

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

MIA

Ein cloud-basierter Marktplatz für Informationen und Analysen auf dem deutschsprachigen Web



neofonie*



Technische Universität Berlin



Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages



MIA Konsortium



Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Ausgangslage

Das deutschsprachige Web

- mit derzeit *mehr als sechs Milliarden Webseiten*
- bietet ein außerordentliches Potenzial für zahlreiche Anwendungen (Apps) zum Beispiel in den Bereichen
 - **Markt- und Trendforschung,**
 - **Vertrieb von Nachrichten**
 - und allgemein in der **„Business Intelligence“** (Unternehmenssteuerung).

Wann und wie oft wird
mein Produkt in den
Nachrichten erwähnt?

Startet irgendwo ein
Shitstorm gegen eines
meiner Produkte?

Mit welchen anderen
Produkten wird mein
Produkt in Blogs in
Verbindung gebracht?

Wie kann ich mein
Marketing Budget
optimieren?

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Jedoch => breites Wissen erforderlich und hohe Kosten!!!

- Wissen in den Bereichen:
 - sprachtechnologischer Vorverarbeitung, Informationsextraktion und Annotation von Daten
 - skalierbarer Datenverarbeitung und effizienter Speicherung
 - Kosten:
 - zur Sammlung, Bereitstellung und Auswertung dieser Datenbasis,
 - für Aufbau, Betrieb und Wartung der nötigen IT-Infrastruktur.
- ***große Hürde speziell für kleine und mittelständische Unternehmen (KMUs)***

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Ziel des MIA - Projektes

Projektziel ist die **Entwicklung eines Marktplatzes**

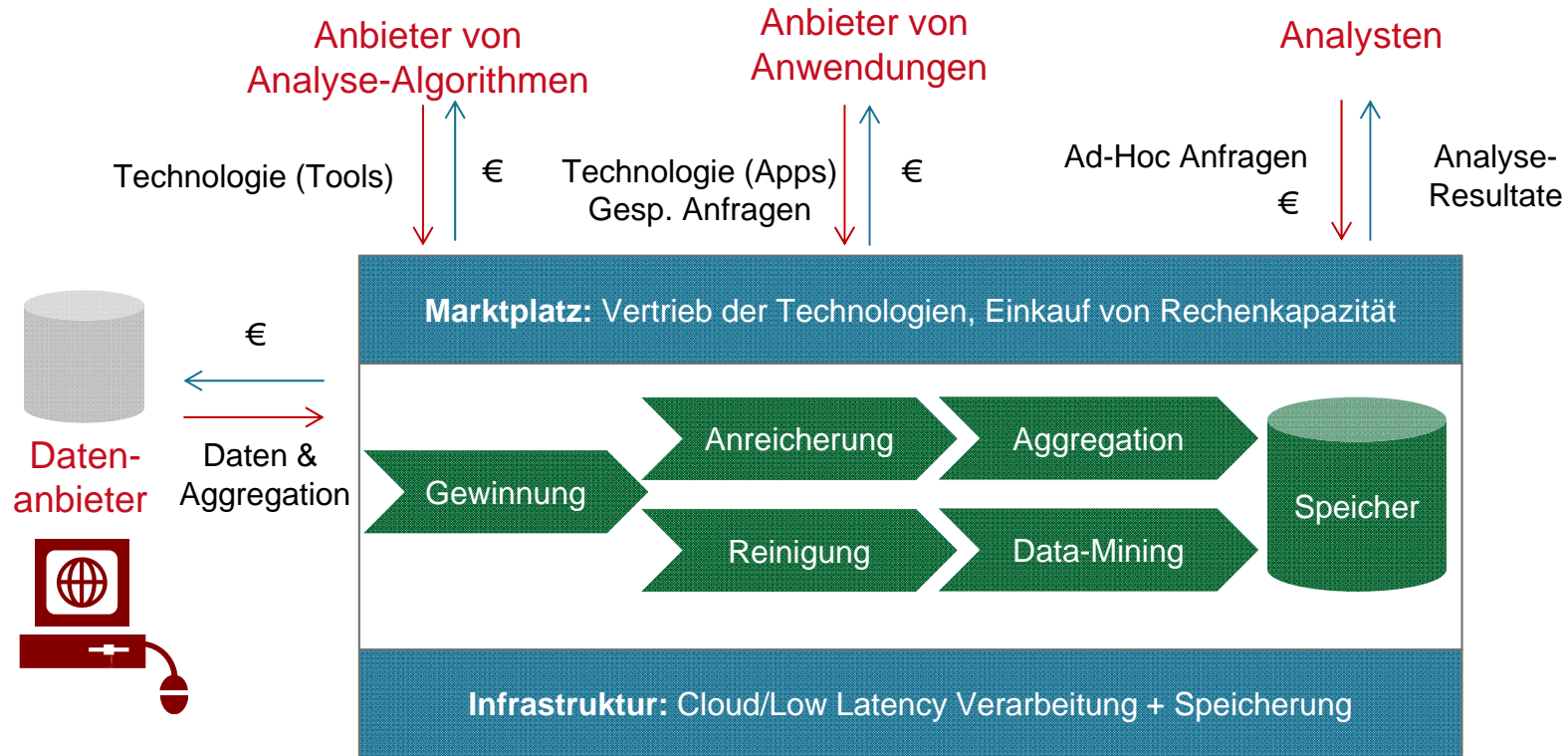
- für den nachhaltigen Betrieb
 - einer *zuverlässigen*,
 - *skalierbaren* und
 - *vertrauenswürdigen* Plattform
- zur Gewinnung, Vorhaltung und Nutzung des Datenbestandes des
 - *deutschsprachigen Webs*
 - und *anderer freier Informationsquellen*.

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

MIA – ein cloud-basierter Marktplatz für Informationen und Analysen

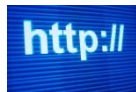


Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Datenkatalog / (Vor-)Verarbeitungsschritte



Webcrawl:

424.922.496 indizierte Seiten

Webseiten:

Klassifikation:

nach Nachrichtentopics und Erkennung medizinischer
Dokumente



Nachrichten:

115.623.652 Einträge

Nachrichten:

Satzerkennung, Tokenisierung, Lemmatisierung, POS-tagging,
Eigennamenerkennung (NER), Dependenzparsing

Teile der Nachrichten: Eigennamendisambiguierung



Blogs:

17.753.993 Einträge

Reuters Korpus (80.000 Dokumente, EN):

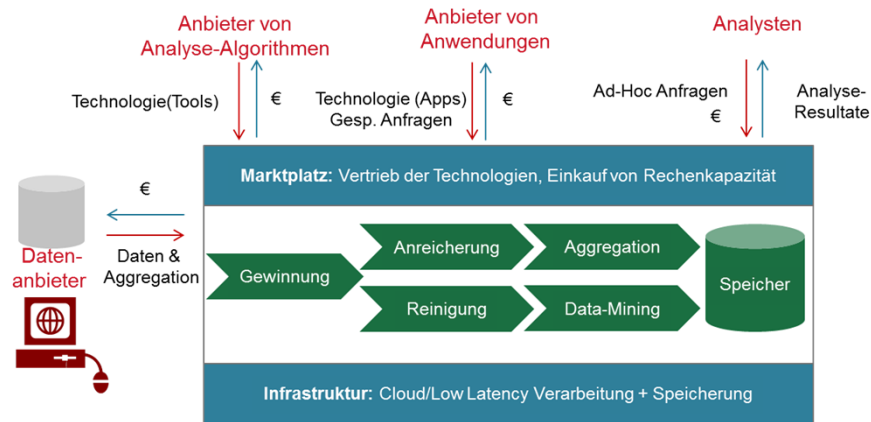
Satzerkennung, Tokenisierung, Lemmatisierung, POS-tagging, NER,
Dependenzparsing, Konstituentenparsing, Relationsextraktion
(Reverb)

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

MIA Analyseplattform und Marktplatz



realisiert durch

Über MIA Marktplatz Anfrage erstellen Modul erstellen Log-in



10 Quellen im MIA Marktplatz
Was ist der MIA Marktplatz? - Hilfe zeigen

Quellen Funktionsmodule Alle Kategorien

Automobile 2012
Tabelle mit Daten von über 500 verschiedenen Automobilen.

Common Crawl
Importiert Common Crawl Archive-Dateien im WARC-Format.

Dynamisch erzeugte Dokumente
Container für dynamisch erzeugte Dokumente aus der automatisierten MIAq-
Prozessierung.

SOCIAL MEDIA
Foren
Forenbeiträge aus einer Vielzahl deutschsprachiger Internetforen.

impressum
Impressen
Extrahierte Impressumseiten, in denen Firmendaten annotiert wurden.

http://www
Medixin-Subset vom Web-Crawl
Ein Teil (2 Mio. Dokumente) des Crawl des deutschsprachigen Internets, die mit dem bisherigen Klassifikator als Medizin-Dokumente eingestuft wurden.

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Entwicklung von Demonstratoren für Kundengewinnung (Auswahl)

Freepal/INDREX:

Im Rahmen von MIA wurde im Bereich Informations Extraktion, im speziellen an der automatischen Relationsextraktion, gearbeitet.

Dashboard interaktive Medienanalyse:

Diese Demo zeigt eine interaktive Analyse auf gecrawlten und prozessierten Marktplatzdaten mittels ParStream.

MIA-SchulApp:

Die auf Webservices basierende MIA-Webanwendung „Schullotse“ ermöglicht es, Berliner Schulen auf Basis statistischer Auswertungen zu vergleichen

WATT :

Das interaktive Web-Annotations-Tool (WATT) analysiert beliebige Texte nach semantischen Gesichtspunkten.

Zeitmaschine:

Die „Zeitmaschine“ ermöglicht die interaktive Visualisierung des gesamten Nachrichtenarchivs der Wochenzeitung DIE ZEIT.

INDREX/ParStream:

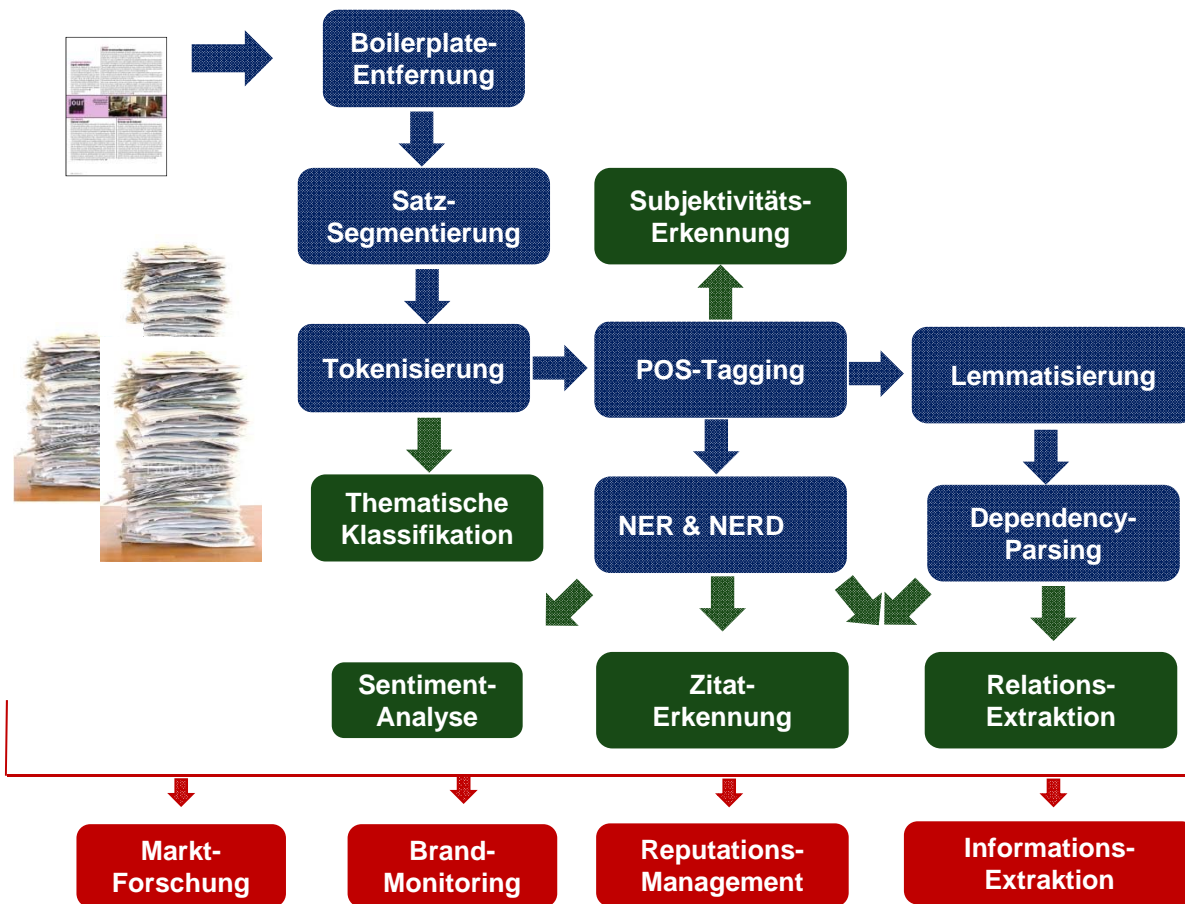
Im Rahmen von MIA wurde ebenfalls untersucht in wieweit traditionelle SQL-Datenbanken erweitert werden können, um direkt innerhalb der Datenbank Textanalyseanfragen auszuführen

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Entwicklung von Tools und Showcases

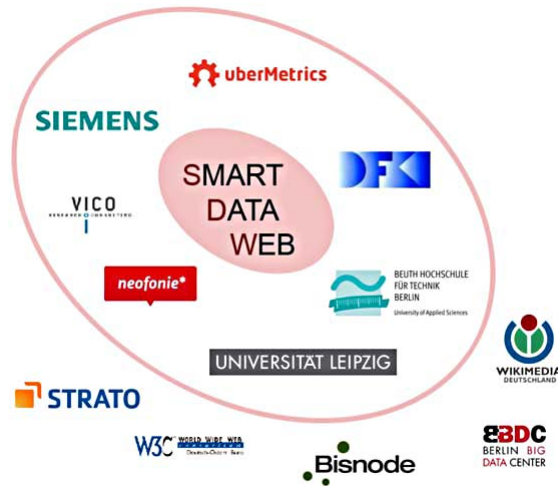


Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Wissenstransfer in Nachfolgeprojekte



Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Realisierung eines Marktplatzes inkl. deklarativer Anfragesprache



15 Quellen im MIA Marktplatz

[Was ist der Mia Marktplatz? - Hilfe zeigen](#) ↓

The screenshot shows the MIA Marktplatz interface with a grid of data sources. The interface has a top navigation bar with 'Quellen' and 'Funktionsmodule' tabs, and a dropdown menu for 'Alle Kategorien'. The grid contains the following items:

- REUTERS**: Annotierte Reuters Dokumente. Includes a download icon.
- Automobile 2012**: Tabelle mit Daten von über 500 verschiedenen Automobilen. Includes a download icon.
- Clueweb09 Freebase Annotations**: Tabelle mit Sätzen aus Clueweb09 mit Freebase Annotationen. Includes a download icon.
- Common Crawl**: Importiert Common Crawl Archive-Dateien im WARC-Format. Includes a download icon.
- Dynamisch erzeugte Dokumente**: Container für dynamisch erzeugte Dokumente aus der automatisierten Miaql-
Dokumentation. Includes a download icon.
- Foren**: Forenbeiträge aus einer Vielzahl deutschsprachiger Internetforen. Includes a download icon.

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages



Danksagung

Das Projekt MIA wird
vom Bundesministerium für Wirtschaft und Energie (BMWi)
im Rahmen des Technologiewettbewerbs "Trusted Cloud"
unter der Projektnummer 01MD11014A gefördert.

Kontakt: Dr. Holmer Hensen, TU-Berlin, FG DIMA