Armchair Auditing of Insolvency Procedures

submitted in partial fulfilment for the degree of master of science

Tom Akkermans
11323671

master information studies
data science
faculty of science
university of amsterdam

2018-08-18

	First Supervisor	**Second Supervisor**
Title, Name	Dr Maarten Marx	
Affiliation	UvA, FNWI, IvI	
Email	maartenmarx@uva.nl	.

Armchair Auditing of Insolvency Processes

Tom Akkermans
University of Amsterdam
tom.akkermans@student.uva.nl

ABSTRACT

[todo: distilleer aan het einde]

CCS CONCEPTS

• **Information systems** → *Document filtering*;

KEYWORDS

keyword1, keyword2, keyword3

ACM Reference Format:
Tom Akkermans. 2018. Armchair Auditing of Insolvency Processes. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

When a company is declared bankrupt by the court, a court committee appoints a administrator to settle the bankruptcy. The administrator's task is to cash in the company's estate and use it to settle the creditors claims. A court's judge must supervise the administrator in order to to act in the best interest of the creditors.

The supervisory function of the judge, the conflict of interest between the administrator and creditors and the appointment process of the administrator are processes which have been the subject of research[8], about which media articles have appeared [9, 10, 12] and which have led to legal proceedings. The involved parties ask for more transparency and information about these processes. Supervisory judges working in a reactive mode under the work pressure could benefit from data driven supervision. Access for the general public and journalists to this process information could provide an extra check and thus further improvement of insolvency processes.

The Dutch government started in 2005 publishing insolvency data[2] according to the insolvency law [3]. It provides an on-line search form to retrieve a single insolvency case and provides open data sources with technical APIs that provide court publications and administrator reports. However, the information from a single insolvency case is limited as it does not provide aggregated information, the administrator reports are unstructured and the collection not searchable and not all interested parties can deal with the offered raw data APIs.

Instead of open data, there is a need for **open analysis** to enable 'armchair audits'[11] of insolvency processes. In this thesis a prototype of an information system will be presented that enables such audits and search for non=technical users. For this, the system processes large amounts of structured and unstructured data of insolvency processes using open and publicly available data sources. From this data, the system:

- extracts insolvency process flow information.
- builds a fully linked, clean entity structure of insolvents, administrators, judges, courts as well as administrator reports and court publications.
- extracts the text of administrator reports and indexes sections and parameters by imposing structure on the content.

A web GUI on top of the data models provides the user interface for audit and search. We describe the implementation challenges and show that the system can provide new insights to the stakeholders who can use a non-technical interface to investigate the insolvency processes and answer specific questions.

1.1 Parties involved in the insolvency process

There are many parties involved in the insolvency process. Directly involved are the the administrator, the insolvent, the judge and the creditors. Indirectly involved but interested are a.o. the organisations of Recofa and Insolad and journalists.

1.1.1 The Administrator (De Curator). The administrator's task is to liquidate the bankrupt firm by selling the assets and from the proceeds pay off the creditors. A second task is to investigate the default and see if there is a case of mismanagement or fraud.

1.1.2 The Insolvent. (*De Failliet*) The insolvent is declared bankrupt by a creditor or declared himself bankrupt. He might be interested in continuing the business and would not want to be persecuted for fraud or malpractice.

1.1.3 The Bankruptcy Judge (De Rechter-Commissaris). The supervisory judge exercises supervision over the administrator which is appointed by the court and is entitled to grant him or her permissions for certain actions.

1.1.4 The Creditors. (*De Schuldeisers*) The creditor has a claim on the insolvent. There is a ranking of creditors from preferred creditors to unsecured creditors.

1.1.5 Insolad. Insolad is the association of lawyers in insolvency law (administrators). It provides up-to-date knowledge on the insolvency practise and the development of laws and regulations in the field.

1.1.6 Rechters-Commissarissen Insolventies (Recofa). Recofa is the consultative body under the council of justice, *raad van de rechtspraak*, specifically for judges working in insolvency law. It sets out policy guidelines for the courts and instructions (e.g. on reporting) for the administrators. The council of the justice system is tasked with improving the quality and efficacy of the jurisdiction. It performs research and shares research findings. It also supplies IT resources.

1.1.7 The Investigative Journalist. The investigative journalist is interested in informing the public on the balance of powers with

Conference'17, July 2017, Washington, DC, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the insolvency law and typically writes critically about cases where this balance is disrupted.

1.1.8 The Economic Journalist. The economic journalist is interested on the impact of the insolvencies and trends therein on the economy .

2 RELATED WORK

[todo]

2.1 Research Questions

The main research question is: **Is is possible to build a useful, complete and correct structured information system based on unstructured CIR data?**

[define measures for success: completeness, accuracy, utility]

- **useful**: the system is useful when it answers questions of the parties involved. We define so-called Persona to represent archetypical users of the information system and define specific questions they have. These questions are distilled from the insolvency law, news articles, research papers, court cases [refs] and interviews.
- **complete** the system's entity structure is complete when all main entities and their inter relations are found. Completeness is also defined for specific parameters needed to answer the user's questions.
- **accuracy** The system's identity structure is accurate when entities refer to their actual real life counterparts.

3 METHODOLOGY

3.1 Description of data sources

The data used by the information system is sourced from three public registers:

1. The Central Insolvency Register (*Centraal Insolventie Register or CIR*)
2. The Register of lawyers (*NOvA's Tableau*)
3. The Register of ancillary positions of judges. (*Register van nevenfuncties van rechters*)

The CIR provides the bulk of the data. The other two registers are used for the entity resolution of administrators and judges.

3.1.1 Central Insolvency Register.

Introduction. The CIR [2] contains company insolvency data supplied by the courts and the administrators. Courts are obliged to supply the insolvency data and free consultation of thereof according to the insolvency law, article 19 [3]. CIR started the register on the 1st of January 2005 and retains case related data until six months after the ending of the insolvency.

SOAP web service. CIR operates a web service using the HTTP SOAP 1.2 protocol which returns information in XML format. Using the web service we can request the new and updated entities of:

- Insolvents
- Publications, by the Court
- Reports, by the Administrator (meta data only)
- Judges
- Administrators
- Courts

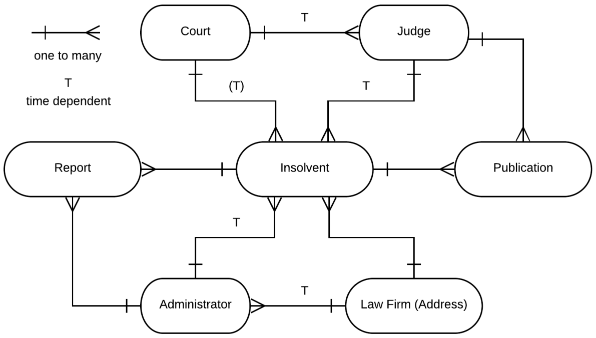

Figure 1: Insolvency entity relations.

The returned XML data is structured: the entities are connected in the XML response by composition as defined in XSD schemas provided by CIR. Unique identifiers exist for the Insolvent, Publication and Report entities and they can easily be stored in normalized SQL tables.

Database normalization. Not all the data is normalized. The entities Judges and Administrators have no identifiers but consist of free text fields for their name parts. This leads to unwanted duplication as names can be written in many different forms and typos can be introduced, e.g. one judge's name appears as:

- "mr. W.J. Geurts - de Veld"
- "mr. W.J. Geurts-deVeld"
- "mr. W.J. Geurts-de Veld"
- "mr. W.J.Geurts-de Veld"
- "mr.W.J. Geurts-de Veld"
- "mr. W.J. Geurts-de Veld (Rotterdam)"
- "mr W.J.Geurts-de Veld"
- "W.J.Geurts-de Veld"

We need to define so-called master data for judges and administrators containing the real world entities. The two data sources in the sections 3.1.2 and 3.1.3 are chosen for this purpose. CIR entity names are first normalized for de-duplication and are subsequently linked to the master data records on their normalized name.

PDF report web service. A second web service operated by CIR provides Administrator Reports in PDF format by HTTP request. There are two types of reports:

1. progress reports
2. financial attachments

Recofa has published templates for both report types[6]. These reports hold much of the unstructured data.

CIR data contents. Table 1 shows the content of the CIR register data as of date [2018-08-12] and the average monthly size of new additions.

The number of current defaults is about the lowest of the century [ref] [graph of declining defaults]

Table 1: Actual number of records and monthly average additions per entity.

Entity	no. of records	monthly add.
Insolvent	17,331	280
Report	146,865	4408
... progress report	87,430	
... financial attachment.	56,611	
Publication	37,031	1447
Administrator (distinct)	2329	
Judge (distinct)	451	
Court	11	

3.1.2 Register of lawyers, NOvA Tableau. The NOvA tableau is the official register for lawyers and maintained by the *Nederlandse Orde van Advocaten (NOvA)*[5]. Lawyers are obliged to be registered in the tableau by the lawyer's law (*advocatenwet*, article 1 [4]). NOvA offers an on-line search form where key word search and filters can be applied to search for a lawyer. This data source is used to collect the master data for Administrators.

3.1.3 Register of judges, Nevenfuncties van rechters. The Register for ancillary positions for judges is made available by *de Rechtspraak*[1]. It offers an on line form and returns the name, current and historical occupation and ancillary positions. This data source is used to collect the master data for Judges.

3.2 Description of the system and data flow

3.2.1 System components. Figure 2 gives an overview of the system components for extracting, enriching and integrating the sourced data and making the resulting structured and higher level information available to the user.

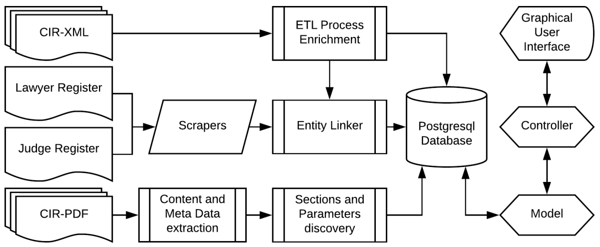

Figure 2: System overview.

The ETL process and Enrichment. This component extracts the entities with selected data fields from the XML. The data is enriched and annotated after which it is stored in the Postgresql database. An example of the enrichment are the start and end date of the insolvency which are derived from court publications.

The Entity Linker. This component is responsible for linking judges and administrators in CIR to real life entities found in the judge and lawyer registers. It does this by normalization of the CIR names and resolving them to the register entities using similarity functions. Unresolved entities can be linked manually to guarantee completeness. Established links are stored in an association table. After an initial run, the register scrapers are called upon only when new judges or administrators entries appear in CIR.

PDF processors. PDF reports are processed to extract the textual content and meta data. The text sections, as defined in the progress report template, and key data parameter are discovered in a subsequent process. This data is available for search later on.

Model-View-Controller (MVC). This is a well established pattern of three subcomponents working together for a GUI. The user operates a graphical interface, prototyped in Jupyter notebooks, to query the data or interact with data visualisations or tables. The interface is the View in the MVC component. On user command, the Controller asks the Model to prepare the necessary data and then passes this data to the View to update the interface.

3.3 Implementation challenges

3.3.1 Entity De-duplication, Scraping and Linking.

Deduplication: normalizing names. CIR provides administrator names in four parts: title, initials, middle part and family name.

The initials were normalized into single characters using periods and no spaces. In Dutch, initials for names like *Theodoor* and *Christiaan* are written with two or three characters as *Th.* or *Chr.*. These names are derived from the Greek where the initial *Th(eta)* for example is written as one character. Initials for double names like *Albert-Jan* may be written as *A-J.* or *A.-J.*.

The middle part, e.g. 'van der' or 'ter' was sometimes supplied in the middle part field and sometimes in the family name (and sometimes in both). To normalize the middle part it was merged into the family name making sure it did not appear twice.

The family name was stripped from academic and noble (not uncommon in this profession) titles. Misplaced initials and CIR specific comments were removed. Maiden names are connected to the surname with a hyphen and no spaces.

Furthermore, all parts are made lower case, surrounding whitespace characters are stripped, accents are removed and multiple spaces are replaces with a single space.

De-duplication by normalizing the administrator names reduced the number of distinct administrator names from 2329 to 1559, a 33% reduction.

Judge names are provided in a single field. This sometimes leads to initials or a middle name part that is stuck together to the family name and was separated using regexm e.g. in 'C. vanSteenderen' and 'mr. W.J.Geurts-de Veld'. CIR often adds the court between parenthesis which was also removed in the normalization.

De-duplication by normalizing the judge names reduced the number of distinct judge names from 565 to 277, a 51% reduction.

Administrator name scraping. NOvA unfortunately does not provide complete lists of registered lawyers to which we could link the administrators. Instead we link the administrator by using the site's search form. An underlying REST API returning JSON data was discovered which was used instead of the form. The API (or form) can not handle initials or stop words as 'de' and 'van' and only the family name without stop words was used. The best candidate in the search results, which are sorted on the relevance provided by NOvA, is found by (partially) matching the initials until a match is found. The API enables the use of filters and we filter from narrow to broad until a candidate is found, filtering on:

(1) legal specialism: administrator
(2) legal branch: insolvency law
(3) no filter

NOvA only supplies the actual lawyer register but CIR also lists administrators previously working on a case therefore a substantial amount of administrators cannot be linked. Another website, www.advocatenzoeken.nl that sources data from NOvA but keeps historical data was scraped as well to complement NOvA.

Administrator name linking. From the normalized name list we linked 73% to NOVa registered lawyers and another 18% administrators from the www.advocatenzoeken.nl site. The remaining 10% unfound names are linked to a special 'unknown' lawyer.

Table 2: Administator entity linking results

Source	no. linked	correct	incorrect
NOvA	1134	1133	1
AdvocatenZoeken	280	279	1
Not found	149		
Total	1563	1412	

Correctness is assumed when the names match. When an obvious typo is found in the family name. When the initials either match, or are contained in the master record or are permutated or are missing.

[discuss not found results: common name, place in name]

Judge name scraping. The CIR register for ancillary functions for judges was scraped for judges from courts and higher courts. The total set contains 3691 judges and should be a superset of active case judges. The website was driven by the Angular javascript framework and could not easily be scraped with simple HTTP requests. Selenium was used to mimic user browsing behaviour to invoke the javascript.

Judge entity linking. A judge is linked by to a judge master record with the most similar normalized name where similarity was calculated using the Levensthein or edit distance. 88% of the normalized names were matched. 4 of the incorrectly 12% matched names could be set manually and the remaining 8% was linked to a special 'unknown' judge.

Table 3: Judge entity linking results

Source	no. linked	correct	incorrect
Ancillary positions register	277	246	32
+Manual lookup of the 32		10	-10
Total	277	256	22

Correctness is defined similar to the administrator normalized name correctness.

[discuss not found (initially) results: maiden name, not a judge anymore]

3.3.2 Process Mining.

3.3.3 PDF content extraction. The PDF reports can be split into two types depending on how the PDF was created:

(1) Scanned from paper using a scanner or copier.
(2) Converted by software from another format.

Scanned PDFs. Scanned PDFs contain images only. To convert the images to text we used the Ghostscript library for image extraction and the Tesseract OCR engine for the character recognition. Tesseract supports the Dutch language which is paramount to our application. It is very important to supply Tesseract with good quality images. We used the Ghostscript settings of a *tiffgray* output device with a *300 DPI* resolution.

Converted PDFs. Converted PDF content can be extracted with a number of packages such as PDFMiner, PyPDF2, pyPoppler and pdftotext. We chose the latter after comparison because it maintains the layout which is important for section and parameter extraction and being build in C++ it is very fast.

Third type. A third type appeared: PDFs that were scanned and subsequently OCR-ed by a copier. They contain both text and images. The OCR quality is often poor. We retried re-OCR-ing the images with Tesseract which solved some errors but introduced others. Meta data about a.o. the scanner type was obtained which could improve post-process text extraction in future work.

3.3.4 PDF report section and parameter extraction. [todo]

3.4 Methods

3.4.1 RQ1.

4 EVALUATION

dedup lawyers: Phonetic similarity https://www.quora.com/What-is-a-good-algorithm-service-for-fuzzy-matching-of-companies-names-for-de-duplication

5 CONCLUSIONS

5.1 Acknowledgements

REFERENCES

[1] [n. d.]. Beroepsgegevens en nevenfuncties van rechters. https://namenlijst.rechtspraak.nl

[2] [n. d.]. Centraal Insolventieregister. https://insolventies.rechtspraak.nl

[3] [n. d.]. Insolvency Law - Faillissementswet. http://wetten.overheid.nl/BWBR0001860/2018-07-01

[4] [n. d.]. Lawyer Law - Advocatenwet. http://wetten.overheid.nl/BWBR0002093/2018-07-25

[5] [n. d.]. Nederlandse Orde van Advocated. https://www.advocatenorde.nl/over-de-nova

[6] [n. d.]. Recofa-richtlijnen. https://www.rechtspraak.nl/Voor-advocaten-en-juristen/Reglementen-procedures-en-formulieren/Civiel/Insolventierecht/Paginas/Recofa-richtlijnen.aspx

[7] 2017. *Klantwaardering Rechtspraak 2017.* Technical Report.

[8] Burak Bölük. 2011. *Is de benoeming van de curator door de rechtbank en het toezicht op de curator door de rechter-commissaris aan een verbetering toe?* Master's thesis.

[9] Dennis Meneer. 2017. Curatoren vluchten uit lege boedels. https://www.ftm.nl/artikelen/curatoren-slaan-op-de-vlucht-voor-lege-boedels Accessed on 2018-06-14.

[10] Dennis Meneer. 2018. Dit is de schadelijkste wet van Nederland. https://www.ftm.nl/artikelen/faillissementswet-schadelijkste-wet-van-nederland Accessed on 2018-06-14.

[11] Daniel E. O'Leary. 2015. Armchair Auditors: Crowdsourcing Analysis of Government Expenditures. *JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING* 12 (jul 2015), 71–91. https://doi.org/10.2308/jeta-51225

[12] Jan-Hein Strop. 2015. De schimmige benoeming van curatoren. https://www.ftm.nl/artikelen/de-schimmige-benoeming-van-curatoren Accessed on 2018-06-14.

6 APPENDIX: PERSONA

The main Persona are:

- **The Judge** representing the side of Recofa, *Raad van de Rechtspraak* and The Bankruptcy Judge (*Rechter-Commissaris*)
- **The Administrator** representing the side of Insolad (Vereniging Insolventierecht Advocaten) and The Administrator (Curator)
- **The Insolvent** representing the owner(s) of the defaulted company.
- **The Creditor** representing the Unsecured Creditor (*De Concurrente Schuldeiser*).
- **The Journalist** representing the investigative and economic journalist.

In the following sections each Persona will be described and their questions to the system stated.

6.1 The Judge

The Persona of Judge is interested in the adherence to the insolvency law and to the Recofa policy guidelines. It is also interested in specific trends in the insolvency processes and operational issues such as work pressure for judges. On an individual judge level it is interested in supervising its active cases.

Questions:

- How many cases does a judge supervise at a certain point in time (now)
  - … distribution over all judges at a certain court
  - … distribution over all judges at all courts
- What is the process flow of cases through court:
  - What is the case time distribution from begin to end
  - How long does it take to set the verification meeting (law says < 14 days)
  - How long does it take to publish the plan of final distribution.
  - What percentage of cases:
    * end early as there are no assets
    * end by paying all creditors
    * contain an agreement between insolvent and creditors (akkoord)
    * are following an simplified settlement (no meeting of creditors)
    * are filed by the insolvent vs creditor
  - What are the experience factor (*jaren praktijkervaring*) and estate factor (*grootte actief*) for the cases determining the administrators hourly wage.
- Are administrators reporting according to the instructions:
  - How often is the progress reporting deadline breached.
  - How often is the financial attachment omitted for all reports in a case.
- Are administrators using the supplied template for progress report:
- How often does insolvency fraud occur
- *What are the issues in progress reports that need the judge's attention*

6.2 The Administrator

The Persona of Administrator is interested in the administrator appointment process of the courts and on issues that threaten its business or leave it powerless.

from [7]:

"Volgens de curatoren is meer transparantie van belang bij de verdeling van faillissementen: ze ervaren deze verdeling nu als een black box. [...] Daarnaast maken de curatoren zich flinke zorgen over het hoge verloop onder rechtercommissarissen en de griffie en de gevolgen daarvan."

Questions:

- Which curators are on the court's short list for appointment
- Are insolvency cases distributed fairly over the curators on the list
- How long are judges working in the insolvency field, how often do they rotate.
- How often do empty estates occur (an empty estate implicated work without pay)
- To what degree do Banks claim all the proceeds from the assets.

6.3 The Insolvent

The Persona of Insolvent could be interested in restarting the business. It also is afraid of being made personally liable for the bankruptcy claims.

- - How often are insolvent made personally liable (when there is little estate)
- - How often do insolvency restarts and prepacks occur

6.4 The Creditor

The Persona of Creditor is interested in the recovery rate of its claim and the time of payout and the factors that influence those.

- What is my expected recovery rate and time
- - How high is the administrator's salary, is he eating up the proceeds.

6.5 The Journalist

The Persona of Journalist is writing an article on a certain insolvency phenomenon. An economic journalist might periodically write the same story on say the number of defaults in the last quarter compares to a previous period, usually for business readers. The investigative journalist writes for a wide audience, therefore the article contains both general trends as well as the personal individual story. For this the investigative journalist needs query functionality to dig for evidence and pull individual records.

- J2: (topic: incapable administrators) Which administrators have often been taken of their cases by the judge.

- J1: How many defaults occurred over the last three months compared to a year earlier. [viz:cumulative lines of two periods]