Tool Criticism

Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman

Information Access

CWI

Koninklijke Bibliotheek
National Library
of the Netherlands

COMMIT/

SEALINC
Media

First mention of …

… in the OCRed newspaper archive of the KB?

Digital Humanities

- digital archives / large data collections
  - collection bias / digitization policy
  - digital representation ≠ physical object
  - tools are imperfect
- source cannot be considered independent from tools
  - need methods to detect tool-induced bias

Source criticism

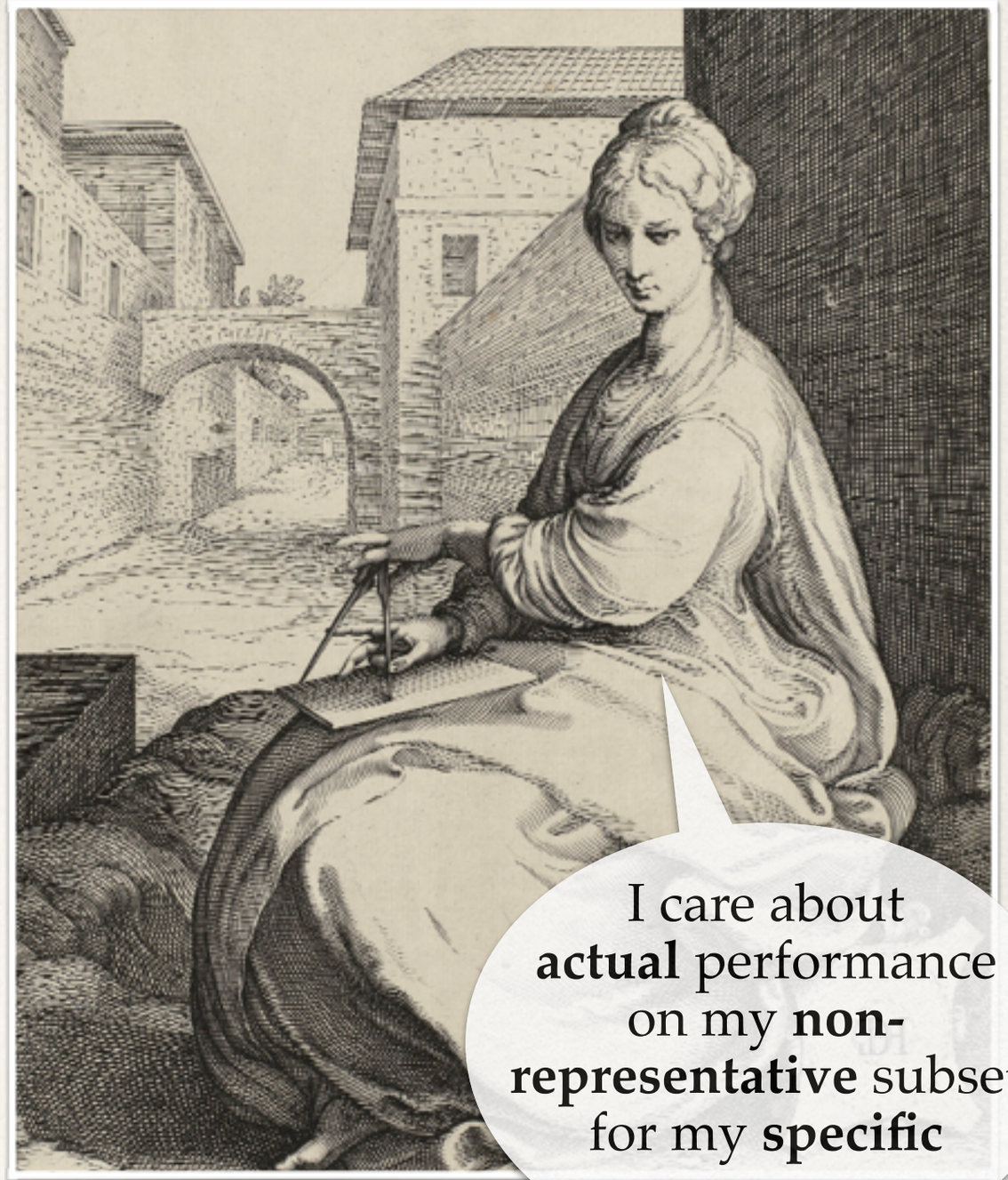
- well established method in the humanities to detect bias in a source
  - Who is the author?
  - Is the information current?
  - Is the information objective and credible?…

- need a similar approach for tool-induced bias:

  tool criticism

Methodology

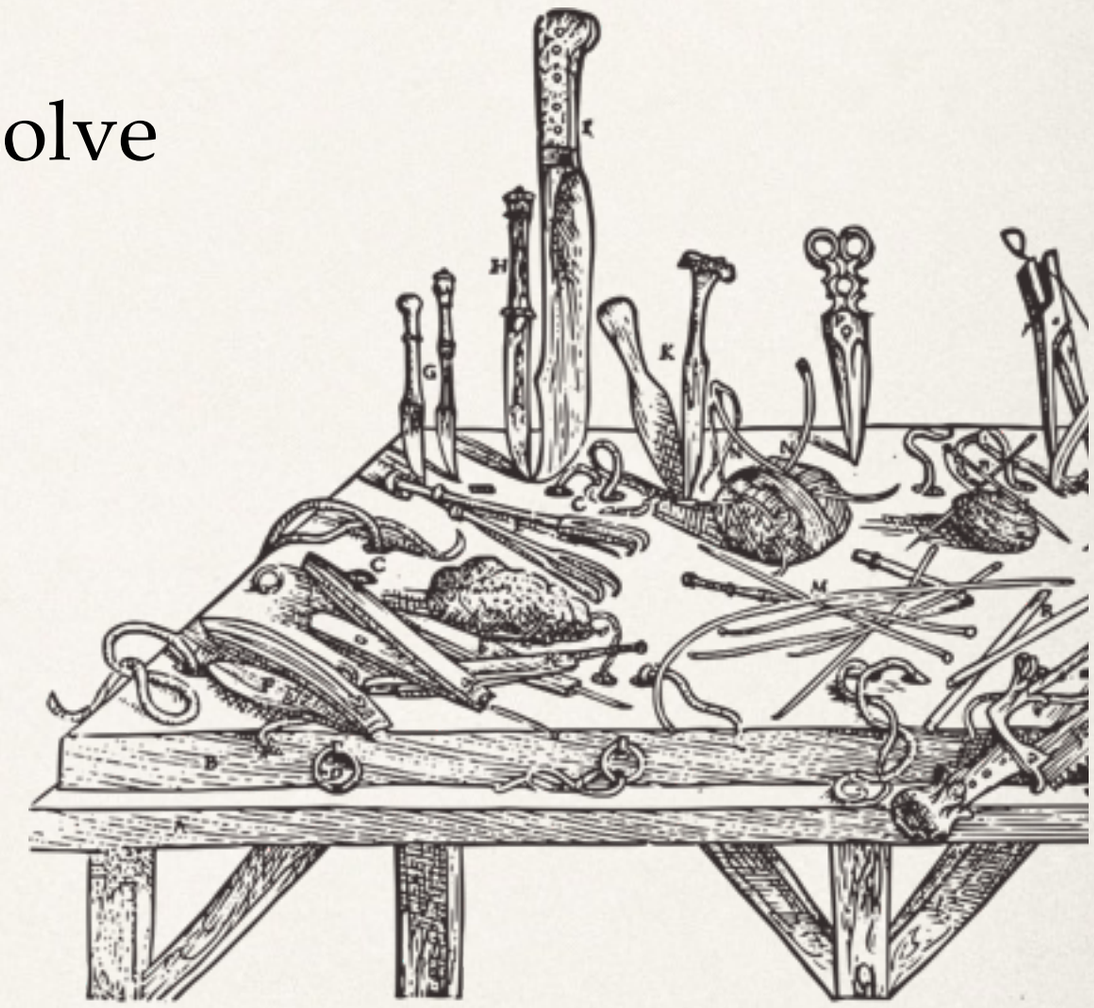
- interviews with humanities scholars
  - classification of common research tasks
  - lack of trust blocks progress
- use case: digital newspaper archive of KB
  - no formal OCR evaluation
  - useful for scholars?
  - mismatch between two perspectives



Two different perspectives of quality evaluation

No silver bullet

- we propose novel strategies that solve part of the problem:
  - critical attitude (awareness and better support)
  - transparency (provenance, open source, documentation, …)
  - alternative quality metrics (taking research context into account)

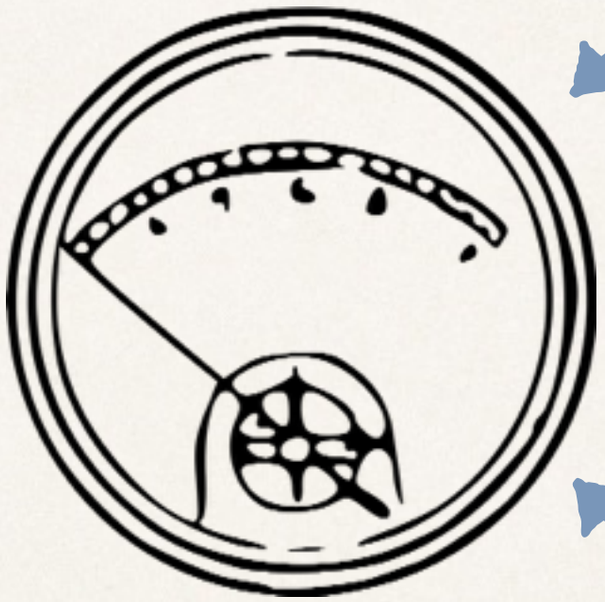


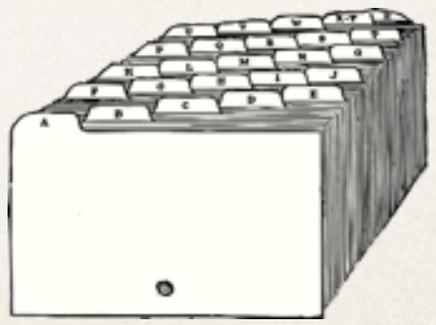
M. Traub, J. van Ossenbruggen, L. Hardman,
Impact Analysis of the OCR Quality Problem in Digital Archives, TPDL2015 (under review)

Context-aware quality indicators

query

task





data

Future work

- What strategies should a tool support to help scholars discover and dealing with bias?
- What is a good way of estimating uncertainty for a specific task?
- Can we crowdsource (part of the data for) better estimates?
- What is a good way of conveying the estimated impact to scholars?