Digitizing Large Volumes of Historic Information and Interpretation by OCR

Rik WOUTERS, the Netherlands

Key words: archive, historic, cadastre map, field sketch, OCR, field sketch

SUMMARY

Since a long time the Dutch Land Registry, Cadastre and Mapping Agency (in short Kadaster) delivers digital information to customers. Already in the early 90-ties 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 was provided.

At present Kadaster is scanning old paper documents: field sketches and historic cadastre maps. In total some 9 million documents are scanned. In 2010 this kind of historical information can be made available through the internet.

An other project deals with the retrieval of information concerning encumbrances, restrictions, servitudes, and the like. By means of text recognition tools more then 95 of this information is extracted form old deeds. Also this information is available on the internet. The paper describes what procedures, approach and techniques have been used to make a next step in e-services provided by Kadaster.

Digitizing Large Volumes of Historic Information and Interpretation by OCR

Rik WOUTERS, the Netherlands

1. INTRODUCTION

1.1 General

When buying or selling registered properties one is legally obliged to register the accompanying notarial deeds in the cadastral system. This system ensures that the source of our details and – accordingly – the information are kept up-to-date at all times. The Agency for Land registry, Cadastre and Mapping in short Kadaster, keeps registers by law. These registers consist (among other things) of notarial deeds relating 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 that indicate the rights that are related to the registered properties (legal status).

The most important details from the deeds referred to above that 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 look up the original deed in the public registers, or to have this done. The cadastral register also functions as an index for the public registers. It provides a clear overview for each parcel of, for example, the rights related to a parcel, who the owner is 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 virtually directly available. This is of great importance when we consider that in excess of one million real estate transactions take place every year

1.2 Objectives concerning historic information

The land information system reached a high standard of quality and meets by the time largely what is asked for in society: a reliable, transparent and internet-based cadastral information system. Because of the level reached new opportunities for delivering services to the society emerge. The 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
- centrally organised IT-infrastructure

The main challenge of the Agency is now to make 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 cadastre maps (see chapter 2). There is a growing demand for information from historical deeds, more specifically information on easements, servitudes and the like. For that old paper deeds are scanned and the text is searched for easements: important information for the buyer as it comes to the purchase of a real estate object. This information will be available in the cadastre registration from 2010 (see chapter 3).

2. Digital archiving

2.1 Introduction

In 2005 Kadaster started the archiving project. Prime objective of the project is to open up and improve completeness, improve conservation to secure an optimal use. This is to avoid bottlenecks, arising from analog archives to be resolved.. Digital access should be location-independent with sufficient speed and performance. A paper archive is according to law not anymore required. The handling of the original documents must be arranged in good agreement with the National Archives and the Regional Archive Centres (RAC).

In this project a total of 5 million historic parcel maps and 6 million field sketches (see figure 1) will be digitize. By 2010 all archived documents should be in digital format available. The total available budget is € 22 million.

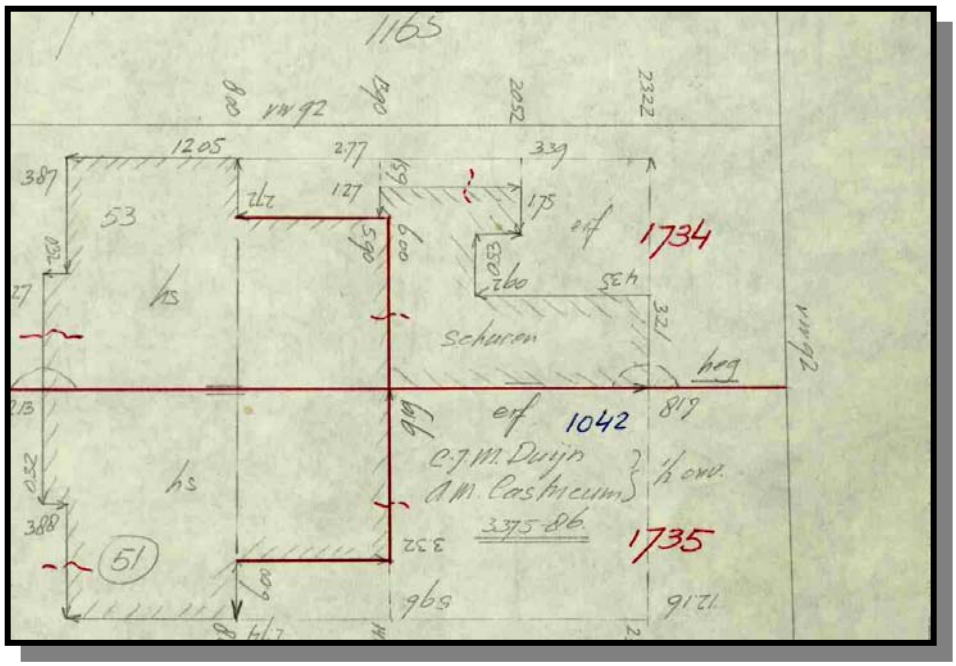

Figure 1: old field sketch

2.2 Pilot

Digitizing of large volumes of documents is a large, time critical and costly operation. A pilot was therefore conducted, to gain experience, develop indicators and the to choose the best method for an adequate digitization, at the lowest possible cost. This pilot was conducted in the office of Alkmaar, with the cooperation of two company's renown because of their ability. Castricum was the pilot municipality from which town all field sketches and historic parcel maps were prepared, cleaned and then scanned, indexed and verified. The pilot proofed to be very valuable and has lots of useful information added to the approach of the project, but also for any digitization of the other archives. Conclusions include that use of surveyors for the upgrading of data is necessary to achieve the desired quality. It further includes that check list are indispensable for checking completeness of outsourcing work. Further it became clear that text recognition technology is not usable for interpretation of information on the documents. In the local offices a supervisor / archive manager is required for overall coordination, uniformity of procedures and knowledge of local exceptions. Outsourcing of scanning and indexing implies less risk (it is an only one time activity, lack of knowledge / experience in Kadaster) and will be more efficient.

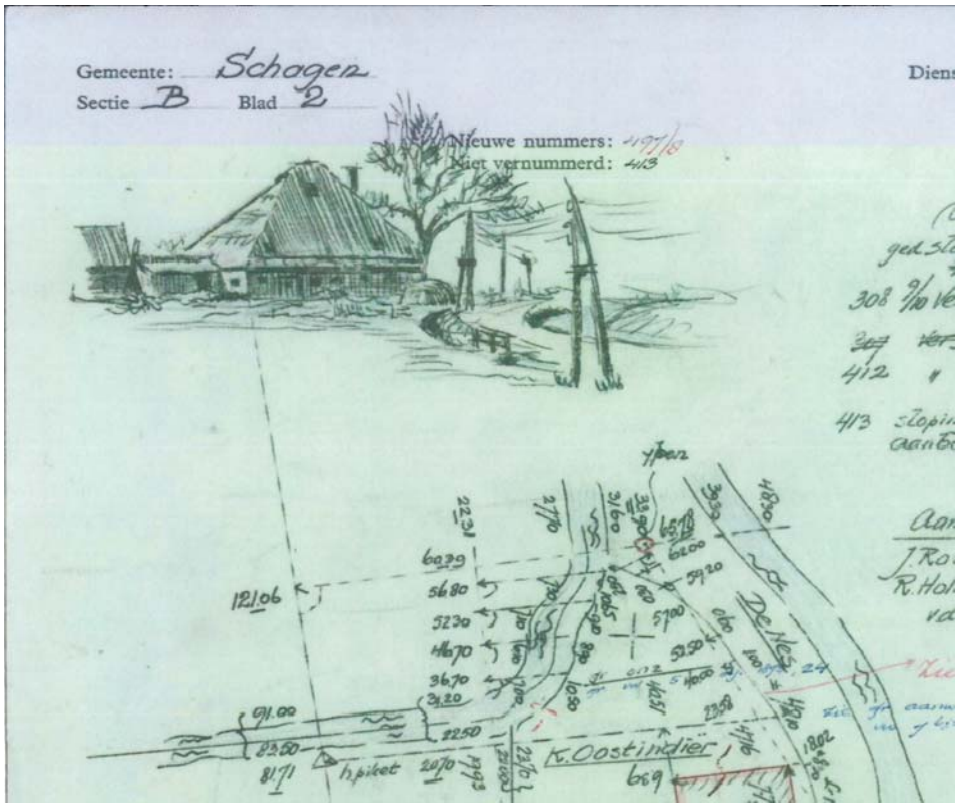

Figure 2: Part of original file sketch.

During the pilot instructions have been developed for the preparation and cleanup, and to provide solutions for the types of exceptions that were found in the pilot and as result of the checking of the archives. In addition, guidelines for scanning and indexing were defined. In addition to instructions, standards and norms were developed, used for calculating the required capacity and for control / management of various activities.

For the consultation of the digital field sketches and historic parcel maps one need a good information system like good performance for querying and print performance, dynamic zooming, etc.

The quality is also set in an instruction. The idea is to randomly conduct quality checks to be carried out on image quality of the scans and the indexing by a fixed designated team. In addition, automated completeness checks are performed based on the inventory lists. In the pilot is considered the method of cleansing: a) in advance with limited administrative staff, followed by an indexation surveying employees or b) full cleanup in advance by surveyors and indexing by administrative forces. The latter appeared to preferable because it is cheaper, less surveying capacity required, a more complete archive is obtained, and allows a better control on the outsourcing.

Figure 3: moderate quality of historic paper maps

The indexing itself and the control on the indexing, implies a lot of screen manipulation and thus a risk of RSI. This requires the necessary attention. By outsourcing these activities, this risk will be moved to the scanning company.

Project Approach

Based on the experiences of the pilot a well-defined process has been described that consists of three phases: a) preparation/cleanup, b) digitizing/indexing, c)monitoring/implementation. The first phase is done in-house by surveyors and (temporary) administrative forces, whereas the second phase is done by outsourcing and the last stage again in-house by (temporary) administrative staff.

The preparation/cleanup includes an inventory, followed by ordering, making complete and making ready to scan the documents. This step ends with a completeness check. To ensure a smooth process of scanning and indexing, it is crucial that this step is done with great care and is carried out by our own experienced geodetic staff, where possible, supplemented by (temporary) administrative forces. After the cleanup is a strict archive management must be in place. This first phase requires the most internal capacity, which will create a tension with the available surveying capacity. For each branch office a separate implementation plan will be established, including all relevant aspects (volume of work, available staff, etc).

By outsourcing it is expected that the digitizing (scanning) and indexing (including the control on the indexing) will be more efficient, because of competition in the market and the

opportunity for transfer of work to low-wage country. Later on it proofed that the activities were most efficiently (cheap and high quality) performed by the National Tax Authorities. Additional advantage was that a European procurement was not required.

Figure 4: scanning and indexing facility

The monitoring/implementation include a manual random monitoring of image quality and the indexing. In addition, automated checks for completeness of documents and parcel numbering are done. Implementation can be done with (hired) administrative staff at a central location. Experience learns that, despite the attention to quality of this process, some time after-care will be needed. Essential is a fast and direct control, for subsequent delivery to the RAC's to make it possible to create documents quickly and to have them digitally available for regular work. For the transfer to the RAC's directives will be agreed with the National Archives on the method of transfer. For a back-up option it is desirable to retain the paper field sketches and historic parcel maps in the National Archives. This must be laid down in a gentleman's agreement between Kadaster and National Archive. After certainty is obtained about the quality and an effective query-system is in place, the specific filed sketch can be destroyed.

2.3 Organization of the project

The implementation of activities is organised a project. A steering group will assess the plans and advise the Executive Board in decision-making during the project. The director of the GEO-department is project responsible. The implementation of the work is in the hands of a national project manager, assisted by two team leaders. He coordinates all project activities and is accountable to the project manager and the steering committee. In each regional office

contact persons are appointed. For the control of regional bound activities, local archive managers are assigned with specific knowledge of the archive. These managers together with the team leader will manage the daily work in the cleanup and preparation of records for scanning. In order to guarantee the quality of the work, the implementation must be done with sufficiently qualified and available resources. From the central office domain and ICT knowledge staff must be available for the project. They should receive sufficient priority from their department to support the project. The control and the preparation for delivery to National Archive and RAC's will be assigned to a central support team.

2.4 Developments

For the future, the Kadaster is the manufacture of digital field sketches and digital cadastre maps. Until than, the paper archive of field sketches and historic parcel maps will still grow. This growth will be included as an addition after the first scan round. Unlike the field sketches, the digital historic parcel maps field, is already available (parcel history), but the disclosure of the maps has yet to be arranged. It was essential for the success of the project, that an effective query system was available from the start of scanning process.

The coming years will critical with respect to the deployment of internal surveying capacity for multiple projects and regular work. The Director GEO will have to choose for which activities own geodetic staff will be deployed. From the project it was strongly requested to use own surveyors because of the desired quality.

Planning

The planning is based on the standards that were developed in the pilot, and the time embedded in the timetable for the removal of FB. Especially the step digitizing and indexing asks many actions and capacity in a short time frame. In this period, the required capacity of the geodetic staff is greatest. In addition to project management and the domain and ICT experts (central department), some 114 man years are needed in each of the regional offices. Assumption is that total volume of documents amounts to 3.5 million field sketches and 6 million historic parcel maps. The planning has also in a financial component. The estimate reaches an amount of almost € 21 million. For monitoring progress and the actual cost, a time accountability model has been developed, where insight is given for each activity and for each regional office. To ensure progress timely and appropriate adjustments will be made.

2.5 Risks and measures

In the project a number of risks have been identified. There is a possible link with the digitization of other analogue archives. Therefore, in case of outsourcing an activity, a clause is included in the contract for extension of activities.

The availability of ICT facilities and resources is a prerequisite for a successful implementation. The Director Services (which includes ICT-department) and the Steering Committee are informed about this critical issue. Provide own surveying capacity is

particularly critical for the quality of the process in the field. This should be in good agreement with the GEO Director and regional directors.

Outsourcing of the working packages is important. The logistics are important to. The whole operation of outsourcing will be organized and monitored in a core team.

Timely availability of an effective query system is a very critical point. It is a condition for acceptance in mainstream work and assures that documents do not need to be return to the regional offices.

During the project creation of a separate scan team is for any growth at the end of this project, or the operation of a system with digital assistance cards and / or field work. The transfer of archives to the RAC's is subject to strict requirements. The undersigning of a new Gentleman's agreement will straiten this out. It should have the technical requirements that both digital and analogue supplies must meet, are included.

Communication

A good base to create acceptance is to have involved all stakeholders and to inform them as much as possible. It is important to communicate through various channels. On the Intra-net a dedicated project page is available as well as a special forum for the project out reach. This in principle creates a dialogue with readers. The project has its own e-mail account available for comment, questions and the like for internal staff of Kadaster. Finally, on regular basis articles are publish in the Kadaster internal magazine and meetings are hold for the exchange of knowledge and experience. External communication is only foreseen as the project is well under way and results and experience are available to be shared. Presentations and publications in scientific journals are an appropriate means.

3. TRACING EASEMENTS

3.1 Introduction

When purchasing or valuating a piece of land or a dwelling it is of interest to know whether rights of third parties on the parcel rest. For example, a right of way or a right of view. An easement is a burden that a property ("the serving yard") is loaded for another property ("the ruling yard"). Regularly questions are asked by clients about easements, encumbrances, servitudes, etc. In that case the client can request Kadaster to conduct an easement investigation a so call "erfdienstbaarhedenonderzoek".

In case of an easement investigation, Kadaster examines if there are documents, containing information intending to establish an easement on the serving yard. It is also possible that the Kadaster examines also whether there are easements in which case the parcel is not the serving, but as ruling yard is involved.

The result of the investigation is a statement that declares that all the documents which relate to the plot as serving yard (possibly in combination with a plot as ruling yard) to a certain date were examined. It also includes the results passages in the deed(s) that may relate to the requested easement. If a deed is partly readable, the text of that the easement can not be typed. In that case a free full copy of this deed is provided.

The research into easements is time consuming and means a lot of manual work, reading old deeds. The work is location bound because in all cases one has to refer to the paper archives. Especially notaries devote a lot of time to this kind of research.

The main objective of the project is to extract the easements from the deed and register the easements in the cadastre registration in order to make them easy accessible. The new system will improves quality and reduces processing time of notarial deeds

Challenge is to extract the required information automatically, without (a lot of) human interference. That is why Optical Character Recognition (OCR) techniques were introduced.
The consequences of the wrong tracking or losses of information on easements are very large. The Quality Charter of the Kadaster is also included that the quality of this information has a high reliability.

3.2 Project

In 2006 a project has been started build a fully automated system to retrieve all the easements from the deed and register them in cadastre registration. A separate retrieval system for easement-information was developed.

In the preparatory phase a lot attention was paid to the overall process. In about a dozen use cases the elementary functionality was described. Because of the large volume of documents

(over 15 million), the in many cases bad readability of the deed, and the painstaking work involved a automated solution and the use of OCR was required

Within this project, Unisys is responsible for processing 15 million records containing references to easements. An easement is a burden which a property - under the yard - for the benefit of another property - the ruling yard - has been burdened. The easement, as recorded in Book 5 BW (Dutch Law), is a limited right: a right which is derived from a more comprehensive law (e.g. property, leasehold or building). Because of a load for and charged to a yard to speak, is expressed that the easement called upon law, which separated from the yard can be, and that right passes with ownership of the dominant and serving yard.

By the use of text recognition technology, it is possible for 50 percent of the deeds, to exclude them from a manual process with a quality rate of 99.5 percent. This solution will save Kadaster a lot of cost. In a later stage, these digital data can be used to inform customers about easements in significantly faster manner.

3.3 Proof of concept

At the end of June 2009, Kadaster finalized successfully the "proof of concept" period. In the recent period it was demonstrated that the solution with text recognition of easements, meets the requirements of Kadaster.

Figure 5: example of a deed with section on easements

It approved to be a real cost saving process compared with the manual reading of documents, reducing processing time and increase quality in the search of the deed tracing easements with a success rate of 99.5 percent.

During the pilot already some 10,000 deeds were processed successfully. Even deeds which are not of high quality could be processed (see figure 5). In additional research it appeared that the OCR-solution proved to trace easements with a higher accuracy than manual processing by reading and checking the deed.

The Kadaster has, based on the outcome of the 'proof of concept' confidence in the solution. This solution also showed the possibility to build easements system and thus increase legal certainty in the social movement in real estate promotion. In the Netherlands the first time that such a project on this scale is performed.

This solution offers many opportunities for other organizations where large conversion have been postponed because the solution is not cost effective or of to low quality.

3.4 Technical component

For this project, we make use of a solution in which two software components are implemented: Text Kernel and Autonomy. Text Kernel provides solutions for automation of data-entry processes to, for rapid checking of relevancy from large quantities of texts and to classify these texts to relevancy. These texts may be in databases or in archives or from coming from the internet. In case of the Kadaster the source are scanned images of microfilm made for text recognition through OCR. Autonomy Corporation plc. is a supplier of infrastructure software and is advance in the development of Meaning Based Computing. Autonomy's technology recognizes the concept and context of any form of digital data,

including unstructured information.

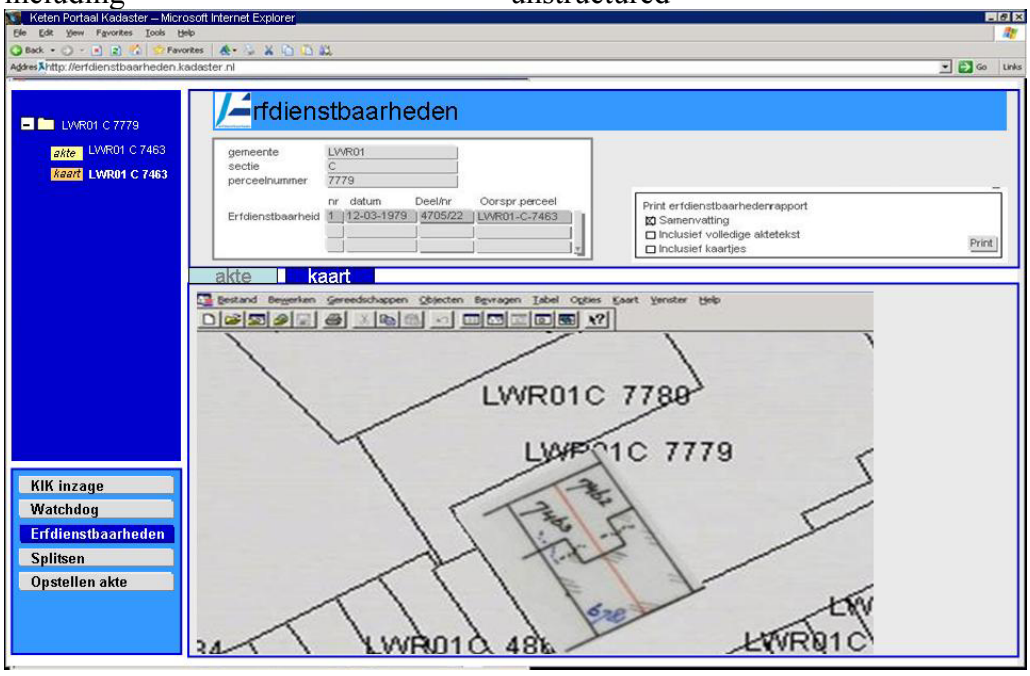

Figure 6: Screen shot of retrieval system

By implementing a solution based on both Text Kernel and Autonomy, we are capable to achieve exceptionally high quality in the automatic recognition of deeds, which includes no reference to easements and from difficult-readable documents. This quality is crucial for the realization of an automated easement system which supports an even better service of Kadaster to its customers.

REFERENCES

Information was obtained from

1. www.kadaster.nl
2. Santema, Arjen: Know your limitations (2009)
3. Volkskrant: Unisys bewijst de waarde van Tekstherkenning (article) (2009)
4. Koning, Herman de: Project Digitizing Historic Parcel Maps en Field Sketches (2009)

BIOGRAPHICAL NOTES

Mr. Wouters gained his MSc in Agricultural Sciences at Wageningen University (Netherlands) in 1982. After his study he worked 5 years for the 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. From 1996 he joined the Kadaster and was responsible for large and complex IT-projects. His last project dealt with the renewal of major parts of the land registration system. In April 2006 he became regional manager for Kadaster International, where he is responsible for the regions Central and Eastern Europe and Asia.

CONTACTS

Ir. H.J. Wouters
Openbare Registers en Kadaster
Hofstraat 110
7311 KZ Apeldoorn
NETHERLANDS
Tel. + 31 55 5285748
Fax + 31 55 3557362
Email: rik.wouters@kadaster.nl
Web site: www.kadaster.nl