New analysis features of the CRExplorer for identifying influential publications

Andreas Thor [1], Lutz Bornmann [2]
Werner Marx [3], Rüdiger Mutz [4]

[1] University of Applied Sciences for Telecommunications Leipzig, thor@hft-leipzig.de
[2] Administrative Headquarters of the Max Planck Society, bornmann@gv.mpg.de
[3] Max Planck Institute for Solid State Research, w.marx@fkf.mpg.de
[4] ETH Zürich, mutz@gess.ethz.ch

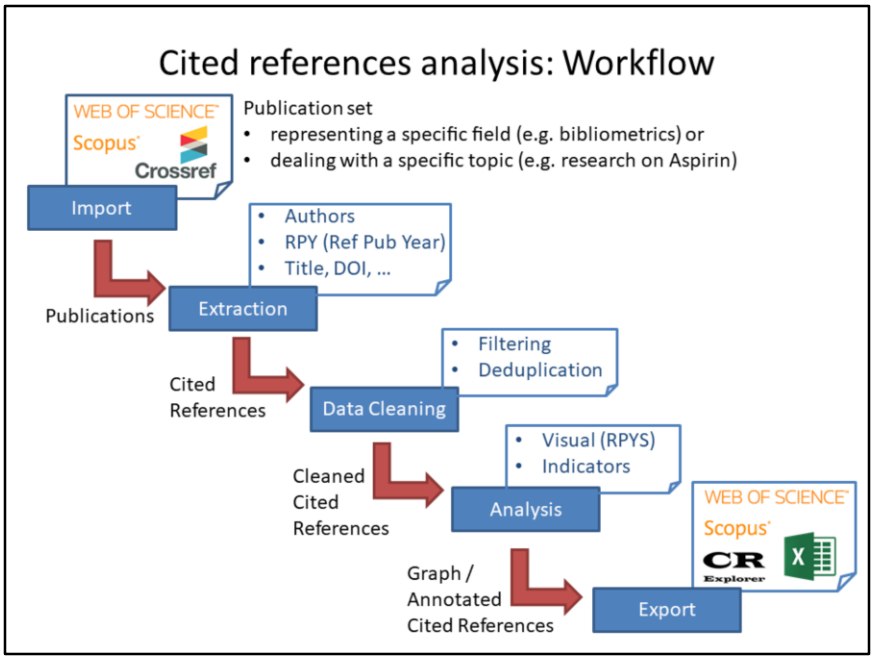
23rd International Conference on Science and Technology Indicators (STI 2018), 12-14 September 2018, Leiden

CRExplorer: A Tool for Cited References Analysis

- What are the **most important** publications in a research field? On which shoulders does the research stand?
- Identify those publications in a research field or on a specific topic …
  - … which have been **influential over many years** in the past
  - … were highly cited over a longer **time period or at certain time points** (shortly or several years after publication)

- **Cited References Explorer**
  - Workflow-based Data Analysis (GUI + Script language)
  - Import / export data formats: Scopus, Web of Science, CrossRef, CSV
  - Automatic Data Extraction and Data Cleaning
- Different types of analysis
  - Visual (Reference Publication Year Spectroscopy)
  - Indicator-based (Top-N, Sequence)

- http://www.crexplorer.net
  - Run (Java Web Start) or download (JAR), Guide + Datasets, Papers (e.g., use cases)


Cited references analysis: Workflow
WEB OF SCIENCE
Scopus
Crossref
Publication set
• representing a specific field (e.g. bibliometrics) or
• dealing with a specific topic (e.g. research on Aspirin)
Import
• Authors
• RPY (Ref Pub Year)
• Title, DOI, ...
Publications
Extraction
Cited References
• Filtering
• Deduplication
Data Cleaning
Cleaned Cited References
• Visual (RPYS)
• Indicators
Analysis
WEB OF SCIENCE
Scopus
CR Explorer
Graph / Annotated Cited References
Export

Pre-Processing

- **Data Extraction** of bibliographic information of Cited References (Strings)
  - Regular Expressions (Patterns), Integrity checks

CR	RPY	AU_L	J_N	VOL	PAG	DOI
Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569, DOI 10.1073/pnas.0507655102	2005	Hirsch	P NATL ACAD SCI USA	102	16569	10.1073/PNAS.0507655102
Kosmulski M., 2006, ISSI NEWSLETTER, V2, P4	2006	Kosmulski	ISSI NEWSLETTER	2	4	
Jin BH, 2007, CHINESE SCI BULL, V52, P855, DOI 10.1007/s11434-007-0145-9	2007	Jin	CHINESE SCI BULL	52	855	10.1007/S11434-007-0145-9

CR	RPY	AU_L	J_N	TI	VOL	PAG
Bornmann, L., Daniel, H.-D., What do citation counts measure? A review of studies on citing behavior (2008) Journal of Doc...	2008	Bornmann	Journal of Documentation	What do citation counts me...	64	45
Bornmann, L., Scientific peer review (2011) Annual Review of Information Science and Technology, 45, pp. 199-245	2011	Bornmann	Annual Review of Inform...	Scientific peer review	45	199
Hirsch, J.E., An index to quantify an individual's scientific research output (2005) Proceedings of the National Academy of S...	2005	Hirsch	Proceedings of the Natio...	An index to quantify an indi...	102	16569

- **Data Filtering**
  - by Citing Publication Year (PY)
  - by Reference Publication Year (RPY)
  - by Number of Citations (N_CR)
- **Data Cleaning (Deduplication)**
  - Detecting and merging duplicates is important for high-quality data analysis

2x Hirsch-CR aus unterschiedlichen Quellen
Vor Filterung nach N_CR am besten Deduplication

Deduplication (Disambiguation): Clustering + Merge

- Different **variants** of the same Cited Reference
  - due to typos, missing bibliographic information, different abbreviation styles, ...
- **Clustering** based on string similarity (author, title) and year
  - Configuration: Threshold (e.g., 80%) + use of DOI, Volume and Page Number

ID	CR	RPY	N_CR
95	Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569, DOI 10.1073/pnas.0507655102	2005	155
8664	Hirsch J. E., 2005, P NATL ACAD SCI, V102, P16569	2005	1
8898	Hirsch J. E, 2005, P NATL ACAD SCI USA, V102, P16569	2005	1
6465	Hirsch J. E., 2005, P NATL ACAD SCI USA, V102, P16569	2005	1
8896	Jin B. H, 2007, CHINESE SCI BULL, V52, P855	2007	1
13	Jin BH, 2007, CHINESE SCI BULL, V52, P855, DOI 10.1007/s11434-007-0145-9	2007	37
65	Kosmulski M., 2006, ISSI NEWSLETTER, V2, P4	2006	16
8453	Komulski M., 2006, ISSI NEWSLETTER, V2, P4	2006	1

- **Merging**: Cluster representative + Accumulation of N_CR

ID	CR	RPY	N_CR
95	Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569, DOI 10.1073/pnas.0507655102	2005	158
13	Jin BH, 2007, CHINESE SCI BULL, V52, P855, DOI 10.1007/s11434-007-0145-9	2007	38
65	Kosmulski M., 2006, ISSI NEWSLETTER, V2, P4	2006	17

Syntaktische Unterschiede

- Fehlende Angaben (z.B. DOI) bei #8664
- Punctuation (authors‘ initials)
- Heteorgenous Journal name (USA #8664 vs. #8898)
- Typos (Ko*s*mulski)

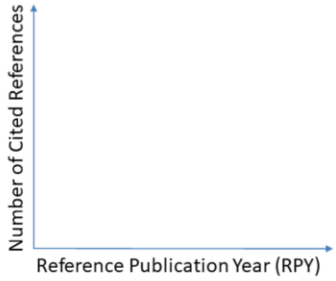
Cluster representative = Cited Reference of a cluster with the highest number of N_CR

Reference Publication Year Spectroscopy (RPYS)

- Method to analyze historical roots based on cited references within single research fields
- Analysis of the frequency with which references are cited in the publications in terms of the publication years of these CRs.

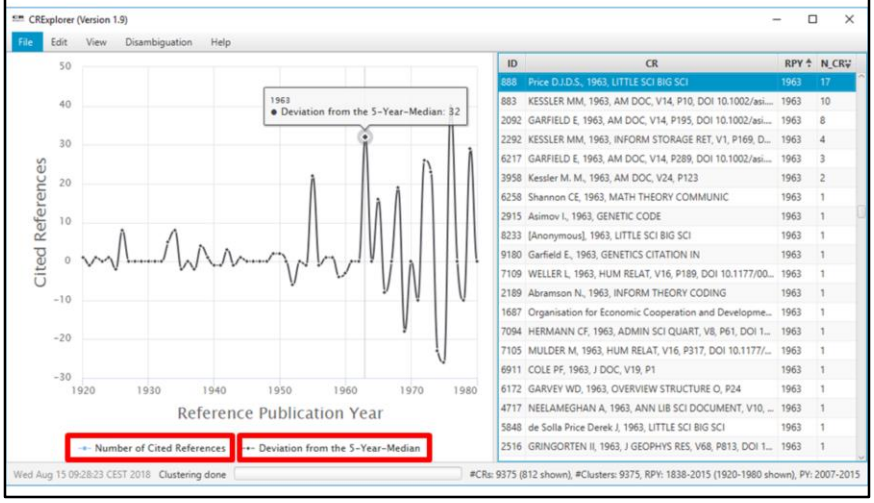
- **Spectrogram**

- Origins show up in the form of more or less pronounced peaks mostly caused by individual publications that are cited particularly frequently



RPYS: Example

- Scientometrics papers (2007-2015) → 9,375 Cited References



ID	CR	RPY	N_CR
888	Price D.J.D.S., 1963, LITTLE SCI BIG SCI	1963	17
883	KESSLER MM, 1963, AM DOC, V14, P10, DOI 10.1002/asi...	1963	10
2092	GARFIELD E, 1963, AM DOC, V14, P195, DOI 10.1002/asi...	1963	8
2292	KESSLER MM, 1963, INFORM STORAGE RET, V1, P169, D...	1963	4
6217	GARFIELD E, 1963, AM DOC, V14, P289, DOI 10.1002/asi...	1963	3
3958	Kessler M. M., 1963, AM DOC, V24, P123	1963	2
6258	Shannon CE, 1963, MATH THEORY COMMUNIC	1963	1
2915	Asimov I., 1963, GENETIC CODE	1963	1
8233	[Anonymous], 1963, LITTLE SCI BIG SCI	1963	1
9180	Garfield E., 1963, GENETICS CITATION IN	1963	1
7109	WELLER L, 1963, HUM RELAT, V16, P189, DOI 10.1177/00...	1963	1
2189	Abramson N., 1963, INFORM THEORY CODING	1963	1
1687	Organisation for Economic Cooperation and Developme...	1963	1
7094	HERMANN CF, 1963, ADMIN SCI QUART, V8, P61, DOI 1...	1963	1
7105	MULDER M, 1963, HUM RELAT, V16, P317, DOI 10.1177/...	1963	1
6911	COLE PF, 1963, J DOC, V19, P1	1963	1
6172	GARVEY WD, 1963, OVERVIEW STRUCTURE O, P24	1963	1
4717	NEELAMEGHAN A, 1963, ANN LIB SCI DOCUMENT, V10, ...	1963	1
5848	de Solla Price Derek J, 1963, LITTLE SCI BIG SCI	1963	1
2516	GRINGORTEN II, 1963, J GEOPHYS RES, V68, P813, DOI 1...	1963	1

<u>Übergang zur nächsten Folie</u>: N_CR in Tabelle … aber wie verteilt sicht das?

Number of Publication Years (PYs)

- **N_PYEARS** = No. of PYs in which the CR has been cited
- N_PY_{RPY} = No. of PYs in which a CR from RPY has been cited
- **PERC_PYEAR** = N_PYEARS / N_PY_{RPY}

Scientometrics papers (1978-2016)
→ 39 publication years

In how many publication years has Lotka (1926) been cited?

… has not been cited in 1978, 1979, and 1983

Cited reference	N_CR
Lotka (1926) *The frequency distribution of scientific productivity*	155
Garfield (1979) *Citation indexing: its theory and application in …*	151
Small, H., (1973) *Co-citation in the scientific literature: A new measure …*	162
Katz, J.S., Martin, B.R., (1997) *What is research collaboration?*	171

$$\frac{N\_PYEARS}{N\_PY_{1926}} = \frac{36}{36}$$

$$\frac{N\_PYEARS}{N\_PY_{1979}} = \frac{34}{37}$$

$$\frac{N\_PYEARS}{N\_PY_{1973}} = \frac{33}{39}$$

$$\frac{N\_PYEARS}{N\_PY_{1997}} = \frac{20}{20}$$

Top-N

- **Relative comparison** of Cited References w.r.t. the Reference Publication Year (RPY) and the Publication Year (PY) of citing publications
- **N_TOP10** = Number of publication years in which the CR has been in the Top-10% of all CRs of the same RPY

Scientometrics papers (1978-2016)
→ 39 publication years

Cited reference	N_CR	N_PYEARS
Lotka (1926) *The frequency distribution of scientific productivity*	155	36
Garfield (1979) *Citation indexing: its theory and application in …*	151	34
Small, H., (1973) *Co-citation in the scientific literature: A new measure …*	162	33
Katz, J.S., Martin, B.R., (1997) *What is research collaboration?*	171	20

Was Lotka (1926) a higly cited paper in each citing publication year?

If the Cited Reference has been cited in a publication year, it was in the top 10% of all cited references of its RPY.

Sequence Types

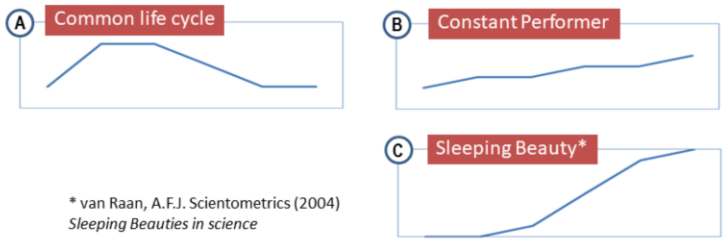
- **SEQUENCE** and **TYPE**: Distribution of citations over time and identification of common time-series patterns

Is the number of citations per publication year increasing or decreasing over time?

- Example: Number of citations per publication year …

	2001	2002	2003	2004	2005	2006	Σ
A	2	6	6	4	2	2	22
B	2	3	3	4	4	5	21
C	0	0	1	4	7	8	20
Σ	4	9	10	12	13	15	63

- … induce time sequence patterns for cited references



* van Raan, A.F.J. Scientometrics (2004) *Sleeping Beauties in science*

Sequence Computation

- **Observed** values (number of citation per publication year)

	2001	2002	2003	2004	2005	2006	Σ
A	2	6	6	4	2	2	22
B	2	3	3	4	4	5	21
C	0	0	1	4	7	8	20
Σ	4	9	10	12	13	15	63

- **Expected** values

$$22 \cdot \frac{9}{63} \approx 3.1$$

	2001	2002	2003	2004	2005	2006	Σ
A	1.4	3.1	3.5	4.2	4.5	5.2	22
B	1.3	3.0	3.3	4.0	4.3	5.0	21
C	1.3	2.9	3.2	3.8	4.1	4.8	20
Σ	4	9	10	12	13	15	63

- **z-value**: Standard Normal Distribution (mean=0, std. dev.=1)

$$\frac{6 - 3.1}{\sqrt{3.1}} \approx 1.6$$

	2001	2002	2003	2004	2005	2006
A	0.5	1.6	1.3	-0.1	-1.2	-1.4
B	0.6	0.0	-0.2	0.0	-0.2	0.0
C	-1.1	-1.7	-1.2	0.1	1.4	1.5

Sequence Types

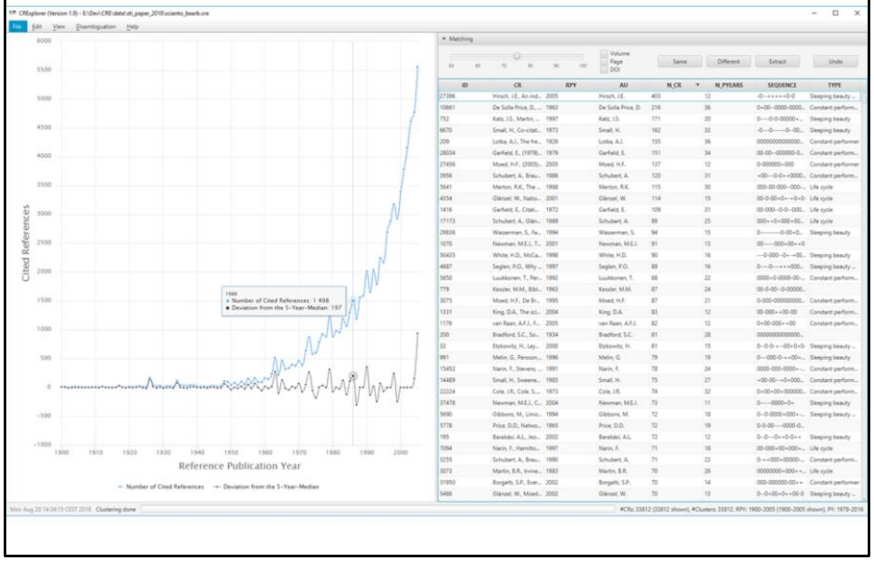
- Classification of Cited References based on z-value patterns

	2001	2002	2003	2004	2005	2006
A	0.5	1.6	1.3	-0.1	-1.2	-1.4
B	0.6	0.0	-0.2	0.0	-0.2	0.0
C	-1.1	-1.7	-1.2	0.1	1.4	1.5

- Scientometrics dataset (1978-2016)

CR	RPY	N_CR	SEQUENCE	TYPE
Lotka, A.J., The frequency distribution of scientific productivi...	1926	155	0000000000000000000000000000-0000000000	Constant performer
Garfield, E., (1979) Citation Indexing: Its Theory and Applicati...	1979	151	00-00--000000-000+0++00000-0+00+0+00-0	Constant performer + Life cycle
Small, H., Co-citation in the scientific literature: A new meas...	1973	162	-0---0------0--0000-00-0000++00+0+++0++	Sleeping beauty
Katz, J.S., Martin, B.R., What is research collaboration? (1997)...	1997	171	0----0-0-00000++0+++	Sleeping beauty

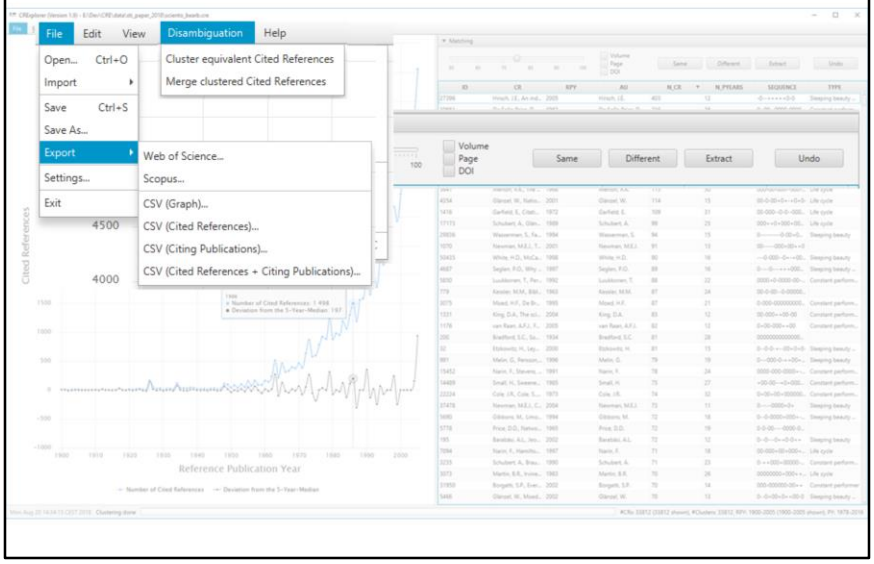
CRExplorer: User Interface



Übergang von vorheriger:

- This is how the result looks like in CRExplorer
- To achieve this a user performs several steps in the workflow …

Interactive Workflow using GUI



Übergang zu nächster:

* Instead of a sequence of clicks a user can encode the workflow definition in a script program

CRExplorer's Script Language

- Script language for workflow automation

- **Reproducibility** of results
- Same analysis procedure for different publication sets
- Processing **large** files

```
importFile (
   dir: "E:/data/input/",
   type: "WOS",
   sampling: "RANDOM",
   maxCR: 1000
)

cluster(
   threshold: 0.8,
   volume: true,
   page: true,
   DOI: false
)
merge ()

removeCR (RPY: [0, 1995])

exportFile (
   file: "E:/data/output.csv",
   type: "CSV_CR"
)
```

Summary + Future Work

- CRExplorer: A Tool for Cited References Analysis

- Data Extraction + Data Cleaning (Deduplication)
- Reference Publication Year Spectroscopy (RPYS)
- Indicators (TOP-N, Sequence)
- Script-based Automation

- New import / export formats
- User-defined indicators
- Help wanted ☺

Website: http://www.crexplorer.net
Source: https://github.com/andreas-thor/cre

Thank you!