

White Paper

Beyond Document 2.0: The Future of Documents.

Francois Ragnet
Managing Principal Technology Innovation
Xerox Global Services
January 2007

Sophie Vandebroek
Chief Technology Officer and President,
Xerox Innovation Group
January 2007

Table of Contents

2	The future of documents and why it matters to your business.
2	The document is an agent for innovation.
3	The Smart Document: a key to business improvement.
3	Leaving Document 1.0.
4	Building a Document 2.0 infrastructure.
4	Automating work process through Document 2.0 technologies.
6	Document 2.0 services for a mobile work force.
6	The intelligent use of paper.
7	Moving toward Document 3.0.
7	The journey starts here.
8	About the authors

The future of documents and why it matters to your business.

We spend a lot of our time thinking about how to make information-based work easier, more efficient, and more effective.

To us, that means unleashing the full potential of what we at Xerox call Smart Documents.

No longer just a dumb, static container of information, today, when properly leveraged, the document is a powerful, intelligent vehicle for driving business processes.

The document is an agent for innovation.

To demonstrate the point that the document is an agent for innovation in any enterprise, let's start with an example from human resources:

- A new job application comes in via email or a job fair kiosk, in response to a specific job posting.
- As resumes come in, they are automatically categorized.
- A duplicate detection service notes that the candidate has already applied for a similar position and has a resume on file.
- The document enriches itself by spotting discrepancies between the two resumes. It also proactively goes out on the Web and fact checks some of the applicant's claims on his resume. It also consolidates his "records" in forums, publications, etc.
- Meta-data are extracted from the resume, such as previous employer names (experiences), people's names (references), dates and skills (technical jargon) and highlights them in different colors.
- The full document is transformed in conformance with a specific industry-standard, in this case, HR-XML, and extracts a semantic representation of the document, which also highlights any contradictions and discrepancies.
- An alert is sent to the manager's smartphone, providing him an interactive view of the resume's key elements (based on his criteria) and of the analysis. Upon review, the manager decides not to pursue the applicant.

- The resume is now a “Smart Document,” and after connecting with the ERP, it performs an intelligent search and identifies other job offers that match the resume. The resume personalizes and customizes its content, reprioritizing objectives and focusing on the most relevant experiences based on the varying requirements.
- It also redacts itself, obfuscating personal information that should not be visible to all recipients.
- The resume then forwards multiple personalized instances of itself to the various managers who had corresponding offers and waits for their reactions. One manager quickly triggers an interview and the applicant is promptly hired.
- The other versions of the resume validate that their peer’s status has changed to “closed” because the hiring transaction was completed, and automatically back out of the hiring workflow.
- A master copy of the resume is stored on the corporate ERP system.
- Six months later, the resume decides to refresh itself. It enriches itself by looking on the corporate network for new experiences that the employee has accumulated, and automatically updates itself.

The resume is transformed into a self-managing, intelligent agent, an integral part of the human resources hiring workflow.

Maybe some of you have already reached this higher ground. But I suspect that for most, the example above reads more like science fiction than a typical day in the HR department.

Let’s take a closer look at the evolution of the Smart Document and see how you can use them to transform the way your organization works.

The Smart Document: a key to business improvement.

The hiring process is of course only one example. Smart Documents can positively impact every critical aspect of your organization and your most important business processes.

Recent studies have shown that companies spend **anywhere from 10 to 15% of their revenue on “document-related costs.”** This is as much as three times the average amount they invest in R&D. Businesses produce more than 4.5 trillion hard copy

pages a year, supporting **80% of knowledge workers’ activities**, and **90% of customer communications**. Knowledge workers spend 20% of their day looking for information in documents – and 50% of that time, they can’t find what they’re looking for. The need has never been greater for new ways to capture, manage, and deliver information.

Fortunately, document technology and capabilities are evolving at an amazing rate, and the potential this evolution has for business improvement is huge.

To put this in perspective, consider the evolution of the Web from Web 1.0 to Web 2.0. From its inception as a collection of discrete sites, the Web has evolved into a more organic, collaborative, participatory exchange of information and viewpoints.

At Xerox, we believe the document has had a similar evolution, moving from Document 1.0 to **Document 2.0**, and now preparing for Document 3.0.

As we saw in the HR example earlier, with today’s technology, it is possible to add significant amounts of intelligence to the container – create a smarter document – to make the information inside easier to find, more dynamic, more secure and more useful to the organization, and allow the information to trigger automated work processes.

Organizations that harness the power of Document 2.0 will have a **tremendous competitive advantage** over those still in 1.0. And those who master 2.0 will be ready to quickly incorporate the new and still emerging capabilities of Document 3.0 to unlock the full potential of Smart Documents to improve the speed, efficiency and effectiveness of their operations.

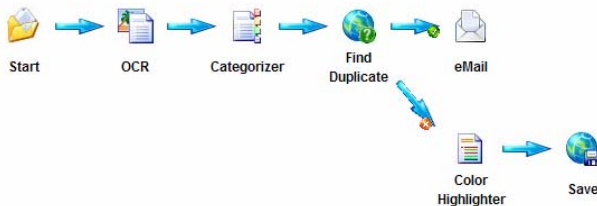
To help you find where you are on this evolutionary path, here is a brief overview.

Leaving Document 1.0.

Once the document became digital and networked, a new set of challenges appeared. How to manage paper and digital information? How to provide secure access? How to cost-effectively enable the use of color output? How to provide document services to support a mobile and wireless workforce? How to create information-based workflows that are faster and more automated?

Organizations can meet these challenges with a variety of, software, services and hardware such as color multifunction devices, scanning with OCR, rules-based printing, authentication and accounting tools and content management systems, etc. These are just some of the technologies that enable Document 2.0.

When combined as part of an overall document strategy, they lay the foundation for what we at Xerox call Smart Document Management.



It begins with a current-state assessment that looks at the entire document landscape – all of the documents, all of the document devices, all of the information workflows – and charts a document evolution path for the entire enterprise.

Building a Document 2.0 infrastructure.

Fully leveraging Document 2.0 requires **the right infrastructure** and the right **document services**.

This infrastructure will start with **an intelligent Document 2.0 repository system** such as DocuShare, which can not only provide versioning, tracking and collaborative editing, but also **intelligent workflow capabilities** and other Document 2.0 capabilities.

These Document 2.0 technologies will be **distributed on the network** and orchestrated intelligently through the infrastructure in a services oriented architecture (SOA). Accessible as services, they will either be tightly integrated for production-oriented environments where personalized communications need an infrastructure tightly linked to marketing databases and digital asset management systems. Or the SOA can be loosely integrated with Web services to support the daily casual activities of knowledge workers. In either case, the Document 2.0 infrastructure must be **connected with your own, existing IT infrastructure**, whether an ERP back-office or a CRM system, and allow asynchronous and complex document tasks.

All of these Document 2.0 technologies enable documents to be integrated into complex workflows, involving both **manual and automatic processing**. When combined with additional process evaluation and improvement services, such as Lean Six Sigma, as well as an in-depth observation of work practices, the results are a streamlined, efficient, highly automated business process.

This Document 2.0 infrastructure will also require high quality devices to support the **fluid integration of paper documents**. Typically, this will include some combination of office scanners and decentralized capture servers (e.g., [Xerox SMARTsend](#)) which support ad hoc scanning in branch offices, or high-speed, on- or off-site, high volume devices for paper-intensive processes and/or large amounts of legacy documents. Together, they are the on-ramps and off-ramps in the Document 2.0 world.

Automating work process through Document 2.0 technologies.

Whether paper or digital, the document needs to be **integrated into the business process it impacts**. Extracting its content and acting on it automatically, if possible, is crucial to the improvement of business processes.

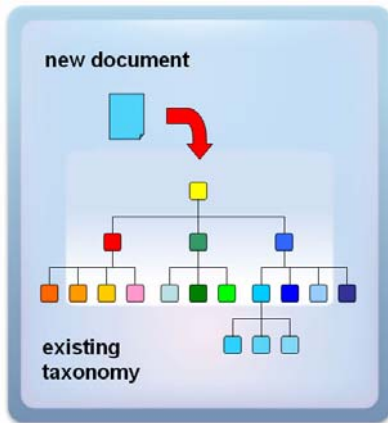
Consider the daily flood of incoming hard copy and electronic mail that must be sorted and reviewed before it is acted on. It's currently a time-consuming process that involves people reading the document and processing it, extracting the relevant information and ensuring that all other required information is available before taking action.

In the case of customer correspondence, document types may include a contract or a request for a change in personal information. Relevant information may be the customer's contract number and a new address. Other required information may be proof of a new address. When all of this processing is completed, the worker updates the customer database with the new information.

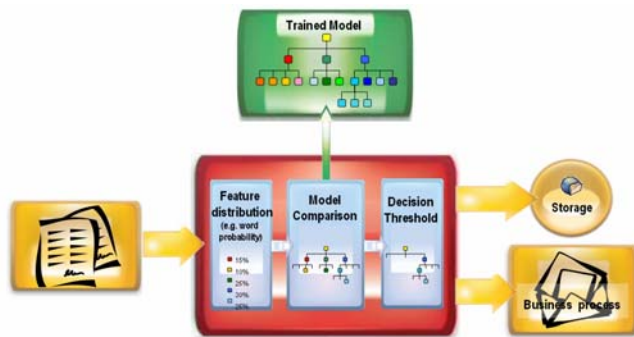
Scanning incoming hard copy mail is a key step in automating this process for correspondence.

But scanning alone won't achieve true process automation. Even scanned mail still requires an operator to manually review and process it.

Document 2.0 technologies allow the gradual automation of these types of business processes.

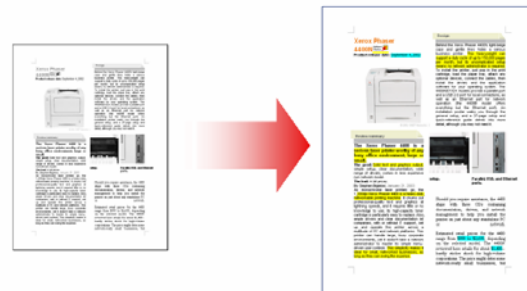


- **Classification**, whether based on the textual content or image content of the document, *allows the quick filtering or routing of the document to the right business process, archive or individual.* For example, keywords such as “cancel” or “cancellation” identify the incoming mail as a cancellation letter. But in general, one single word is not sufficient to determine the category of a document; in this case, it is necessary to provide more in-depth processing of its overall content to be able to decide its category with a high degree of confidence.



- **Advanced meta-data and fact extraction** uses both linguistic- or layout-based analysis, and can be used for indexing. In the mailroom example, analysis may include extracting customer ID, contract number, and/or address. Or it could be used for more advanced applications such as spotting interesting relationships between entities in a document, i.e., identifying the expiration of insurance policies in contracts.

- **Duplicate detection** avoids redundancy (and the associated waste of time and space) by identifying very early on whether a document or previous version already exist, such as mail that has already been received, reviewed and acted on, and therefore does not need to be processed. In the earlier HR example, two subsequent versions of the same applicant’s resume were identified and differences were highlighted.
- **Color highlighting** brings the most relevant information in the document to the attention of the user, if a person is required in the processing step. This could include contract expiration dates, especially if the contract has already expired. More generally, this optional highlighting can be used as a reading aid for documents, letting the user spot the interesting elements of that document.



All of these technologies and applications learn from human intervention and feedback to refine and improve their accuracy over time.

These techniques already exist for electronic documents. Ensuring they also work for hard copy documents is another difficulty. Indeed, scanned paper documents are already subject to Optical Character Recognition (OCR) errors, but more importantly, handwriting remains a very significant portion of these documents.

- **Handwritten “word spotting”** software allows a specific word to be sought for in documents, to quickly identify a document type, such as a cancellation letter, even in a handwritten document.
- **Annotation lifting** can separate the handwritten marks from the underlying form of the document, to identify what parts of the documents have been modified by a human in forms processing, validation and document type recognition.

Document 2.0 services for a mobile work force.

To support mobile workers and enable them to perform traditional work in any environment, *it is increasingly important to provide access not only to documents, but also to document services and collaborative tools.* These document services have to be readily available and easy to discover in an unfamiliar environment through wireless technologies.

More and more, “document services kiosks” allow consumers to use readily-available wireless infrastructures (3-4G, Wifi, and WiMax) to discover available local printing, scanning and other document services, and let hot spot users use them securely. These Mobile Document Services are aided by specialized, short-range protocols, such as Bluetooth, UWB, WiBree, or Wireless USB and FireWire.

The mobile user can be aided by a multi-purpose mobile device, typically a smartphone, which becomes the user’s “new office,” allowing him or her to print, scan and execute complex secure workflows from a personal device.

Even though they make most sense outside an enterprise, printing and scanning securely within a firewall continues to grow as a very important need, and is aided by small devices that can identify the user, such as SmartCards, USB flash drives, or others.

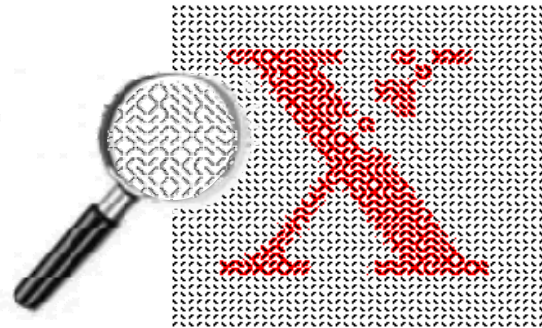
The intelligent use of paper.

Document 2.0 is **not only about electronic documents.** It is also about all of that paper that is still filling up file drawers, cluttering our desks, and piling up in device output trays. Paper reduction strategies are certainly a critical cost-cutting measure that all organizations must undertake, not to mention their importance from an environmental perspective. But the paperless office is unfortunately still far from reality. In addition to trying to replace paper documents entirely, *bringing more intelligence to the paper document might also be a viable option in the short- to mid-term.*

While creation and storage of documents has gone electronic, paper is still very popular for presenting information at meetings, for editing, reviewing, etc. Much of this paper output is for one-time, one-day use.

Low-cost RFIDs, for example, allow the tracking of a document in a business process as it goes through steps of a workflow, whether physical or virtual, thus allowing that paper document to be followed from the beginning to the end of its lifecycle.

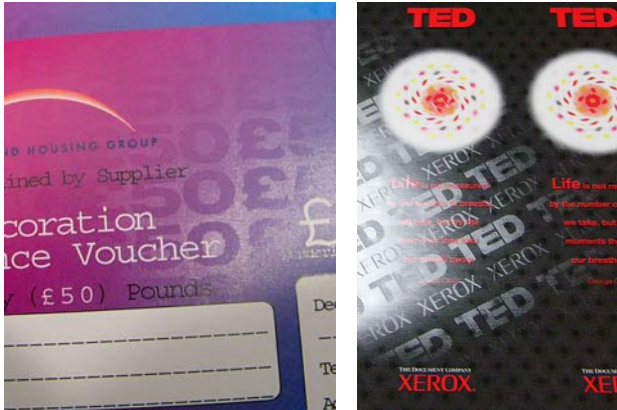
Similarly, **printed marks on paper** can convey machine-readable information which, while totally transparent and unobtrusive to the user, can help bridge the gap between the paper and electronic worlds. This information can be stored with very high density, for example with [DataGlyphs](#), and can carry meta-data, which becomes readily available to a computer system when scanning the document back in.



Alternatively, other types of technologies can be used to secure the content of a hard copy document, by adding invisible marks or watermarking patterns that can help recover these documents easily, or detect a fake.

Two examples are **Glossmarks**, which add a differential gloss to your color printouts that is visible under normal light; and **microprinting**, which allows the printing of information in extremely small fonts.

Both methods are very efficient and effective techniques to prevent counterfeiting.



The truth is, despite its shortcomings, when compared to electronic documents, *paper is still a crucial media and will continue to be central to document work for some time.*

For the time being, the focus must be on making paper documents smarter too – preserving their full content as they make their round trips between paper and digital worlds.

Moving toward Document 3.0.

Document 2.0 is here. Organizations in a wide range of industries are leveraging these capabilities to dynamically change the way their people work with information. They spend less time looking and more time doing. This is just the beginning: Smart Documents in a 3.0 architecture is the next frontier.

With Smart Documents, content and format will become more **dynamic and responsive**. They will take XML to the next level, providing customizable, semantic views of their content, tailored to the recipient.

Smart Documents will be **evergreen**. *They will be self-aware, capable of validating their content or proactively enriching themselves with other relevant content, becoming an organized “mash-up” of content collected from various locations. They will be time-conscious, capable of deciding when to refresh themselves with new content, or retire.*

When brought into a workflow, Smart Documents will **drive the process** or even adapt it, routing to appropriate reviewers or providing automated decision-making capability.

Smart Documents will become a collection of content provided by a network of contributors, **cross-fertilizing and capturing their knowledge.**

They will be capable of finding other “relevant” documents, creating a social network between documents and their authors, allowing the creation of knowledge from unstructured, heterogeneous information.

They will be **secure**, making their content available only to authorized viewers and will intelligently redact content based on personalized viewing permission rules.

When distributed, they will **dynamically recompose for presentation to multiple recipients** or output devices, based on each recipient’s known preferences, permissions and requirements. And Smart Documents will embrace new media. Whether rewritable e-Paper, multimedia e-book or 3D avatars, Smart Documents promise much more interactivity and customizability.

Although used more responsibly, **hard copy will still be a medium of choice**. When printed, Smart Documents will make the user aware of the latest version available, and suggest the most cost-efficient device to print on.

To make the most efficient use of paper resources, **“transient documents”** – documents whose lifespans are less than a day – will be printed using specially treated paper and an advanced marking process. The printed information will “disappear” after a few hours, letting the paper be reloaded into the output device, instead of the recycling bin.

This is not science fiction. This is a world taking shape right now. *The question you have to ask is... how long before you put your organization on its document evolution path?*

The journey starts here.

What’s the first step? You’ve already taken it by reading this paper. No matter where you are in this evolution, the business improvement possibilities go far beyond cost savings. When organizations begin to leverage the power of Smart Documents, they put themselves on an evolutionary path that leads to an optimized IT environment, and more importantly, unlocks the potential of their human environment.

With greater and faster access to dynamic information, people can communicate and collaborate more effectively, generate new ideas that create revenue and drive an enterprise forward.

About the authors



As Managing Principal, Technology Innovation, within Xerox Global Services, Francois Ragnet leads a team charged with transferring novel technologies into mainstream Xerox solutions offerings. Current initiatives focus on text-, image-, or feature-based categorization of documents, as well as identifying deeper semantic analyses which will enable Smart Document generation from traditional legacy formats or paper. His team also focuses on enhancing current offerings within the office environment to improve the efficiency of current products and streamline support processes.

Previously, Ragnet served as program manager and senior project leader for the Xerox Research Centre in Europe where his team provided innovative technologies in support of next-generation Xerox offerings. Their goal was to provide solutions that allow users to filter, configure and extract information from documents across advanced platforms that bridge production, printing and scanning. François was also a project leader in wireless technologies—specifically mobility and wireless (Bluetooth), content management, security, print and infrastructure management.

While at the Research Centre of Europe in Grenoble, France, Ragnet was a founding member of the Technology Showroom—a showcase of experimental technologies that hosts international events with customers from all over Europe.

Prior to joining Xerox, Ragnet was a researcher at the National Institute of Standards and Technologies in Gaithersburg, MD, where he focused on creation of a demonstration platform for state-of-the-art collaborative work technologies.

He holds a master's degree in telecommunications, from the Institut National des Telecommunications, Paris, France.

To learn more about Francois' work visit,

www.xerox.com/thoughtleadership_Ragnet

About the authors, cont.



Dr. Sophie Vandebroek is chief technology officer and president of the Xerox Innovation Group for Xerox Corporation. She was named to this position in January 2006, and became a corporate vice president in February 2006.

Vandebroek is responsible for overseeing Xerox's worldwide research centers and for maximizing the company's multimillion-dollar investment in research and technology.

Most recently, she was chief engineer of Xerox Corporation and vice president of the Xerox Engineering Center. As chief engineer, a position she assumed in 2002, Vandebroek was responsible for coordinating Xerox's engineering efficiency and effectiveness, a period during which Xerox refreshed more than 95 percent of its office product line and launched its flagship iGen3™ Digital Production Press.

Prior to that, she served as chief technology officer at Carrier Corp. From 1991 until 2000, Vandebroek held a number of increasingly responsible roles at Xerox including technical advisor to Xerox's chief operating officer and director of the Xerox Research Centre of Canada.

Vandebroek is a Fellow of the Institute of Electrical & Electronics Engineers and served as an elected member on the IEEE Administrative Committee. She is also a Fulbright Fellow and a Fellow of the Belgian-American Educational Foundation. She holds 12 U.S. patents.

Vandebroek has received awards from Xerox, IBM, HP, Monsanto, the Belgium National Science Foundation, Semiconductor Research Corporation, IEEE and Cornell University. Vandebroek served as a judge for MIT's Technology Review Young Innovators awards, the Wall Street Journal Innovation awards and the FIRST Lego and Robotics competition regional awards. She currently serves on several university and professional advisory boards.

Xerox Global Services

Xerox Global Services (XGS) delivers measurable results for companies in industries from aerospace and the financial services to healthcare, government and retail by looking at business challenges in a whole new way. Our more than 15,000 consultants and service delivery experts specialize in managing office assets and output, reengineering document-driven processes and optimizing print production environments.

We employ Smarter Document ManagementSM technologies to add intelligence and structure to both paper and digital documents and activate the content they contain. And we combine our extensive industry expertise with tools like Lean Six Sigma to create a powerful portfolio of services that adds real value to enterprises worldwide.

For more information on Xerox Global Services, visit: www.xerox.com/globalservices

* * *

©2007 XEROX CORPORATION. All rights reserved. Xerox[®], Smarter Document ManagementSM and There's a new way to look at it[®] are registered trademarks of Xerox Corporation.