Data driven supervision in insolvency procedures

**Student: Tom Akkermans**

**Supervisor: Dr. Maarten Marx**

**Abstract**

After a company files for bankruptcy at the court, a court committee will appoint a curator to settle the bankruptcy. The curator dissolves the company by monetizing the company's estate and redeeming its creditors. A court’s judge supervises the curator. Both these processes: curator appointment and supervision, have been topic of research[1-2], have raised concerns in the media[13-21] and have lead to court cases for not being transparent and/or not being adequate.

In this thesis an information system prototype will be constructed to shed light on these processes. The system takes in the structured and unstructered data of the insolvency procedures using open and publicly available data sources[22-28]. From this data the system builds and annotates a complete linked entity structure.

Using a web application interface, a stakeholder can query and explore the networks of linked entities and insolvency case data in order to provide transparency on the insolvency processes.

# Personal Details

UvA Student ID: 11323671

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>

# Research problem and corresponding sub questions

The main goal of the thesis is to design a prototype information system that provides transparency of insolvency processes through query and exploration functionality. In particular it should answer the question: with which level of correctness and completeness can the information system answer the following questions:

1. **On the adequacy of supervision of insolvency cases**:
   1. What is the high level process flow of the insolvency cases. [refrase]
   2. What is the number of cases per supervisory judge though time.
   3. Are the insolvency procedures followed in accordance with the guidelines.
   4. Is the curator hourly rate in accordance with the RECOFA guidelines[30].
2. **On the appointment of the curators**:
   1. Which curators are on the short-list (over time) used by the court.
   2. Do curators get a fair mix of cases with and without proper income
   3. Do curators get appointed to enough cases to make their specialisation worthwile
   4. Are some curators favored over similar others by the committee.
   5. What is the relationship between the court location and the curator location. Are curators always sourced locally?

The design of the information system that answers the research questions leads to the following sub questions:

1. Can the system discover and build a complete network of linked entities of courts, curators, law firms, judges and insolvency cases through time?
2. The insolvency reports hold much of the unstructured, untapped and untagged insolvency case data. These reports are either scanned or converted to PDF and can be in document or spreadsheet format. Some of the questions are based on this obscured information. Can the system effectively extract the relevant data from the PDF insolvency reports?
3. Can the system user effectively query and explore the constructed data in order to answer the main research questions?

# Methodology

## Resources

The central data source that will be used is the *centraal insolventie register (CIR) [*22] which is made available by the Dutch government[27] through a webservice that delivers XML files for court publications and specific updates and delivers insolvency case reports in PDF format.

The CIR registry currently holds over 49,631 insolvency cases, 310,913 PDF reports, 1895 unique curators (free text field, dirty), 561 unique judges (free text field, dirty), 1211 unique law firms (address, no name) and 22 courts.

Other data sources used to determine the golden master records for entities are the *nevenfunctie register* [23] or *openstate* [24] website for judges, *nederlandse orde van advocaten* [25] for curators and law firms. This knowledge base needs to be scraped from the web. An entity relatinship diagram can be found in the appendix.

### Data acquisition status

The CIR XML data has been back filled into a Postgres SQL database and is kept up to date on a daily basis. Field updates in insolvency cases [such as ..] are overwritten, historical information is not kept. The original XML response files are stored so these field changes could be retrieved if to be found of interest.

The PDF reports have been downloaded to a local disk and new reports are downloaded on a daily basis to a server. The reports need to be moved to a available and secure storage such as Amazon’s S3 storage.

The data used for the golden master records still needs to be scraped and for this purpose scrapers must be build.

### Data lineage

As multiple data sources are used it will be necessary to records the data sources for derived fields to establish the data lineage for any search results.

## Data linkage

The CIR data contains free text fields for the entities Curator, Judge and Law Firm. The names in these fields are often written in many different ways and can contain typos. Other data sources mentioned above will be used to create the ‘golden master records’. The data from CIR must be linked to these master records.

Research must be done on finding effective routines for the record linkage[3-6] of the different data sets where disambuigation, normalization and de-duplication will be a part of. The linking field types are person name with and address. The person name for the judge is a single field, the curator has fields for title, initials, middle part and family name. Each entity type can benefit from different approaches.

The research and resulting process should yield a complete set of linked entities for insolvency cases, curators, law firm, judges and courts without requiring many human computation cycles. The linked entities form time dependent graphs that can be queried.

Named entities and relationships can either be stored in a graph database like Neo4J or in a graph datastructure like NetworkX depending on the size and speed requirements.

This data could be made accessible for semantic search using the RDF[28] model if the query interface would benefit from this.

## Report data extraction

Dark data as coined by Gartner[10] is information collected but not used for the purpose of deriving insight or decision making. This applies to the PDF reports which are stored for compliance purposes and that can be individually read by a user but cannot be processed by computers for information retrieval.

The reports contain the progress of the cases, the findings of the curator and the full financial story of the unwinding. Extracting this information would greatly increase the amount of valuable information.

Insolvency reports come in two forms: progress reports and financial reports. The first is based on a Word document, the latter on an Excel spreadsheet. RECOFA has published templates for both types but the adoption thereof is unknown. With templates, structural information is present in the form of standardized headings and grid structure. A special type of progress report is the final report where the final outcome and numbers are reported.

The curator can either convert the Word/Excel file to PDF or scan a printed version to obtain a PDF. In the latter case the file needs to be OCR-ed first and this will most likely lead to a loss of data and structural information where the best solution must be researched. The initial candidates are the Tesseract OCR engine[8] and ABBYY’s OCR cloudservice[9].

The split for progress vs. financial report and converted vs. scanned reports is currently as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Converted | Scanned | Unknown |
| Progress | 25.1% [78554] | 16.6% [51564] | 22.5% [69858] |
| Financial | 7.2% [22542] | 9.7% [30177] | 19.0% [59207] |

NLP techniques will be researched to extract the relevant data from the text.

Reports could be stored in the cloud on S3 or on hard disk and extracted information will be kept in Postgres and Solr datastores. The current disk size of all reports is 114GB.

### Fields of interest in the [Final] Progress Report

* verslagperiode
* bestede uren in verslagperiode
* bestede uren in totaal
* oorzaak faillissement
* personeel
* paulianeus handelen, onderzoek
* crediteuren:
  + Fiscus
  + UWV
  + Andere preferente crediteuren
  + Concurrente crediteuren
* Termijn afwikkeling faillissement

### Fields of interest in the [Final] Financial Report

# Query interface

A web application with a query interface and visualized results will be designed and build that enables the user to answer the research questions and explore the data by himself.

# Literature and references

**Main problem**

[01] De onafhankelijkheid van de faillissementscurator – H2.2-2.3 Prof. mr. B. Wessels, 111p – May 2013

[02] B. Bölük – ‘Is de benoeming van de curator door de rechtbank en het toezicht op de curator door de rechter-commissaris aan een verbetering toe?’ - UvA Master Thesis – Jan 2011

**Data linkage**

[03] Elmegarmid, Ipeirotis and Verykios Duplicate Record Detection: A Survey - IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007

[04] A Survey of Entity Resolution and Record Linkage Methodologies – Brizan, Tanzel - Communications of the IIMA 2006 Volume 6 issue 3

[05] Navarro, Gonzalo (2001). "A guided tour to approximate string matching". ACM Computing Surveys. 33 (1): 31–88.

[06] Preparation of name and address data for record linkage using hidden Markov models - Churches, et al. BMC Medical Informatics and Decision Making 2002, 2:9

[07] T. Berners-Lee. Putting Government Data online. <https://www.w3.org/DesignIssues/GovData.html>

**Extracting structured data from PDF records**

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

[09] ABBYY Cloud OCR SDK - <https://www.abbyy.com/en-eu/cloud-ocr-sdk/>

[10] Gartner Glossary - <https://www.gartner.com/it-glossary/dark-data>

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

[12] Shedding Light on Dark Data – Waters technology

**Public demand for transparency**

[13] Follow The Money (FTM) - <https://www.ftm.nl/artikelen/faillissementswet-schadelijkste-wet-van-nederland> - Feb 2018

[14] BNR - <https://www.bnr.nl/incoming/10148410/bnr-juridische-zaken-toezicht-op-curatoren-moet-beter> - Jun 2013

[15] Blog Michel Knapen - <https://mknapen.wordpress.com/2014/02/20/curator-uit-de-hoge-hoed/> - Feb 2014

[16] FTM - <https://www.ftm.nl/artikelen/het-vreemde-geval-van-de-vluchtende-curator> - Mar 2015

[17] FTM - <https://www.ftm.nl/artikelen/de-schimmige-benoeming-van-curatoren> - Aug 2015

[18] FTM - <https://www.ftm.nl/artikelen/waarom-curatoren-nooit-nee-te-horen-krijgen> - Aug 2015

[19] FTM - <https://www.ftm.nl/artikelen/hoe-de-rechter-commissaris-danst-aan-de-touwtjes-van-de-curator> - Aug 2015

[20] FTM - <https://www.ftm.nl/artikelen/het-extreme-verdienmodel-van-curatoren> - Sep 2015

[21] Advocatie - <http://www.advocatie.nl/grote-kantoren-bedanken-voor-curatorenwerk> - Apr 2017

**Data sources**

[22] Centraal insolventieregister - <https://insolventies.rechtspraak.nl/>

[23] Nevenfuncties van rechters - <https://namenlijst.rechtspraak.nl/>

[24] Open State - <http://ors.openstate.eu/index.php/content/page/opendata>

[25] Nederlandse orde van Advocaten - <https://www.advocatenorde.nl/vind-een-advocaat>

[26] Insolad - <https://www.insolad.nl/leden-zoeken/>

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

[28] Resource Description Framework (RDF) - <https://www.w3.org/RDF/>

**Insolvency law and guidelines**

[29] Insolvency law – <http://wetten.overheid.nl/BWBR0001860/2018-01-01>

[30] RECOFA guidelines - <https://www.rechtspraak.nl/Voor-advocaten-en-juristen/Reglementen-procedures-en-formulieren/Civiel/Insolventierecht/Paginas/Recofa-richtlijnen.aspx>

[31] RECOFA openbaar faillissementsverslag rechtspersoon

# Risk assessment

The following risks and resolutions are being considered:

**Risk**: OCR extraction turns out to produced very poor quality data and structure.

**Backup**: A significant part of the reports are converted and do not need to be OCR-ed, this subset can be used.

**Risk**: The building of a ‘production ready’ public accessible user interface will take to long to build.

**Backup**: The query functionality will be build in a Jupyter notebook.

**Risk**: Entity resolution routines cannot make a complete network

**Backup**: Manual addition of the missing entities as total entity numbers are not that high.

**Risk**: The data needed to answer the questions cannot be extracted or not in enough cases. For the curator appointment, two questions depend on the ability to extract curator fees from the final insolvency report.

**Backup**: The answers to the questions must carry a disclaimer on the limited information availability or quality. In the worst case the questions cannot be answered.

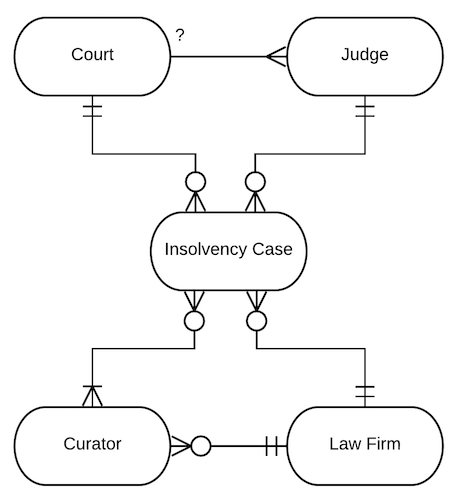
# Project plan

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

|  |  |  |
| --- | --- | --- |
| Period | Task | Deliverable |
| Week 1-2 | Collect data from additional sources by scraping the websites and registry. | Knowledge base of Curators, Judges and Law Firms. |
| Week 3-4 | Perform entity linking between CIR fields and knowledge base. | A complete network of linked entities.  Report of linking/dedup/normalizing routine performance. |
| Week 5-6 | Design and build network query interface | Web application and interface. Answers to related research questions. |
| Week 7 | Explore insolvency reports and compare OCR solutions | Report on PDF extraction results, report structure and choice of OCR engine. |
| Week 8-10 | Build OCR extraction and case annotation | Insolvency case annotations. Performance report over all reports. |
| Week 11 | Design and build insolvency case interface | Query Interface and answers to related research questions. |
| Week 12 | Wrapping up development and writing final report | Final code base and final report. |
| Week 13 | Thesis defense | - |

# Appendix A: Entity Relation Diagram

The entities and their relationships at a specific point in time. Each relationship holds for a period of time. The multiplicity of the relationship is given in Crow’s foot notation.



# **Appendix B: Openbaar faillissementsverslag rechtspersoon**

In de template van het faillissementsverslag[31] worden o.a. de volgende relevante onderdelen en velden gedefinieerd:

1. Algemeen

* Activiteiten onderneming
* Personeel gemiddeld aantal
* Saldo einde verslagperiode
* Verslagperiode
* Bestede uren in verslag periode
* Bestede uren Totaal

1. Inventarisatie
2. Personeel
3. Activa
4. Debiteuren
5. Bank / Zekerheden
6. Doorstart / voortzetten onderneming
7. Rechtmatigheid

* 7.5 Onbehoorlijk bestuur
* 7.6 Paulianeus handelen [onderzoek ja/nee]

1. Crediteuren

* 8.1 Boedelvorderingen
* 8.2 Preferente vorderingen fiscus
* 8.3 Preferente vorderingen UWV
* 8.4 Andere preferente crediteuren
* 8.5 Aantal concurrente crediteuren
* 8.6 Bedrag concurrente crediteuren
* 8.7 Verwachtewijzeafwikkeling

1. Procedures
2. Overig

* 10.1 Termijn afwikkeling faillissement

Notities:

* het verslag kan van geconsolideerd afgewikkelde zaken zijn, danwel gelijkluidend uitgebracht worden
* items kunnen verwijzen naar eerdere verslagen
  + ‘Zie verslag nr. 2 van 1 mei 2006’
  + ‘Zie vorig verslag’
  + ‘Zie verslag 17’
  + ‘verwezen wordt naar het [eerste/tweede/..] verslag’
  + ‘Curator verwijst naar het [eerste/tweede/..] verslag’
* .. of naar andere item nummers:
  + ‘Curator verwijst naar [7.5] van dit verslag