Exploratory search in insolvency jurisdiction

By T. Akkermans – 11323671

Supervisor: Dr. Maarten Marx

**abstract**

The Dutch government is taking general steps to disclose data financed by public funds. The law states that jurisdiction should be public but in practise often only selected court decisions are published. There are subfields such as insolvency law where defined court actions must be published. The government supplies minimal functionality to retrieve these court publications and accompanying curator reports *if* the curator filed them. There is lacking structure in this data to facilitate exploratory search. Furthermore much of the data resides in PDFs and text fields which makes binary keyword search and full-text search unavailable.

In this thesis project a complex information system prototype will be proposed to 1. retrieve obscured information and knowledge from the open Dutch insolvency data and 2. add structure to this data to facilitate exploratory search and advanced text search for purposes of exploration, retrieval, analytics and process mining.

# Personal Details

Email: [toak@quantleap.nl](mailto:toak@quantleap.nl)

Supervisor email: [m.j.marx@uva.nl](mailto:m.j.marx@uva.nl)

Github: <https://github.com/quantleap/is-ds-thesis>

# Debate on disclosure of jurisdiction

Rechtspraak.nl has been publishing jurisdiction since december 1999 but only a tiny fraction (0.9% in 2004 and 2% in 2011 [a1]). The criteria for pucblication have been formulated in 2012 [a2] but courts are responsible for the publications but can determine the criteria for publication themselves.

According to the Dutch Insolvency Law [a3], courts are obliged to inform the general public about companies insolvency and suspension of payment and at defined moments. Next to the *Staatscourant* this is done in the *Centraal Insolventieregister* (CIR).

Curators decide if they upload their reports to the CIR. There are many cases however where curators make the report available on their own sites but not in CIR.

[a1] mentions both arguments in favour of publication: scrutiny of the judicial system by the general public for all types of cases and availibility of information for research, as well as against publication: the additional workload in screening and anonymising of verdicts and that the increase of verdicts causes search to return too many valid results. The arguments against publication were stated by the *Nederlandse Vereniging voor Rechtspraak* (NVvR).

[a1] <http://www.rechtsvordering.nl/rechtspraaknl-moet-alle-uitspraken-publiceren>

[a2] <https://www.rechtspraak.nl/Uitspraken-en-nieuws/Uitspraken/Paginas/Selectiecriteria.aspx>

[a3] <http://wetten.overheid.nl/BWBR0001860/2017-04-01>

# Programme to modernise practising law

Quite recent, in July 2016, a law was passed [b1] to simplify and digitise civil law and administrative procedural law. The corresponding programme is called *Programma Kwaliteit en Innovatie* (KEI) [b2]. A secured webportal named *Mijn Rechtspraak* will provide involved parties access to digital case information and progress. Curators and judges will start communicating digitally using this portal [b3].

[b1] <https://www.rechtspraak.nl/Voor-advocaten-en-juristen/modernisering-rechtspraak/Paginas/Programma-Kwaliteit-en-Innovatie-van-de-Rechtspraak.aspx>

[b2] https://www.rechtspraak.nl/SiteCollectionDocuments/Tijdlijn-KEI.pdf

[b3] <https://www.rechtspraak.nl/Voor-advocaten-en-juristen/Reglementen-procedures-en-formulieren/Civiel/Insolventierecht/Paginas/Digitaal-toezicht.aspx>

# Research question

The goal of the thesis is to build a prototype information system that can uncover various forms of relevant information and knowledge from the open insolvency data that consists of XML and PDf reports.

The relevance of the information will be validated by potential users – the stakeholders in the process of company insolvency, ranging from creditors to the judicial system participants, courts and curators, to the general public.

The main research question is how we can employ NLP techniques to uncover hidden structures in the PDF reports and text fields in XML to retrieve relevant user information.

Subquestions are:

* How are insolvency cases processed by the judicial system?
* What are their predicted outcomes?
* Who are the stakeholders and what are their interrelationsships?
* How is balance of the countervailing powers of courts and curators and does it impact the debtor and creditors interests ?
* How can stakeholders be better served by the judicial information system?

# Query functionality

The existing web query platform[[1]](#footnote-1) of the Centraal Insolventie Register (CIR) is meant to retrieve insolvencies by:

1. Insolvency case properties, either:
   1. Id and court
   2. Publication id, or
   3. Period, court and publication type
2. Company properties, either:
   1. Name,
   2. KvK number (Chambers of Commerce), or
   3. Postal code and house number

There is no way of aggregating results, filtering on other properties or on entities other than insolvency case or company.

The following additional entities and properties could be filtered and grouped by:

1. Insolvency Case
   1. Stage: begin date – surseance – acknowledged creditors - end date
2. Company
   1. SBI number (Standaard Bedrijfsindeling, source: CBS)
   2. Location (address to lat/lon, radius search)
   3. Legal entity
3. Report
   1. Nr of pages
   2. Type: [progress | end | financial] report
4. Court
5. Administrator
   1. Member insolad
6. Administrator office
   1. address
7. Bankruptcy judge
   1. Court (only one court per judge – they rotate courts/disciplines)
8. Creditors

A graph structure can be build out of Court, Administrator, Admin. Office and Case to investigate the allocation of cases from judges to administrator (offices).

# OCR engines

The entity extraction and text search depend strongly on the quality of the OCR-ed reports. This section describes several available programmable OCR engines and their test results.

The number of pages to scan must be assessed in case of a commercial option. Support for the Dutch language is mandatory.

1. Apache License
   1. Tesseract
   2. OCRopus
2. Commercial, student perks:
   1. ABBYY: <http://ocrsdk.com/for_students/>
   2. Adobe?

To Excel:

1. ABBYY

<https://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software>

# Literature and web research

This chapter summarized literature and web content related to the thesis project.

## Web

[1] <http://www.rechtsvordering.nl/rechtspraaknl-moet-alle-uitspraken-publiceren>

## Literature

[2] Extracting Databases from Dark Data with DeepDive - Zhang, Ce and Shin, Jaeho – Stanford University.

[2] An Overview of the Tesseract OCR Engine - Ray Smith Google Inc.

[4] <http://www.project-consult.de/files/AIIM_IW_Dark_Data_in_Capture_Processes_April_2016.pdf>

[5] “PST files and ZIP files account for nearly 90% of dark data” - White paper by IDC Estimates.

[6] Shedding Light on Dark Data – Waters technology

# References

[1] Dataportaal van de Nederlandse overheid <https://data.overheid.nl/>

[] De Rechtspraak – zoeken naar uitspraken. https://www.rechtspraak.nl/

[2] Faillissementswet - <http://wetten.overheid.nl/BWBR0001860/2017-04-01>

[] PoliticalMashup - <http://www.scienceguide.nl/201301/het-woord-is-aan-de-kamer.aspx>

# Methodology

The data that will be used is the *insolventie register* data [1] which is made available through a webservice. A datastore will be set up to hold the XML files as well as the PDF reports. The PDF reports will be OCR-ed and full text will be stored. Additional relevant data sources, such as maps, KVK, court and curator information will be coupled.

Analytics will be developed to extract higher order information and knowledge from the data. NLP routines will be used for the full text reports. Graphs will be build for the parties involved, both companies as well as curators. Process mining will be done on the court cases to see how the cases develop over time and to provide predictions on their settlement.

The knowledge and information obtained will be fed back to the user in the form of reports and visualisations. The results will be assessed in accordance with an informatin relevance study done with potential users.

# Risk assessment

The following risks and resolutions are being considered:

Risk: There are no potential users willing to cooperate in interviews

Backup: Information relevance will be inferred from literature and web

Risk: Open source OCR will not be good enough

Backup: Commercial software can be used – in last resort on a subset of the data.

Risk: The system complexity becomes too large with many features.

Backup: There can be focus on one or two information structures.

Risk: The user interface may eat up too much time from the information measures.

Backup: There can be focus on certain measures. Or a certain visualizations if they provide more information / knowledge.

# Project plan

The following plan is a broad step schedule of tasks and deliverables:

|  |  |  |
| --- | --- | --- |
| Period | Task | Deliverable |
| Week 1 | User interviews | User information relevance study |
| Week 2 | Design system architecture | System design |
| Week 3 | Data analysis and enrichment | Available data overview |
| Week 4 | Data and process mining, building predictive analytics. | Knowledge retrieval based on selected |
| Week 5 | data and processes |
| Week 6 | Predictive models based on selected |
| Week 7 | data |
| Week 8 | Evaluating measures and visualisation of information | User information retrieval scores |
| Week 9 | D3 visualisations |
| Week 10 | Wrapping up development and writing final report | Final code base |
| Week 11 | Final report |
| Week 12 | Thesis defense | - |

# References

[1] <http://insolventies.rechtspraak.nl/>

[2] <https://data.overheid.nl/>

[3] <http://opendatahandbook.org/>

# Feedback op Proposal

1. Ik stel voor dat je meteen die hele site leegtrekt, en dat eigenlijk nu al doet en kijkt of je compleet bent. Je gaat daar toch niet mee wachten tot de 3de of 4de week? Als het nou niet lukt?

Ik heb middels webservices alle data opgehaald en voeg er dagelijks de nieuwe data aan toe. Een regelmatige compleetheid check moet nog worden toegevoegd.

1. vreemde dingen, zoals OCR. Wat ik aan pdf's op de site vond was allemaal prima doorzoekbaar met control F, dus wat is het OCR probleem?

Ik heb een check gemaakt hoeveel rapporten zijn gescanned en hoeveel geconverteerd. Een gescanned rapport is een collectie TIFs en kan niet doorzocht worden op tekst. Een geconverteerd rapport is van bv Word formaat omgezet naar PDF en bevat wel tekst elementen. De ratio converted/scanned is stijgende maar grote portie is nog steeds gescanned. Er is ook een onbekend deel wat verder uitgezocht en geclassificeerd moet worden.

NB

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sum of count** | Column Labels |  |  |  |
| quarter | unknown | converted | scanned | Grand Total |
| 2015-01 | 14% | 39% | 46% | 100% |
| 2015-02 | 13% | 38% | 49% | 100% |
| 2015-03 | 13% | 35% | 51% | 100% |
| 2015-04 | 16% | 33% | 51% | 100% |
| 2015-05 | 14% | 37% | 49% | 100% |
| 2015-06 | 12% | 39% | 48% | 100% |
| 2015-07 | 13% | 36% | 50% | 100% |
| 2015-08 | 13% | 38% | 49% | 100% |
| 2015-09 | 14% | 39% | 47% | 100% |
| 2015-10 | 13% | 40% | 47% | 100% |
| 2015-11 | 15% | 42% | 43% | 100% |
| 2015-12 | 12% | 40% | 47% | 100% |
| 2016-01 | 14% | 45% | 41% | 100% |
| 2016-02 | 12% | 46% | 41% | 100% |
| 2016-03 | 14% | 47% | 39% | 100% |
| 2016-04 | 15% | 45% | 40% | 100% |
| 2016-05 | 22% | 47% | 30% | 100% |
| 2016-06 | 13% | 52% | 35% | 100% |
| 2016-07 | 14% | 52% | 34% | 100% |
| 2016-08 | 12% | 53% | 35% | 100% |
| 2016-09 | 13% | 55% | 33% | 100% |
| 2016-10 | 14% | 56% | 30% | 100% |
| 2016-11 | 13% | 60% | 27% | 100% |
| 2016-12 | 23% | 55% | 22% | 100% |
| **Grand Total** | **14%** | **44%** | **41%** | **100%** |

1. De evaluatie is zwak, mede omdat het zeer onduidelijk is wat je gaat maken. Een paar interviews is echt niet voldoende.
2. Je moet ook kwantitatief evalueren. Dat kan heel makkelijk. Bijvoorbeeld:
   1. Ben ik volledig (heb jij alles in je prachtige systeem dat de overheid publiceert?), en haal je dan ook alles eruit?
   2. Ben ik precies? Dus verander je geen informatie tijdens je "open" maken
   3. Dit soort precision en recall maten zijn belangrijk en het begin. Daarna ga je naar het "nut" kijken.

Welke search in welke prioriteit te ontwikkelen? Eigen prioriteit of die van potentiele gebruikers?

Te doen: kwantitatieve evaluaties.

4A: Kan de volledigheids check worden gedaan obv insolventie kenmerk (volgnummer) ?

1. Wat je gaat doen blijft zeer onduidelijk, ook in je planning. Ik zou dat veel concreter opschrijven. Dat is ook fijn voor jezelf.
2. Je literatuur is niet serieus. Wat linkjes is geen gronding in de literatuur.

Terechte opmerking, ik was meer gefocussed op praktische toepassingen en minder op research.

# Entity Extraction and Search:

In het CIR register zijn een aantal entiteiten gedefinieerd:

* Rechtbanken
  + Rechtbankadres
* Rechter-Commissarisen
* Curatoren
* Curator Verslagen
  + Voortgangsverslag
  + Financieel Verslag
  + Eindverslag
* Rechtbank Publicatie
  + Soort code
* Insolventen (E)
  + Naam
  + KvK nummer (FK)
  + Bedrijfsadres

Met behulp van externe data van Insolad en Mr-Online kunnen daar nog de volgende entiteiten en/of eigenschappen aan worden toegevoegd:

* Insolad status Curator
* Kantoor Curator (E)
* Begin/eind loopbaan (P)

Het KvK kan de volgende informatie leveren:

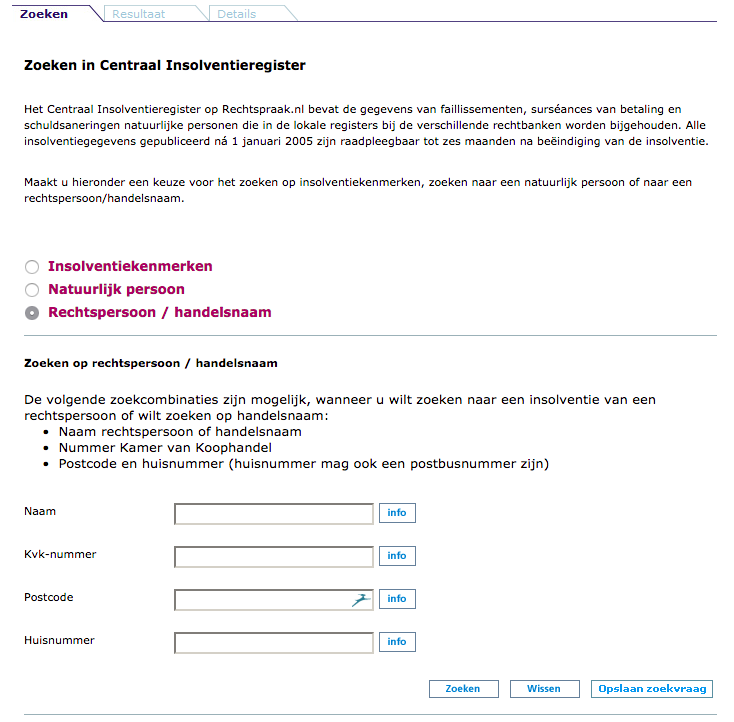
* Insolventen – Bedrijfssector in de vorm van SBI code (P)

Met behulp van rechtbanken , curatoren of debiteuren (allen moeilijk) zijn de volgende verrijkingen mogelijk:

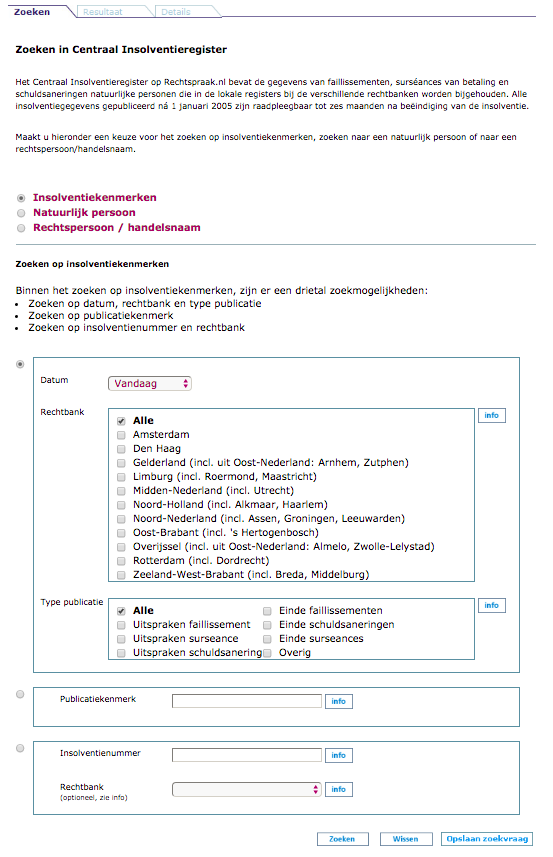
* Debiteur
* Vorderingen (E) (van Debiteur naar Curator)
  + Bedrag, Btw, Preferent (T/F)
* Urenverantwoording (van Curator naar Rechter-Commissaris)

# Huidige zoekfunctionaliteit

* 1. Op rechtspersoon / handelsnaam



* 1. Op Insolventiekenmerken



1. http://insolventies.rechtspraak.nl/ [↑](#footnote-ref-1)