Implementation of the Document Capture Solution

Realizace systému pro digitalizaci dokumentů

Bc. Petr Zeman

Diploma thesis
2014

Tomas Bata University in Zlín
Faculty of Applied Informatics

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2013/2014

ZADÁNÍ DIPLOMOVÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Petr Zeman**
Osobní číslo:	**A12489**
Studijní program:	**N3902 Inženýrská informatika**
Studijní obor:	**Informační technologie**
Forma studia:	**kombinovaná**
Téma práce:	**Realizace systému pro digitalizaci dokumentů**

Zásady pro vypracování:

1. **Vypracujte rešerši na téma Digitalizace dokumentů.**
2. **Analyzujte a diskutujte požadavky na digitalizační systém fungující formou SaaS (Software as a Service).**
3. **Navrhněte a realizujte aplikace sloužící k digitalizaci a vytěžování strukturovaných a polostrukturovaných dokumentů.**
4. **Proveďte funkční a výkonové testování navrhnutých aplikací.**
5. **Diskutujte dosažené výsledky a závěry práce.**
6. **Naznačte další vývojové směry předloženého řešení.**

Rozsah diplomové práce:

Rozsah příloh:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **ZHU, Jackie, Ben ANTIN, Moe BRYAN, Patrick CHESNOT, Ben DAVIES, Tom STUART a Michael VAHLAND. Implementing Imaging Solutions With IBM Production Imaging Edition and IBM Datacap Taskmaster Capture. Vervante, 2011, p. IBM Redbooks, SG24-7969-00. ISBN 07-384-3607-0.**
2. **IBM CORPORATION. Application Development Guide: using IBM Datacap Taskmaster Capture v8.1. IBM Corp., 2013. Dostupné z: http://www-05.ibm.com/e-business/linkweb/publications/servlet/pbi.wss?PAG=C11&SSN=13L4H0002634086338&TRL=3251-05&LST=ALL&RPP=10&submit=Go**
3. **KUNSTOVÁ, Renata. Efektivní správa dokumentů: co nabízí Enterprise Content Management. Praha: Grada Publishing, 2009, 204 s. ISBN 978-80-247-3257-2.**
4. **FOWLER, Martin. UML distilled: a brief guide to the standard object modeling language. 3rd ed. Boston: Addison-Wesley, c2004, xxx, 175 p. ISBN 978-0321193681.**
5. **PATTON, Ron. Software testing. 2nd ed. Indianapolis: Sams Publishing, 2006. ISBN 06-723-2798-8.**

Vedoucí diplomové práce: **Ing. Tomáš Sysala, Ph.D.**
Ústav automatizace a řídicí techniky

Datum zadání diplomové práce: **21. února 2014**

Termín odevzdání diplomové práce: **20. května 2014**

Ve Zlíně dne 21. února 2014

prof. Ing. Vladimír Vašek, CSc.
děkan

L.S.

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky

doc. Mgr. Roman Jašek, Ph.D.
ředitel ústavu

ABSTRACT

Document capture is currently becoming more and more important ICT component in the management of the business processes in the private companies and in public institutions. Currently the world market offers several technologies that are capable of processing any structured, semi-structured and unstructured document. In this thesis I use and present the IBM Datacap Taskmaster Capture technology and demonstrates its possibilities on a sample project for a mid-sized company that solves the problem of processing incoming invoices and internal delivery notes between branch offices.

The first chapter of this thesis depicts the Enterprise Content Management in general and presents the Capture component and its possibilities. The second chapter introduces well-known and important technologies that currently exist in the world market. The third chapter describes in detail the architecture and functioning of the IBM Datacap Taskmaster Capture. The fourth chapter describes the requirements for a software which should be able to run in the SaaS cloud model. The fifth chapter discusses the requirements specified in the fourth chapter in the context of the deployment of IBM Datacap Taskmaster Capture in the SaaS model and also proposes a possible deployment architecture. The sixth chapter describes collection and analysis of requirements for the document capture of incoming invoices and delivery notes between the branch offices. This chapter also presents the application design based on the analyzed requirements and describes the application implementation itself. The last, seventh chapter shows the results of functional and performance testing conducted on the proposed application. The achieved results and proposal for further development directions are discussed in the conclusion.

Keywords: Document capture, IBM Datacap Taskmaster Capture, SaaS

ABSTRAKT

Technologie skenování a vytěžování dat z dokumentů se v současné době stává stále významnějším ICT prvkem v oblasti řízení business procesů v soukromých společnostech, i státních institucích. V současné době je na světovém trhu několik technologií, které jsou schopny vytěžovat jakékoli strukturované, polo-strukturované i nestrukturované dokumenty. V rámci této práce využívám technologii IBM Datacap Taskmaster Capture,

na které demonstruji ukázkový projekt pro středně velkou společnost, která řeší problém zpracování příchozích faktur a interních dodacích listů v rámci oblastních poboček.

První kapitola této práce se zabývá obecně problematikou Enterprise Content Managementu a představuje komponentu Capture a její možnosti. Druhá kapitola představuje nejznámější a významné technologie, které se v současné době vyskytují na světovém trhu. Třetí kapitola detailně popisuje architekturu a funkci systému IBM Datacap Taskmaster Capture. Čtvrtá kapitola charakterizuje požadavky na software, který by měl být schopen provozu v cloudovém modelu SaaS. Pátá kapitola diskutuje požadavky stanovené ve čtvrté kapitole v souvislosti s nasazením systému IBM Datacap Taskmaster Capture v modelu SaaS. Zároveň navrhuje možnou architekturu nasazení systému. Šestá kapitola se zabývá sběrem a analýzou požadavků na digitalizaci faktur a dodacích listů v rámci společnosti a na základě těchto požadavků popisuje návrh aplikace, která tyto požadavky řeší. V rámci této kapitoly je rovněž popsána samotná realizace této aplikace. Poslední sedmá kapitola znázorňuje výsledky funkčního a výkonového testování, které bylo provedeno na navržené aplikaci. Funkční testování je provedeno na základě definovaných testovacích scénářů, které jsou sestaveny na základě případů užití. Aplikace splňuje stanovené případy užití. V rámci výkonového testování jsou testovány části systému, které nevyžadují uživatelskou interakci, vyjma procesu skenování. Pro test jsou použity tři různá rozlišení dokumentů a pro každý rozlišení je změřen čas zpracování v rámci kroku procesu. Zároveň je vyhodnocena úspěšnost vytěžení údajů v závislosti na zvoleném rozlišení dokumentu. Z výsledků vyplývá, že nejlepších výsledků lze dosáhnout při použití rozlišení 300 DPI. V závěru práce jsou diskutovány dosažené výsledky a navrženy další vývojové směry aplikace.

Klíčová slova: Digitalizace a vytěžování dokumentů, IBM Datacap Taskmaster Capture, SaaS,

I would like to thank my supervisor Ing. Tomáš Sysala, Ph. D. who provided me necessary information, valuable advice, comments and steering to create this thesis. I would also like to thank my family for their lasting support during its arising.

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně …….

podpis diplomanta

CONTENTS

INTRODUCTION

Nowadays many companies have a need to optimize their process of handling the paper documents. The pressure for the minimizing the archived paper documents is becoming the main company concern. System that provides the ability to capture the documents and recognize the desired data faster and more effectively is the right solution for these companies.

The main objective of this thesis is to present the IBM Datacap Taskmaster Capture system that delivers automated process for capturing any kind of the paper of electronic documents. The solution provided in this thesis is for the sample mid-sized company that needs to capture the incoming invoices and delivery notes between the branch offices. Another objective is to analyze the requirements of this company, propose the right solution and deploy the final solution. Thesis also discuss and proposes the possible deployment of the IBM Datacap Taskmaster Capture in the SaaS cloud model.

The thesis focuses on the document capture solution that opens up another possibilities for preserving the exported documents and their data. The preservation and another subsequent processes are not described in this thesis.

I. THEORY

1 DOCUMENT CAPTURE AS PART OF ENTERPRISE CONTENT MANGEMENT

Document scanning, document capture, imaging, data capture, all these terms represent process of transforming paper documents into electronic form. The differences between these terms are in the level of functionality and system integration but they are part of Enterprise Content Management.

1.1 Enterprise Content Management

Association for Information and Image Management (AIIM) is internationally recognized authority that first defined the term: Enterprise Content Management (ECM). The acronym ECM has been reinterpreted and redefined many times during last ten years.

Since 2003, the AIIM has defined ECM as follows:

The technologies used to capture, manage, store, deliver, and preserve information to support business processes.

In year 2005 there was change in definition which unfortunately cut out the process component in definition:

Enterprise Content Management is the technologies, tools, and methods used to capture, manage, store, preserve, and deliver content across an enterprise.

The last change in definition was in year 2010 when AIIM added more detail into definition:

Enterprise Content Management is the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists. (1)

There are several other definition of the ECM such as:

Enterprise Content Management (ECM) spans a broad category of systems, strategies and tools designed to improve organizational processes that involve the capture, management, preservation and delivery of information. (2)

Enterprise content management (ECM) is a set of defined processes that allow a corporation, agency or organization to obtain, organize, store and deliver information crucial to its operation in the most effective manner possible. (3)

Above mentioned definitions are not in conflict with the original AIIM definition. They even more develop the original definition and show the entire wide of the ECM.

When considering the topic of content management we have to distinguish ECM between general category Content Management (CM) as well as two specific things, the Web Content Management (WCM) and Enterprise Content Management (ECM). They have different origins, different functions and go very apart in the claim. (4)

Content management (CM) is a repeatable method of identifying all content requirements up front, creating consistently structured content for reuse, managing that content in a definitive source, and assembling content on demand to meet the customers' needs. (5)

From a process point of view, CM is a process for **collecting**, **managing**, and **publishing** content.

- **Collection**: You either create or acquire information from an existing source. Depending on the source, you may or may not need to convert the information to a master format (such as XML). Finally, you aggregate the information into your system by editing it, segmenting it into chunks (or components), and adding appropriate metadata.
- **Management**: You create a repository that consists of database records and/or files containing content components and administrative data (data on the system's users, for example).
- **Publishing**: You make the content available by extracting components out of the repository and constructing targeted publications such as Web sites, printable documents, and e-mail newsletters. The publications consist of appropriately arranged components, functionality, standard surrounding information, and navigation. (5)

Many people think of content management as Web content management—and it is about Web content management—but Web content may be only one of the types of information you need to manage. Most organizations need to manage both paper and Web, and they often also manage common content between the mediums. Sometimes this is known as Enterprise Content Management. (5)

Although there are several ECM concepts (based on the specific system), this chapter describes ECM basics based on the AIIM principles. The reason is because AIIM describes ECM in general point of view.

1.1.1 Enterprise Content Management architecture

ECM can be understood as three solutions: as a middleware, as independent services and as uniform repository. Refer to the **Figure 1** for the general ECM architecture in context with other applications.

- **Enterprise Content Management as integrative middleware.** ECM is used to overcome the restrictions of former vertical applications and island architectures. The user is basically unaware of using an ECM solution. ECM offers the requisite infrastructure for the new world of web-based IT, which is establishing itself as a kind of third platform alongside conventional host and client/server systems. Therefore, Enterprise Application Integration (EAI) – will play an important role in the implementation and use of ECM. ECM is an essential component of service-oriented applications (SOA). (1)

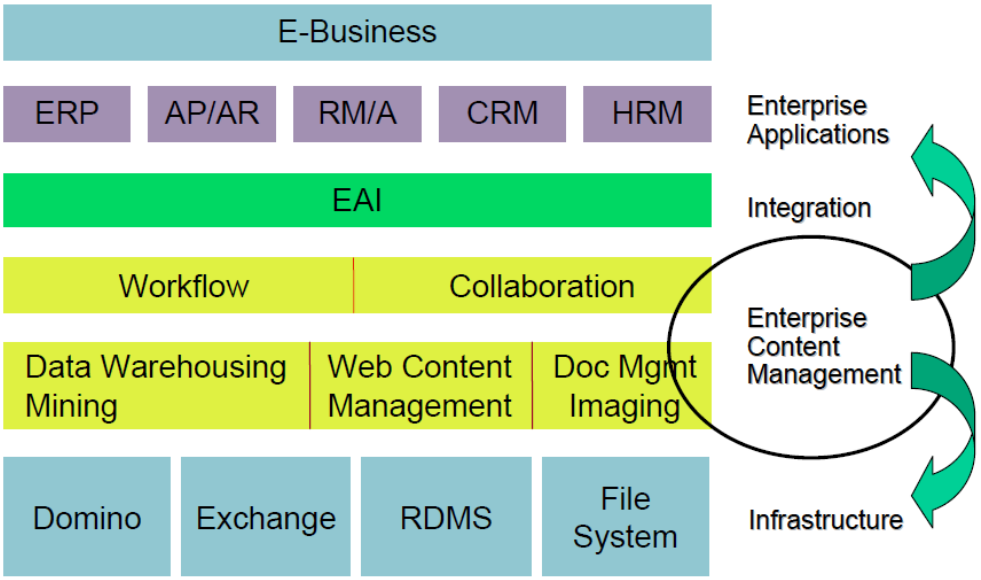

Figure 1 – ECM as vertical information system infrastructure (1)

- **Enterprise Content Management components as independent services**. ECM is used to manage Information without regard to the source or the required use. The functionality is provided as a service that can be used from all kinds of applications. The advantage of a service concept is that for any given functionality

only one general service is available, thus avoiding redundant, expensive and difficult to maintain parallel functions. (1)

- **Enterprise Content Management as a uniform repository for all types of information.** ECM is used as a content warehouse (both data warehouse and document warehouse) that combines company information in a repository with a uniform structure. Expensive redundancies and associated problems with information consistency are eliminated. All applications deliver their content to a single repository, which in turn provides needed information to all applications.

ECM thus is a collection of infrastructure components that fit into a multi-layer model and include all Document Related Technologies (DRT) for handling, delivering, and managing poorly structured data. As such, Enterprise Content Management is one of the necessary basic components of the overarching E-Business application area. ECM also sets out to manage all the information of a WCM and cover archiving needs as a universal repository.

Figure 2 describes ECM high level architecture with ECM parts on the left side, example of the business applications described on the right side and connection between ECM and business applications.

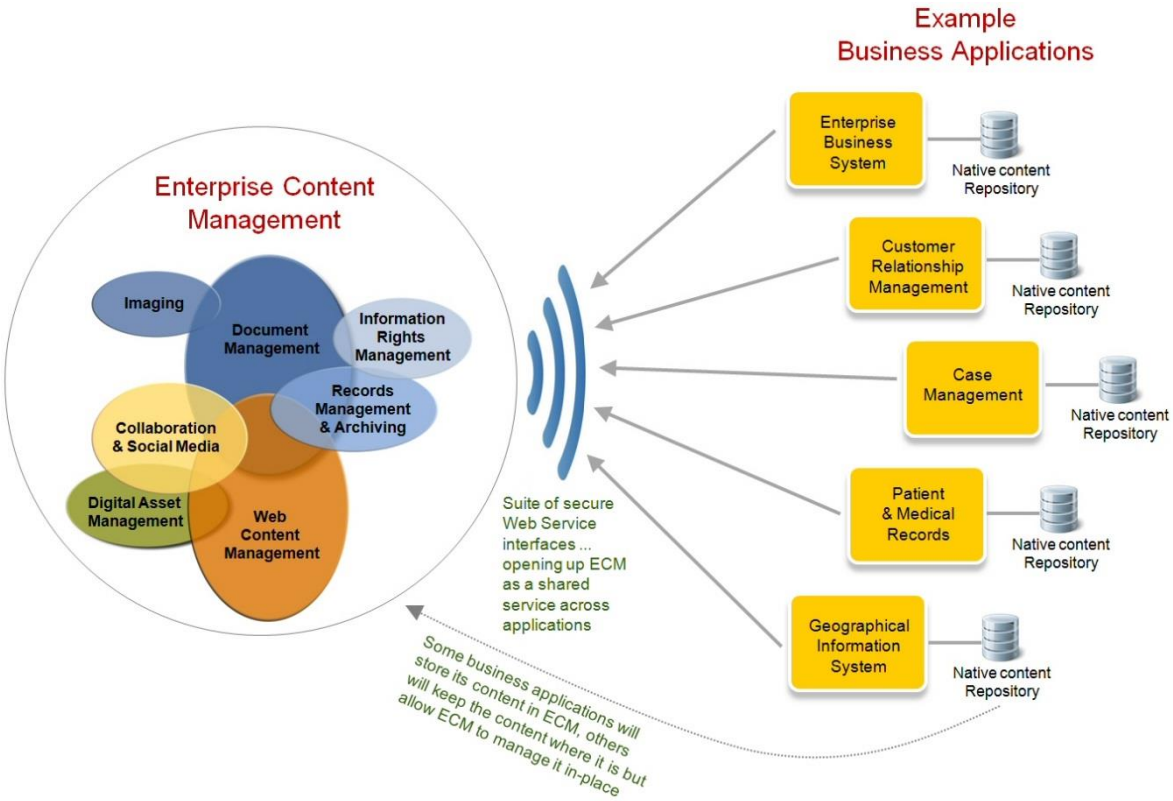

Figure 2 – High level ECM system architecture (6)

1.1.2 The Five Components of Enterprise Content Management

Enterprise Content Management solutions combine a wide variety of technologies and components, some of which can also be used as stand-alone solutions without necessarily being incorporated into an enterprise-wide system. (7)

These ECM components and technologies can be categorized as:

- **Capture** contains functionalities and components for generating, capturing, preparing and processing analog and electronic information. There are several levels and technologies, from simple information capture to complex information preparation using automatic classification. Capture components are often also called "Input" components. (7)
  Refer to chapter 1.2 for detailed description.
- **Manage** components are for the management, processing, and use of information. They incorporate: *Databases for administration and retrieval* and *Access authorization systems*. The goal of a closed ECM system is to provide these two components just once as services for all "Manage" solutions such as Document Management, Collaboration, Web Content Management, Records Management and Workflow / Business Process Management. (7)
- **Store** components are used for the temporary storage of information which it is not required or desired to archive. Even if it uses media that are suitable for long-term archiving, **Store** is still separate from **Preserve**. (7)
- **Preserve** components of ECM handle the long-term, safe storage and backup of static, unchanging information, as well as temporary storage of information that it is not desired or required to archive. (7)
- **Deliver** components of ECM are used to present information from the **Manage**, **Store**, and **Preserve** components. They also contain functions used to enter information in systems (such as information transfer to media or generation of formatted output files) or for readying (for example converting or compressing) information for the **Store** and **Preserve** components. The functionality in the **Deliver** category is also known as "output". (7)

This model is based on the five lead categories of AIIM International (**Figure 3**).

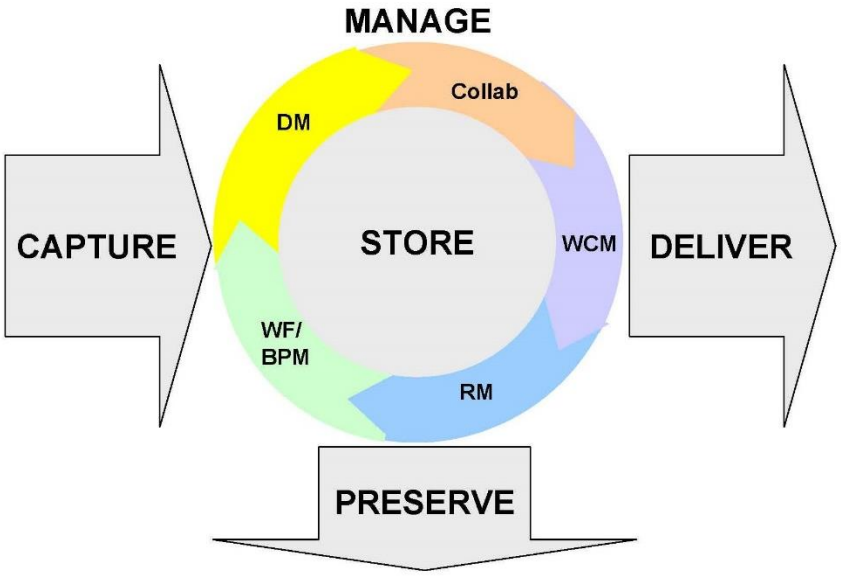

Figure 3 – The 5-component-model of ECM (1)

The traditional ECM application areas are

- Document Management (DM)
- Collaboration (of supporting systems, groupware)
- Web Content Management (WCM)
- Records Management (RM)
- Workflows / Business Process Management (BPM)

These areas form the "manage" components that connect Capture, Store, Deliver and Preserve and can be used in combination or as alternatives. While Document Management, Web Content Management, Collaboration, Workflow and Business Process Management are more for the dynamic part of the life cycle of information, Records Management takes care of information which will no longer be changed. (1)

The utilization of the information is paramount throughout, whether through independent clients of the ECM system components, or by enabling existing applications that access the functionality of ECM services and the stored information. The integration of existing technologies makes it clear that ECM is not a new product category, but an integrative force.

The scope of this thesis is mainly from the ECM Capture part. Next chapters provide detailed information about ECM Capture, document capture technologies and implementation of the ECM capture solution.

1.2 Enterprise Content Management Capture

The Capture contains functionalities and components for generating, capturing, preparing and processing analog and electronic information. There are several levels and technologies, from simple information capture to complex information preparation using automatic classification. Capture components are often also called "Input" components. (7)

1.2.1 Basic terminology

This sub-chapter describes fundamental terminology associated with the ECM Capture.

Document

The concept of document has been defined as:

Any concrete or symbolic indication, preserved or recorded, for reconstructing or for proving a phenomenon, whether physical or mental. (8)

The Czech Archives Law (Zákon 499/2004 § 2 bod d) defines document as,

Každý písemný, obrazový, zvukový, elektronický nebo jiný záznam, ať již v podobě analogové či digitální, který vznikl z činnosti původce.

In English translation:

Each written, visual, aural, electronic or any other record, whether in the form of analog or digital, which originated from the work of the originator.

The document is categorized into three categories based on the content:

- **Structured document** – A document for which both the type (number, letter, check mark, etc.) and location of data is known before scanning. For example, the data field for line 35 of IRS tax form 1040, positioned on the lower right corner of the page, will always contain a number. (9)
- **Semi-structure document** – A document that includes known types of data, but where on the page this data is positioned is not known. An example is an invoice. It's known that an invoice must include an amount and date due, but since every company is free to create their own invoices, there's no way to know where they might position this information on the page. (10)
- **Unstructured document** – A type of document for which, prior to scanning, both the type and location of the information it contains is unknown. Documents that

cannot be identified as structured or semi-structured are assigned to this category. They could be virtually any type of document: correspondence, petitions, advertisements, manuals, brochures or annual reports. (11)

Metadata

Metadata describes other data. It provides information about a certain item's content. For example, an image may include metadata that describes how large the picture is. (12)

In connection with document, metadata contains information about captured data from the document.

Batch

A group of one or more documents processed in a single scan operation. (13)

1.2.2 Parts of the ECM Capture

Figure 4 describes input documents, processes and technologies that are considered as a part of ECM Capture

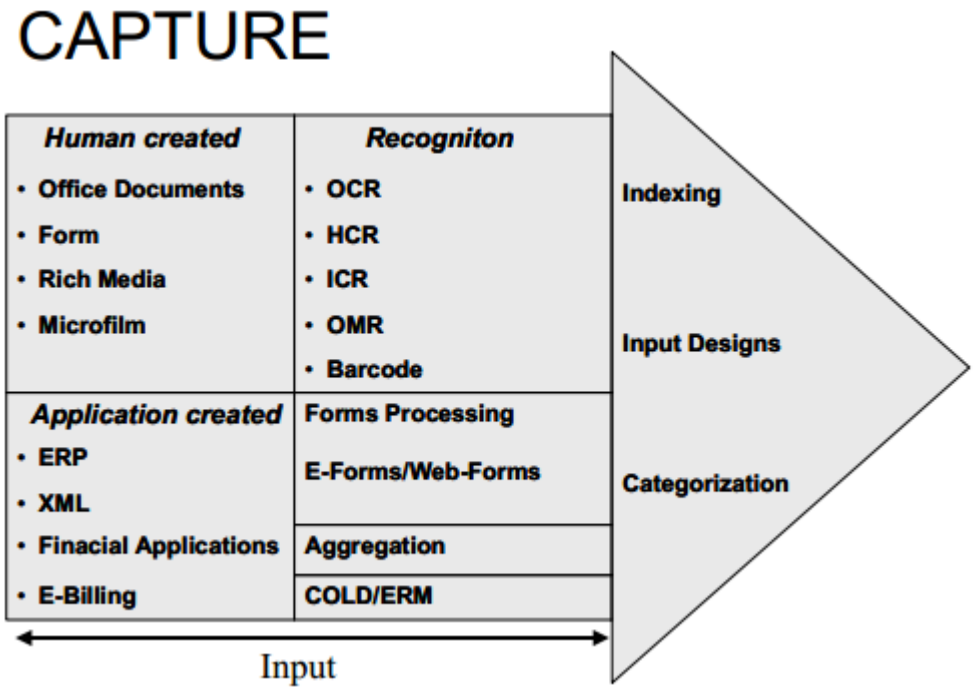

Figure 4 – ECM Capture components (1)

Manually generated and captured information (Human and application created)

Manual capture can involve all forms of information, from paper documents to electronic office documents, e-mails, forms, multimedia objects, digitized speech and video, and

microfilm. Automatic or semi-automatic capture can use EDI or XML documents, business and ERP applications or existing specialist application systems as sources. (1)

Technologies for processing captured information (Recognition)

- **OCR** (Optical Character Recognition) This method converts image information into machine-readable characters. OCR is used for printed text.
- **HCR** (Handprint Character Recognition) This refinement of OCR converts handwriting or lettering into machine characters, but does not yet give satisfactory results for running text. However, for defined field content it has become very reliable. (**Figure 5**)
- **ICR** (Intelligent Character Recognition) ICR is a further development of OCR and HCR that uses comparison, logical connections, and checks against reference lists and existing master data to improve results. (**Figure 5**)
- **OMR** (Optical Mark Recognition) OMR, as used for barcodes for example, reads special markings in predefined fields with very high accuracy. It has proven its value in questionnaires and other forms. (**Figure 6**)
- **Barcode** Barcodes on mailed forms allow for the automatic recognition and filing of returns. Nowadays technologies recognize 1D and 2D barcodes (CODE39, CODE128, Data Matrix, QR Code etc.) (**Figure 7**) (1)

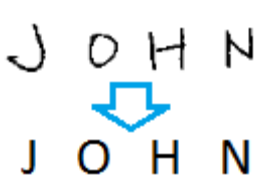

Figure 5 – HCR/ICR

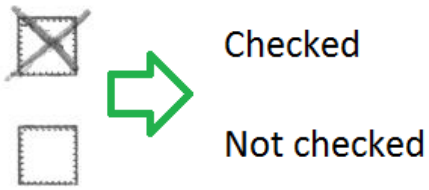

Figure 6 - OMR

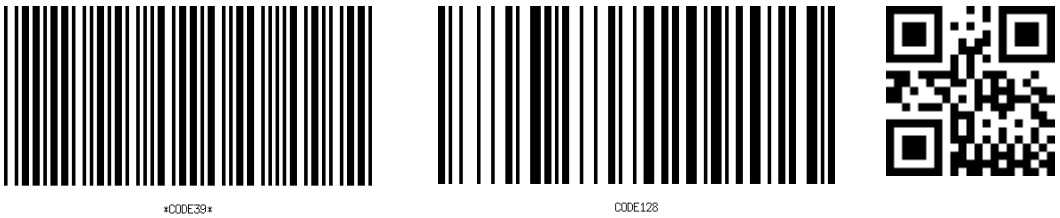

Figure 7 – Examples of 1D and 2D barcodes (Code39, Code128, QR code)

Document imaging processing techniques are used to show scanned images, and also allow legibility enhancement for capture. Functions like “despeckling,” which removes isolated

pixels, or "adjustment," which straightens images from sheets that feed in at an angle, improve the results of recognition technologies. (1)

Forms Processing

In forms capture, there are two groups of technologies, although the information content and character of the documents may be identical.

- **Forms Processing** means the capture of industrially or individually printed forms via scanning. Recognition technologies are often used here, since well-designed forms enable largely automatic processing.
- **E-Forms / Web-Forms** Automatic processing can be used to capture electronic forms as long as the layout, structure, logic and contents are known to the capture system (7)

Aggregation

Aggregation is a process of combining data entries from different creation, capture, and delivery applications. The goal is to combine and unify data from different sources, in order to pass them on to storage and processing systems with a uniform structure and format. (7)

Computer Output on Laser Disk (COLD/ERM)

COLD/ERM are technologies for the automatic processing of structured entry data and is still in use although laser disks have not been on the market for years. The acronym ERM stands for Enterprise Report Management. In both, supplied output data is processed based on existing structure information in such a way that it can be indexed independently of the origination system, and transferred to a storage component that can be dynamic (Store) or an archive (Preserve). (7)

Components for subject indexing of captured information

Systems incorporate further components for subject indexing and getting captured digital information to the appropriate recipients. These include:

- **Indexing** (Manual). In English parlance, indexing refers to the manual assignment of index attributes used in the database of a "manage" component for administration and access.
- **Input Designs** (profiles). Both automatic and manual attributing can be made easier and better with preset profiles. These can describe document classes that

limit the number of possible index values, or automatically assign certain criteria. Input designs also include entry masks and their logic in manual indexing.

- **Taxonomy.** The taxonomic processing facilitates a formal order of information according to the respective needs of an enterprise. Here nomenclatures, thesaurus and file systematics play a role.
- **Categorization** (automatic classification or categorizing). Based on the information contained in electronic information objects, whether OCR converted faxes, office files or output files, automatic classification programs can extract index, category, and transfer data autonomously. These systems can evaluate information based on predefined criteria or in a self-learning process. (1)

The objective of all "Capture" components is the provision of information to the "Manage" components for further processing or archiving. (1)

2 DOCUMENT CAPTURE TECHNOLOGY OVERVIEW

Nowadays many companies are focusing on the document capture, recognition or imaging technologies. This chapter describes the world-wide known companies and technologies in document capture. All companies that are mentioned in this chapter aim on many IT areas, but this chapter is focusing only on the document capture related technologies.

Major document capture companies are (alphabetically ordered):

- ABBYY
- EMC^2
- IBM
- Kofax
- Nuance
- OpenText
- ReadSoft

2.1 ABBYY

ABBYY is a leading provider of document conversion, data capture and linguistic software. The key areas of ABBYY's research and development include document recognition technologies and applied linguistics.

Today, ABBYY is an international company with over 1250 regular and 900 freelance employees worldwide. ABBYY products are being sold in more than 130 countries around the world through an extensive network of regional and international partners. The overall number of users of ABBYY products, which range from end user applications for PC and mobile devices to enterprise solutions and development tools, exceeds 30 million people according to internal research. (14)

Major document capture technologies:

- **ABBYY Recognition Server** is a flexible and easy-to-deploy server-based OCR solution designed for high-volume document processing. (15)
- **ABBYY FlexiCapture** is the next generation of intelligent, accurate and highly scalable data capture and document processing software. It provides a single entry point to automatically transform streams of different forms and documents of any structure and complexity into business-ready data. (16)

2.2 EMC[2]

EMC is a global leader in enabling businesses and service providers to transform their operations and deliver information technology as a service (ITaaS). Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect and analyze their most valuable asset — information — in a more agile, trusted and cost-efficient way. EMC employs approximately 60,000 people worldwide. (17)

Major document capture technologies:

- **EMC Captiva QuickScan Pro** is an out-of-the-box desktop scanning and document capture solution. (18)
- **EMC Captiva Capture** enables organizations to capture documents and data from paper, electronic files, and other sources, transforming it into digital content and delivering it into content management systems and business processes. (19)
- **EMC Captiva Advanced Recognition** automatically recognizes large amounts of paper documents, captures the data, and delivers the information to the right systems, processes, and people. It enables businesses to identify all kinds of documents using intelligent recognition technologies, and automatically capture machine- and hand- printed data using optical character recognition (OCR), intelligent character recognition (ICR), and barcode recognition. (20)

2.3 IBM

IBM (International Business Machines) is one of the two (second is Hewlett-Packard) world's largest information technology company in terms of revenue ($104 billion in 2012). IBM products include hardware and software for a line of business servers, storage products, custom-designed microchips, and application software. Increasingly, IBM derives revenue from a range of consulting and outsourcing services. With the advent of the low-cost microchip, the personal computer, distributed computing, open rather than proprietary standards, and the Internet. At the end of 2012, IBM had over 434246 employees. (21)

Major document capture technology:

- **IBM Datacap Taskmaster Capture** is one of the industry’s most flexible three-tier client server capture platforms. Driven by a highly configurable procedural

rules engine, it combines advanced capture functionality with superior flexibility. Clients use Datacap Taskmaster Capture to mix thin client and thick client implementations to create highly scalable, distributed capture applications with multiple types of recognition, document identification and automated validations. (22)

2.4 Kofax

Kofax® Limited is a software provider based in Irvine, California. It provides smart process applications that simplify the business critical First Mile™ of information-intensive customer interactions. Kofax combines market leading capture, process management, analytics and mobile capabilities that enable organizations to increase their responsiveness to customers, provide better service, gain a competitive advantage and better grow their businesses while reducing operating costs. (23)

Major document capture technologies:

- **Kofax Express** enables companies to control the flow of information entering their enterprise by converting stacks of paper into electronic actionable and managed business content that can be stored and archived. (24)
- **Kofax Capture** automates and accelerates business processes by capturing all types of paper and electronic documents and forms, transforming them into accurate and actionable information, and delivering it all into your core business applications, processes and workflows. (25)
- **Kofax Transformation Modules™ (KTM)** is an integrated platform of applications that streamline the transformation of different document types into structured electronic information, ready for delivery into business systems and processes. (26)

2.5 Nuance

Nuance Communications is an American multinational computer software technology corporation, headquartered in Burlington, Massachusetts, United States, that provides speech and imaging applications. Current business products focus on server & embedded speech recognition, telephone call steering systems, automated telephone directory services, medical transcription software & systems, optical character recognition software, and desktop imaging software. (27)

Major document capture technologies:

- **OmniPage** turn high volumes of paper and digital documents into files you can edit, search and share in the format of your choice. (28)
- Nuance has a several technologies that are connected with MFP like **DigiDocFlow**, **eCopy**, **SimplifyScan**, **SmartOffceScan** etc.

2.6 OpenText

OpenText Corporation is headquartered in Waterloo, Ontario, Canada and is Canada's largest software company. It produces and distributes Enterprise Information Management (EIM) software solutions for large corporations across all industries. OpenText solutions are aimed at addressing information management requirements, including the management of large volumes of content compliance with regulatory requirements, and mobile and online experience management. OpenText employs over 5,000 people worldwide. (29)

Major document capture technologies:

- **OpenText Imaging** – A complete solution for capturing and displaying the complete range of business documents.
- **OpenText Capture Center** – Automatically captures and interprets paper documents, scanned images, email, and faxes using sophisticated document and character recognition software.
- **OpenText Tempo Box** – Simplifies the content management experience, and allows users to easily sync, and share information across multiple devices, without sacrificing the records management rigor and security demanded by your organization's policies and regulations. (30)

2.7 ReadSoft

ReadSoft is a company that develops markets and supports software that automates the processing of documents, such as invoices, in different business processes and ERP environments within organizations. ReadSoft was founded by two university students in Linköping, Sweden, in 1991, both of which are still actively involved in the management of the company. Today ReadSoft employs ca 500 employees across the globe. Its headquarters is in Helsingborg, Sweden. ReadSoft has subsidiaries in 16 countries and partners in an additional.

ReadSoft is also a name for the document capture technology. ReadSoft focuses mainly on invoice and forms processing.

The main focus of this thesis is to design, develop and perform application testing based on IBM Datacap Taskmaster Capture technology. The next chapters of this thesis are focused on this technology.

3 IBM DATACAP TASKMASTER CAPTURE ARCHITECTURE

IBM® Datacap Taskmaster Capture is a complete solution for document and data capture. Taskmaster scans, classifies, recognizes, validates, verifies, and exports data and document images quickly, accurately and cost-effectively. By combining recognition engines for OCR, ICR, OMR, and bar codes with libraries of hundreds of script-based and code-based (.NET) actions, Taskmaster accurately captures data from any type of structured, highly variable, or unstructured documents. Taskmaster Capture can capture machine print, hand print, bar codes, and check box data. By using the Taskmaster flexible rules engine, data capture can be tailored to fit the most demanding business requirements and can be changed quickly when business needs change. For indexing applications, Taskmaster streamlines the manual data entry of index entries by using recognition to automatically identify the index values on each document and to automate the document identification process. (31)

3.1 Taskmaster application architecture

Taskmaster applications are designed to scan, process, and verify the data in documents. Although each Taskmaster application is different, most include seven basic steps.

- **Page input** – Scan a batch of hardcopy pages or import electronic documents into application. The output from this stage is a batch of individual TIFF image files. Each page is initially assigned the page type “Other”.
- **Page identification** – Perform image enhancement to improve the image quality. Then, determine each page type, automatically or by displaying it to an operator for manual identification if necessary. The goal is to identify the page type, but not a variant (for example, an airline ticket, but not a ticket from a specific airline).
- **Document assembly** – Organize the individual page files into a document according to predefined document definitions (for example, a form might have two required pages and an optional attachment). Run document integrity confirmation to ensure that each document satisfies the rules for that document type.
- **Data recognition** – On each page, locate the data fields for that page type (for example, an airline ticket contains a passenger name, a departure airport). Then, use a Taskmaster recognition engine to obtain the character data for each field. The recognition engine indicates the degree of confidence for each character.

- **Data validation** – Check the validity of specific fields. For example, you can check for valid dates, valid field formats, and correct totals. You can also complete searches to ensure that a state abbreviation is valid, or a purchase order number matches an item in a purchase order database.
- **Data verification** – Display low-confidence data and fields that failed validation to an operator for verification, correction, and exception handling. When the operator submits the batch, the application runs the validation rules again to ensure that all data satisfies the validation criteria.
- **Data export –** Export the data or document images to a text file, an XML file, a database, a Document Management system, or the next stage in a workflow. (32)

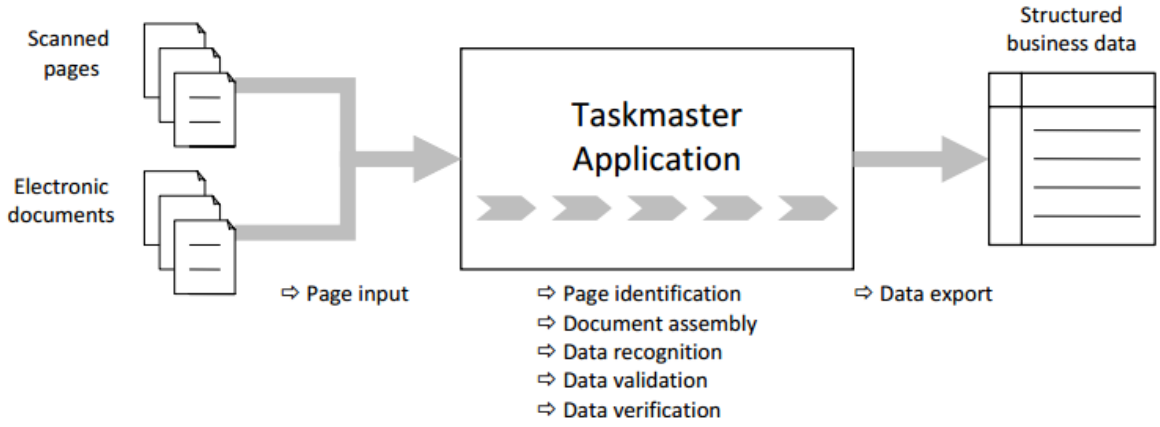

Figure 8 – General Taskmaster application architecture (33)

3.1.1 Page input

Taskmaster works primarily with TIFF image files. The first activity in any Taskmaster workflow is to convert the documents to TIFF format and insert the documents into an input repository.

Documents can be **hardcopy** or **electronic**. If the documents are hardcopy, they must be scanned and moved the resulting files to the application repository. Electronic documents can come from various sources in various formats. (32)

3.1.2 Page identification

Taskmaster supports several methods for page identification, including but not limited to:

- **Fingerprint matching** – Taskmaster generates a "fingerprint" that describes each incoming page. The fingerprint can include information about the relative densities

of different regions of the page or the location of text on the page. Taskmaster then compares the new fingerprint to a library of fingerprints for known page types. When it finds a match it assigns the corresponding page type.

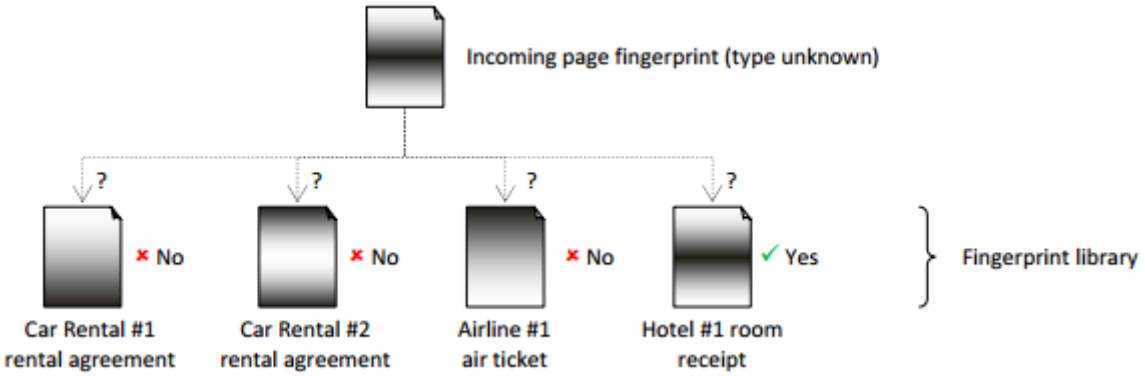

Figure 9 – Fingerprint matching (33)

- **Structure-based identification** – Structure-based identification uses the position of a page within the batch to determine its type. If application handles only one page type, or if the document structure is consistent (for example, all documents are two pages with a main page and a trailing page), it can assign page types based on position within the batch.
- **Text matching** – To complete page identification by using text matching, Taskmaster must first complete a full page recognition. Taskmaster then searches the recognition results for a string that is unique to each page type.
- **Manual page identification** – The page identification techniques described so far all identify pages automatically. It's also possible to configure application to display unrecognized pages to an operator for manual identification

Additionally, if application supports only a single page type, it can simply assign a static page type to all incoming pages. (32)

3.1.3 Document assembly

Taskmaster identifies incoming pages and assigns the correct page type by using fingerprint matching or one of the other identification methods. The next step assembles the batch of individual pages into documents according to the rules that are defined within the document hierarchy.

The **document hierarchy** (DCO) describes the structure of the documents that application is designed to process. The levels within the hierarchy are **batch**, **document**, **page**, and **field**. (32)

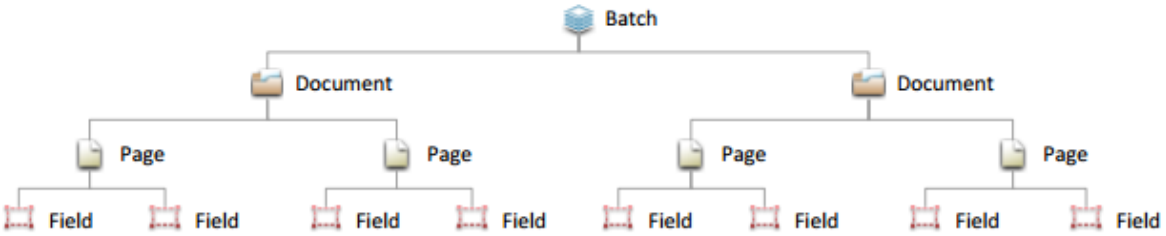

Figure 10 – Document hierarchy (33)

At the top of the document hierarchy is the batch, which refers to all pages of all document types. Beneath the batch level, the document hierarchy defines:

- The **document** types the application can process. An application can process only one document type, or multiple document types.
- The **page** types within each document type. Each document can contain only one page type or multiple page types.
- The number and order of pages within each document type. Pages can be required or optional.
- The data **fields** within each page type. Data fields can be required or optional. (32)

3.1.4 Data recognition

Data recognition is the stage during which Taskmaster locate the fields that we want to capture and then convert the fields into character-based data. The data that is obtained from **recognition** is stored in the page data files that Taskmaster set up in the document assembly stage.

Full **page recognition**, as the name suggests, uses the text and location of text on the page. Taskmaster includes three optical character recognition (OCR) engines, plus one intelligent character recognition (ICR) engine that it can be used to do full page recognition:

- **OCR_A** – ABBYY FineReader OCR engine.
- **OCR_S** – Nuance (formerly ScanSoft) OmniPage OCR engine.
- **OCR_SR** – Newer implementation of the Nuance OmniPage OCR engine.
- **ICR_C** – Open Text RecoStar ICR engine.

Other ICR engines are available as options. (32)

3.1.5 Data validation

Data validation determines whether captured data complies with the data integrity rules that are defined in your business requirements. A validation failure does not necessarily mean that the original page contains invalid data. It might mean that the recognition engine failed to recognize one or more characters correctly.

Taskmaster contains many choices how to validate the data. From simple data format validity, calculation the captured data to using the external data sources during validation (e.g. ERP system).

3.1.6 Data verification

During verification, Taskmaster displays pages to an operator for manual checking and possible correction. There are three primary reasons to display pages to an operator:

- The batch failed document integrity checking.
- A page contains one or more characters or OMR fields that were marked low confidence by the recognition engine.
- A page does not pass a validation rule because there is a problem with the integrity of the data. (32)

3.1.7 Data export

Taskmaster can export data to a text file, an XML file, a database, a Document Management system, or a custom business process. The default output format is a text file, but you can use some actions to export data to a database and an XML file. (32)

3.1.8 The Taskmaster workflow and rule execution

During the data capture process, documents go through a workflow that consists of several tasks, including page identification, character recognition, field validation, verification, and export. Some tasks require operator intervention, while other tasks run automatically.

A **workflow** contains **jobs** and **tasks**. Furthermore, tasks are associated with **task profiles** that contain **rules** and **actions** that are applied by the tasks while a job is processing a batch.

A job consists of one or more tasks. To process a batch of documents, it must run the batch through each task in the selected job. Some tasks (for example, Export) run without operator intervention, whereas others (for example, Verify) require an operator.

Each task is linked to a task profile that includes one or more rulesets. A ruleset consists of one or more rules. The rule itself is defined by the programmed **functions** and actions within it.

3.2 Taskmaster platform architecture

IBM® Datacap Taskmaster Capture provides a flexible and scalable architecture for distributing tasks across machines according to the anticipated processing load.

At one end of the spectrum is the single machine configuration, where all Taskmaster software components are installed on the same machine. This configuration is typically used for providing product demonstrations, in a proof of concept environment, or during initial product evaluation.

At the other end of the spectrum is the client/server configuration, where the various Taskmaster software components are installed on dedicated machines, such as web servers and database servers. This configuration can support hundreds of simultaneous users, and uses centralized application management and shared databases.

The Taskmaster system includes twelve components. Refer to **Figure 11** for Taskmaster system high level architecture:

- Taskmaster server
- Taskmaster databases
- Taskmaster File server
- Rulerunner service
- Taskmaster Web Server
- Taskmaster Web Client
- Taskmaster thick client
- Datacap Studio
- NENU
- RV2
- Business applications or databases

- Lightweight Directory Access Protocol (LDAP) or Active Directory (34)

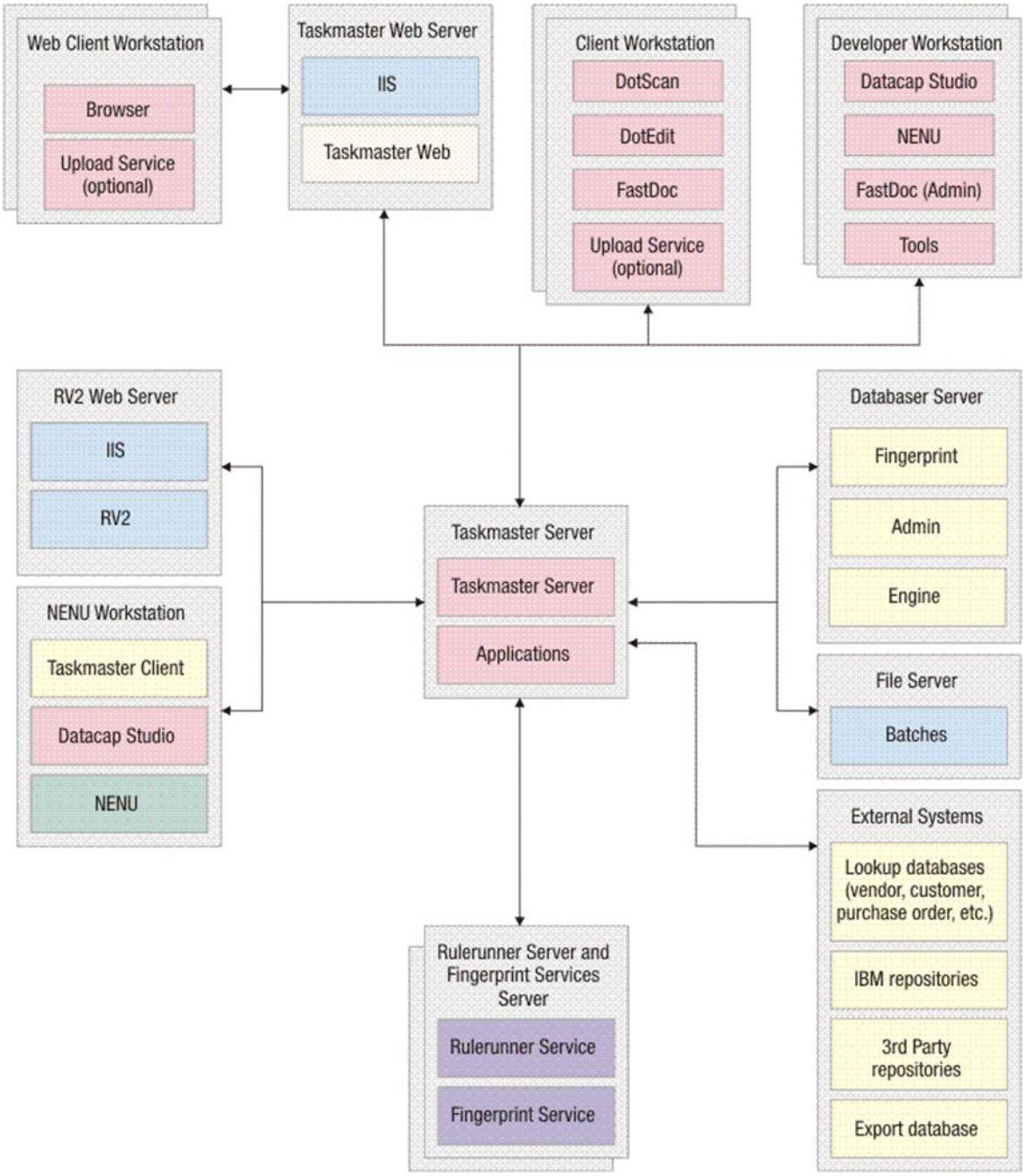

Figure 11 – Taskmaster system high level architecture (35)

3.2.1 Taskmaster Server

The Taskmaster Server component provides the core functions of the Taskmaster system. It manages and serves batches to workstations and users. It also orchestrates the tasks according to the workflow of the Taskmaster application. It provides user authentication

and access control, assigns batch IDs, and controls batch queuing, and controls access to the Taskmaster databases.

All communications between the Taskmaster Server and its clients, or the other core Taskmaster Server components; use the Datacap Taskmaster socket protocol. For communicating with the databases, it uses Microsoft Object Linking and Embedding for Database (OLE DB). It also uses the Common Internet File System (CIFS) interface to mount the file share that is required to access batches. Taskmaster Server also uses Active Directory Service Interfaces (ADSI) or LDAP to communicate with the Directory Service for user authentication. (34)

3.2.2 Taskmaster databases

For its operation, a Taskmaster application relies on relational databases. In the Taskmaster sample and add-on applications, Microsoft Access databases are used for portability reasons but must not be used in production. In a production system, Taskmaster databases are hosted in Microsoft SQL Server or Oracle.

The system uses following databases:

- **Admin:** The Administrator database stores information about users, groups, workstation, auditing, functional security, and application configuration. It also stores workflow configurations.
- **Engine:** The Engine database stores information about batches, statistics, and queue states.
- **Fingerprint:** The Fingerprints database manages the pointers to the fingerprints that are used in a given application. Each application has its own set of self-contained databases.
- **External databases** (optional)**:** In many cases, Taskmaster applications need access to external, non-Taskmaster databases during processing. Taskmaster applications can perform lookups (to validate data such as Vendor IDs, Purchase Order Numbers, zip codes, Customer IDs, etc.), and can also export data and images to line of business (LOB) systems and databases. (34)

3.2.3 Taskmaster file server

A file server hosts image files, extracted data and control files, and files that are required for running various applications such as the fingerprint files and document hierarchy

definition files. The file server must be shared across all the components that need to process the batches. (34)

3.2.4 Rulerunner Service

Datacap Rulerunner Service runs as a Windows service and runs batch processing tasks that do not require operator interaction, such as recognition and export. In a typical production environment, Rulerunner is configured to run the page identification and recognition tasks automatically. After verification and submission, Rulerunner detects that the batch is ready for export and runs the export task automatically. (34)

3.2.5 Taskmaster Web server

Taskmaster Web interfaces with the Taskmaster Server to serve web pages and documents to web users and to upload documents that are scanned or imported remotely. It is configured as a virtual directory (or site) in Internet Information Services (IIS) for Windows Server. It handles all communications with the back-end services through the Taskmaster Server. (34)

3.2.6 Taskmaster web client

Taskmaster Web supports browser-based Taskmaster clients. The web client (also known as the thin client) provides similar functionality to the Taskmaster Client but does not require additional software installed on the workstation. When verifying a batch using the web client, verification rules run on the Taskmaster Web machine. Taskmaster Web also lets you configure an application's workflow and perform administrative tasks such as setting up Taskmaster users. The Taskmaster Web installation option includes a restricted version of the Taskmaster Application Manager. (36)

3.2.7 Taskmaster thick client

The Taskmaster Client component is a set of programs that provide user access to Taskmaster applications. The DotScan, DotEdit, and FastDoc user interfaces are Taskmaster software components that run under Windows and provide the ability to run the end-user tasks such as scanning and verification that are part of a software application such as APT. When a user verifies a batch, verification rules run on the Taskmaster Client Workstation. The Taskmaster Client installation option includes DotScan, DotEdit,

FastDoc, Datacap Studio, NENU, the default action libraries, and a restricted version of the Taskmaster Application Manager. (34)

3.2.8 Datacap Studio

Datacap Studio is used to configure the Taskmaster applications, by defining and assembling the document hierarchies, recognition zones and fields, rules, and actions. It requires access to the file server and the Taskmaster databases. (34)

3.2.9 NENU

New Enhanced Notification Utility (NENU) is used to automate system health and housekeeping tasks. Such tasks include batch monitoring, status notification, and automatic deletion of completed batches. Tasks are scheduled by using the Microsoft Windows Scheduler. (34)

3.2.10 RV2

RV2 is a web application that is used to display Taskmaster reports on system activities such as batch status, station activity, or problem batches. (34)

3.2.11 Business applications or databases

Typically, connection to a business application or database is through Open Database Connectivity (ODBC). For example, a customer, vendor, or purchase information can be queried against a database and used for image process verification purposes. In addition, information that is extracted from an image can be exported to business applications and databases. (34)

3.2.12 Lightweight Directory Access Protocol (LDAP) or Active Directory

An LDAP or Active Directory service is also often a part the configuration for Taskmaster users to authenticate. (34)

4 DOCUMENT CAPTURE AS SOFTWARE AS A SERVICE

Software as a Service (SaaS) is one of the cloud computing service models. Cloud computing is a style of computing in which dynamic, scalable and virtual resources are provided over the Internet. Cloud computing refers to services that provide common business applications online, which are accessed from a Web browser, while the software and data are stored on the servers. (37)

Cloud computing specifically refers to incorporating **software as a service** (SaaS), **platform as a service** (PaaS), and **infrastructure as a service** (IaaS). Users do not need to have knowledge of, expertise in, or control over the technology infrastructure in the "cloud" that supports them. (37)

Figure 12 shows three mentioned cloud computing models with application examples provided in each model.

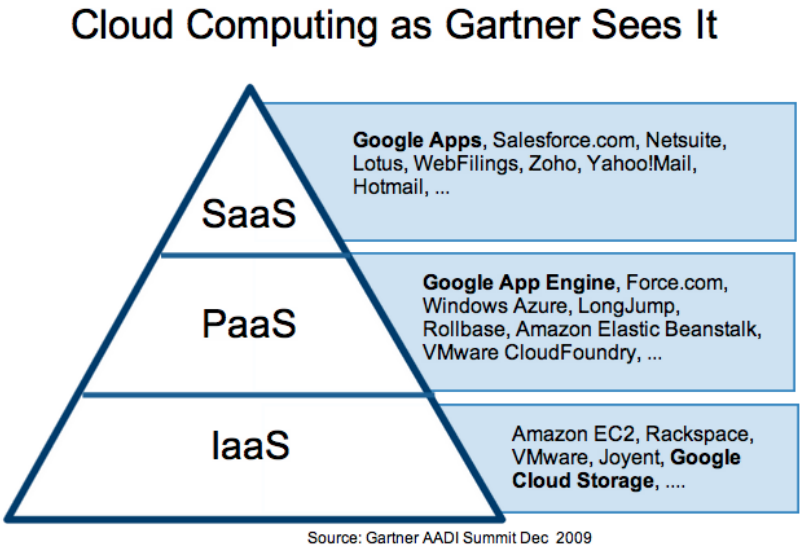

Figure 12 – Cloud computing service levels (38)

4.1 Software as a service

SaaS describes any cloud service where consumers are able to access software applications over the internet. The applications are hosted in “the cloud” and can be used for a wide range of tasks for both individuals and organizations. Google, Twitter, Facebook and Flickr are all examples of SaaS, with users able to access the services via any internet enabled device. Enterprise users are able to use applications for a range of needs, including

accounting and invoicing, tracking sales, planning, performance monitoring and communications.

Example of benefits when using SaaS:

- **No additional hardware costs**; the processing power required to run the applications is supplied by the cloud provider.
- **No initial setup costs**; applications are ready to use once the user subscribes.
- **Pay for what you use**; if a piece of software is only needed for a limited period then it is only paid for over that period and subscriptions can usually be halted at any time.
- **Usage is scalable**; if a user decides they need more storage or additional services, for example, then they can access these on demand without needing to install new software or hardware.
- **Updates are automated;** whenever there is an update it is available online to existing customers, often free of charge. No new software will be required as it often is with other types of applications and the updates will usually be deployed automatically by the cloud provider.
- **Cross device compatibility**; SaaS applications can be accessed via any internet enabled device, which makes it ideal for those who use a number of different devices, such as internet enabled phones and tablets, and those who don't always use the same computer.
- **Accessible from any location;** rather than being restricted to installations on individual computers, an application can be accessed from anywhere with an internet enabled device. (39)

4.2 Platform as a service

PaaS allows users to create software applications using tools supplied by the provider. PaaS services can consist of preconfigured features that customers can subscribe to; they can choose to include the features that meet their requirements while discarding those that do not. Consequently, packages can vary from offering simple point-and-click frameworks where no client side hosting expertise is required to supplying the infrastructure options for advanced development.

Some of the features that can be included with a PaaS offering:

- Operating system
- Server-side scripting environment
- Database management system
- Server Software
- Support
- Storage (40)

  etc.

4.3 Infrastructure as a service

IaaS provides access to computing resource in a virtualized environment, "the Cloud", across a public connection, usually the internet. In the case of IaaS the computing resource provided is specifically that of virtualized hardware, in other words, computing infrastructure. The definition includes such offerings as virtual server space, network connections, bandwidth, IP addresses and load balancers. Physically, the pool of hardware resource is pulled from a multitude of servers and networks usually distributed across numerous data centers, all of which the cloud provider is responsible for maintaining. The client, on the other hand, is given access to the virtualized components in order to build their own IT platforms.

Example of benefits when using IaaS:

- **Scalability**; resource is available as and when the client needs it and, therefore, there are no delays in expanding capacity or the wastage of unused capacity
- **No investment in hardware**; the underlying physical hardware that supports an IaaS service is set up and maintained by the cloud provider, saving the time and cost of doing so on the client side
- **Utility** style costing; the service can be accessed on demand and the client only pays for the resource that they actually use
- **Location independence**; the service can usually be accessed from any location as long as there is an internet connection and the security protocol of the cloud allows it
- **Physical security of data center locations**; services available through a public cloud, or private clouds hosted externally with the cloud provider, benefit from the physical security afforded to the servers which are hosted within a data center

- **No single point of failure**; if one server or network switch, for example, were to fail, the broader service would be unaffected due to the remaining multitude of hardware resources and redundancy configurations. For many services if one entire data center were to go offline, never mind one server, the IaaS service could still run successfully. (41)

4.4 Requirements for Document Capture software provided as SaaS

The requirements for Document Capture software are not other from any application provided as SaaS. All these application must consider following requirements:

1. Security
2. Privacy
3. Data Governance
4. Availability
5. Performance
6. Interoperability
7. Compliance (42)

Figure 13 – Essential Standard Requirements Areas fo SaaS (42)

II. ANALYSIS

5 IBM DATACAP TASKMASTER CAPTURE PROVIDED AS SAAS

Previous chapter describes elementary principles of the cloud computing. This chapter focuses on requirements analysis and technical proposal for IBM Datacap Taskmaster Capture provided as SaaS.

Any software provided as SaaS must consider following requirements:

1. Security
2. Privacy
3. Data Governance
4. Availability
5. Performance
6. Interoperability
7. Compliance

5.1 System architecture proposal

All above mentioned requirements, that are analyzed, are based on the system architecture how the system can be deployed as SaaS. (**Figure 14**)

The architecture is designed with one input a one output channel through HTTPS. The webserver with Taskmaster Web is the only point accessible from the Internet. Export web service is not standard IBM Datacap Taskmaster Capture component, it is custom part that is responsible for exporting data into the client DMS (or any other) system. The web service supports only one way communication in terms of sending data to the service from the Taskmaster Server. It means the service cannot be called from the Internet.

All system components are in internal network and they are hidden behind the internal firewall. Communication between Web Server and other system components are managed by Taskmaster Server through TCP port number 2402 (standard Taskmaster Server communication port). Taskmaster Server is the core of the system manages connection between all system components.

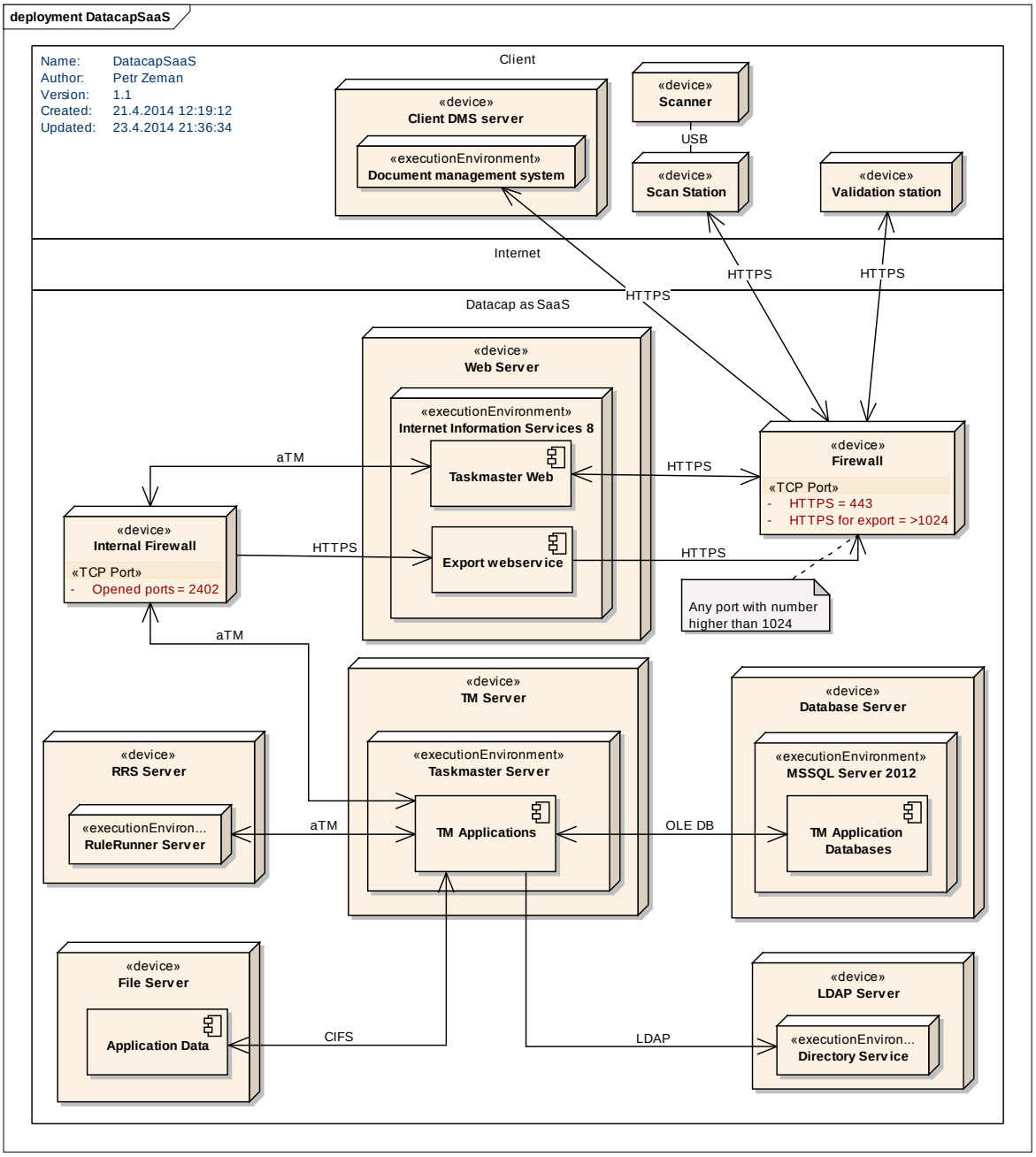

Figure 14 – IBM Datacap Taskmaster Capture SaaS architecture

5.2 Requirements analysis

This sub-chapter provides proposals how the IBM Datacap Taskmaster Capture can be successfully deployed in SaaS model and fulfill the requirements that an imposed in this cloud computing model.

5.2.1 Security and Privacy

SaaS security and privacy has been the number one concern of any application running in this model. While SaaS Privacy depends on SaaS Security, how the secured SaaS data is created, used, and destroyed is dictated by Privacy considerations. Any application must meets several security responsibilities as:

- User authentication and access control
- Protection against unauthorized access
- Physical and personal security
- Malware detection and remediation
- Rights to audit the SaaS vendor's operations and access log information
- External thread prevention

ISO/IEC 27001 standard provides techniques to accomplish above mentioned responsibilities and many other with systematical procedure.

The technical proposal's security can be divided into two parts:

- **Internal** security consists of data storage and communication between system components. System does not provide long-term preservation, it means all data that is stored on the File server is only temporary and is part of the data is deleted in regular cycles. The temporary data is encrypted using AES encryption. IBM Datacap Taskmaster Server uses proprietary encryption for communication between system parts. This encryption meets the FIPS 140-2 Level 4 Federal Information Processing Standard (FIPS).
- **External** security consists of user authentication, protection against external threads and protection against unauthorized access. Every input operation, such as scanning or validation is provided through HTTPS protocol. The system is designed to use LDAP server as a user authentication system.

5.2.2 Data Governance

Data governance represents a convergence of data quality, data management, data policies, business process management, and risk management surrounding the handling of data in the system.

Since this system serves as an input to any client DMS, the data governance lies on processes used in client's organization. Any company will not lose control of data because all data, exported from the system will be stored in their own DMS. All temporary system data are deleted from the system in regular cycles.

5.2.3 Availability

Right after information security, one of the top concerns among SaaS is system availability, or uptime. Availability is a characteristic of a system and is defined in percent.

In general terms, business-critical applications need to be available between 99.9% (8.76 hours/year downtime) to 99.95% (4.38 hours / year downtime).

There are three principles of high availability engineering. They are

1. Elimination of single points of failure. This means adding redundancy to the system so that failure of a component does not mean failure of the entire system.
2. Reliable crossover. In multithreaded systems, the crossover point itself tends to become a single point of failure. High availability engineering must provide for reliable crossover.
3. Detection of failures as they occur.

Above mentioned system architecture proposal does not initially cover this requirement. This requirement can be easily achieved in two additions to the initial technical proposal.

1. Any component of the system can be duplicated. This means the system can consist of several Web Servers, Taskmaster Servers or Rulerunner Servers etc. When one of the server fails, another takes place and ensures the system does not stop working.
2. Geographically different server farms ensures system availability in case of data center failure in one location. When failure occurs, network component ensures redirection to another location.

5.2.4 Performance

See chapter 7.2 Application Performance Testing for application performance testing performed on the system based on the system architecture proposal.

5.2.5 Interoperability

Interoperability between SaaS systems, other Cloud applications, and legacy applications is a major concern of the Enterprise. With SaaS, there is no ability to access the data directly – all access is controlled by the SaaS application. Access to the data may be provided through SOA, or a lighter weight and increasingly more common RESTful API.

System architecture proposal expects two types of interoperability – **export** and **validation**. Export interoperability is simple – if the target system has any standard interface, such as SOA, export web service can communicate with this system. Validation interoperability can be achieved again through the standard communication interface or the data can be loaded into internal database server.

5.2.6 Compliance

Compliance, like security has a higher impact on the SaaS software (and internet) vendor than the on premise software vendors since they have operational responsibility for the system and its data.

Compliance divides into three major areas:

1. **Industry specific compliance** requirements such as HIPPA, FINRA, SOX, and various DoD requirements (U.S.) – these industry standards are unique to each geopolitical area.
2. **Geopolitical compliance** requirements such as the EU Data Protection Directive (Directive 95/46/EC), Korea/Taiwan Personal Information Protection Act (PIPA), India Information Technology Act (ITA) etc.
3. **Enterprise specific** requirements to meet the corporation's internal IT standards such as Single Sign-On (SSO) and Authentication standards.

Compliance issues are related to Security and Privacy requirements, but have very specific governmental requirements.

6 IBM DATACAP APPLICATION DESIGN AND IMPLEMENTATION

6.1 Business Requirements and Application Architecture

Every Datacap application needs to be developed based on the business requirements. This chapter describes example of company that needs to capture supplier invoices and internal delivery notes between company branch offices. This company name is Meebootix spol. s r.o. and all documents are in Czech language.

The following analytical steps are required before starting an application development:

- Business process requirements
- Business process use cases
- Required document structure and fields for each page type
- Permissible field values and business validation rules
- Data export format
- Application process architecture

6.1.1 Business process requirements

Defining the business requirements involves examining the documents that will be processed, determining which fields to capture, and deciding what to do with captured data.

The requirements are divided into seven groups based on the general Taskmaster application architecture (3.1). **Figure 15** displays list of the requirements for each Taskmaster application step. Every step is described in separate diagram and every requirement provides detailed information what is required.

- Page input (R001) – **Figure 16** and **Table 1**
- Page identification (R002) – **Figure 17** and **Table 2**
- Document assembly (R003) – **Figure 18** and **Table 3**
- Data recognition (R004) – **Figure 19** and **Table 4**
- Data validation (R005) – **Figure 20** and **Table 5**
- Data verification (R006) – **Figure 21** and **Table 6**
- Data export (R007) – **Figure 22** and **Table 7**

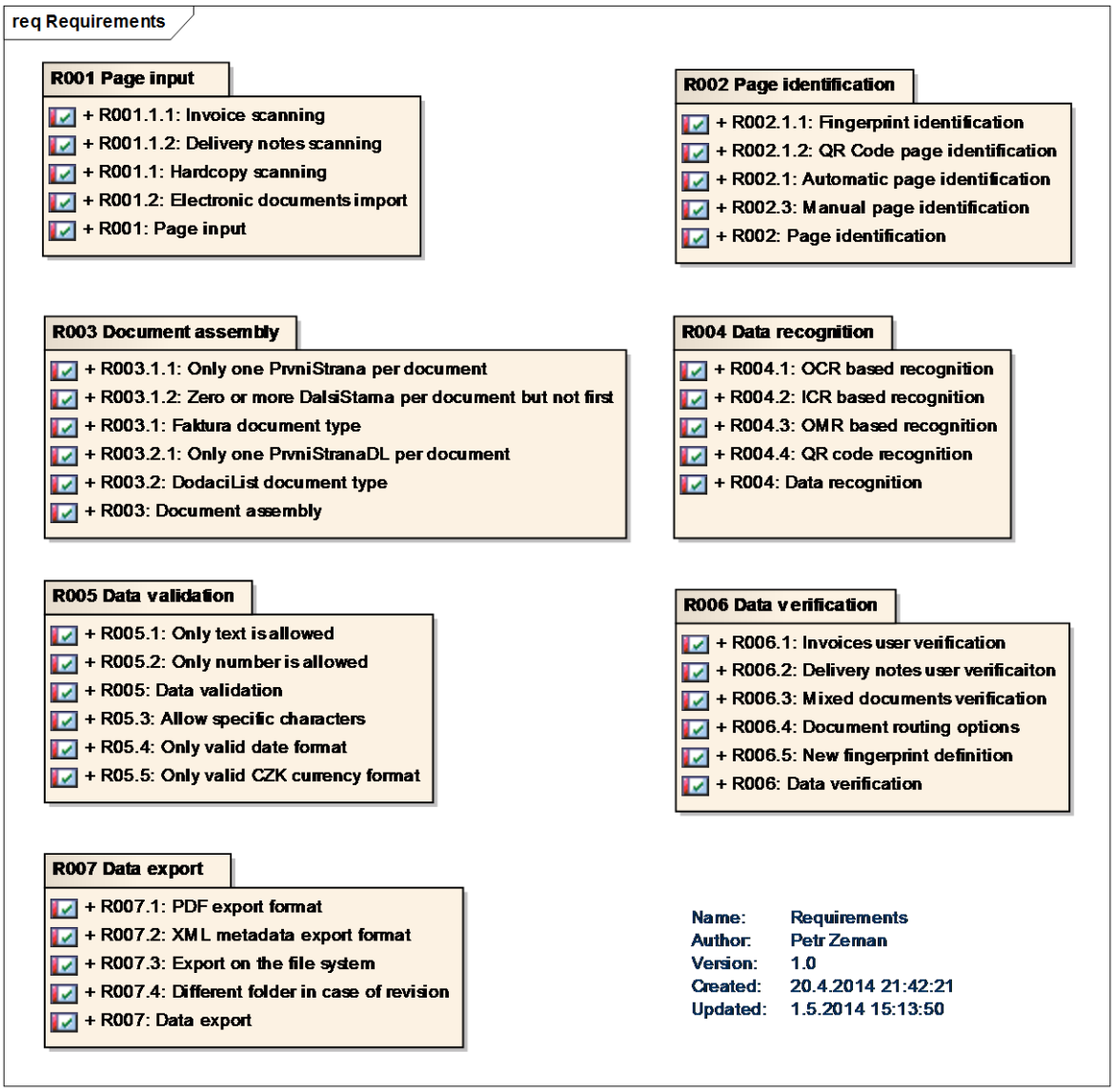

Figure 15 – General requirements diagram for every Taskmaster application's step

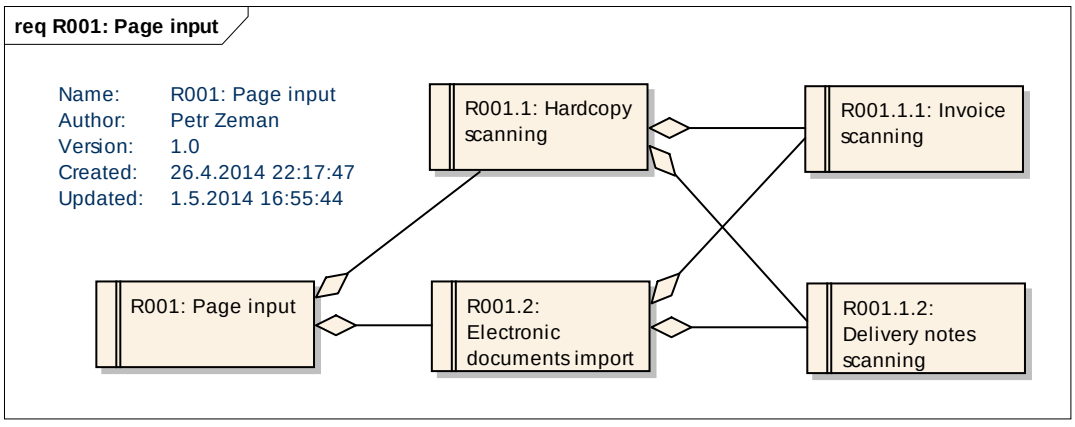

Figure 16 – Page input requirements diagram

Table 1 – Page input requirements diagram description

Element name	Description
R001: Page input	System must be able to scan multiple documents. The documents must be in Czech language
R001.1.1: Invoice scanning	System must accept invoices as input documents
R001.1.2: Delivery notes scanning	System must accept delivery notes as input documents
R001.1: Hardcopy scanning	System must accept input pages through scanner
R001.2: Electronic documents import	System must be able to import pages through the virtual scanning

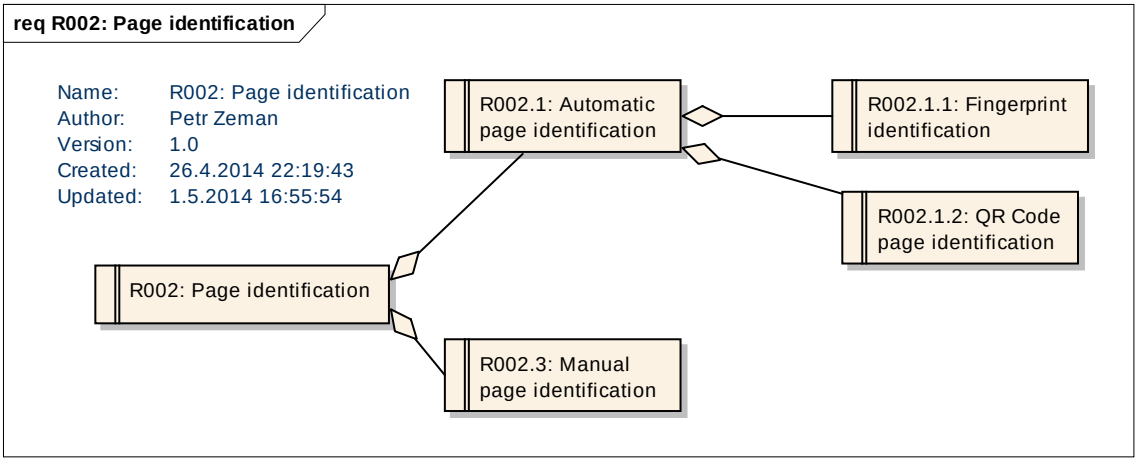

Figure 17 – Page identification requirements diagram

Table 2 – Page identification requirements diagram description

Element name	Description
R002.1: Automatic page identification	Documents must be identified automatically based on the specified technology
R002.3: Manual page identification	Documents can be identified manually in case the fingerprint identification fails. (This means the input page is new type of page not known by the fingerprint engine)

Element name	Description
R002.1.2: QR Code page identification	Delivery notes are identified based on the QR code
R002.1.1: Fingerprint identification	Invoices are identified base on the fingerprint technology

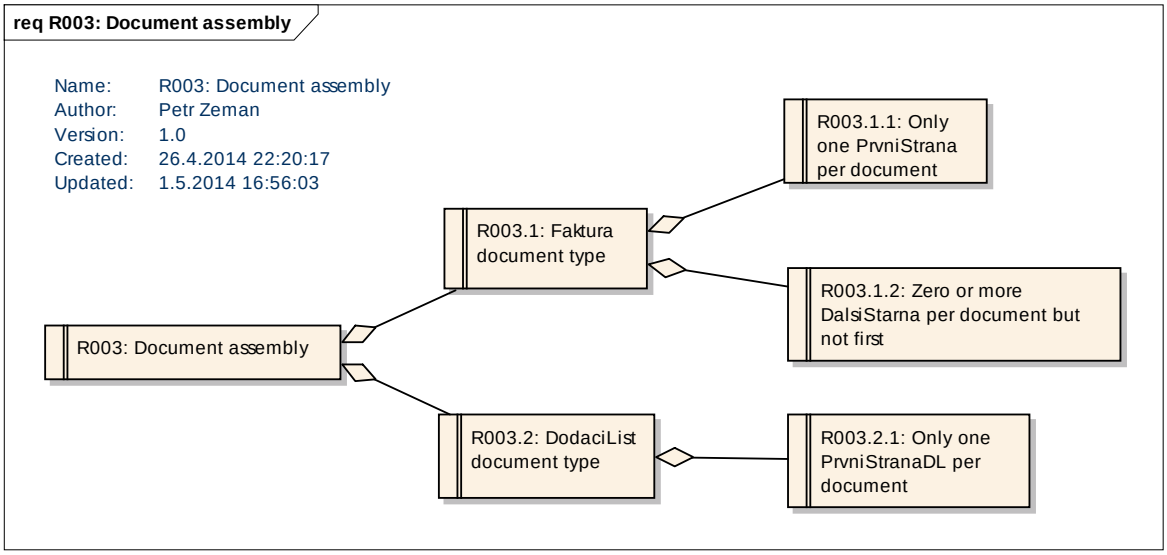

Figure 18 – Document assembly requirements diagram

Table 3 – Document assembly requirements diagram description

Element name	Description
R003.1: Faktura document type	System assembles pages into Faktura document type based on the predefined document structure
R003.2: DodaciList document type	System assembles pages into DodaciList document type based on the predefined document structure
R003.1.1: Only one PrvniStrana per document	Every Faktura must contain only one PrvniStrana page type
R003.1.2: Zero or more DalsiStarna per document but not first	Every Faktura may contain zero or more DalsiStrana page type. This type cannot be first in document.

Element name	Description
R003.2.1: Only one PrvniStranaDL per document	Every DodaciList must contain only one PrvniStranaDL page type

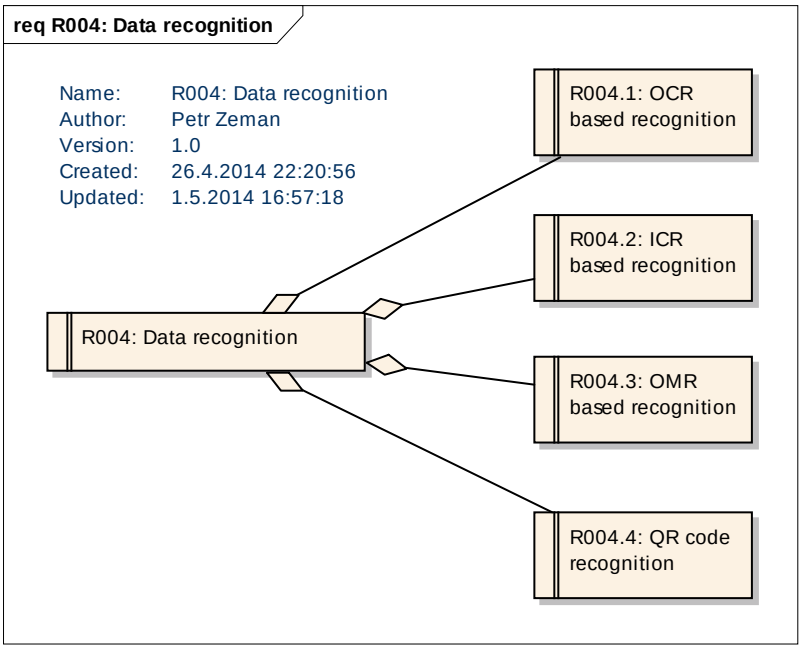

Figure 19 – Data recognition requirements diagram

Table 4 – Data recognition requirements diagram description

Element name	Description
R004.1: OCR based recognition	Application must be able to capture data using OCR technology
R004.2: ICR based recognition	Application must be able to capture data using ICR technology
R004.3: OMR based recognition	Application must be able to capture data using OMR technology
R004.4: QR code recognition	Application must be able to capture data using QR code recognition.

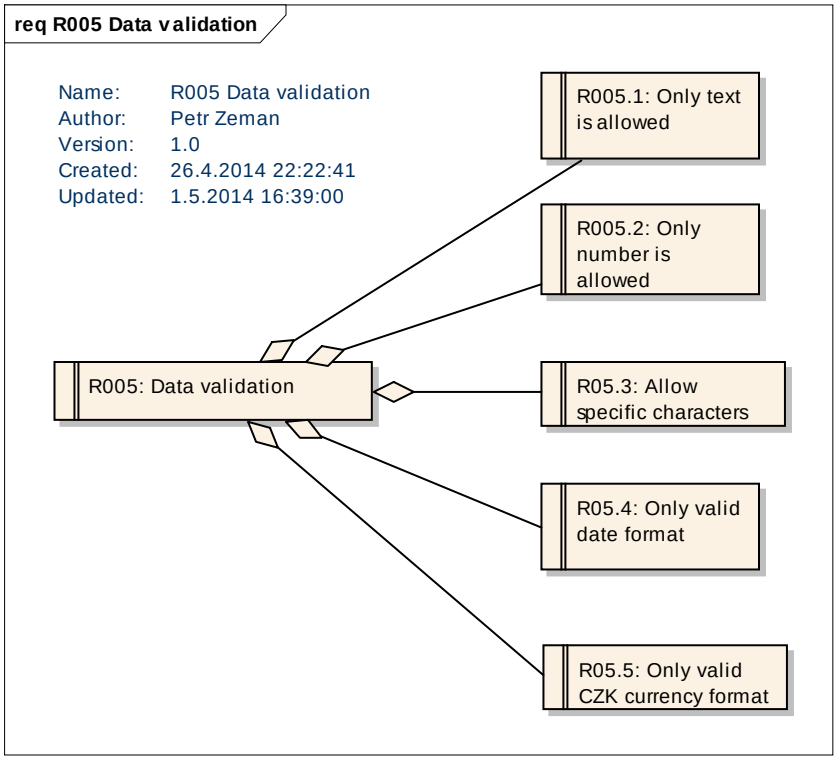

Figure 20 – Data validation requirements diagram

Table 5 – Data validation requirements diagram description

Element name	Description
R005.1: Only text is allowed	Application can accept only text characters for specific fields.
R005.2: Only number is allowed	Application can accept only number characters for specific fields.
R05.3: Allow specific characters	Application can accept only explicitly specified characters for specific fields.
R05.4: Only valid date format	Application can accept only date that are in valid Czech format. All other formats are rejected.
R05.5: Only valid CZK currency format	Application can accept only correct CZK currency format. All other formats are rejected.

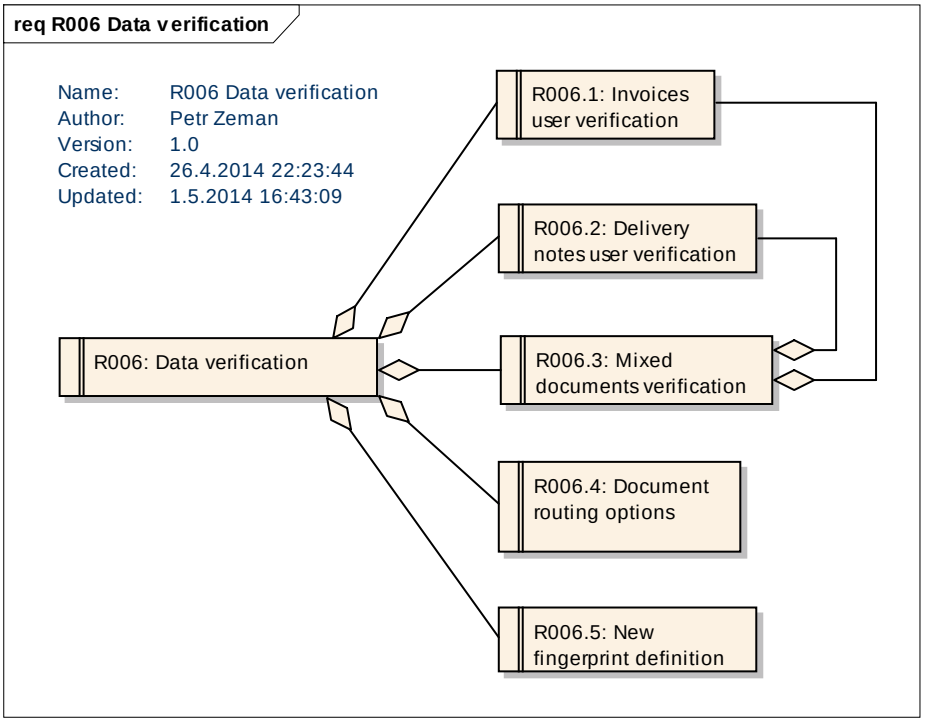

Figure 21 – Data verification requirements diagram

Table 6 – Data verification requirements diagram description

Element name	Description
R006.2: Delivery notes user verification	Validation user can validate delivery notes - DodaciList document type
R006.1: Invoices user verification	Validation user can validate input invoices - Faktura document type
R006.3: Mixed documents verification	Validation user can multiple documents from each document type
R006.4: Document routing options	Validation user can choose routing options for every document. These routing options include these actions: review, delete
R006.5: New fingerprint definition	Validation user can define new fingerprint for invoices

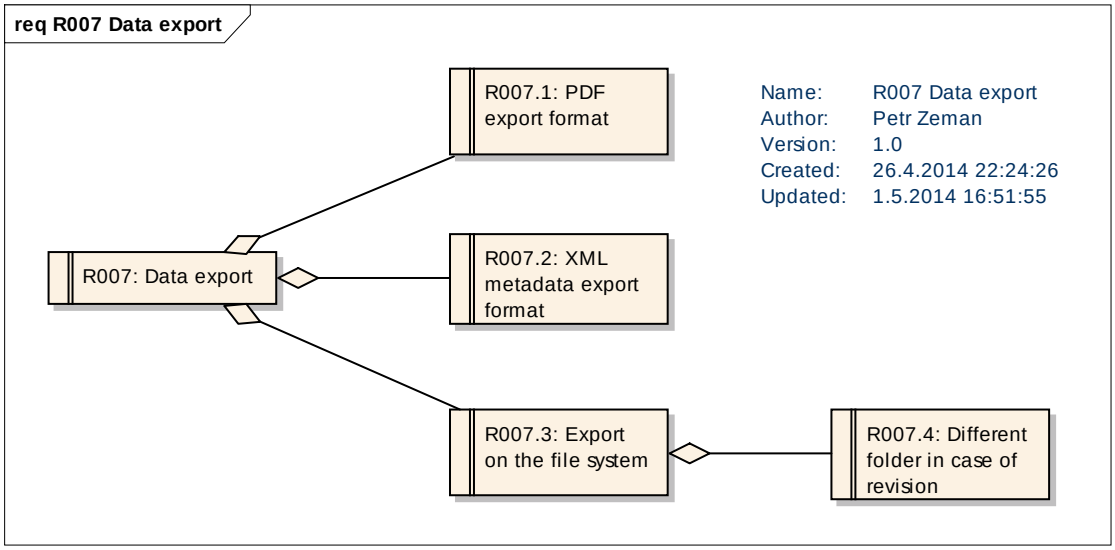

Figure 22 – Data export requirements diagram

Table 7 – Data export requirements diagram description

Element name	Description
R007.1: PDF export format	System must export documents in PDF file format with OCR layer
R007.2: XML metadata export format	System must export captured metadata in XML file format
R007.3: Export on the file system	System must export all data to the file system
R007.4: Different folder in case of revision	System must export data to different folder if document is selected for revision

6.1.2 Business process use cases

Use case is a list of steps that defines interaction between a role and a system, to achieve stated objective. In case of developing business process requirements and use cases I do not consider standard system functionality provided by the IBM Datacap Taskmaster Capture. Use cases are developed directly to accomplish requirements stated in previous chapter (6.1.1). These use cases are application specific, this means different application will have different use case diagram.

Figure 23 shows use case diagram developed for the Meebootix company based on the business requirements. The diagram description is shown in **Table 8**. This diagram presents two types of actors (users):

- **Scan User** – Performs tasks related with document input, document scanning and Manual Page identification in case an invoice is not recognized.
- **Validation User** – Performs tasks related with document validation and document routings. User is also responsible for new fingerprint definition and it is also able to perform Manual Page identification in case an invoice is not recognized.

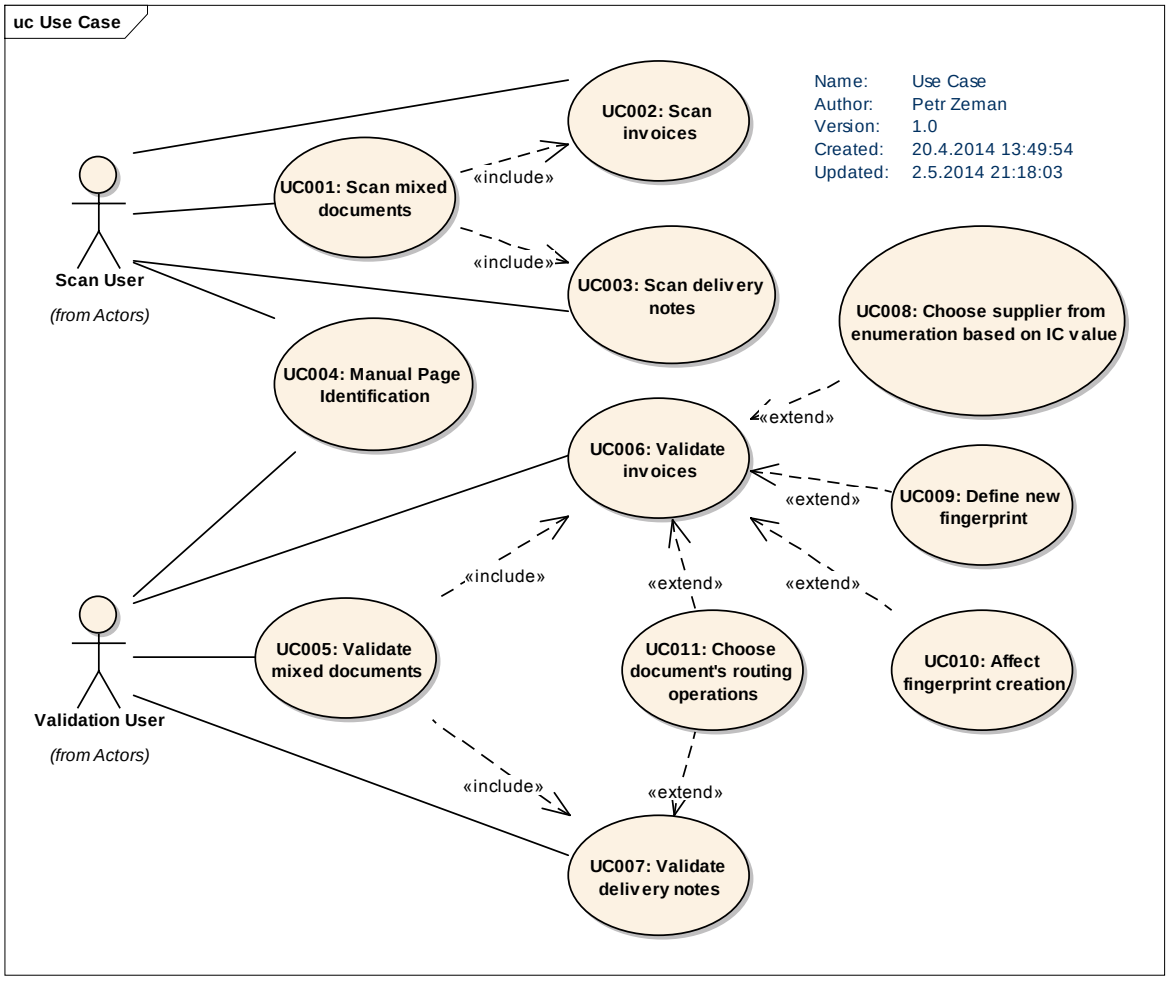

Figure 23 – Use case Taskmaster application diagram

Table 8 – Use case Taskmaster application diagram description

Element name	Description
UC001: Scan mixed documents	User can scan invoices and delivery notes in one batch
UC002: Scan invoices	User is able to scan invoices
UC003: Scan delivery notes	User is able to scan delivery notes
UC004: Manual Page Identification	In case the invoice is not recognized, user is able to identify invoice manually
UC005: Validate mixed documents	User is able to validate invoices and delivery notes in one batch
UC006: Validate invoices	User is able to validate invoices
UC007: Validate delivery notes	User is able to validate delivery notes
UC008: Choose supplier from enumeration based on IC value	User is able to pick supplier name from the enumeration based on IC. If the IC is inserted, the supplier name will be entered automatically.
UC009: Define new fingerprint	User is able to define new fingerprint for invoices. New fingerprint is automatically created if Manual Page Identification is performed or user wants to create new one.
UC010: Affect fingerprint creation	User is able to decide and affect if the fingerprint is created or not
UC011: Choose document's routing operations	User is able to select additional routing options. These options are: Review, Delete or none

Each use case has its own scenario how the use case is performed. These scenarios are placed in Appendix A I. The application functional testing (7.1) is performed based on these scenarios.

6.1.3 Document and page types

Meebootix requires application for processing supplier invoices headers in Czech language and delivery notes between company branch offices, these notes are also in Czech language. The document types and page types are summarized in the **Table 9**. Because this application is for Czech company, names of the all internal objects (documents, pages and fields) are in Czech. **Table 9** also provides English translation of all objects.

Table 9 – Document and page types

Taskmaster names		**English translation**	
Document type	**Page types**	**Document type**	**Page types**
Faktura	PrvniStrana DalsiStrana	Invoice	FirstPage TrailingPage
DodaciList	PrvniStranaDL	DeliveryNote	FirstPageDL

The application has to able to handle structured documents (delivery notes) and semi-structured documents (invoices). See **Figure 25** and **Chyba! Nenalezen zdroj odkazů.** for examples of the documents. Please note the shown invoices (**Figure 25**) are real and therefor they are anonymized.


Dodací list č.: 21423400

Dodavatel:
Meebootix, spol. s r.o.
Šikmá 15
586 01 Jihlava

IČO: 12345678 DIČ: CZ12345678

Příjemce:
Meebootix, spol. s r.o.

Pobočka č.:AX8856908

Datum vystavení: 25. 4. 2014

Informace o dodávce

Celková hmotnost: 55 kg
Rozměry: Kategorie B

27. 4. 2014
Datum předání

Podpis příjemce

Figure 24 – Example of delivery note

Figure 25 – Examples of supplier invoices

6.1.4 Required document structure and fields for each page type

Every IBM Datacap application has its own document structured based on document and page types. The correct document structure is important part of document assembly. Invoices generally have multiple pages, Meebootix delivery notes has only one page. **Table 10** summarizes the structure of each document type. Every *Faktura* document has to begin with only one *PrvniStrana* and can be followed with multiple *DalsiStrana* pages. Because *DodaciList* has only one page, every page creates new document.

Table 10 – Required document structure

Document Type	**Page Type**	**Number**	**Required?**	**Order**
Faktura		Any number per batch	No	Any position within batch
	PrvniStrana	One per document	Yes	Must be first in document
	DalsiStrana	Any number per document	No	Cannot be first in document
DodaciList		Any number per batch	No	Any position within batch
	PrvniStranaDL	One per document	Yes	Must be first in document

The assumption for this application is that input batches contain mixed documents with multiple, consecutive pages that is in the correct order. If the invoice pages are not in correct order, the application will add them to incorrect document.

Every document has several fields that need to be captured based on the business requirements (6.1.1). **Table 11** and **Table 12** summaries Taskmaster field names for every page type. Tables also contain description for each field. *PrvniStrana* and *PrvniStranaDL* also contain system fields (**Table 13**) which are used for affecting subsequent processing.

Table 11 – Fields for PrvniStrana page type

Faktura – PrvniStrana		
Taskmaster field name	**Description**	**Populated from**
Dodavatel	Supplier name	Database
IC	Supplier Taxpayer identification number	OCR
DIC	Supplier VAT identification number	OCR
CisloFaktury	Invoice number	OCR
VariabilniSymbol	Variable symbol	OCR
BankovniUcet	Bank account number	OCR
DatumVystaveni	Invoice date	OCR
DatumSplatnosti	Maturity date	OCR
DUZP	Taxable supply date	OCR
DatumPrijeti	Delivery date	OCR
CastkaCelkemBezDPH	Total price without VAT	OCR
CastkaCelkemSDPH	Total price with VAT	OCR

Table 12 – Fields for PrvniStranaDL page type

DodaciList – PrvniStranaDL		
Taskmaster field name	**Description**	**Populated from**
CisloDodacihoListu	Delivery note number	QR code
DatumVystaveni	Supplier name	OCR
Podpis	Marks if delivery notes is signed or not	OMR
DatumPredani	Delivery date	ICR

Table 13 – System fields

System fields	
Taskmaster field name	**Description**
NovaSablona	Indicates whether new fingerprint for an invoice is created or not. Two possible values ANO/NE (YES/NO).
Operace	Provides operation for verification. Possible values: • Nic (None) – nothing happens • Smazat (Delete) – Marked document is deleted • Revize (Revision) – Marked document is routed to different export path for revision.

Relations between all elements (batch, document types, page types, fields) are shown on Document hierarchy diagram (**Figure 26**). This diagram shows that Faktura or DodaciList cannot exist without a batch. The same rule is applied on page types. Each page types graphically shows related fields and system fields.

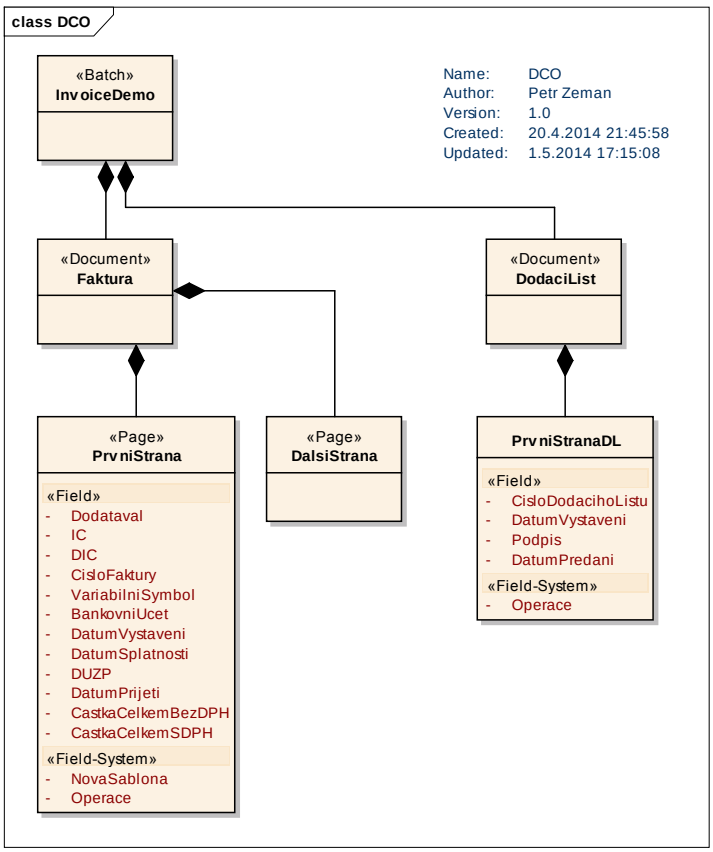

Figure 26 – Document hierarchy

6.1.5 Permissible field values and business validation rules

Every field has limited values that are acceptable. **Table 14** and **Table 15** describes permissible values for every captured field and also provides detail of the validation rules. These rules are based on the business requirements (6.1.1).

Table 14 – Permissible field values and validations for PrvniStrana page type

Faktura – PrvniStrana		
Taskmaster field name	**Permissible values**	**Validation rule**
Dodavatel	Any text	Value has to be populated from list of suppliers stored in the system database. The value is populated based on the IC value.
IC	Any number with maximum length 8 characters	Value cannot be empty
DIC	Any number with maximum length 10 characters with 2 chars prefix: "CZ"	Value has to match regular expression CZ\d{10}
CisloFaktury	Any text	Value cannot be empty
VariabilniSymbol	Any number	-
BankovniUcet	Any number with '/' or '-' delimiter	-
DatumVystaveni	Any valid date format	-
DatumSplatnosti	Any valid date format	-
DUZP	Any valid date format	-
DatumPrijeti	Any valid date	Value is automatically prefilled

Faktura – PrvniStrana		
Taskmaster field name	**Permissible values**	**Validation rule**
	format	with date of scanning the batch.
CastkaCelkemBezDPH	Any valid CZK currency format	Value has to be in valid CZK currency format. Value has to be only numeric.
CastkaCelkemSDPH	Any valid CZK currency format	Value has to be in valid CZK currency format. Value has to be only numeric.

Table 15 – Permissible field values and validations for PrvniStranaDL page type

DodaciList – PrvniStranaDL		
Taskmaster field name	**Permissible values**	**Validation rule**
CisloDodacihoListu	Any number	QR code on each delivery note contains CisloDodacihoListu value.
DatumVystaveni	Any text	-
Podpis	True/False	Value must be true. Otherwise the delivery note is routed to the Revision
DatumPredani	Any valid date format	Value must be same day or later than DatumVystaveni

A validation failure does not necessarily mean that the original page contains invalid data. It might mean that the recognition engine failed to recognize the input data. The validation ensures that all errors are caught and invalid data are correctly handled, corrected and routed.

6.1.6 Data export format

The data will be exported in combination of PDF and XML. PDF file is created for each document and contains all captured images with OCR layer. XML contains all captured field data in structured form. All exported data are stored in predefined folders.

Application exports combination PDF and XML for every document with following naming rules:

- Document type: Faktura – IC_CisloFaktury
  - Example: 123456789_25424215.pdf , 123456789_25424215.xml
- Document type: DodaciList – CisloDodacihoListu
  - Example: 34568745.pdf , 34568745.xml

Documents are exported in following relative folder structure:

Table 16 – Export destination for document types

Relative folder path	Description
..\Faktura	Export path for correct documents from Faktura document type
..\Faktura\Revize	Export path for documents from Faktura document type that are marked for revision
..\DodaciList	Export path for correct documents from DodaciList document type
..\DodaciList\Revize	Export path for documents from DodaciList document type that are marked for revision

Figure 27 and **Figure 28** shows XSD definition for XML exported from the application. All exported XML files will match these XSD definitions based on the current document type.

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="data">
    <xs:complexType>
      <xs:sequence>
        <xs:element type="xs:string" name="dodavatel"/>
        <xs:element type="xs:int" name="ic"/>
        <xs:element type="xs:string" name="dic"/>
        <xs:element type="xs:int" name="cisloFaktury"/>
        <xs:element type="xs:string" name="variabilniSymbol"/>
        <xs:element type="xs:string" name="bankovniUcet"/>
        <xs:element type="xs:string" name="datumPrijeti"/>
        <xs:element type="xs:string" name="datumVystaveni"/>
        <xs:element type="xs:string" name="datumSplatnosti"/>
        <xs:element type="xs:string" name="datumZdanitelnehoPlneni"/>
        <xs:element type="xs:string" name="castkaCelkemBezDPH"/>
        <xs:element type="xs:string" name="castkaCelkemSDPH"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Figure 27 – XSD definition for Faktura export

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="data">
    <xs:complexType>
      <xs:sequence>
        <xs:element type="xs:int" name="cisloDodaciholistu"/>
        <xs:element type="xs:string" name="datumVystaveni"/>
        <xs:element type="xs:boolean" name="podpis"/>
        <xs:element type="xs:string" name="datumPredani"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Figure 28 – XSD definition for DodaciList export

Please note that this is only example scenario and shows possibilities of the system. The documents will be directly exported in DMS or ERP system in real production environment.

6.1.7 Application process architecture

The last analytical step is to design process Taskmaster application workflow. This workflow is shown in **Figure 29** and it encapsulates all previously designed functionality.

The workflow steps are:

- **Scan** – This step contains actions of hardcopy and electronic document scanning performed by the user.
- **PageID** – Page identification is automatic task performed by the system. Pages are identified based on the defined rules. Invoices are identified using fingerprinting and delivery notes are identified using QR code recognition.
    - **ManualPageID** – In case the invoices are not automatically recognized, user has to manually identify first and trailing pages of all unidentified invoices.
- **Profiler** – Automatic task performed by the system is responsible for assembling the documents, data recognition and server based captured data clean and validation. This task also determine if the document needs to be identified or can be directly exported.
- **Verify** – Low-confidence data and fields that fail validation are shown to the user for verification, correction and exception handling.

- **Export** – Last step of the workflow exports data in defined format to the file system. This step is fully automated by the system.

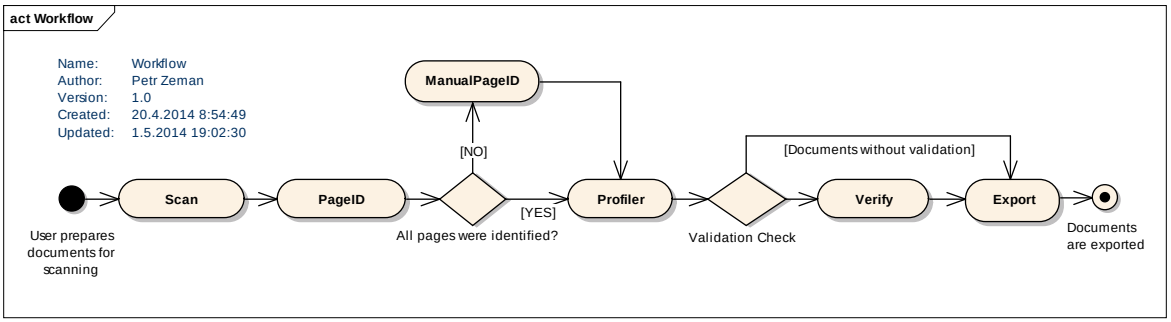

Figure 29 – Taskmaster application workflow

6.2 Application Implementation and User Interface

This chapter is based on the application architecture that is presented in the chapter 6.1. The application process architecture (6.1.7) presented the application workflow. Following sub-chapters describes every workflow task.

6.2.1 Scan

The Scan is first step of the application workflow. This part involves the user interaction. In case of virtual scanning (scanning electronic documents) this part can be fully automatized. **Figure 30** shows the implementation of the Scan in the Meebootix application.

Scan user prepares document for scanning, this action involves following actions:

- Removing all the staples or paper clips that connects document pages together
- Check if all papers are not damaged (torn, crumpled a lot). Repair the page if it is severely damaged.
- Ensure that document page order is correct.

After the documents are ready for scanning, user logs into the system and initialized the scanning client (see **Figure 31** for scan client user interface). System in this part creates a new batch and waits for pages to be scanned. User inserts prepared pages into the scanner and starts the scanning. After the scanning is completed, user is able to verify if all the pages were scanned correctly. If not, user is able to rescan the wrong pages. When the batch is correct, user submits all the documents into the system.

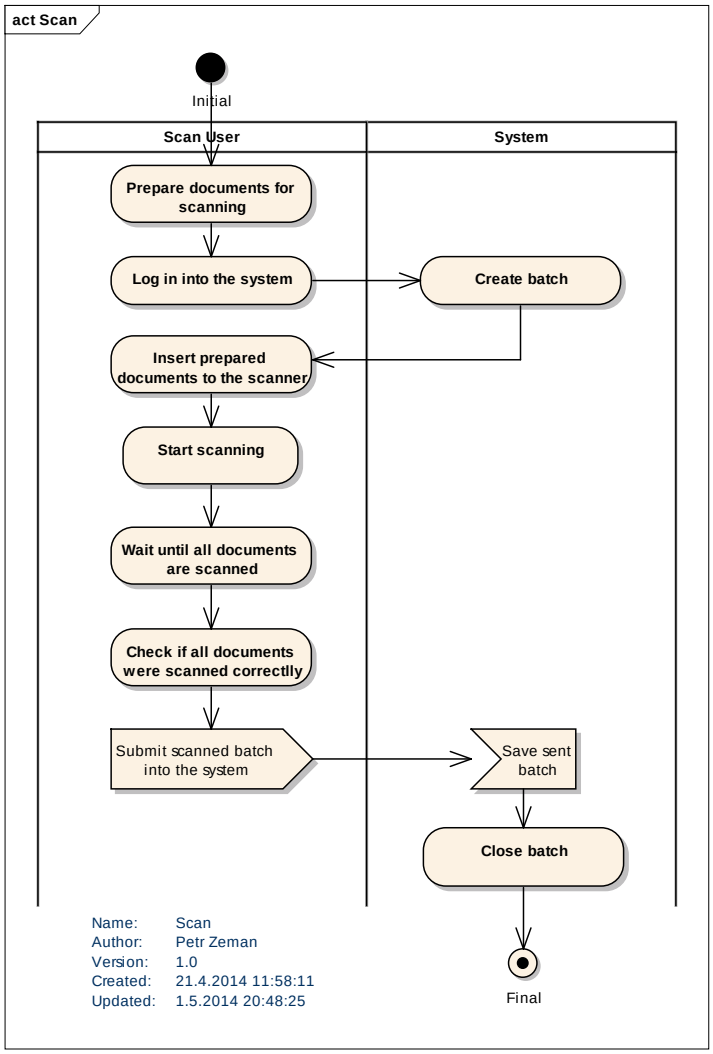

Figure 30 – Scan activity diagram

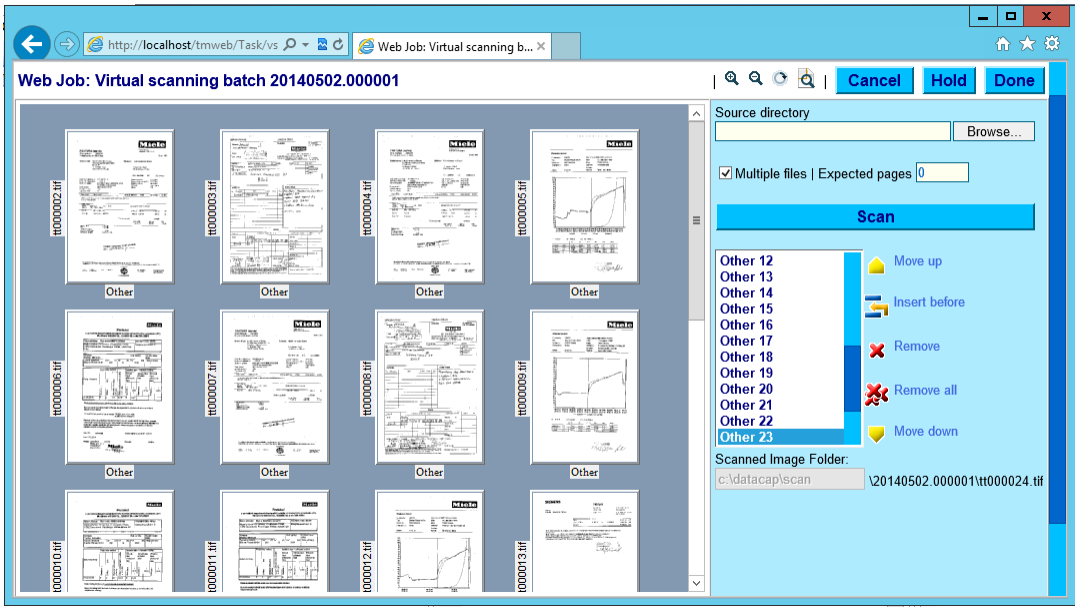

Figure 31 – Scan user interface

6.2.2 PageID

Page identification is the second step in the application workflow. PageID is responsible for correct page identification, image enhancement and blank page deletion. It is a fully automated process that works sequentially with the batch pages. **Figure 32** displays activity diagram for PageID process.

When the batch is opened, the page is picked by the system. To ensure the maximum task effectivity, the first action performed by the PageID is blank page detection and deletion. This ensures the following actions are not slowed down because of running the task on the blank pages.

In the context of the application requirements the application first reads all the QR codes located on the page. Pages that contains the right (with required value) QR code are marked and the value of the QR code is saved in the page variables. Next the image enhancement is performed, this task performs actions such as page rotation, deskew, line removal, character repair etc. The task is executed after the QR code recognition because it can damage or even deletes the QR code from the page. After the image enhancement, all the pages that contains the right QR code are automatically recognized as the *PrvniStranaDL* page type.

If the page does not contain the correct QR code, then it is possible this page is either the *PrvniStrana* or *DalsiStrana* page type. For these pages the fingerprint identification is performed, if the identification is successful, the system sets the page type based on the fingerprint settings. If the page is not identified, then the page is marked for the ManualPageID (6.2.3). The above described process is done for each page within the batch.

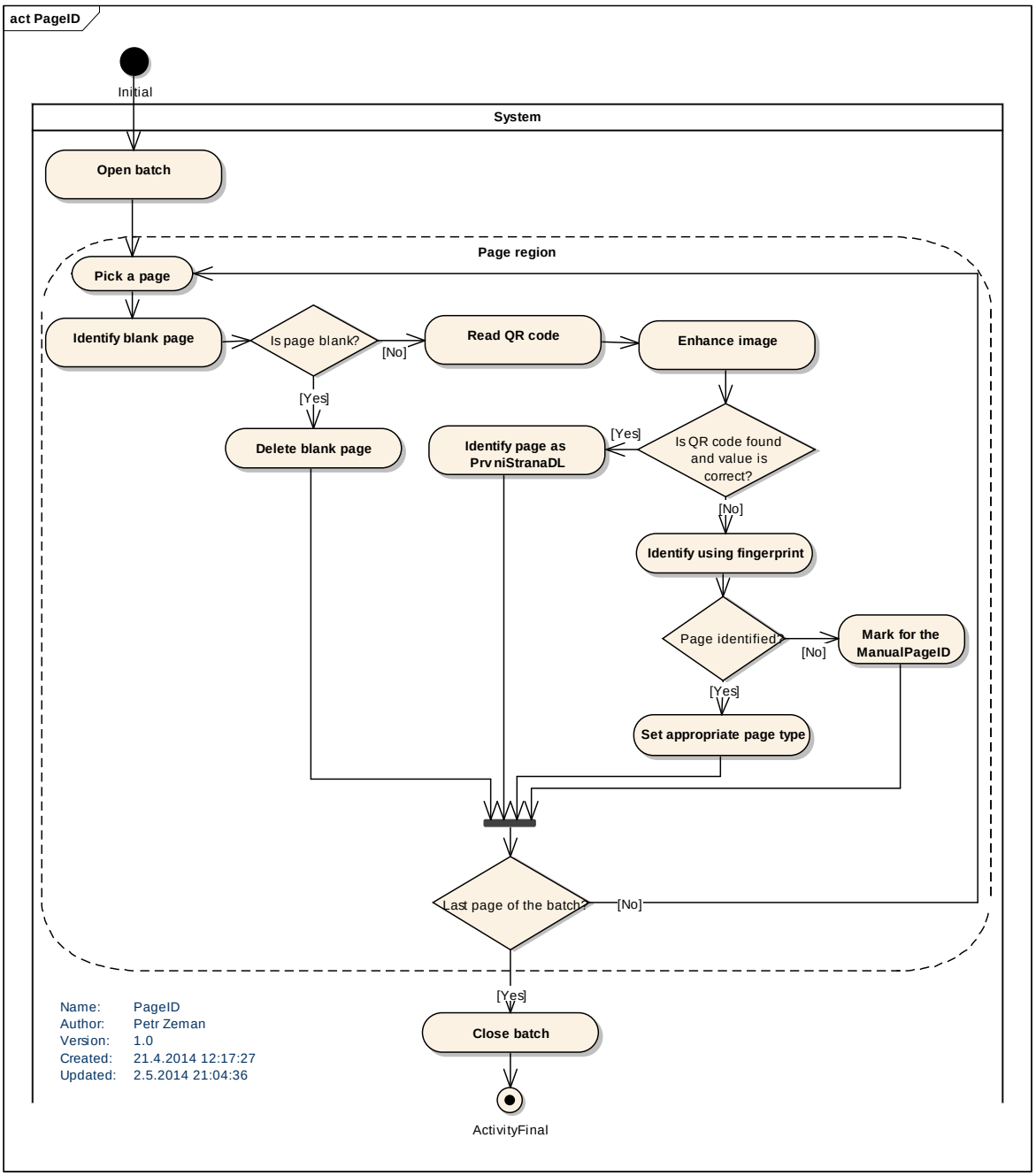

Figure 32 – PageID activity diagram

6.2.3 ManualPageID

Manual page identification is an optional task performed only if the batch contains page (or pages) that was not correctly identified. This task is invoked when fingerprint identification fails and application needs a user interaction. **Figure 33** shows the ManualPageID activity diagram.

User needs to log into the system and starts the ManualPageID client. The system opens the batch and presents all the batch pages to the user. The identified pages have prefilled page types, unidentified pages are consider as the *Other* page type and also highlighted for faster manual page identification. See **Figure 34** for the ManualPageID user interface. When the all pages are identified, user ends the task and the system closes the batch and sends it to the Profiler task (6.2.4).

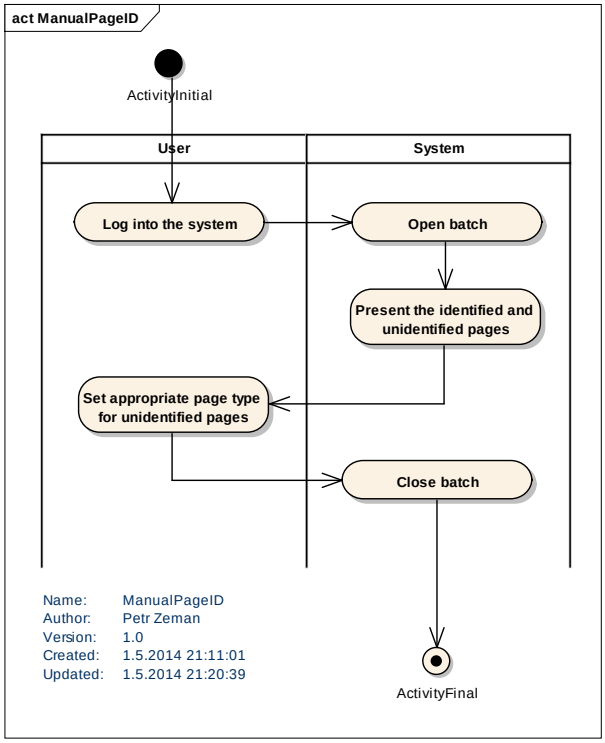

Figure 33 – ManualPageID activity diagram

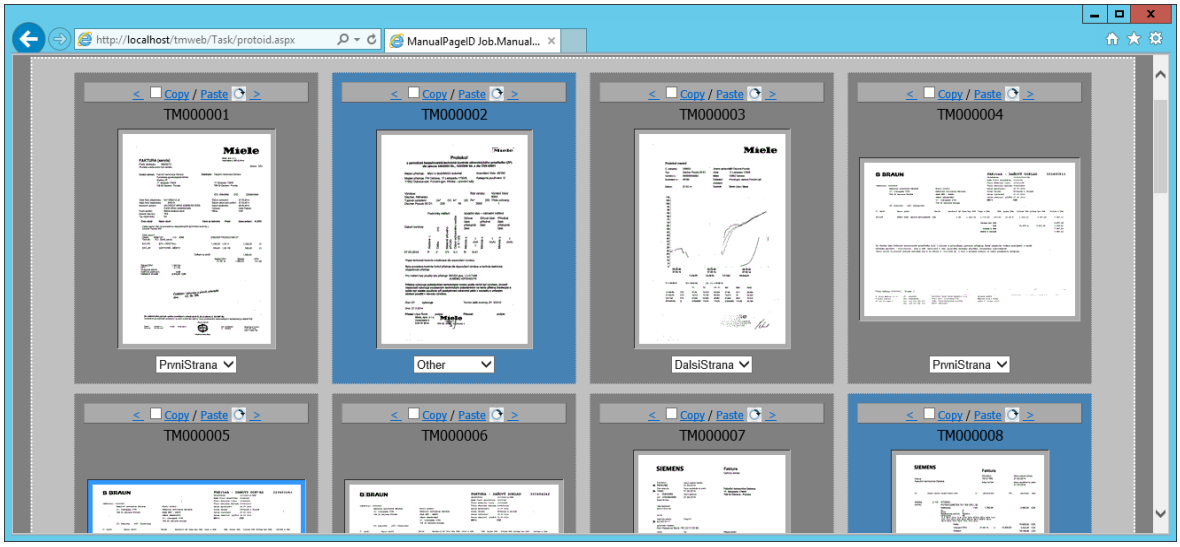

Figure 34 – ManualPageID user interface

6.2.4 Profiler

Profiler task is responsible for the data recognition, captured values clean and server based validation. The task is fully automated and the individual actions are performed based on the business application architecture (6.1). **Figure 35** shows the Profiler activity diagram.

The system opens the batch, assembles identified pages into the documents and create fields for each page type within the document. Next steps are performed for each document, system picks the document and determine next actions based on the document type.

- ***DodaciList*** – Because this document type is structured the system draws the zones for the data capture to the known locations. Although *DatumVystaveni* uses different algorithm, this field is recognized using the text matching method and it can be found anywhere on the page. When the zones are placed, the recognition engine recognized desired data – *DatumPredani* using the ICR engine, *Podpis* using the OMR engine.
- ***Faktura*** – The date recognition differs on the type of page identification. If any of the document pages were successfully identified using fingerprint technology, the fingerprints provides location (zones) of the data to be captured. When these zones are loaded the OCR recognition engine captures all the data. If the document pages were not successfully identified using fingerprints the data are located using the regular expression location method. This method uses lists of predefined values that could be considered as valid input data. For example: If the system is looking for the *IC* value, it will try to find the "IČ" or "IČO" word on the page. If the word is found it will look in all direction around these words if any matches the target value.

The next steps are the same for both document types. The captured data are cleaned based on the defined rules (**Table 14** and **Table 15**) and also the data is subjected to the server validation (**Table 14** and **Table 15**). If document succeeds the server validation, it is marked for No Verify. Documents marked for No Verify are automatically moved to the export without the user verification. This process are performed on all document within the batch.

Before the batch moves to the Verify task, documents that were marked for No Verify are split from the batch (creates a new batch) and directly moved to the Export task.

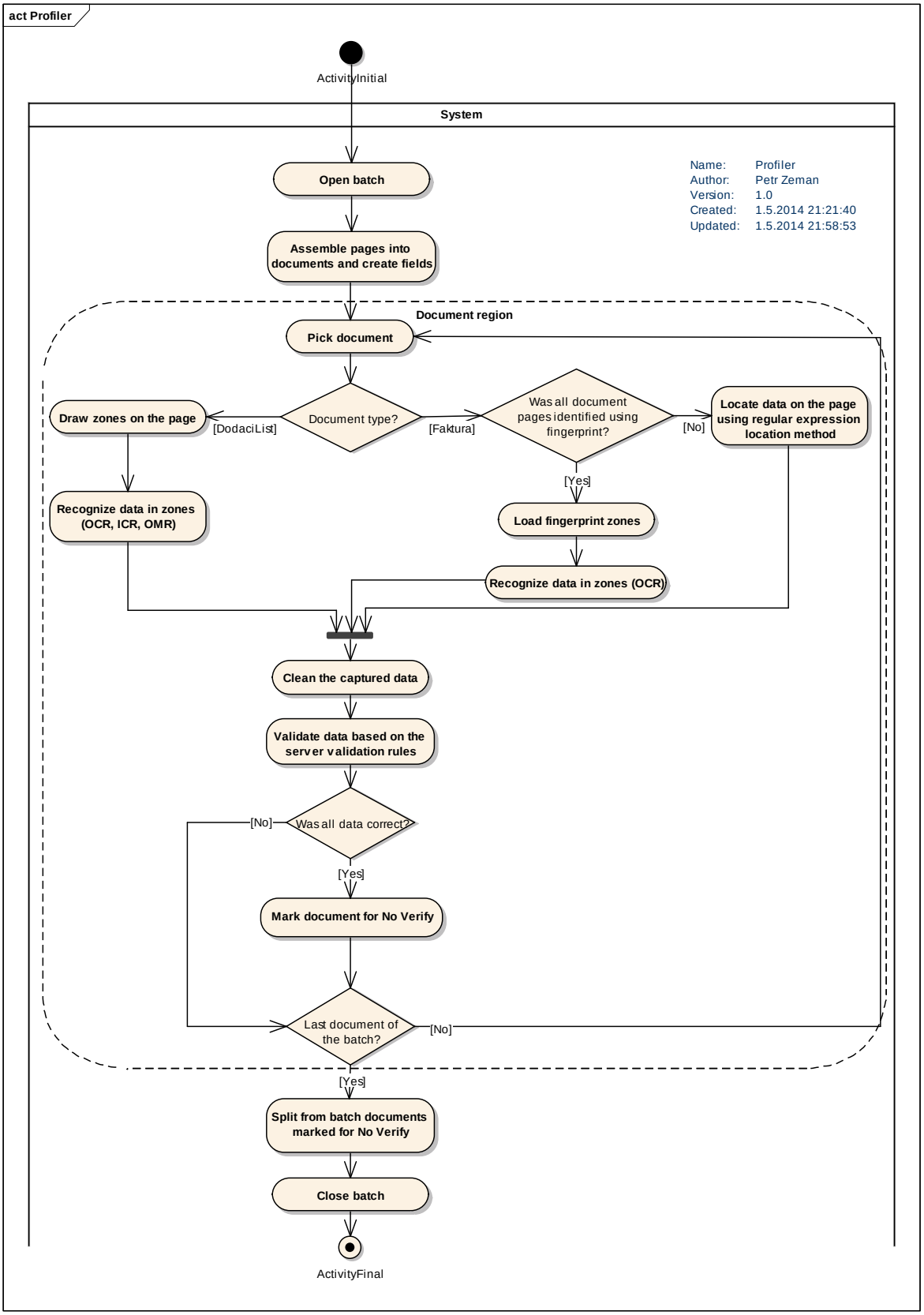

Figure 35 – Profiler activity diagram

6.2.5 Verify

The Verify is the last task that needs the user interaction. **Figure 36** shows the Verify activity diagram. User need to log into the system and run the verify client. The system opens the batch and presents: batch structure, fields for the selected document and page preview for selected page. See **Figure 37** for the Verify user interface. User needs to validate all the documents within the batch, if there is something wrong with the document (bad scan quality, business based problem etc.) user can mark the document for additional processing – delete the document or mark the document for revision. When all documents are verified the system closes the batch and moves it to the Export task.

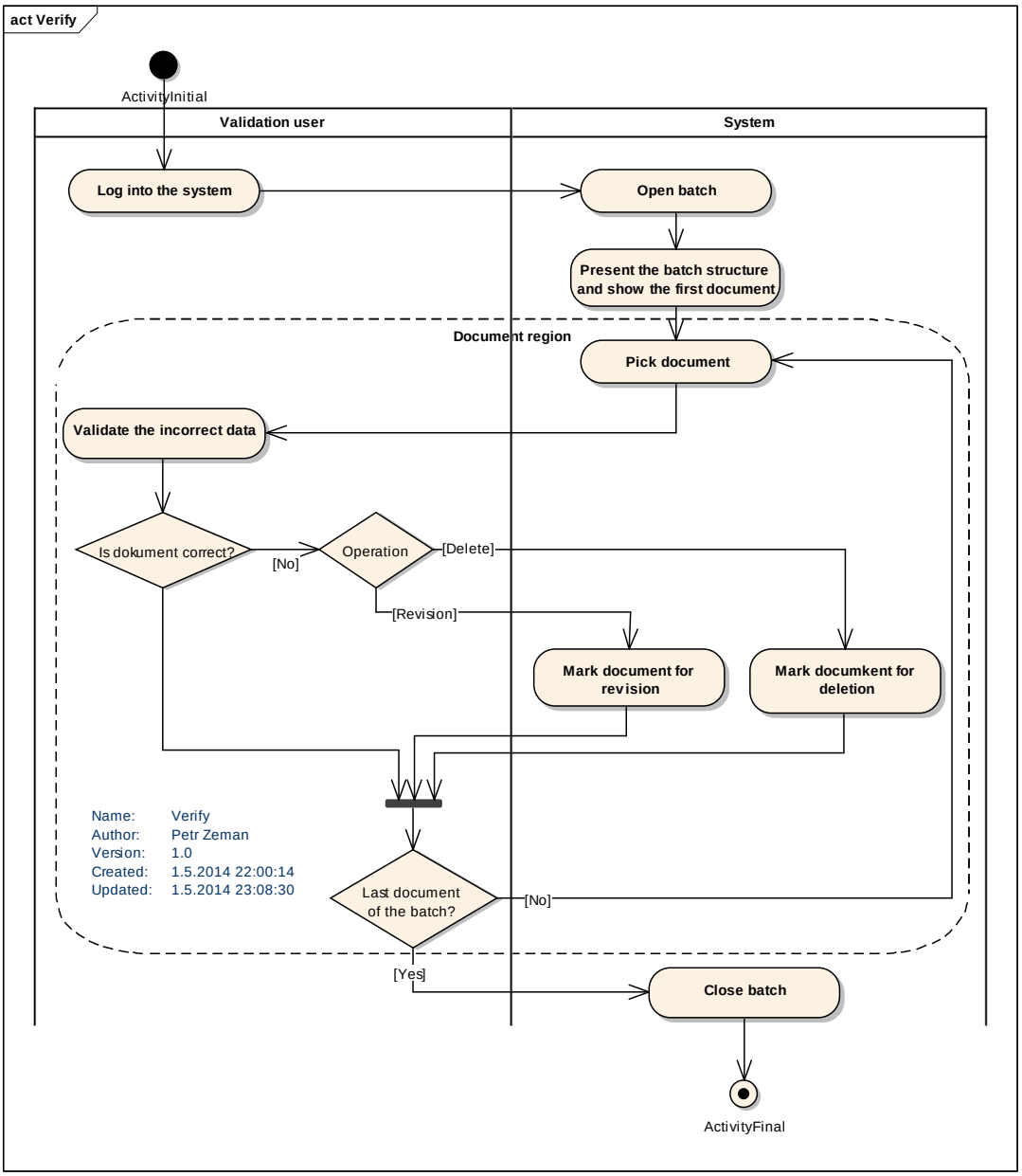

Figure 36 – Verify activity diagram

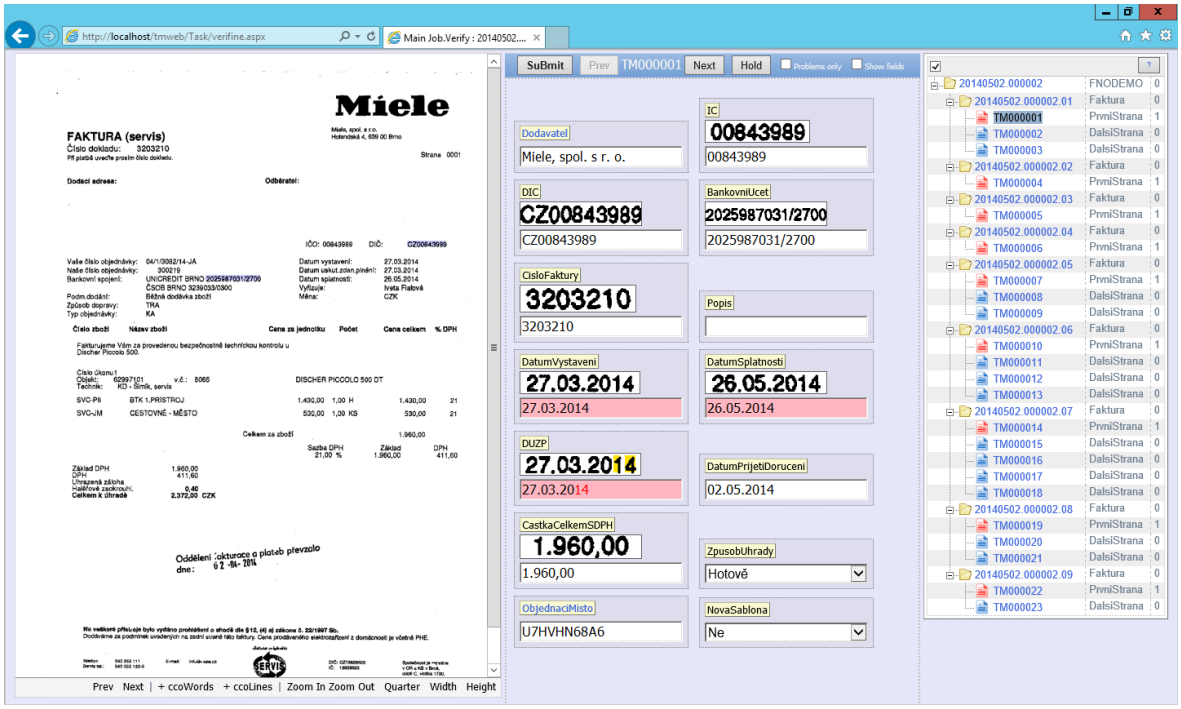

Figure 37 – Verify user interface

6.2.6 Export

Export is the last task within the workflow. **Figure 38** shows the Export activity diagram. The batches are processed from the Verify task or directly from the Profiler task if the documents meets the server validation requirements. Export task works with the documents, when system opens the batch and picks the documents, the PDF file is generated. This file contains all pages that were in the document. After the PDF is completed, the XML file is generated. This XML is created based on the XSD specification (**Figure 27** and **Figure 28**) for each document type.

If the document type is *DodaciList* the files are exported to the export location (:

Table 16). The *Faktura* document type is also exported to the export location. But if the document is marked as New Fingerprint the new fingerprints are generated for every marked page. The pages that were manually identified are marked for New Fingerprint. Also the validation user can determine if the document needs to be considered as a New Fingerprint, then all pages in the document are marked as a New Fingerprint.

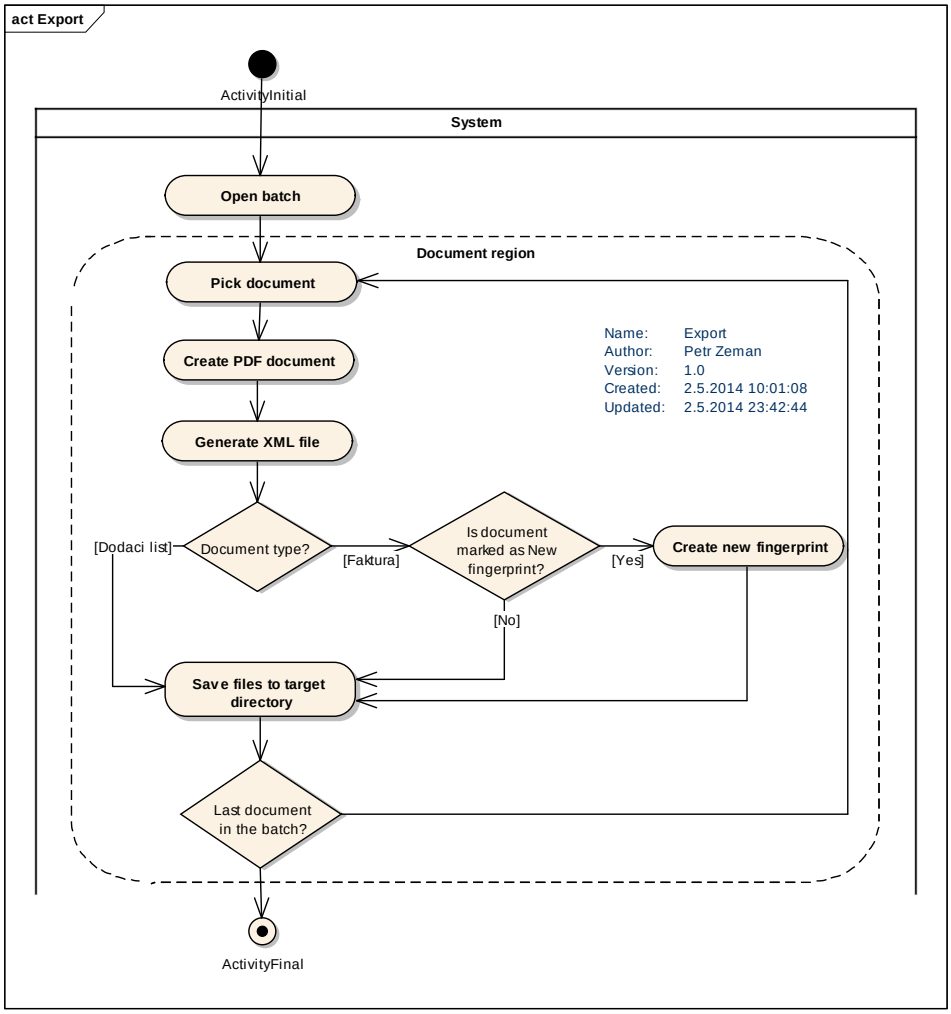

Figure 38 – Export activity diagram

7 IBM DATACAP APPLICATION TESTING

After the application is designed and implemented the final step is to test the application. Software testing provides several methods how to test the software and what to test. In case of the IBM Datacap application testing I consider the following tests as the most important:

- **Application functional testing** – provides a proof that all established requirements were accomplished.
- **Application performance testing** – provides valuable information if the application is more effective and faster than the previous company process.

7.1 Application Functional Testing

The IBM Datacap application functional testing is based on the Use case testing. The use cases were defined in the chapter 6.1.2. Every use case has its own scenario how the use case can be accomplished. These scenarios are provided in the Appendix A I.

Scenario structure:

- **Main success scenario** – Describes steps that needs to be done in order to prove the use case is completed.
- **Extensions** – Describes steps that extends the main success scenario steps in case of the failure of the current step.
- **Step** – Number of the current scenario step. The steps are ordered.
- **Actor** – U (user), S (system)
- **Description** – Describes step

Each use case was tested on the designed application and the results are presented in the tables below. Every table contains information about tested use cases:

- **Test case** – The sequence of steps from the test scenario
- **Notes** – Application functionality notes identified during the testing.
- **Test status** – Test case status – possible values: FAILED or PASSED.
- **Overall test status** – Overall test status for the use case based on the test cases.

Use cases and testing tables:

- UC001, UC002, UC003 – **Table 17**
- UC004 – **Table 18**
- UC005 – **Table 19**
- UC006, UC007 – **Table 20**
- UC008 – **Table 21**
- UC009 – **Table 22**
- UC010 – **Table 23**
- UC011 – **Table 24**

Table 17 – UC001, UC002, UC003 functional testing details

UC001: Scan mixed documents **UC002: Scan invoices** **UC003: Scan delivery notes**		
Test case	**Notes**	**Test status**
1-2-2a	User can enter incorrect password infinite times	PASSED
1-2-3-4-5-6-7-7a	Simulated scanner malfunction – disconnected USB	PASSED
1-2-3-4-5-6-7-8-9-9a-9b-9c-7-8-9-10-11	-	PASSED
1-2-3-4-5-6-7-8-9-10-11	-	PASSED
Overall test status		**PASSED**

Table 18 – UC004 functional testing details

UC004: Manual Page Identification		
Test case	**Notes**	**Test status**
1-2-2a	User can enter incorrect password infinite times	PASSED
1-2-3-4-5-6-7	-	PASSED
Overall test status		**PASSED**

Table 19 – UC005 functional testing details

UC005: Validate mixed documents		
Test case	**Notes**	**Test status**
1-2-2a	User can enter incorrect password infinite times	PASSED
1-2-3-4-5-6-7	-	PASSED
Overall test status		**PASSED**

Table 20 – UC006, UC007 functional testing details

UC006: Validate invoices **UC007: Validate delivery notes**		
Test case	**Notes**	**Test status**
1-2-2a	User can enter incorrect password infinite times	PASSED
1-2-3-4-5-6-7-8-9	Missing information what is wrong with the data. Only provided information is background color	PASSED
1-2-3-4-5-5a-6-7-8-9	Missing information what is wrong with the data. Only provided information is background color	PASSED
Overall test status		**PASSED**

Table 21 – UC008 functional testing details

UC008: Choose supplier from enumeration based on IC value		
Test case	**Notes**	**Test status**
1-1a-1b	-	PASSED
1-1a-2-3-4	-	PASSED
1-1a-2-3-3a		PASSED
1-1a-2-3-3a-3b-3c	Supplier list does not provide additional data filtering.	PASSED
1-1a-2-3-3a-3b-3c-3d	Supplier list does not provide ability to add not existing supplier to the list.	PASSED
Overall test status		**PASSED**

Table 22 - UC009 functional testing details

UC009: Define new fingerprint		
Test case	**Notes**	**Test status**
1-2-3-4-5	-	PASSED
Overall test status		**PASSED**

Table 23 - UC010 functional testing details

UC010: Affect fingerprint creation		
Test case	**Notes**	**Test status**
1-2-2a-3-4-5-6-7	-	PASSED
1-2-3-4-5-6-7	-	PASSED
Overall test status		**PASSED**

Table 24 - UC011 functional testing details

UC011: Choose document's routing operations		
Test case	**Notes**	**Test status**
1-2-3	Missing ability to mark the document for rescan.	PASSED
Overall test status		**PASSED**

7.2 Application Performance Testing

The application performance testing is performed on the system parts that are fully or partly automated (Scan). The user performance tests are not included in this testing.

The solution designed in previous chapter was deployed on the system architecture that is based on the reference IBM Datacap Taskmaster Capture SaaS architecture (**Figure 14**) and all tests are performed on this system. The whole solution runs in the fully virtualized environment. Every test measures the performance on three different input resolutions (200 DPI, 300 DPI and 400 DPI).

7.2.1 Scan

The Scan test measures the duration of the hardcopy page input. Tests are performed in two scanning modes – Simplex (one side scanning) and Duplex (both side scanning). Fujitsu fi-6130 is used as the testing scanner. The test results contains values only for the system operation (scanning, importing pages to the system etc.), user interaction is not included (inserting documents into the scanner, checking whether the documents were scanned correctly etc.).

Table 25 summarizes test results for both scanning modes. **Figure 39** provides the graphical representation of the results. Testing was performed on the 20 batches and each batch contained 50 pages. The batches were scanned in black and white color mode.

Table 25 – Scan performance test results

	Simplex			Duplex		
	Mean scan time [s]	Variance [s]	Standard deviation [s]	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	1,753	0,002	0,0554	1,837	0,002	0,0549
300 DPI	2,267	0,002	0,0516	2,363	0,002	0,0526
400 DPI	4,733	0,003	0,0499	4,913	0,003	0,0501

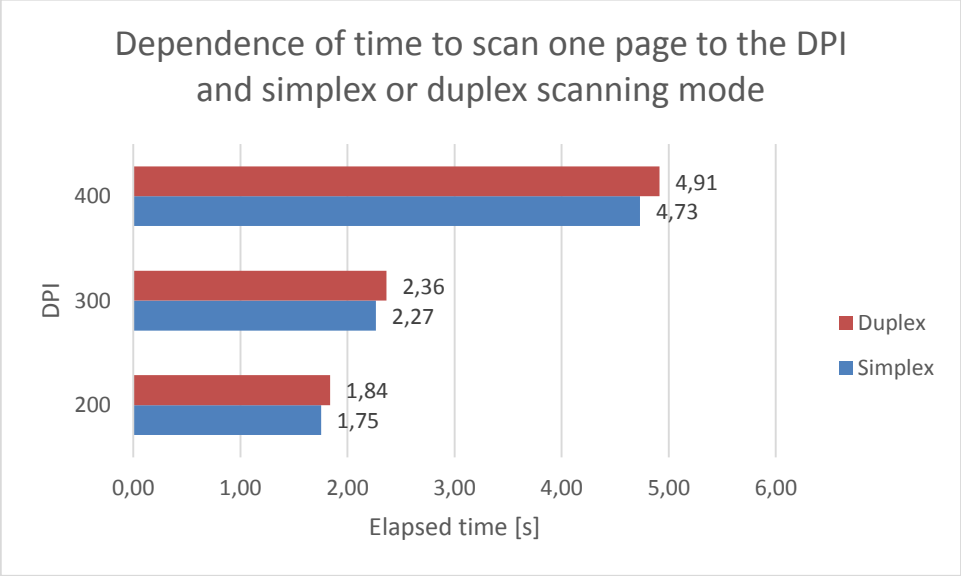

Figure 39 – Scan performance test graph

Another part of the Scan testing is the average file size of the page image file. The images are saved into the TIFF file format with CCIT Group 4 compression. **Table 26** summarizes the average page file size based on the scanned image resolution.

Table 26 – Average page file size based on the resolution

	Mean file size per page [KB]	**Variance [KB]**	**Standard deviation [KB]**
200 DPI	30,8	16,56	4,07
300 DPI	44,6	28,64	5,32
400 DPI	73	46,8	6,84

7.2.2 PageID

PageID is responsible for identifying the input pages. The test is divided into two main groups:

- **Faktura document type testing** – Faktura uses the fingerprint therefore the test is performed with two types of input pages – known and not known page type.
    - Fingerprint – The system already knows this page type
    - New page type – New page type, system needs to make sure no fingerprint matches this page.
- **DodaciList document type testing** – DodaciList uses the QR code based identification.

Table 27 and **Table 28** summarize test results for the both testing groups. **Figure 40** provides the graphical representation of the results. Testing was performed on the 20 batches per group (10+10 in case of Faktura document type). Every batch contained 10 documents.

Table 27 – PageID performance test results for Faktura document type

	Fingerprint			**New page type**		
	Mean scan time [s]	Variance [s]	Standard deviation [s]	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	4,79	0,002	0,049	4,83	0,002	0,052
300 DPI	5,42	0,003	0,042	5,50	0,003	0,049
400 DPI	6,69	0,003	0,052	6,83	0,003	0,055

Table 28 – PageID performance test results for DodaciList document type

	QR code		
	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	2,54	0,001	0,027
300 DPI	2,87	0,001	0,022
400 DPI	3,54	0,002	0,031

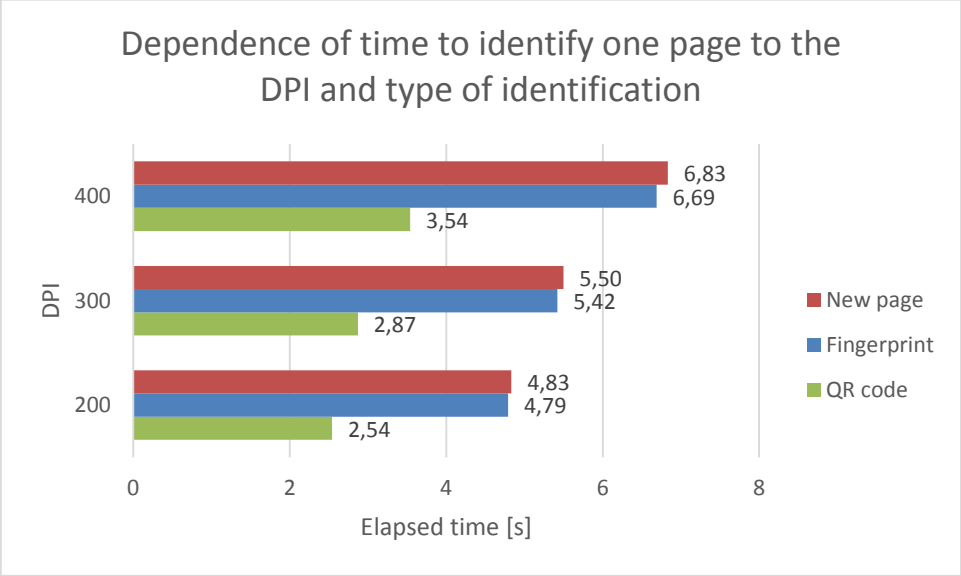

Figure 40 – PageID performance test graph

7.2.3 Profiler

Profiler is responsible for the data recognition, data clean and server based validation. The test uses the same testing groups as in PageID testing (7.2.2):

- **Faktura document type testing** – Faktura uses the fingerprint therefore the test is performed with two types of input pages – known and not known page type.
  - Fingerprint – The system uses the fingerprint based zones
  - New page type – Data are located using the regular expression location method.
- **DodaciList document type testing** – Data are captured using zones.

Table 29 and **Table 30** summarize test results for the both testing groups. **Figure 41** provides the graphical representation of the results. Testing was performed on the 20 batches per group (10+10 in case of Faktura document type). Every batch contained 10 documents.

Table 29 – Profiler performance test results for Faktura document type

	Fingerprint			**New page type**		
	Mean scan time [s]	Variance [s]	Standard deviation [s]	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	0,20	0,001	0,0003	0,74	0,002	0,0004
300 DPI	0,27	0,002	0,0004	0,86	0,002	0,0004
400 DPI	0,35	0,002	0,0003	0,91	0,003	0,0005

Table 30 - PageID performance test results for DodaciList document type

	DodaciList		
	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	0,94	0,002	0,0005
300 DPI	1,01	0,002	0,0004
400 DPI	1,25	0,002	0,0005

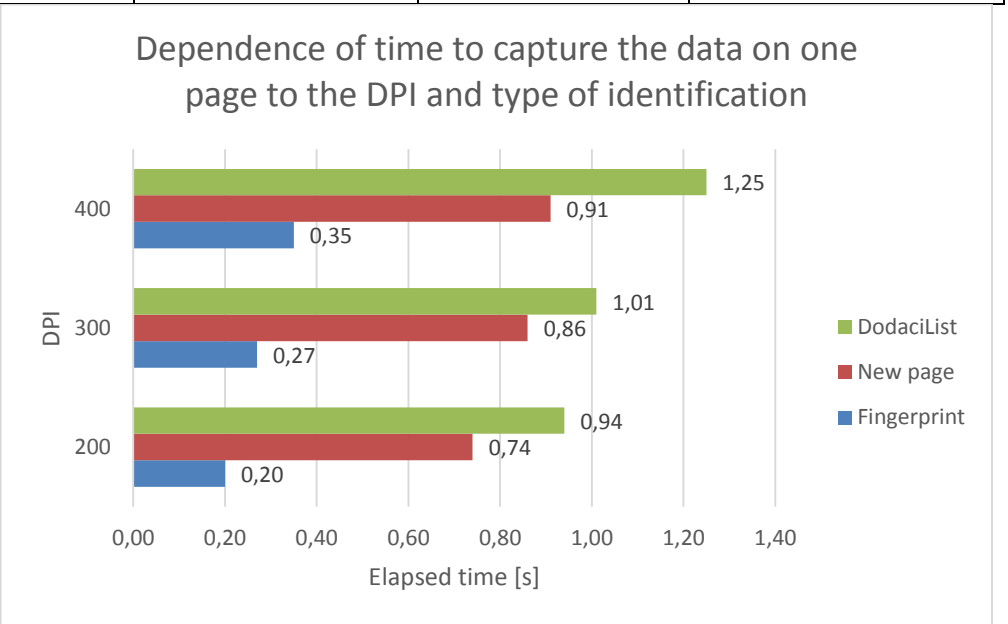

Figure 41 – Profiler performance test graph

Another test performed in the Profiler testing was to determine the recognition success depending on the DPI. **Table 31** summarizes the recognition results. Test methodology included the following steps:

- Automatic fingerprint based recognition for Faktura document type. DodaciList document type was identified using the QR recognition.
- Comparison correctly and incorrectly extracted fields
- Faktura document type contains 11 fields, DodaciList contains 3 fields.

Figure 42 provides the graphical representation of the results. Testing was performed on the 20 batches per group. Every batch contained 10 documents.

Table 31 – Profiler data recognition success

	Faktura			**DodaciList**		
	Mean capture success rate [%]	Variance [%]	Standard deviation [%]	Mean capture success rate [%]	Variance [%]	Standard deviation [%]
200 DPI	71,72	19,38	3,75	72,50	17,50	3,06
300 DPI	84,18	9,77	0,7	85,00	12,25	1,50
400 DPI	86,53	8,34	0,95	87,50	12,50	1,56

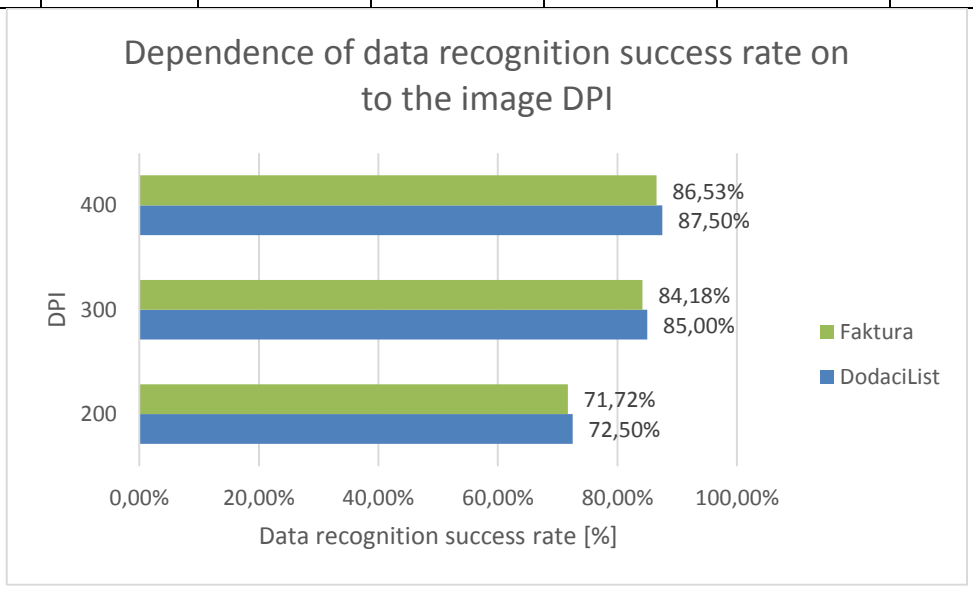

Figure 42 – Data recogniton success rate graph

The results shows big difference in data recognition success between 200 and 300 DPI, almost 13% more data was correctly recognized for invoices and 12.5% for delivery notes. The difference in data recognition between 300 and 400 DPI is not so significant.

Figure 43, **Figure 44**, **Figure 45** shows differences in quality of recognition based on the scanning resolution (DPI).

Figure 43 – Captured text at 200 DPI

Figure 44 – Captured text at 300 DPI

Figure 45 – Captured text at 400 DPI

7.2.4 Export

Export is the last step in the application workflow. PDF and XML files for the each document are exported.

Table 32 summarizes test results for the both testing groups. **Figure 46** provides the graphical representation of the results. Testing was performed on the 20 batches per group. Every batch contained 10 documents.

Table 32 – Export performance test results for Faktura document type

	Faktura			**DodaciList**		
	Mean scan time [s]	Variance [s]	Standard deviation [s]	Mean scan time [s]	Variance [s]	Standard deviation [s]
200 DPI	2,75	0,157	0,024	2,14	0,098	0,019
300 DPI	2,81	0,162	0,027	2,31	0,101	0,024
400 DPI	2,94	0,159	0,021	2,5	0,099	0,023

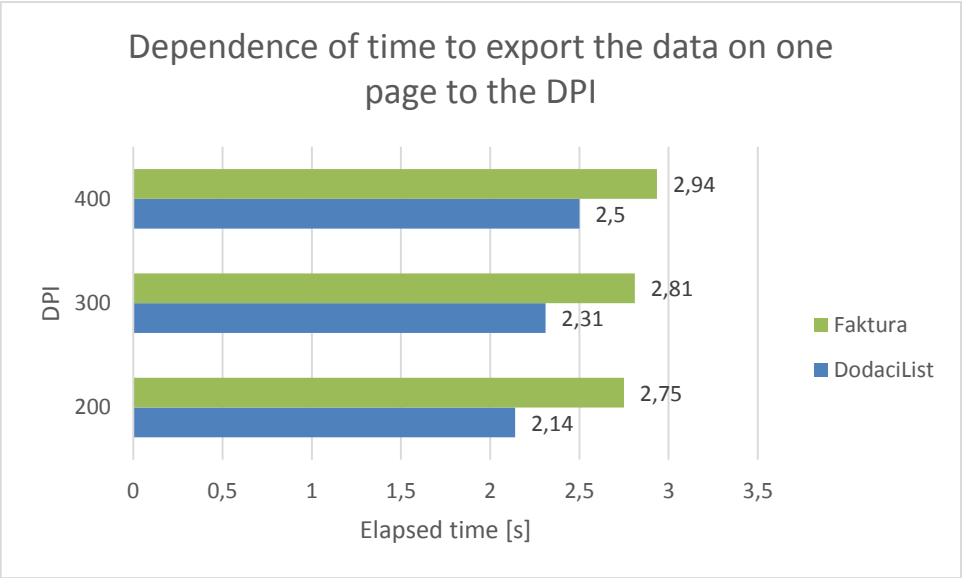

Figure 46 – Export performance test graph

CONCLUSION

The thesis provides an overview of the Enterprise Content Management and detailed information about the ECM Capture part. Theory part presents the most known ECM Capture technologies. The thesis objectives were to present the IBM Datacap Taskmaster Capture technology and show the possibilities that this technology provides.

The analysis provides the sample implementation of processing an input invoices and internal delivery notes between branch offices. Analysis copies the real deployment process from gathering the customer requirements to application deployment and application testing.

Meebootix is a sample company that provided several requirement for the future system. I created the use cases based on these requirements and used them for the functional testing. The application passed all the functional testing scenarios and test cases. The general problem in the functional testing was the missing protection against the multiple incorrectly inserted user name or password. This is not the problem of the application but the system. Functional testing also discovered a few deficiencies such as missing information in user Verify that only shows the wrong data, but not provides additional information what is wrong with the data. Also the supplier selection list could provide additional functionality such as filtering in the existing list or possibility to add a new supplier. Another further improvement could be additional choice in the routing operations to rescan the document.

The designed application was also subjected to the performance testing. The testing was performed only on the system parts, this means the user interaction was not measured. The reason why the performance testing was done only on the system parts is because the user performance differs from person to person and the results could be very misleading. Also the results show what application processes and does not need to be done by the user. The testing was done on the three types of image resolution (200, 300 and 400 DPI). I measured the processing time of the each page within the batch of documents. The measurement result shows the time depending on the image DPI in every workflows task. The Profiler testing also provides results how much data was correctly recognized depending on the image DPI. The results shows big difference in data recognition success between 200 and 300 DPI, almost 13% more data was correctly recognized for invoices

and 12.5% for delivery notes. The difference in data recognition between 300 and 400 DPI is not so significant, but the processing time of the 400 DPI image is up to 28% longer.

The application is ready for the demonstration proposes and with additional improvements is also ready for the production environment. For the production I would recommend additional improvements in the data validation such as enhanced IC validation using the IC verification algorithm. Another good improvement can be connection with the ARES system provided by the Ministry of the interior of the Czech Republic. This system provides official information from the Czech commercial register and this information could be used in the verification to ensure the input data is correct.

BIBLIOGRAPHY

1. **AIIM.** What is Enterprise Conternt Management (ECM)? *AIIM.* [Online] [Cited: Feburary 9, 2014.] http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.

2. **M-Files.** *The Business Case for Enterprise Content Management.* Dallas : M-Files, 2013.

3. **Rouse, Margaret.** enterprise content management (ECM). *SearchWindowsServer.* [Online] September 13, 2013. [Cited: March 22, 2014.] http://searchwindowsserver.techtarget.com/definition/enterprise-content-management-ECM.

4. **Kampffmeyer, Ulrich.** ComputerWoche. *ECM - Herrscher über Informationen.* [Online] September 24, 2001. [Cited: March 23, 2014.] http://www.computerwoche.de/a/cw-exktrakt-herrscher-ueber-informationen,523803.

5. **Boiko, Bob.** *Content Management Bible, 2nd Edition.* Indianapolis : Wiley Publishing, Inc., 2005. 0-7645-7371-3.

6. **McGrath, Adrian.** Making the "E" in Enterprise Content Management actually mean something. *The Information Management Pulse.* [Online] April 5, 2010. [Cited: March 23, 2014.] http://mcgratha.wordpress.com/2010/04/05/ecm-erp/.

7. **Kampffmeyer, Ulrich.** *Trends in Record, Document and Enterprise Content Management.* Hamburg : PROJECT CONSULT GmbH, 2004.

8. **Buckland, Michael Keeble.** *Information and information systems.* Westport : ABC-CLIO, 1991. 0-313-27463-0.

9. **Kofax, Inc.** Glossary - Strctured Document. *Kofax.* [Online] [Cited: April 1, 2014.] http://www.kofax.com/glossary/structured-document/.

10. **Kofax, Inc.** Glossary - Semi-Structured Document. *Kofax.* [Online] [Cited: April 1, 2014.] http://www.kofax.com/glossary/semi-structured-document/.

11. **Kofax, Inc.** Glossary - Unstructured Document. *Kofax.* [Online] [Cited: April 1, 2014.] http://www.kofax.com/glossary/unstructured-documents/.

12. **TechTerms.com.** Metadata. [Online] 2014. [Cited: April 1, 2014.] http://www.techterms.com/definition/metadata.

13. **Kofax, Inc.** Glossary - Batch. *Kofax.* [Online] [Cited: April 1, 2014.] http://www.kofax.com/glossary/batch/.

14. **ABBYY.** ABBYY - Company Overview. *ABBYY.* [Online] [Cited: April 12, 2014.] http://web.archive.org/web/20091215111857/http://www.abbyy.com/company.

15. **ABBYY.** Overview. *ABBYY Recognition Server.* [Online] [Cited: April 12, 2014.] http://www.abbyy.com/recognition_server/.

16. **ABBYY.** Overview. *ABBYY FlexiCapture.* [Online] [Cited: April 12, 2014.] http://www.abbyy.com/data_capture_software/.

17. **EMC2.** Corporate Profile. *EMC2.* [Online] [Cited: April 12, 2014.] http://www.emc.com/corporate/emc-at-glance/corporate-profile/index.htm.

18. **EMC2.** EMC Captiva QuickScan Pro. *Captiva Family.* [Online] [Cited: April 12, 2014.] http://www.emc.com/enterprise-content-management/captiva/quickscan-pro.htm.

19. **EMC2.** Captiva Capture. *Captiva Family.* [Online] [Cited: April 12, 2014.] http://www.emc.com/enterprise-content-management/captiva/capture.htm.

20. **EMC2.** EMC Captiva Advanced Recognition. *Captiva Family.* [Online] [Cited: April 12, 2014.] http://www.emc.com/enterprise-content-management/captiva/advanced-recognition.htm.

21. **SearchITChannel.** IBM. *IT Channel resources.* [Online] November 2006. [Cited: April 12, 2014.] http://searchitchannel.techtarget.com/definition/IBM-International-Business-Machines.

22. **IBM Corporation.** Datacap taskmaster Capture. [Online] 2011. [Cited: April 12, 2014.] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=BR&htmlfid=IMB14111USEN.

23. **Kofax, Inc.** About Us. *Kofax.* [Online] [Cited: April 12, 2014.] http://www.kofax.com/company/.

24. **Kofax, Inc.** *Kofax Express.* [Online] 2013. [Cited: April 12, 2014.] http://www.kofax.com/downloads/datasheets/ds-kofax-express-en.pdf.

25. **Kofax, Inc.** Overview. *Kofax Capture.* [Online] [Cited: April 12, 2014.] http://www.kofax.com/document-capture-software/.

26. **Kofax, Inc.** Overview. *Kofax Transformation Modules.* [Online] [Cited: April 12, 2014.] http://www.kofax.com/forms-processing/.

27. **Wikipedia.** *Nuance communications.* [Online] April 9, 2014. [Cited: April 2014, 2014.] http://en.wikipedia.org/wiki/Nuance_Communications.

28. **Nuance.** OmniPage Ultimate. [Online] [Cited: April 12, 2014.] http://www.nuance.com/for-business/by-product/omnipage/index.htm.

29. **Wikipedia.** *Open Text Corporation.* [Online] April 10, 2014. [Cited: April 13, 2014.] http://en.wikipedia.org/wiki/Open_Text_Corporation.

30. **OpenText.** Featured Products. *Capture.* [Online] [Cited: April 13, 2014.] http://www.opentext.com/What-We-Do/Products/Enterprise-Content-Management/Capture.

31. **IBM Corporation.** IBM Datacap Taskmaster Capture. *FileNet P8 Platform Information Center.* [Online] [Cited: April 12, 2014.] http://pic.dhe.ibm.com/infocenter/p8docs/v5r2m0/index.jsp?topic=%2Fcom.ibm.p8.sysoverview.doc%2Fp8sov140.htm.

32. **IBM Corporation.** Application Development Guide: Using IBM Datacap Taskmaster Capture v8.1. [Online] 2013. http://www-05.ibm.com/e-business/linkweb/publications/servlet/pbi.wss?CTY=US&FNC=SRX&PBL=SC19-3251-05.

33. **IBM Corporation.** Application Development Guide: Using IBM Datacap Taskmaster Capture v8.0.1. [Online] 2011. http://www.elink.ibmlink.ibm.com/publications/servlet/pbi.wss?CTY=US&FNC=SRX&PBL=SC19-3251-01.

34. **Zhu, Jackie, et al., et al.** *Implementing Imaging Solutions with IBM Production Imaging Edition and IBM Datacap Taskmaster Capture.* New York : IBM Corporation, 2011.

35. **IBM Corporation.** Taskmaster architecture. *IBM Datacap Information Center.* [Online] November 2013. [Citace: 14. April 2014.] http://pic.dhe.ibm.com/infocenter/datacap/v8r1m0/index.jsp?topic=%2Fcom.ibm.dc.install.doc%2FTM_Intro_What_is_IBM_Datacap_Taskmaster_Capture.htm.

36. **IBM Corporation.** Taskmaster software components. *IBM Datacap Information Center.* [Online] November 2013. [Citace: 14. April 2014.] http://pic.dhe.ibm.com/infocenter/datacap/v8r1m0/index.jsp?topic=%2Fcom.ibm.dc.install.doc%2Fdcpov005.htm.

37. **Netlingo.** Cloud computing. [Online] [Cited: April 14, 2014.] http://www.netlingo.com/word/cloud-computing.php.

38. **Chun, Wesley.** What is Cloud Computing. *Google Developers Academy.* [Online] June 2012. [Cited: April 14, 2014.] https://developers.google.com/appengine/training/intro/whatiscc.

39. **interoute.** What is SaaS. *Cloud Articles.* [Online] [Cited: April 15, 2014.] http://www.interoute.com/what-saas.

40. **interoute.** What is PaaS. *Cloud Articles.* [Online] [Cited: April 15, 2014.] http://www.interoute.com/what-iaas.

41. **interoute.** What is IaaS. *Cloud Articles.* [Online] [Cited: April 15, 2014.] http://www.interoute.com/what-iaas.

42. **Key, Dave.** Requirements for Enterprise SaaS Applications. *Cloud Strategies.* [Online] [Cited: April 21, 2014.] http://www.cloudstrategies.biz/requirements-for-building-enterprise-saas-applications/.

LIST OF ABBREVIATIONS

ADSI	Active Directory Service Interfaces
AES	Advanced Encryption Standard
AIIM	Association for Information and Image Management
BPM	Business Process Management
CIFS	Common Internet File System
CM	Content Management
COLD	Computer Output on Laser Disk
DCO	Document hierarchy
DMS	Document Management System
DPI	Dots per inch
DRT	Document related Technologies
EAI	Enterprise Application Integration
ECM	Enterprise Content Management
ERM	Enterprise Report Management
ERP	Enterprise Resource Planning
FIPS	Federal Information Processing Standard
HCR	Handprint Character Recognition
HTTPS	Hypertext Transfer Protocol Secure
IaaS	Infrastructure as a Service
ICR	Intelligent Character Recognition
IEC	International Electrotechnical Commission
IIS	Internet Information Services
ISO	International Organization for Standardization
IT	Information Technology

ITaaS	Information Technology as a Service
LDAP	Lightweight Directory Access Protocol
OCR	Optical Character Recognition
OLE DB	Object Linking and Embedding for Database
OMR	Optical Mark Recognition
PaaS	Platform as a Service
PDF	Portable Document Format
QR Code	Quick Response Code
RM	Records Management
SaaS	Software as a Service
SOA	Service-Oriented Application
TCP	Transmission Control Protocol
TIFF	Tagged Image File Format
USB	Universal Serial Bus
WCM	Web Content Management
XML	eXtensible Markup Language
XSD	XML Schema Definition

LIST OF FIGURES

LIST OF TABLES

APPENDICES

APPENDIX A I. USE CASE SCENARIOS

APPENDIX A II. IC VERIFICATION ALGORITHM

APPENDIX A I. USE CASE SCENARIOS

UC001: Scan mixed documents			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the scan client
	4	S	Create the batch
	5	U	Prepare documents for scanning - mix invoices and delivery notes documents together
	6	U	Insert documents into the scanner
	7	U	Start scanning
	8	S	Imports scanned images
	9	U	Check if document were scanned correctly
	10	U	Submit the batch
	11	S	Receive the batch and close
Extensions	2a	S	User credentials are incorrect - return to step 1
	7a	S	Scanner malfunction - check and repair the scanner and return to 7 or EXIT
	8a	S	Scanned images cannot be imported - Internal system error EXIT
	9a	U	Document are no scanned correctly
	9b	U	Delete incorrect pages
	9c	U	Insert the deleted pages to the scanner and return to 7
	11a	S	Batch cannot be received - Internal system error EXIT

UC002: Scan invoices			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the scan client
	4	S	Create the batch
	5	U	Prepare documents for scanning – only invoices
	6	U	Insert documents into the scanner
	7	U	Start scanning
	8	S	Imports scanned images
	9	U	Check if document were scanned correctly
	10	U	Submit the batch
	11	S	Receive the batch and close
Extensions	2a	S	User credentials are incorrect - return to step 1
	7a	S	Scanner malfunction - check and repair the scanner and return to 7 or EXIT
	8a	S	Scanned images cannot be imported - Internal system error EXIT
	9a	U	Document are no scanned correctly
	9b	U	Delete incorrect pages
	9c	U	Insert the deleted pages to the scanner and return to 7
	11a	S	Batch cannot be received - Internal system error EXIT

UC003: Scan delivery notes			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the scan client
	4	S	Create the batch
	5	U	Prepare documents for scanning – only delivery notes
	6	U	Insert documents into the scanner
	7	U	Start scanning
	8	S	Imports scanned images
	9	U	Check if document were scanned correctly
	10	U	Submit the batch
	11	S	Receive the batch and close
Extensions	2a	S	User credentials are incorrect - return to step 1
	7a	S	Scanner malfunction - check and repair the scanner and return to 7 or EXIT
	8a	S	Scanned images cannot be imported - Internal system error EXIT
	9a	U	Document are no scanned correctly
	9b	U	Delete incorrect pages
	9c	U	Insert the deleted pages to the scanner and return to 7
	11a	S	Batch cannot be received - Internal system error EXIT

UC004: Manual Page Identification			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the ManualPageID client
	4	S	Present the batch pages
	5	U	Set the correct page types for pages identified as Other page type
	6	U	Submit the batch
	7	S	Close the batch
Extensions	2a	S	User credentials are incorrect - return to step 1
	4a	S	Pages cannot be presented - Internal system error EXIT
	5a	U	Page type cannot be selected because the page type is not in combo box - Internal server error EXIT

UC005: Validate mixed documents			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the verify client
	4	S	Open the batch and present document that needs to be verified
	5	U	Verify documents - UC006 and UC007
	6	U	Submit the batch
	7	S	Close the batch
Extensions	2a	S	User credentials are incorrect - return to step 1
	4a	S	Batch cannot be opened - Internal server error EXIT

UC006: Validate invoices			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the verify client
	4	S	Open the batch with invoices and present document that needs to be verified
	5	U	Check fields that needs verification
	6	U	Correct the wrong input data
	7	U	Click the submit button
	8	S	Validate and clean the data
	9	S	Close the batch
Extensions	2a	S	User credentials are incorrect - return to step 1
	4a	S	Batch cannot be opened - Internal server error EXIT
	5a	U	Fields contain wrong data, Red background indicates failed validation, Yellow background indicates low confidence characters - return to step 6
	6a	U	Data cannot be corrected - Internal server error EXIT
	7a	U	Button cannot be pressed - Internal server error EXIT
	8a	S	Run clean and validation rules on input data - Validation fails - return to step 4

UC007: Validate delivery notes			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Log in to the system
	2	S	Verify user credentials
	3	U	Run the verify client
	4	S	Open the batch with delivery notes and present document that needs to be verified
	5	U	Check fields that needs verification
	6	U	Correct the wrong input data
	7	U	Click the submit button
	8	S	Validate and clean the data
	9	S	Close the batch
Extensions	2a	S	User credentials are incorrect - return to step 1
	4a	S	Batch cannot be opened - Internal server error EXIT
	5a	U	Fields contain wrong data, Red background indicates failed validation, Yellow background indicates low confidence characters - return to step 6
	6a	U	Data cannot be corrected - Internal server error EXIT
	7a	U	Button cannot be pressed - Internal server error EXIT
	8a	S	Run clean and validation rules on input data - Validation fails - return to step 4

UC008: Choose supplier from enumeration based on IC value			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Check if IC field is filled
	2	U	Press the submit button
	3	S	Find the Supplier name in database based on the IC field value
	4	S	Insert found Supplier name into the field
Extensions	1a	U	IC is not filled - Insert IC value from the invoice
	1b	U	IC is not present on the invoice - Mark document for revision EXIT
	3a	S	Supplier name cannot be found - leave the field value empty
	3b	U	Click on the Supplier button
	3c	U	Select the Supplier name from the list
	3d	U	Supplier name is not in the list - Mark document for revision EXIT

UC009: Define new fingerprint			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Select document that needs to be defined
	2	U	Select the field that needs to be defined
	3	U	Draw a zone on the page where the data are located
	4	S	Snap the OCR data to the field value
	5	U	Repeat steps 2-4 for all fields

UC010: Affect fingerprint creation			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Select document where the fingerprint creation will be affected
	2	U	Check if NovaSablona value is set to ANO
	3	U	Select field that will be defined
	4	U	Draw a zone on the page where the data are located
	5	S	Snap the OCR data to the field value
	6	U	Repeat steps 2-4 for all fields that will be defined
	7	U	After the batch is exported check if the new fingerprint was created
Extensions	2a	U	NovaSablona value is set to NE - Change to ANO

UC011: Choose document's routing operations			
Main success scenario	**Step**	**Actor**	**Description**
	1	U	Select document that needs special routing
	2	U	Change the routing options to desired value; Possible values: Review, Delete
	3	U	Check if the required action was performed after the batch is exported

APPENDIX A II. IC VERIFICATION ALGORITHM

Example IC = 69663963

1. The first to the seventh digit is multiplied with the numbers 8, 7, 6, 5, 4, 3, 2 and the results of multiplication are summed

   *SUM = 6*8 + 9*7 + 6*6 + 6*5 + 3*4 + 9*3 + 6*2 = 228*

2. Count the division remainder of a number eleven.

   MODULO = SUM % 11

   MODULO = 228 % 11 = 8

3. Following rules has to be valid for the eight number (marked as **c**) of the Example IC
   a. If the MODULO is 0 or 10, then c = 1
   b. If the MODULO is 1, then c = 0
   c. In other cases c = 11 – MODULO

   c = 11 – 8 = 3

   Last number of Example IC is 3, c = 3 -> IC is valid!