet very personal process of what has become known as knowledge working.

The notion of the notebook (the blank page as opposed to the published page) is also different from what the Web has become while still obviously being on the same continuum of work in progress towards some kind of sharing/publication. This blending of personal use with the Semantic Web's potential for automatic association of associated resources (whether personal or published, local or global) is a significant shift in how most of us have been thinking about the Semantic Web. Let me frame that last statement. There have been projects thinking about the Semantic Web desktop - using the Semantic Web as a personal or local server layer for data.[10] The projects foreground that there is value in applying Semantic Web protocols to the local context. There have also been projects like myTea[11] which have imagined using Semantic Web technologies to maintain transparent context histories [25] as a way to generate a dynamic, annotable bioinformatics experiment record (if not lab book) to track and record bioinformatics experiments as they develop acroos the variety of local and web tools used. The bioinformatician does not have to make a record of each step they take with their digital data; the system creates the record for them. At any point they can annotate or interlink the record of actions carried out.

What is proposed here as a model for the Semantic Web not as Desktop, not as an over-arching environment but as Notebook + Memex goes in a somewhat different direction as a model for the Semantic Web than what is written on Wikipedia. We have already said that the page cannot reflect the rich associative possibilities of what the Semantic Web promises so one may ask, how could the analogue of a researcher's notebook which is so idiosyncratic support this concept? The notebook in this context is meant to force several concurrent concepts. First, there is the focus on lightweight data capture. It is critical that we re-investigate input methods, which means that we must also re-investigate data storage. Right now the needs of the system to have structure captured manually have forced dreadful form-based user interfaces. We have the knowledge to do better. From filling is exactly the kind of repetitive task that a machine is well suited to carry out and leave us to the creative process. If we want light weight data capture and rich data structure, this is a challenge we must address. Second, the notebook is an active repository: notes, images, pictures are frequently taped into them as are references to other documents. The semantic backing of the "+ Memex" components of the notebook enables the possible interconnections – the lines between notes, the calculations across points, the paths across domains – to be developed and maintained. Likewise, the single data soup of the Memex repository means that data can be shared easily among a rich variety of representations. Tim Berners-Lee's Tabulator [2] attempts to provide just such a flexible set of views on RDF sources that have been brought together and queried: the results can then be represented in whatever view is most appropriate: table, calendar, map, or in time, hybrid views.

One of the core attributes of this notion of the Semantic Web as notebook + Memex is that it situates the Semantic Web conceptually within the realm of human engagement where we are actively "extending the record." Right now, very few Semantic Web tools, whether mSpace, Haystack or Tabulator support direct authoring. With a Semantic Web (or Memex) – backed Notebook, we can imagine the Semantic Web components regularly seeking out associations to support the researcher's process. Where the mighty Tinderbox works to develop these connections among the local Tinderbox-specific entries, a Semantic Web enabled notebook could draw across any local data source (associating active documents with working emails and appointments, for

[10] http://www.semanticdesktop.org/xwiki/bin/view/Main/

[11] http://mytea.org.uk

instance) with related (Semantic) Web sources. This local/personal focus is a compelling kind of inversion of the usual models of the (Semantic) Web. Instead of an emphasis on publishing for the World Readable Web, we are emphasizing the pre-publishing, ingesting, personal activities of work, of active personal process rather than finished, public end. By this approach, we include the whole continuum of activity, not just the end point of the processes Bush clearly imagined in leading up to the public "extension of the record."

7. CONCLUSIONS AND OPORTUNITES

In this paper I have suggested that we need a tractable model of the Semantic Web in order to enable people to imagine not only how it can work for them, but how they will want to design tools to support that vision. The proposal is that we can look to the Web's analogue as a model for framing one for the Semantic Web. The Web has been postulated as a familiar technology with a new technology: the printed page + links. I have argued that a similar formulation for the Semantic Web is a Notebook + the Memex. In both the familiar notebook, and the more visionary Memex, the emphasis is on engaging with information, developing it, working with it, as *work in progress*. While the Semantic Web can be seen to provide the protocols to enable the Memex to support dynamic and automatic associations across inter-related domains, the notebook emphasizes both the more writerly and the more personal side of engaging with information.

I have also suggested that this personally informed conceptualization of the the Semantic Web has the potential to lead to a different computing paradigm that may be more effective for human interaction, and may take better account of how we should by now be able to engage with computers, rather than computers forcing us to suit them (yes, this is a call to kill the form, and be liberated from it). Another way to imaging the paradigm proposed is partially captured by the interaction with the Computer on Star Trek, Next Generation**.** It is conversational: it is an ebb and flow of generating and validating ideas with the Computer, and merging these into new answers that are then shared with others (members of the Enterprise still go to conferences and present papers). Except for the voice interface, this model of computer interaction is very much like what the Memex describes with its scientist in the field, and what is proposed here as the Notebook+Memex: the personal working out and evolving of ideas towards a solution. The difference between Star Trek and the Memex is that the Computer is more actively engaged in assisting with data retrieval and calculations. This level of assistance is becoming possible via the logical structures supported in the Semantic Web's protocols. Another critical observation of these two models, both Star Trek and Memex, is that forms are only implicit. For instance, on Star Trek, no one says "Open calendar: date, march 3, event: meeting with Cmd. Riker, start time: 1300, end time 1400." At most they provide tags, saying, "Captain's Log" for instance, to initiate an entry. Likewise who makes the entry is captured from the context of the voice and location of the speaker. Captain's logs are then able to be pulled together on demand, to support queries such as "what else was going on in Sick Bay when I made my log?"

The one thing missing from these visions of the future computer is the social networks of data sources that are of current and of pressing interest to many considering the shape of the Web [3]. In a way, the Memex was sensitive to the social in its consideration of the numbers of people who would contribute trails through data, sharing their associations for reuse and re-interrogation. This social immediacy enabled by the internet is fostering perhaps a new paradigm for both computing itself and what may constitute "publication" at earlier stages, that supports models to which sharing work in progress. We already have a form of this intermediary publishing of results in the e-Science space: chemists are publishing crystal structures as they are generated in eBank[12]; bioinformaticians likewise daily add to databases of genes. Each source is used regularly as a key resource by other scientists. Little of the data in these repositories has first been published in formal journal papers. The role of direct experimental results being available for comparative consideration is taking on a bold new prominence in science work, above and beyond the formal primary research presentation of a peer-reviewed paper.

If we believe that this intermixing of voices and intermixing of idea generation represents an important set of axes and continuums to support, then our vision will need to be for tools to support these kinds of interactions – interactions we carry out regularly in the physical world, but that are less well supported in the digital space. Again, therefore, tools to support the in-process generation of ideas, to support the ready inter-relation of concepts, are critical for the next model of interaction with these systems. This interest in new models of computing, or of interacting with computers also emphasizes creativity as a necessary component to support in the design of the interaction. As Shneiderman has pointed out [28] we currently have little understanding about how to support creativity directly: what exactly in a tool set improves achieving an "ah ha" moment? How do we evaluate the strength of this feature? And yet creativity, the achievement of an insight that provides a new path to solve a problem, is a fundamental part of the scientific, process, or any research enterprise. One might postulate that the freeform nature of the notebook is an established tool in the support of creativity in the discovery process. If that is so, which attributes? How can they be understood to be directly and effectively supported digitally?

A question of moment may be, therefore, do we want to challenge ourselves to take as a fundamental goal designing systems not just to support a particular task, but to support creativity? Such a challenge takes most of us out of our comfort zone of known approaches for design, validation and the perceived role of computers for "productivity." Surely, though, these are the kinds of challenges we are now ready to ask of the systems we develop, whether at a high level of formal hypertext models or on the ground of embodying, for instance, Semantic Web enabled systems. Perhaps such challenges will become part of the agenda that Web Science will embrace. Perhaps the Notebook + Memex = Semantic Web is one approach to help us get there.

8. ACKNOWLEDGMENTS

Thanks to the kind folks at the School of Information, University of Texas at Austin who first engaged with me on this topic. Thank you as well to the Semantic Web User Interaction Steering Committee who provided feedback on the blog version of this paper, and to the thoughtful reviewers of this HT paper whose comments have been key in sharpening the current presentation. I hope this version is closer to their vision of the work.

[12] http://www.ukoln.ac.uk/projects/ebank-uk/

This paper is an outcome of work under the aegis of the Web Science Research Initiative, supported via both an EPSRC Overseas Travel Grant, EP/E035930, and a Royal Academy of Engineering Global Research Award.

9. REFERENCES

[1] Bansky, "Another Crap Advert" Image held at Art of The State Web site, http://www.artofthestate.co.uk/Banksy/banksy_another_crap_advert.htm. Accessed from Jan-May 2007.

[2] Berners-Lee, T. Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D. Tabulator: Exploring and Analyzing linked data on the Semantic Web. SWUI06, swui.semanticweb.org/swui06.

[3] Berners-Lee, T., Hall, W., Hendler, J. Shadbolt, N., Weitzner, D., Creating a Science of the Web. SCIENCE 313.5788 (August 11, 2006): 769 – 771.

[4] Berners-Lee, T., Hendler, J., Lassila, O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*. May 2001.

[5] Bernstein, Mark. An Apprentice that Discovers Hypertext Links. Hypertext: Concepts, Systems and Applications, Proc. of ECHT'90. A. Rizk, N. Streitz and J. Andre, Cambridge Univ. Press, 1990: 212-223.

[6] Bernstein, Michael., Van Kleek, M., Karger, D. and schraefel, m.c. (2007) Information Scraps: How and Why Information Eludes our Personal Information Management. Working Paper. http://eprints.ecs.soton.ac.uk/14231/.

[7] Brody, T., Harnad, S. and Carr, L. Earlier Web Usage Statistics as Predictors of Later Citation Impact. Journal of the American Association for Information Science and Technology (JASIST) 57.8(2006) pp. 1060-1072.

[8] Bush, V. As We May Think. *Atlantic Monthly* July 1945. http://www.theatlantic.com/doc/194507/bush.

[9] Dix, A. Deconstructing Experience - pulling crackers apart. Funology: From Usability to Enjoyment. Kluwer Academic Publishers, Dordrecht, 2003, 165-178.

[10] Gray, J. Turing Award Lecture: What Next? A dozen remaining IT problems. Turing Award Lecture, 1999. http://research.microsoft.com/~gray/talks/Gray_Turing_FCRC.pdf

[11] Halaz. F. Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems. Hypertext 87 and CACM 31.7(July 1988):846-855.

[12] Houston, R.D. Harmon, G. Vannevar Bush and Memex, ARIST 41(2007): C2.

[13] Hughes, G., Mills, H., de Roure, D., Frey, J., Moreau, L., schraefel, m. c., Smith, G. and Zaluska, E. The semantic smart laboratory: a system for supporting the chemical eScientist. Organic and Biomolecular Chemistry 2 (2004): 1-10.

[14] Jones, W. Personal Information Management. ARIST 41(2007): C10.

[15] Jourknow project site. http://projects.csail.mit.edu/jourknow/

[16] Kalnikaité, V. Whittaker, S. Capturing life experiences: Software or wetware?: discovering when and why people use digital prosthetic memory. CHI 2007: 71-80.

[17] *Liber Chronicarum*, George Khuner Collection. The Metropolitan Museum of Art (on line): Timelines of Art History, http://www.metmuseum.org/toah/hd/prnt/ho_1981.1178.29.htm. Accessed Jan 2007.

[18] Marshall, C., Shipman, F., Combs, J. VIKI: spatial hypertext supporting emergent structure. HT 94: 13-23.

[19] "The Most Famous Poster." American Treasures of the Library of Congress, Memory Exhibit, Online. http://www.loc.gov/exhibits/treasures/trm015.html, accessed Jan., 2007.

[20] Nelson, T. Literary Machines, Mindful Press, Sausalito, California, 1981.

[21] Quan, D., Huynh, D., and Karger, D. Haystack: A Platform for Authoring End User Semantic Web Applications. Proc ISWC 2003.

[22] Richer, J.P. The Notebooks of Leonardo Da Vinci, Vol 1. New Ed (edition) Dover Publications, 1989.

[23] schraefel, m. c., Zhu, Y., Modjeska, D., Wigdor, D. Zhao, S. Hunter Gatherer: Interaction support for the creation and management of within-web-page collections. Proc WWW (2002): 172-181.

[24] schraefel, m. c., Hughes, G., Mills, H., Smith, G., Payne, T. and Frey, J. Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment. In Proc. of CHI 2004, 25-32.

[25] schraefel, m. c., Brostoff, S., Cooke, R., Stevens, R. and Gibson, A. Transparent interaction; dynamic generation: context histories for shared science. In Proc of ECHISE 2005, Munich, Germany.

[26] Seniors Online Increases. Sec: Seniors Statistics. *Seniors Journal.com: Senior Citizens Information and News*. Feb 4, 2003. http://www.seniorjournal.com/NEWS/SeniorStats/3-02-04SnrsOnline.htm. Accessed Jan 2007.

[27] Shadbolt, N., Hall, W., Berners-Lee, T. The Semantic Web Revisited. IEEE Int. Systems. May-June (2006):96-101.

[28] Shneiderman, Ben. Leonardo's Laptop: Human Needs and the New Computing Technologies. Cambridge: MIT Press, 2003.

[29] Siegle, David. Creating Killer Web Sites. 2nd Ed. New York: Hayden Press, 1997.

[30] Stevens, R., Tipney, H.J., Wroe C., Oinn, T., Senger, M., Lord, P. Goble, C., Brass A, Tassabehji, M. Exploring Williams-Beuren Syndrome Using myGrid. Intelligent Systems for Molecular Biology (ISMB), 2004.

[31] Van Kleek, M., Bernstein, M., Karger, D. and schraefel, m.c. (2007) GUI— Phooey! : The Case for Text Input. UIST, 2007, Rhode Island, USA (in press).