MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

ASSIST project: processing humanities documents with text mining tools

Davy Weissenbacher, NaCTeM

Davy.Weissenbacher@manchester.ac.uk

MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

Assist project, overview

- **Goal**: evaluation of a tuned and enhanced ASSERT system for processing educational and mass-media documents
- Improve the research of documents in the Educational Evidence Portal: Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI)
- Help for the sociologist work of the Frame Analysis: National Centre for e-Social Science (NCeSS)

MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

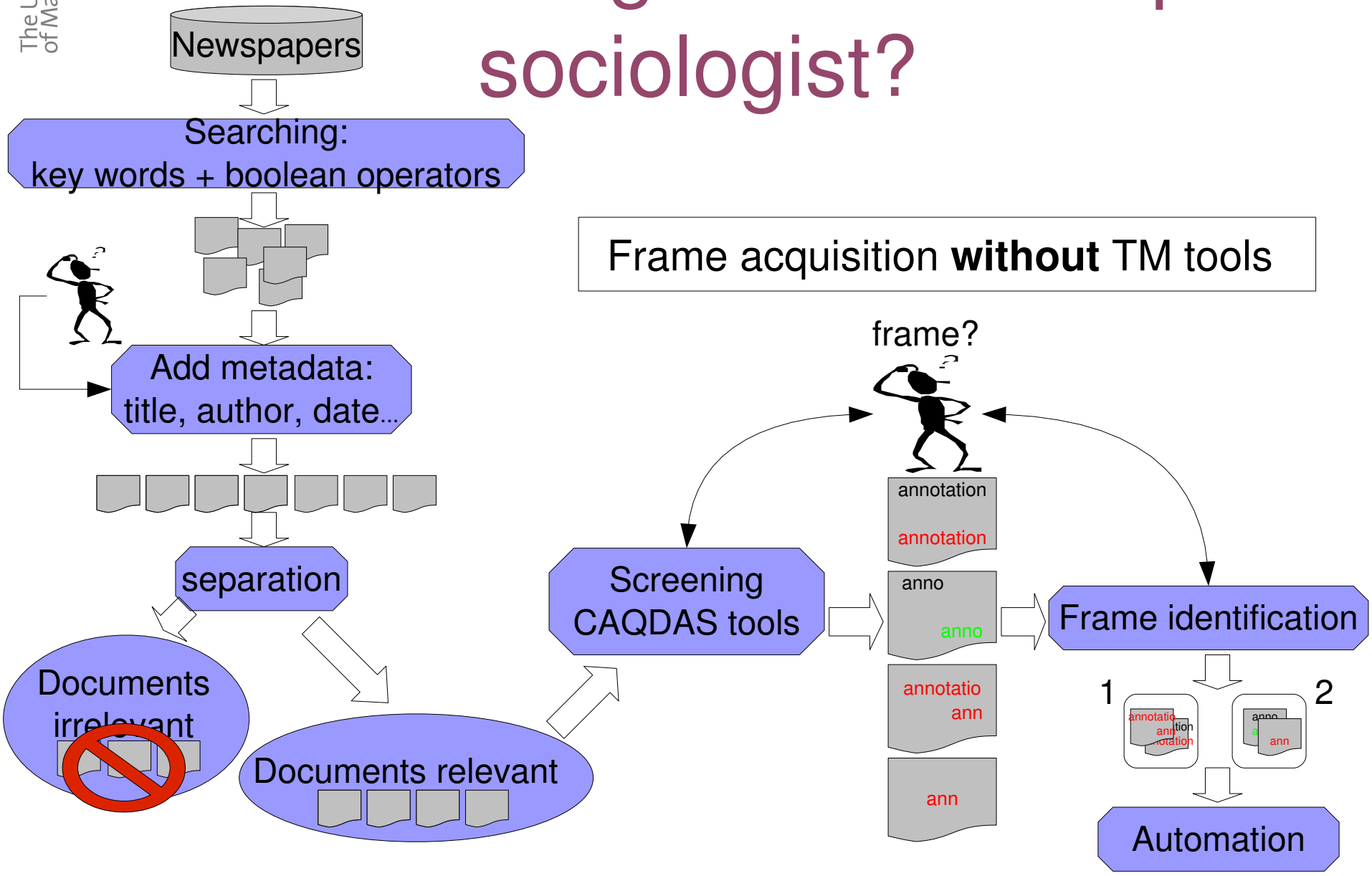
Assist project, overview

- **Goal**: evaluation of a tuned and enhanced ASSERT system for processing educational and mass-media documents
- Improve the research of documents in the Educational Evidence Portal: Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI)
- **Help for the sociologist work of the Frame Analysis: National Centre for e-Social Science (NCeSS)**

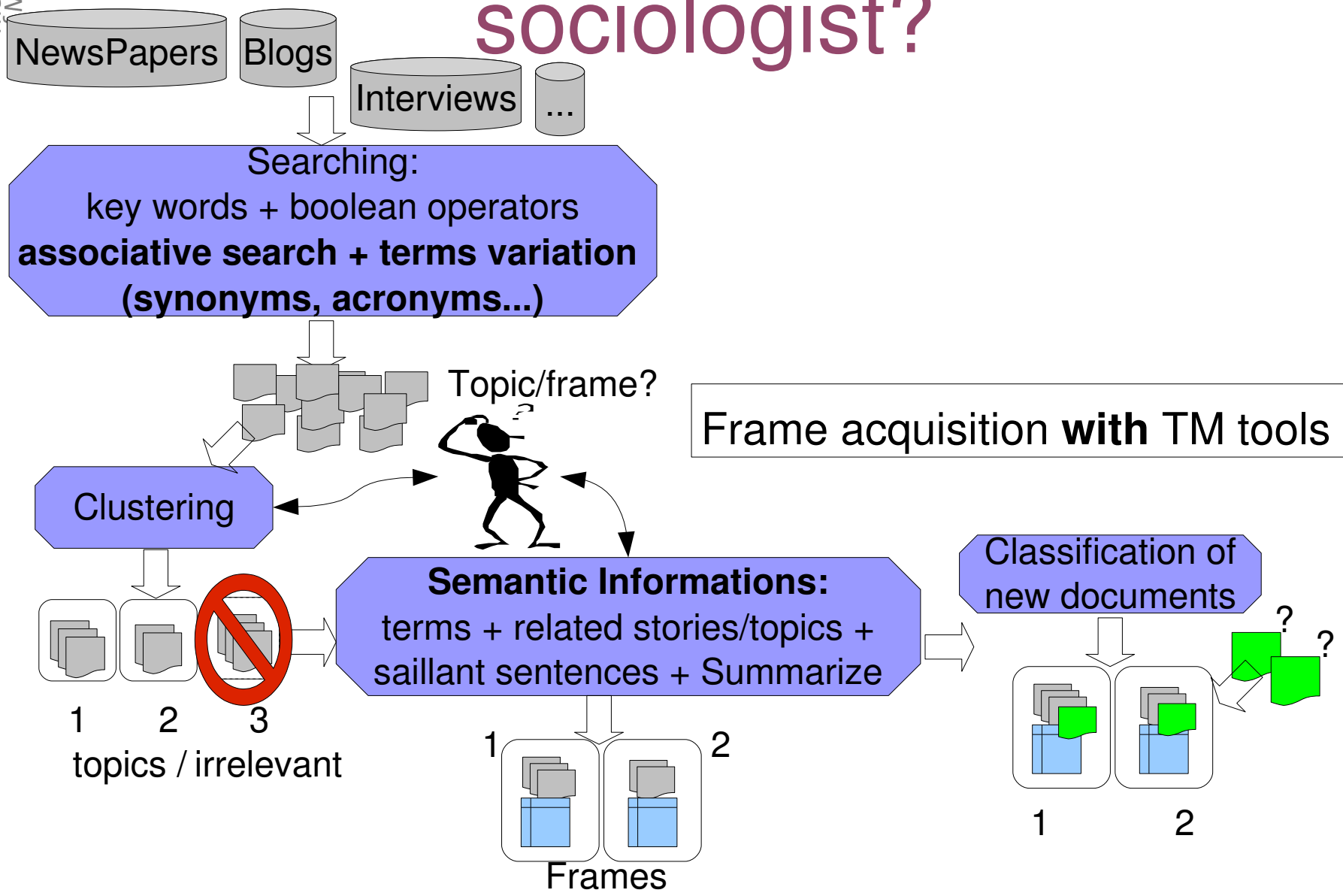
Frame definition

- 'In a communicating text, a frame is a way to promote a problem definition, causal interpretation, moral evaluation, and/or treatment recommendation'
  - problem: gun control in U.S.
  - moral evaluation:
    preserve individual rights **vs** increasing safety people
  - treatment recommendation:
    authorize **vs** forbid

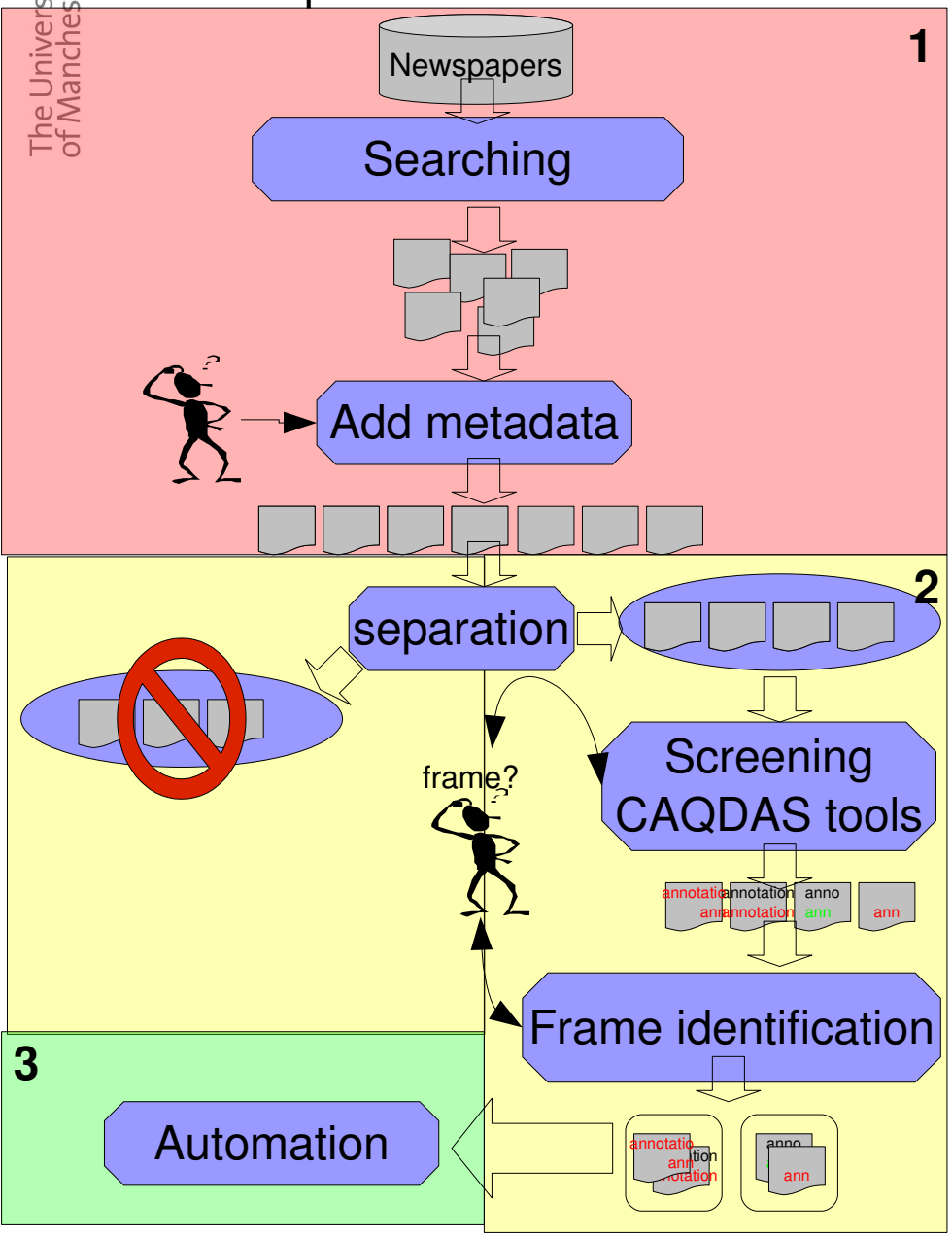
How text mining tools can help the sociologist?



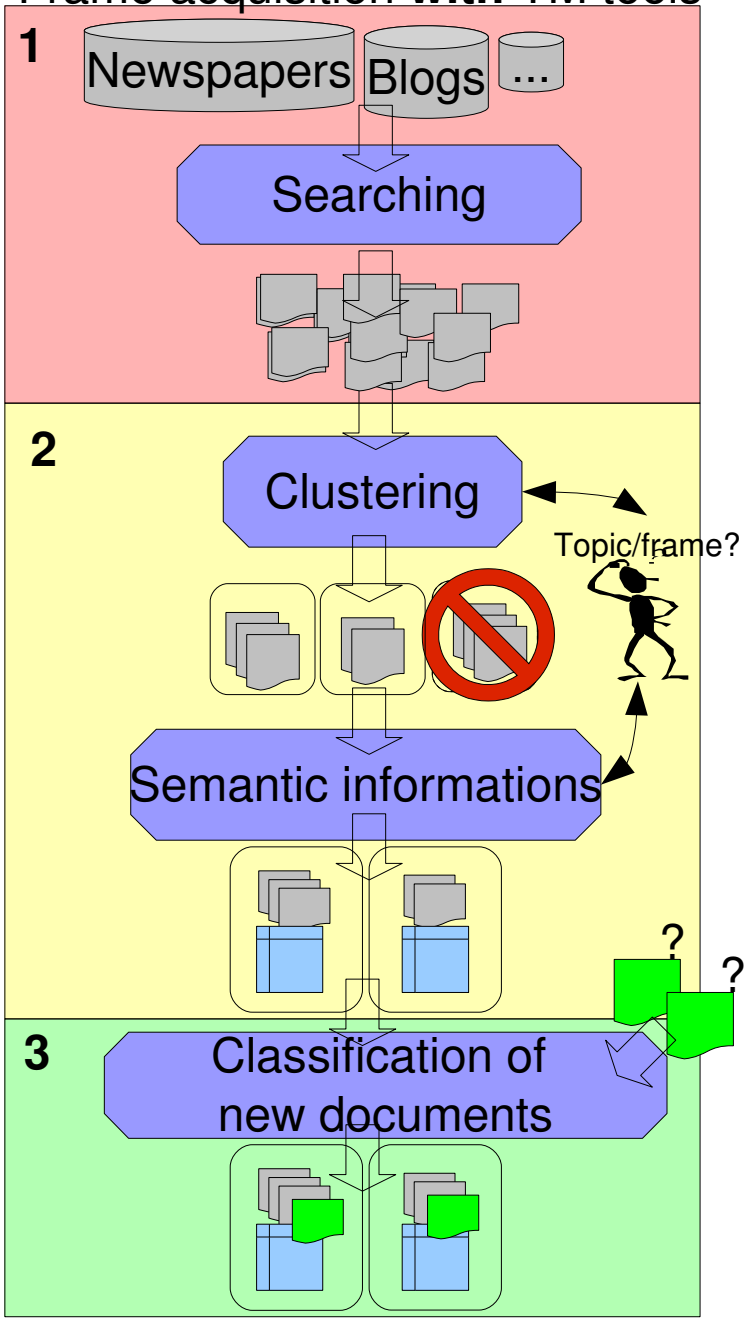
How text mining tools can help the sociologist?



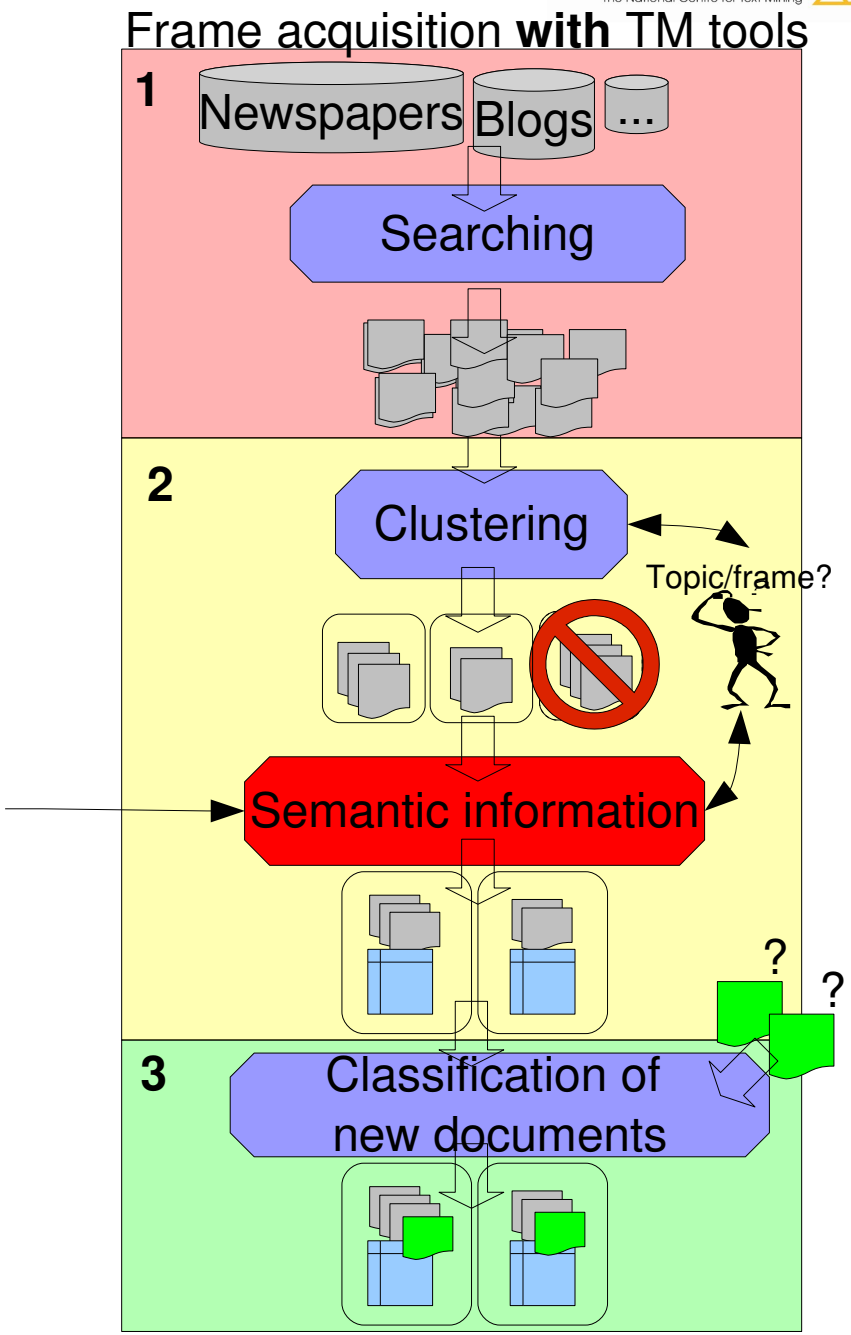
Frame acquisition **without** TM tools



Frame acquisition **with** TM tools



Which pieces of information are needed to help the frame analysis?



Which semantic information are really relevant?

- Named Entity Recognition:
  - Person [real/fiction/Nicknames...], Organization [company/institute/government...], ...
- Anaphora resolution:
  - ex. 'Gordon Brown' often refers to 'the prime minister' or to the pronoun 'he'
- Term:
  - ID card frame: 'big brother', 'illegal immigrant', 'dna database'

Which semantic information are really relevant?

- lexical chains
  - [economy/sector/economic system/...], [Rome/capital/city/...]
- summarize
  - extraction of the most important sentence in a document/set of document
- opinion mining
  - 'While Ian Angell is **correct** to point out the risks of identity theft we should [...]', 'It would be **wrong** to test first on young people and students'

Which semantic information are really relevant?

- design by user according to the studied frame
  - frame: ratchet up testing in high school
    - person working at a school, at state of government, name of exams...

MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

Conclusion

- Present the ASSIST project
- Show how to use the Text Mining tools to support the frame analysis work of sociologist
- Open a discussion around the semantic informations required to facilitate the access of the content of mass-media documents