Seed+Expand

aggregating the scientific output of
the Netherlands, 2000-2010

Linda Reijnhoudt, Rodrigo Costas, Ed Noyons,
Katy Börner, Andrea Scharnhorst

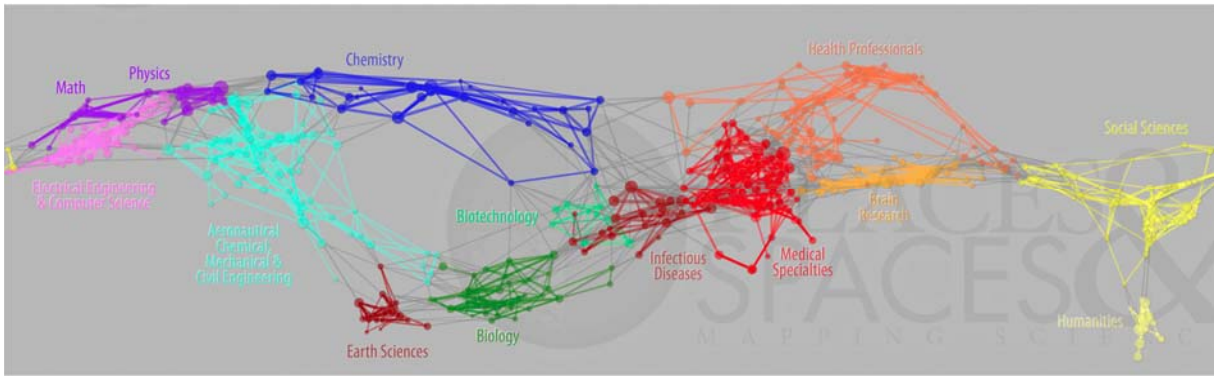
CWTS Meaningful metrics

CNS Cyberinfrastructure for Network Science Center

Data Archiving and Networked Services DANS

goal

to study the dynamics on the output of
Dutch professors (2001-2011)



but, lack of data on
the output of full professors!

the problem

given a Dutch professor in the NARCIS system

find all his/her publications

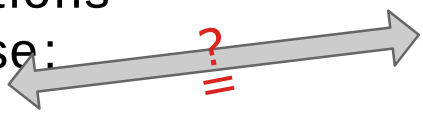
How to connect bibliographic data from CWTS with the NARCIS system?

context

CWTS

Bibliometric publications database:

- author
- author-order
- email (sometimes)
- affiliation (sometimes)
- journal



DANS

NARCIS

dutch scholars:

- name, initials
- DAI affiliations
- organisation
- email

non trivial I

- misspelled names
  - *Van Knienberg* instead *Van Knippenberg*
- different initials / first name
  - *Johannes* and *Hans*
- different formats in the data across sources
  - Prefixes separated in the NARCIS system
    - P.M.P. | van | Bergen en Henegouwen
  - Made initials or concatenated in WoS
    - Henegouwen, PMPVE (Henegouwen, Paul M. P. van Bergen En)

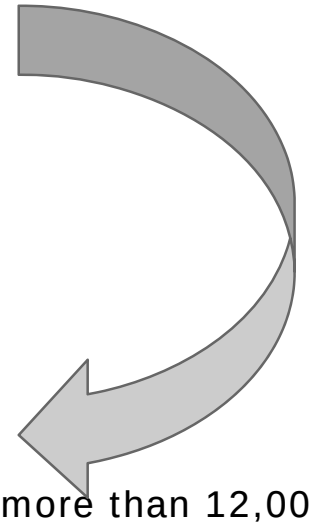
non trivial II

- multiple scholars have the same author name (homonymy)
- the same scholar with multiple author names (synonymy)
  - changes over time, e.g., due to marriage

the raw data

NARCIS database (DANS)

- 8378 Dutch full professors
  - affiliation to dutch organizations
  - name, initials
  - email
  - DAI



CWTS bibiometric data system

- close to 23 million publications in more than 12,000 journals
- no unique author identifier for all authors

the Gold Standard

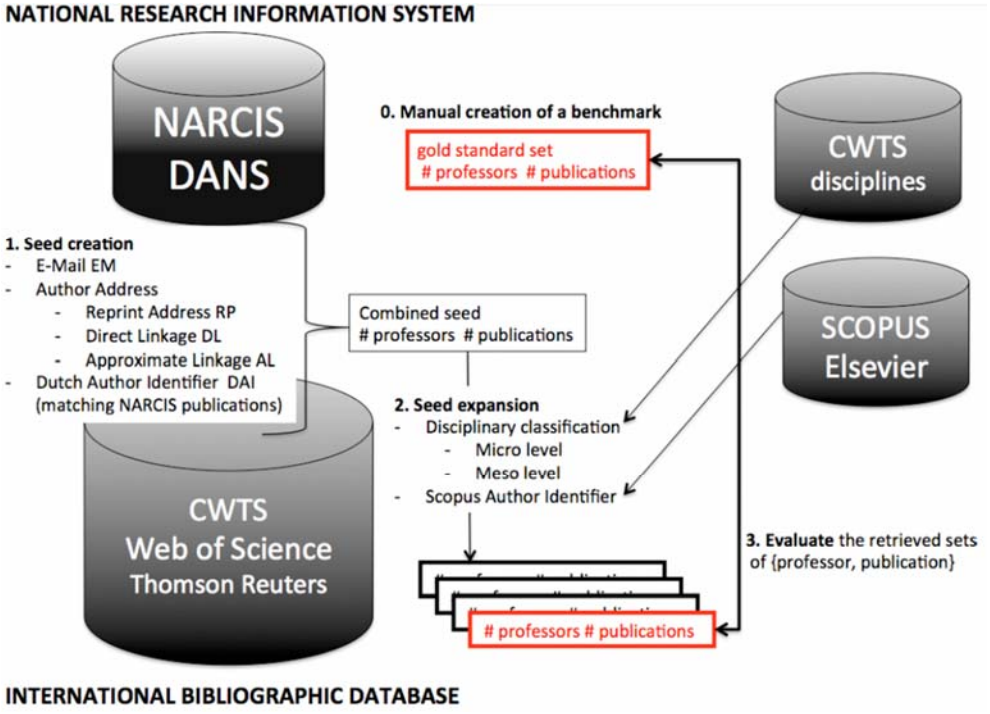
we already know the complete *oeuvre* of 1400 Dutch full professors, due to manually verified publication lists by CWTS

USEFUL TO VALIDATE OUR METHODOLOGY

the 1400 of the 8376 (17%) full professors who already appear in this list:

the Gold Standard

the sources & main overview



Slide 9

1 @linda: need better image
tjreijnhoudt,

2 I think so!
Rodrigo Costas,

Seed+Expand main concept

- seed creation, precision
  - given a full professor, {initials, name, email, affiliations}
  - find one or more publications that are **most likely** authored by this professor
- seed expansion, recall
  - given these 'seed' publications,
  - find publications by the **same author**
    1. publication-based classifications
    2. Scopus Author Identifier

seed creation

1. Email seed (EM)
2. Author Address approaches (*)
   a. Reprint Author (RP)
   b. Direct linkage author-addresses (DL)
   c. Approximate linkage author addresses (AL)
3. Digital Author Identifier seed (DAI)

(*) For these seeds, very common names have been excluded

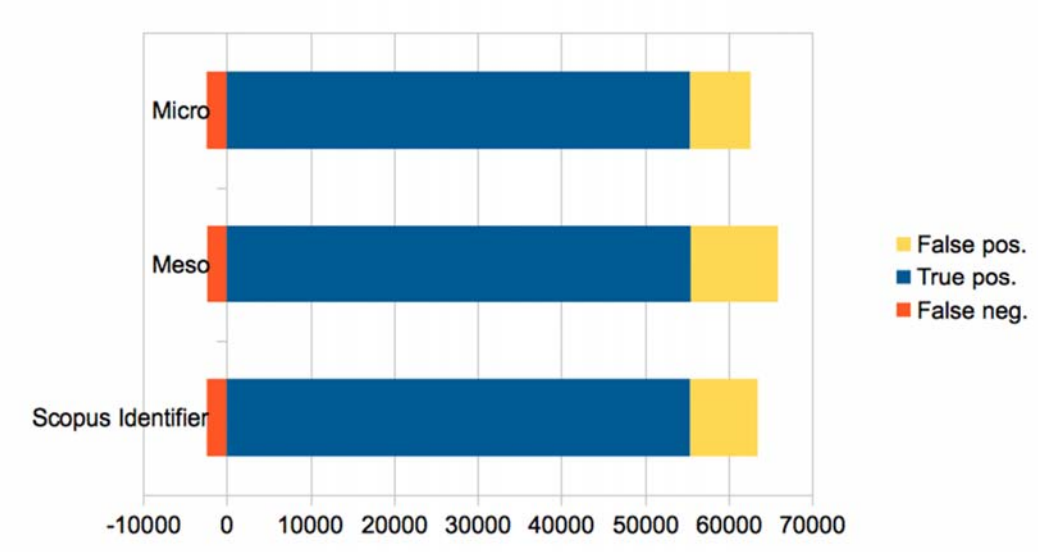
seed expansion

1. CWTS Paper-Based Classification (2001-2011)
   - based on citation relationships of publications
   - 672 meso, over 20K micro disciplines
   - micro: +23% unique papers over seed
   - meso: +34% unique papers over seed

1. Scopus Author Identifier Approach (1996-2011)
   - +69% unique papers over seed

evaluation

Gold standard: *2001-2010*



Results

- 80% of Dutch professors detected
- Micro-disciplines: highest precision (88.5)
- Scopus Author id & micro disciplines: same recall (95.9)
- This methodology can be applied to other sets and author identity schemes (ORCID, VIVO, etc.)
- Further research on disciplinary differences and improvements

General Discussion

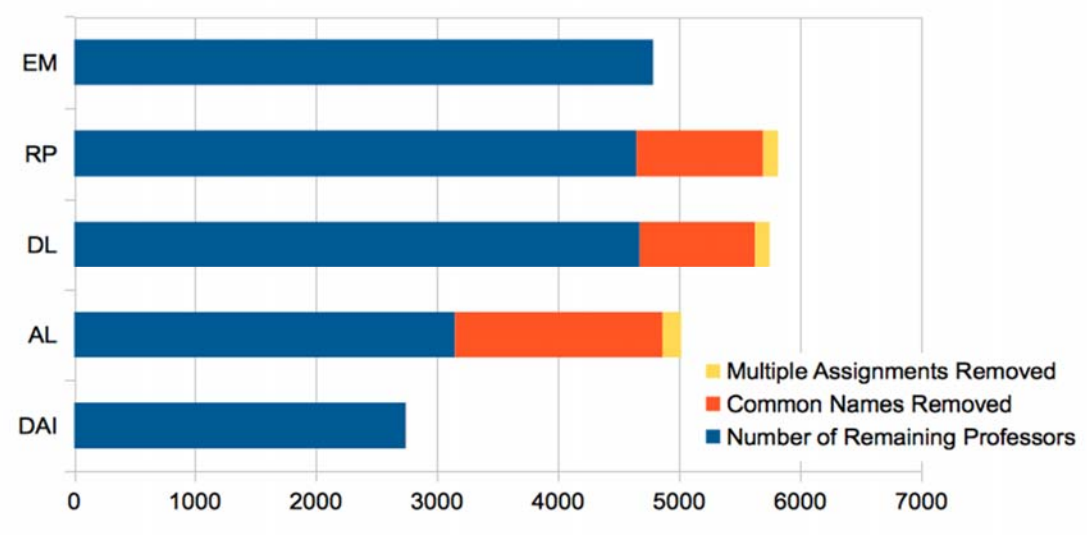
- increasing bibliographic data sources but still lacking author disambiguated data!!
- lack of research on how to connect databases
  - repositories
  - bibliographic databases (WoS, Scopus, etc.)
  - altmetrics
- e-mail data and DAI/ORCID-like identifiers are powerful linking elements across systems

the end ...

Thank you very much for your attention!
Questions?
Comments?

five seeds

combined: 6753 of 8376 full professors found



Slide 17

1 Don't you want to include this slide?
Rodrigo Costas,