Armchair Auditing of Insolvency Processes

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
2019-03-20

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

Universiteit van Amsterdam

Amsterdam Data Science

Contents

Armchair Auditing of Insolvency Processes

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

ACM Reference Format:
Tom Akkermans. 2019. Armchair Auditing of Insolvency Processes. In *Proceedings of 2 (University of Amsterdam).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

When a company is declared bankrupt by the court, a court committee appoints an administrator to settle the bankruptcy. The administrator's task is to liquidate the company's estate and use the proceeds to settle the creditors claims. A supervisory judge ensures that the administrator is acting in the best interest of the creditors.

Involved parties have been demanding more transparency into insolvency processes. 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[1], about which media articles have appeared [8, 7, 16] and which have led to legal proceedings.

citeer

In 2005 the Dutch government started the digital register of insolvency data[12]. An on-line search form [14] is provided to retrieve a single insolvency case. Web services are provided to retrieve court publications in XML format and administrator reports in PDF format.

However, the information from a single insolvency case is limited as it does not provide aggregated and linked information. The administrator reports are unstructured and not searchable as a collection. Furthermore most involved parties lack the technical skills to use the web services. Instead of open data, there is a need for **open analysis** to enable 'armchair audits'[10] of insolvency processes.

In this thesis we investigate whether [RQ] **it is possible to build a complete and correct structured information system (IS) based on open and public data that can satisfactorily provide transparency to parties involved**. The IS comprises a web GUI, data model and several data collection and extraction parts. The main research question can therefore be broken up into the following sub questions:

- RQ1 Can we construct a complete, cleaned [deduped] and fully linked entity network of insolvency cases, administrators, judges and courts.
- RQ2 Can we correctly and completely label insolvency case state data with state data [start/end date] in order to mine the insolvency process.
- RQ3 Can we correctly extract specific fields [paulianeus handelen] of interest from unstructured documents to classify insolvency cases.
- RQ4 Can involved stakeholders use the IS to provide the requested transparency on insolvency processes.

University of Amsterdam, 3,
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

This thesis describes the building of such a system that takes in large amounts of open and publicly available data in structured and unstructured form, extracts and enriches useful facts and makes it consumable for analysis to provide insights into the insolvency processes via a web GUI.

2 RELATED WORK

Describe related work section in FINAL THESIS QUALITY (WORK IN PROGRESS).

2.1 RQ1: Constructing the Entity Network

Entity deduplication. Describe deterministic methods for deduplication. Describe entity linking using edit/Levenshtein distance measures.

2.2 RQ2: Process mining insolvency case data

Describe Wil van Aalst - Process Mining - Play in. Describe methods using log files to deduce the process model.

2.3 RQ3: Text mining unstructured PDF files

Describe text mining techniques.

2.4 RQ4: IS Evaluation by involved parties

Describe evaluation measures for information systems.

Shannon and Weaver [15] state that output of an IS can be defined at different levels

(1) **technical level**: the accuracy and efficiency of the system.
(2) **semantic level**: the success of information conveying the intended meaning.
(3) **effectiveness level**: the effect of information on the receiver

This implies that factors contributing to a successful IS can also be defined at different levels. DeLone and McLean[2, 3] have gathered many factors from literature and grouped them into six distinct but interdependent categories, see figure 1 below.

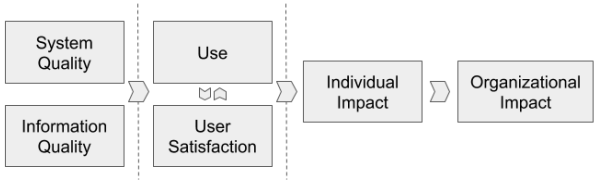


Figure 1: DeLone and McLean Information Success Model.

Each category or aspect of the I/S success measures still contains variables that contribute to success:

- **System Quality**: measures the IS itself on the technical level. Example variables are data currency, completeness, and ease of use.
- **Information Quality**: measures the IS output such as reports at the semantic level. Example variables are accuracy, precision and relevance.
- **Information Use**: measures the actual use of the IS which is where the effectiveness level starts.
- **User Satisfaction**: measures satisfaction from the perspective of the user, usually on an interval scale. This category is the mostly used as a single success measure and is fairly subjective.
- **Individual Impact**: measures the information effect on the behaviour of the recipient.
- **Organizational Impact**: measures the information effect on the performance of the organization.

3 METHODOLOGY

3.1 Description of Data Sources

add informative visuals for data sources description

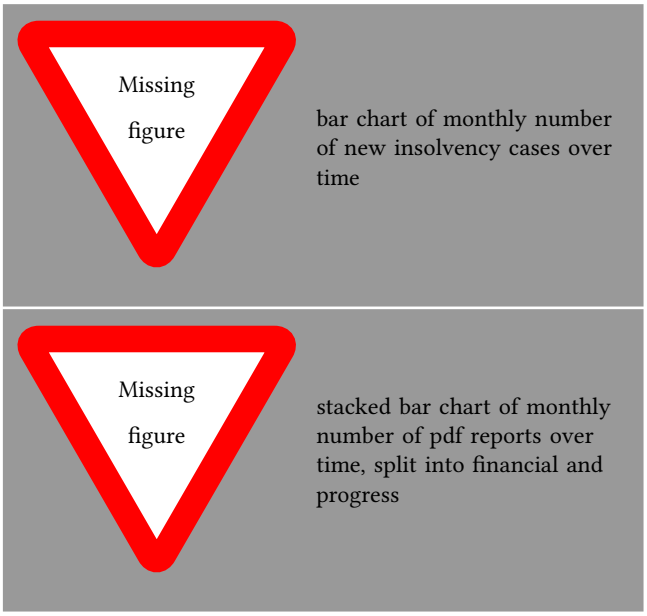
3.1.1 *Central Insolvency Register.*

Data suppliers. The CIR contains company insolvency data supplied by the courts and the administrators. Courts are obliged to supply the insolvency data and free consultation thereof according to the insolvency law, article 19 [5]. CIR contains insolvency cases from the 1st of January 2005 and retains these until six months after their end date. CIR also contains other data such as personal debt restructuring (*schuldsanering*), personal insolvency and company's failure to pay (*surseance*) but this data is out of scope.

Entity records. The CIR register contains the following entities in numbers of records (as of 2019-03-21):

Table 1: number of entity records.

Entity	no. of records
Court	11
Supervisory Judge (distinct names)	580
Insolvency	51,392
Administrator (distinct names)	58,201
Publication	142,172
Report	357,803
... progress report	237,657
... financial attachment.	120,146



Publications on an insolvency case are done by the court and include the initial declaration of bankruptcy. Administrators periodically submit progress reports as well as financial attachments to the CIR.

Entity identifiers. CIR entity data is made available by a SOAP web service returning XML responses. The XML is semi structured data and contains entities by composition which are extracted using a parser. It provides natural unique identifiers for Insolvency Cases, Publications and Reports so they can be easily stored in normalized SQL tables and linked. The other entities: Courts, Judges and Administrators have no identifiers but consist of free text fields for their name parts. These entities must be de-duplicated en linked to a master data record. It can be easily observed in table 1 that this is certainly needed for administrators

state estimated number of administrators

.

Entity relations. Figure 2 below shows the relationships between the entities including their cardinality. Note that some relationships are time dependent, e.g. a judge can be replaced during the lifetime of an insolvency case. Since 1-1-2019 there can be two judges appointed to one case.

Administrator reports. A second web service operated by CIR provides administrator reports in PDF format. These reports hold much of the unstructured data. Recofa has published templates for both progress and financial attachment reports [13] which provide a certain structure to the contents.

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)*[9]. Lawyers are obliged to be registered in the tableau by the lawyer's law (*advocatenwet*, article 1 [6]). NOvA offers an on-line search form where keyword search and filters can be applied to search for a lawyer. This data source was chosen to collect the master data for Administrators.

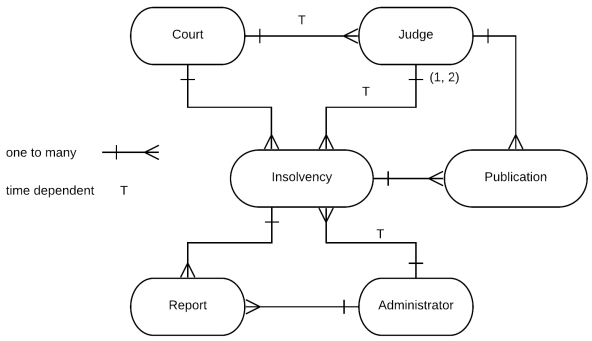


Figure 2: Insolvency entity relations.

3.1.3 Register of judges, Nevenfuncties van rechters. The Register for ancillary positions for judges is made available by *de Rechtspraak*[11]. It offers an on-line form and returns the name, current and historical occupation and ancillary positions. This data source was chosen to collect the master data for Judges.

3.2 Information System Description

Figure 3 gives an overview of the system components for sourcing, extracting, enriching and integrating the data and making the resulting structured and higher level information available to the user's analysis.

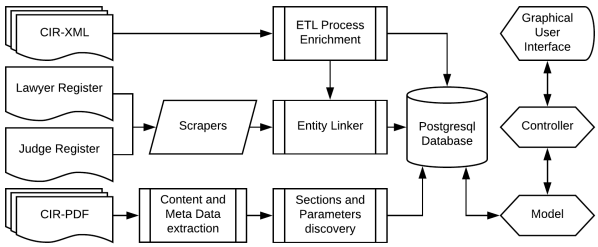


Figure 3: System overview.

Data flows from left to right through the following components:

Data Sources. Data is sourced from three public registers:

(1) The Central Insolvency Register (*Centraal Insolventie Register or CIR*). CIR exposed both an XML and PDF file web service.
(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.

ETL and Enrichment. This component loads entities with selected data fields from the CIR XML data. The data is cleaned and enriched after which it is stored in a relational database.

Entity Linker. This component is responsible for linking judges and administrators in the CIR XML data to real life entities found in the judge and lawyer registers.

PDF Processors. These components processes the CIR PDF reports to extract 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 and loaded into the relational database.

Database and File Storage. Entity data is stored in a relational Postgresql database. Administrator PDF reports are stored in Amazon's S3 object storage.

Model-View-Controller (MVC). This well established pattern of subcomponents works together as a graphical interface for the user to analyse the data.

3.3 Methods

Hoe je je vraag gaat beantwoorden. Dit is de langste sectie van je scriptie. Als iets erg technisch wordt kan je een deel naar de Appendix verplaatsen. Probeer er een lopend verhaal van te maken. Het is heel handig dit ook weer op te delen nav je deelvragen

3.3.1 RQ1: Constructing the Entity Network.

the Entity Network. The requested CIR XML data will go through an Extract Transform Load (ETL) process. Here each entity data is extracted, de-duplicated and identified and linked to the other entities to construct the entity network. For each entity a so-called master data table is created that contains the unique and agreed upon instances or 'golden records' of the entities.

Insolvency Cases. Insolvency cases are supplied with a natural identifier (*insolventienummer*), for example F.19/13/123, which consist of the insolvency type (F), a system number (19), year of insolvency ([20]13) and a serial number (123).

Duplicate insolvency cases exist for a number of reasons:

(1) data errors: padded zeros on the serial number, e.g. '0123'
(2) redefinition of courts[1]: cases administered both with the new case number format as well as the old case number format without the system number.
(3) cases handed over to another court.

To de-duplicate these cases, lineage of erred or transferred cases was established and publications and reports re-linked to the correct case. The duplicate cases could then easily be filtered out.

Judges. We define the master data of judges by scraping the names of judges found in the register of ancillary positions of judges[11]. The register names and CIR judge names are normalized after which the CIR name is linked to the most probable register judge name using the Levenshtein distance

describe what happens in case of a tie

.

[1] in 2013 courts were consolidated (*Herziening Gerechtelijke Kaart*, see [4]).

CIR judge name normalization and de-duplication Judge names in the CIR data are provided as single free text field and can contain many duplicates. For example 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"

This data needs to be de-duplicated and we do this by cleaning and normalizing the name in a number of sequential steps where the order is of importance:

(1) add missing spaces between name parts
(2) change the name to lower case and strip surrounding white-space characters
(3) remove accents, academic and nobility titles, additions in brackets, periods and double spaces
(4) normalize the use of hyphens

register judge normalization The normalization of register judge names slight differs from the CIR name. The number of removed nobility and academic titles is larger and ...

simple matching using Levenshtein distance does not place the same weight on initials and surname, describe matching

Administrators

Measuring CIR Data Completeness. To measure the completeness of the data obtained from the CIR XML web service we make use of its following methods[17]:

(1) **searchModifiedSince** (cutoff 2012-01-01 / 2005-01-01) : returns the publication ids of modified insolvents
(2) **searchReportsSince** (cutfoff 2010-07-01? / 2005-01-01) : returns the report ids of added reports.

The publications contain the fields for the ids or names of the other entities. By checking these fields form missing values and the values against the database we measure completeness of the data.

(re)measure completeness using dates above

Extracting Information from Entity Data.

write

Extracting Information from Unstructured Reports.

write

3.3.2 *RQ2: Process mining insolvency case data.*

3.3.3 *RQ3: Text mining unstructured PDF files.*

3.3.4 *RQ4: IS evaluation by involved parties.*

4 RESULTS

subsectie per deelvraag. Is/in hoeverre is de vraag beantwoord. Treffende Visualisaties.

4.0.1 *RQ1: Constructing the Entity Network.*

4.0.2 *RQ2: Process mining insolvency case data.*

4.0.3 *RQ3: Text mining unstructured PDF files.*

4.0.4 *RQ4: IS evaluation by involved parties.*

4.1 System Quality

move this into RQ4 above

4.1.1 Completeness. Completeness of Publication and Report data gathered from CIR was checked using the CIR SOAP requests. Within the publications we checked for completeness of Insolvency Case, Administrator, Judge and Court data:

- Publication: 100%. CIR contained 142173 publications of insolvents that were all present in the database.
- Report: 99.99%. CIR contained 142330 reports of which 9 were missing in the database.
- Insolvent: 100.00%. CIR publications contained 51392 insolvent cases of which none were missing in the database.
- Administrator: 98.65%. 695 insolvent cases had no administrator data.
- Judge: 99.46%. 275 insolvent cases had no supervisory judge data.
- Court: 99.94% 28 insolvent cases had no court data.

4.1.2 Handling Duplicates.

Courts. Courts are specified in the XML response as a free text field but are uniquely identified by their name and do not need further processing:

Court name	Cases
Rechtbank Amsterdam	4601
Rechtbank Den Haag	4851
Rechtbank Gelderland	5947
Rechtbank Limburg	3277
Rechtbank Midden-Nederland	5580
Rechtbank Noord-Holland	3543
Rechtbank Noord-Nederland	4556
Rechtbank Oost-Brabant	5123
Rechtbank Overijssel	3789
Rechtbank Rotterdam	5638
Rechtbank Zeeland-West-Brabant	4459
None	28

Consistency: there are no duplicates. Completeness: 28 records have missing data.

5 CONCLUSIONS

[Hierin beantwoord je jouw hoofdvraag op basis van het eerder vergaarde bewijs.]

5.1 Acknowledgements

REFERENCES

[1] Burak Bölük. "Is de benoeming van de curator door de rechtbank en het toezicht op de curator door de rechter-commissaris aan een verbetering toe?" MA thesis. Jan. 2011.

[2] W.H. DeLone and E.R. McLean. "Information System Success: The Quest for the Dependent Variable". In: *Information Systems Research* 3.1 (1992), pp. 60–95.

[3] W.H. DeLone and E.R. McLean. "The DeLone and McLean Model of Information System Success: a Ten-Year Update". In: *Journal of Management Information Systems* 19 (Apr. 2003), pp. 9–30.

[4] *Herziening gerechtelijke kaart*. URL: https://www.om.nl/organisatie/herziening/.

[5] *Insolvency Law - Faillissementswet*. URL: http://wetten.overheid.nl/BWBR0001860/2018-07-01.

[6] *Lawyer Law - Advocatenwet*. URL: http://wetten.overheid.nl/BWBR0002093/2018-07-25.

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

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

[9] *Nederlandse Orde van Advocated*. URL: https://www.advocatenorde.nl/over-de-nova.

[10] Daniel E. O'Leary. "Armchair Auditors: Crowdsourcing Analysis of Government Expenditures". In: *JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING* 12 (July 2015), pp. 71–91. DOI: 10.2308/jeta-51225.

[11] De Rechtspraak. *Beroepsgegevens en nevenfuncties van rechters*. URL: https://namenlijst.rechtspraak.nl.

[12] De Rechtspraak. *Centraal Insolventieregister*. URL: https://insolventies.rechtspraak.nl.

[13] De Rechtspraak. *Recofa-richtlijnen*. URL: https://www.rechtspraak.nl/Voor-advocaten-en-juristen/Reglementen-procedures-en-formulieren/Civiel/Insolventierecht/Paginas/Recofa-richtlijnen.aspx.

[14] De Rechtspraak. *Zoeken in het Centraal Insolventieregister*. URL: https://insolventies.rechtspraak.nl/#!/zoeken/index.

[15] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois, 1949.

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

[17] *Technische documentatie webservice Centraal Insolventieregister*. Jan. 2019.