Original research / Artículo original / Pesquisa original - Tipo 1

Free software for digitalization and management of electronic documents at official entities

Javier López Martínez / jlopez358@unab.edu.co

Universidad Autónoma de Bucaramanga, Bogotá-Colombia

ABSTRACT Taking into account the current regulations in Colombia about document management and best practices, a technological solution for digitization and document management in public sector entities was designed, following the Research Applied method. The solution includes the integration of two Web systems focused on free software. The first one, called FuidXel, is a development in PHP language, original of the present project, that includes a tool next to the client for digitization, and a Web application, next to the server, for the conformation of the document, that consists on libraries of tools for the treatment of images and PDF files. The second system is an enterprise content manager for managing electronic documents, called Alfresco. FuidXel integrates with Alfresco through CMIS protocol for the sending of PDF documents, made up of the product images of the digitization and management, information and traceability metadata.

KEYWORDS Web; digitalization, ECM; document; PDF; free software.

Software libre para la digitalización y gestión de documentos electrónicos en entidades oficiales

RESUMEN Teniendo en cuenta la normativa vigente en Colombia sobre gestión documental y sus mejores prácticas, se diseñó una solución tecnológica para la digitalización y gestión documental en entidades del sector público, siguiendo un método de investigación aplicada. La solución comprende la integración de dos sistemas Web enfocados al software libre. El primero, llamado FuidXel ,es un desarrollo en lenguaje PHP, original del presente proyecto, que comprende una herramienta al lado del cliente para digitalización; y una aplicación Web al lado del servidor, para la conformación del documento, que consta de librerías de herramientas para el tratamiento de imágenes y archivos PDF. El segundo sistema es un gestor de contenido empresarial para la gestión documentos electrónicos llamado Alfresco. FuidXel se integra a Alfresco mediante un protocolo CMIS para el envío de los documentos PDF conformados por las imágenes producto de la digitalización y los respectivos metadatos de gestión, información y trazabilidad.

PALABRAS CLAVE Web; digitalización; ECM; documento; PDF; software libre.

Software livre para a digitalização e gestão de documentos eletrônicos em entidades estatais

RESUMO Atendendo às normas vigentes na Colômbia sobre gestão de documentos e as suas melhores práticas, foi desenhada uma solução tecnológica para a digitalização e gestão de documentos em entidades do setor público, seguindo um método de pesquisa aplicada. A solução inclui a integração de dois sistemas web dirigidos ao software livre. O primeiro, chamado FuidXel, é um desenvolvimento em linguagem PHP, original deste projeto, que inclui uma ferramenta do lado do cliente para a digitalização; e uma aplicação Web do lado do servidor, para a formação do documento, que consta de bibliotecas de ferramentas para o processamento de imagens e arquivos PDF. O segundo sistema é um gestor de conteúdo corporativo para a gestão de documentos eletrônicos chamados Alfresco. FuidXel integra-se com o Alfresco através de um protocolo CMIS para o envio de documentos PDF formados por imagens produto da digitalização e os metadados associados de gestão, informação e rastreabilidade.

PALAVRAS-CHAVE Web; digitalização; ECM; documento; PDF; software livre.

Received / Recepción: March 11, 2017 - Accepted / Aceptación: March 31, 2017 *doi:10.18046/syt.v15i40.2392*

I. Introduction

This project starts from the problems that exist in the public sector regarding the application of the current legislation in the processes related to the massive digitization of documents and the management of electronic documents. In the public sector entities in Colombia, the congestion in the handling of paperwork is evident, in some cases, due to the use of paper documents in large volumes. Although there may be well-structured mechanisms for archiving such documents, their high quantity can collapse physical storage spaces, affecting the times of customer service and the quality of life of officials who operate with those documents.

Although the tools to perform the digitization of documents are well known, their implementation in the public sector is complex because, in addition to supporting the massive process of documents, they must take into account the application of norms established by the National Government. It is imperative that public sector entities comply with national regulations in relation to document management, more precisely those regarding to the management of electronic documents, because in addition to being legal systems and state regulations, as stated in the Presidential Directive 04 of 2012, their non-implementation would avoid the improvement of the management processes and would affect the normal operation of the procedures that the citizens should carry out.

The general objective of the research project was to design a technological solution that allows digitization in the management of electronic documents for public administration institutions in Colombia, using free software tools. To achieve it, information was collected in a theoretical framework and the applied research method was used; as a result of the investigative process, the need to follow an iterative and incremental software development life cycle was evidenced. The conclusions of this process correspond to the solution produced and to the evidences found that are summarized at the end of this article.

II. Document management

Document management is an essential part of the operation and purpose of public sector entities. Physical and electronic documents are indispensable in the administrative, technical and technological, human resources, judicial and financial areas, among others. The management of these documents is the responsibility of all the employees and officials of the entity. The volume of documents, in most public sector entities, is high, requiring physical and electronic spaces suitable for storage and preservation (Larrañaga, 2008).

I. Introducción

Este proyecto parte de la problemática que existe en el sector público respecto de la aplicación de la normativa vigente en los procesos relacionados con la digitalización masiva de documentos y la gestión de documentos electrónicos. En las entidades del sector público en Colombia es evidente la congestión en la atención de los trámites, en algunos casos, debido a la utilización de documentos en papel en grandes volúmenes. Aunque pueden existir mecanismos bien estructurados para archivar dichos documentos, su cantidad elevada puede colapsar los espacios físicos de almacenamiento, lo que afecta los tiempos de atención al público y la calidad de vida de los funcionarios que operan con esos documentos.

Aunque ya son bien conocidas las herramientas para realizar la digitalización de documentos, su implementación en el sector público es compleja ya que, además de apoyar el proceso masivo de documentos, deben contar con la aplicación de normas establecidas por el Gobierno Nacional. Es imprescindible que las entidades del sector público cumplan con la normativa nacional en relación con la gestión documental, más precisamente con la que tiene que ver con la gestión de documentos electrónicos, porque además de ser ordenamientos jurídicos y ordenamientos estatales, como lo indica la Directiva Presidencial 04 de 2012, su no aplicación evitaría el mejoramiento de los procesos de gestión y afectaría el normal funcionamiento de los trámites que deben realizar los ciudadanos.

El objetivo general del proyecto de investigación fue diseñar una solución tecnológica que permita la digitalización en la gestión de documentos electrónicos para instituciones de administración pública en Colombia, usando herramientas de software libre. Para lograrlo, se recopiló información en un marco teórico y se utilizó el método de investigación aplicada; como resultado del proceso investigativo, se evidenció la necesidad de seguir un ciclo de vida de desarrollo software iterativo e incremental. Las conclusiones de este proceso corresponden a la solución producida y a las evidencias encontradas que se resumen al finalizar este artículo.

II. Gestión documental

La gestión documental es una parte esencial del funcionamiento y la finalidad de las entidades del sector público. Los documentos físicos y electrónicos son indispensables en las áreas administrativas, técnicas y tecnológicas, de recursos humanos, judicial y financiera, entre otras. La gestión de dichos documentos es responsabilidad de prácticamente todos los servidores y empleados de la entidad. El volumen de documentos, en la mayoría de las entidades del sector público, es elevado, por lo que se requiere de espacios, tanto físicos, como electrónicos, adecuados para su almacenamiento y preservación (Larrañaga, 2008).

De acuerdo con la normativa colombiana, la gestión documental es el

> (...) conjunto de actividades administrativas y técnicas tendientes a la planificación, manejo y organización de la documentación producida y recibida por las entidades, desde su origen hasta su destino final con el objeto de facilitar su utilización y conservación. (Ley 594 de 2000)

Las entidades oficiales en Colombia deben tener en cuenta la aplicación de un extenso número de normas en los procesos de gestión documental. Como referencia más amplia se puede revisar el marco legal del Programa de Gestión Documental del Ministerio de las Tecnologías de la Información y las Telecomunicaciones [MINTIC] (Márquez & Chacón, 2014). A continuación se presentan las más relevantes relacionadas con la gestión de documentos electrónicos.

- la Ley 527 de 1999, que reglamenta el uso de los mensajes de datos, el comercio electrónico y las firmas digitales, y establece las entidades de certificación;
- la Ley 594 de 2000, ley general de archivos, conformación, organización y preservación de archivos públicos, que se refiere a la incorporación de tecnologías de avanzada en la administración y conservación de archivos;
- el Decreto 2609 de 2012, mediante el cual el Archivo General de la Nación [AGN] y el MINTIC establecen los lineamientos generales que regulan el expediente electrónico;
- la Circular Externa 005 de 2012 del AGN, que presenta recomendaciones para llevar a cabo los procesos de digitalización y las comunicaciones oficiales electrónicas en el marco de la iniciativa de cero papel;
- el Decreto 019 de 2012, ley anti-tramites, por el cual se dictan normas para suprimir o reformar regulaciones, procedimientos y trámites innecesarios de la administración pública e incluye, entre las disposiciones más importantes, la exigencia a las entidades del sector público de adoptar tecnologías que permitan agilizar los trámites;
- la Directiva Presidencial 04 de 2012, mediante la cual se solicita a las entidades públicas la reducción de papel mediante el uso de tecnologías para el uso de medios electrónicos en la gestión documental del Estado; y
- el Acuerdo 005 de 2013 del AGN, por el cual se establecen los criterios básicos para la clasificación, ordenación y descripción de los archivos en las entidades públicas y privadas que cumplen funciones públicas.

El AGN y el MINTIC, dando alcance a la normativa nacional, sugieren u ordenan seguir normas internacionales para establecer un estándar en los diferentes aspectos que aborda la gestión documental, respecto de las tecnologías de la información, esto es:

According to Colombian regulations, document management is the

> (...) set of administrative and technical activities tending to the planning, management and organization of the documentation produced and received by the entities, from their origin to their final destination in order to facilitate their use and conservation. (Law 594 of 2000)

Official entities in Colombia must take into account the application of an extensive number of regulations in document management processes. As a broader reference, the legal framework of the Document Management Program of the Ministry of Information Technologies and Telecommunications [MINTIC] can be reviewed (Márquez & Chacón, 2014). The following are the most relevant regulations related to the management of electronic documents.

- Law 527 of 1999, which regulates the use of data messages, e-commerce and digital signatures, and establishes the certification entities;
- Law 594 of 2000, general law of archives, conformation, organization and preservation of public archives, which refers to the incorporation of advanced technologies in the administration and conservation of archives;
- Decree 2609 of 2012, by means of which the General Archive of the Nation [AGN] and the MINTIC establish the general guidelines that regulate the electronic document;
- AGN External Circular 005 of 2012, which presents recommendations for carrying out digitization processes and official electronic communications within the framework of the zero-paper initiative;
- Decree 019 of 2012, anti-procedural law, which establishes rules to suppress or reform unnecessary regulations, procedures and formalities of public administration and includes, among the most important provisions, the requirement to public sector entities of adopting technologies to streamline procedures;
- Presidential Directive 04 of 2012, which requires public entities to reduce paper through the use of technologies for the use of electronic media in the document management of the State; and
- AGN Agreement 005 of 2013, which establishes the basic criteria for the classification, planning and description of archives in public and private entities that perform public functions.

The AGN and the MINTIC, reaching the national regulations, suggest or order to follow international regulations to establish a standard in the different aspects that addresses document management, regarding information technologies, that is:

- the Model requirements for the management of electronic records [MoReq2], a document published by the European Union which meets the requirements for an electronic document management system (EC-IDABC, 2004); and
- the General International Standard Archival Description of the International Council of Archives [ICA] (2011), which establishes parameters, structures and metadata for describing documents in archives.

III. Support tools

At the international level, document management support tools are identified by the acronym ECM [Enterprise Content Management] or EDMS [Enterprise Document Management System], that are oriented to: the management of enterprise contents; storage, preservation, search, display of electronic documents in different formats; workflows, metadata assigning, user management, document version management, and document editing with office tools.

The main tools to support document management in the area of free software, described below are: Alfresco, Nuxeo and OpenKM Community (Jiménez, 2016, Franco & Rosas, 2015).

- Alfresco. Developed in the JAVA programming language, includes a content manager, a web portal to manage and access content, performs searches in the Apache Solr-Lucene engine, and workflows with JBPM standard. Although it has enterprise versions that allow subscription to the direct support of the company, the Community Edition has a LGPL [Library General Public License] that defines it as free software.
- Nuxeo. ECM [Enterprise Content Management] similar to Alfresco allows content management collaboratively, workflows and integration with MS-Office and Open Office; it is developed in JAVA programming language and its configuration and installation is simple compared to Alfresco.
- OpenKM Community. Document and records management; has a web 2.0 user interface based on the GWT [Google Web Toolkit] framework and a security layer that centralizes access management for users.

- el *Model requirements for the management of electronic records* [MoReq2], documento publicado por la Unión Europea que reúne los requerimientos para un sistema de gestión de documentos electrónicos (EC-IDABC, 2004); y
- la Norma Internacional General de Descripción Archivística, del Consejo Internacional de Archivos [ICA] (2011), donde se establecen parámetros, estructuras y metadatos para la descripción de documentos en los archivos.

III. Herramientas de apoyo

A nivel internacional las herramientas de apoyo a la gestión documental se identifican por las siglas ECM [*Enterprise Content Management*] o EDMS [*Enterprise Document Management System*], se orientan a: la gestión de contenidos empresariales; el almacenamiento, preservación, búsqueda, visualización de documentos electrónicos en diferentes formatos; flujos de trabajo, asignación de metadatos, administración de usuarios, gestión de versiones de documentos, y edición de documentos con herramientas ofimáticas.

Las principales herramientas de apoyo a la gestión documental en el ámbito del software libre, descritas a continuación, son: Alfresco, Nuxeo y OpenKM Community (Jimenez, 2016; Franco & Rosas, 2015).

- Alfresco. Desarrollado en lenguaje de programación JAVA, incluye un administrador de contenidos, un portal web para administrar y acceder al contenido, realiza búsquedas con el motor Apache Solr-Lucene y flujos de trabajo con estándar JBPM. Aunque cuenta con versiones empresariales que permiten la subscripción al soporte directo de la empresa, la versión *Community Edition* tiene una licencia LGPL [*Library General Public License*] que la define como software libre.
- Nuxeo. ECM [*Enterprise Content Management*] similar a Alfresco, permite la gestión de contenido de manera colaborativa, flujos de trabajo y la integración con MS-Office y Open Office; está desarrollado en lenguaje de programación JAVA y su configuración e instalación es sencilla en comparación con Alfresco.
- OpenKM Community. Administrador de gestión documental y de registros; tiene una interfaz de usuario web 2.0 basada en el *framework GWT* [*Google Web Toolkit*] y una capa de seguridad que centraliza la gestión de acceso a los usuarios. Permite autenticación LDAP [*Lightweight Directory Access Protocol*], CAS (*Centralized Authentication Service*) o por medio de base de datos; maneja un motor de flujo de trabajo JBPM.

La extensibilidad e interoperabilidad son condiciones necesarias en las herramientas de apoyo a la gestión documental, debido a que afectan procesos de forma transversal y a que normalmente se requiere que se adapten a otros sistemas o se adecuen a nuevas características para

un proceso en particular. Las herramientas de apoyo a la gestión documental regularmente tienen plataformas de interoperabilidad con diferentes clases de tecnologías, a continuación, se mencionan las más usadas e importantes.

- CMIS [*Content Management Interoperability Services*], estándar abierto que permite el acceso de manera única a los sistemas de gestión de contenidos, establece mecanismos para manejar contenidos, metadatos de contenidos, contenidos de carpetas, asociaciones y transferencia de archivos a nivel de aplicación; existen dos enlaces del protocolo, usando SOAP y, por otro lado, REST, utilizando la convención AtomPub.
- REST [*Representational State Transfer*], originalmente se refiere a un conjunto de principios de arquitectura para transferencia de datos por medio de protocolo HTTP, actualmente se usa en un sentido más amplio, para describir cualquier interfaz que utilice dice protocolo para intercambio y operaciones sobre los datos sin utilizar abstracciones en patrones adicionales de intercambio de mensajes –como lo hace SOAP–. En REST se puede utilizar cualquier formato (XML, JSON, etc.), cuenta con unos parámetros funcionales clave, un protocolo cliente servidor sin estado y operaciones bien definidas, y usa hipermedios.
- SOAP [*Simple Object Access Protocol*], especificación de un protocolo que permite la estructuración de información en la implementación de servicios web dentro de redes computacionales, se basa en la capa de protocolos de aplicación y es usado regularmente en protocolos HTTP y SMTP. Está conformado por una pila de protocolo de servicios web basada en XML y consiste de tres partes: envoltura, la cual define la estructura del mensaje; un conjunto de reglas que expresan instancias de aplicación y tipos de datos; y una convención para definición de procedimientos de llamadas y respuestas. Un mensaje SOAP es un XML que contiene los elementos: envoltura, encabezado, cuerpo y error.

IV. Digitalización

Dentro de la gestión documental, los procesos de digitalización dan origen a los documentos electrónicos a partir de los documentos físicos. El uso de documentos electrónicos busca hacer más eficiente la gestión documental ya que se disminuye el uso de documentos físicos y sus problemas derivados, como el consumo de espacio en las oficinas, el transporte interno de papel en grandes cantidades y el riesgo de daño de las hojas. La digitalización se define como el "el procedimiento tecnológico por medio del cual se convierte un soporte análogo (papel) o electrónico, en una imagen digital" (MINTIC, 2012a).

En el ámbito de la gestión documental existen diferentes tipos de digitalización, de acuerdo con su finalidad. Los procesos de digitalización se clasifican en dos grandes

Allows authentication in LDAP [Lightweight Directory Access Protocol], CAS [Centralized Authentication Service] or through database; manages a JBPM workflow engine.

Extensibility and interoperability are necessary conditions for support tools in document management, because they affect processes in a cross-cutting way and are normally required to adapt to other systems or adapt to new characteristics for a particular process. The support tools in document management regularly have interoperability platforms with different kinds of technologies, the most used and important are mentioned below.

- CMIS [Content Management Interoperability Services] is an open standard that allows uniquely access to content management systems, establishes mechanisms to manage content, content metadata, folder contents, associations and file transfer at an application level; there are two protocol links, using SOAP and, on the other hand, REST, using the AtomPub convention.
- REST [Representational State Transfer], originally refers to a set of architecture principles for data transfer through HTTP protocol, it is currently used in a broader sense, to describe any interface that uses protocol for exchange and operations on the data without using abstractions in additional patterns of message exchange –as SOAP does–. In REST any format (XML, JSON, etc.) can be used, it has a few key functional parameters, a non-state server client protocol and well-defined operations, and it uses hypermedia.
- SOAP [Simple Object Access Protocol], is the specification of a protocol that allows the structuring of information in the implementation of web services within computational networks, it is based on the protocol layer of application and is normally used in HTTP and SMTP protocols. It is comprised by an XML-based service protocol stack and consists of three parts: wrapper, which defines the message structure; a set of rules that express application instances and data types; and a convention for the definition of call and answer procedures. A SOAP message is an XML that contains the following elements: wrapper, header, body, and error.

IV. Digitization

Within document management, the digitization processes give origin to the electronic documents from the physical documents. The use of electronic documents intends to make document management more efficient since it reduces

the use of physical documents and its derived problems, such as the consumption of office space, the internal transport of paper in large quantities and the risk of damage to the documents sheets. Digitization is defined as "the technological process by means of which an analogue (paper) or electronic support is converted into a digital image" (MINTIC, 2012a).

In the field of document management there are different types of digitization, according to their purpose. The digitization processes are classified into two large groups, according to the type of electronic document resulting: digitization without removing the original analogue document and digitization with replacement of the analogue support.

- Digitization without removing the original analogue document. In this category can be found, digitization with control and processing purposes, generally used in correspondence offices, where a large number of documents are received; digitization for archival purposes, where it is necessary to comply with archival standards and regulations issued by the AGN (Law 527 of 1999); and digitization for contingency purposes and business continuity, which is done in case any eventuality affects the original analogue support.
- Digitization with replacement of the analogue support, certified digitization. In this type of digitization, the electronic document replaces the analogue support, that is, the physical document. However, in order to achieve this objective, a standard previously established by competent bodies and by the AGN, which is endorsed by an authorized body, must be used. It should not be understood as a process where digital signature of the documents is necessarily carried out, but as a process that must comply with certain standards that can be certified by the same entity, in accordance with the standards issued by the competent bodies or by an authorized third party (MINTIC, 2012a).

As a result of digitization, it is included the concept of electronic document that is defined as "information generated, sent, received, stored or communicated by electronic, optical or similar means" (MINTIC, 2012b). It is also considered the electronic document file as the one produced by an entity or person by reason of their activities, which must be treated according to the archival principles and processes. Electronic documents can be classified by their form of creation, origin or format, and must comply with characteristics of authenticity, integrity, reliability and availability.

The authentic copy is a new electronic document issued by an accredited entity to do so, which has the same pro-

grupos, de acuerdo con el tipo de documento electrónico resultante: digitalización sin eliminación del documento original análogo y digitalización con sustitución del soporte análogo.

- Digitalización sin eliminación del documento original análogo. En esta categoría se puede encontrar la digitalización con fines de control y trámite, utilizada generalmente en las oficinas de correspondencia, donde se recibe una gran cantidad de documentos; digitalización con fines archivísticos, donde se debe cumplir con normas y estándares archivísticos expedidos por el AGN (Ley 527 de 1999); y digitalización con fines de contingencia y continuidad del negocio, que se realiza en caso que alguna eventualidad afecte los soportes análogos originales.
- Digitalización con sustitución del soporte análogo, digitalización certificada. En este tipo de digitalización el documento electrónico sustituye al soporte análogo, es decir al documento físico. Sin embargo, para alcanzar este objetivo, se deben utilizar un estándar previamente establecido por organismos competentes y por el AGN, que es avalado por una instancia u organismo autorizado. No debe entenderse como un proceso donde necesariamente se realice firma digital de los documentos, sino como un proceso que debe cumplir con ciertos estándares que pueden ser certificados por la misma entidad, de conformidad con las normas que expidan los organismos competentes o por un tercero autorizado (MINTIC, 2012a).

Como resultado de la digitalización se encuentra el concepto de documento electrónico que se define como "la información generada, enviada, recibida, almacenada o comunicada por medios electrónicos, ópticos o similares." (MINTIC, 2012b). También se considera el documento electrónico de archivo, como aquel que es producido por una entidad o persona en razón de sus actividades, el cual debe ser tratado conforme a los principios y procesos archivísticos. Los documentos electrónicos se pueden clasificar por su forma de creación, origen o formato, y deben cumplir con características de autenticidad, integridad, fiabilidad y disponibilidad.

La copia autentica es un nuevo documento electrónico expedido por una entidad acreditada para hacerlo, que tiene el mismo valor probatorio que el original. Su autenticidad se acredita a partir de su comprobación de igualdad con el original y produce los mismos efectos sobre organizaciones e interesados (MINTIC, 2012a). Existen varios tipos de copias auténticas: copia electrónica autenticada con cambio de formato, copia electrónica autenticada de documento de papel (digitalización certificada) y copia electrónica parcial auténtica.

Dentro de los métodos apropiados y legales para establecer la confiabilidad de un documento electrónico se encuentran la firma electrónica y la firma digital. En el

proceso de conformación del documento electrónico se le asignan metadatos y marcas adicionales, como la firma digital y la estampa cronológica. En el ámbito de la gestión documental en documentos electrónicos se encuentran las clases de metadatos de información, gestión, seguridad, trazabilidad, firma y estampado cronológico.

Para garantizar la eficacia del proceso de digitalización, las entidades del sector público deben definir un plan de gestión de calidad. En un proceso de control de calidad principalmente se debe revisar el resultado de las imágenes; si su cantidad es muy limitada, tiene sentido que se pueda revisar imagen por imagen. Por lo regular, los proyectos de gestión documental se deben realizar sobre una gran cantidad de imágenes, casos en los que es recomendable que se realice sobre un muestreo equivalente a un 10% de las imágenes de forma aleatoria por cada uno de los dispositivos de captura. Dentro de un control de calidad se debería establecer, en un programa institucional, definir el alcance del control de calidad, determinar si se realizará manualmente o de forma automática, definir cuándo se haría una nueva digitalización, y reglamentar un seguimiento continuo.

En la digitalización, el medio tecnológico comúnmente usado para convertir los documentos en soporte análogo en documentos electrónicos es el escáner, equipos que pueden tener accesorios de alimentación automática de papel o una bandeja plana donde se coloca una hoja de papel manualmente. Para la operación del escáner a través de un ordenador hay estándares a nivel de aplicación que aseguran la compatibilidad con el software, entre ellos se encuentran: ISIS, utilizado a nivel industrial o empresarial; TWAIN, originalmente destinado a computadores personales, actualmente usado para grandes volúmenes de documentos; y SANE [*Scanner Access Now Easy*], un API que provee acceso estándar a hardware de escaneo de imágenes orientado a entornos UNIX, incluyendo GNU/Linux.

Los escáneres pueden diferenciarse en el mercado de acuerdo con características técnicas, como: resolución máxima, medida en puntos por pulgada [DPI]; la velocidad de captura, expresada en páginas por minuto [PPM]; la interfaz, puertos ISIS, USB, puerto Ethernet; y el tamaño de hoja permitido, carta, A4, oficio, etc.

Los gestores de contenido empresarial (ECM y EDM) descritos, por lo general no cuentan con módulos para la digitalización y deben integrarse con herramientas externas destinadas para tal fin, las que, como se puede apreciar en la **Tabla 1** –donde se mencionan algunas de esas herramientas–, no están enfocadas específicamente en gestión documental para entidades oficiales.

Las funcionalidades que debería tener un módulo de digitalización integrado a una herramienta de apoyo de gestión documental deberían incluir opciones de optimización de imágenes, como cambios en los valores de brillo, contraste, color, limpieza de manchas, compresión y cambios de formato entre otras, las mismas que deberían

bative value as the original. Its authenticity is confirmed from the verification of equality regarding the original and produces the same effects on organizations and stakeholders (MINTIC, 2012a). There are several types of authentic copies: authenticated electronic copy with change of format, authenticated electronic copy of paper document (certified digitization) and authentic partial electronic copy.

Within the appropriate and legal methods to establish the reliability of an electronic document are the electronic signature and the digital signature. In the process of forming the electronic document, is assigned additional metadata and marks, such as the digital signature and chronological stamp. In the field of document management in electronic documents are the information metadata classes, management, security, traceability, signature and chronological stamp.

To ensure the efficiency of the digitization process, public sector entities must define a quality management plan. In a process of quality control, it is important to check the results of the images; if their quantity is very limited it makes sense that can be reviewed image by image. Generally, document management projects must be carried out on a large number of images, in which cases it is recommended to perform a sampling equivalent to 10% of the images at random by each of the capture devices. Within a quality control it should be established, in an institutional program, the definition of the scope of the quality control, determine whether it will be performed manually or automatically, define when a new digitization would take place, and regulate a continuous monitoring.

In digitization, the technological means commonly used to convert documents in analogue support into electronic documents is the scanner, equipment which may have automatic paper feed accessories or a flat tray where a sheet of paper is manually placed. For the operation of the scanner through a computer, there are standards at an application level that ensure compatibility with the software, these include: ISIS, used at an industrial or business level; TWAIN, originally intended for personal computers, currently used for large volumes of documents; and SANE [Scanner Access Now Easy], an API that provides standard access to image scanning hardware targeted to UNIX environments, including GNU/Linux.

The scanners can be differentiated in the market according to technical characteristics, such as: maximum resolution, measured in dots per inch [DPI]; capture rate, expressed in pages per minute [PPM]; interface, ISIS and USB ports, Ethernet port; and the allowed sheet size, letter, A4, legal, etc.

Table 1. Capture or digitizing tools / Herramientas de captura o digitalización

Tool	License	Compatibility	Description
Document and content capture with auto-categorization	Licensed by Alfresco Server (not free) / *Licenciado por Alfresco Server (no libre)*	Alfresco	It allows access through the Alfresco portal, it has functionalities for large-volume scanning and quick metadata assignment; it has a user interface for monitoring and displaying lists of documents. Integrates workflows to the incoming documents (Alfresco, 2017a) / *Permite el acceso por el portal de Alfresco, tiene funcionalidades para escanear en gran volumen y asignar metadatos de forma rápida; cuenta con una interface de usuario para monitorear y desplegar listas de documentos. Integra los flujos de trabajo a los documentos entrantes (Alfresco, 2017a)*
Document indexing module for Alfresco share	Property of the manufacturer/ *Propiedad del fabricante*	Alfresco	Module that can perform the capture from scanners, multifunction devices, printer, copier. It allows easy indexing queues for the user; it can be integrated to the Alfresco workflows (Alfresco, 2017b) / *Modulo que puede realizar la captura desde escáneres dispositivos multifuncionales impresora, escáner, copiadora. Permite colas de indexación fáciles para el usuario, se puede integrar a los flujos de trabajo de Alfresco (Alfresco, 2017b)*
Ephesoft	Corporate version (not free); Community version (Open source), only for Linux (not free)/ *Versión empresarial (no libre); versión Community (Open source), solo para Linux (no libre)*	Alfresco and Nuxeo	Capture of physical documents via scanner, from e-mail or fax. Optical character recognition [OCR] and handwriting text or barcode reading. Import processed content by sorting field keys included as metadata (Ephesoft, 2017)/ *Captura de documentos físicos a partir de escáner, provenientes de correo electrónico o fax. Reconocimiento de texto [OCR] y de texto de escritura a mano o lectura de código de barras. Importación de contenido procesado mediante clasificación de llaves de campos incluidos como metadatos (Ephesoft, 2017)*

The described business content managers (ECM and EDM), generally do not have modules for digitization and must be integrated with external tools intended for this purpose, which, as can be seen in **Table 1** –where some of these tools are mentioned– are not specifically focused on document management for official entities.

The functionalities that should have a digitizing module integrated with a document management support tool should include options for optimizing images, such as values changes in brightness, contrast, color, stain cleaning, compression and format changes, among others; the same ones that should be used to obtain a readable document, not to modify the original content. ImageMagick, LibTIFF and Netpbm (described below) are three free software tools that can be used for image processing.

- ImageMagick is a suite of tools to create, edit and compose bitmap images, which supports a large number of formats, including: PNG, JPEG, GIF, TIFF and PDF. In the following console run example, a cut of a specific size is made to a series of images

$convert '*.jpg' -crop 120x120+10+5 thumbnail%03d.png

- LibTIFF is a library of tools to make simple manipulations of TIFF (Tag Image File Format) images, available for Linux, BSD, Solaris and MacOS X platforms. The following example converts G3 to G4 encoding (Group4) over TIF images.

$ tiffcp−c g4 −r 10000 g3.tif g4.tif

ser utilizadas para obtener un documento legible, no para modificar el contenido original. ImageMagick, LibTIFF y Netpbm (descritas a continuación), son tres herramientas software libres que pueden servir para el tratamiento de imágenes.

- ImageMagick es una suite de herramientas para crear, editar y componer imágenes de mapas de bits, que soporta una gran cantidad de formatos, ente ellos: PNG, JPEG, GIF, TIFF y PDF. En el siguiente ejemplo de ejecución en consola, se realiza un corte de un tamaño específico a una serie de imágenes.

$convert '*.jpg' -crop 120x120+10+5 thumbnail%03d.png

- LibTIFF es una librería de herramientas para hacer simples manipulaciones de imágenes TIFF (*Tag Image File Format*), disponible para plataformas Linux, BSD, Solaris y MacOS X). En el siguiente ejemplo se realiza la conversión de codificación G3 a G4 (Group4) sobre imágenes TIF.

$ tiffcp−c g4 −r 10000 g3.tif g4.tif

- Netpbm es un conjunto de herramientas para manipular imágenes que incluye la conversión y compresión de distintos formatos; soporta cerca de cien formatos en unas trescientas herramientas separadas. En el siguiente ejemplo En este ejemplo se extrae la página 3 de un documento PDF y se guarda como imagen en formato PNG.

$ pdftoppm -f 3 -l 3 -png origen.pdf > destino_pag3.png

V. Método

La investigación se enmarcó en un enfoque cualitativo, lo que indica que, bajo la perspectiva de un observador, se realiza una evaluación de experiencias después de recolec-

tar información de las observaciones siguiendo un conjunto de técnicas o métodos.

El proceso del enfoque cualitativo está representado por cuatro grandes fases: preparatoria, trabajo de campo, analítica e informativa (Monje, 2011). El tipo "investigación aplicada", elegido para este trabajo de investigación, es el que más se adapta a la problemática planteada, ya que es un proyecto que se "caracteriza porque busca la aplicación o utilización de los conocimientos adquiridos, a la vez que se adquieren otros, después de implementar y sistematizar la práctica basada en investigación" (Vargas, 2008).

La base poblacional de observación son las normas relacionadas con la gestión documental electrónica que rigen a las entidades públicas en Colombia y los requerimientos funcionales para el desarrollo de software derivados de ellas. Al ser tan extenso el panorama normativo, se realizó el proceso de investigación con una muestra representativa de la normativa y de los requerimientos de mayor prioridad, de acuerdo con las necesidades expuestas por el MINTIC y el AGN, que como se mencionó, son los organismos encargados de expedir los lineamientos generales que regulan la gestión documental electrónica.

Dado que se trata de un enfoque cualitativo, las unidades de análisis no son parte de un análisis numérico exacto, sino de la base de observaciones en un proceso de desarrollo de software fundamentado en la experiencia y los conocimientos. En la **Tabla 2**, para cada unidad de análisis, se presentan sus indicadores y valores probables; en todos los casos, la técnica de recolección es la revisión documental, y la técnica de análisis, el análisis documental.

Como se puede observar en las referencias, es necesario, por un lado disponer de un sistema para llevar a cabo

- Netpbm is a set of tools to manipulate images that includes the conversion and compression of different formats; supports about one hundred formats in about three hundred separate tools. This example extracts page 3 of a PDF document and saves it as an image in PNG format.

```
$ pdftoppm -f 3 -l 3 -png origen.pdf > destino_pag3.png
```

V. Method

The research was framed in a qualitative approach, indicating that, from the perspective of an observer, an evaluation of experiences is performed after collecting information from the observations following a set of techniques or methods.

The process of the qualitative approach is represented by four major phases: preparatory, fieldwork, analytical and informative (Monje, 2011). The "applied research" type chosen for this research work, is the one that best adapts to the problem raised, since it is a project that is "characterized because it aims the application or use of the acquired knowledge, while acquire others, after implementing and systematizing research-based practice "(Vargas, 2008).

The population base of observation are the rules related to electronic document management that govern public entities in Colombia and the functional requirements for the development of software derived from them. Since the normative panorama is so extensive, the research process was carried out with a representative sample of the regulations and the

Table 2. Analysis units / Unidad de análisis

Unit of analysis	Indicators	Values
Current regulation with regard to digitization and electronic document management / *Normativa vigente con respecto a la digitalización y gestión documental electrónica*	Relevance to implement / *Relevancia para implementar*	High / Medium / Low / *Alta / Media / Baja*
	Complexity to implement / *Complejidad para implementar*	High / Medium / Low / *Alta / Media / Baja*
Application requirements derived from the regulations found / *Requerimientos de aplicación derivados de la normativa encontrada*	Level of compliance / *Nivel de cumplimiento*	Accomplished / Partially accomplished / Not accomplished / *Cumplido / Parcialmente cumplido / No cumplido*
Original developed software functionalities / *Funcionalidades software originales desarrolladas*	Adaptation to requirements / *Adecuación a los requerimientos*	Appropriate / Moderately appropriate / Partially appropriate / Not appropriate / *Adecuada / Medianamente adecuada / Parcialmente adecuada / No adecuada*
Reused tools with licensing related to free software / *Herramientas reutilizadas con licenciamiento relacionado al software libre*	Adaptation to requirements / *Adecuación a los requerimientos*	Appropriate / Moderately appropriate / Not appropriate / *Adecuada / Medianamente adecuada / No adecuada*
Electronic documents resulting from the digitization process / *Documentos electrónicos resultantes del proceso de digitalización*	Quality / *Calidad*	High / Medium / Low / *Alta / Media / Baja*

highest priority requirements, in accordance with the needs presented by MINTIC and AGN, which, as mentioned, are the organisms responsible for issuing the general guidelines that regulate electronic document management.

Since this is a qualitative approach, the units of analysis are not part of an exact numerical analysis but the basis of observations in a software development process based on experience and knowledge. In **Table 2** are presented indicators and probable values for each unit of analysis; in all cases, the collection technique is the documentary review, and the analysis technique is the documentary analysis.

As can be seen in the references, it is necessary, on one side to have a system to carry out the digitization of physical documents, and on the other, to have a content manager [ECM] that provides services for the management of electronic documents. According to previous analyses, the ECM Alfresco Community is the most recommended document management support tool (Jiménez, 2016).

It was concluded that the digitization module should be developed because the market does not find a free software tool focused on the official entities in Colombia, and was named as FuidXel the new tool that would carry out the digitization of physical documents.

FuidXel is a server client system that must have a component for running the scanner and displaying the resulting images; therefore, it is specified that the component is a client-side application, external to the main service application. The client-side application is designed and developed for a Windows environment, because this is the most used operating system in the public sector.

Due to FuidXel is a free software project, it is necessary to find a methodology according to a process that can be increased on a recurring basis and in accordance with the agile method of development (Beck et al., 2001; Paz, Castañeda, & Arboleda, 2011; Navarro, Fernández, & Morales, 2013; Britto, 2016). It is decided that the most appropriate is the iterative method, since it has incremental characteristics, which can support a growing understanding of the requirements and the development is divided in a planned way into smaller parts, called iterations (**Figure 1**).

It is observed the concordance between the steps of the applied research and the methodology of development. The initial planning phase described the problem situation to be intervened or improved, justified by the normative analysis; the theory was selected to expose it, with its central concepts in relation to available and applied technologies; and the instruments were used to describe and classify the relevant

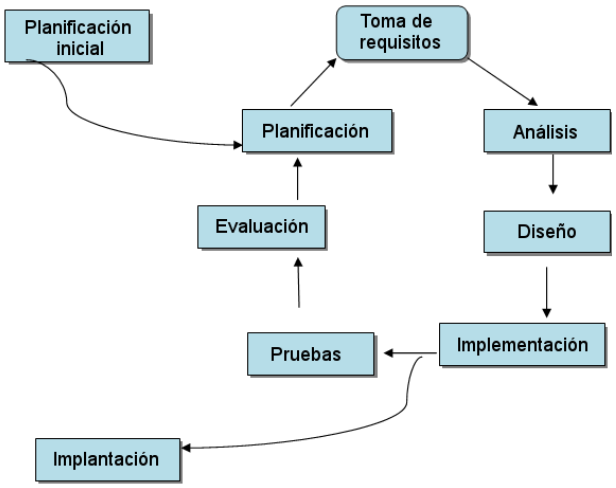

Figure 1. Iterative model of software life cycle (Mas, Megías, Ginestà, & Peña, 2005) / Modelo Iterativo de ciclo de vida de software (Mas, Megías, Ginestà, & Peña, 2005)

la digitalización de los documentos físicos, y por el otro, disponer de un gestor de contenidos [ECM] que brinde servicios para la gestión de documentos electrónicos. De acuerdo con los análisis previos, el ECM *Alfresco Community* es la herramienta de apoyo a la gestión documental más recomendada (Jimenez, 2016).

Se concluyó que el modulo de digitalización debería ser desarrollado porque en el mercado no se encuentra una herramienta de software libre enfocada a las entidades oficiales en Colombia, y se nombró como FuidXel a la nueva herramienta que llevaría a cabo la digitalización de documentos físicos.

FuidXel es un sistema cliente servidor que debe contar con un componente para la ejecución del escáner y la visualización de las imágenes resultantes; Por lo tanto, se especifica que ese componente es un aplicativo al lado del cliente, externo a la aplicación de servicio principal. La aplicación al lado del cliente se diseña y desarrolla para un ambiente Windows, por ser este el sistema operativo más utilizado en el sector público.

Debido a que FuidXel es un proyecto de software libre se busca una metodología acorde con un proceso que pueda incrementarse de manera recurrente y que esté de acuerdo con el método ágil de desarrollo (Beck et al., 2001; Paz, Castañeda, & Arboleda, 2011; Navarro, Fernández, & Morales, 2013; Britto, 2016). Se decide que el más acorde es el método iterativo, ya que tiene características incrementales, que pueden soportar una comprensión creciente de los requerimientos y el desarrollo se divide de manera planeada en partes más pequeñas, llamadas iteraciones (**Figura 1**).

Se observa la concordancia entre los pasos de la investigación aplicada y la metodología de desarrollo. En la fase de planificación inicial se describió la situación problema que se va intervenir o mejorar, justificada con el análisis

normativo; se seleccionó la teoría para exponerla, con sus conceptos centrales en relación con las tecnologías disponibles y aplicadas; y se diligenciaron los instrumentos para describir y clasificar la normativa relevante, de acuerdo con su viabilidad respecto de su complejidad. Luego, se procedió a examinar la situación problema comprendida en la normativa sobre la teoría seleccionada de los mecanismos tecnológicos, se definieron los requerimientos derivados de la normativa; se diligenció el instrumento para listar los requerimientos; y se derivó el prototipo de acción con la metodología de desarrollo de software que se soluciona en las iteraciones. Con base en los requerimientos resultantes, se definió además la documentación inicial de los diseños del desarrollo software.

Cada iteración cuenta con fases de análisis, diseño, desarrollo, pruebas y evaluación, que corresponden a los dos últimos pasos de la investigación aplicada. En las fases de análisis, diseño y desarrollo se examina la situación problema para derivar el prototipo de acción, y en la fase de pruebas se realiza el procedimiento para ensayar y probar el prototipo de acción.

VI. Resultados

De acuerdo con la metodología seleccionada el desarrollo del proyecto se inició con una planificación y luego se realizaron las iteraciones que resultaron de ella.

Planificación

Para realizar la planificación inicial fue necesario partir del marco teórico, a partir del análisis de la normativa encontrada se derivó el documento de especificación de requerimientos, y posteriormente se definió el cronograma de actividades con el formato del método de trabajo *Scrum* (ProductBacklog). Posteriormente se definió un conjunto de diseños de desarrollo de software básicos que representan la implementación de las soluciones.

Las guías institucionales del MINTIC y el AGN son referencia inequívoca a las normas establecidas, lo que quiere decir que sirven de pauta base para la definición de los requerimientos para cumplir los objetivos. Siguiendo las guías del MINTIC (2012 a; 2012b), la Circular Externa 005/2012 y el Acuerdo 005/2013 del AGN, y las normas internacionales sugeridas o determinadas por estos organismos, se tiene la documentación suficiente para definir los requerimientos. Como se indicó en la sección anterior, el presente proyecto no pretendió abordar exhaustivamente los requerimientos normativos en relación con la gestión documental sino brindar herramientas de apoyo a los procesos de digitalización y consulta posterior de documentos electrónicos observando la normativa implementada para las entidades del sector público.

Iteración 1

El modelo de infraestructura se puede observar en la **Figura 2**, donde se modela la distribución de máquinas servidoras. Con respecto a este modelo, se generó un ambiente virtualizado para emular el funcionamiento de los

regulations, according to their feasibility with respect to their complexity. Then, it was proceed to examine the problem situation included in the regulation on the selected theory of technological mechanisms, the requirements derived from the regulations were defined; the instrument was submitted to list the requirements; and the prototype of action was derived with the methodology of software development that is solved in the iterations. Based on the resulting requirements, the initial documentation of software development designs was also defined.

Each iteration has phases of analysis, design, development, testing and evaluation, which correspond to the last two steps of applied research. In the analysis, design and development phases, the problem situation is examined to derive the prototype of action, and in the testing phase, the procedure to test and check the prototype of action is performed.

VI. Results

According to the selected methodology, the development of the project began with a planning and then the resulting iterations from it were made.

Planning

In order to carry out the initial planning it was necessary to start from the theoretical framework, based on the analysis of the normative found, it was derived the document of specification of requirements, and later the schedule of activities was defined with the format of the Scrum (ProductBacklog) working method. Subsequently, a set of basic software development designs that represent the implementation of the solutions were defined.

The institutional guides of MINTIC and AGN are an unequivocal reference to the established norms, which means that they serve as the basic guideline for the definition of the requirements to fulfill the objectives. Following the guidelines of MINTIC (2012 a; 2012b), External Circular 005/2012 and AGN Agreement 005/2013, and the international norms suggested or determined by these agencies, there is sufficient documentation to define the requirements. As indicated in the previous section, this project did not intend to address, in an exhaustively way, the regulatory requirements in relation to document management, but to provide tools to support digitization processes and further consultation of electronic documents, observing the regulations implemented for public sector entities.

Iteration 1

The infrastructure model can be observed in **Figure 2**, where the distribution of server machines is modeled. With

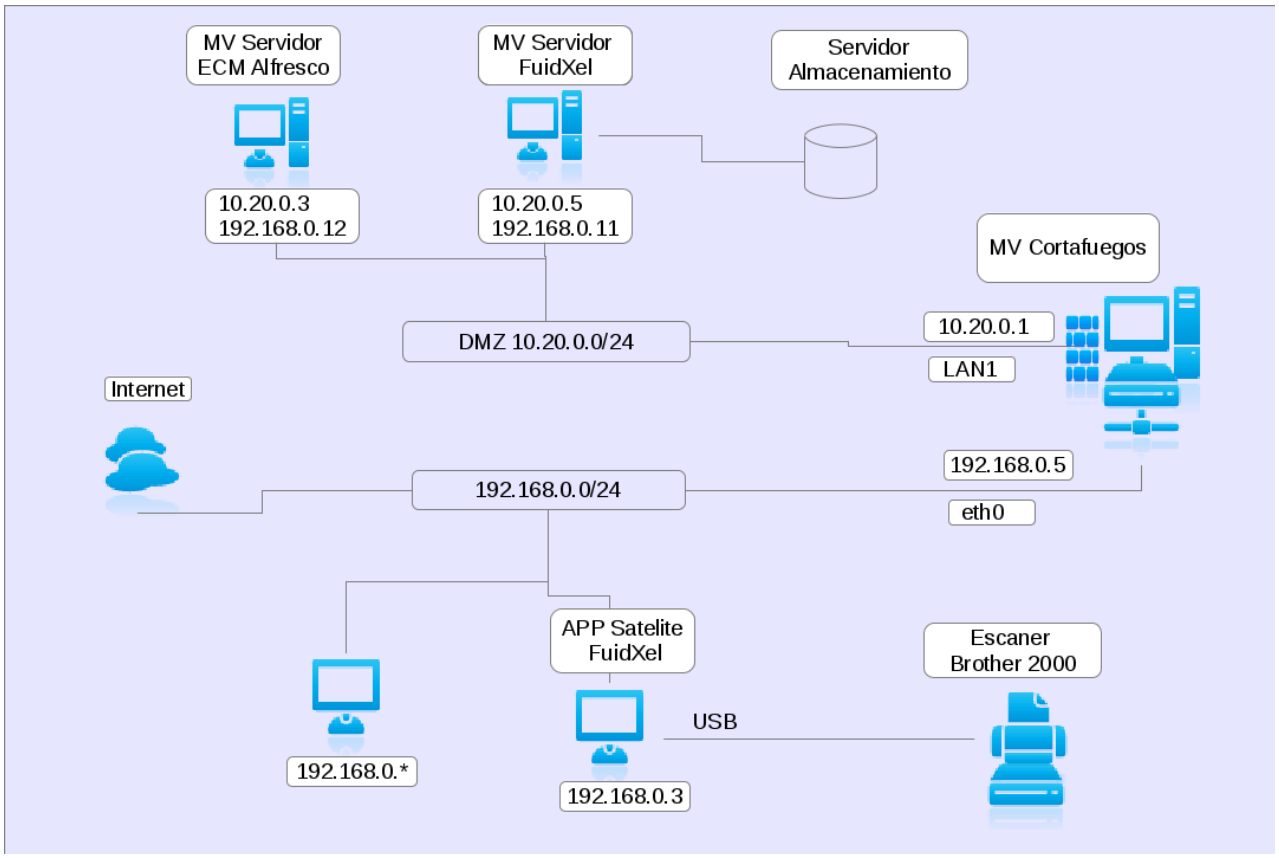

Figure 2. Adaptation architecture in development environment / Adaptación arquitectura en ambiente de desarrollo

respect to this model, a virtualized environment was generated to emulate the operation of the main components and thus to find a way for the development of the application of digitization. A virtual machine was adapted for the operation of Alfresco Community, another virtual machine for the operation of the system to support the digitization, a third one to emulate a firewall and, finally, the host machine that will serve as a client to the application services.

Iteration 2

For this purpose, FuidXel input and management modules were developed. For the input module, the authentication form was developed using the LDAP protocol, and for the graphical interface model it was used a framework or set of HTML, CSS and Javascript tools, called Bootstrap, which has aspects of responsive web design that is used to make the interface change its size and behavior according to the size of the window. The main menu is dynamic regarding the database.

Iteration 3

The user interface for scanning and displaying images was implemented. In **Figure 3**, it can be observed the appearance of the developed tool. It corresponds to the client-side application, since the scan is performed on computers that

componentes principales y así conseguir un medio para el desarrollo de la aplicación de digitalización. Se adaptó una máquina virtual para el funcionamiento de Alfresco Community, otra máquina virtual para el funcionamiento del sistema de apoyo a la digitalización, una tercera para emular un cortafuegos y, por último, la maquina anfitriona que servirá como cliente a los servicios de aplicación.

Iteración 2

Para ella se desarrollaron los módulos de entrada y administración de FuidXel. Para el módulo de entrada se desarrolló el formulario de autenticación, mediante el protocolo LDAP, y para el modelo de interfaz gráfica, se utilizó un Framework o conjunto de herramientas HTML, CSS y Javascript, llamado Bootstrap, que cuenta con aspectos de diseño web adaptable, que sirven para que la interfaz cambie de tamaño y comportamiento de acuerdo con el tamaño de la ventana. El menú principal es dinámico respecto de la base de datos.

Iteración 3

Se implementó la interfaz de usuario de escaneo y visualización de imágenes. En la **Figura 3** se puede observar el aspecto de la herramienta desarrollada. Ella corresponde a la aplicación del lado del cliente, ya que el escaneo se realiza en equipos que pueden ser externos al servidor. Primero se abordó la parte de visualización, donde fue necesario definir parámetros de guardado local de las imágenes en

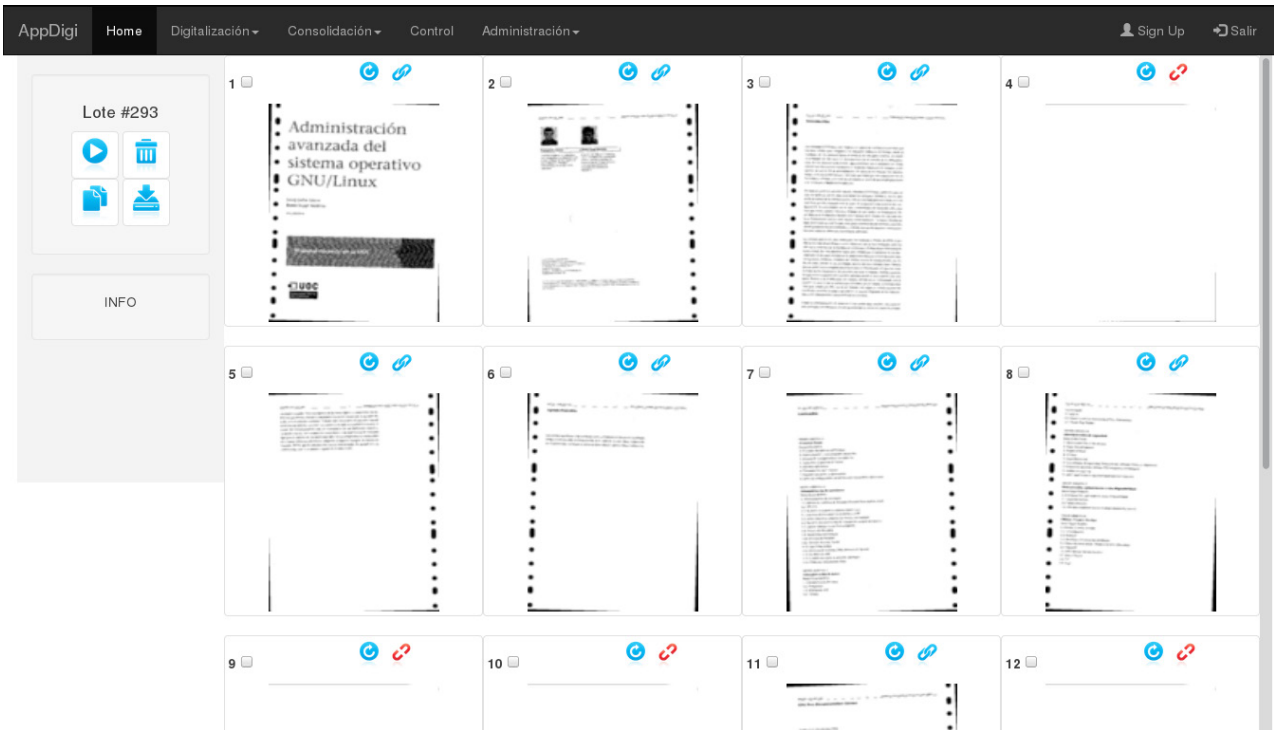

Figure 3. Scan and preview user interface / Interfaz de usuario de escaneo y pre-visualización

lote. La pre-visualización de imágenes requiere un formato compatible con cualquier navegador, se encontró que PNG era el mejor soportado. El proceso interno de conversión genera las miniaturas de imágenes para pre-visualización en formato PNG y las imágenes definitivas en formato TIFF.

Se implementó el proceso de escaneo, para realizarlo, fue necesario desarrollar un programa en Python utilizando el módulo de Twain con licencia GPL llamado Twainmodule. Python puede recibir parámetros como un comando en consola, de esta manera, por vía de ejecución del sistema en PHP, se envía la información parametrizada de la ejecución del escaneo al programa en Python lo que genera la captura de imágenes a través del escáner. A continuación se puede observar un ejemplo de la ejecución de dicho programa como un comando de consola:

$ejecucionEscaner.pyw -d TW-Brother -c none -a False -u True -b 1 -p bitonal -f bmp -r 300.0 -F C:\Digitalizacion\00000

Iteración 4

La indexación es el proceso posterior a la captura y transferencia de imágenes. Se deben asignar las imágenes a los metadatos descriptivos o de gestión del documento al que pertenezcan. Para brindar unicidad al documento se le debe asignar un índice. En el ámbito de las instituciones del sector público regularmente el índice corresponde a un número de radicado asociado a un trámite.

Para la interfaz de indexación se desarrolló un paquete de clases, que sirven para la generación automática de formularios de diligenciamiento a partir de la parametrización en base de datos. Esto fue necesario porque, tanto

can be external to the server. First, the visualization part was addressed, where it was necessary to define parameters of local saving of the images in batch. The preview of images requires a format compatible with any browser, it was found that PNG was the best supported. The internal conversion process generates thumbnails of images for preview in PNG format and final images in TIFF format.

It was implemented the scanning process, to achieve it, was necessary to develop a program in Python using the module of Twain licensed GPL called Twainmodule. Python can receive parameters like a command in console, so, by means of execution of the system in PHP, the parameterized information of the execution of the scanning is sent to the program in Python that generates the capture of images through the scanner. The following is an example of running the program as a console command:

$ejecucionEscaner.pyw -d TW-Brother -c none -a False -u True -b 1 -p bitonal -f bmp -r 300.0 -F C:\Digitalizacion\00000

Iteration 4

Indexing is the process after capturing and transferring images. Images must be assigned to the descriptive or document management metadata to which they belong. To provide uniqueness to the document, an index must be assigned. In the field of the institutions of the public sector, regularly the index corresponds to a number of residing associated to a procedure.

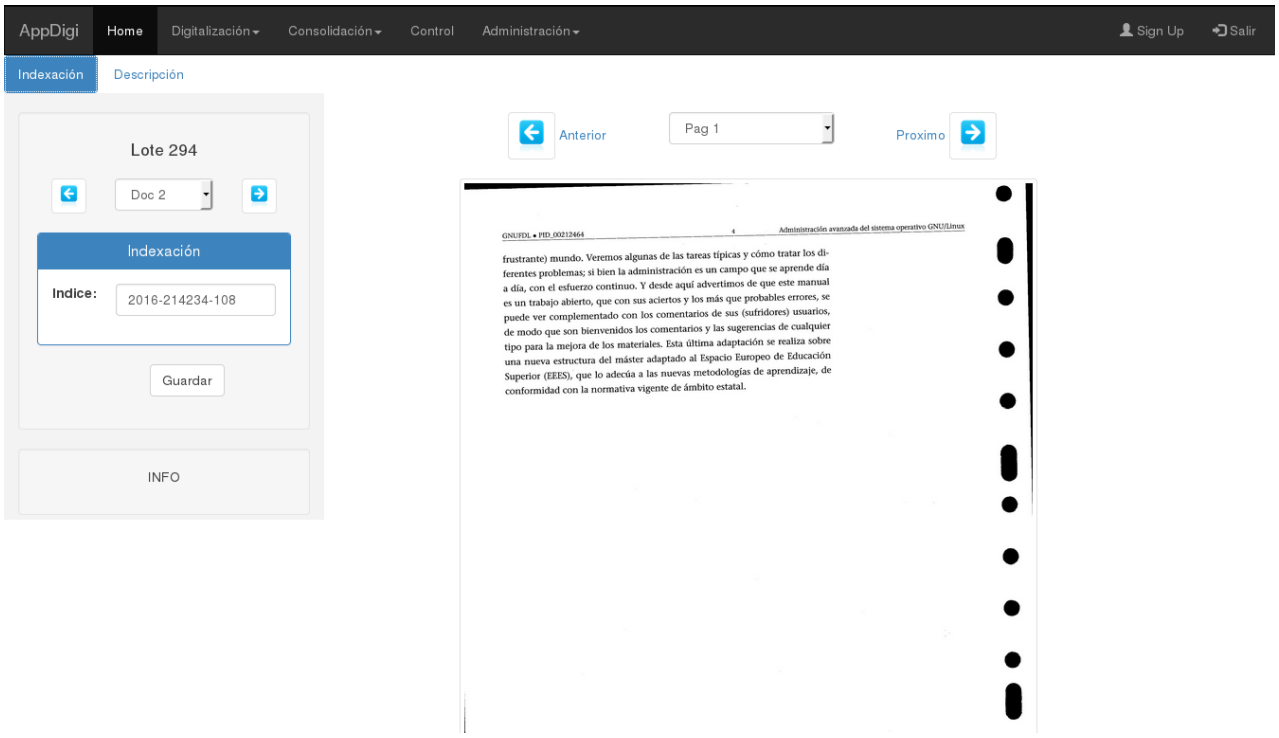

Figure 4. Indexing of documents main index form / Indexación de documentos formulario índice principal

For the indexing interface, it was developed a package of classes, which are used for the automatic generation of fulfillment forms from the parameterization in database. This was necessary because both the main index of the document and the document description form may vary according to the public institution; with this implementation it was possible to configure the number of different forms by means of the database and to integrate them by calling the functionality and defining parameters as the form code.

The indexing interface (**Figure 4**) was divided into two tabs: in the first, it can be observed the images of the document, along with the form to assign the index; in the second one, is the documentary description form that serves to fill out the fields related to the ISAD (G) standard.

Iteration 5

In this iteration, the ECM Alfresco Community Edition was observed and analyzed. The ECM for the Electronic Document Management System [EDMS] provides support as the main repository for documents resulting from digitization. Alfresco has a wide range of functionalities, among which are: workflow configuration, configuration of catalogs of labels, hierarchical organization of directories, configuration of taxonomic models for documentary description, search of documents by information and generation of sites by user groups.

el índice principal del documento, como el formulario de descripción documental, pueden variar de acuerdo con la institución pública, con esta implementación fue posible configurar el número de formularios distintos por medio de base de datos e integrarlos llamando la funcionalidad, definiendo apenas parámetros como el código del formulario.

Se dividió la interfaz de indexación en dos pestañas (**Figura 4**): en la primera se pueden observar las imágenes del documento, junto con el formulario para asignar el índice; en la segunda se encuentra el formulario de descripción documental que sirve para el diligenciamiento de los campos relacionados con la norma ISAD(G).

Iteración 5

En esta iteración se realizó la observación y el análisis del ECM Alfresco Community Edition. El ECM para el Sistema de Gestión de Documentos Electrónicos [SGDE] brinda apoyo como repositorio principal de los documentos resultantes de la digitalización. Afresco cuenta con una amplia gama de funcionalidades, dentro de las cuales se encuentran: configuración de flujos de trabajo, configuración de catálogos de etiquetas, organización jerárquica de directorios, configuración de modelos de taxonomías para descripción documental, búsqueda de documentos por información y generación de sitios por grupos de usuarios.

El mapeo de los metadatos relacionados con el documento electrónico se configuró en la herramienta de administración de modelos de aspectos y tipos personalizados. Se creó un nuevo modelo y en él se crearon los aspectos de acuerdo con las secciones de la norma de descripción do-

cumental. Para cada aspecto se configuraron las propiedades correspondientes a los campos de diligenciamiento en cada sección, de esta manera se implementó la configuración de descripción documental para el ECM Alfresco Community. Estos metadatos corresponderán a los desarrollados en el módulo de indexación de FUIDXEL, que serán transferidos para la integración de los dos sistemas.

Iteración 6

En los últimos requerimientos donde interviene los dos sistemas FUIDXEL y Alfresco Community fue necesario establecer el mecanismo de integración. Dentro de los disponibles por Alfresco se encuentra la combinación de REST y CMIS. La integración con Alfresco se realizó con la ayuda de la librería "CMIS PHP Client" que es un proyecto de *Apache Software Foundation* [ASF]. Esta librería contiene funcionalidades para iniciar una conexión a la plataforma de servicios que requiere una autenticación con usuario y clave, lo cual corresponde a una cuenta con permisos en Alfresco. Para complementar el documento se agregó un componente de reconocimiento de texto OCR, por medio de un script desarrollado en SHELL. El programa utiliza varios comandos que debieron ser instalados en el servidor. Las herramientas utilizadas son TESSERACT y ExactImage para el reconocimiento del OCR y la conformación del documento PDF; también fueron utilizados comandos de herramientas de tratamiento de imágenes y documentos PDF como PDFtk, Netpbm, ImageMagick y Poppler-utils. La conversión del documento PDF a PDF "buscable", es decir con el texto reconocido, es un proceso transparente para el usuario ya que se ejecuta en segundo plano, de manera asíncrona.

Se dispuso en el portal web GITHUB para descarga la solución tecnológica para digitalización bajo licencia de software libre GPLv3 (ver López, 2017).

VII. Discusión y conclusiones

En la documentación encontrada no se caracterizó una herramienta con licenciamiento compatible con alguna versión GPL de la Free Software Foundation, la que define al software libre, que sirva de apoyo a la digitalización en la gestión de documentos electrónicos destinada a entidades oficiales en Colombia.

En el diseño de una propuesta tecnológica para la digitalización en la gestión de documentos electrónicos en instituciones de administración pública en Colombia se debe contemplar la integración con otras herramientas para la transferencia de documentos y el protocolo CMIS, ya que es el estándar de integración más usado en los gestores de contenido empresarial.

En las pruebas funcionales se pudo comprobar el proceso de generación de documentos electrónicos en formato PDF/A con texto parcialmente reconocido a partir de la digitalización de documentos físicos almacenados en lotes de imágenes.

The mapping of metadata related to the electronic document was configured in the custom types and aspects model management tool. A new model was created and the aspects according to the sections of the norm of documentary description were created on it. For each aspect, the properties corresponding to the fulfillment fields in each section were configured, thus the documentation description configuration for the Alfresco Community ECM was implemented. These metadata will correspond to those developed in the indexing module of FUIDXEL, which will be transferred for the integration of the two systems.

Iteration 6

In the last requirements where the two FUIDXEL and Alfresco Community systems intervened, it was necessary to establish the integration mechanism. Among those available by Alfresco, there is the combination of REST and CMIS. The integration with Alfresco was carried out with the help of the library "CMIS PHP Client" which is an Apache Software Foundation [ASF] project. This library contains functionalities to initiate a connection to the service platform that requires authentication with user and password, which corresponds to an account with permissions in Alfresco. To complement the document an OCR text recognition component was added, through a script developed in SHELL. The program uses several commands that must be installed on the server. The tools used are TESSERACT and ExactImage for the recognition of OCR and the conformation of the PDF document; commands from image processing tools and PDF documents such as PDFtk, Netpbm, ImageMagick and Poppler-utils were also used. The conversion of the PDF document to "searchable" PDF, that is to say with the recognized text, is a transparent process for the user since it runs in the background, asynchronously.

On the web portal, it was available GITHUB for downloading the technology solution for licensed digitization of GPLv3 free software (see López, 2017).

VII. Discussion and conclusions

In the documentation found, a tool was not characterized with licensing compatible with some GPL version of the Free Software Foundation, which defines free software that supports digitization in the management of electronic documents for official entities in Colombia.

In the design of a technological proposal for digitization in the management of electronic documents in public administration institutions in Colombia, integration must be considered with other tools for document transfer and the CMIS

protocol, since it is the most widely used integration standard by enterprise content managers.

In the functional tests, it was possible to verify the process of generating electronic documents in PDF/A format with partially recognized text from the digitization of physical documents stored in batches of images.

In the iterations within the software development cycle that was followed for the application of the research method, checks were made on the requirements defined in the initial planning phase, which resulted in its compliance with the sector of guidelines issued by the MINTIC and AGN, which were also selected in the initial planning phase.

Now, a synthesis of what is found during the development of the project is made in relation to the specified requirements.

The requirement related to the document scanning process was addressed in the third iteration of the development cycle. In the testing phase of this iteration, it was found that color and monochrome images can be scanned. The scan settings depend on the parameters defined in the management tool. The scan is performed with a module for Python that supports the TWAIN standard, because in PHP language was not found a library or module that was supported. For this case, it was observed that the scanner used has a memory limit that influences the support to the number of sheets; after this limit, the scanner loses efficiency or stops the scanning process. The development of related functionality meets the definition of the requirement, but its efficiency can be improved with a better scanner.

The optimization of images was addressed in the third iteration, where tools were developed for the conversion of images from the scan. For this task, the ImageMagick library for the PHP language was used. Although this conversion can be done with system commands, it was observed that it is better and more efficient to develop it through the API. Image conversion is done from the BMP format to TIFF and PNG. Other scanners support different formats that can also be implemented in FuidXel by changing the predefined constants.

In the conversion were implemented compression parameters to reduce the size of the images coming from the scanner, since by default they exceed 1 MB, if the images are left in this size and a large number of them are handled, when transferring the files may turn out to be unmanageable because it would consume large amounts of network resources and storage.

En las iteraciones dentro del ciclo de desarrollo de software que se siguió para la aplicación del método de investigación, se realizaron comprobaciones de los requerimientos definidos en la fase de planeación inicial, que dieron como resultado su cumplimiento con respecto del sector de lineamientos emitidos por el MINTIC y el AGN, los cuales fueron también seleccionados en la fase de planeación inicial.

A continuación, se realiza una síntesis de lo encontrado durante el desarrollo del proyecto en relación a los requerimientos especificados.

El requerimiento relacionado con el proceso de escaneo de documentos fue abordado en la iteración tres del ciclo de desarrollo. En la fase de pruebas de esta iteración, se comprobó que pueden ser escaneadas imágenes a color y monocromáticas. La configuración del escaneo depende de los parámetros definidos en la herramienta de administración. El escaneo se realiza con un módulo para Python que soporta el estándar TWAIN, debido a que en lenguaje PHP no se encontró una librería o modulo que fuera soportado. Para este caso se pudo observar que el escáner utilizado tiene un límite de memoria que influye en el soporte al número de hojas; después de este límite, el escáner pierde eficiencia o detiene el proceso de escaneo. El desarrollo de la funcionalidad relacionada cumple con la definición del requerimiento, pero su eficiencia se puede mejorar con un mejor escáner.

La optimización de imágenes fue abordada en la iteración tres, donde se desarrollaron herramientas para la conversión de imágenes provenientes del escaneo. Para esta tarea se utilizó la librería ImageMagick para el lenguaje PHP. Aunque esta conversión se puede realizar con comandos del sistema, se pudo observar que es mejor y más eficiente desarrollarlo por medio del API. La conversión de imágenes se realiza desde el formato BMP a TIFF y PNG. Otros escáneres soportan formatos diferentes que también pueden ser implementados en FuidXel cambiando las constantes predefinidas.

En la conversión se implementaron parámetros de compresión para disminuir el tamaño de las imágenes provenientes del escáner, ya que por defecto sobrepasan a 1 MB, si las imágenes se dejan en este tamaño y se maneja un gran número de ellas, a la hora de transferir los archivos el resultado puede llegar a ser inmanejable, ya que consumiría grandes cantidades de recursos de red y de almacenamiento.

En relación con el control de calidad, se desarrollaron las actas de inicio y finalización de almacenamiento de imágenes en las iteraciones tres y seis. En la fase de pruebas de la iteración seis, se comprobó el funcionamiento de la generación del acta de finalización, la cual se pudo comparar con el acta de inicio y así contrastar cuántos documentos esperaban ser digitalizados y cuántos finalmente fueron digitalizados. En el acta de finalización se puede constatar además información como el formato, el medio

de digitalización, el número de imágenes y el tamaño de los archivos resultantes.

La configuración de cuadros de clasificación documental fue posible de realizar mediante el uso y la organización de directorios en Alfresco, asignándoles metadatos descriptivos. La funcionalidad se comprobó con respecto al formato que se puede encontrar en la página del AGN. Con esta configuración se puede observar que es posible clasificar los documentos electrónicos con respecto al cuadro de clasificación documental, y utilizar flujos de trabajo para destinar los documentos de acuerdo con su clasificación en un proceso administrativo.

Se configuró un flujo de trabajo básico en Alfresco, para hacer un proceso de revisión en la carpeta de entrada de los documentos electrónicos provenientes de la digitalización. Con esta configuración, se aborda el requerimiento SGDE_ECM_03, que corresponde al del control de calidad en el gestor de contenido empresarial. La comprobación de esta funcionalidad se realizó en la iteración 6, donde se pudo observar que los usuarios pueden ingresar al directorio a revisar los documentos, uno por uno, y clasificarlos dependiendo de si están o no correctos. El proceso carece un poco de eficiencia, ya que después de la clasificación no se encontró la forma para que aparezca el siguiente documento automáticamente, sino que este se debe buscar manualmente.

Como elemento de seguridad se desarrolló una interfaz para el envío de las imágenes provenientes del proceso de escaneo desde la aplicación satélite al servidor FuidXel, donde es necesaria la autenticación con usuario y contraseña para ejecutar programas de integración que se encuentran en el servidor. Por otro lado, en la integración entre FuidXel y Alfresco Community Edition, se requiere una autenticación para el envío de documentos electrónicos y metadatos descriptivos que viene por defecto en el protocolo CMIS, utilizado para el envío de los mensajes. S&T

In relation to the quality control, the initiation and completion reports of image storage were developed in the iterations three and six. In the test phase of iteration six, it was verified the operation of the generation of the completion report, which could be compared with the initiation report and thus to compare how many documents were expected to be digitized and how many were finally digitized. In the completion report it is also find information such as the format, digitization means, number of images and the size of the resulting files.

The configuration of document classification tables was possible through the use and organization of directories in Alfresco, assigning them descriptive metadata. The functionality was checked with respect to the format that can be found in the page of the AGN. With this configuration it can be observed that is possible to classify the electronic documents with respect to the document classification table, and to use workflows to assign the documents according to their classification in an administrative process.

A basic workflow was configured in Alfresco to make a revision process in the input folder of the electronic documents from the digitization. With this configuration, the SGDE_ECM_03 requirement is addressed, which corresponds to the quality control in the enterprise content manager. The verification of this functionality was carried out in iteration 6, where it was observed that users can enter to the directory to review the documents, one by one, and classify them depending on whether or not they are correct. The process lacks a bit of efficiency, since after the classification there was no way for the following document to automatically appears, instead this one must be searched manually.

As an element of security, an interface was developed for sending the images coming from the scanning process from the satellite application to the FuidXel server, where user authentication and password are required to execute integration programs that are located on the server. On the other hand, in the integration between FuidXel and Alfresco Community Edition, authentication is required for sending electronic documents and descriptive metadata that comes by default in the CMIS protocol, used for sending the messages. S&T

References / Referencias

Alfresco (2017a). *Document and content capture with auto-categorization*. Retrieved from: https://www.alfresco.com/partners/solutions/document-and-content-capture-auto-categorization

Alfresco (2017b). *Document indexing module for alfresco share.* Retrieved from: https://www.alfresco.com/partners/solutions/document-indexing-module-alfresco-share

Archivo General de la Nación [AGN]. (2012). *Circular Externa 005 de 2012.* Retrieved from: http://www.archivogeneral.gov.co/sites/all/themes/nevia/PDF/Transparencia/CIRCULAR_05_DE_2012.pdf

Archivo General de la Nación [AGN]. (2013). *Acuerdo 005 de 2013.* Retrieved from: http://www.archivogeneral.gov.co/sites/all/themes/nevia/PDF/Transparencia/ACUERDO_05_DE_2013.pdf

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001). *Manifesto for agile software development.* Retrieved from: http://agilemanifesto.org/

Britto, J. (2016). Comparación de metodologías ágiles y procesos de desarrollo de software mediante un instrumento basado en CMMI. *Scientia Et Technica, 21*(2), 150-155. doi:10.22517/23447214.9249

Decreto 019 de 2012 (2012, enero 10). *Diario Oficial No. 48.308.* Bogotá, Colombia: Imprenta Nacional.

Decreto 2609 de 2012. (2012, diciembre 17). *Diario Oficial No. 48.647.* Bogotá, Colombia: Imprenta Nacional.

Ephesoft (2017). *Extracting meaning from unstructured content.* Retrieved from: http://www.ephesoft.com/products

European Commission - IDABC Project [EC-IDABC]. (2004, January). *Moreq: model requirements for the management of electronic records.* Retrieved from: http://ec.europa.eu/idabc/en/document/2303/5927.html

Franco, T. & Rosas, L. (2015). *Desarrollo e implementación de un sistema de gestión documental para uso interno de soproma (Generación y Digitalización de Documentos)* [tesis]. Universidad Central del Ecuador: Quito.

International Council of Archives [ICA]. (2011). ISAD(G): *General International Standard Archival Description* [2a ed.]. Retrieved from: http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

Jiménez, G. (2016). *Modelo de evaluación para la selección de herramientas de software libre en el proceso de gestión documental* [tesis]. UNAB: San Juan de Pasto, Colombia.

Larrañaga, U. (2008). *Metodología de digitalización de documentos.* Álava, España: Sociedad Informática del Gobierno Vasco. España.

Ley 527 de 1999. (1999, agosto 21). *Diario Oficial No. 43.673.* Bogotá, Colombia: Imprenta Nacional.

Ley 594 de 2000. (2000, julio 20). *Diario Oficial No. 44.093.* Bogotá, Colombia: Imprenta Nacional.

López, J. (2017). FuidxelServer. Sistema de apoyo a la digitalización para la gestión documental. Available at: https://github.com/javeeto/FuidxelServer

Márquez, B. & Chacón, Q. (2014). *Programa de gestión documental*. Bogotá, Colombia: MINTIC.

Mas, J., Megías, D., Ginestà, M., & Peña, A. (2005). *Ingeniería del software en entornos de software libre.* Barcelona, España: Fundació per a la Universitat Oberta de Catalunya.

Ministerio de las Tecnologías de la Información y las Comunicaciones. [MINTIC]. (2012b). Ministerio de las Tecnologías de la Información y las Comunicaciones. [MINTIC]. (2012a). *Cero papel en la administración pública. Guía No 3: Documentos electrónicos*. Bogotá, Colombia: MINTIC.

Ministerio de las Tecnologías de la Información y las Comunicaciones. [MINTIC]. (2012a). *Cero papel en la administración pública. Guía No 5: Digitalización certificada de documentos.* Bogotá, Colombia: MINTIC.

Monje, C. (2011). *Metodología de la investigación cuantitativa y cualitativa: guía didáctica.* Neiva, Colombia: Universidad Surcolombiana.

Navarro, A., Fernández, J., & Morales, J. (2013). Revisión de metodologías ágiles para el desarrollo de software. *Prospectiva*, 11(2), 30-39.

Paz, A., Castañeda, L., & Arboleda, H. (2011). Agile methodology for small teams using Microsoft platforms. *Sistemas & Telemática, 9*(18), 83-99. doi:10.18046/syt.v9i18.1078

Presidencia de la República de Colombia. *Directiva Presidencial 04 de 2012.* Retrieved from: https://www.mintic.gov.co/portal/604/articles-3647_documento.pdf

Vargas, Z. (2008). *La investigación aplicada: una forma de conocer las realidades con evidencia científica. Revista Educación, 33*(1), 155-165.

CURRICULUM VITAE

Javier López Martínez Systems Engineer, he worked as an analyst programmer at the Universidad el Bosque (Bogotá, Colombia) and as development software engineer contractor at the Superintendencia de Industria y Comercio (Bogotá, Colombia). Her main areas of expertise are: academic management, documents management, PHP programming languages, Javascript, HTML, and GNU Linux platform management / Ingeniero de Sistemas, trabajó como analista programador en la Universidad el Bosque (Bogotá, Colombia), y como ingeniero de desarrollo de software contratista en la Superintendencia de Industria y Comercio. Sus áreas de experiencia son: la gestión académica, la gestión documental, los lenguajes de programación PHP, Javascript y HTML, y la administración de plataformas GNU/Linux.