MULTICHANNEL PUBLISHING AND AUTOMATED UPDATING OF ONLINE NEWSPAPERS

Alex Jonsson* and Nils Enlund*

Keywords: electronic publishing, newspaper, on-line, WWW, Internet, scripting, automation, HTML, parallel publishing

Abstract: There is a growing trend among newspapers toward the parallel use of news and advertising for online distribution. The preferred medium is the World Wide Web. Online publishing requires additional editing and updating as well as conversion of the material to the HTML format. This must be done with a minimum of manual effort.

We have developed automatic methods for the formatting and continuous updating of newspaper information in a hierarchical WWW server structure. Using background processing functions, automaticaly initiated by the addition of new information in specific folders/subdirectories on the server, text and HTML-tagged files can be converted and merged with predefined static text and inline graphics.

The pages of an online newspaper can thus be kept topical by simple drag-and-drop techniques while still maintaining a consistent appearance to the reader.

The methods have been implemented and are in continuous use producing the online version of the newspaper *Svenska Dagbladet.*

---

* KTH/GT, Royal Institute of Technology, Division of Graphic Arts Technology, Stockholm, Sweden (alexj@gt.kth.se, nilse@gt.kth.se)

1. BACKGROUND

1.1. Multichannel newspaper publishing.

New digital distribution methods and the rapid growth of an information technology infrastructure covering both the home and the office market has created both new business opportunities and new competition for the traditional newspaper publishers. Most newspapers seem to consider the digital techniques as merely an additional channel for distributing the information that is collected and processed for publication in the printed newspaper.

Electronic publishing is seen as a means of repurposing information and generating additional revenue through parallel publishing using multiple distribution channels [Katz 1994]. Reading newspaper pages from a display screen will not, however, in the long run be an attractive proposition for the customers [Diller 1995]. An electronic distribution medium will require a new business concept and a new product design in order to become viable and attract both readers and advertisers — the traditional newspaper paradigm cannot be directly transferred from paper to screen [Enlund 1993].

1.2. Rapid growth

Newspapers have rapidly adopted online services as a means of distributing new electronic products. Currently, there are more than 900 online newspaper services in operation worldwide. Half of them are published in the United States and one fourth has European origin. There are also an estimated 670 magazines and 320 TV/radio broadcasters active in online publishing [Editor & Publisher Interactice, New York, Jan 1996.]. By the end of 1996 there will be an estimated 2,000 electronic newspapers worldwide [Outing 1996].

Internet is the preferred distribution channel. At the end of 1995, more than 90% of the electronic services were accessible over Internet and the World Wide Web. This figure is likely to increase. The closed commercial online carriers (AT&T Interchange, Prodigy, Compuserve, America Online, etc.) are losing the position they initially had as online users and information providers are moving to Internet. Also the Bulletin Board Sys-

tems (BBS) that have held a significant share of the online newspaper market are losing ground.

1.3. The rationale for electronic publishing

Currently, the development of electronic newspaper products is in an experimental phase. The explosive spread of Internet and especially of the World Wide Web service has made this the preferred platform for newspapers interested in exploring the intricacies of electronic publishing. The rationale for publishing an experimental electronic edition of a printed newspaper on the Internet can be summarized as follows:

- **A natural business extension.** Newspaper publishing companies have a long tradition and much experience regarding the collection, filtering and presentation of information. They also have the necessary technical equipment available since the prepress environment is highly suitable also for digital publishing. In an age of information overflow, the proven quality of information produced by a newspaper publisher will have a high market value.

- **New business opportunities in a new market.** Publishing on computer networks is an unexplored new market with unknown business potential. There is a possibility that traditional channels will lose market shares to new competitors in the new medium. For many traditional publishing companies, ventures into digital publishing are means of securing all bases.

- **Advertisement.** Some types of newspaper advertising, notably classified advertising, are clearly well suited to interactive digital media. Traditional display advertising in an electronic publication is, however, a considerably more complicated issue. Newspaper publishers and advertisers will have to experimentally investigate what forms interactive advertising can take — layered ads, hyperlinks, etc. — as well as how and how much the publisher can charge for advertising space.

- **Additional services — new sources of revenue.** A newspaper has to select and filter its information to fit the printed product. Of all news items considered for publication and of all images received and processed, only a small fraction actually reaches the printed form. Using complementary, digital distribution channels, more information can be made available to the customers that wish to receive it. Additional data formats, such as video, audio, and animation sequences can also be used.

- **Knowledge and experience.** Pilot projects increase knowledge and experience of digital publishing and related topics such as multimedia, interactive advertising and electronic methods of payment.
- **Image value.** By being active on the Internet, a publishing company exhibits an image of technological prowess. This may improve its position also on the traditional market.

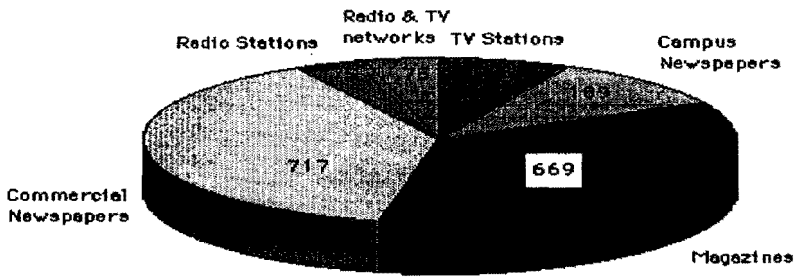

Figure 1. Note that the newspaper publishing companies are not alone in this business and on the Internet, everyone looks the same to the user. The pie chart shows the number of web sites originating from each company type respectively, as of Jan 1996. [Source Editor & Publisher, March 1996].

1.4. Concepts

Some of the central information technology terms and concepts referred to in this report are the following [Fluckiger 1995.].

Internet. Internet is a structureless, multiconnected, international network of computers — servers and clients. Every computer connected to the network has a unique network address, comparable to a telephone number including nation and area codes. In 1996, the number of computers connected to Internet is estimated to be between 40 and 60 million. The computers communicate using the TCP/IP network protocol [Fluckiger 1995].

Transmission Control Protocol/Internet Protocol, TCP/IP. TCP/IP is the network protocol of Internet. Data to be transmitted is divided into packets of fixed size, consisting of a header and the actual payload. The packet header contains information about the source,

destination, priority when travelling over the network, sequence number in a data stream, ex-piry date, and contents of the packet's payload. The protocol can run on almost any physical network.

The World Wide Web, WWW. WWW is, after electronic mail, the most popular service on the Internet, providing an information access structure suitable for electronic publishing. WWW uses hypertext and multimedia techniques to make "pages" of information easy to retrieve, browse and contribute to. Adressable "pages" — containers of multimedia informa- tion — are stored on WWW servers distributed on the Internet. These "pages" can be accessed and inspected using computer programs called browsers. Many people falsely consider WWW to be equivalent to the In-ternet, whereas the former is only one of many available applications on the worldwide computer network. The WWW and its protocols are actu- ally independent of the Internet and can be used on visually any other type of computer network.

Hypertext Mark-up Language, HTML. HTML is the standard way to mark-up documents intended to be read by a WWW browser application. Compact marked-up documents are interpreted by the WWW browser which add typography and layout for presentation. HTML is a DTD (Doc-ument Type Description) of SGML (Structured Generalized Mark-up Language) [Fluckiger 1995.]. Unfortunately, there exist several dialects of HTML.

Hypertext Transfer Protocol, HTTP. HTTP is the transportation protocol, for distributing HTML documents on Internet.

File Transfer Protocol, FTP. FTP is another, somewhat older transportation protocol, for transferring files between computers both locally and over the Internet. When you download a file from another computer over a network, you make a download request and if you are properly authorized, the server software sends you the file in a stream or in fixed-size packets.

Likewise, when you upload a file you tell the server what folder/directory you wish to be the file's destination and you then perform the actual upload. Often, a username and/or password is required when interacting with the server.

1.5. Publishing strategies

Two slightly different electronic publishing strategies are described in this report. The publishing methodology is, however, in most aspects similar for both strategies and many newspaper companies use a hybrid form. The strategies differ in the frequency of updating the contents of the electronic product.

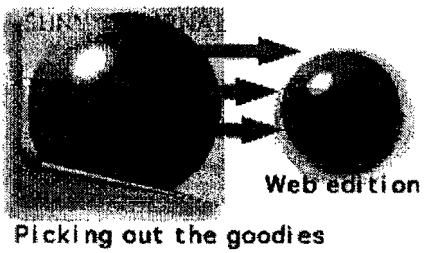

Figure 2 The digital edition is based solely upon material in the physical paper edition, and exported from the make-up system to HTML.

A. **Regular electronic editions.** The electronic edition is updated once a day with contents from the printed paper. The new material is prepared together with the contents of the printed newspaper. It is ex- ported for digital distribution directly from the newspaper production data base or from the digital pages produced using a page make-up sys- tem. In this strategy, there is a one-to-one relationship between the edi-tions of the printed version and of the digital version of the newspaper.

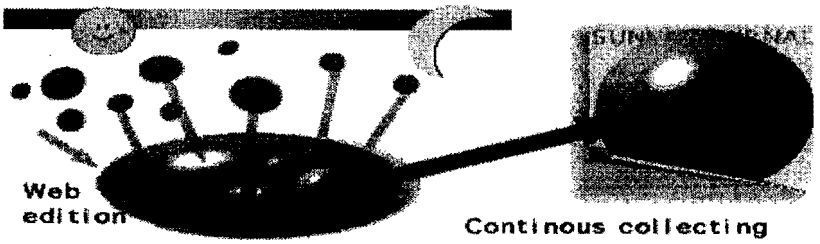

Figure 3 The web edition is continuously updated throughout the working day.

B. **Continuous updating.** News and articles prepared for the printed paper edition are exported to the digital product continuously during the day. The updating may be incremental or happen in bursts

at regular intervals. The electronic newspaper becomes an editionless publication, always containing fresh information. The information may or may not, in some version, be included in the printed newspaper.

In practice, all shades between the two strategies are used. If an event of such magnitude that it would have stopped the presses occurs, it seems only natural to immediately insert it as an update also in a static electronic edition.

Regardless of the strategy used, the publisher will greatly benefit from automatic and semiautomatic functions for purging old information from the electronic edition and for adding topical information.

2. AN ELECTRONIC PUBLISHING PROJECT

In February 1995, the Swedish daily national newspaper *Svenska Dagbladet* initiated a research and development project aiming at producing an electronic edition on the World Wide Web. The electronic edition was intended primarily as an experiment, providing a vehicle for testing production tools and routines, contents and structure, and business concepts. It was, however, stated at the outset that a successful experiment would have to be rapidly transformable into a commercially viable online publishing service.

The simple updating of the electronic edition was identified as a critical factor in a successful project as well as in a regular publishing service. The electronic edition has to be created primarily using the material intended for the printed newspaper and as a byproduct of producing the daily newspaper. The amount of additional manual labour has to be minimized and the publishing activities have to be integrated in the normal newspaper prepress production processes. The creation of automatic and semiautomatic updating functions thus became a focus area of the project that was carried out in cooperation with the Division of Graphic Arts Technology at the Royal Institute of Technology.

The electronic edition was implemented in steps of improved contents and increasing functionality. We started off in early 1995 by publishing static information about the project using an experimental server that was closed to the public. The next step was to experiment with a hierarchical

file structure with hyperlinked index files. A large number of previously published reviews of films, records, opera, ballet, and books were found suitable for this purpose.

On June 23,1995, the World Wide Web edition of *Svenska Dagbladet* was opened to the general public. We had then implemented the file structure, performed basic programming, designed a workflow for daily publishing, and trained the staff involved. Since then we have carried out various experiments using animations, JPEG compressed images, QuickTime videoclips, audio sequences, as well as hyperlinking to external WWW sites carrying information related to published news items and current events. The electronic edition is now a stable publishing product, produced daily by the newspaper editorial and advertising staff using automatic and semiautomatic publishing tools for daily updates.

The service is continuously being enhanced and improved. Sound files, short video sequences and links to information related to news topics have helped to enhance the depth and richness of the news articles. From being a byproduct, the digital edition has become a full-fledged service (figure 4).

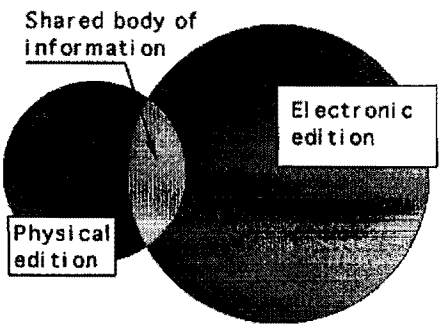

Figure 4 The two editions share some news headlines and advertisements. They are today two compliementary, commercial services at *Svenska Daybladet.*

During 1996 we are experimenting with including active, downloadable applications in the service, using the Java language from Sun Microsystems and adding animation with the aid of Macromind's Shockwave. In the following, however, we concentrate on the concepts and tools developed for producing the basic WWW service.

3. IMPLEMENTATION OF A WWW NEWSPAPER

In exploring the implementation of the WWW edition of a daily newspaper, we will look separately at the problems of transferring text from the newspaper data base, creating structure, including graphics, and creating routines for the automatic updating the digital publication.

Since the structure of the *Svenska Dagbladet* WWW edition is fairly complex, we will use as an example an imaginary newspaper called the *Sunny Journal.*

3.1. Text material

Newspaper editorial text processing systems normally transfer text material either to a proprietary publishing system or to an off-the-shelf page make-up program, such as QuarkXpress. However, the document structure and file format of these programs are often complex and cannot easily be exported to HTML for use on the World Wide Web. Special export filters exist for some programs, e.g., QuarkXpress [Anon. 1996].

However, not all material in a newspaper originates at the same source, and forcing all material through a specific page make-up program in order to publish it electronically seems unnecessary. It would be better to find a simple common denominator, a neutral text format that can be utilized both by page make-up programs and HTML converters. The main candidates are:

- ASCII text with line breaks.
- RTF, Rich Text Format, used by Microsoft Word
- SGML
- Proprietary formats like those used by Microsoft Word, WordPerfect, etc.

Since text material arrive at a newspaper in many different formats and mark-up styles, the proprietary formats proved to be too numerous to be an acceptable alternative. SGML is very rich, extensive and flexible but will require two conversions: first from the original format into SGML and then from SGML into HTML for the final publishing. The RTF format is

a semi-standard cross-platform format that conserves text styles, e.g., bold typefaces, italics and underlines. Not all programs, however, have the capability to convert from its own proprietary format into RTF.

We therefore decided to use the simplest of formats — text with line breaks — mainly in order to avoid multiple conversion filters when translating into HTML.

3.2. Document structure

A basic problem in digital publishing is the question of how to best organize the contents for user appeal and easy access. A good model is to be found in the structure and organization of a modern printed newspaper, with its clear logic of sections and departments. If we copy this structure to the WWW edition, every section of the paper can be represented by a folder/subdirectory in the digital edition (figure 5).

Let us assume that our small fictitious newspaper, *Sunny Journal,* carries five editorial departments as follows:

- Arts & Entertainment (arts)
- Domestic & local news (domestic)
- Economic dep. (economics)
- International news (international)
- Science and knowledge (science)

Figure 5 The five subdirectories for the *Sunny Journal's* digital edition.

Transferring this department structure to the WWW edition will result in a simple file structure in the WWW server, as in figure 5. Within this

structure, each independent story, or "page", is represented by a document file.

When creating the structure, we must distinguish between the physical file structure — which reflects the logical structure in its traditional, hierarchical form — and the hyperlinked structure. The power of the World Wide Web and the basic idea of hyperlinking is the possibility of going beyond the hierarchical structure by creating associative links between information items.

Since the number of files in a WWW publishing system tends to have a very high growth rate, it is convenient, from a file management standpoint, to arrange the hyperlinked files on the WWW server in a simple hierarchical fashion. Within each editorial department folder/subdirectory, the file structure should be kept as flat as possible in order to avoid long document location descriptions.

Another important navigational aspect is that the reader, regardless of which document he is currently accessing, never should be more than one or two mouseclicks away from the main index page. Every page should also include a mail link to the document creator or to the system operator.

3.3. Graphics

Graphic elements give a WWW service much of its identity and character. Graphic elements are stored in separate files on a WWW server. When, in an HTML document, a reference is made to a graphic object, the browser will retrieve the file and place it in position on the user's screen. The first time a graphic object is downloaded from a WWW server, it is cached, i.e., temporarily stored, in the computer of the user. The next time the same image is requested, there is no need to retrieve it over the telecommunications network. It is therefore of great importance, due to speed considerations, to whenever possible reuse the same graphic objects on several WWW pages.

It is noteworthy, that during the first ten months of the *Svenska Dagbladet* experimental service, more than 80% of the total network traffic consisted of transfers of graphic elements. Since large graphics require the transmission of large files, we have chosen to keep images small in size, with few colors and with a high compression ratio.

The graphic elements are preferably small in file size. If a larger graphic element (> 50 kB) has to be presented because part of its value lies in large physical size or great detail, a smaller, thumbnail-type image is placed on the page with a hyper-link to the larger image file, along with information about file size of the larger image for convenience. This measure is also an effort to allow consistent layout style since the size of graphic elements thus can be predicted regardless of the format of the original image. For the experimental service, we designed one identifying header banner for each department. The design is consistent in style (figure 6).

News graphics and other non-permanent graphics are added to the server folders/subdirectories and linked to individual documents using the original file name of the text file as an identifier. For example: we have a text file, containing an article about the lynx. It is appropriately named lynx and the converted HTML-file is named *lynx.html* (see figure 6). If an image file with the document name *lynx.gif,* or alternatively *lynx.jpeg,* is resident in the same folder/subdirectory on the server, it is automatically linked to the file and positioned, for instance, underneath the story heading. The default placement of news graphics is aligned to the right margin of the page. This can naturally be altered manually if called for.

In a case with several graphics that are to be linked to the same story these will have to carry unique names. There is a built-in feature which supports multiple reserved names. If the story's title is *foo* then *foo.gif, foo_t.gif, foo_m.gif and foo_b.gif* are reserved and can be searched for ty the deamon. The suffix is related to the image placement: _t for top, _m for middle and _b for bottom.

Figure 6 An example page from the *Sunny Journal.* Notice the inline graphic and the static banner at the top of the "page". .

When using a very strict filing structure coupled with automated updating, visual richness and depth of content in the electronic service must be created by other means. In-line graphic and navigation tools must be carefully designed and implemented. These elements alone give the WWW edition of a newspaper its special character and the users a feeling of consistency and familiarity.

Some newspaper publishers, e.g. the *Washington Post,* have chosen to give their web edition very little similarity in appearance compared to the physical paper edition. Others, like the Swedish newspaper *Aftonbladet,* claim that there are benefits in the synergy between the two editions and strive to let the reader get a feeling of recognition on their first visit to their web edition.

3.4. The publishing daemon

Daemon is the Unix term for a resident program without user interface that runs in the background on a computer. We can design a daemon for our electronic newspaper that performs the following tasks (figure 7):

- Regularly scan the department folders for newly added text files, for instance, every other minute.
- Convert all new files to HTML.
- Examine the folder for images with matching names.
- Add the appropriate header with the correct banner image and a footer with links back to the department index page as well as to the main index page.
- Update the index page of the department folder with a link to the new document, or, when an outdated story is removed from a folder, remove the corresponding link.

If, as in the case of *Svenska Dagbladet,* the daemon is resident on the web server together with the server software, added speed and simplicity is obtained by not having to transport the files back and forth over the local area network. Text files are prepared in the appropriate format by purging them of all non-text attributes and placing them in the appropriate folder/subdirectory to await conversion to HTML. Note, that any corresponding graphic file must be placed in the same folder prior to activating

the daemon in order for it to be automatically added to the converted-HTML document.

In our experimental project, the source folders and the target publishing folders were the same.

In a remote access situation using an external web server where updates are performed only a few times daily over a modem line, the daemon could be invoked at will at discrete intervals. This gives good user control and avoids unnecessary file transfers. In this case, the daemon can be located on any workstation with local area network access.

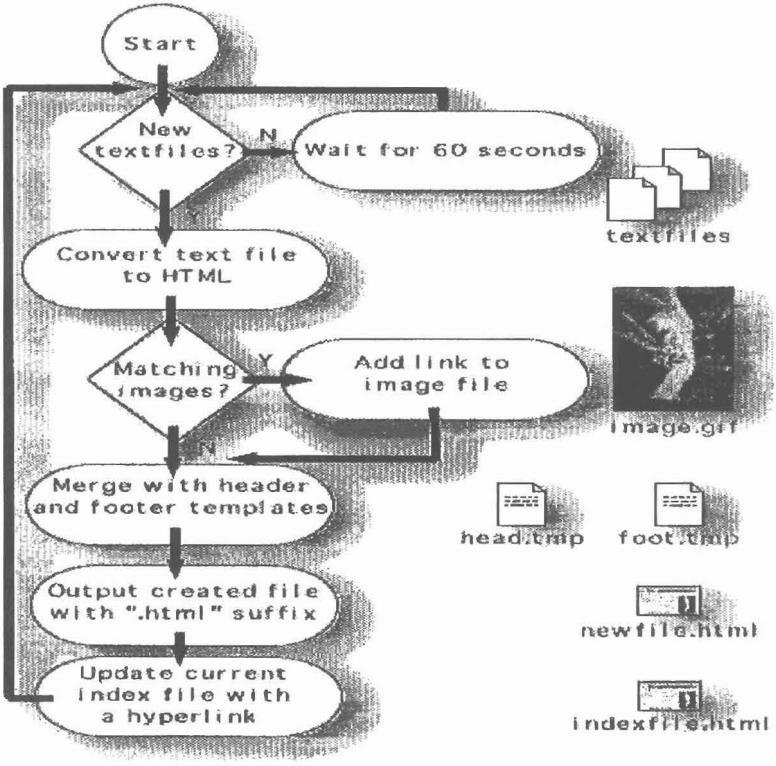

Figure 7 The publishing daemon. The file icons represent text files and graphic elements that are prepared in advance. The daemon runs at set intervals or upon request from the user.

Scripts are the background-running programs required for performing the automated tasks of the daemon. Dependent on the system environment, the scripts can be either true, interpreted scripts on the operating system level or independent compiled programs [Trinko 1994, Brenner 1996].

In a case where the publishing daemon resides on the web server itself, multi-tasking capability is required. In our project, the operating system choice fell on the Unix platform. The experiments were carried out on a Sun SPARC 2 workstation running under the Sun Solaris Unix operating system.

The publishing system consists of:

- The compiled C++ daemon program.
- Prepared header and footer templates for the WWW pages, one header and one footer for each editorial department.
- A conversion table for special national characters.
- Pointers to the source folders.
- A configuration file with miscellaneous settings, e.g. folder scan frequency, image import capability on/off toggle, etc.
- On-line documentation for the users.

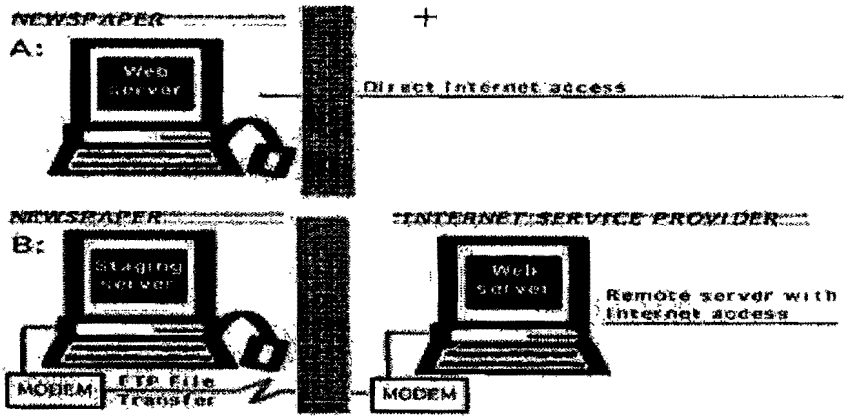

Figure 8 Two publishing methods. A has its own Internet access, while B rents space at a local internet service provider's web server.

In order for the publishing daemon to be able to automatically update the index files, all HTML documents must be resident in their appropriate folders at all times. This may call for the use of a so called staging server, a mirrored server where the document structure can be tested before the actual publishing (see figure 8).

In a small publishing set-up with only a few persons working on the contents, the risk of misplaced files is not critical. When working with an external web server the staging server becomes more important. The material on an external server is usually accessed with the FTP protocol.

A better solution than to directly access the external server and kepping all documents there is to have a local mirror of the external server. A new document file can be added to this mirror site and letting the index file be updated locally without any file download from the WWW server. This is also a convenient way to check for errors and typos.

Upon uploading the new HTML document (along with its images, if any) and the altered index files, the exernal publishing WWW server is immediately updated as well.

When working with the daemon, images for individual files must be prepared in advance. This calls for collaborate measures between the editors of the electronic service and the photo department. In a scenario with continuous publishing, the updating of both news text files and graphics must be synchronized.

3.4 Experiences regarding implementation

The WWW server at *Svenska Dagbladet* log approximatly 60,000 external connections per week and offer over 6000 HTML-documents to the public. This corresponds to 300 — 400 megabytes transmitted from the web server every day. Many documents have similar structure and are well suited for automated publishing. The daily update is performed by three employees who all also work parttime with the physical edition. About 20 employees at *Svenska Dagbladet* are involved in the digital edition in all. Lately, the sports section have grown interest in publishing sports results, interviews and comments regarding sports events on the server. Their material is easily added to the web server's structure and is today updated on a daily basis. This supports the theory that may be wise to incorporate

few sections of the physical paper at an early stage, continuously develop the WWW service using these sections to a presentable level and expand to the other departments.

4. CONCLUSIONS

There are a number of routine tasks in multichannel digital publishing that can be automated using scripting tools. With careful planning of document structure and a well structured graphic design, a rich World Wide Web edition of a newspaper can be created and updated without the need for extensive manual maintenance and support.

A variety of tasks are well suited for automation, e.g., file conversion, the adding of static HTML code, insertion of graphics, and updating of the hyperlink structure. A publishing daemon, inspecting file additions and deletions in a designated folder/subdirectory, can handle these tasks. However, automation requires careful preparation in order to be efficient. When planning the electronic WWW service, it is worthwhile to use a simple, flat file structure with the possibility for expansion for additional publishing sections.

REFERENCES

Anon.: "Electronic delivery — Internet and CD-ROM publishing", Seybold Special Report, Vol. 4, Nr. 8, March 1996, pp. 9-21.

Brenner,S E: "Introduction to CGI/PERL", M & T Books, New York 1996

Diller, B.: "Don't Repackage — Redefine!", Wired, February 1995, pp. 82-84.

Enlund, N.: Elektroninen sanomalehti—ongelmat ja tutkimustarpeet ("Electronic newspapers — problems and research issues", in Finnish), Report TKO-B110, Helsinki University of Technology, Otaniemi, 1993.

Fluckiger, F.: Understanding networked multimedia — applications and technology, Prentice Hall, London, 1995, 620 p.

Katz, J.: "Online or not, newspapers suck", Wired, September 1994, pp. 50-58.

Outing, S.: "Hold on(line) tight", Editor & Publisher, February 17, 1996, pp. 41-61.

Outing, S: "Where is the money?", Editor & Publisher, February 17, 1996, p 81.

Wolf, G.: "Steve Jobs, the next insanely great thing", Wired, February 1996.

Savola, T et al: "Using HTML", Que Corporation, Indianapolis, US, 1995

Trinko, T: "Applied Mac Scripting", New York. 1994