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

Armchair Auditing of Insolvency Processes

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

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.

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[9], about which media articles have appeared [10, 11, 13] and which have led to legal proceedings. The involved parties demand more transparency of these processes. Supervisory judges with significant work load could benefit from data driven supervision. Information access to the general public and journalists to this processes would provide additional checks and balances. We derive so-called Personas to represent these parties and articulate their information needs.

In 2005 the Dutch government started a digital register of insolvency data[2]. An on-line search form [8] is provided to retrieve a single insolvency case and there are web services to retrieve court publications and administrator reports in XML and 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 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'[12] of insolvency processes. In this thesis we investigate **whether it is possible to build a complete and correct structured information system based on open and public data that is useful in that it enables such audits and search for non-technical users.**

For this we investigate whether we can build a system that:

- constructs a complete, cleaned and fully linked entity network of insolvency cases, administrators, judges, etc.
- fully extracts the information of interest from entity data.
- fully extracts the information of interest from unstructured documents.
- provides the persona with easy to use functionality to answer its information need.

We describe the steps in building such a system that takes in large amounts of open and publicly available data in structured and unstructured data 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

add literature related to the research questions: on information system quality, entity linking, de-duplication, tobit model

3 METHODOLOGY

3.1 Description of Data Sources

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 [4]. CIR contains insolvency cases from the 1st of January 2005 and retains these until six months after they end.. 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

89   state estimated number of administrators

90   .

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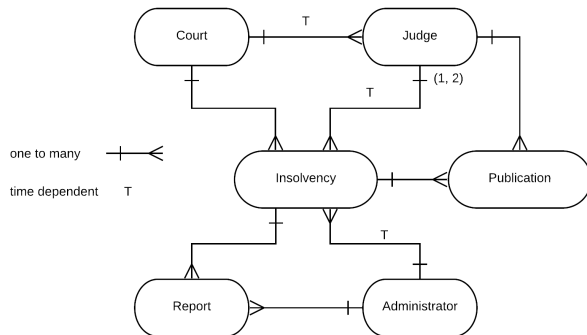95   appointed to one case.

Figure 1: Insolvency entity relations.

96   *Administrator reports.* A second web service operated by CIR
97   provides administrator reports in PDF format. These reports hold
98   much of the unstructured data. Recofa has published templates for
99   both progress and financial attachment reports [7] which provide a
100 certain structure to the contents.

101 *3.1.2   Register of lawyers, NOvA Tableau.* The NOvA tableau is the
102 official register for lawyers and maintained by the *Nederlandse Orde*
103 *van Advocaten (NOvA)*[6]. Lawyers are obliged to be registered in
104 the tableau by the lawyer's law (*advocatenwet*, article 1 [5]). NOvA
105 offers an on-line search form where keyword search and filters can
106 be applied to search for a lawyer. This data source was chosen to
107 collect the master data for Administrators.

108 *3.1.3   Register of judges, Nevenfuncties van rechters.* The Register
109 for ancillary positions for judges is made available by *de Recht-*
110 *spraak*[1]. It offers an on-line form and returns the name, current
111 and historical occupation and ancillary positions. This data source
112 was chosen to collect the master data for Judges.

3.2   Information System Description

114 Figure 2 gives an overview of the system components for sourc-
115 ing, extracting, enriching and integrating the data and making the
116 resulting structured and higher level information available to the
117 user's analysis.
118   Data flows from left to right through the following components:

119   *Data Sources.* Data is sourced from three public registers:

120 (1) The Central Insolvency Register (*Centraal Insolventie Register*
121   *or CIR*). CIR exposed both an XML and PDF file web service.
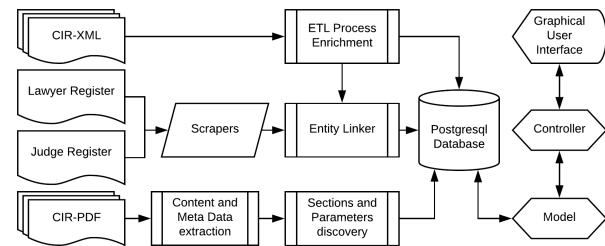122 (2) The Register of lawyers (*NOvA's Tableau*).

Figure 2: System overview.

123 (3) The Register of ancillary positions of judges. (*Register van*
124   *nevenfuncties van rechters*)

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

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

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

133   *PDF Processors.* These components processes the CIR PDF re-
134 ports to extract textual content and meta data. The text sections as
135 defined in the progress report template and key data parameter are
136 discovered in a subsequent process and loaded into the relational
137 database.

138   *Database and File Storage.* Entity data is stored in a relational
139 Postgresql database. Administrator PDF reports are stored in Ama-
140 zon's S3 object storage.

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

3.3   Methods

145 *3.3.1   Constructing the entity network.* Here we describe how each
146 entity is cleaned, standardized, de-duplicated and identified after
147 which it can be linked to other entities to construct the entity
148 network. For each entity a so-called master data table is created
149 that contains the agreed upon instance or 'golden record' of the
150 entities that can be used as a common reference point.

151   *Courts.* Courts are specified in the XML response as a free text
152 field. A master table is created using the given names.

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.

Insolvency cases. Insolvency cases are supplied with a natural identifier (*insolventienummer*), for example F.19/13/123, which consist of the insolvency type, a system number, year of insolvency and a serial number. A number of issues introduce unwanted insolvent duplicates:

(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 solve for this data errors were corrected and duplicate cases were linked to the correct cases using scripts, annotating lineage and re-linking publications

Judges.

Administrators.

3.3.2 Extracting information from entity data.

3.3.3 Extracting information from unstructured reports.

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.]. Herziening gerechtelijke kaart. https://www.om.nl/organisatie/ herziening/
[4] [n. d.]. Insolvency Law - Faillissementswet. http://wetten.overheid.nl/ BWBR0001860/2018-07-01
[5] [n. d.]. Lawyer Law - Advocatenwet. http://wetten.overheid.nl/BWBR0002093/ 2018-07-25
[6] [n. d.]. Nederlandse Orde van Advocated. https://www.advocatenorde.nl/ over-de-nova
[7] [n. d.]. Recofa-richtlijnen. https://www.rechtspraak.nl/ Voor-advocaten-en-juristen/Reglementen-procedures-en-formulieren/ Civiel/Insolventierecht/Paginas/Recofa-richtlijnen.aspx
[8] [n. d.]. Zoeken in het Centraal Insolventieregister. https://insolventies. rechtspraak.nl/#!/zoeken/index
[9] 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.
[10] 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.
[11] 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.
[12] 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
[13] Jan-Hein Strop. 2015. De schimmige benoeming van curatoren. https://www.ftm. nl/artikelen/de-schimmige-benoeming-van-curatoren Accessed on 2018-06-14.

---

[1]in 2013 courts were consolidated (*Herziening Gerechtelijke Kaart*, see [3]).