Print-n-Link: Weaving the Paper Web

Moira C. Norrie
Dept. of Computer Science
ETH Zurich
8092 Zurich, Switzerland
norrie@inf.ethz.ch

Beat Signer
Dept. of Computer Science
ETH Zurich
8092 Zurich, Switzerland
signer@inf.ethz.ch

Nadir Weibel
Dept. of Computer Science
ETH Zurich
8092 Zurich, Switzerland
weibel@inf.ethz.ch

ABSTRACT

Citations form the basis for a web of scientific publications. Search engines, embedded hyperlinks and digital libraries all simplify the task of finding publications of interest on the web and navigating to cited publications or web sites. However the actual reading of publications often takes place on paper and frequently on the move. We present a system Print-n-Link that uses technologies for interactive paper to enhance the reading process by enabling users to access digital information and/or searches for cited documents from a printed version of a publication using a digital pen for interaction. A special virtual printer driver automatically generates links from paper to digital services during the printing process based on an analysis of PDF documents. Depending on the user setting and interaction gesture, the system may retrieve metadata about the citation and inform the user through an audio channel or directly display the cited document on the user's screen.

Categories and Subject Descriptors

D.2.11 [**Software Engineering**]: Software Architectures; H.4.m [**Information Systems Applications**]: Miscellaneous; H.5.m [**Information Interfaces and Presentation**]: Miscellaneous

General Terms

Design, Algorithms, Experimentation

Keywords

Interactive paper, document integration, digital library, citation management

1. INTRODUCTION

Nowadays we are all familiar with the notion of hyperlinks as a means of creating interconnected document spaces that can be easily searched and navigated with the help of search engines and web browsers. The ease with which we can not only find documents on the web, but also follow links to related documents is considered to be one of the main advantages of working with digital documents as opposed to paper equivalents. Yet many studies have shown that paper is still the preferred medium when it comes to reading documents for reasons of comfort, mobility and also ease of annotation.

In this paper, we show how emerging technologies for interactive paper enable a bridge to be built between paper and digital document spaces that allows users to travel back and forth between the two using the medium most appropriate to the current task and situation. Users can print documents for reading and, by activating hyperlinks on paper with a digital pen, they can not only access digital information about related documents, but also initiate web searches to retrieve copies of those documents. The retrieved documents may in turn be printed for reading with automatically generated embedded hyperlinks on the paper. Mobility is supported by providing disconnected users on the move with audio information about related documents and the option of adding them to a request list to be processed when reconnected.

As an application, we take the everyday academic task of reading scientific publications. Whether the reading activity is for the purpose of knowledge discovery or reviewing, readers regularly look up citations to other publications by flicking to the reference list at the back of the paper. In some cases, a search for the cited publication may follow and, if found, often it is printed and added to a pile of documents to be read. Citations are the established means of providing the links that form the basis of the scientific web of publications that spans the digital and paper worlds. Just as citations within digital documents can be used to automatically generate hyperlinks to cited documents, we can automatically generate links within paper documents through an analysis of the PDF file when the document is printed.

Section 2 presents a more detailed look at the activity of reading scientific publications and a discussion of related work. Section 3 describes the digital pen and paper technologies and how these can be used to support interaction between paper and digital services. We then present the general functionality of our enhanced reading system in Section 4 and the overall architecture in Section 5. Details of the virtual printer component used to analyse PDF documents for automatic link generation as well as printing are given in Section 6. Concluding remarks follow in Section 7.

DocEng'06, October 10–13, 2006, Amsterdam, The Netherlands.

2. READING OF PUBLICATIONS

While there have been dramatic increases in the use of digital technologies for the storage, processing and delivery of information over the last two decades, the affordances of paper have ensured its retention as a key medium for reading and annotating documents. Paper has many advantages over digital media in terms of how people can work with it, both individually and in groups. It is portable, light, cheap, flexible and robust. Furthermore, various forms of paper-based collaboration and interaction are nearly impossible to support in digital environments [12].

A set of reading-related affordances of paper documents are pointed out by Sellen and Harper in their book *The Myth of the Paperless Office* [18]. First, paper allows for quick and flexible navigation through a document. The size of a document acts as a rough indicator for the amount of information stored in it and provides a spatial orientation for readers while flicking through the pages. A second affordance of paper is the possibility of marking up a document while reading by using different forms of annotations. A highlighter pen may be used to mark important sections or comments may be written in the margins. These free-form annotations can help readers structure their thoughts or be used during a review process to note comments for the authors. Studies have shown the difficulty of supporting the same richness of annotations in digital applications [15].

Sellen and Harper also mention an affordance which is based on the fact that the information on paper is fixed but still the paper documents remain mobile. It is possible to read across more than one document at the same time by placing multiple documents next to each other and thereby defining a spatial order on a work space. Furthermore, paper supports the seamless interweaving of hybrid activities such as reading and writing. By placing a document next to a notebook we can take notes while reading the document.

Reading scientific publications is a major activity of academics that is associated with several tasks such as knowledge acquisition, collaborative authoring and the reviewing of student reports and papers. For the reasons stated above, it is common for publications to be edited and distributed electronically, but that the actual reading process takes place on paper, enabling the documents to be easily annotated with comments or corrections. It is also common for academics to use travelling time to carry out these tasks and papers, reports and theses are often printed before embarking on longer or shorter journeys on buses, trains and planes.

A feature of scientific publications is the importance that citations play in providing links to related work and background material. Frequently readers interrupt the linear reading process to lookup a citation embedded in the text to check if they know the cited work or to point them to further reading. While some styles of citations give a good indication of the work cited by including the authors' names and the year, simple numeric references are often used, giving no hint about the work cited. If a reference is particularly interesting, the reader may actually search for a copy of the cited paper and temporarily switch to reading that paper, position it alongside the original paper for purposes of comparison or add it to a pile for later reading.

One can consider that citations are the basis for a scientific web of publications and, of course, this leads naturally to the use of web technologies to navigate this web by converting citations and bibliography items into hyperlinks as part of a process of publishing documents on the web. Digital document spaces created by publishing on the web therefore offer advantages in terms of both search and navigation while paper documents are preferred for reading. Our goal is to obtain the best of both worlds by effectively extending the web to span paper and digital document spaces, allowing users to freely navigate between them using whichever medium is appropriate for a given task.

It has long been recognised that linear paper documents are usually easier to read than deeply nested and interlinked web pages and a number of research projects have investigated ways of linking paper to digital resources so that paper documents can be used for linear reading while serving as access points to textual or non-textual digital information. For example, the *Books with Voices* project [11] provides a paper-based user interface for fast random access to interviews stored in the form of digital videos. To link the paper documents to the videos, the transcripts are tagged with linear barcodes. As we will see later, it is important that the definition of these links can be generated automatically.

The *Origami* project [17] developed a paper-based user interface to the World Wide Web. Content-intensive web pages are often printed as it is more convenient to read larger texts on paper rather than to study them on a screen. However, by printing out a web page we lose some of its functionality and can no longer directly access hyperlinks available in the digital version. The goal of the Origami project was to bridge the gap between paper and web pages by also making the hyperlinks on the printed documents interactive. The application is based on an extended version of Wellner's *DigitalDesk* [23] where cameras and projectors are fixed above a desk and used to detect user actions and project digital information onto the desk, respectively. When a web page is printed out, an association between the printed document and the digital version is added to a registry. If a user later uses the paper document on the DigitalDesk and points to a specific hyperlink, the digital document copy will be retrieved based on the registry entry and the selected position is forwarded to the digital web page and interpreted as if it would have been a mouse selection. Finally, the linked information is accessed and the corresponding web page loaded. A major advantage of this approach is that no explicit link authoring is required and therefore no additional user effort is necessary to make the hyperlinks in a paper document interactive. However, the disadvantage of the DigitalDesk is that it assumes a fixed, physical environment and mobility is not supported.

Recent technological developments in digital pen and paper solutions allow for the possibility of encoding links on paper in a way that is non-disruptive to the content and more flexible than a simple tagging based on barcodes. Since interaction is based on the document itself rather than its physical environment, interaction can take place in any environment including on the move. In the following sections, we describe these technologies and show how they can be used to realise interactive paper. In particular, we present our Print-n-Link architecture which supports the automatic authoring of interactive paper versions of scientific publications.

3. INTERACTIVE PAPER

Our interactive paper framework is based on a classic client-server architecture as shown in Figure 1. On the client side, a digital pen connected to a computing device such as a regular desktop PC or a PDA is used to detect (x,y) coordinates within a paper document. In addition, the input device has to identify the document used and the page number within this document. The document's identifier and page number together with the positional information are transmitted via an HTTP request to the server components responsible for further data processing.

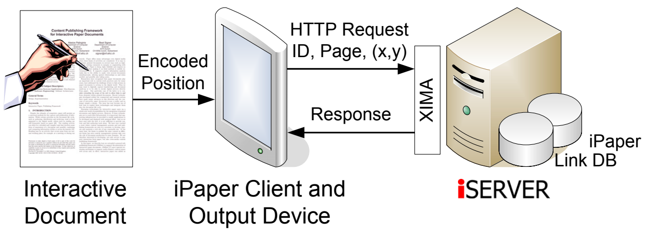

Figure 1: Interactive paper architecture

The server components consist of the eXtensible Information Management Architecture (XIMA), a framework for multi-channel access, and iServer, a cross-media link server. iServer and its iPaper plug-in are responsible for mapping a particular pen input to the appropriate information resource or digital service. XIMA then transforms the information output into the appropriate format, e.g. HTML for a regular web browser or VoiceXML for voice applications, and sends it back to the client device [20]. The transformation is based on a dynamically generated XML representation of the information stored within the iServer platform and all client-specific rendering is realised using the Extensible Stylesheet Language Transformation (XSLT). It is important to note that all operations on the server side are completely independent of a particular paper, printing or input device technology and we have implemented handlers for a range of input devices in the past [13].

The Print-n-Link system described in this paper is based on the *Digital Pen and Paper* technology [8] for high resolution paper-based position tracking developed by the Swedish company Anoto. The Anoto technology encodes information directly on each piece of paper using a special pattern of tiny printed dots. These dots are arranged with a small fixed horizontal or vertical displacement relative to the intersection points of a virtual grid covering the document page. Each dot encodes a two bit sequence as defined by its horizontal and vertical displacement and the digital pen's camera processes 6 by 6 arrays of these dots. The sequence of these 36 dots, each encoding two bits, results in a unique 72 bit sequence which then can be mapped to a global position within the Anoto virtual document space using a special transformation function. Note that the printing of the Anoto pattern results in a slightly grey document background with minimal interference with the printed artwork. Some issues of the printing process are discussed later in Section 6.

Several digital pens based on Anoto functionality are commercially available. For the realisation of the Print-n-Link prototype, we used the latest pen generation, the *magicomm G303* [14] produced by Hitachi Maxell, that can not only store information in the pen's memory but also provides built-in streaming functionality. The collection of additional data such as the pressure of the pen nib, orientation and tilt can be used to enhance the rendering process and create images that realistically recreate the look and feel of handwritten text and sketches. For the application presented in this paper, we used the magicomm G303 in streaming mode and it continuously streams its positional information to the iPaper Client using Bluetooth technology.

The iServer platform is a very general cross-media information platform that allows associations to be defined between arbitrary digital or physical resources. The core of the iServer architecture includes fundamental cross-media information concepts for linking, creating multiple virtual layers on top of resources and user management. Based on a resource plug-in mechanism, new resource types can easily be integrated into the iServer platform. iPaper is the plug-in responsible for handling pen input and the generality of iServer allows active areas defined on paper to be linked to or from any other media resources for which a plug-in has been implemented such as images, video, sound or HTML documents. A special feature of the iServer platform used in Print-n-Link is the possibility to also link to *active content* in the form of active components. While regular links just return an associated resource of information such as a video, active components are bound to a piece of Java program code. As soon as a linked active component is accessed, the corresponding active component is instantiated on the server (logic) as well as on the client side (stub). For example, for Print-n-Link, we implemented active components to access citation metadata and search for cited documents when the corresponding link is selected. More generally, if a user points to a paper-based button or active area, the corresponding instantiated active component may download a web page, contact a database, directly communicate with a running application or implement any other functionality. It is out of the scope of this paper to describe iServer and the iPaper plug-in in detail. Further information about the interactive paper framework and the underlying cross-media information infrastructure can be found in [19].

4. ENHANCED READING SYSTEM

To enhance the reading of scientific publications on paper, Print-n-Link takes into account different settings. For example, as two extremes, we will consider the cases where users are on the move with minimal equipment and no network access and where they are in the office with a desktop computer connected to the Internet, a monitor in front of them and a printer available. Even if these two settings are very different in terms of resources, our system supports the delivery of information relevant to the current document, deferring tasks which require network connectivity or a printer.

A user currently on the move can easily get access to the bibliographic information normally located at the end of the paper via an audio channel as depicted in Figure 2. This avoids them having to continuously interrupt the linear flow of reading by looking up a citation in the bibliography, possibly causing the reader to lose their position within the text and forcing them to re-read the paragraph.

With the help of our system, a reader may simply point with the digital pen to a citation embedded in the text and access details about it. Print-n-Link uses the internal structure of the document to retrieve the metadata of the selected reference listed in the bibliography. Normally, the system delivers information about the authors of a publication, its

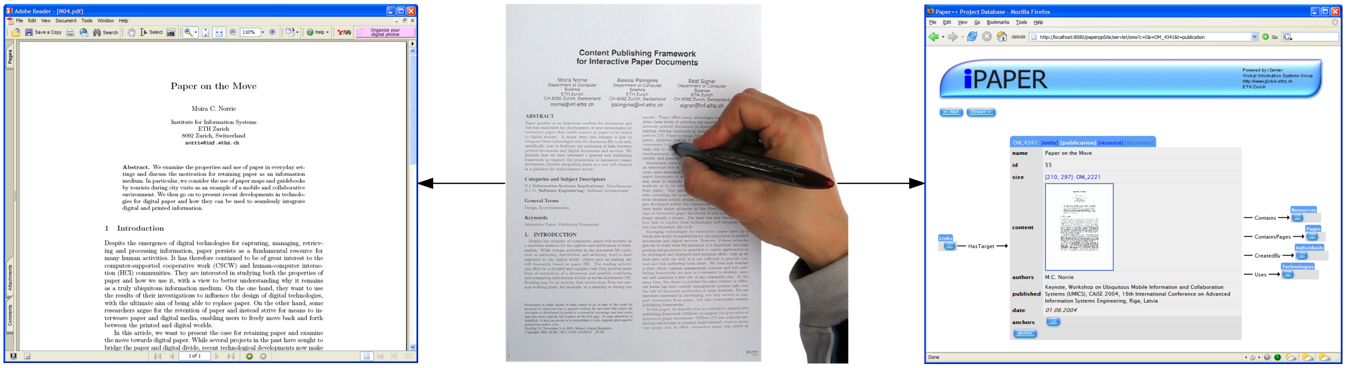

Figure 3: Retrieving a referenced document

title and the year, but a user may set general preferences or indicate the specific information of interest using written gestures as commands e.g. writing an 'a' to specify authors. This information is delivered using a text-to-speech engine via an audio channel. In mobile environments, this information is delivered using an earpiece. On-going developments in digital pens where the pen itself contains an integrated loudspeaker as used in the FlyPen educational toy [7] could provide alternative options suitable for use in non-public spaces.

In the office setting, the user has the possibility of accessing more information about selected references by retrieving the documents in question. This would normally be done manually using a digital library or search engine with a query containing some metadata of the publication found in the bibliography. Print-n-Link performs these tasks automatically for the user. A different interaction gesture with the digital pen, such as a double touch, tick gesture or writing a 'p' command, generates a request for a digital version of the cited document. Depending on the user configuration, either a PDF of the requested document will be displayed directly or a summary of its metadata together with a link to the PDF as illustrated in Figure 3.

If the user decides that they want to access a referenced publication while on the move and disconnected from the network, the reference is stored in a special bookmark list.

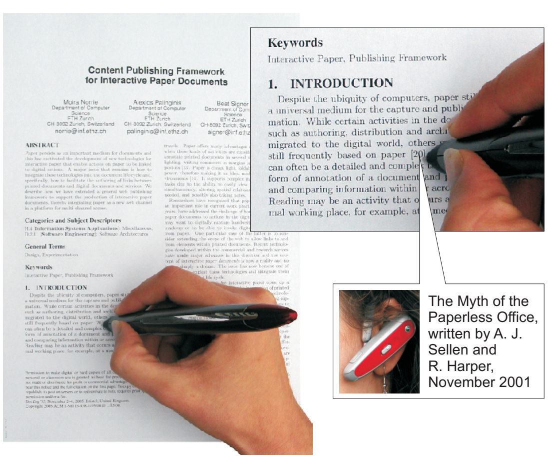

Figure 2: Delivery of audio information

Back in the office, the reader may review this list and select referenced documents to be retrieved.

It often happens that the reader quickly examines a cited publication and decides that they are actually interested in reading it and the next step is to print it out. Print-n-Link is able to directly produce an interactive publication from the retrieved PDF document during the printing process, thereby allowing the user to interact with that document in the same way as the original document. The user simply selects the Print-n-Link virtual printer and the system will analyse the PDF document and generate links automatically from the citations and bibliography as part of the printing process. In this way, our system dynamically builds a web that spans the paper and digital document spaces.

5. ARCHITECTURE

As already outlined, Print-n-Link is able to retrieve information about citations within a publication directly from the paper version of the document. A user can activate a simple citation retrieval service which delivers bibliographic information about the cited document through an audio channel using a text-to-speech engine. A reader using the system in an environment providing network connectivity, a display and a printer can also make use of a more complex document lookup service which is able to automatically retrieve the cited document based on the metadata in the bibliography. This document may be displayed on the user screen or printed out. The printed document then provides the same interactivity and functionality as the citing document. Figure 4 shows an outline of the overall architecture of Print-n-Link.

The system is composed of the following four main components: *iPaper* and *iServer* previously introduced in Section 3, the *Print-n-Link Engine* and *iDoc*. Depending on the current user's environment, these components are combined in different ways. Print-n-Link provides a mobile and a desktop version supporting the two different types of interaction described earlier.

The mobile client and the desktop application access the citation retrieval service in the same way. The iPaper client component receives position information from the digital pen, transforms it to the appropriate coordinate space and forwards this data to the iServer platform. The interactive paper plug-in performs a lookup in the database and, if the position delivered by the digital pen is bound to a citation link, iServer contacts the Print-n-Link engine which then

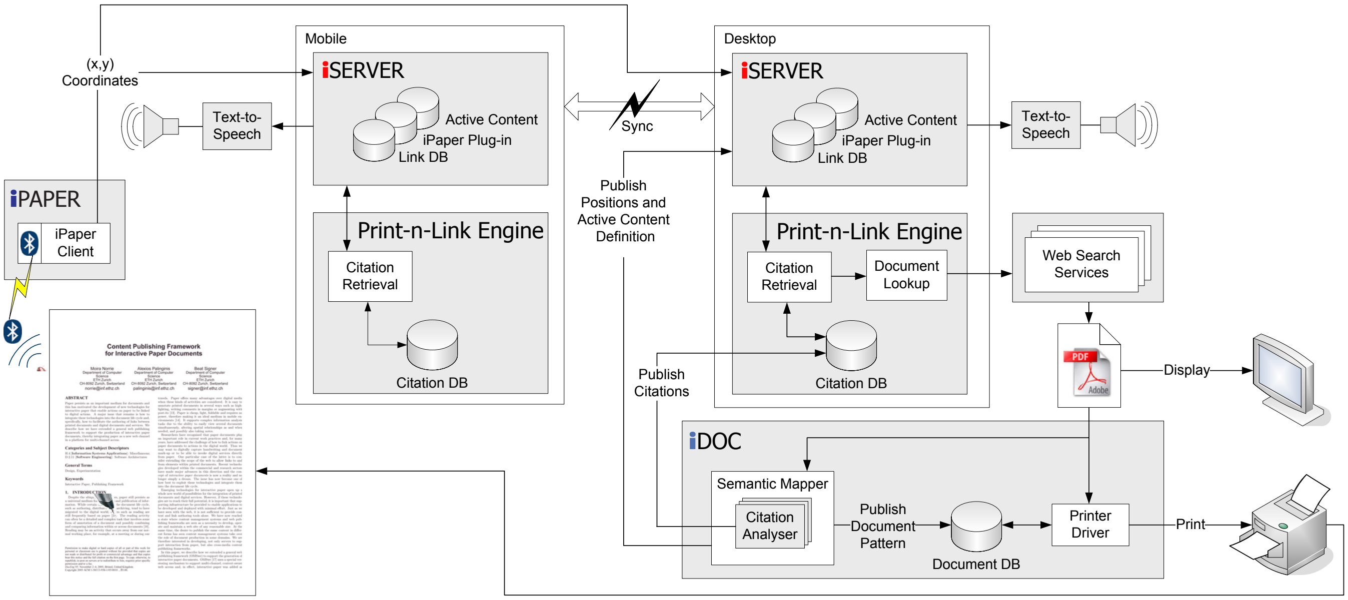

Figure 4: Print-n-Link architecture

retrieves the information about the citation. To get this information, Print-n-Link makes use of the citation retrieval service which searches a special citation database that stores information about the citations extracted from all of the PDF documents that have been analysed and printed by the system. This information includes details of the authors, the title, the year of publication, how it was published and optionally any direct link (URI) to a digital document version if available in a paper's bibliography.

If a copy of the cited publication is requested within the desktop application, Print-n-Link first checks the available metadata for a URI pointing directly to the document to be retrieved, for example in the form of a Digital Object Identifier (DOI). In the case that such a direct link is available, the corresponding digital document copy is downloaded and returned. Otherwise, Print-n-Link calls the internal document lookup service which may either query an external digital library or a search engine using the bibliographic metadata and returns the requested document if found. Another possibility would be to build an OpenURL based on the available metadata and forward it to an external OpenURL resolver to retrieve the most appropriate document copy. Other projects have investigated different means for locating scientific documents online [10]. Our document lookup service is freely configurable for the case that users may want to use any of these services or to query their personal, or any third-party, digital library. We are currently using a combination of public services such as CiteSeer and Google together with our own digital library of publications related to our research.

The search services return a link to a PDF version of the document or to a web page from where a link to the PDF document can be extracted. The retrieved document can then be displayed or printed as an interactive document using a special virtual printer driver which is part of iDoc. The iDoc component is a publishing framework responsible for managing the semantic content of a document. It enables the automatic authoring and publishing of links on paper and, in collaboration with iServer and iPaper, forms the basis for authoring, printing and accessing interactive documents. When a PDF document is sent to the virtual printer, its content is analysed for citation information and the document will be printed on the special Anoto pattern to enable its interactivity.

The iDoc component has two main parts: a printer driver, responsible for merging the source PDF document with the Anoto pattern and physically printing them out, and a semantic mapper which analyses the document in order to map the required semantic structure to the definition of active areas on the paper. The semantic mapper is based on a collection of plug-ins to analyse different types of documents. For Print-n-Link, we developed a citation analyser plug-in, which retrieves the position of each citation within a document and publishes the various types of information related to the citation to the appropriate server. Once the document enters iDoc, the printer driver contacts the document database within iDoc, selects the particular Anoto pattern space to be used for the printing process and prints the interactive document. At the same time, the semantic mapper analyses the document and publishes data about active areas corresponding to citations to iServer. The information extracted from the bibliography at the end of the document is published to the citation database within the Print-n-Link engine.

After these steps, the desktop application used to lookup the document, and print out the interactive version of it, contains information about its citations. If the user goes mobile, the citation database of the Print-n-Link engine and the iServer databases should also be updated on the mobile device. The user therefore has to synchronise the mobile client with the desktop application. In this way, the user may also access information about the documents printed through the desktop application in a mobile environment. Note that the mobile version of the system has a light version of the Print-n-Link engine where the document lookup service simply records the request for later processing and it has no iDoc component.

In a first implementation, single and double touches with the pen on a citation link are used to activate citation retrieval and document lookup services, respectively. We also plan to experiment with the use of different gestures—such as writing a tick on top of the citation to request a copy. Similarly, we are investigating what document metadata should best be provided via the audio channel and the use of different gestures to specify exactly what information is required.

6. VIRTUAL PRINTER

The printing of interactive documents requires authoring of the active areas and a special technology for merging the Anoto pattern with the source document. The iDoc component introduced in Section 5 is a publishing framework enabling the automatic generation and printing of interactive documents on the basis of either existing documents or information stored in a content database. In this section, we first describe how Print-n-Link uses iDoc to print documents based on a virtual printer service and then how a semantic mapper automatically creates links from citations in the printed document to digital services.

6.1 Printer Driver

The production of Anoto-enabled documents was originally aimed at large companies printing interactive paper forms to be used for the automatic capture and digitalisation of handwritten information. In order to print an Anoto-enabled document, a special license is needed, which defines the part of the Anoto pattern that should be used. There exist different licenses for asynchronous functionality, where the pen is placed in a cradle to transfer data, and synchronous communication where the pen sends data to the client continuously in streaming mode. Depending on the application, a developer will select one of the two license types. Parts of the pattern associated with the selected license, i.e. specific pages, are assigned to the Anoto-enabled document, which then may be printed out. Most of the applications based on Anoto technology rely on specialised printing companies taking over the task of printing the Anoto pattern on predefined source documents and currently there are few applications trying to print Anoto-enabled documents on-demand.

The printing of such a document normally consists of four steps as illustrated in Figure 5: producing the source document ①, creating a PDF version of it ②, merging the document with the Anoto pattern into a PostScript (PS) file ③, and printing out the PS file on a PS-enabled printer ④. While the first two steps require no special tools, the most critical task of merging the source document with the pattern requires software developed by Anoto. The tool normally used for this task is the Form Design Toolkit(FDT) [2], which is essentially an Adobe Acrobat plug-in enabling the production of Anoto-enabled PS files based on a PDF source document and an Anoto license.

To simplify this process, projects requiring on-demand printing of dynamic Anoto-enabled documents have investigated other strategies. For example, PADD [9] loads pre-printed pages with Anoto pattern into the printer tray prior to the printing of augmented paper documents. In this way,

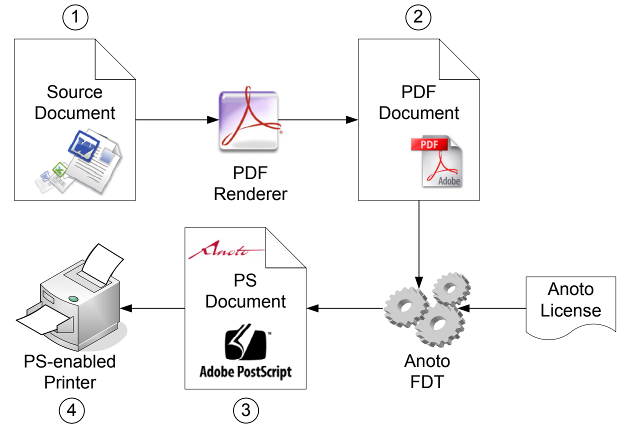

Figure 5: Anoto document production process

printing becomes a two-step process. This approach raises two other problems: a) the printing of the final documents requires an exact calibration of the pre-printed page, both in the orientation and in the positioning within the printer, and b) the pattern used must be selected in advance preventing checks for the unique allocation of pattern to documents at print time. By using such a system, we therefore lose part of the dynamic process of printing on-demand.

A second problem that often arises when using Anoto technology is the use of black colour in the printing of artwork. The Anoto pen is based on a micro camera which captures snapshots of the Anoto pattern. Light is emitted by the built-in infrared (IR) LED, reflected by the paper surface and captured by the camera. Since the Anoto dot pattern is printed with IR-absorbing black toner, the dots absorb the infrared light and the pattern will appear as white dots on the captured camera images. If the amount of black toner particles used for printing the artwork exceeds a certain limit, the printing will interfere with the pattern and the camera can no longer detect the dot pattern. This implies that no black colour should be used in areas that are to be interactive. Since most documents make use of black, this is a common problem. There are mainly two approaches to solving this problem: one is software-based and the other one deals directly with the printer cartridges.

In the *EdFest* project [4], we used a simple software approach where a document's black component was replaced with another colour (i.e. pure blue) during the authoring process. In the *Lost Cosmonaut* project [22], we instead substituted the black colour with a mixture of the three fundamental colours by setting the K component of the CMYK colour model to zero. This approach is basically a manipulation of the source document with a graphic editing tool which defines colours that look like black but do not contain a black component. Using this approach, care must be taken not to use a "smart" printer which will replace the CMY combination with a black component to save toner. Anoto also supports this approach in the FDT tool. The advantage is that the entire process is automated and does not require any user intervention.

The hardware approach consists of replacing the normal black cartridge of the printer with a special colour which does not interfere with the Anoto pattern. Anoto is proposing Anoto substitute black, which is a special non-infrared absorbing black to be used to print the artwork. The special black is currently available for offset printers only, but may also be produced for laser or ink-jet printers. Another approach was used in [9] where the black ink cartridge was simply removed from the ink-jet printer causing it to enter a reserve mode and automatically print black as a mix of the three other colours. A good printer may even support this as a real feature. For example, Oki printing solutions, gives the user the choice of using true black, based on the K cartridge only, or composite black, as a mix of the C, M and Y components. However, it is important to note that these solutions do not solve the problem of on-demand printing where real black is needed for the pattern. A software solution is therefore the only feasible approach when printing Anoto-enabled documents on-demand.

The iDoc printer driver enables one-step publishing of the interactive paper publication from any PDF source document. Once the PDF enters iDoc, the printer driver contacts the document database and checks which Anoto licenses have already been used and the relative pages assigned. It then selects a license accordingly and defines a new set of pages to be used. At the same time, the PDF is sent to the semantic mapper which is responsible for analysing its content. The printer driver is based on the Anoto Paper SDK [3], a Microsoft Windows library enabling the generation and printing of Anoto patterns. The core of the library, the Pad and Print Generation Module (PPGM), is outlined in Figure 6.

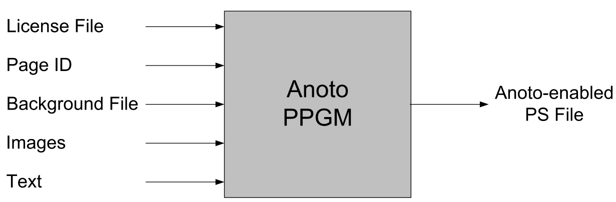

Figure 6: Pad and Print Generation Module

The iDoc printer driver uses only parts of the functionality defined by PPGM. The license and page information retrieved from the document database is used as an input for the PPGM and the source PDF file is processed by the PPGM page by page as a background file. Information about the pattern used is added as text and optional parameters can be set in order to enable the experimental black reduction and to set the size of the printed dots. The PPGM produces a PS file containing the combination of one page of the source PDF document and the corresponding Anoto pattern. The iDoc printer driver is responsible for first splitting the PDF file into single pages and then joining the resulting PS files into a single output document. The last step is to print out the interactive document. The printer driver allows the user to select the preferred printer and ensures that it is a PS-enabled printer.

6.2 Semantic Mapper

The authoring and publishing of interactive paper documents requires the definition of active areas on the paper document which are then linked to particular functionalities. For static documents, this can be done manually using our link authoring tool for interactive paper documents shown in Figure 7. With the help of the graphical link authoring tool, active areas in the form of arbitrary geometrical shapes can be defined based on an image generated from the PDF document and linked to the appropriate digital service or content. Furthermore, the specification of active areas can also be done manually by drawing them on paper. Our interactive paper framework also allows the definition of links to be imported based on an XML specification of the active areas.

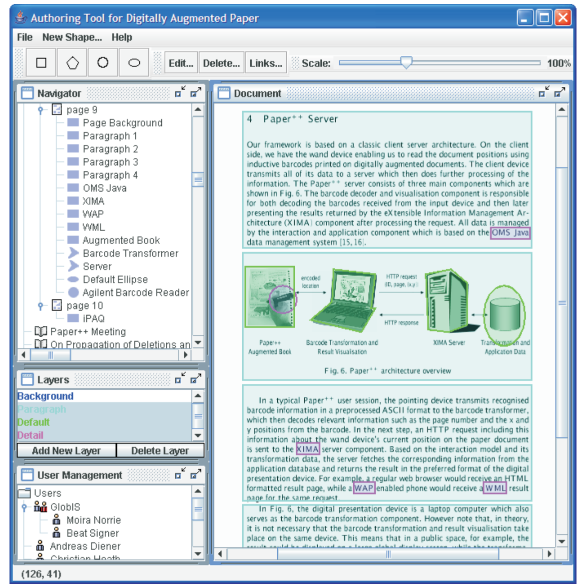

Figure 7: Link authoring tool

If a large number of active areas need to be defined or the authoring process is bound to the document publishing process in a dynamic way, tools are required to support the automatic authoring of links. For example, in the *Active Book* project [21], each physical page of the Active Book is represented as an HTML clickable map with links to digital resources. In this case, the authoring process is reduced to the production of such clickable maps. However, even if the process may be automated, this approach only allows for the authoring of a very limited amount of information.

In the *Open Journal* project [16] existing links stored in a separate database are added in a dynamic way to a PDF document. However these links were previously authored in a separate process and the dynamic authoring of links is not taken into account.

Our semantic mapper component is responsible not only for the dynamic inclusion of links into a PDF document, but also for the dynamic authoring of them during the publishing process. Since the nature of interactive documents varies greatly according to the application, the semantic mapper has a common core with a number of plug-ins developed for the analysis and authoring of different types of documents. For example, the *iPublish* plug-in of the semantic mapper was used to author interactive event brochures for a festival from a content database [4].

The semantic mapping of Print-n-Link is based on the citation analyser component that parses the PDF document in order to find citations within the document. The semantic mapper then calculates the position of a citation within a page. As outlined in Figure 4, the positional information is exported to the link server and associated with the Print-n-Link functionality. We now describe how this analysis is done in order that we can determine the precise coordinates of the area that should be made active.

The citation analyser investigates the structure of the text entries according to the PDF reference model [1]. Figure 8 shows a simple extract of a PDF document, where the position and the format of the word 'pdf' are encoded. Note that the PDF encoding format may vary based on the distiller component used to generate the PDF document.

```
1 BT
2   /F13 12 TF
3   288 720 TD
4   (pdf)TJ
5 ET
```

Figure 8: Text encoded in PDF format

The `BT` and `ET` markers set the beginning and end of the text object, respectively. The expression on line 2 specifies the font: `F13` is the Helvetica font, as described within the font subdictionary of the current resource dictionary, while `12` is the font size defined in terms of the current base unit (e.g. point). The position of the text with respect to the bottom and the left edges is defined on line 3. In this case, the text is positioned 720 points (25.4 cm) from the bottom edge and 288 points (10.16 cm) from the left edge.

Within a text string, the different characters are displayed through glyphs, which are specific graphical character renderings defined for each font family and stored in special purpose languages such as *Type 1* or *TrueType*. The rendering of each glyph is defined according to the standard width of each character within the text block. However, special layout requirements such as the text justification, require this distance to be adjusted independently for each text block. Figure 9 outlines the coordinate system of a typical PDF glyph, which forms the basis for the metrics between the different characters.

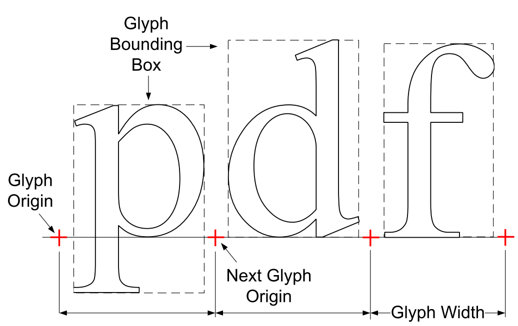

Figure 9: PDF glyph metrics

The `glyph origin` sets the position (0,0) where the glyph will be drawn. The origin of the next glyph is then set to a certain distance (x,y) from the origin. This forms the `next glyph origin`. Even if a displacement is possible in both the `x` and `y` directions, standard western authoring tools are based on the Latin alphabet, and therefore define a positive x-displacement and a y-displacement of zero.

The PDF format outlined in Figure 8 shows a simplified version of the PDF encoding. Since many PDFs are produced using Acrobat Distiller, we describe how the iDoc Semantic Mapper maps textual content to its position to paper in such documents. Figure 10 shows the encoding of one line of a PDF document, containing the reference '`[14]`' (at line 9) in the Acrobat Distiller format generated on a Microsoft Windows platform. Here an entire page is encoded between the `BT` and `ET` tags and every line, defined by a different `TJ` operator, is placed between square brackets. Before listing the content of the `TJ` tag, the font (`F4`) is defined.

```
1 BT
2 ...
3 /F4  1 TF
4 8.9664 0 0 8.9664 316.8 561.08 TM
5 0 -1.1644 TD
6 -0.0039 TC
7 [
8     (v)-10.2(i)0.5(ron)-8.4(m)-4(en)18.3(t)-5.7(s)
9     -414.9([14].)-695.4(I)-7.7(t)-407.2(s)0(u)-8.4(p)
10    -8.4(p)-35.2(ort)-5.7(s)-414.9(com)-4(p)-8.4(lex)
11    -425.1(i)0.5(n)-8.4(f)2.3(orm)-4(a)1.4(t)-5.7(i)
12    0.5(on)-436.7(an)-8.4(aly)-10.2(s)0(is)
13  ]TJ
14 ...
15 ET
```

Figure 10: Acrobat Distiller format

Line 4 defines the position of the text line within the page using the `TM` (transformation matrix) operator. This matrix contains 6 elements: the first and the third element (i.e `8.9664`) define the horizontal and vertical scale factor with respect to the base unit defined earlier. The last two numbers represent the position of the text block relative to the origin of the coordinate system, normally placed at the bottom-left of the page. The second and third numbers are normally set to zero for text, because they deal with more complex graphical effects such as rotating and skewing. The example shows a line placed 316.8 points (11.18 cm) from the left and 561.08 (19.79 cm) from the bottom of the page. Even if the transformation matrix already defines a position, some correction to it may be defined afterwards for each single `TJ` operator: the `TD` operator on line 5 moves the text position a little bit down, but leaves it at the same horizontal position. The spacing between the words, which is normally defined globally, may also be corrected: the `TC` operator on line 6 specifies the spacing with respect to the standard spacing, between the different words: `-0.0039` points.

The real text to be analysed is represented by the `TJ` operator. The text here is separated depending on the relative space between the single glyphs. If the space remains the same, a sequence of glyphs is represented within the same text block between parentheses. If the distance between the glyphs is not standard, it is specified by a value written between glyphs, as shown in lines 8–12.

The outlined PDF structure allows the extraction of relevant information from the PDF document. Previous projects investigating text extraction from PDF documents have focussed mainly on the content and general layout of the document rather than on the position of the text elements within a page, e.g. [5, 6]. By taking into account single text blocks, iDoc dynamically calculates the position of the text within the PDF document at different levels of granularity (paragraph, word, character), thereby enabling a semantic mapping between the position on paper and the digital version of single text blocks. Based on the outcome of this process, the citation analyser can search for any standard references like '[...]' within the PDF structure and retrieve the position of the text block. The reference '[14]', for example, is positioned at 102 mm from the top and 129 mm from the left edge. The success rate of detecting bibliographic references within a text document depends on multiple factors. The current implementation deals with the most common formats for references and their variations (e.g. a space before the reference, numerical characters only, commas between the square brackets, etc.) to detect bibliographic references. Note that some PDF renderers make use of special packages to automatically link the reference to the bibliographic listing. For example, the use of LaTeX in combination with the hyperref package results in *PDF links* which can be followed in the digital document version thereby increasing the chance that a matching pattern is a reference. By combining the analysis of the common patterns with any available PDF link metadata, Print-n-Link is able to find most references. However, we plan to experiment with more complex citation patterns in order to increase the detection rate.

```
<rectangle id="n_umics04_shape" resource="10639_norrie">
  <name>norrie_umics_04</name>
  <upperLeft>
   <point>
    <x>129</x><y>102</y>
   </point>
  </upperLeft>
  <size>
   <width>4</width><height>4</height>
  </size>
</rectangle>

<activeComponent id="n_umics_04_ac">
  <name>norrie_umics_04</name>
  <properties>
   <parameter>
    <key>org.ximtec.iserver.ac:oid</key>
    <value>norrie_umics_04</value>
   </parameter>
  </properties>
  <identifier>
   org.ximtec.ipaper.app.printnlink.CITATION_REQUEST
  </identifier>
</activeComponent>

<link id="n_umics_04_link"
sources="n_umics_04_shape"  targets="n_umics_04_ac">
  <name>norrie_umics_04</name>
</link>
```

Figure 11: Positions and active content

Once a reference has been detected and its position computed, the citation analyser has to retrieve the necessary information about the cited document from the bibliography that is normally located at the end of the document. After a text block has been recognised as a potential entry in the bibliography, the system checks its validity and computes a score based on parameters such as its position within the document and some predefined patterns such as the order of metadata in bibliography entries, the format of the text block and the position of the reference number relative to the text. A PDF link pointing from a reference to such a text block (i.e. the surrounding bounding box) increases the likelihood that the text block represents the desired bibliography entry. Finally, the text block with the highest score is retrieved and analysed in order to extract metadata about the author, the title of the citation, the year of publication etc. This information is transmitted to the citation database which stores the necessary metadata for each printed document. At the same time, iDoc contacts iServer and publishes the definition of new active content bound to the position of the citation within the PDF document. An XML representation of the information published to iServer is shown in Figure 11.

7. CONCLUSIONS

We have presented a system that can automatically print copies of scientific publications where the citations can be activated using a digital pen to retrieve information about these documents or initiate searches for the documents themselves. The approach is based on the use of Anoto functionality to activate links within documents and the analysis of PDF to generate the definitions of these links in terms of active areas. This is just one example of an application where there is a natural means of linking paper and digital document spaces and the approach is general and could be applied in many other applications. For example, it could be used to link between printed and digital versions of web pages or to link between printed and on-line editions of newspapers to mention just two.

8. ACKNOWLEDGEMENTS

We thank Monica Frisoni for implementing parts of the Print-n-Link prototype including the citation analyser and citation retrieval components.

9. REFERENCES

[1] Adobe Systems inc. *PDF Reference, Adobe Portable Document Format*, fifth edition, February 2006. Version 1.6.

[2] Anoto AB. *Form Design Tool User Guide*, February 2006. Version 2.1.

[3] Anoto AB. *Paper SDK Specification and Description*, February 2006.

[4] R. Belotti, C. Decurtins, M. C. Norrie, B. Signer, and L. Vukelja. Experimental Platform for Mobile Information Systems. In *Proceedings of MobiCom 2005, 11th Annual International Conference on Mobile Computing and Networking*, pages 258–269, Cologne, Germany, August 2005.

[5] J.-L. Bloechle, M. Rigamonti, K. Hadjar, D. Lalanne, and R. Ingold. XCDF: A Canonical and Structured Document Format. In *Proceddings of DAS 2006, 7th IAPR Workshop on Document Analysis Systems*, pages 141–152, Nelson, New Zealand, February 2006.

[6] H. Chao and J. Fan. Layout and Content Extraction from PDF Documents. In *Proceddings of DAS 2004, 6th IAPR Workshop on Document Analysis Systems*, pages 213–224, Florence, Italy, September 2006.

[7] Fly Pentop Computer, LeapFrog Enterprises, Inc., Emeryville, USA, http://www.flypentop.com.

[8] B. Forster. Writing to the Future. *Computerworld*, February 2001.

[9] F. Guimbretière. Paper Augmented Digital Documents. In *Proceedings of UIST 2003, 16th Annual ACM Symposium on User Interface Software and Technology*, pages 51–60, Vancouver, Canada, November 2003.

[10] G. Hoff and M. Mundhenk. Finding Scientific Papers with HomePageSearch and MOPS. In *Proceedings of SIGDOC '01, ACM Conference on Design of Communication*, pages 201–207, Santa Fe, USA, October 2001.

[11] S. R. Klemmer, J. Graham, G. J. Wolff, and J. A. Landay. Books with Voices: Paper Transcripts as a Tangible Interface to Oral Histories. In *Proceedings of CHI 2003, ACM Conference on Human Factors in Computing Systems*, pages 89–96, Fort Lauderdale, USA, April 2003.

[12] D. M. Levy. *Scrolling Forward: Making Sense of Documents in the Digital Age*. Arcade Publishing, October 2001.

[13] P. Luff, C. Heath, M. C. Norrie, B. Signer, and P. Herdman. Only Touching the Surface: Creating Affinities Between Digital Content and Paper. In *Proceedings of CSCW 2004, ACM Conference on Computer Supported Cooperative Work*, pages 523–532, Chicago, USA, November 2004.

[14] Magicomm G303, http://www.magicomm.co.uk/.

[15] C. C. Marshall. The Future of Annotation in a Digital (Paper) World. In *Proceedings of the 35th Annual GSLIS Clinic: Successes and Failures of Digital Libraries*, Urbana-Champaign, USA, March 1998.

[16] S. Probets, D. F. Brailsford, L. Carr, and W. Hall. Dynamic Link Inclusion in Online PDF Journals. In *Proceedings of EP '98, 7th International Conference on Electronic Publishing Document Manipulation and Typography*, pages 550–562, Saint-Malo, France, April 1998.

[17] P. Robinson, D. Sheppard, R. Watts, R. Harding, and S. Lay. Paper Interfaces to the World-Wide Web. In *Proceedings of WebNet '97, World Conference on the WWW, Internet & Intranet*, Toronto, Canada, November 1997.

[18] A. J. Sellen and R. Harper. *The Myth of the Paperless Office*. MIT Press, November 2001.

[19] B. Signer. *Fundamental Concepts for Interactive Paper and Cross-Media Information Spaces*. PhD thesis, ETH Zurich, 2005.

[20] B. Signer, M. C. Norrie, P. Geissbuehler, and D. Heiniger. Aural Interfaces to Databases based on VoiceXML. In *Proceedings of VDB6, 6th IFIP Workshop on Visual Database Systems*, pages 235–249, Brisbane, Australia, May 2002.

[21] I. Siio, T. Masui, and K. Fukuchi. Real-world Interaction Using the FieldMouse. In *Proceedings of UIST '99, 12th Annual ACM Symposium on User Interface Software and Technology*, pages 113–119, Asheville, USA, November 1999.

[22] A. Vogelsang and B. Signer. The Lost Cosmonaut: An Interactive Narrative Environment on Basis of Digitally Enhanced Paper. In *Proceedings of VS 2005, 3rd International Conference on Virtual Storytelling*, Strasbourg, France, December 2005.

[23] P. Wellner. Interacting with Paper on the DigitalDesk. *Communications of the ACM*, 36(7), July 1993.