DESTATIS
Statistisches Bundesamt

WEBSCRAPING NOTICES OF INSOLVENCY PROCEEDINGS WITH R

Using publicly available data to enhance survey response quality

uRos2019, 20./21. May 2019, București

Insolvency Proceedings and Official Statistics

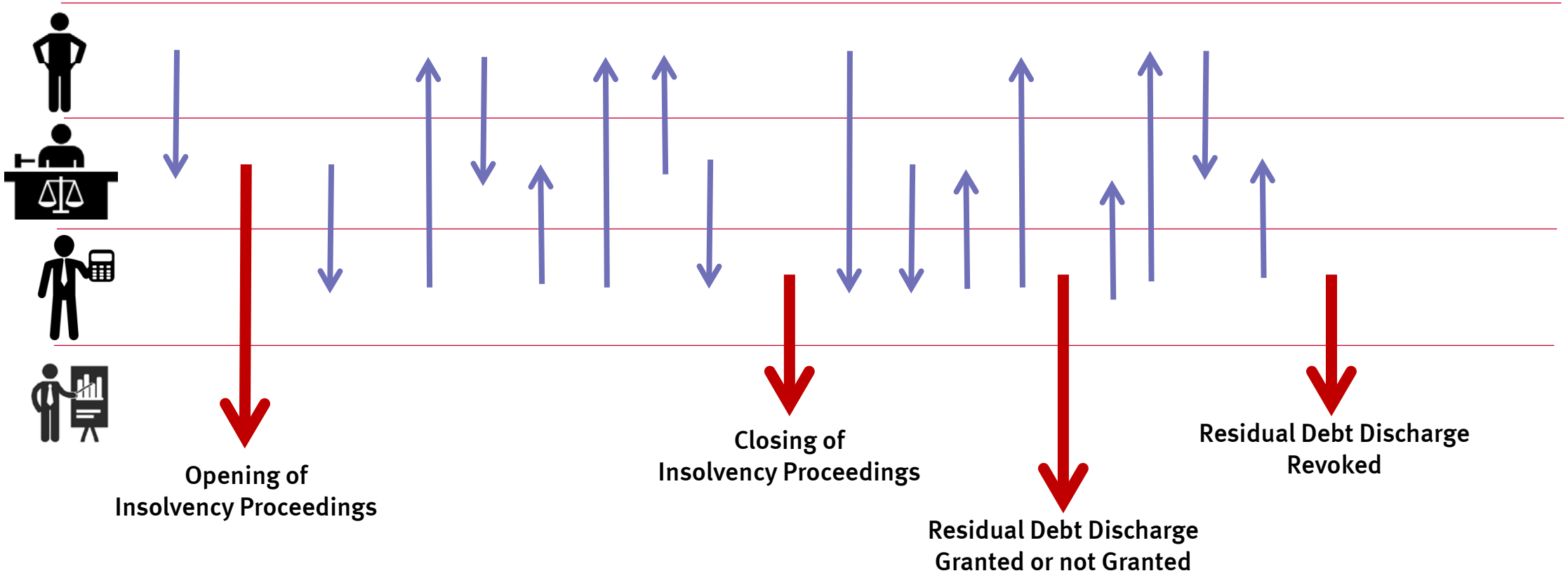
Debtor

Insolvency Court

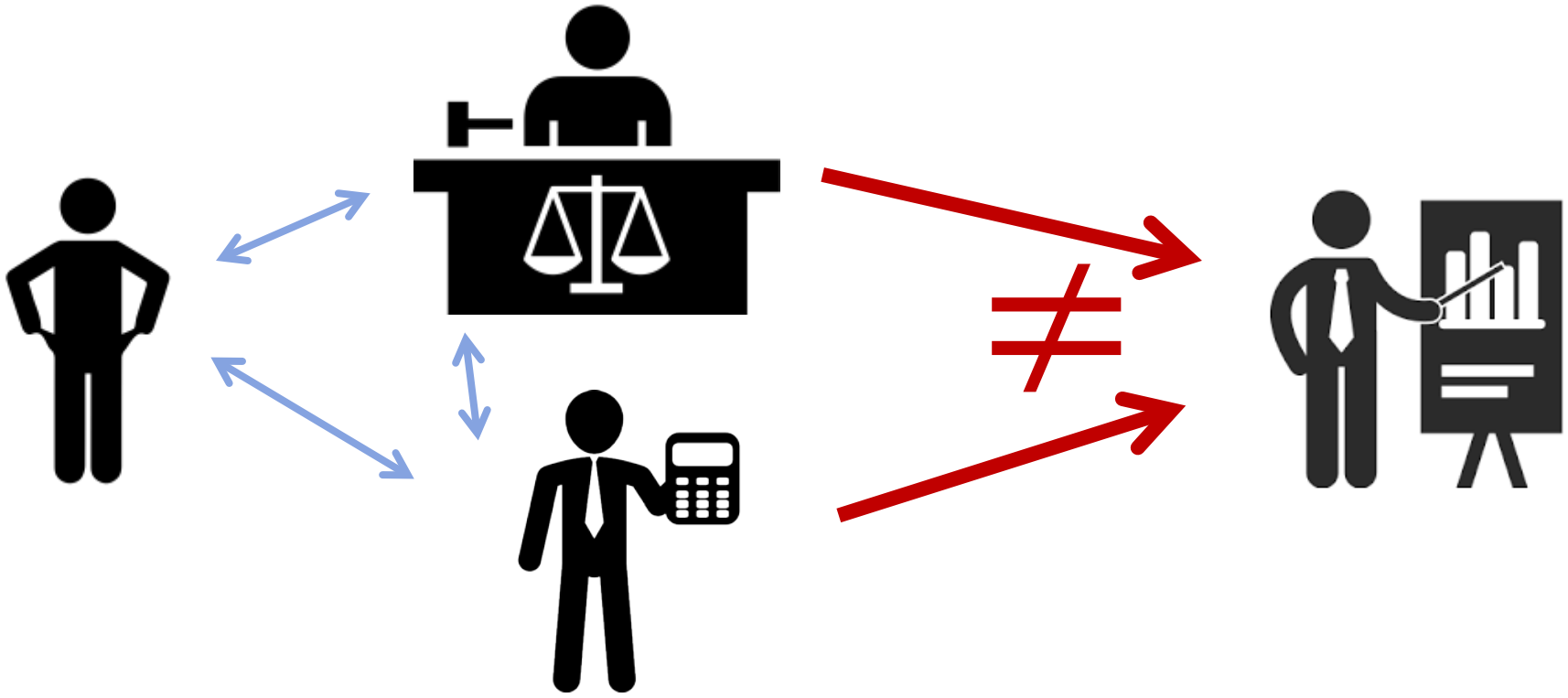
Insolvency Administrator

Official Statistics

Insolvency Proceedings and Official Statistics

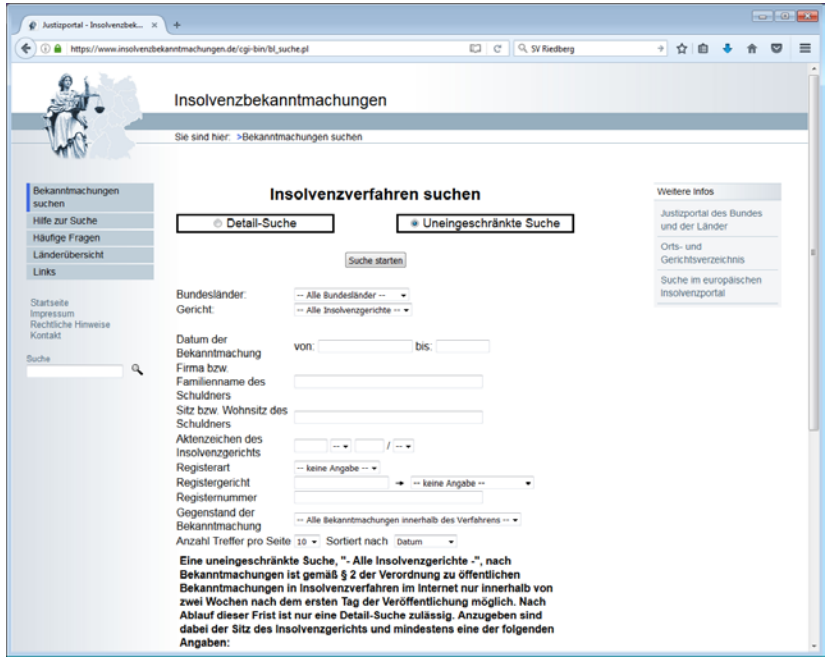


Problem: Two Respondents on One Case

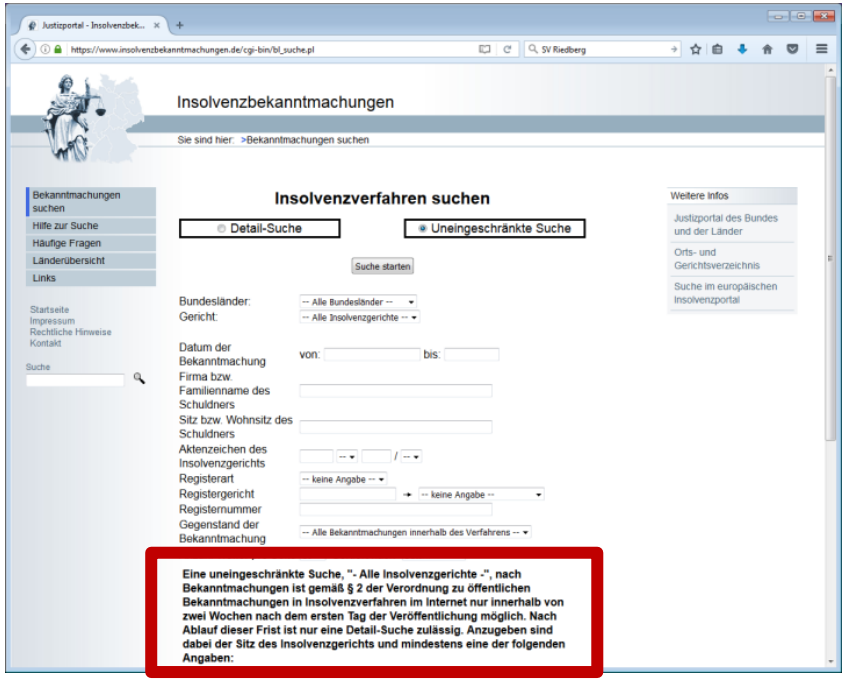


A Further Source:
Publication Obligations of Insolvency Courts

www.insolvenzbekanntmachungen.de



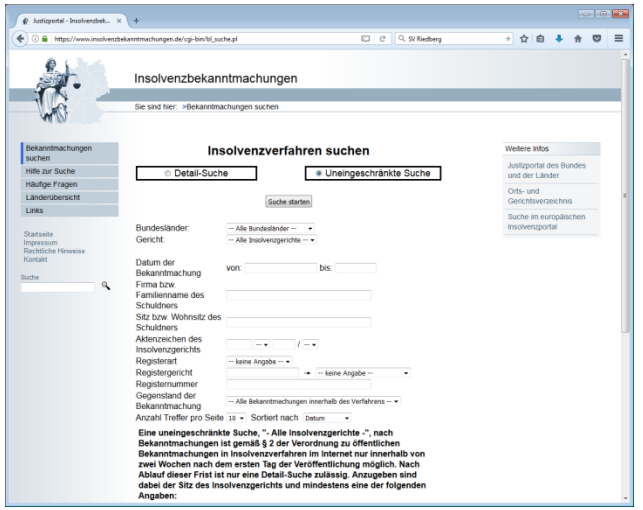
Unrestricted Search



An unrestricted search, "- All insolvency courts -", by announcements is possible according to § 2 of the regulation to public announcements in insolvency proceedings in the Internet only within two weeks after the first day of publication. After expiry of this period, only a detailed search is permitted.

Unrestricted Search

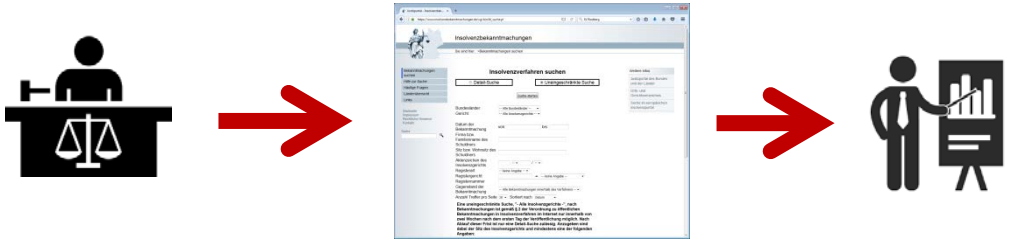
An unrestricted search, "- All insolvency courts -", by announcements is possible according to § 2 of the regulation to public announcements in insolvency proceedings in the Internet only within two weeks after the first day of publication. After expiry of this period, only a detailed search is permitted.



To achieve a complete picture of the opened proceedings, one needs to search every two weeks

⇨ **Cron Job**

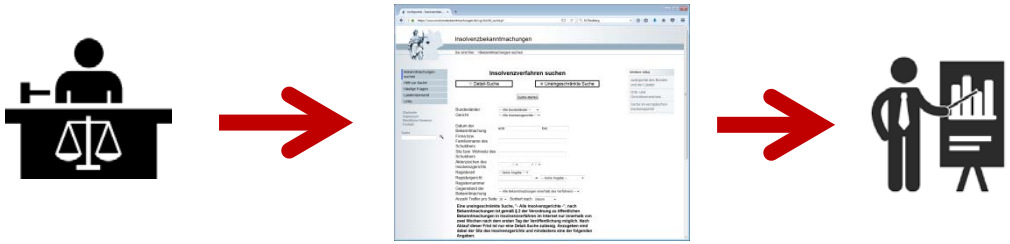
Data Collection



Objective: Complete list of identifiers (case number, court) of opened proceedings, as well as of proceedings with statistically relevant decisions.

Path: R-script that, within the two-week period, searches the texts of the notices in the unrestricted search for certain keywords and notes the corresponding case/court identifiers.

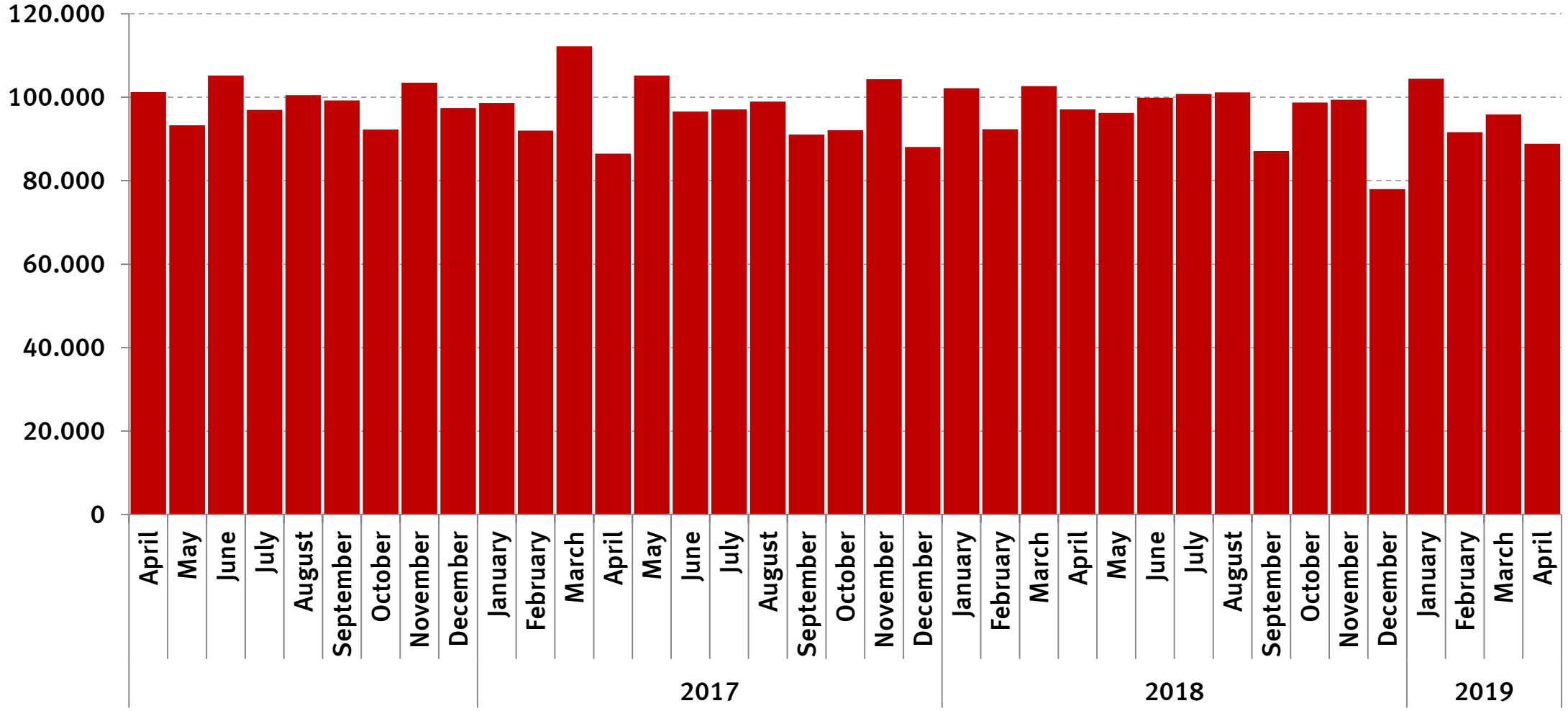
Data Collection



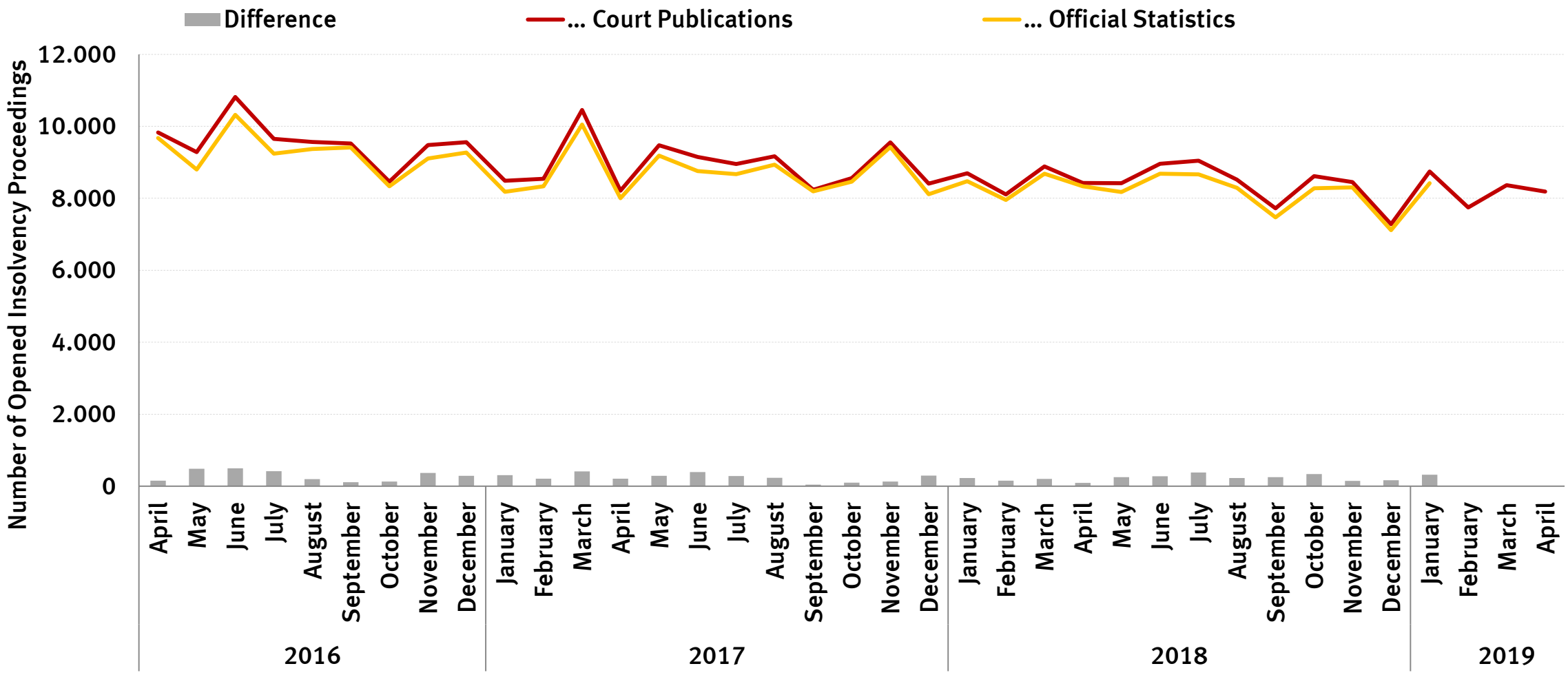
First collection in March 2016. Since then, full picture of openings and monthly lists of relevant procedures for which notifications should be received.

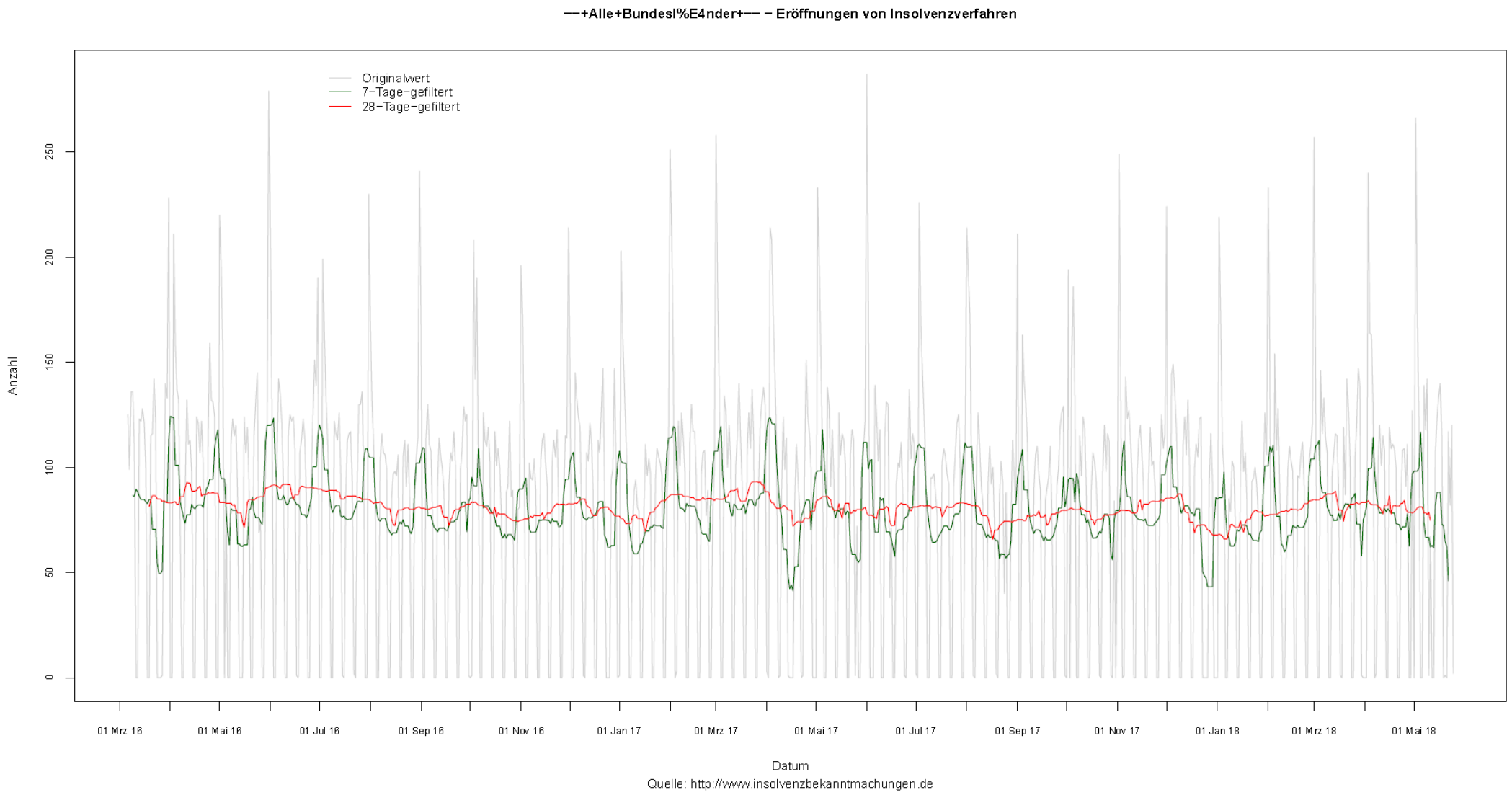
Currently about 3.7 million decisions have been analysed …

Number of Court Publications



Number of Opened Insolvency Proceedings ...




--+Alle+Bundesl%E4nder+-- – Eröffnungen von Insolvenzverfahren
Originalwert
7-Tage-gefiltert
28-Tage-gefiltert
Anzahl
0
50
100
150
200
250
01 Mrz 16
01 Mai 16
01 Jul 16
01 Sep 16
01 Nov 16
01 Jan 17
01 Mrz 17
01 Mai 17
01 Jul 17
01 Sep 17
01 Nov 17
01 Jan 18
01 Mrz 18
01 Mai 18
Datum
Quelle: http://www.insolvenzbekanntmachungen.de

What to do with the Data?

Main Goal:

Use data to get information about court decisions that should trigger a report by the insolvency administrators …

… so we can remind them that a report is due.

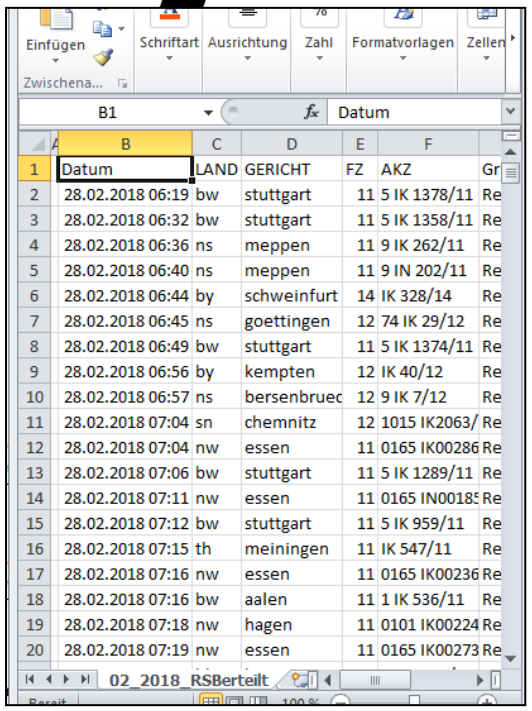
What to do with the Data?

(.*)

Identification of relevant court decisions is done by applying a set of regular expressions to the full text of the decisions.

Resulting lists are reasonably short and can be matched with incoming reports.

Staff can use the information to contact insolvency administrators efficiently.



	B	C	D	E	F
1	Datum	LAND	GERICHT	FZ	AKZ
2	28.02.2018 06:19	bw	stuttgart	11	5 IK 1378/11
3	28.02.2018 06:32	bw	stuttgart	11	5 IK 1358/11
4	28.02.2018 06:36	ns	meppen	11	9 IK 262/11
5	28.02.2018 06:40	ns	meppen	11	9 IN 202/11
6	28.02.2018 06:44	by	schweinfurt	14	IK 328/14
7	28.02.2018 06:45	ns	goettingen	12	74 IK 29/12
8	28.02.2018 06:49	bw	stuttgart	11	5 IK 1374/11
9	28.02.2018 06:56	by	kempten	12	IK 40/12
10	28.02.2018 06:57	ns	bersenbruec	12	9 IK 7/12
11	28.02.2018 07:04	sn	chemnitz	12	1015 IK2063/
12	28.02.2018 07:04	nw	essen	11	0165 IK00286
13	28.02.2018 07:06	bw	stuttgart	11	5 IK 1289/11
14	28.02.2018 07:11	nw	essen	11	0165 IN0018
15	28.02.2018 07:12	bw	stuttgart	11	5 IK 959/11
16	28.02.2018 07:15	th	meiningen	11	IK 547/11
17	28.02.2018 07:16	nw	essen	11	0165 IK00236
18	28.02.2018 07:16	bw	aalen	11	1 IK 536/11
19	28.02.2018 07:18	nw	hagen	11	0101 IK00224
20	28.02.2018 07:19	nw	essen	11	0165 IK00273

What we cannot do?

Privacy matters.

Court decisions contain personal data of debtors and insolvency administrators. Therefore full texts of court decisions are not stored.

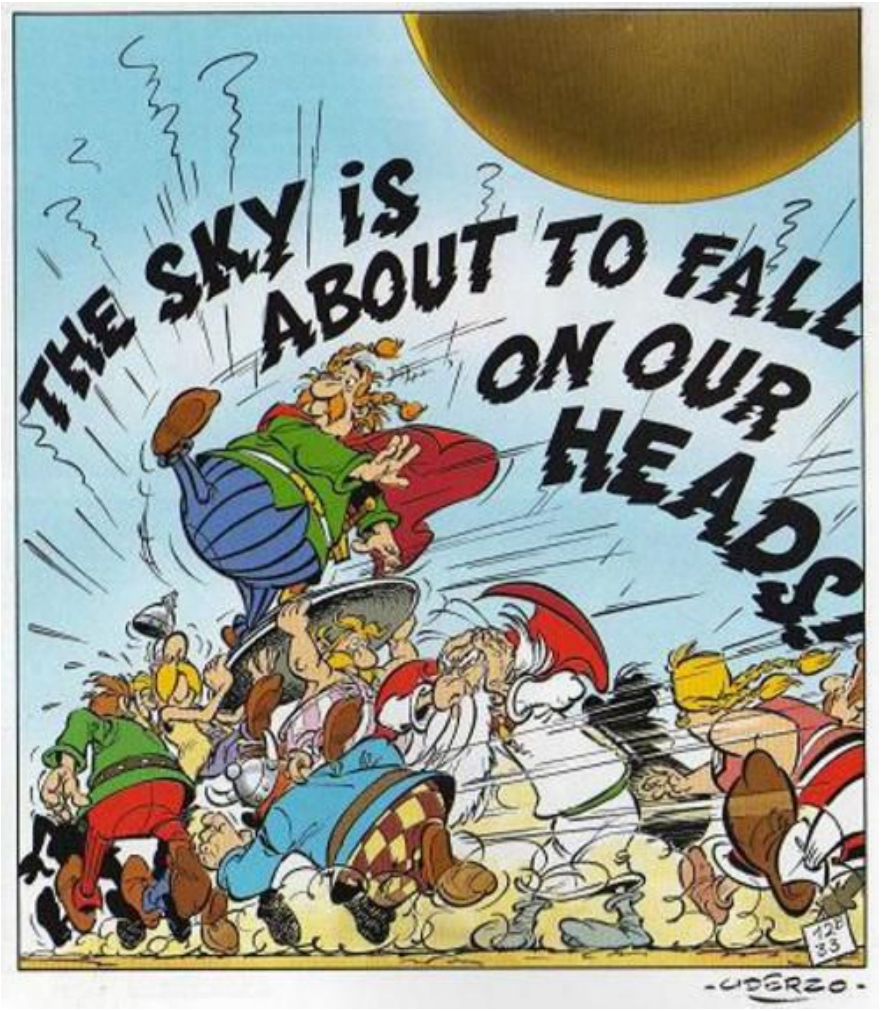
Only Case-IDs as well as type and time of decision are stored.

Emojis by Noto Color Emoji

Pit Falls

Q: What if one day the service is relaunched?

A: A legal basis for direct data access needs to be formulated.



Picture by Albert Uderzo

Possible Improvements

Keyword search currently via regular expressions.
Are there any smarter ways to analyze text?

Since we have collected a training set, could it be done with supervised learning approaches?

Emojis by Noto Color Emoji

Contact

Joerg Feuerhake

joerg.feuerhake@destatis.de

Tel.: 0049 611 75 4116