Utrecht University

The Netherlands

Universiteit Utrecht

Research

WAHSP

Towards a flexible and stable CLARIN-supported web-application for historical sentiment mining in public media

Universiteit Utrecht

DESCARTES CENTRE

KB

Huygens Instituut

UNIVERSITEIT VAN AMSTERDAM

WAHSP

New opportunities for historical research:

The user's perspective

PROBLEMS IN HISTORICAL RESEARCH

History as science is based on a critical investigation of sources.

The validity of conclusions is strengthened by increased representativeness of the data.

PROBLEMS IN HISTORICAL RESEARCH

Available data are too extended and/or too scattered for comprehensive study and analysis.

If indices on data are available, the researcher is dependent on the author of these indices.

They are almost always incomplete and, over larger periods of times, inconsistent.

EXAMPLE: RESEARCH IN THE HISTORY OF DRUGS

S.Snelders, F.J.Meijman & T. Pieters. 'Heredity and Alcoholism in the Medical Sphere: The Netherlands, 1850-1900', *Medical History* 51 (2007) 219-236;

S. Snelders, F.J. Meijman & T. Pieters, 'Alcoholism and Hereditary Health in Dutch Medical Discourses, 1900-1945', *Social History of Alcohol and Drugs* 22 (2008) 130-143.

Use of drugs (medication) in medical (psychiatric) journals 1850-1945

Relatively small data collections (6-7 journals)

Indices of inferior quality

Changing terminology

How do we get more than qualitative indications?

KB Newspaper database

Databank Digitale Dagbladen

Journals in the collection of the Royal Library (KB) in The Hague (Krantenmagazijn Koninklijke Bibliotheek)

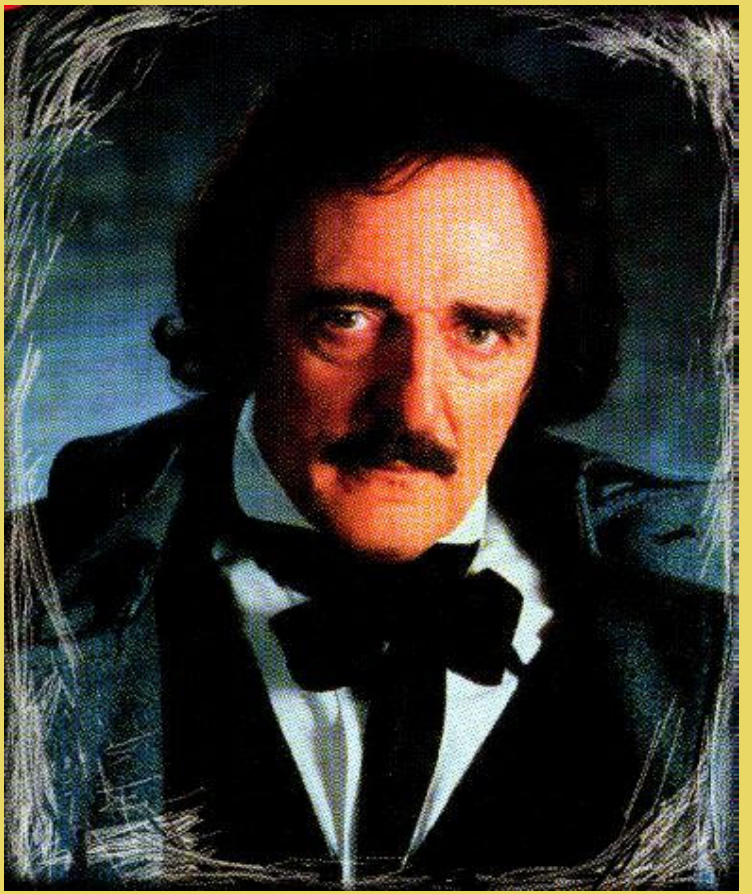
The project Databank Digitale Dagbladen digitalizes on a large scale Dutch national.regional, local and colonial newspapers and make these online and free available.

Online available
Since 28 May 2010 first results on the webservice *Historische kranten*
http://kranten.kb.nl
One and a half million newspaper pages 1618-1945.
Will be augmented in stages, up until 1995.

Edgar Allan Poe's detective

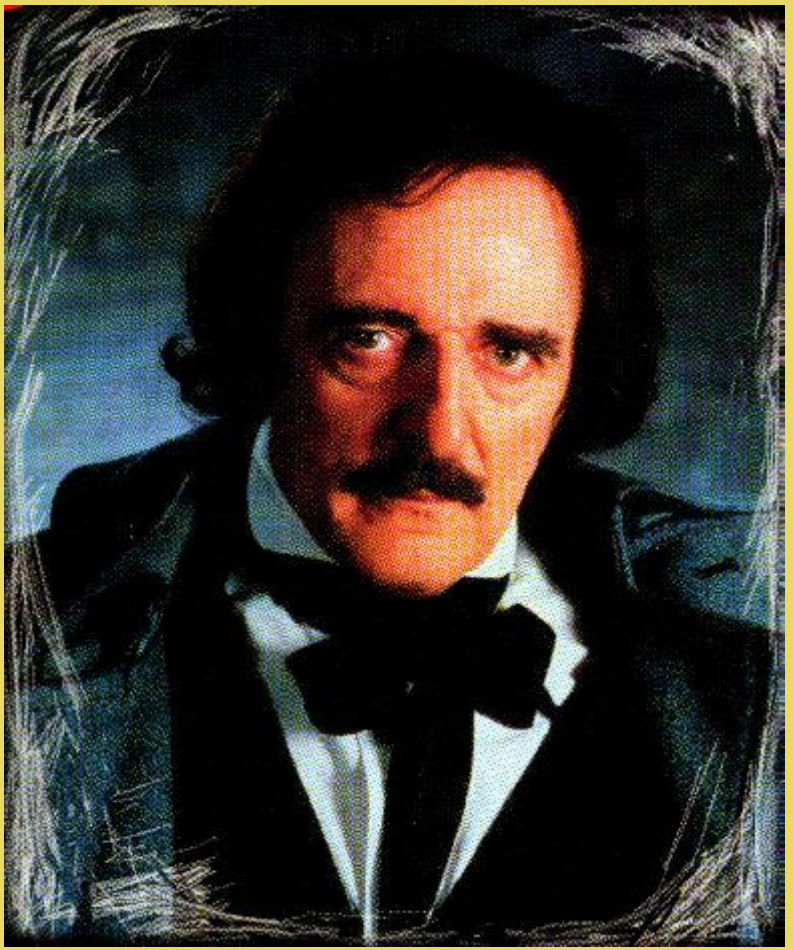
'Experience has shown, and a true philosophy will always show, that a vast, perhaps the larger portion of truth arises from the seemingly irrelevant.'

(*The Mysterie of Marie Roget, 1842)*



Poe's detective finds the truth by using data in those newspaper articles that *do not* concern the murder.

In a similar way we will find sentiments in those newspaper articles that are at first sight irrelevant.



What does the user of WAHSP want?

- Auxiliary tool for observing and analyzing trends and patterns

- Interactive tool with the possibility to adapt the original lexicon

- Possibility of production and analysis of subsets of data

- Identification of individual key documents

WE NEED:

A semi-automatic and interactive application that extracts relevant data from a mass of seemingly irrelevant data.

An application that does not replace, but supports the intuition and insights of the researcher.

Fietstas$^{\alpha}$

Text Analysis Service

Web service for text information processing:

- **Extraction: terms, names, sentiment clues,...**
- **Cross-document name normalization and linking**
- **Data analysis: compare, track dynamic changes**

Interfaces and protocols

- **RESTful web service**
- **XML, XHTML, JSON, XSLT**
- **SOAP, XML-RPC (in progress)**

Public and open-source:

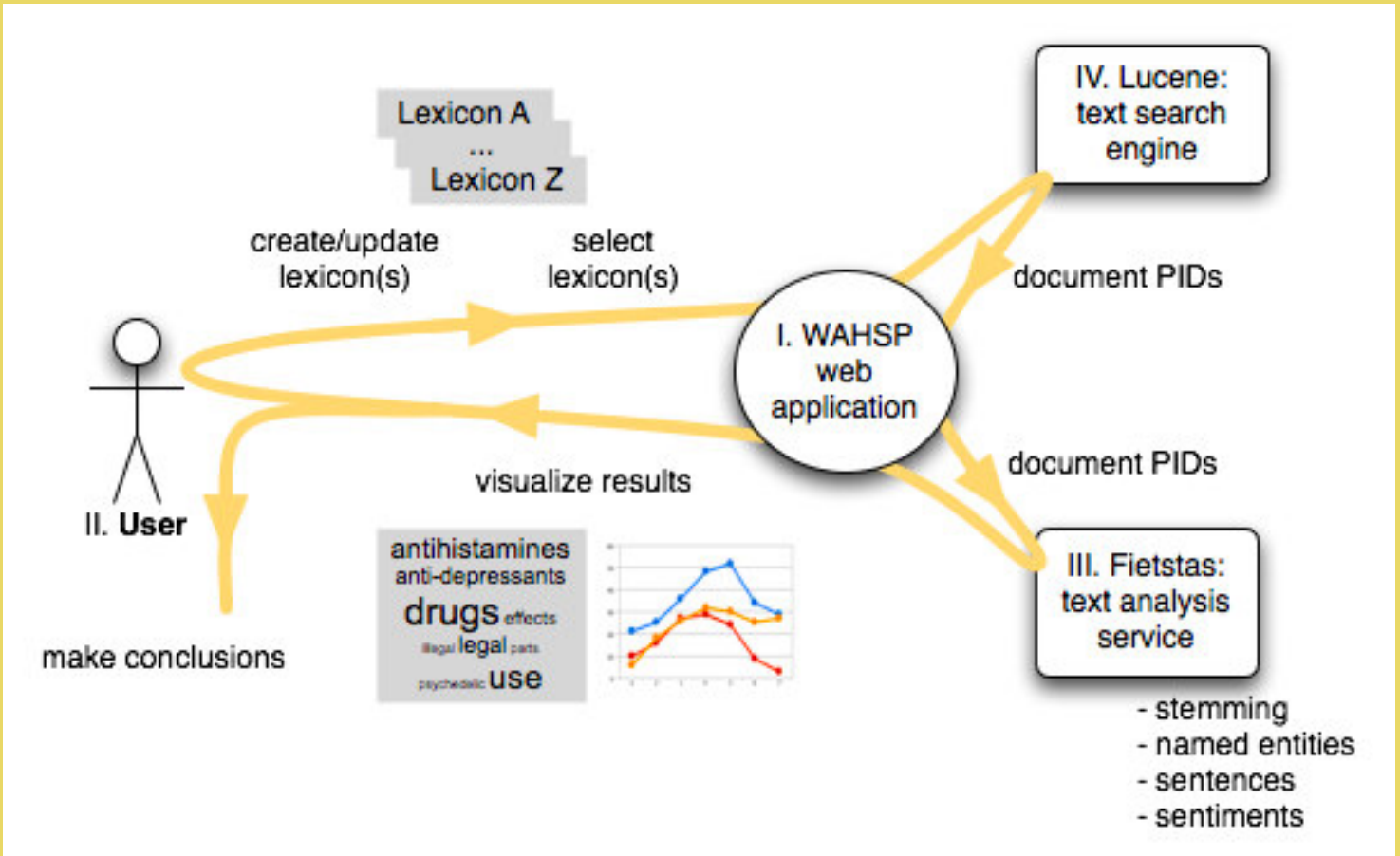
http://fietstas.science.uva.nl

UNIVERSITEIT VAN AMSTERDAM

Huygens Instituut

KB

WAHSP



Universiteit Utrecht

DESCARTES CENTRE

UNIVERSITEIT VAN AMSTERDAM