Responsible Research & Evaluation

Responsible Indicators?

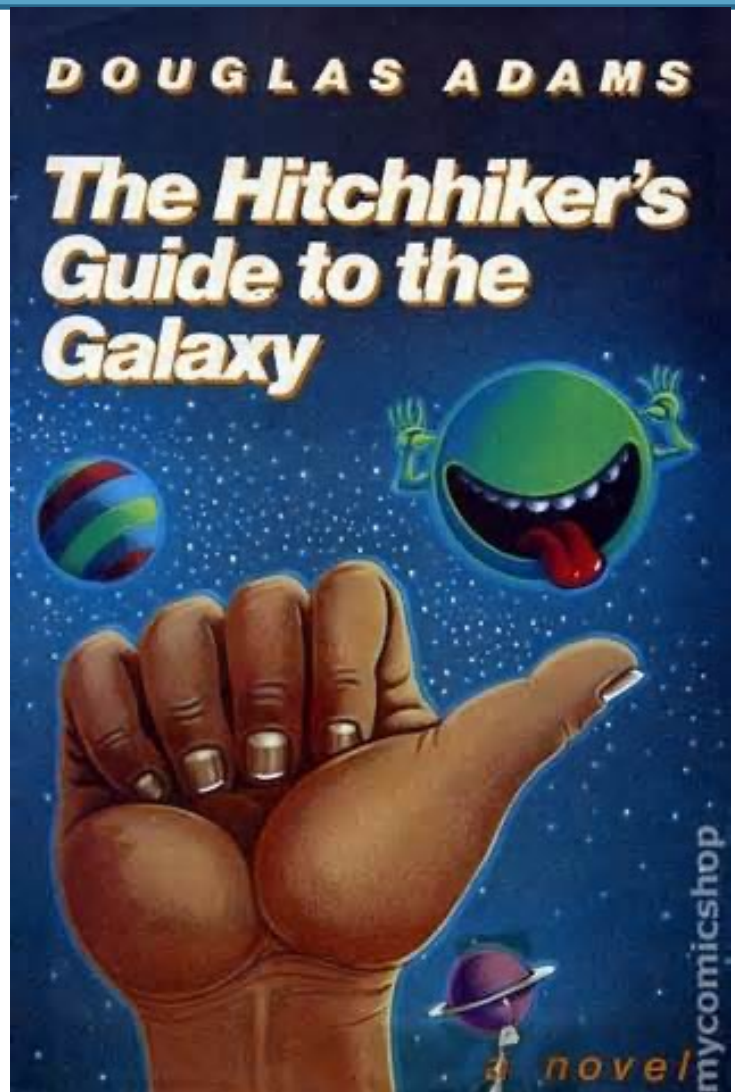
Michael Ochsner

Context

- Responsible Research and Innovation
  - Important concept in policy, not so clear in practice
  - Evaluation and RRI → How to evaluate research responsibly? Or only evaluate (irresponsibly) how responsible research is?
- Information is key
  - Information is not neutral
  - Numbers are objective (?)
  - Benchmarking: Data quality!
- My questions: What is Responsible Evaluation of Research?
  And: Are there Responsible Indicators?

Numbers...: Metrics and Concepts

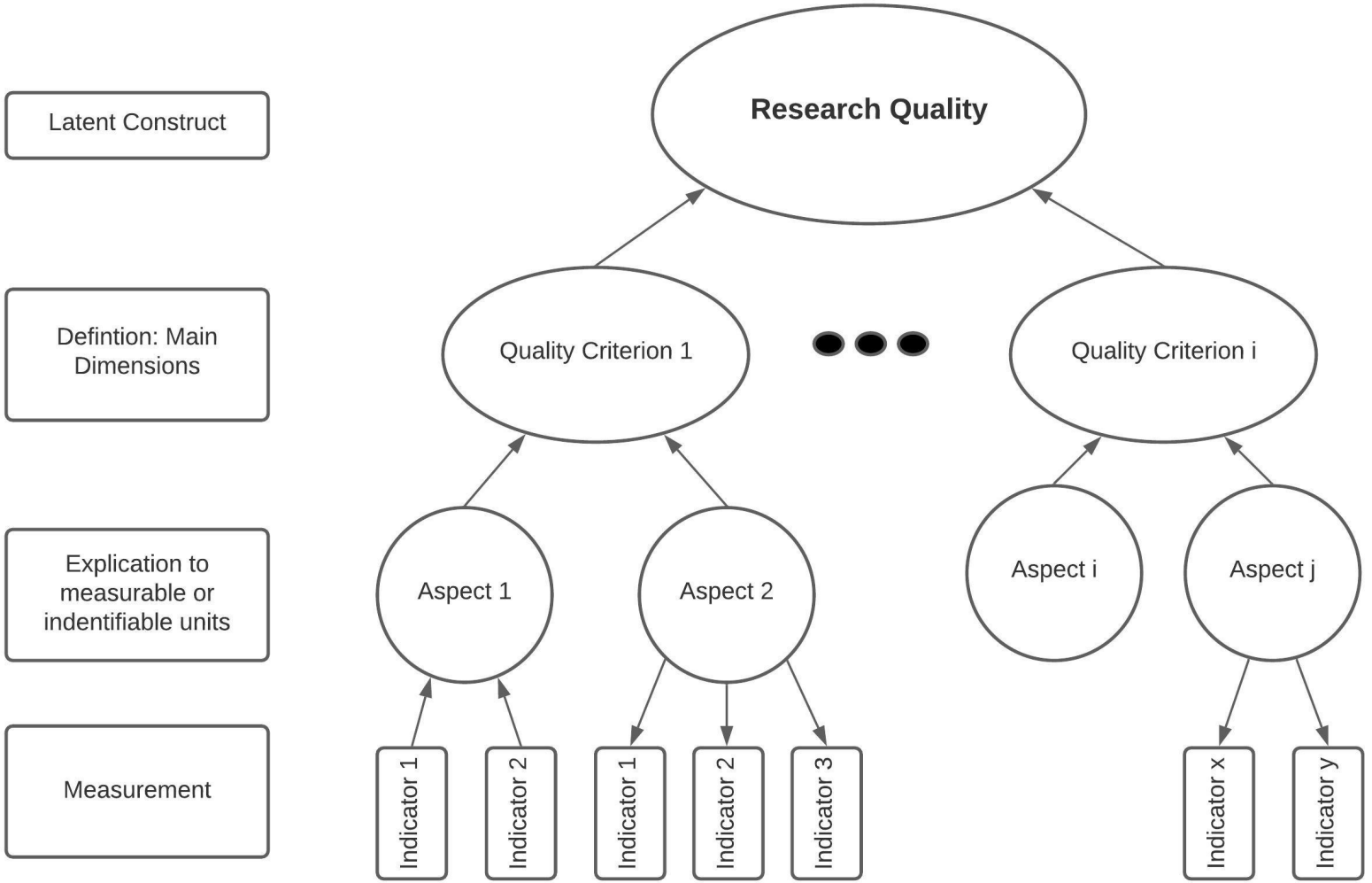
- Novel/Radio play by Douglas Adams
  Hitchhiker's Guide to the Galaxy
  (1979)
- Deep Thought:
  The ultimate question of life,
  the universe and everything
  - 7.5 million years to compute and check
  - The answer was.... **42**
- answer is meaningless – because the question was stupid:
  - did not specify the form of answer
    nor did they really know what they asked for



Validity

- Numbers should reflect something
- „Quality“, “Performance“, „Societal Impact“ are latent concepts
- Validity is the extent to which a measure (i.e., an indicator) actually measures what it purports to measure (i.e., a concept) (Borsboom et al., 2004, p. 1061)
- Scientometrics is data-driven: „measuring what can be measured“ endangers validity, mostly reducing it to *correlation*.
- Thunder correlates highly with lightning (and there is even a causal relationship). However, lightning cannot measure thunder.

Measurement Model



Research Quality (Humanities)

- Valid measures for research quality?
  orange: three disc.; blue: two disc.; ***bold and italic***: commonly used

1. Scholarly exchange
2. Innovation, originality
3. Productivity
4. Rigour
5. Fostering cultural memory
6. Recognition
7. Reflection, criticism
8. Continuity, continuation
9. Impact on research community
10. Relation to and impact on society
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. Relevance

Measurement

- What do indicators measure that are often used in evaluation?

Table 1: Frequently used indicators and criteria they can potentially measure

Indicators	Criterion
Citations	Recognition; impact on research community; relevance
Prizes	Recognition; impact on research community; relevance
Third party funding	Recognition; impact on research community; relevance; relation to and impact on society
Collaborations	Scholarly exchange; recognition
Transfers to society and economy	Relation to and impact on society
Publications	Scholarly exchange; productivity
Board memberships	Scholarly exchange; recognition; impact on research community
Recruitment	Continuity, continuation

Research Quality (Humanities)

- Measured by commonly used indicators (***bold and italic***)

1. ***Scholarly exchange***
2. Innovation, originality
3. ***Productivity***
4. Rigour
5. Fostering cultural memory
6. ***Recognition***
7. Reflection, criticism
8. ***Continuity, continuation***
9. ***Impact on research community***
10. ***Relation to and impact on society***
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. ***Relevance***

Research Quality (Humanities)

- English Literature, German Literature and Art History
- Consensual Indicators (orange: all three; blue: in two disciplines)

1. *Scholarly exchange*
2. Innovation, originality
3. *Productivity*
4. Rigour
5. Fostering cultural memory
6. *Recognition*
7. Reflection, criticism
8. *Continuity, continuation*
9. *Impact on research community*
10. *Relation to and impact on society*
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. *Relevance*

So what?!

- Criteria are consistent across fields (SS & H) as well as evaluation situations (general evaluation or grants for young scholars)
- About 50% of relevant criteria not measurable with indicators
- Indicators measure the less important criteria

→ Validity issue! We do not measure what we want to measure but what we can

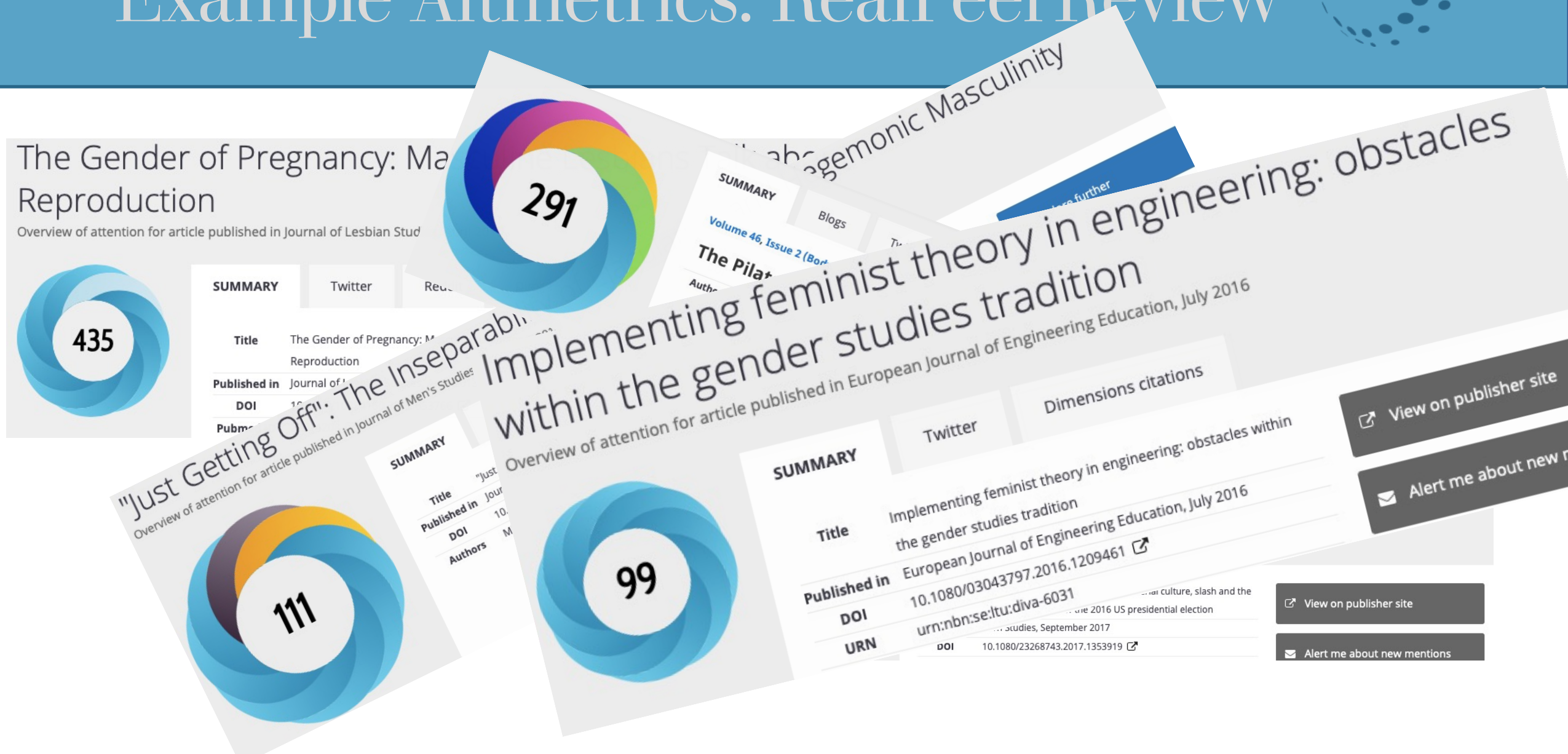
Example Altmetrics

- Indicator is present almost everywhere
- We do not know what it measures nor is it stable (Gumpenberger, Glänzel, Gorraiz, 2016)
- It is seen as measure for societal impact → but it's driven by researchers (Ke, Ahn, Sugimoto, 2017)
- Based on Twitter data but also other social media → but Tweets correlate with >0.9
- Strongly dependent on single accounts (institutional; fun)

New Real Peer Review

@RealPeerReview

Example Altmetrics: RealPeerReview



Example Altmetrics: RealPeerReview

- Random selection of RPR-articles and control group

	@RPR					Control Group			
	Obs	Mean	Median	Min	Max	Mean	Median	Min	Max
AAS	67	50	23	3	440	9	2	0	226
Tweets	67	73	29	5	948	10	1	0	293
Percentile	67	90	94	49	99	42	40	0	99
PP Journal	67	86	92	40	100	37	30	0	99
PP Similar Age	67	87	90	62	99	42	48	0	99

So what?!

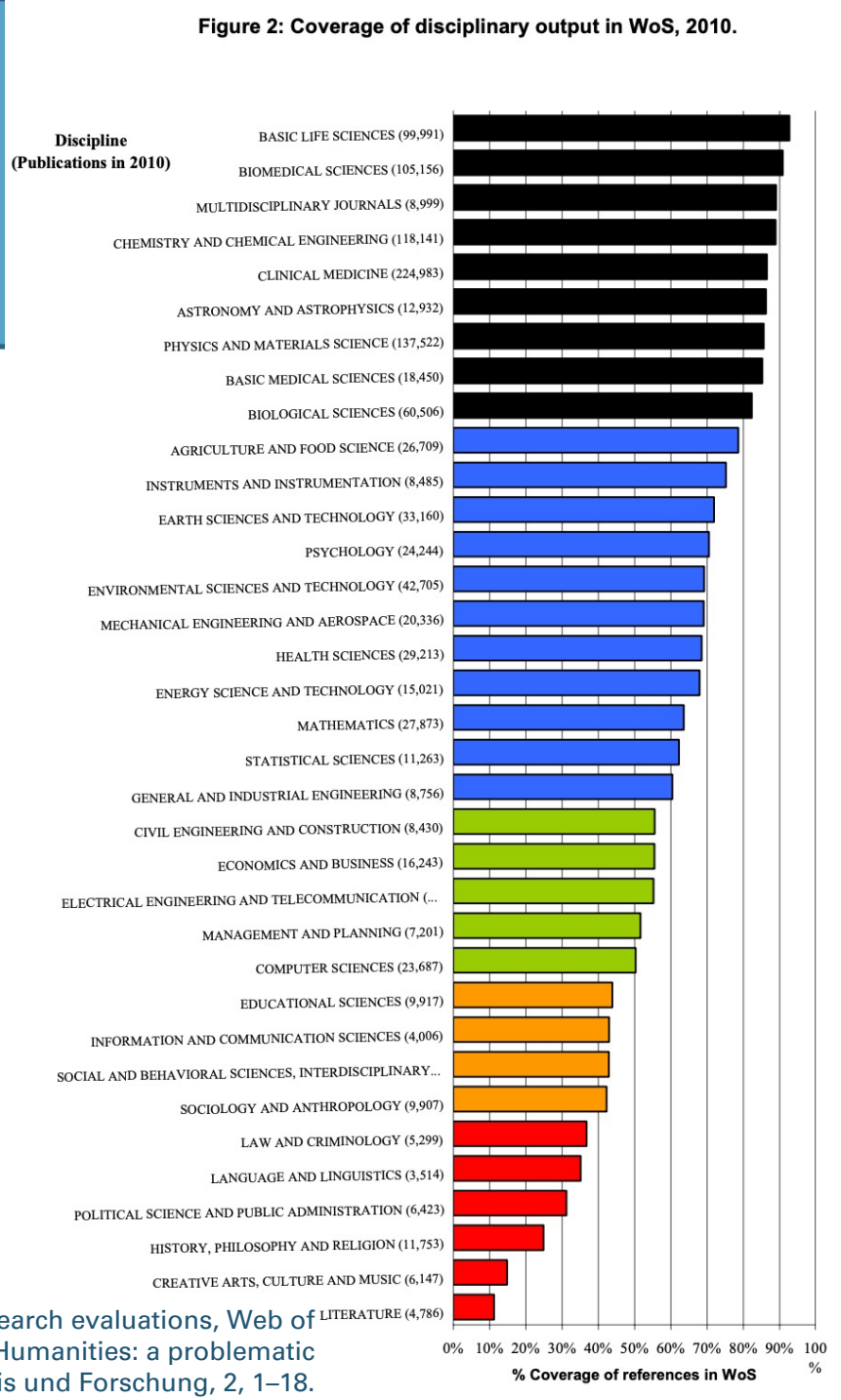
- We have objective numbers
- But not really meaningful results

- Still, it is visible everywhere → It has an impact
- If used in evaluation, the following incentives are made:
  - Have a Twitter account and tweet all your articles
  - Have an institutional account and retweet → already in 6th decile
  - Choose a funny, provoking title for your article
  - Study porn, feminist theory, funny sports or drugs and reference US presidents
- Are these the incentives to be promoted?

Data Quality

- Let's assume, we have a correct indicator, measuring what we want
- Still, data quality issue → something that is missing from any discourse

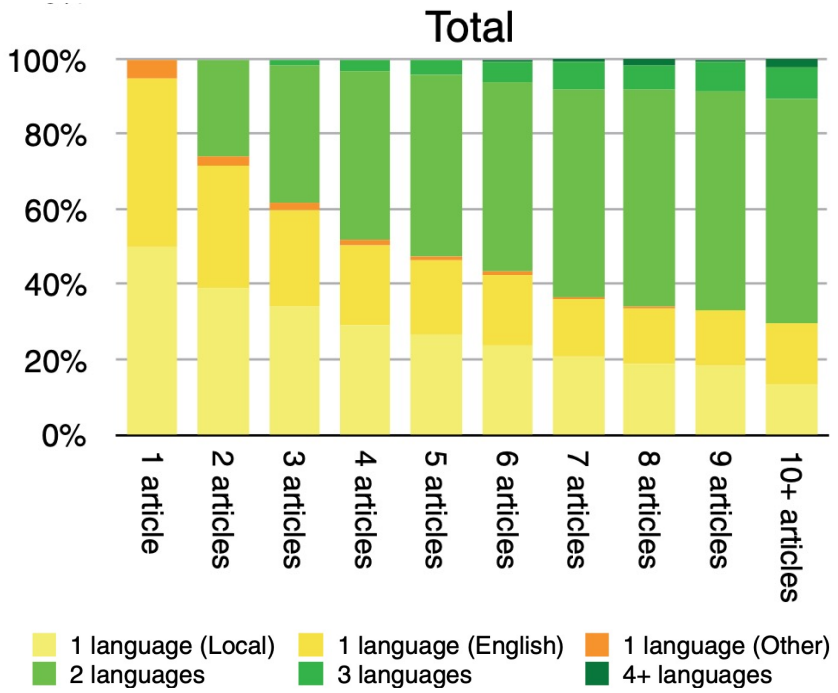
- If many or even most citations from WoS go to non-WoS articles, what is then the meaning of a citation score based on WoS data?



van Leeuwen, T. N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? Bibliometrie - Praxis und Forschung, 2, 1–18.

Data Quality

- If it is systematic bias, like language or local topics?



Kulczycki, E., et al. (2020). Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study. Journal of the Association for Information Science and Technology, 26(1), 41. http://doi.org/10.1002/asi.24336

Table 2 Shares of document types in political science in different countries

Countries	Sources (Span)	WoS articles (%)	Journal articles (%)	Books as author (%)	Books as editor (%)	Book chapters/ articles in books (%)	Proceeding papers (%)
Germany	Publication lists of two institutes (2003–2007)	7	22	4.4	7.5	39	15
Norway	CRIStin (2005–2009)[a]	28	46	4[b]		50[c]	–
Flanders, Belgium	VABB-SHW (2000–2009)[d]	17	79	1.7	2.6	16	0.3

Chi, P.-S. (2015). Changing publication and citation patterns in political science in Germany. Scientometrics, 105(3), 1833–1848. http://doi.org/10.1007/s11192-015-1609-3

Responsible Metrics?

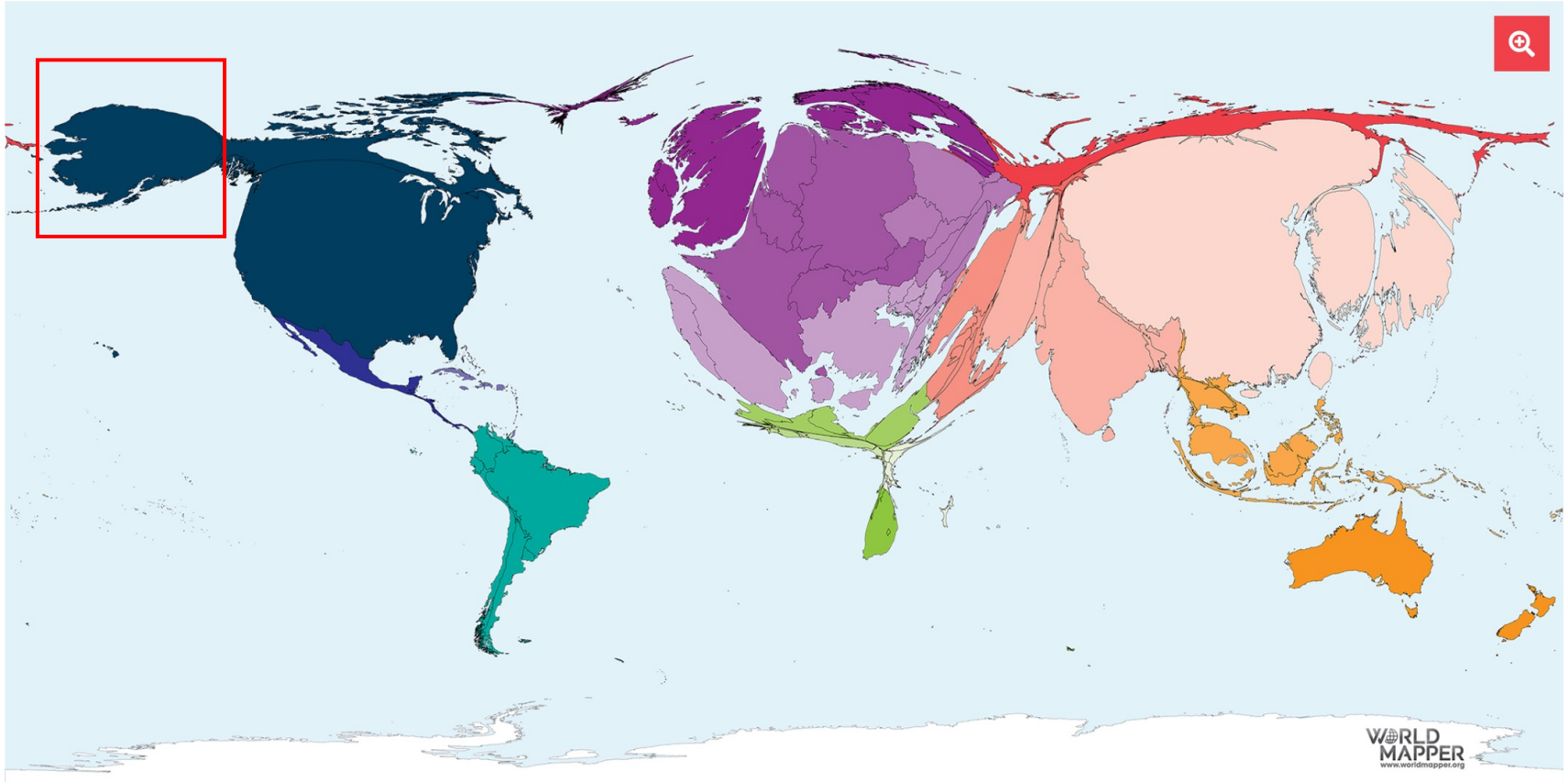
- Metrics often **do not** cover concept encompassingly
- Missing out important information
  - Leads to an **invalid** measurement
  - Leads to **side-effects**
  - Leads to **changes in behavior** (de Rijke & Rushforth, 2015; de Rijke et al., 2016)
    - Not „perverse“ or „unintended“ effects but wrong incentives
    - Not wrong behavior but wrong policy intentions
    - Pay 1$ per dead rat. People will start to breed rats.

Responsible Metrics?

- Policy information tools:
- Worldmapper:

  Science papers published

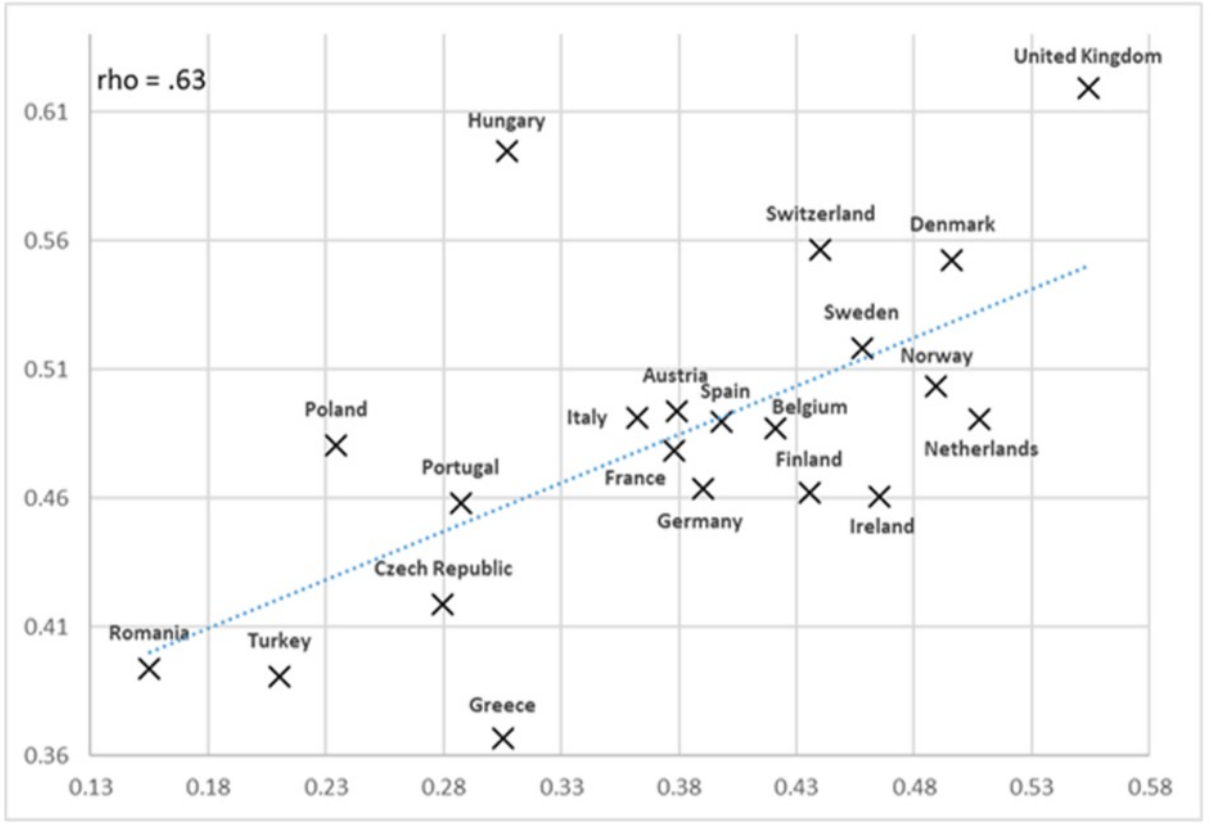
- Indicators are often Misinformation or even Disinformation



Responsible Metrics?

- Typical example
  - Open Access as publication to wider audience
  - Twitter as societal impact
- But
  - We know that knowledge transfer is difficult (see effect of OA on Covid-Discussions: Doubts about scientific knowledge!)
  - We know that Tweets are mainly driven by scientists
  - Open Access promoters are likely to be Tweeters
- What then is the information in the Graph?
  - Is is responsible to show data without any validation of the measurement and data?
- Responsible is about *use* of indicators

Figure 4. Percentage of open access (vertical axis) and tweeted (horizontal axis) publications by country (top) and topic (bottom)



Conclusion: The answer is 42

- Deep Thought created a new solution including beings that will resolve the question of all questions:
  Planet Earth, directed by white lab mice
  - Calculating time: 10 million years.
  - Earth destroyed before the result was ready by Psychiatrists who feared loss of their careers
- Metrics are never responsible
  - Users are responsible, those who present the metrics
- Sketch of Responsible Use of Metrics (be it evaluation or Covid)
  - Assure that indicators validly measure the concept
  - Assure the data quality („representation“, error, reliability)
  - Interpret within the boundaries of measurement and data quality

Research Evaluation

- Research Evaluation Must Correspond to Research Practices
- Involve all Stakeholders of Research Evaluation
- Acknowledge Diversity of Evaluation Practice
- Include a Broad Range of Evaluation Criteria
- Combine Different Evaluation Methods
- Carefully Evaluate Interdisciplinary Research

→ENRESSH Policy Brief on Better Adapted Procedures for Research Evaluation
https://doi.org/10.6084/m9.figshare.12049314.v1