Extracting Information from Deeds by Optical Character Recognition (OCR) and Text Interpretation

Rik WOUTERS, Gert MEIJERINK and Jakub ZAVREL the Netherlands

Key words: archive, deed, OCR, text recognition, text interpretation, Textkernel

SUMMARY

Since a long time the Netherlands' Cadastre, Land Registry and Mapping Agency (in short Kadaster) delivers digital information to customers. Already in the early 1990s information to public notaries was disseminated through IBM-global network. In 2002 Kadaster opened the internet shop KOL (Kadaster-on-line), through which legal ownership information on real estate is provided.

At present Kadaster is scanning 15 million deeds which up to now were stored on microfilm. The accessibility of the deeds had its limitations, especially now Kadaster has become an organisation were all processes are organised in a centralised way and at a national level.

Part of the project deals with the retrieval of information concerning servitudes, easements and the like. By means of text recognition tools this information is extracted from recorded deeds. In addition this information has become? available on the internet. This paper describes which procedures and approaches have been used to make a next step in e-services provided by Kadaster. The paper also dedicates attention to a special techniques for interpretation, that was applied by the intelligence services in Romania.

The system saves around 75% of manual processing costs during the coding of semantically complex information from an existing 15 million archive deeds. Furthermore, the system uses a combination of a highly accurate (>99.8% precision) single class classifiers for negatives, and a domain specific region of interest detection module. The recognition modules are integrated in a web based workflow system supporting large scale distributed manual coding.

Extracting Information from Deeds by Optical Character Recognition (OCR) and Text Interpretation

Rik WOUTERS, Gert MEIJERINK and Jakub ZAVREL the Netherlands

1. INTRODUCTION

1.1 General

When buying or selling registered properties, in the Netherlands, one is legally obliged to register the accompanying notarial deeds in the cadastral system. This system ensures that the source of cadastral data and – accordingly – the information are kept up-to-date at all times. The Netherlands' Cadastre, Land Registry and Mapping Agency, in short Kadaster, keeps registers by law. These registers consist (among other things) of notarial deeds related to the registered properties. In most cases, these are deeds of conveyance (when transferring property from the buyer to the vendor) and mortgage deeds. The public registers contain details which indicate the rights that are related to the registered properties (legal status).

The most important details in the deeds, referred to above, which relate to immoveable property are incorporated in the cadastral register. The section in which, and the number under which the deed is listed in the public registers enables the user to find the original deed in the public registers, or to have someone do this for the user. The cadastral register also functions as an index for the public registers. It provides a clear overview of each parcel of, for example, the rights related to a parcel, ownership and the purchase price.
In the event of a dispute arising between the public registers and the cadastral register, the public registers take precedence for establishing the legal status of registered properties. Because most civil-law notaries, estate agents and other parties involved are directly affiliated with the cadastral system, the registered information is directly available. This is of great importance when one considers that more than one million real estate and mortgage transactions take place every year in the Netherlands.

1.2 Objectives concerning historic information

The land information system reaches a high quality standard and meets currently demands of society: a reliable, transparent and accessible cadastral information system. Because of the high level reached, new opportunities for delivering services to the society emerge. These opportunities concentrate on the delivery of services to third parties. The "selling point" is the fact that the cadastral registration has favourable characteristics for third parties:
- nation-wide coverage of data;
- high quality registrations;
- state of the art web-portal; and
- centrally organised IT-infrastructure.

The main challenge of Kadaster today is to make also available historic information through the internet, having the portal for national distribution of data in place. One of the most important historic information sources is the archive of field sketches and historic cadastral maps [1]. There is however, also a growing demand for information from historical deeds, more specifically information on easements, servitudes and the like. To that purpose recorded paper deeds are currently scanned and the text related to for mentioned easements is searched for: this is important information for the buyer when it comes to the purchase of a real estate object. This type of information will be on line available in the cadastre registration from 2010 onwards.

2. GLOBAL CONCEPT

2.1 Introduction

In many organisations, very large archives of electronic or paper documents exist, whose contents are crucial for the core tasks of the organisation. In the best case, the documents are available in a full text information retrieval system. However, the information that is essential for the organisation is often hidden in the concepts, meanings and relationships which are used in the text rather than in clear keywords. Often the concepts are complex, they can be expressed in text in many different ways and require domain expertise to be recognized by a reader. The ideal situation would be to code the semantic information as metadata in a structured information repository. However, the cost associated with converting the existing document archives by manual coding are huge, and are often considered an obstacle. Automatic classification methods are not considered applicable because their recognition and interpretation accuracy does not meet the information quality standards of the organisation. In this paper we discuss a text mining approach that can achieve significant cost reduction in this process, while at the same time guaranteing the required information quality standards.

The above situation holds true for the case of Kadaster in the Netherlands. The document archive contains 15 million tranfer deeds that are contracts concerning the transfer of ownership of real estate. The more recent parts of this archive is already available as PDF documents, the older parts of the archive as images which are scanned from microfilm. The documents can be considered “noisy text” because a large part of the text has been obtained by OCR from images with very poor quality. The legal right that is to be coded for the whole archive is the concept of “Erfdienstbaarheid” (explained in more detail below). In a pilot project it was determined that a human coding operator needs a) an average of three minutes per document to determine whether it contains the concept (positive class) and if yes b) to code which lot numbers are in that particular relationship, and c) identify all text fragments that specify the concept (land/right). Hence the goal of the project was to save cost on a budget which was estimated to cost five hundred thousand to one million man hours.

2.2 Registration of easements

When purchasing or valueing a piece of land or a dwelling it is of interest to know whether rights of third parties are vested on the parcel. For example: a right of way or a right of view. An “Erfdienstbaarheid” is an easement comprising a burden on a certain property ("the restricted parcel”) is loaded for another property (“the ruling parcel”) like a “right of way” providing access to the public street over an adjacent property. Frequently questions are asked by clients about the occurrence of easements, encumbrances, servitudes, etc. In that case the client can request Kadaster to conduct an easement investigation: in a so-called “*erfdienstbaarhedenonderzoek*”.

In case of an easement investigation, Kadaster examines whether exists registered documents (deed), which contains information intending to establish an easement on the serving yard. It is also possible that the Kadaster examines whether there are easements, in which case the parcel is involved asthe serving but as ruling yard.

The result of the investigation is an official statement which sums up the easements are found based on the examination to a certain specific date of all documents which relate to the plot as ‘serving parcel’ (possibly in combination with the plot as ‘ruling parcel’). It also includes the texts in the deed(s), that might relate to the requested easement. When a deed is only partially readable, the specific text of that the easement can not be selecetd., in which case a full copy of this deed is provide for free.

The research into easements is time consuming and means a lot of manual work, because all recorded deeds have to be assesssed. The research activity is bound to the location of storage because in all cases one has to access the paper archives. Especially notaries and notary clercs devote a lot of time to this kind of research.

The main objective of the project was to find and select the easements in the deeds and to record the detemined easements in the digital cadastral system in order to make them easily accessible. The new system will improve dataquality and reduces processing time of drawing up and recording notarial deeds.

The challenge now was to find and select the required information automatically, without (a lot of) human interference. That is why Optical Character Recognition (OCR) techniques and text processing applications were introduced.

The consequences of the wrong tracking or losses of information on easements are substantial. The buyer can blame the notary or Kadaster, who are liable in case of mistakes, and can request for indemnification in case of fiancial losses. The Quality Charter of the Kadaster promises alse for this type of information the normal high reliability.

2.3 Work flow

The "*Erfdienstbaarheden*"-programme had a budget of 20 M€. It comprises a large number of projects, including numerous activities. The total project was controlled by a work flow application. Figure 1 shows the basic steps.

The process consists of six basic steps, starting with the scan of deeds (micro films, XML-files and paper deeds). The second step is the activity where text is transformed into readable text with OCR- techniques (see section 3.1). The third step is the most important activity where the text is interpreted and the relevant information searched for is extracted (see section 3.2). After that, the manual checking takes place: the as "positive" selected deeds will be checked and relevant data stored in the database. All results will be checked independently. If the result does not meet the required standard the production of the whole day is disregarded. Finally the deed is archived.

Overall activities pertain to project management and control. In this programma adequate project management was crucial. The project was complex (new technology), had many stakeholders and subcontractors and the quality to be achieved was crucial. For that process information and mangement information were produce on a daily basis.

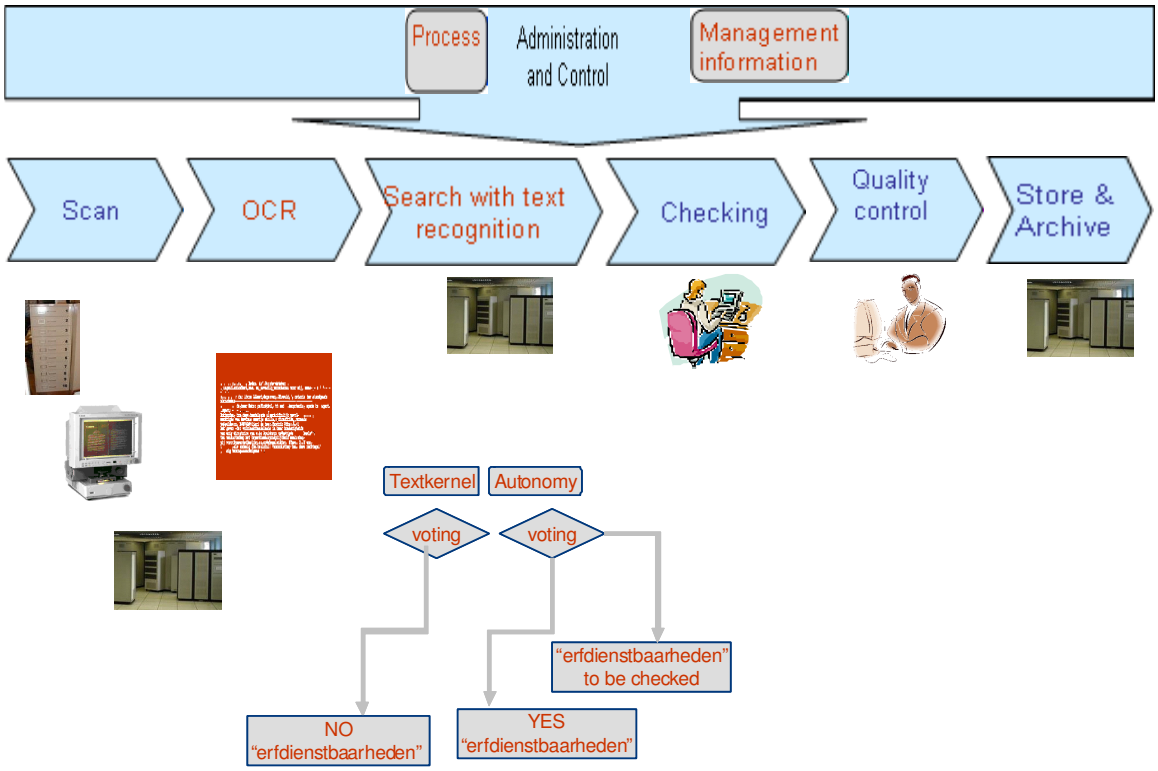

Figure 1. General scheme of project steps

3. BASIC APPLICATIONS FOR OCR AND INTERPRETATION

3.1 Character recognition

With the OCR-technology as explained, digitised text is generated. In the production street two OCR-applications are applied. On the left text page of Figure 2 one can see the result of the first application. It produces a result built up by several words (all relevant) but leaves a lot of words out. On the right page one can see the result of the second application that builds up a lot of words, not all of them correct, and misses out a few words. Now these two results are merged together to obtain a far better end result. This large amount of text archived by the OCR-application is needed to be successful in the next step of the process: i.e. text interpretation.

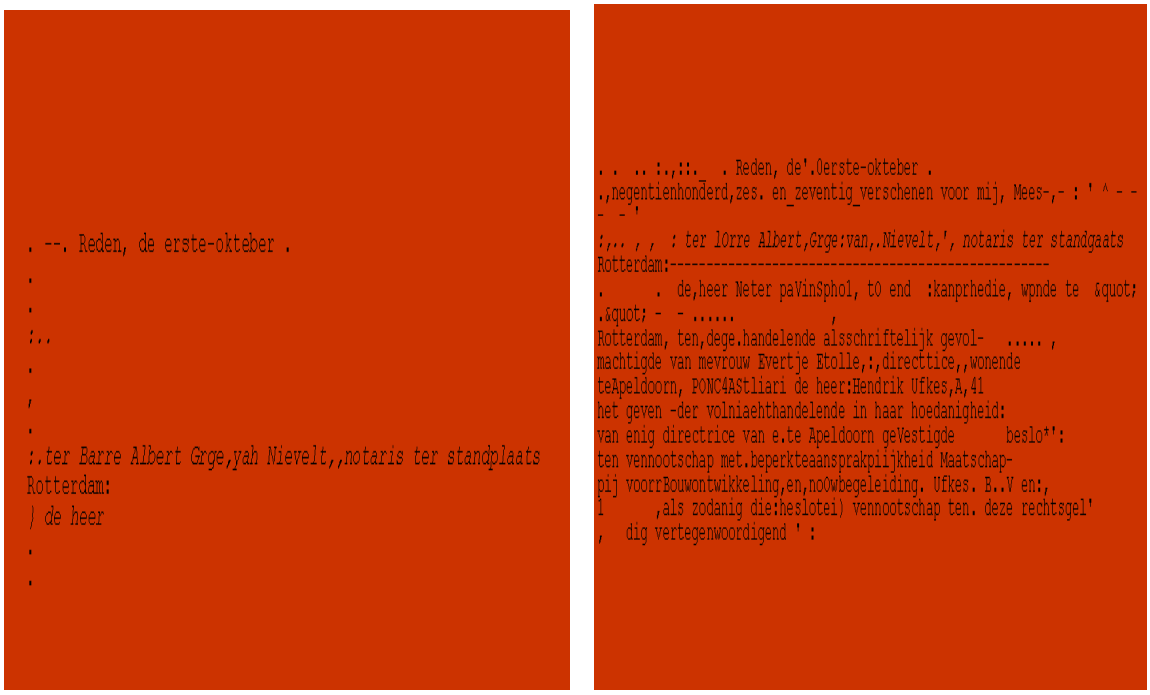

Figure 2. Left page text after first step and right page after second step

3.2 Text interpretation

The concept of easement as explained concerns the right of the owner of one parcel? of land (the dominant tenement) to restrict the freedom of another person's use of land (the servient tenement), or to guarantee his own use of it. It is a right (and obligation) that is treated as part of the property itself, and as such it passes from one owner of a parcel to the next. It is only established or abandoned in a deed approved by the notary that is available in the Kadaster archive (positive class). Kadaster codes the concept as a quadruple <document id, date, dominant tenement id, servient tenement id, text fragment>. The specific vocabulary

specifying the right and its establishment or abandonment is highly variable. The table below gives a few examples of indicative phrases:

Erfdienstbaarheden ("easements")
Erfdienstbaarheid ("easement")
Vestiging Erfdienstbaarheden ("establishment easements")
Vestiging Erfdienstbaarheid ("easement")
Dienend erf ('serving parcel')
Lijdend erf ('restricted parcel')
Rechten van overpad ('rights of path')
Recht van overpad ('right of path')
Recht van weg ('right of way')
Recht van uitweg ('right of exit')
Recht van pad ('right of foothpath')
Wordt bij deze gevestigd ('in this deed will be estblished…')
Bij deze wordt ten laste van (' herewith will be established on…')
Bij deze wordt ten behoeve van ('herewith will be established for…')
Gevestigd het recht van erfdienstbaarheid ('establisment of a right…')

It is not feasible to arrive at a complete list of all possible easements and their formulations in text without reading the whole archived deed. An additional complication in recognition of the concept is that all deeds about a parcel that succeed a positive class document in time quote the relevant fragments from the positive class document verbatim. And a final non-trivial complication is the presence of high amounts of character recognition noise. Some examples are given below (clerical errors in writing or misinterpretation by computer):

ErTdienstbaarheden,
lijdende erTdienstbaar-
lijdende er f d ien s tbaarheden
orfdionstbaarheden
et.rfdienstbaarheden,l
eredienstbuarhnden
glijdende arfdienstbaarheden,
e r f dien stbaarheden
crfdienstbaarhcdcn.
e r f dienstbaartieden
lijdende e r r
lijden d e eri'dienstbaarheden,
*orfdlenstbaarhedön^*
erldienatbaarheden
erraienstbaiirheden ^
. erCdionstbaarheoen
die crTdienstbaarheden
crfdicnstbaarhedcn,
irfdienstbaarheden,
lijdende erxdienstbaarheuen,

Cfjdicnstbaarhedcn
lijdende zrfddienspbaahheden
lijdende ereeienstbaerheden.

3.3 Accurate single class classifier for negatives

The approach used to detect easement descriptions in the documents is a string matching approach based on the nearest neighbour classifier [2]. We collected instances of positive fragments from a small manually annotated training set (4000 documents). The company Textkernel has a very fast string matching application, based on q-gram matching and string similarity computation [3], that has been optimised for millisecond retrieval in text databases with millions of records using techniques inspired by database cache optimisation [4], called FuzzServer. FuzzServer is able to exhaustively measure the similarity of all fragments up to a certain string length from the document against the text fragments from the training set. This makes the approach at the same time robust to linguistic variation and large amounts of OCR noise.

The main opportunity for time savings within the project, was the fact that only 10-20% of the documents are of the positive class. For the positive class documents the system identifies the fragments where easements might be mentioned. The manual coding of the parcel numbers to the easement is still needed. However, it is very difficult to distinguish between truly positive documents and false positives (e.g., containing verbatim citations of positive documents). We did, therefore, focused on creating a very accurate classifier for the negative class. We call this the problem of being sure about what is not there. This classifier has not only features about matches of positive fragments in the training set, but also about negative matches and about the total completeness and quality of the underlying OCR results. By tuning the thresholds of this classifier Kadaster was able to achieve a precision on the negative class of over 99.5% at a recall of 55%. This means that close to one half of the total document archive no longer needs to be read by human beings, because one can be certain that it does *not* contain easements.

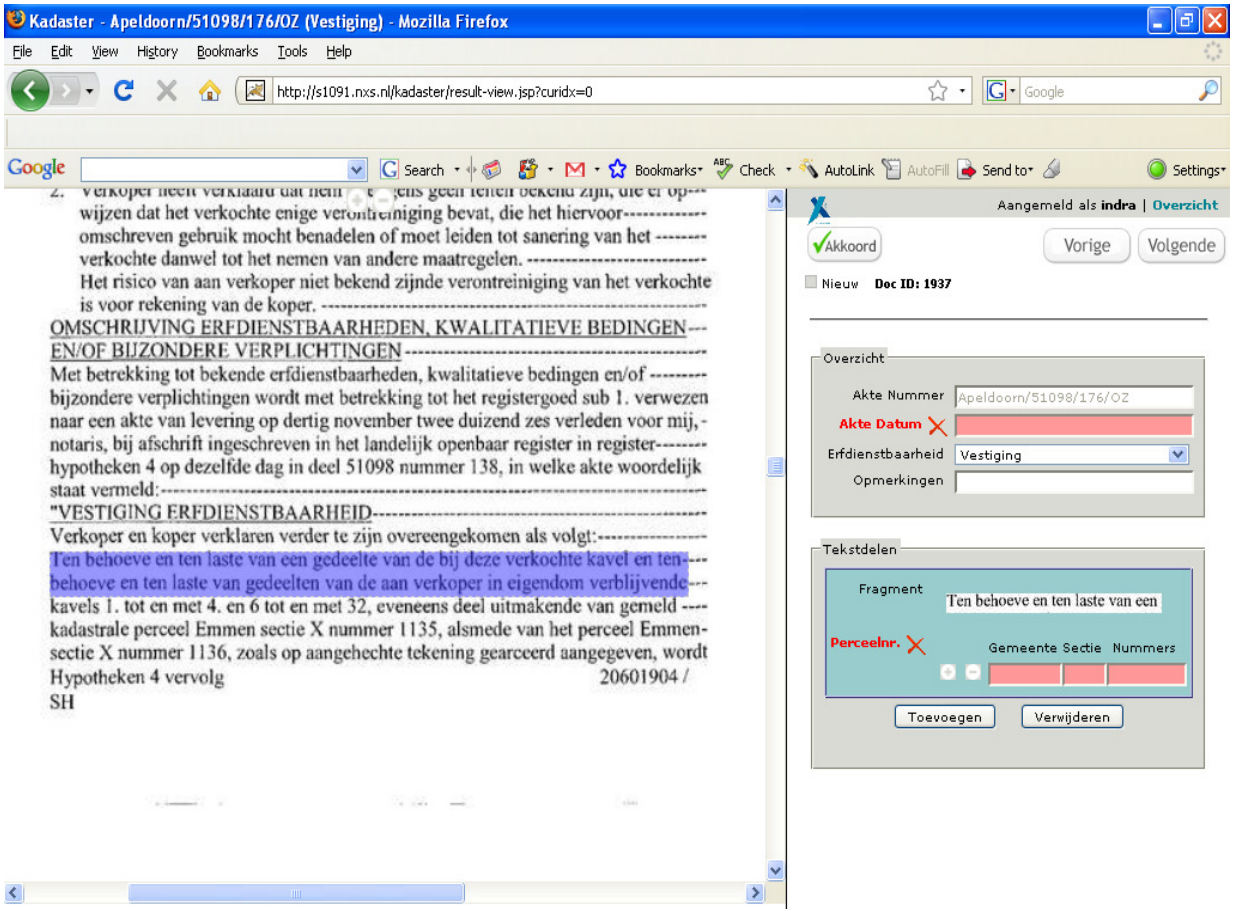

Figure 3. Web page: application for registration of easement relevant text found by computer

The suggestion of the computer, indicating a potential easement is analysed by the operator. For that purpose, a specific application was developed (see Figure 3). The easement can be linked to the appropriate parcel.

4. TECHNICAL CONCEPT

4.1 System architecture

The solution adopted by Kadaster was set up as a pipeline system consisting of Workflow Agents. The agents log the changes made to the pipeline state in a central document database to track the flow of documents through the system. Documents are loaded in batches into the system's spool directory, where the Pipeline Filler agent picks them up and sends them to the OCR and recognition engines. The OCR and recognition engines work in parallel on 56 CPU cores to deliver the needed throughput. The results are delivered into the spool directory for manual coding. Kadaster operators work with Textkernel's web based workflow application, called Sourcebox [5]. In Sourcebox, documents are grouped into small sets for operators in

order to manually code and check. After manual coding, results are checked and quality assurance is performed on a daily basis by blind evaluation of system and human coded output samples to ensure an accuracy of 99.5% on the whole document stream.

Figure 4: picture of the Unisys computer centre: large numbers of servers needed

The workflow in Sourcebox is based on a number of roles: operator, manager, QA-operator, and QA-manager. A regular operator receives only those documents which have been classified as positive by the automatic classifiers. For these documents, the system highlights the fragments with positive matches in the image of the document, allowing fast orientation in the document (see figure 4). The workflow system achieves a further 50% reduction in the time needed for manual coding, bringing the total reduction to approximately 75% of the person hour budget [6].

4.2 Technical system concept

For the project a substantial extension of the technical infrastructure was needed. This is a temporary facility, mainly hosted by Unisys [7]. Especially with gray-scale scanning the bulk of the computer capacity is needed for OCR. This applies to both the processing as well as for the storage capacity. During the production phase, which lasts for approximately four years, the production street is not part of the technical infrastructure of Kadaster but is being hosted by Unisys, including the technical system. The functional management for the (central)

activities in the field of designing and optimizing the word-processing software is the responsibility of Unisys. The functional workflow management, planning and user support is the responsibility of Kadaster.

After these four years a small part of the configuration was left in place to be used to support the primairy task of electronic deed processing. It concerns a small number of servers and relatively little storage capacity.

In Figure 5 is shown the infrastructure needed for externally hosted production line, the various application components are positioned and the facilities required for integrating the internal and external (Unisys) functional environments.

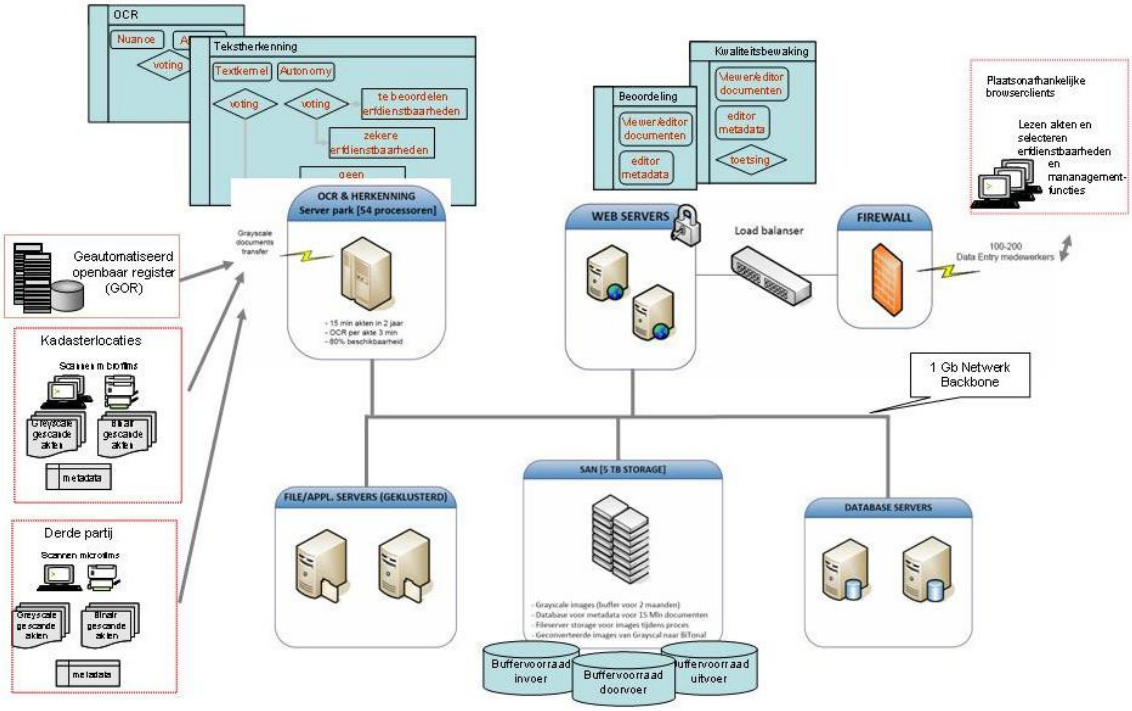

Figure 5: Design of the technical infrastructure (in Dutch language)

All end user applications are available as WEB-applications. For employees in the regional offices this means that the applications can come trough the standard on the Citrix platform present, through Internet Explorer or any other internet browser. This also applies to the management applications. For employees outside the office a normal browser-implementation on a PC with Internet access is sufficient.

5. PARTNERSHIP

The project is almost finished now. The project was a joined effort between Kadaster and Unisys. In a partnership, where mutual trust was as important as a formal contract, both parties had very strong interests in the success of the project: Kadaster's interest is to safe a lot of time and money to digitise large numbers of deeds and to extract valuable information from these documents and Unisys because of acquiring a good reference for future work and the profit.

The first step in 2006 was a pilot which was mainly meant to prove the potential performance of modern applications to recognize characters/text fragments and to extract meaningful information. This step was financially fully covered by Unisys. After the positive result, Kadaster tendered the implementation project, according to EU-tender regulations. Unisys won the tender and a contract was signed. The contract price was based on a specific charging model, where the benefits for Kadaster were the prime input. The charging model based on "click charge": For every document that is processed there was a basic charge. On top of this, there was a bonus scheme:

- For every document that does not have to pass manual processing, an additional charge is made;
- For every document that does need manual processing, a time saving incentive is introduced; and
- For every minute saving in the new process compared to the old process, a fee is charged (a small percentage of the savings Kadaster makes).

The charm of the model is that both Kadaster and Unisys did try to get the best results as both parties financially benefit from less manual processing and from time savings. The result was that both organisations became true business partners throughout the project, having the same objective.

6. DISCUSSION AND CONCLUSIONS

After several attempts to extract information from deeds in the late 1980s and 1990s, Kadaster succeded to make a software application using OCR and text rinterpretation techniques ready for effective and efficient support to Kadaster's largest archiving project.

The extraction of specific information on easements was successful. The quality levels currently achieved by the system were 99.8% quality; far better than manually can be achieved. Now over 50% of documents are processed automatically. On top of that, the documents that require manual intervention are now processed at a 50% faster speed.

The success of the project is boosted by an excellent partnership between Kadaster and Unisys. The agile charging model supported an efficient cooperation and stimulated mutual interest in making an optimal contribution to the project.

ACKNOWLEDGEMENTS

I would like to thank Prof. ir. Paul van de Molen and Miss ir. Louisa Jansen for reviewing this paper.

REFERENCES

[1] Wouters, Rik *: Digitizing Large Volumes of Historic Information and Interpretation by OCR. October 2009*
[2] B.V. Dasarathy (editor). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* Los Alamitos, CA: IEEE Computer Society Press, 1991.
[3] G. Navarro and R. Baeza-Yates. A practical q-gram index for text retrieval allowing errors. *CLEI Electronic Journal*. 1998.
[4] M. Zukowski, S. Heman, N. Nes, P. Boncz. Super-Scalar RAM-CPU Cache Compression. *22nd International Conference on Data Engineering (ICDE'06)*, page 59, 2006.
[5] Textkernel BV. *Sourcebox Whitepaper*, 2008.
[6] Jakub Zavrel, Remko Bonnema, Martijn Spitters, Gert Meijerink, Gerard Mulder. L*arge Scale Text Mining with Highly Accurate Detection of Negatives.* 2009
[7] Arjen Santema et al. *Project Start Architectuur Xtherkenning.* 2008

BIOGRAPHICAL NOTES

Rik Wouters

Holds a degree (MSc) in Agricultural Sciences from Wageningen University, The Netherlands. He worked for five years for FAO, where he had assignments in watershed management and forestry projects in Africa and Asia. In the Netherlands, he worked over 15 years in IT-projects. In 1996 he joined the Netherlands Cadastre, Land Registry and Mapping Agency and was responsible for large and complex IT-projects among which a project dealing with the renewal of major parts of the land registration system. In 2006 he became regional manager for Kadaster International, where he is responsible for the regions Central and Eastern Europe and Asia. In recent years he carried out many review and advisory missions to ECA-countries for the World Bank, the Dutch Government and other donor-organisations.

Gert Meijerink

Holds a MBA degree (MSM) in Strategic Management from the Lemniscaat Business School, The Netherlands and an engineering degree in Advanced Electronics from Rens and Rens. He worked for five years for DICOM, where he had several assignments in the document imaging industry in the Benelux and France. For BancTec he worked 9 years in several roles including 5 years international product management and sales. In the Netherlands, he worked over 15 years in Electronic Content Management-projects. In 2007 he joined Unisys and was commercially responsible for large and complex IT-projects in the Dutch government and the justice and public safety area. In 2008 he closed the erfdienstbaarheden project with Dutch Kadaster.

Rob Vaandrager

Holds a degree Business Administration from Keele University, the United Kingdom. He works since 1979 at Netherlands Cadastre, Land Registry and Mapping Agency. After seven years he became Manager of a production team in Rotterdam. Over twenty years he was team leader of various teams all over the Netherlands and he moved up to the position of deputy director of the local office of Eindhoven. The last three years he holds the position of Senior Project Manager and in this position he is responsible for the programme of "unlock and update easements".

Jakub Zavrel
Holds MSc-degrees in Artificial Intelligence (Utrecht) and Computational Linguistics (Amsterdam). He worked for several years at the University of Tilburg and University of Antwerp doing research in Machine Learning of Natural Language and published extensively about classification based machine learning methods for sequence labeling. More attracted to applications that to theoretical research, he was one of the founders of Textkernel in 2001. As CEO, he has since then led its growth to one of the leading Dutch language technology companies.

CONTACTS

Ir. Rik Wouters
Netherlands' Cadastre, Land Registry and Mapping Agency
Hofstraat 110
7311 KZ Apeldoorn
THE NETHERLANDS
Tel. +31 88 183 3520
Fax +31 88 183 2074
E-mail: rik.wouters@kadaster.nl
Website: www.kadaster.nl

Gert Meijerink MBA
Unisys Technology, Consulting and Integration Solutions
Tupolevlaan 1
1119 NW Schiphol-Rijk
THE NETHERLANDS
Tel. +31(0) 20 526 7500
E-mail: gert.meijerink@nl.unisys.com
Website: www.unisys.com

Ir. Jakub Zavrel
Textkernel BV
Nieuwendammerkade 28A-17
NL-1022 AB Amsterdam
THE NETHERLANDS
Tel. +31 20 494 2496