Design and Implementation of a fault tolerant form processing application using machine learning

Master's Thesis in Computer Science

submitted
by

Christoph Neubauer

born 23.06.1991 in Homburg (Saar)

Written at

Lehrstuhl für Mustererkennung (Informatik 5)
Department Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg.

in Cooperation with

Universidade Federal do Parana
Curitiba

Advisor: PD Dr.-Ing. habil. Peter Wilke

Second Advisor: Prof. Luiz Eduardo S. Oliveira

Started: 12.09.2016

Finished: 13.03.2017

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Studien- und Diplomarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 12. November 2016

Übersicht

Abstract

Contents

Chapter 1

Introduction

Some general information on the context and setting.

1.1 Motivation

Specific motivation for the problem at hand.

1.2 Task

The task of this master thesis is to evaluate different electornic invoice formats. An application shall be designed and implemented that processes invoice forms and is capable of storing them in the invoice format that suits the most. The application should use machine learning in order to improve processing accuracy over time. Nevertheless, errors during the scan process should be handled by the application itself. The output of this application should be conform with the definition of the electronic invoice format that has been decided in beforehand.

1.3 Related Work

Other relevant academic work and how it differs from this work, for example "textual" citation, as shown in "parenthesis" citation

1.4 Results

What has been achieved in this work?

1.5 Outline

How is the thesis structured and why?

1.6 Acknowledgments

A big thank you for the support to ...

Chapter 2

InvoiceFormats

During the technologization of companies over the world, electronic invoices (also known as e-invoice) have become more and more important.

2.1 Description of leading formats

2.2 Definition of decision criteria

2.3 Comparison and decision finding

Chapter 3

OCR

Chapter 4

Implementation

4.1 Requirements definition

4.2 Definition of modules

4.3 Architectural concept

4.4 Module 1 - OCR

4.5 Module 2 - ANN

4.6 Module 3 - Converter

4.7 Problems during the implementation

Chapter 5

Conclusion

5.1 Result of the application

5.2 Findings

5.3 Outstanding issues

Chapter 6

InvoiceFormats

During the technologization of companies over the world, electronic invoices (also known as e-invoice) have become more and more important. E-Invoicing offers companies the possibility to improve their business processes, making invoicing faster and more efficient and enables a direct connection to other tools like ERP-Software.

To enable companies these benefits and in order to make the communication between companies even possible a comprehensive standard has to be defined. With an invoice standard at hand, companies can use invoices from their business partners and read them into their systems (in case of B2B).

There are several invoice standards in action at the moment. The next section will deal with the most important ones and describes them as well as pointing out the benefits and drawbacks of the format. After that, the next section defines criteria that are relevant for the application and how to measure them. In the last section, these criteria are applied on the formats defined in section 6.1 and compared against each other. Eventually a decision regarding the usage of one of these formats is made.

6.1 Description of leading formats

6.1.1 UN/EDIFACT

EDIFACT is a well-established [Kau15] subset of standards from CEFACT regarding the electronic interchange of structured data. It is developed and maintained by the United Nations Economic Comission for Europe (UNECE) [fE16].

The European Commission states that the UN/EDIFACT INVOIC message has been a cornerstone in electronic invoicing over the past years [**?**, **?**].

There are several subsets of EDIFACT that have been developed for different industries. For instance, the chemical industry uses CEFIC/ESCom[1] as their standard, while automotive industry is in charge with ODETTE/FTP2[2].

EDIFACT has different message types such as ORDCHG for a request to change an order or PAYORD which contains a payment order. In the context of this thesis, the message type INVOIC (containing an invoice) is the most interesting one.

6.1.2 XCBL

The XML Common Business Library is an extension of the CBL which originally has been devloped by Veo Systems Inc. [Cov01]. The company has been bought by Commerce One Inc. in 1999 [CO00], page 29.

xCBL currently exists in version 4.0 (since 2003)[3]. Since the company has gone bancrupt in 2004 **??** it is not very likely that this format gains more interest in the future.

6.1.3 OASIS/UBL

UBL stands for Universal Business Language and is being developed by OASIS. The current version is 2.1 and is normed by the international standardization organization[4].

Several countries developed their own subset of this format. Especially interesting in this case is a project called PEPPOL (Pan-European Public Procurement Online project) that aims at developing a format for public sectors in the whole European Union[5].

Also interesting in the context of invoice interchange is the UBL-based project called *simplerinvoicing* that aims at connecting ERP systems with accounting and e-invoicing software by providing an own invoicing standard[6].

6.1.4 ZugFerd

This invoice format has been published initially in 2014 [**?**]. The name is a german acronym, containing the name of the corresponding forum (FeRD). It can be translated to ”Central User

[1] see also: https://www.cefic.org/Industry-support/Implementing-reach/escom/

[2] see also: https://www.odette.org/services/oftp2

[3] see also: https://www.xcbl.org

[4] see ISO/IEC 19845:2015

[5] see also: https://www.peppol.eu/about_peppol/about-openpeppol-1

[6] see also: www.simplerinvoicing.org/en/

Guide of the Forum for electronic Invoicing in Germany". Although this invoice format is rather young it tries to fulfill the directive 2014/55/EU of the european parliament [?] while still being flexible and simple. This directive states that the use of electronic invoice formats should be adopted by all member states of the european union until the 27. of November 2018 [?, ?].

The approach of the ZugFerd-format enables not only big companies to work with that format, but also smaller and medium companies (SME's) that are in need of such a format but are normally not able to implement a complex electronic invoice standard. Furthermore three levels of conformance are defined: Basic, Comfort and Extended. Each of those levels have a different amount of required information fields, that have to be set in order to be a valid ZugFerd-format. Nevertheless, in all of the three formats, it is possible to define more information in free text fields.

This enables extensibility of the format and the possible business areas in which this standard can be used.

The German Forum for electronic invoice (FeRD) states that this format has been accepted as a core standard in Germany to be used in the future such that every company, that wants to start business relations with a german company, has to use this standard **??**. Furthermore, the possibility to extend this standard to all European Countries is in sight, as stated in **??**.

6.2 Definition of decision criteria

While the standards defined in the section before focus on specific areas or try to combine multiple fields, this section defines the criteria that are most relevant for the application that is developed.

6.2.1 Future Potential

One of the major criteria for a suitable invoice standard should be its future potential. Developing an application that deals with a standard that is not being used 10 years later does not make sense. Therefore, any standard that is going to be replaced should not be considered useful.

6.2.2 Relevance in Germany (and Europe)

As this thesis is being written at a German university, the chosen standard should be relevant in Germany. Even better if it is relevant in Europe as well. On the other Hand, standards that are not of interest for Europe should be excluded.

6.2.3 Extendability to more countries

The possibilities of a standard to be used in other countries will also affect its importance over the next decades. Standards that only suits the requirements of one country are not important enough. The focus lies on standards with a wide (possible) range of countries to be affected, instead.

6.2.4 Extendability of the standard itself

Last but not least, the extendability of the standard itself is an important criterion. The world is changing and new requirements are coming while older ones are getting broken up. A valuable standard should be able to deal with these changes and should be extensible towards new requirements, or special requirements in specific business areas.

6.2.5 Complexity

The complexity of the standard is important for this thesis too. Not only is the development of the application limited by time, but also makes a complex standard it hard to understand it and less error-prone.

6.3 Comparison and decision finding

6.3.1 Application of the criteria

6.3.2 Decision and explanation

Appendix A

Glossary

B2B - Business to Business

ERP - Enterprise Resource Planning

EDIFACT - Electronic Data Interchange For Administration, Commerce and Transport

FeRD - Forum für elektronische Rechnung Deutschland

SME - Small and medium enterprises

UBL - Universal Business Language

xCBL - XML Common Business Library

ZugFerd - Zentraler User Guide des Forums für elektronische Rechnung Deutschland

List of Figures

List of Tables

Bibliography

[Bre02] Thomas M. Breuel. *Two Geometric Algorithms for Layout Analysis*, pages 188–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[Bre03] Thomas M. Breuel. High performance document layout analysis, 2003.

[Bre07] Thomas M. Breuel. The hocr microformat for ocr workflow and results. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 1063–1067, Sept 2007.

[Bre08] Thomas M. Breuel. The ocropus open source ocr system, 2008.

[Che08] Qiang Chen, Quan-sen Sun, Pheng Ann Heng, and De-shen Xia. A double-threshold image binarization method based on edge detector. *Pattern Recogn.*, 41(4):1254–1267, April 2008.

[Che13] Shuhan Chen, Weiren Shi, and Wenjie Zhang. An efficient universal noise removal algorithm combining spatial gradient and impulse statistic. *Mathematical Problems in Engineering*, 2013:1–12, 2013.

[CO00] Inc. Commerce One. Annual report of 2000, 2000. https://www.media.corporate-ir.net/media_files/NSD/CMRC/reports/10_k.pdf, last visited on 09.11.2016.

[Cov01] Robin Cover. Xml common business library (xcbl), 2001. https://www.xml.coverpages.org/cbl.html, last visited on 09.11.2016.

[Den14] Andreas Dengel and Faisal Shafait. *Analysis of the Logical Layout of Documents*, pages 177–222. Springer London, London, 2014.

[fE16] UN Economic Commission for Europe. Introducing un/edifact, 2016. https://www.unece.org/cefact/edifact/welcome.html, last visited on 08.11.2016.

[Ham07] Hatem Hamza, Yolande Belaïd, and Abdel Belaïd. Case-based reasoning for invoice analysis and recognition. In Rosina O. Weber and Michael M. Richter, editors, *7th International Conference on Case-based Reasoning - ICCBR 2007*, volume 4626, pages 404–418, Belfast, United Kingdom, August 2007. Springer Berlin / Heidelberg. The original publication is available at www.springerlink.com , ISBN 978-3-540-74138-1, ISSN 0302-9743 (Print) 1611-3349 (Online).

[Kau15] Achim Kauffmann. 5 punkte, die sie über zugferd wissen sollten, 2015. https://www.basware.de/blog/2015-07-10/ZUGFeRD-5-punkte-die-sie-wissen-sollten, last visited on 09.11.2016.

[Kle04] Bertin Klein, Stevan Agne, and Andreas Dengel. *Results of a Study on Invoice-Reading Systems in Germany*, pages 451–462. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[Men04] Kurt Menges. Commerce one declares bankruptcy: Does this foretell the fate of b2b e-commerce?, 2004. https://www.supplychainmarket.com/doc/commerce-one-declares-bankruptcy-does-this-fo-0001, last visited on 09.11.2016.

[Nag95] George Nagy. *Document image analysis: What is missing?*, pages 576–587. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995.

[Ram12] Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. Layout-aware text extraction from full-text pdf of scientific articles. *Source Code for Biology and Medicine*, 7(1):7, 2012.

[Wan15] Chenyang Wang, Yanhong Xie, Kai Wang, and Tao Li. *OCR with Adaptive Dictionary*, pages 611–620. Springer International Publishing, Cham, 2015.

[Zhu05] Li Zhuang and Xiaoyan Zhu. *An OCR Post-processing Approach Based on Multi-knowledge*, pages 346–352. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.