# 1 Developments since 1990

## 1.1 Print-on-demand

In 1990, when the first high-speed digital copiers came on the market, it appeared as if we were at the dawn of a new era in publishing. The new machines were hailed to bring about a revolution in the publishing industry as radical as Gutenberg's introduction of the movable character press.

The traditional printing industry requires large capital investments, long print runs, large warehousing facilities, and careful printing volume forecasting. Compared to traditional presses, high-speed digital copiers are an order of magnitude cheaper, can print cost-efficiently runs as short as a single copy, have small warehousing requirements for the digital storage media, and eliminate the risk of volume forecasting. The new machines have been very successful in the market place, both for in-house printing and in traditional print shops.

The traditional printing press manufacturers have quickly responded to the new market situation by introducing direct-to-plate presses, where the pre-press and plate-making workflow steps are carried out completely in the digital domain. These machines are finding a fertile market for short print runs, which cover the range from 1000 or less to 10,000 copies, and require faster printing times and more sophisticated finishing processes than printers based on copier technology can offer.

In the last few years there has been considerable technological progress in color POD. However, systems based on this technology have not been as successful on the market as the black-and-white counterparts five years earlier. One reason is that color printing is much more complex than just pipelining four monochrome print stations. The problems are not just controlling registration, a matter of understanding how the mechanical controls on the printer work and of reducing vibrations. Electrophotography is a physically unstable process due to triboelectrical charge effects and complex interactions when a first printed separation influences the next separation through image dependent changes in the electrostatic fields. Controlling such problems requires not only a deeper understanding of color reproduction but also a very robust process.

The market breakthrough of color POD has been hampered by the lack of knowledge dissemination required by the new workflow for digital color reproduction based on colorimetry, careful characterization and calibration of each process stage, and the application of new image processing techniques such as stochastic screening. The pool of experts in the new technology has not achieved a critical mass.

## 1.2 The Internet frenzy

The time when the critical mass is reached may never come. This is because the Internet and the World Wide Web ($W^3$) are bringing about a new paradigm shift before conventional POD has a chance to get established in the color printing business. The

new paradigm is to eliminate the capital costs by having the readers printing the material on their own equipment. The storage costs remain, but distribution costs are paid directly by the readers and the length of a run is no longer an issue because readers print their own copy—the print job itself is distributed to the end user.

This paradigm shift has widespread consequences for the hard copy business. Instead of selling few expensive machines to business people that base purchases on cost/benefit analysis, printer manufacturers will sell many inexpensive machines to professionals and households that base purchases on perceived values. The margins on personal printers are razor thin, but these printers are operated only for a small fraction of the time and are replaced early due to technical obsolescence well before maintenance no longer is economical. Thus, to produce the same amount of printed material, much more print capacity will be deployed.

Not only the print process will be distributed, but also the actual print consumable cost will be distributed among the printing consumers. When a publisher must consider the cost of a run, a page cannot cost more than a penny, but readers are willing to spend a dollar or more on each page produced on *their* printer. Thus, the dollar volume on inks and papers may increase by two orders of magnitude, while at the same time higher profit margins are possible on each consumable unit.

### 1.3 Publication budgets

Companies have substantial budgets to print promotional material like junk mail, especially when they sell a mature product that is different from a competitor's mostly in virtue of a company's image. Since consumers will print a specific brochure only when they have a definite interest in a product, a printed page will have a much higher impact on a company's sale. It is conceivable that companies could start offering coupons not to buy their own products, but to buy printers or consumables on which their literature prints well. Similarly, the publisher of an electronic magazine might give coupons for printer consumables or for connect time to increase the magazine's readership and thus be able to charge more to advertisers.

In the case of books the change might be much slower. On one side readers do probably not have the patience to wait for an entire book to be printed on their slow personal device and the cost of consumables is more visible. On the other side, bookstores are not only places that sell books but are also a social place where people gather and see fellow humans. Experience with video-on-demand field trials show that people prefer to drive to a video rental store where they can see other people, rather than down-load a movie in their family room. For lengthy documents the market is probably more inclined towards faster departmental printers used to print manuals and the like.

### 1.4 Goal of this work

The critical element for this scenario to be successful is the establishment of a viable workflow. From the above it is clear that the new paradigm of individual printing* is so radically different, that the traditional workflow does not apply at all. The problem with color POD is that the required workflow requires too much skill and mastery of the technology. In this paper I will delineate a general methodology that makes printing straightforward, allowing individual users to transfer all information they wish from the digital medium to convenient paper. This proposed technology is disruptive and cannot readily be understood as an evolution from current methods, so I will digress in an attempt to provide as much background information as possible.

The Internet is a network and a number of protocols for transferring data over the network. The various protocols are used for different applications, such as e-mail, file transfer, news groups, multi-user dungeons, the World Wide Web, etc. I will focus on the $W^3$, because this is the application that has become a publication medium for the masses.

Printing long documents is hard because $W^3$ pages are linked but disconnected: on many Web sites the structure is poor. A long document is not posted as one HTML file, but broken up in a hypertext document consisting of many HTML files. To print the entire document, the user must mentally linearize the document and visit each file to print it. We will see later how this problem can be solved elegantly by structuring $W^3$ sites in a way they can be linearized automatically in a natural way.

## 2 1996: the hotting-up of the $W^3$

Originally the $W^3$ was conceived as a collaborative tool, allowing a number of scientists to create and manage a body of knowledge. While in previous network-based hypertext systems all data was encoded using the ASCII standard to achieve minimal common interchange capabilities, the key idea in the $W^3$—and this is the main cause for its success, coupled with a graphical user interface (GUI)—is to use a markup language to allow for text properties and looks, splitting the tasks between author and reader as follows:

- the author decides contents and structure
- the reader decides the appearance

This contrasts to the traditional publishing industry, where the author/publisher decides contents, structure, and appearance. Moreover, appearance is a very important element of a publication, which determines the publication's image and sets it apart

*. The similar term personal printing refers to a local printer used to print documents created on the computer to which the printer is attached. Individual printing refers to distributing a print job to the users for printing at their discretion.

from other publications. In the original embodiment of the W3, collaboration is implemented by each team member maintaining an own personal Web site and creating external hypertext links to the other member's sites in addition to the internal links.

In the traditional publishing model, the author, redactor, or publisher communicates information to the readers; there is no dialogue other than a relatively small number of letters to the editor. As Soren Kierkegaard had discussed with fervent ardor in the middle of last Century when the newspaper publishing industry went through a similar turmoil as the W3 now, the press appeals to people's æsthetic sphere of existence, leveling all information so that nothing is too trivial or too important. This flattening encourages everybody to be interested in everything, to accumulate information and postpone decisions indefinitely. The correspondent in the W3 is to encourage people to surf around and just enjoy the abundance of the new medium.[4]

It is no surprise then, that during 1996 we have seen many efforts to hot-up (in McLuhan's sense[10]) the W3 as a medium. In standardization bodies, such as the Internet Engineering Task Force[9] (IETF) and the World Wide Web Consortium[16] (W3C), the emphasis has been on constructs such as style sheets and font embedding, which allow authors tight control over a document's appearance and to produce graphically compelling high-concept W3 pages.

In the commercial arena, during 1996 we have been spectators to a fierce battle between browser vendors, who have pitched as key differentiators elements for controlling appearance, such as style sheets, fonts, and tables. Another battle ground has been the distinction between the pull model and the push model. In the pull model consumers surf the net and download those pages in which they are interested. In the push model, consumers furnish their address and interest profile to a publisher or provider, which then automatically sends pages to the consumer.[14]

### 2.1 1997: the W3 meets Occam's Razor

With the traditional hot media publishing model, the reader is assumed to be a passive potato-couch surfer (or a person enjoying the æsthetic sphere of existence, to formulate it positively with Kierkegaard) and W3 publishers follow models similar to the print and TV media, which have skilled sales departments for advertisement space. In the new Internet lingo such services are called Web hosting services, and WebTV[15] is a good example of a successful implementation of this model. The Microsoft Network[11] (MSN) is organized in channels offering different contents, similar to the television, two which are magazines (Slate and UnderWire).[13] The MSNBC[12] service has the format of a magazine, with tie-ins to television programs.

This hybrid magazine/news-service/television model is a temporary phase that allows the new providers to become familiar with the medium. It will soon be replaced by much more sophisticated programming that fully exploits the medium. One thing is clear: the Internet (the plumbing for the W3) has matured over 30 years, is well un-

derstood, and many people are familiar with it. This allows large corporations to immediately jump on the technology, bypassing the usual phase of the pioneers. One consequence is that a Web media product can be pitched directly to the broad masses.

Another consequence is that the financial stakes are immediately very high. For example, Corbis[6] is purchasing or licensing a considerable portion of the human visual heritage and selling it over the $W^3$ for a fee. Therefore, an author desiring to use an image created by a great master can often no longer obtain a license from a museum, but for many images has to buy it from Corbis. Although Corbis may offer non-exclusive contracts for those works of art it does not buy, owners may not want to jeopardize they relation with a good customer. The creators of hot Web sites have embraced high-concept production methods, which require high initial investments but then generate a substantial revenue stream.

Finally, because of the simultaneous exploitation of a number of more conventional media technologies, the production requires a very eclectic staff and directors capable to harness this diversity necessary for a high-concept design. The cinematographic industry is the one that possesses to the largest degree this kind of skills. Obviously the big players are looking to Hollywood for contents producers; for example, MSN is ready to spend $300 to $400 million for contents.[2] Although there are legions of creative people that are ready to create for the new medium, the maturity of the technology mentioned above and very big war chests allow the big players to simply go and buy the best content.[3]

With the stakes involved, and interpreting the direction in which the standards are evolving, it appears that this year Occam's Razor will be applied to the $W^3$ and only a few large media giants will survive in the World Wide Web business as a medium. We are seeing a large number of alliances between the giants of the computer industry and the large Hollywood studios, which because of the exclusivity and the possible synergies will rise considerably the entry price for new players. Like during the Gold Rush the sellers of mules, tents, jeans, and shovels made good profits, also in the $W^3$ there will still be many opportunities to make money by keeping the plumbing running.

### 2.2 Impact on color hard copy industry

This trend in the development of the $W^3$ from a collaborative tool for scientists to a hot mass medium has deep consequences for the color hard copy industry. The new medium brings high-concept, very high quality graphical productions in the homes of passive viewers thirsty for information. This contents is fleeting—commercial $W^3$ based media are updated continuously and consumers will not be able to easily retrieve a piece they have seen in a previous session. Even non-commercial pages change rapidly, with the operators of the Alta Vista search engine observing an average lifetime of only 45 days for an URL.

Paradoxically, the new electronic medium will potentially spurn the creation of even more printed material. However, this time the printing no longer occurs in centralized high-efficiency centers such as print shops, but on a myriad of distributed printers in people's homes. For manufacturers this means that the future is not with short-run presses, but instead with very inexpensive convenience printers. Because each consumer prints only few pages on each session, the consumable prices are not critical and profits have to be shifted from service and capital equipment to paper and ink. The high-concept design used in pages created by the media giants will encourage readers to use profusely premium paper and color inks.

The unexpected consequence is that color fidelity as it is understood now will no longer be relevant. The past trend in color reproduction has been towards tight control of color and the aim for absolute fidelity between an original and the reproduction. Due to the pressure on the price of capital equipment, color printers (as well as display controllers and CRTs) can only become more unstable. Moreover, it would be an oxymoron to expect an allegedly passive viewer to spend much time to calibrate devices or even set a correlated color temperature on a display monitor.

The new emphasis in color reproduction must be on color robustness, not color fidelity. This means that color information must not "fall apart" if the conditions change. For example, in a catalog shopping experience, it is not possible to reproduce colorimetrically the color of a sweater. Nevertheless, a shopper must be able to coordinate a skirt with the sweater. The new emphasis will be on color palettes that change in unison, not on single colors that remain static. This will bring a revival to color reproduction paradigms as they were first introduced scientifically with the retinex model.

## 3 Intranets

The $W^3$ is more than a publication medium for the masses. Over the past decades most large companies have built extensive private networks (intranets) to run their businesses. These networks have been built around proprietary protocols, such as SNA and XNS. One result of the Internet frenzy is that companies are switching to the standard TCP/IP protocols to get an instant upgrade to a modern GUI, while at the same time saving costs by using off-the-shelf applications. The developments in this arena are less visible than those in the consumer market, but it is nevertheless a substantial and very profitable market.

Due to their more pragmatic function, intranets will not sport high-concept designs like the mass-market media, but manager's experience with the media as private citizens will nevertheless rise their expectations on the look and feel of their internal Web sites. A number that is often quoted is that on the average it costs $80 to create a Web page and $1 to serve it. In this author's experience a commercial Web site that is updated daily has a full-time staff of 12 to 50 people and commercial Web site updated quarterly requires 6 professionals for 3 months during the creation phase and

$^1/_4$ head-count permanent staff for maintenance. It appears that a more realistic number for the creation and maintenance of each intranet site is likely to be in the range of a quarter million dollars. With a large company having dozens if not hundreds of internal Web sites, sooner or later there will be pressure for lower-cost authoring tools.

For the hard copy aspect, intranets are used to serve a large amount of lengthy and complex documents, such as procedures, forms, manuals, etc. Businesses use faster printers for work groups and there is a pressing issue for employees to be able to print complex documents with as few and as simple as possible commands.

Before presenting solutions to the Web site authoring and maintenance problem and to the printing problem, I want to briefly digress on the possible emergence of new scenarios and usage models of the $W^3$, which will push the emphasis of the proposed solution more to the side of that required for intranets, based on the observation that we human beings are active, not just contemplative.

## 4 Intellectual property

Progress is not linear and cannot be forecast by extrapolation. The above analysis of current events does not mean that we should now resign ourselves to learn debug plug-ins or analyze router traffic, if we want to remain in the Internet business. Two classes of events can change the card game: the understanding of copyrights and the social changes coming about while we are moving at a fast pace into the post-industrial era.

The traditional media companies like Walt Disney[7] as well as the new companies like Corbis have a very high stake in copyrights, because they derive large profits from royalties and merchandising. Copyrights are exclusive rights granted primarily to reproduce and distribute copyrighted work and to prepare derivative works. It is important to understand the difference between an idea and its expression. Ideas and facts cannot be copyrighted, it is the expression that is protected.

The First Amendment and its equivalent in most of the nations's constitutions gives the right to people to freely communicate their ideas, which entails unhampered access to the $W^3$. The Internet, a cellular information infrastructure designed to withstand a total nuclear war, is uncontrollable and uncensorable by its very nature. In the latest intellectual property discussions in Geneva, it was agreed that the act of viewing a document on the $W^3$ does not constitute a copyright infringement.

Consider the particular situation that copyright is based on the concept of expression, or a physical manifestation (container) for ideas. With the disappearance of the physical manifestation as an industrial good and its transformation into a constitutional right, companies based on selling containers for ideas will be in trouble, as we as a society return to the original reasoning behind the copyright law that what has value

are not ideas or their expression, but one can do with them. The $W^3$ renders the utilization of an idea intangible. This is a return to a pre-industrial model, where abundance is more valuable (size of a flock of sheep or harvest), in opposition of the industrial model where value is in scarcity (exclusivity, like diamonds or fewer permanent employees).

The free flow of ideas, coupled with the convenience of paper as a carrier for personal use, has the potential of creating a bonanza for the personal printer market. This bonanza is independent from the outcome of the $W^3$ as a new publication medium like the press, radio, or television.

### 4.1 Social changes and knowledge

The advent of the $W^3$ as a dissemination medium for ideas is not the only deep social change that we might experience with the transition to the post-industrial society. The economic turmoil started with Reagan's and Thatcher's trickle-down economy and progressed in the change of large corporations from diversified conglomerates to mean and lean companies focussed on a core business, while eliminating middle management and assigning production to contractors (buzz-word: virtual corporation), is now resulting in deep change in our society's values. While in the industrial society authority (in the etymological sense of auctoritas—the right of giving orders conferred to those who have the knowledge, skills, or experience) was the main driving value for career advancement and the educational system, in the new post-industrial society authority is replaced with power (connections and networking).

In consequence, the value of achievement is being replaced by the value of celebrity, doubt by certainty, and science by magic. In this new society individuals no longer seek for a career path in a good company, but are now taught by their companies to profess career self-reliance. Instead of a well educated work-force of employees, the post-industrial society consists of a large pool of expert contractors. The difference to the pre-industrial society of independent professionals and artisans might be that in the post-industrial society individuals will not sell their services directly to consumers, but mostly to large virtual companies that own worldwide distribution channels.

This new modus operandi is becoming possible because the $W^3$ allows each individual to utilize the collective human knowledge published on the $W^3$ and in turn contribute the personal knowledge, an act which advertises the individual's skills and mastery, thus attracting customers. This model is robust with respect to the undermining of copyrights as we know them now, because the value is not in the ideas or information, but in the displayed mastery of creating knowledge, which is the utilization of the ideas. In the era of deregulation, privatization, and self-reliance, the $W^3$ will become the ultimate social survival tool.

Some argue that the solution to the levelling problem of the $W^3$ is to invent technologies that allow people to stand out vertically, i.e., to develop information technologies that support strong identities in the sense that they can create high-concept Web pag-

es with little effort. Proponents for this idea encourage the creating of a plethora of plug-ins or to use HTML pages as wrappers for calls to middle-ware or to "webify application suites." In reality, such activities are at an epidermal level and still belong to the æsthetic sphere of existence and do not support strong identities. I believe the important tools are those that help the individual (or a work group or clan for that matter) to distill information into knowledge, for which in this discussion we can use the following traditional definition:

Entity *A* knows *p* if and only if

1. *A* believes that *p*,
2. *A* is justified in believing that *p*,
3. *p* is true.

Based on this view, knowledge is justified true belief. For this discussion belief is based on experience, skills, and mastery, i.e., I am interested in knowledge in its strict form of propositional knowledge—how to do things—rather than an accumulation of facts. Skills and mastery ensure that the belief is acquired by experience, ruling out that a belief meeting these conditions is possessed by sheer accident.

In contrast, the creation of technologies supporting strong identities can be reduced to graphic design tools and the willingness to hire a graphic artist to design one's Web site. In this sense, from the point of view of the social impact of a technology, we can draw a parallel to Aldus Manutius and his creating of the book publishing business.

### *4.1.1 Book publishing*

Indeed, these are times similar to Venice in 1500, when Aldo Manuzio (*alias* Aldus Manutius, 1452-1516) joined a printing business and became a publisher. The adoption of Gutenberg's printing press, the brain drain of scientists from the collapsing Byzantine Empire bringing with them the Greek classics, an educated population that could read Greek, and a flourishing spice business whose profits allowed ordinary people to afford† books; all of these contributed to a vibrant, innovative environment.

Manutius formulated a key idea that made him a main contributor to the Renaissance: he became a publisher instead of a printer.‡ He searched for material and selected what he thought might have the largest readership. When he came across a classic text he thought might appeal to a wide audience, he had it translated from the Greek and published in Latin, and when he thought he might have a best-seller, he would even publish it in Italian, the common people's language.

---

†. The price of Manuzio's books was about a teacher's day's salary. Before, books could be afforded only by princes and wealthy monasteries. Compare this with the price/performance development of computers and software.

‡. The press was owned by an established printer, Andrea Torresano. Manuzio managed the printing shop, selected the texts to be published, made editorial decisions, and arranged for the marketing of the books.

The lesson from Manutius is that although one can become rich by working hard operating a printing press, one can become wealthier by working smart and exploiting new emerging technologies (embracing and extending). Technology is just an enabler; paradigm shifts have more to do with social values. In Manuzio's time, one major problem was the sheer size of books. He came up with some technological solutions, like inventing the italic type style that can be easily read at a smaller size, and folding the paper form (folio) into 16 sheets to reduce the dimensions and make books portable. But with these techniques the books of the time where still too voluminous.

Manutius could have used a technical solution, like publishing each work in several volumes. Instead, he called upon a value judgement. In his time, the largest part of a book was taken by the annotations, which could be several times the number of words in the original text. It was believed that the value of a manuscript depended on the annotations, and on the number and quality of the commentators. Manutius decided that his readers would read the classics for their own intrinsic beauty and the comments would be of interest only to the scholarly. He published only the original text.

Publishing a book stripped of the annotations was not an obvious decision in Manuzio's time, and this is exactly the kind of disruption that a successful technology for the $W^3$ must enable. In 1500 Venice, general education and wealth had reached a critical mass; Manutius recognized the potential of the new market and came up with the critical ideas and technologies to redirect books from an erudite audience to the general public.

The $W^3$ media industry is evolving to the Internet the technologies initiated by Manutius and the methodology introduced by the press in the 1850s; this media industry is not inventing a disruptive technology. In the next section I will present a possible methodology to enable a paradigm shift, but first, for motivation I want to gauge the size of the potential Internet printing market.

### 4.2 Market opportunity for color hard copy

The $W^3$ is the new publishing medium, fulfilling optimally the storage and distribution functions. Paper is the best medium to present written information, it will be the choice medium for the personal usage of information retrieved from the $W^3$ for immediate use. The companies that will be the best in printing information off the $W^3$ will be the ones that will own the digital printing market, not the ones that now excel in the POD market.

| Library | Size in gigabytes |
|---|---|
| Library at Alexandria (400,000 scrolls) | 800 |
| Library of Congress (20 million books) | 20,000 |

**Table 1. Size of some libraries. Source: Brewster Kahle[5]**

| Library | Size in gigabytes |
|---|---|
| Dialog Information Service | 3,000–5,000 |
| Public Web sites (400,000 http nodes) | 2,000 |
| Internet Archive (all material, including news) | 5,000–10,000 |

**Table 1. Size of some libraries. Source: Brewster Kahle[5]**

Table 1 indicates that although large, the size of the $W^3$ is not galactic; it is something that we have been able to manage for at least two millennia. If we consider a site a neighborhood in which individuals ambulate and from which they regularly print, then we must be able to manage well the average 100–300 pages per site (according to Kahle only about 500 sites have more than 10,000 pages). Since we can manage well only chunks of 7±2 items, the main requirement for any technology relating to the Internet must be scalability.

## 5 Solution

I will finally propose a solution that scales well, allows low-cost Web authoring, and solves elegantly the printing problem. We can siege control over the $W^3$ by introducing structure and method in an otherwise chaotic labyrinth of hypertext links. Once we have control, we will be able to add even more information to the system, such as for example the annotations eliminated by Manutius, providing more material that can potentially be printed. We observe that from a mathematical point of view, the $W^3$ is a graph.[8]

Each named paragraph in each file on an exported path on an Internet host is a vertex of a graph labelled by the uniform resource identifier (URI). Each link, e.g., an occurrence of an URI element, is an edge from the vertex in which it is declared to the paragraph into which it points. The links perform the “go to” function and there is no “come from” function, so it is clear that the $W^3$ is a digraph. However, the trail visited in a session can be viewed as an undirected graph when combined with the history trail, because the history trail can be used to backtrack each visited link.

There are two interesting subgraphs. One is the graph corresponding to a web site and one is the subgraph visited during a session. The goal for an author is to design a well-structured site graph, while the goal of a user creating a session is to succeed in obtaining the knowledge sought with the smallest possible effort. [One peculiar property of a session graph is that, because the session history is usually a implemented as a linear list, the browser software can efficiently remove cycles from the session graph creating an acyclic graph, although it is not clear that this behavior induces a good user model.]

We are interested in identifying good methodologies to design site graphs. The session graph can be used to determine the quality of a site graph because usually a short trail is better than a long trail, having contributed to build knowledge more efficiently. A more accurate measure can be obtained by associating weights with the edges in the graph. An example for weights is the character position of the link in the HTML file, which is an indication of the amount of text to be read before the link can be exercised.

## 5.1 Site construction methodology

### *5.1.1 Step one: introduce relations*

Relations allow the use of an order, to sort the information in a Web site. Everybody has noticed that not all links are equal, that following one set of links we can find the knowledge sought faster or more clearly. The weights can be determined automatically by a computer programs, using standard techniques such as

- counting the clock time required to reach a node
- counting and accumulating the number of words from each link to the next up to the final node
- determining the degree of difficulty in the words, sentences, or concepts
- a combination of the above

### *5.1.2 Step two: categorize each node*

The most difficult step in distilling knowledge from information is categorization. This task cannot be automated and is where the authority of an author is proven. Each node is categorized by a number of criteria or facets:

- key words (e.g., image processing, JPEG, rate-control)
- difficulty (beginner, intermediate, advanced)
- audience (student, scientist, buyer, family)
- discourse level (main thread, note, comment, reference, detail, source)
- presentation medium (computer monitor, TV set, printer, communications speed)
- and many more…

### *5.1.3 Step three: find the minimum spanning tree*

For each set of nodes there many possible graphs. This is where authors fail when they try to organize their $W^3$ sites, but fortunately this step can actually be automatized. We note that navigational information is valuable only if data is structured systematically. In fact, it is well known that users get lost in generic graphs and that cycles make it most difficult to stay on course. Trees are a special class of graphs that

is hierachical, has a root, and has no ambiguities (there is exactly one path between two vertices).

It is proven in mathematics that for a given graph the there is a subgraph that is a tree and contains every node of the graph; such a tree is called the spanning tree of the graph. In the first step we introduced weights; the spanning tree with minimum total edge weight is called a minimum spanning tree (MST). Fortunately there is a number of efficient algorithms to compute a MST for a given weighted graph, such as Prim's algorithm and Kruskal's algorithm.[8]

#### *5.1.4 Step four: interweave the trees*

At this point we have a forest consisting of a tree for each category or set of categories. The final step is again one that is carried out manually by the author and is where the richness of a Web site—which makes it unique and compelling—is introduced based on the author's creative skills. In this step the author interconnects the nodes in the various MSTs, whereby the new links express the interconnection of the different categories, such as making available again the annotations discarded by Manutius.

The main property of this methodology is that is scales well, because all the most tedious tasks (find MSTs) is performed by very efficient algorithms, and the other tedious task (weighting links) can also be automatized. The manual steps that require knowledge and creativity can be supported very efficiently with GUIs based on drag and drop paradigms. If HTML is extended to include facets in the URIs,[8] browsers can make use of categorization information, such as display it selectively or give it a different appearance, so it is immediately clear to the reader whether, for example, a particular link from a statement in the proof of a theorem is a link to the lemma where it was proven, a clarification, an illustration, or a reference.

### 5.2 Example application—solving the printing problem

#### *5.2.1 Printing a subtree*

The existence of a tree structure greatly simplifies the printing task, because now the user would not only be able to print a $W^3$ page, but also to print a subtree by enumerating in depth-first order and printing the contents of each node as it is visited. Thus the workflow is extremely simple and can immediately embraced by the masses.

A problem is that the user no longer gets an idea of the size of a print job when a print command is issued. This is not only the case when a subtree is printed, but can occur even in the case of a single page. If only the current page is printed, the fact that $W^3$ printing is not WYSIWYG** prevents the user from getting a feel of the size and scope

**. What You See Is What You Get. HTML is about document structure, not layout, so formatting is adjusted dynamically to the current window size, which is usually not proportional to the printable area on a sheet of paper, especially if the user has a small display monitor.

of a print job, as soon as the page is larger than what fits in the user's browser window.

There are no issues with printing text and vector graphics, all problems have been solved for years with the use of style sheets. Sampled images, instead, should be sent twice, at screen resolution (*e.g.*, 72 dpi or dots per inch) for rapid display and at a resolution matched to the print screen (*e.g.*, 200 ppi or pixel per inch) for adequate quality printing; this problem is solved elegantly with the FlashPix image format.[1] Since images are two-dimensional, the time to transmit an image grows quadratically with height, not linearly. This makes it harder for humans to extrapolate data size from the screen display and the elevator in the scroll bar.

Still in the case of printing a single HTML page, if the page is long and the link is slow, the user may issue a print command before the whole page has been received to print it concurrently. In that case the user has no way to know how long the document print job is. When the user is allowed to print a subtree, it is clear there is no reliable way to gauge the size of a print job. It should be kept in mind, that the bottleneck is not necessarily a slow modem; even when user uses a cable modem connection or a T-3 line, the server or the Internet service provider's (ISP) gateway or even the Internet backbone can be congested.

Humans are much better at making value judgements than computers, therefore it would be a highly desirable HTTP feature if the server notified browser clients of the number of printed pages for each $W^3$ page and subtree, letting the user make print/don't print decisions based on this information. Browsers could generate the information locally without changing the HTTP protocol, but this would require sending the data over the Internet—a huge waste of bandwidth resources when this is done by millions of users.

Scenarios indicating the value of the size hint:

- Some $W^3$ authors put a page size at links to large HTML pages; sometimes they forget to update the size when they change the contents, or make calculation mistakes. Computers are better at this kind of task than humans. Moreover, if the hint is part of the protocol, it would be a universal feature and browser clients might provide a better user interface if the information is encoded in a machine-friendly format.
- The user might have a fast 300 ppm (pages per minute) device set as the default printer and sitting at the root of a large subtree (e.g., a phone list or a large parts catalog). Before anybody notices, many reams of high quality paper can be wasted (this happens more frequently than one might suspect). People tend to waste less paper if a large print job generates a warning with the number of trees being consumed.
- Many people have a slow personal thermal ink jet printer connected locally to their workstations. They have a low tolerance for long print jobs and often start

printing before leaving work in the evening, leaving the printer running unattended for the night.

By mentioning the concept of printing a subtree, I assumed the fact that printing is a disconnected task from browsing. Currently the $W^3$ protocols allow printing only in the form of producing a hard copy of the currently displayed $W^3$ page. It is clear that this simple mechanism will be replaced just as fast as printing documents by sending a display bitmap to a printer was superseded by page description languages.

If printing becomes an independent background task, the $W^3$ protocols can be extended to allow a negotiation of the transmitted document's reproduction parameters, in a similar fashion as this is accomplished in the facsimile transmission protocols. An independent printing task with capabilities negotiation can easily be generalized to incorporate pro tempore licensing or rental schemes for fonts based on nanocommerce, as well as clever downloading schemes that minimize the storage requirements for fonts in spoolers and printers.

### *5.2.2 More benefits for $W^3$ printing*

When the categorization machinery above is seamlessly integrated in a Web site authoring tool, the print job size problem can easily be solved. A category marker can be assigned to carry a print size hint for each node and can be included by the authoring when a file is saved to disk. The browser can use this print size hint and the effective transfer speed to compute the estimated the print job size for a node. If the authoring propagates this information to each node's parent, then the print job size for each subtree can easily be computed.

Until now in this example I have assumed the printed information is the same as the displayed information, except for resolution. As I suggested in Section 5.1.2, "Step two: categorize each node", each node can have a number of facets, and the printed page can be completely different than the displayed page. Such a possibility is very useful for the high-concept material produced by the media business, where the T-1 display faced can contain fancy animations, sound, and video, while the print facet consists of a glossy brochure.

Last but not least, the structuring the information as subtrees allows to easily implement look-ahead functions and allow down-loading and printing to proceed asynchronously. This improves the performance and encourages readers to print even more.

# 6 Other technologies

## 6.1 Search engines

Search engines are useful only for locating pieces of information, they can find a needle in a hay stack. They do not allow the creation of knowledge (why is there a needle

in the hay stack?). For example, if you enter the phrase *to be or not to be* in a search engine, the found set will be empty, because this phrase consists entirely of stop words; until I told you, you would not immediately have been aware of this problem. Even entering the phrase in quotes it will take you a long time to find out what Shakespeare meant when he put it in Hamlet's mouth, if you only use a search engine as your tool.

The old adage that writing a new book is easier than finding an existing one will not be invalidated by the $W^3$ and search engines. This in not only a technological limitation, it is part of human nature because we are active, not just contemplative. We are always dissatisfied with what we have and search for new problems to solve.

### 6.2 Applets

Applets are used in the hot media for special visual effects that get the reader's attention. In the $W^3$ as a cool medium applets are very useful to distribute computation and off-load processes that do not require central resources from the server to the clients.

### 6.3 Agents and Avatars

These tools use a reader's profile to collect on the $W^3$ the information that allegedly interests the reader. This technology has two questionable aspects. The first is that the "philosopher" implementor who decides which information should be gathered in response to a particular profile is a futile dream that cannot exist; to the contrary, history has shown that philosophers have never been good at understanding current events. The second is that people like to read what other people read, which is in contrast to a personalized news service. The only market is competitive intelligence, and that is a small market.

### 6.4 Databases

Databases are a structured storage medium for information. They are very useful when catalogs of all sorts are made available, but have the same limitations as search engines; they deal with information not knowledge.

On the other side, databases are an ideal storage medium of the parts for a Web site and for supporting the categorization task. With tools only slightly more sophisticated than those explained above, a forest of site trees can be generated directly from the information contained in a database if the categories themselves are ordered hierachically.

### 6.5 Nanocommerce and authentication

Nanocommerce refers to transactions worth between 1 ¢ and half a millicent. The Internet is not only a medium, it is a social means. As mentioned in the gold rush analogy above, nanocommerce is one Internet activity that will be very profitable, because it allows individuals to sell small digital items to the flash crowds typical for the $W^3$.

Items sold though this channel can range from clip art, to authoring, programming, image processing, information gathering, etc. The nanocommerce banking function will probably similar to the current copyright clearinghouses.

In many global transactions there is no longer a face to face contact, therefore there is no trust between buyer and seller. Expensive and mass-market items will be produced in contract for the large corporations that own the global distribution networks. Commercial authentication services will allow individuals to sell full custom items directly to anonymous buyers. In contrast to the depressive scenarios depicted by the traditional media, where nobody will ever write a book because no royalties can be collected su to the collapse of copyrights, nanotransactions of all kind of work will flourish; the reason is the same as people mostly do not buy software because they are afraid to go to jail, but because owning registered software it is convenient.

### 6.6 Cryptography

One characteristic of the post-industrial society is self-reliance and the elimination or deregulation of government services. This reduces the need for taxes and politicians are willing to promise tax reductions. Wealthy individuals reduce their taxes by deferring income or by sheltering it in unearned income plans, with are taxed at a lower tax rate if at all. Small business people can solve the problem with creative accounting. For the rest of the society the traditional method to reduce the tax burden has been to barter, which is not of immediate use to people on fixes incomes. With the transition from permanent employment to contractor work, the rest of the society can now shift from a cash society to a limited barter society.

Cryptography is the tool to hide transactions over the Internet that would otherwise be taxable. Economists will be faced with the difficult problem that a large monetary quantity is being created that cannot be quantified statistically. This will make it difficult to gauge measures such as the gross domestic product, because money can dive in one economy and resurface in a completely different economy at the antipodes.

## 7 Conclusions

I have presented a methodology that allows the creation of well-structured Web sites with little effort. This allows individuals and work groups to publish knowledge at a quality equal to that possible for the new media giants. In McLuhan's terminology it allows the coexistence of the $W^3$ both as a cold and as a hot medium. In addition this methodology solves many aspects of the Internet printing problem. This will enable a new paradigm for printing, away from traditional printing and print-on-demand, towards individual printing on local printers. The cost of these printers can be kept low because the print speed can be low and profits can be shifted to consumables.

## References


1. http://image.cup.external.hp.com/ifori/htdocs/b_dlmenu.html
2. http://search.nytimes.com/library/cyber/week/1212msn-hollywood.html
3. http://search.nytimes.com/web/docsroot/library/cyber/week/012297microsoft.html
4. http://shc.stanford.edu/shc/dreyfus.html
5. http://www.archive.org/about-us.html
6. http://www.corbis.com/inside/story.html
7. http://www.disney.com/
8. http://www.hpl.hp.com/personal/Giordano_Beretta/Documents/literature.html#hpl9699
9. http://www.ietf.cnri.reston.va.us/
10. http://www.mcluhan.ca/
11. http://www.msn.com/
12. http://www.msnbc.com/
13. http://www.nytimes.com/library/cyber/compcol/020497compcol-manes.html
14. http://www.pointcast.com/whatis.html
15. http://www.webtv.com/
16. http://www.w3.org/hypertext/WWW/Consortium/