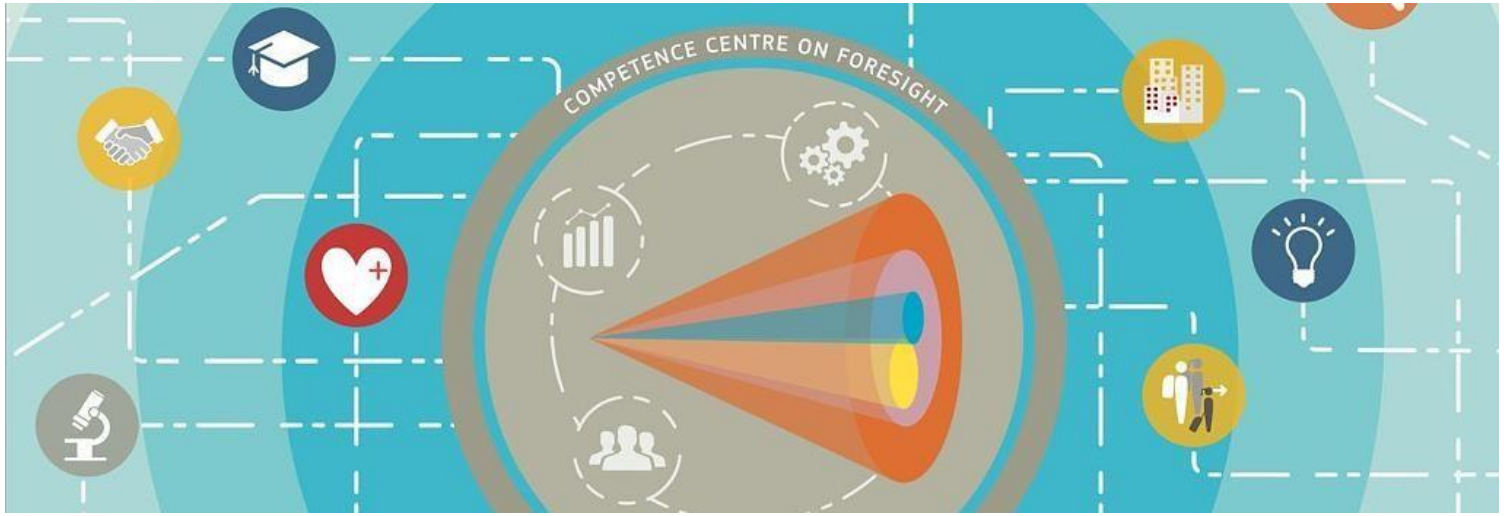
The use of text mining for technology foresight

Software performance in comparison with expert review

20 October 2021

Alberto Moro
Joint Research Centre of the European Commission

Competence Centre on Foresight



The Competence Centre on Foresight **supports EU policy making** by

- providing **strategic** and **future-oriented** inputs,
- developing an **anticipatory culture** inside the European Commission,
- experimenting and developing quantitative and qualitative **methods and tools** for foresight

European Commission

Scope: to identify emerging technologies

to support the Directorate General of Research and Technological Development (DG RTD) of the European Union
In **identifying Future and Emerging Technologies (FETs)** for designing research programs

Case studies in the sectors of:

- Photovoltaics
- Wind power
- Ocean and tidal energy
- Hydropower

http://ec.europa.eu/programmes/horizon2020/en/h2020-section/future-and-emerging-technologies

European Commission

Emerging = low Technology Readiness Level

	TRL 9	Actual system proven in operational environment
System development	TRL 8	System complete and qualified
	TRL 7	System prototype demonstrated in operational environment
	TRL 6	Technology demonstrated in relevant environment
Technology development	TRL 5	Technology validated in relevant environment
	TRL 4	Technology validated in lab (“ugly” prototype)
	TRL 3	Experimental proof of concept
Basic research	TRL 2	Technology concept formulated
	TRL 1	Basic principles observed

https://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level

The Technology Readiness Level (TRL) scale, defined by NASA in the 70s, is now adopted also in research programs to assess the technology/project maturity

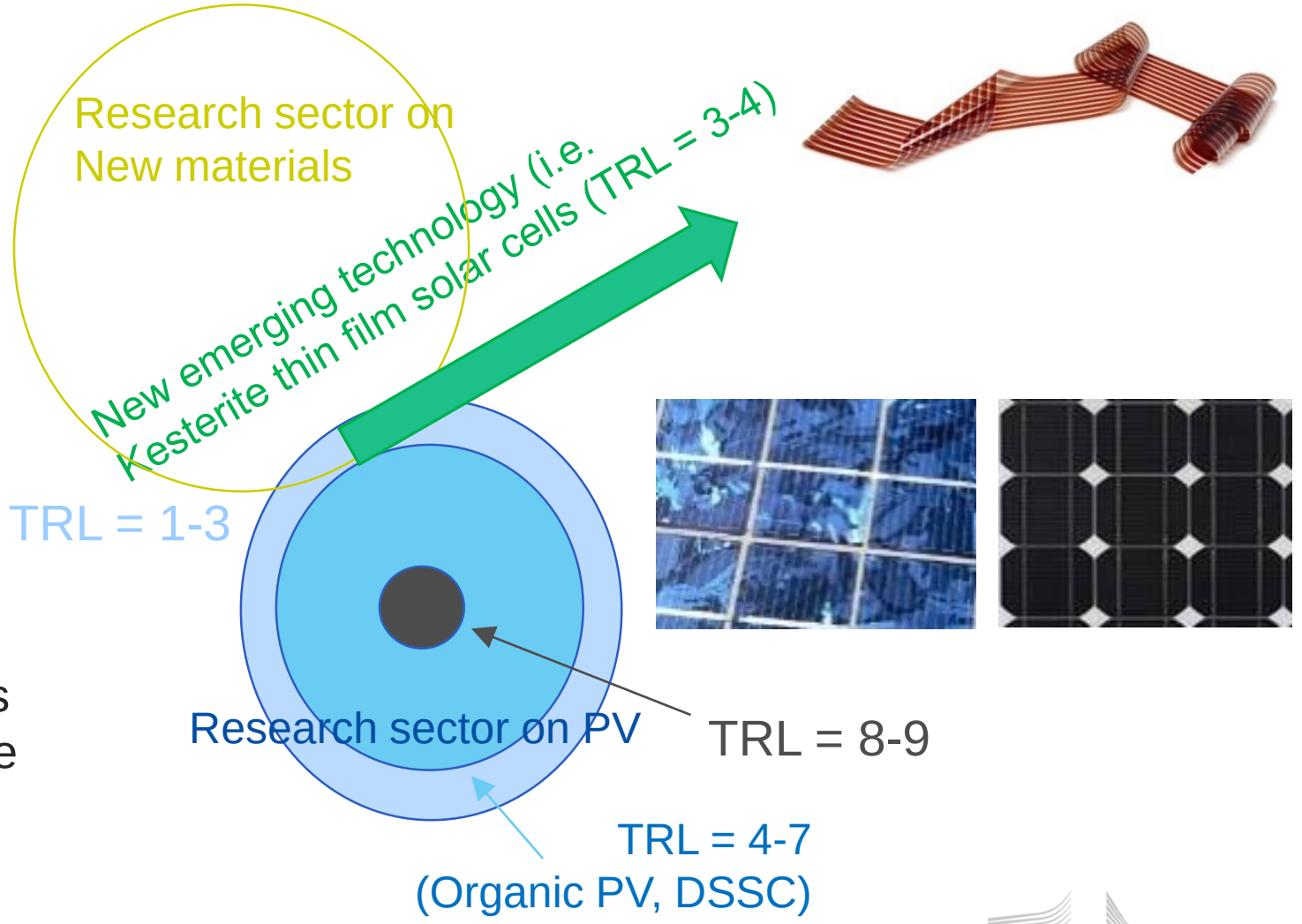
Emerging technologies can be **defined** on the base of their maturity level.
In example: **TRL=1-4**

Example of emerging technology in the PV

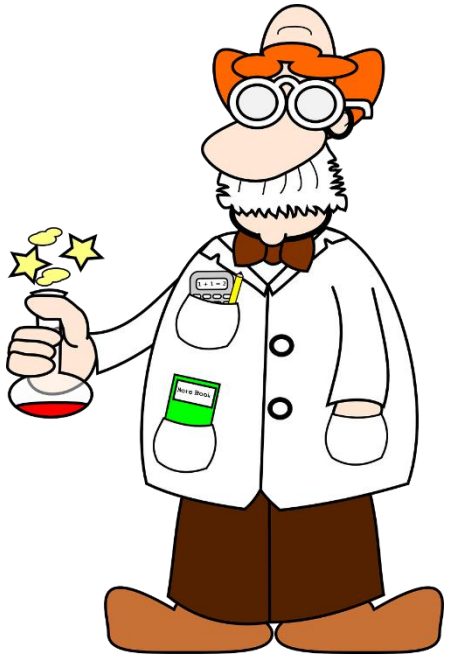
Where to look for emerging technologies? (example for solar photovoltaics – PV)

Relevant research is performed at the intersection of disciplines (cross-fertilisation of research)

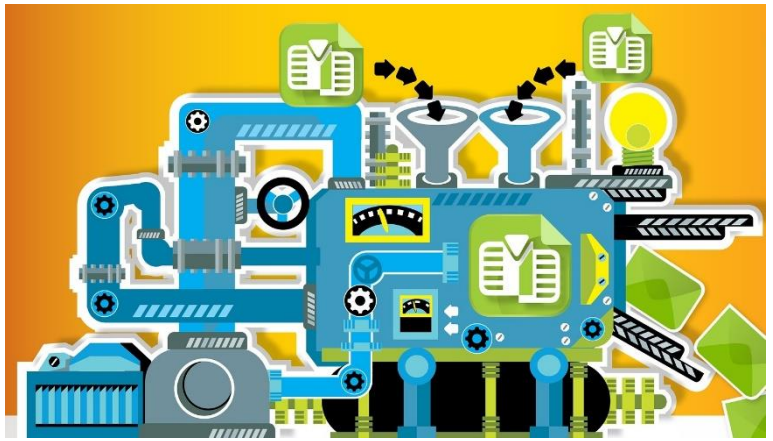
Experts' specialisation:
Most of the mainstream PV experts work on higher TRLs, so can ignore low TRL technologies



Two main methods to identify FETs



"Vs"



Expert review

+ more reliable (IF experts well selected)
+ quality results
- Cost and Time

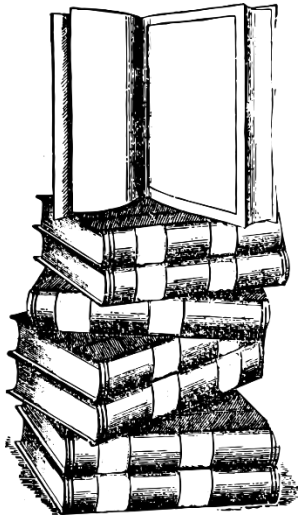
Bibliometric / Text mining software

(can feed an expert review)

+ Less expert support required
+ Cheaper (cost ~ 1/10 of expert review) and quicker
- Less information

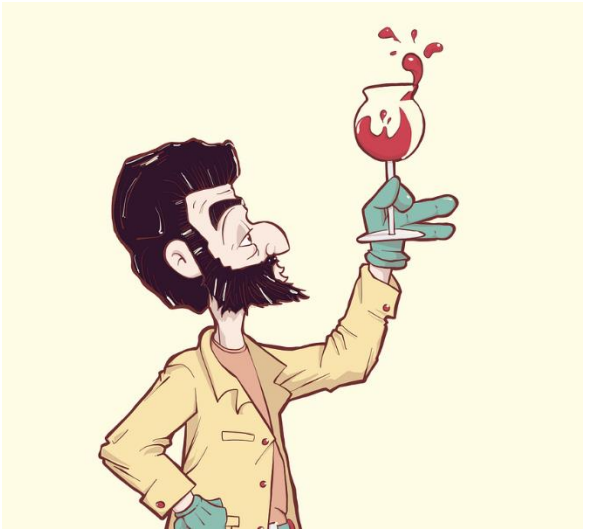
Expert review

1) To identify and/or assess technologies is necessary to **identify and involve**/recruit suitable (usually external) **experts** (requires Time, Cost)



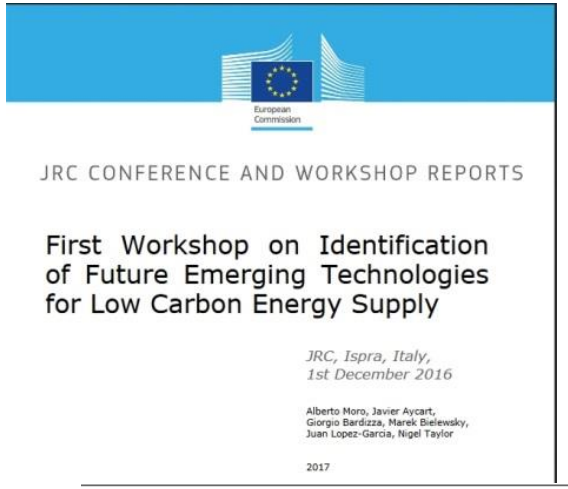
Not easy to find the right experts.
Experts must have:

- **Unbiased** approach (not supporting only his research field)
- **Ability to «sniff»** pro and **cons** of technologies
- **Suitable** background and working experience (TRL)



2) Exercise of collective intelligence merging (e.g. workshop)

Expert review can produce quality outputs



- https://op.europa.eu/en/publication-detail/-/publication/d22ece40-6a9a-11e7-b2f2-01aa75ed71a1
- https://www.sciencedirect.com/science/article/pii/S1364032119304782
- https://www.sciencedirect.com/science/article/pii/S1364032119304575
- https://publications.jrc.ec.europa.eu/repository/handle/JRC112635

European Commission

Text mining - bibliometric software and data

Several bibliometric software (also open source) are available. In example:

TIM Tools for Innovation Monitoring

European Commission
Joint Research Centre

http://www.timanalytics.eu/index.html

vantagepoint Search Technology

VOSviewer

Biblioshiny, CiteSpace

(…)

Bibliometric Software can look into data bases of publications (e.g. Scopus), patents (e.g. Patstat), projects (e.g. CORDIS),…

European Commission

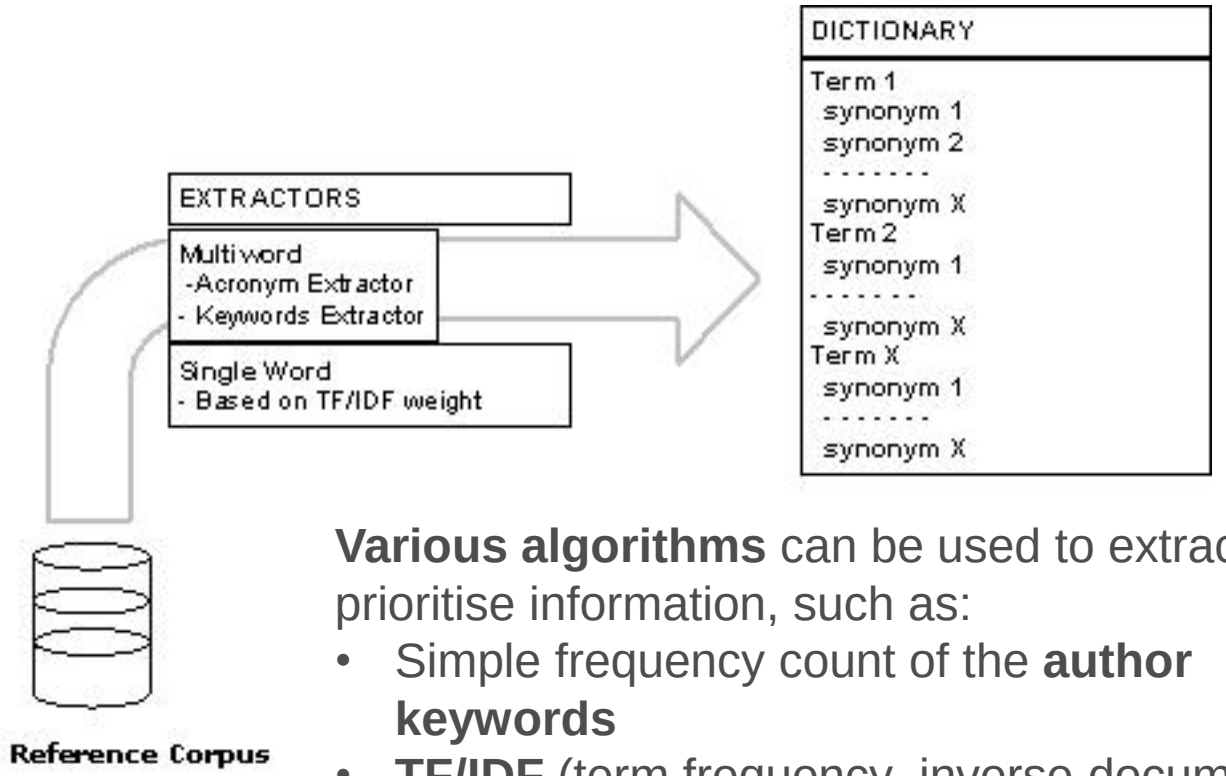
Bibliometric software and text mining

Case studies, here, were analysed with **TIM** (Tools for Innovation Monitoring), the JRC's bibliometric software with text mining features

TIM Tools for Innovation Monitoring

European Commission
Joint Research Centre

http://www.timanalytics.eu/index.html



(mainly) Scopus

Various algorithms can be used to extract and prioritise information, such as:

- Simple frequency count of the **author keywords**
- **TF/IDF** (term frequency–inverse document frequency) on document's title, abstract and author keywords

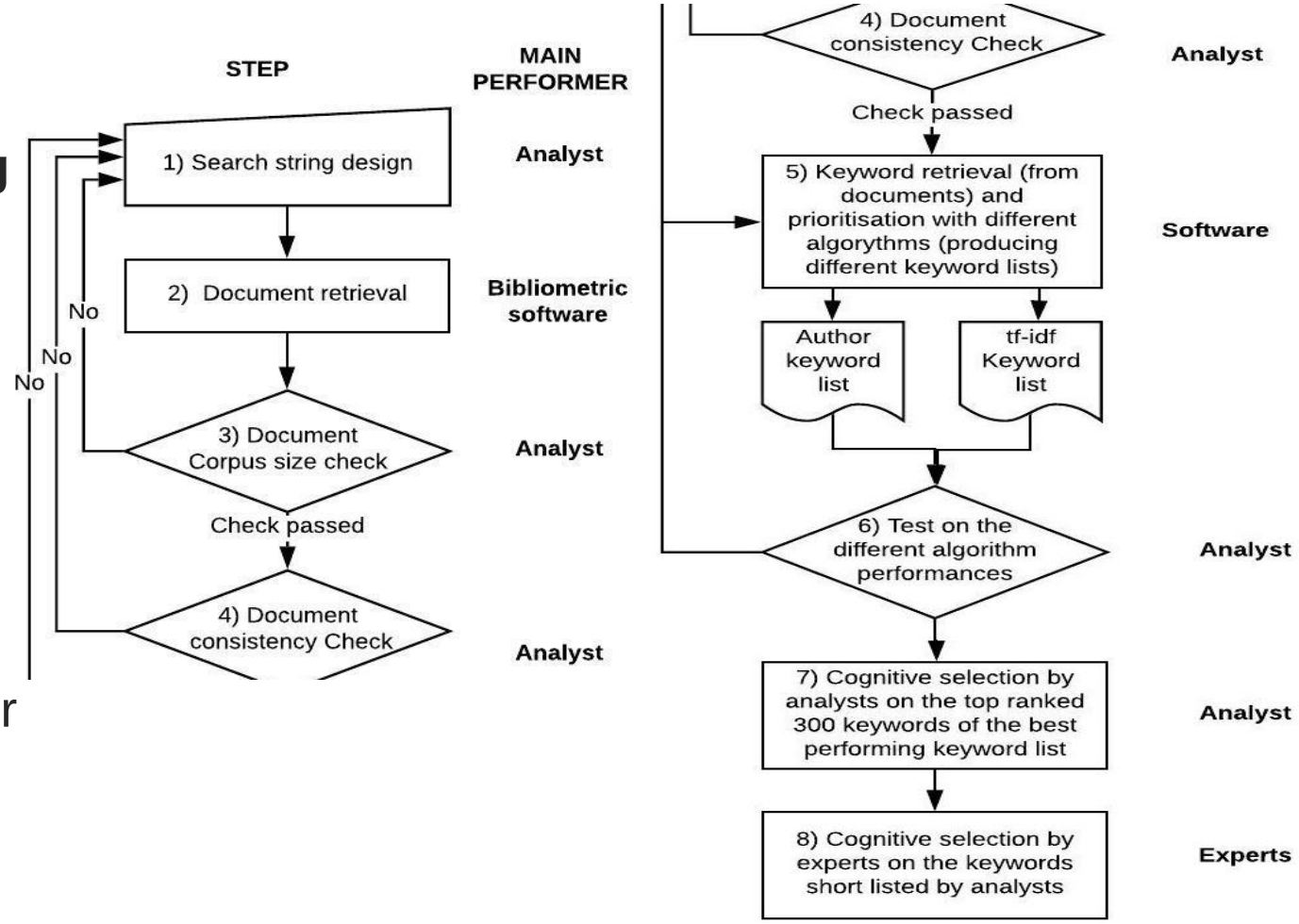
European Commission

Process for identifying technologies

The process *in nuce:*

- The Analyst designs a **Search string**
- The Bibliometric software retrieves **documents** from a data base
- Algorithms extract and prioritise **keywords** from the documents
- Keywords can offer clues for **technology** identification

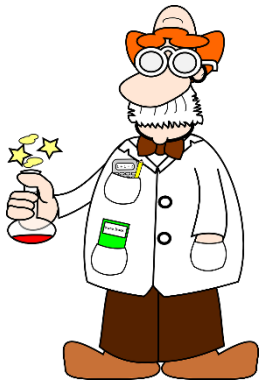
Along the process various Data quality checks by **experienced analysts** and/or experts (quality and quantity of documents), Search string **refining**, different algorithm tests.



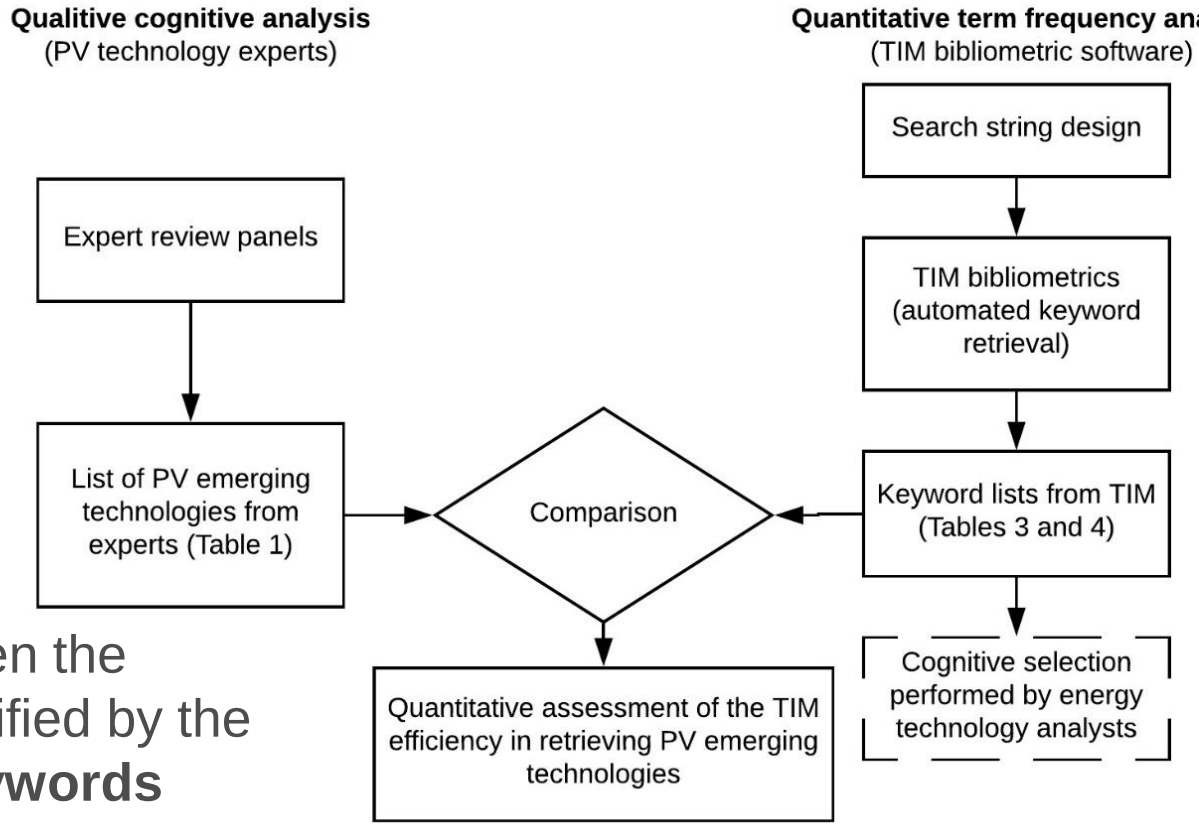
European Commission

Performance comparison: expert Vs software

Two independent exercises: expert review and (Vs) bibliometric software analysis

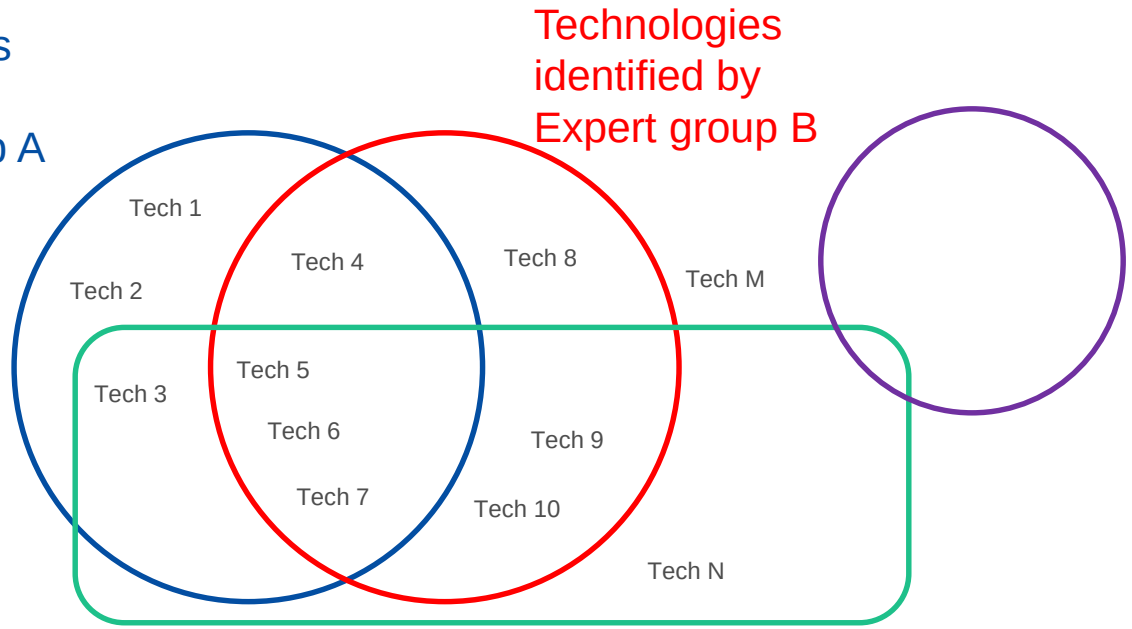


Reference



Comparison between the **technologies** identified by the experts and the **keywords** retrieved by the software

Uncertainty from expert groups



Software recall over expert outputs can vary:

Recall(A) = 4/7 = 57%
Recall(B) = 5/7 = 71%
Recall(AUB) = 6/10
Recall = (6+N)/(10+N+M)

Different technologies can be identified by different expert groups (**number** of experts, expertise/**background**, workshop **dynamics**, **time** available for the discussion…)

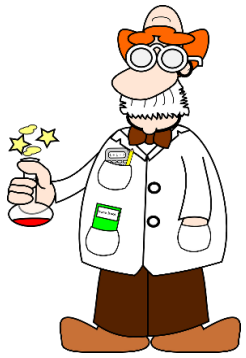
Comparison hypothesis

Calibration of the software Vs technologies identified by the Expert

A number of Master Keywords – MKs is considered as reference.

(In our case studies:

- MK =16 for the PV case study;
- MK = 15 for the wind power case
- MK = 9 for the ocean and tidal case)



"The relevant FETs in the PV sector are:

Kesterite thin film solar cells (or CZTS)
Perovskite thin film solar cells (chalcogenide)
Organic solar cells (OSC)
Dye-Sensitized Solar cells (DSSC)
Intermediate band solar cells (IBSC)
Plasmonic solar cells

Low-cost manufacturing processes, roll-to-roll and flexible substrates

Innovative multi-junction solar cells (also "multi junction")
Thermo-photovoltaics (or Thermal)

Innovative III-V compounds based solar cells (search for "III")

Photoelectrocatalytic devices (also "photocatalytic" also "photoelectrochemical")
Ferroelectric PV
Multiple exciton generation (MEG) solar cells
Hot carrier solar cells
Novel contacts for PV technologies
Solar cells from semiconductor foils

New photovoltaic materials via combinatorial and computational design (also: "modeling")

Bibliometric Noise

Among the keywords extracted by the bibliometric software we can find:

- Useful keywords directly representing **technologies** (e.g. DSSC, OSC)
- obvious terms (“solar cells”) to be discarded
- macro technologies (e.g. “thin film”) **to be discarded**
- clues (e.g. “Exciton” **maybe** meaning MEG) to be further explored,
- synonyms and terms with different spelling which **need cleaning**…

Rank	Clean keyword	Frequency
1	photovoltaics	893
2	solar cells (SC)	367
(…)	(…)	(…)
9	thin film (TF)	130
10	photovoltaic cell	121
(…)	(…)	(…)
13	dye sensitized solar cell (DSSC)	91
(…)	(…)	(…)
16	organic solar cells (OSC)	70
(…)	(…)	(…)
21	quantum dots (QD)	57
(…)	(…)	(…)
27	photovoltaic thermal	49
(…)	(…)	(…)
31	fullerene	43
32	nanostructure	42
(…)	(…)	(…)
38	nanowire	39
(…)	(…)	(…)
42	Perovskite solar cell	38
(…)	(…)	(…)
140	intermediate band (IB)	16
(…)	(…)	(…)
199	pv modelling	12
(…)	(…)	(…)
	kesterite	8
(…)	(…)	(…)

The clumping algorithm

The same concept can be represented by different keywords (synonyms, different spelling or capital letters) in a native author keyword list. In example, “kesterite” in the PV sector:

Rank	Native keywords	Frequency
(…)	(…)	(…)
313	Kesterite	5
(…)	(…)	(…)
3971	kesterites	1
(…)	(…)	(…)
4375	kesterite	1
(…)	(…)	(…)
9737	Kesterites	1
(…)	(…)	(…)

This lowers the **ranking** of a technology

The benefits from a clumping algorithm are twofold:

the clean or clumped keyword “kesterite” has a total of **8 occurrences**.

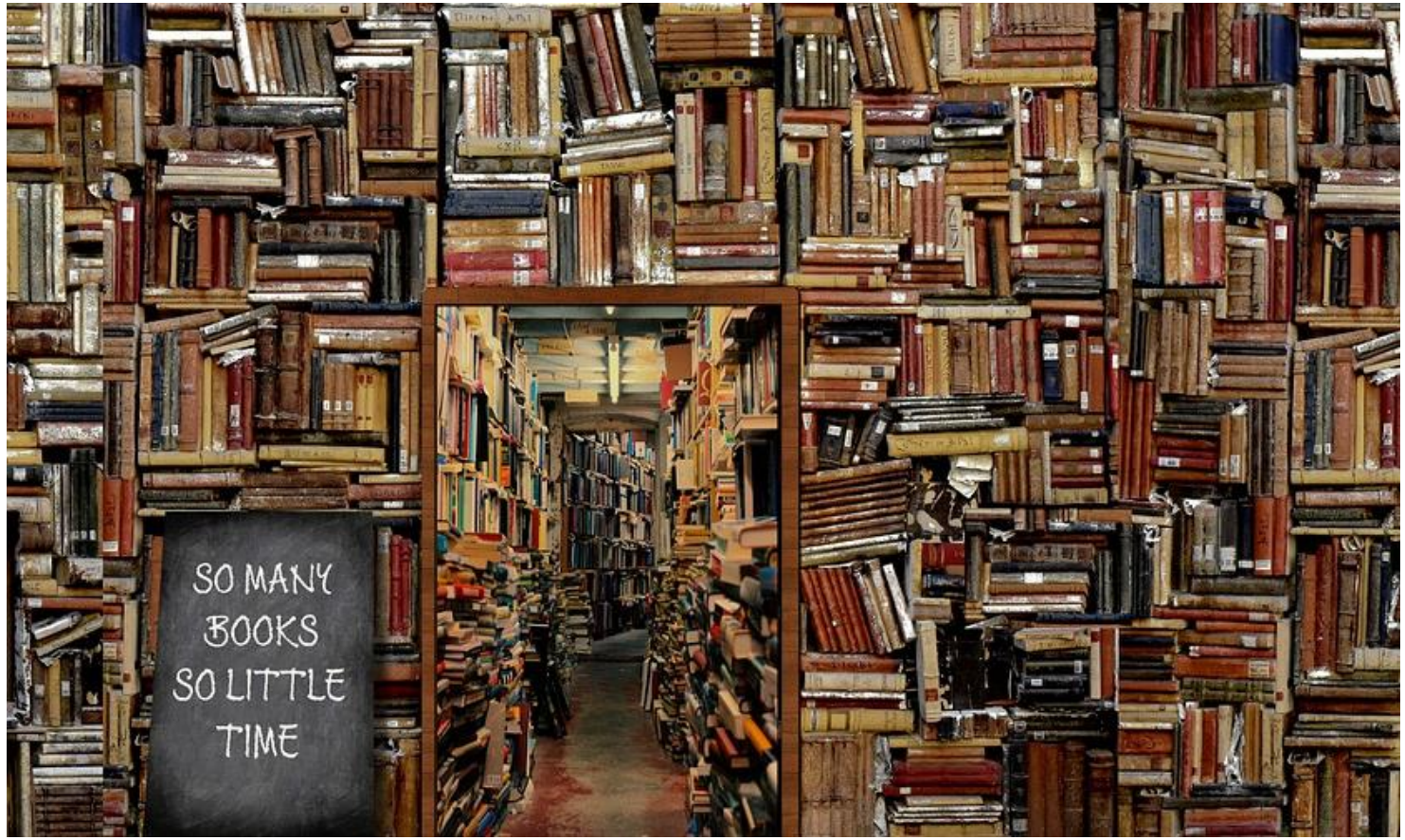
Clumping **reduces the number** the keywords in output to the software to be analysed (e.g. for PV from 131000 to 32000)

The bibliometric software output can be huge

Even if the process is optimised

- Design good search string design
- Clumping

The output of a bibliometric/ text mining software can still be of thousands of keywords (e.g. 32000 for PV)



Resources are limited

Software outputs need **manual analysis** (hypothesis of identifying FETs in a new sector; no machine learning support for one-shot works)

List of keywords can give to skilled **Analysts** and **experts** info or clues to identify **technologies**

How many keywords can be processed in a reasonable interval of time?

Only the top ranked keywords are analysed

Rank		Occurrence
1	wave energy	298
2	wave energy converter	136
3	renewable energy	110
4	tidal energy	91
5	ocean energy	53
6	wave power	44
7	wave energy converters	43
8	oscillating water column	40
9	wave energy conversion	37
(…)	(…)	(…)
295	impulse turbine	3
296	interannual variability	3
297	low crest freeboard	3
298	malaysia	3
299	malta	3
300	marine	3
301	marine currents	3
302	marine technology	3
303	model	3
304	model test	3

From our experience and case study analysis the **N = 300 top ranked keywords** seems a good Time Vs Quality trade-off

(Example for Ocean Energy – see details in References)

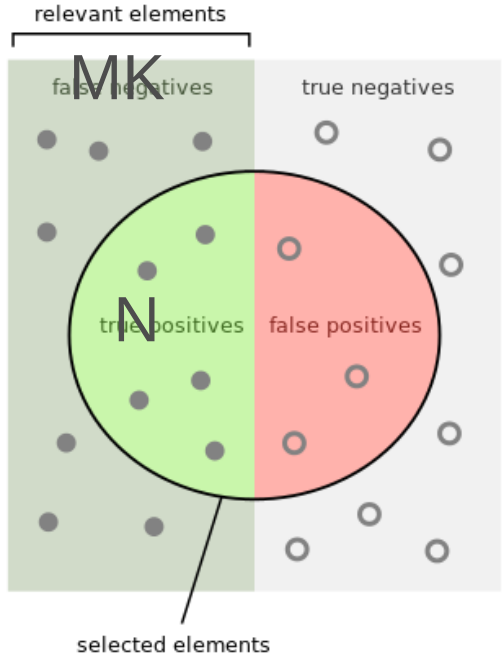
Indicators to compare expert Vs software - 1

We define

$r(N) = n(MK \cap N)$

$Recallrate(N) = (\frac{r(N)}{MK} * 100)$

the percentage of Marker Keywords (MKs) representing experts' technologies present **in the first N-ranked keywords retrieved by the bibliometric software** (under specified search/filtering conditions). This could be also defined "Recall rate at a fixed ranking".
We find relevant **Recallrate**(**300**)



European Commission

Indicators – 2

$$SumRank(MK) = \sum_{i=1}^{n(MK)} Rank(MK_i)$$

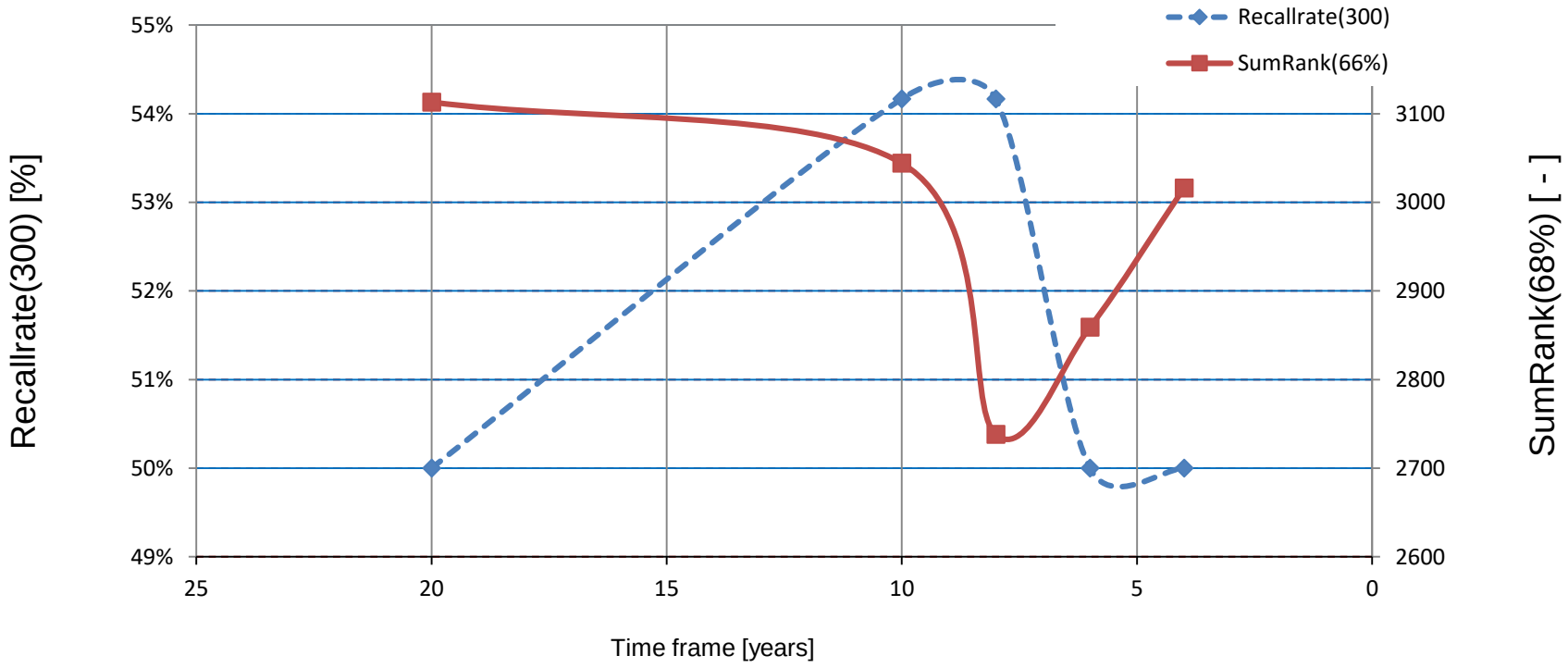
Is the sum of the ranks of all the MKs retrieved by the software under specific search/filtering conditions. This indicator quantifies the success of the software in high-ranking a set of marker keywords: the lower this indicator the higher the efficacy.

This indicator can be calculated only if all the keywords of the MK set are present also in the list produced by the software. By removing the “W” worst-performing marker keywords in the considered keyword list (those with highest values of the rank, or not present in the list):

$$SumRank(x\%) = SumRank(MK - W) = \sum_{i=1}^{n(MK-W)} Rank(MK_i)$$

The number of W MKs should be an amount considered a “reasonable” loss of information (the 68% of them, corresponding to one standard deviation interval in the Gaussian distribution)

Outputs – Optimal Timeframe for FETs

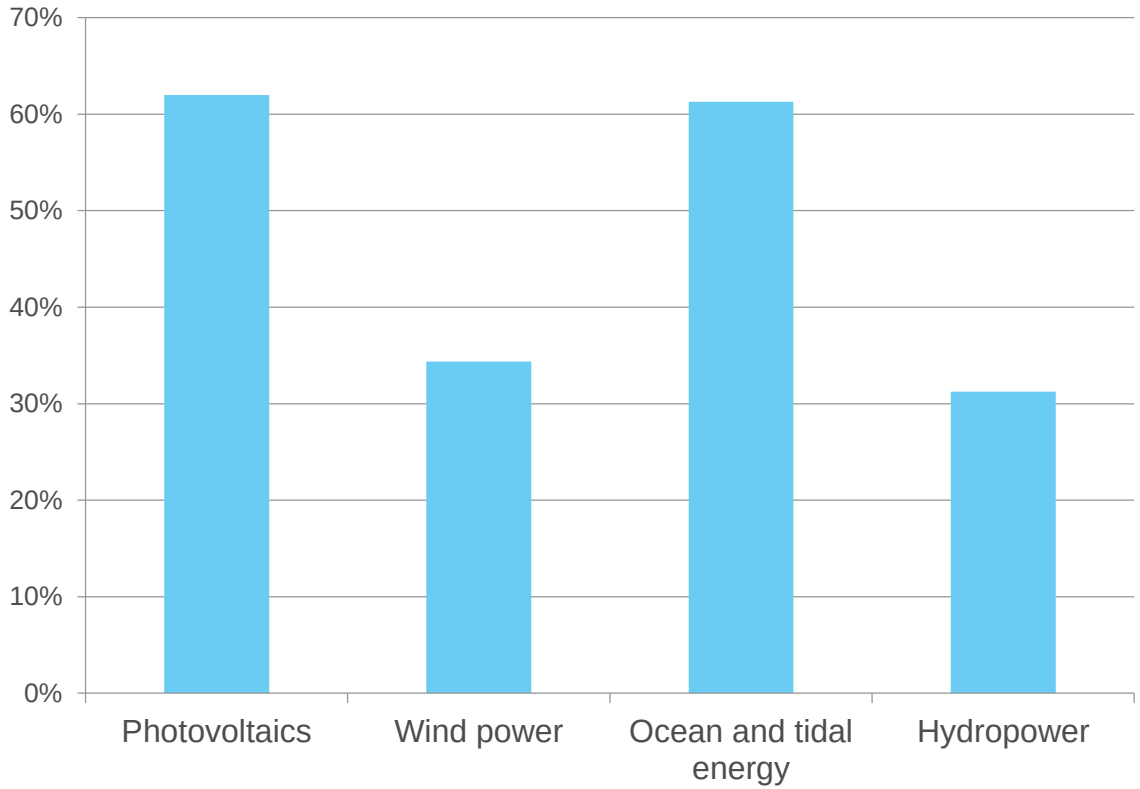


The indicators RecallRate(300) (the higher the better) and SumRank(68%) (the lower the better) both identify in **6-8 years** as the **optimal time interval to identify emerging technologies** in bibliography (e.g. in Scopus)

Outputs – Recall rates of software Vs Experts

For the considered case studies of: Photovoltaic, Wind power, Ocean and tidal energy and Hydropower the TIM **software retrieved 25% to 67%** of the technologies identified by expert groups

In figure: **Recall Rate (300) for the various case studies** Average performance for the algorithms “author keyword and “term frequency-inverse document frequency” TF/IDF



Outputs – Details for Photovoltaics

(16) Technologies identified by the PV expert panel [and related marker keywords]	b) AK NR	c) Unc (AK NR)	d) Tf-idf NR	e) TRL
Measuring Unit	[-]	[-]	[-]	[-]
Dye-Sensitized Solar cells or [DSSC]	**0.09**	0	**0.06**	5 – 6
[Organic solar cells] or [OSC]	**0.24**	0	**0.18**	5 – 6
[Perovskite] thin film solar cells	**0.4**	0	**0.21**	4 – 5
[Plasmonic] solar cells	**1.96**	0.03	**5.35**	3 – 4
Thermo-photovoltaics or [photovoltaic thermal]	**2.58**	0.02	**2.48**	1 – 2
[Transparent conduct]ing materials or [Carrier selective contacts]	**4.13**	0.03	**4.25**	2
Innovative [III-V] compounds based solar cells	**4.76**	0.02	**5.51**	1 – 2
Photoelectrocatalytic devices, [photocatalysis]	**4.85**	0.05	13.55	2
Innovative [multi junction] solar cells	**4.97**	0.05	**3.21**	2 – 3
[Kesterite] thin film solar cells or [CZTS]	**7.48**	0.05	**2.05**	3 – 4
[Intermediate band] solar cells or [IBSC]	**8.7**	0.11	**4.5**	2
Low-cost manufacturing processes, [roll to roll] or [flexible substrate]	13.64	0.28	11.13	n.a.
[Ferroelectric] photovoltaics	15.48	0.24	26.74	1 – 2
New pv materials via [computational design]	17.63	0.54	15.97	n.a.
[Hot carrier] solar cells	28.8	0.72	22.42	1 – 2
[Multiple exciton generation] solar cells or [MEG]	30.65	1.1	17.99	2
Recall rate (300)	**69%**	**-**	**63%**	-
NormalisedSumRank(68%)	**47.8**	**0.25**	**47.59**	-

In green the technologies identified both by the experts and the bibliometric software (amongst the first 300 ranked)

Normalised Ranks in figure (lowest values, better performances, N=300=9.2) – See References

Limitations of the use of bibliometric software to identify emerging technologies

- Bibliometric software is more effective to retrieve technologies with a more consolidated jargon, so higher TRLs (~5) technologies.
  For example, kesterite solar cells have been categorised for several years by their chemical composition, which can vary and can have different acronyms (Cu2ZnSn(S,Se)4, CZTS, CZTSe, CZTSS)
- Not all R&D is published or patented, particularly in strategic fields
- Counts do not distinguish quality

References

Futures

Volume 117, March 2020, 102511

Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software

Alberto Moro [a], Geraldine Joanny [b], Christian Moretti [c]

https://doi.org/10.1016/j.futures.2020.102511

Renewable Energy

Volume 123, August 2018, Pages 407-416

A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review

Alberto Moro, Elisa Boelman, Geraldine Joanny, Juan Lopez Garcia

Show more

https://doi.org/10.1016/j.renene.2018.02.016

Get rights and content

Open Access funded by Joint Research Centre

Thank you very much!

alberto.moro@ec.europa.eu

European Commission

Thank you

European Commission