

# DOCTRACK: AUTOMATIC PRINTED DIGITAL DOCUMENT TAGGING AND REMOTE RETRIEVAL

Thomas Holland and Aaron Quigley<sup>1</sup>

## **Abstract**

*We present the foundations of DocTrack, a digital and paper document identification, distribution, augmentation and interaction system. Printed documents are tagged with an encoded document identifier, allowing the digital version of the document content to be remotely recalled from its paper counterpart using a webcam or mobile device with a camera. We present results regarding the available document space, discuss the relevance of this approach in the light of existing research and outline the proposed applications of the system and future work planned to enable this.*

## **1. Introduction**

The original notion of the paperless office is a myth [5]. Printed digital documents retain an important role in modern business processes and will remain in the pervasive computing environment until technology becomes widely available that surpasses the benefits of printed digital documents at an appropriate cost.

Digital documents (such as PDF files) have many benefits over their paper versions, in terms of their low cost and ease of replication, distribution and storage. In addition, their ability to be quickly searched, easily edited and transformed is appealing. However printed documents on paper still maintain unrivalled abilities in terms of readability, portability and ease of interaction.

Printed digital documents suffer a decoupling from their original digital version once printed. This decoupling has been described as the paper-digital divide [3]. With manual intervention, the digital version used to print the paper version can be recalled and hence some benefits of the digital version regained. However, this relies on the careful and time consuming association of printed documents to their digital sources by those responsible for their printing.

## **2. DocTrack**

The goal of DocTrack is to explore document identification and retrieval without requiring additional hardware or major changes to the existing document printing process. This presented a series of novel research and engineering questions, which we decomposed into three stages.

---

<sup>1</sup>Department of Computer Science, University College Dublin, Ireland, email: {thomas.holland,aquigley}@ucd.ie. This work is partly funded under The Embark Initiative of the Irish Research Council for Science, Engineering and Technology.

## 2.1. Interception

When a document is printed, the print data is intercepted by a CUPS filter, and a globally unique document identifier added as an additional page to the print stream: the original print data is retained locally and (optionally) securely pushed to a remote storage system. The remote storage mechanism is envisioned to be globally distributed, replicated and fault tolerant, the likes of which are now becoming commercially available through services such as Amazon's S3. An additional authorisation layer would be added above the storage level to marshal requests for documents. The use of a CUPS filter allows for use of any postscript accepting printer without driver modification.

## 2.2. Identifier generation

The globally unique document identifier (Figure 1) is currently based on encoding 96 characters into a two dimensional barcode using the ISO/IEC 16022:2006 Data Matrix open standard (preferred over the patented QR Code). The first 64 characters form a machine-user identifier (composed of a SHA-256 hash of the primary MAC address of the machine concatenated with a private user key); the remaining 32 characters form a local document identifier (based on RFC 4122).

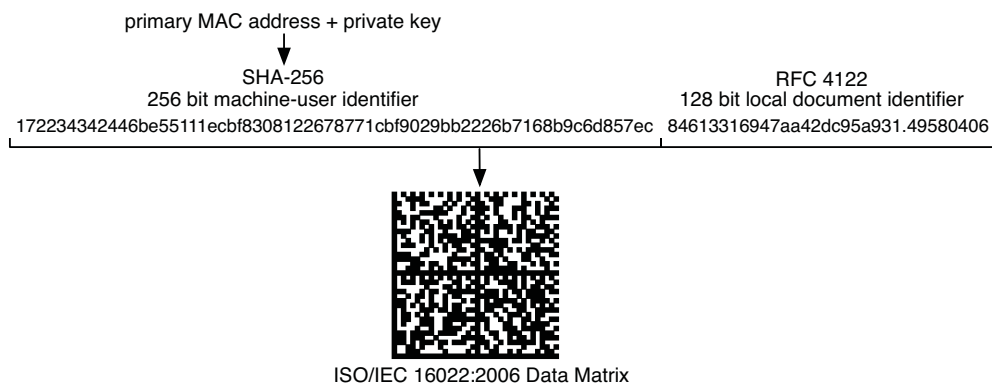


Figure 1. DocTrack globally unique document identifier generation

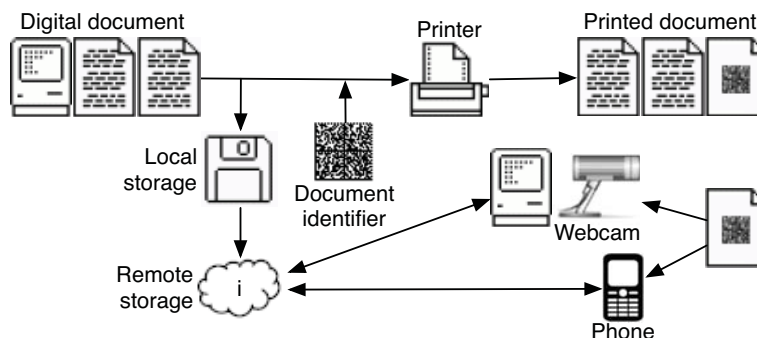


Figure 2. Flow of document content within the DocTrack system

## 2.3. Identifier decoding and document retrieval

Both a mobile and fixed camera video feed can capture document identifiers, which once decoded, can allow the original digital document to be retrieved. If the unique machine-user identifier from the document matches that of the current user and machine, the digital document can be retrieved locally.

If the document was not printed from that machine (or by that user), a document request is submitted to the remote storage system, where the access privileges for the user and requested document are compared and the document served if permitted. Figure 2 illustrates the overall flow of document content within the system.

### 3. Results

Combining a MAC address ( $2^{48}$  combinations, which the IEEE envision being adequate until the next century) with a private user key of any length as the input to a SHA-256 hash to form the machine-user identifier allows for a globally unique identifier to be locally generated (without having being allocated by a central authority, as is required in such proposed standards as Digital Object Identifier ISO/CD 26324). Rather than using a pseudo-randomly generated 256-bit identifier, a structured generation enables a user to associate several machines with themselves in the remote storage system and to be able to add to these as required, as they use new machines over time. Previous identifiers do not lose their validity and can still be used to recall documents long after the machine from which the original document was printed becomes defunct.

Successful decoding of an identifier is influenced by several factors, including the amount of data encoded, the physical size of the identifier and external factors which may hamper decoding (such as rotation, surface angle, damage, obstruction (including shadows), motion and low-light conditions). In staged conditions, assuming an undamaged and unobstructed barcode printed  $56\text{mm}^2$ , with standard office lighting, if both the document and camera are static and directly facing one another at a distance of 300mm, an identifier can be correctly decoded in 100% of captures (in the current prototype, capture of a still image for identifier decoding from the video feed must be manually initiated). In the same conditions, if the camera remains static, but a user holds the document, the chance of successful decoding drops to around 90%; if a user holds both the camera and the document, this drops to around 1 in 5 captures being decoded successfully.

### 4. Related work

Recent research efforts around bridging the paper-digital divide have focused on the capture of annotations from a printed digital document and how this can be used to augment the printed document with digital features [4]. The Anoto platform consists of a proprietary non-repeating pattern of dots some 4.6 million  $\text{km}^2$  in area that can be divided up into approximately 73 trillion letter-sized sheets of paper. By recording which pages of a digital document are printed onto which portion of the Anoto pattern, the original digital document can be determined. To match printed documents with their digital originals, Guimbretière [1] relied on use of pre-printed Anoto paper being loaded into the printer. The system has to maintain an accurate record of the particular parts of the Anoto pattern space with regard to the sheets of paper in the printer, which may cause synchronisation issues.

Use of physical digital pointers to enable physical objects to be given a reference to a digital resource is a technique with many possible implementations. Want et al. [6] used RFID tags to enable the identification of objects (including documents and books); the authors envisioned that the associated digital document would be displayed when the computer detects a tag. Use of barcodes to identify documents has already been commercialised by Wiziway, but at the requirement of a specialised device to read the document identifier. With the rise of the Web, URLs have also been visually encoded: Ljungstrand et al. [2] used one dimension barcodes printed onto stickers to encode URLs,

decoded by a barcode reader attached to a computer. UpCode and Semacode both offer commercial services for use in visual advertising campaigns, in which URLs are encoded using a two dimensional bar code and are designed to be captured using a mobile phone camera and deciphered with a Java program installed on a user's handset.

## 5. Conclusions and future work

We have presented the core functionality of DocTrack: a system designed to automatically tag printed digital documents with a unique visual identifier and enable the digital content to then be remotely accessible through use of a suitable device equipped with a screen and a camera. The results presented concern the available documents space, which is addressable through a globally unique machine-user identifier combined with a local document identifier, and issues surrounding the reliable decoding of the visual document identifier in real world scenarios.

Given the current groundwork in place, there are three key areas we will investigate. The first is support for a three-tier permissions system consisting of users (private documents), groups (shared documents) and others (public documents). With the current model, documents must be either private (only the original owner can retrieve them) or public (anyone can retrieve them). The practical uses of this system would be greatly enhanced by documents also being publishable to one or more groups of selected users. The second area is augmentation: once the digital content can be re-associated with its paper counterpart, what additional information could be presented through the display of the decoding device to aid the user (reference citation counts, author information, versioning details, annotations). The third area is interaction: what digital document features could be made available to the user through the decoding device (content search, digital distribution, group collaboration).

## References

- [1] François Guimbretière. Paper augmented digital documents. In *UIST '03: Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 51–60, New York, NY, USA, 2003. ACM.
- [2] Peter Ljungstrand, Johan Redström, and Lars Erik Holmquist. Webstickers: using physical tokens to access, manage and share bookmarks to the web. In *DARE '00: Proceedings of DARE 2000 on Designing augmented reality environments*, pages 23–31, New York, NY, USA, 2000. ACM.
- [3] Paul Luff, Christian Heath, Moira Norrie, Beat Signer, and Peter Herdman. Only touching the surface: creating affinities between digital content and paper. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 523–532, New York, NY, USA, 2004. ACM.
- [4] Moira C. Norrie, Beat Signer, and Nadir Weibel. Print-n-link: weaving the paper web. In *DocEng '06: Proceedings of the 2006 ACM symposium on Document engineering*, pages 34–43, New York, NY, USA, 2006. ACM.
- [5] Abigail J. Sellen and Richard H.R. Harper. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA, USA, 2003.
- [6] Roy Want, Kenneth P. Fishkin, Anuj Gujar, and Beverly L. Harrison. Bridging physical and virtual worlds with electronic tags. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 370–377, New York, NY, USA, 1999. ACM.