Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Trusted Cloud

MIA

Ein cloud-basierter **M**arktplatz für **I**nformationen und **A**nalysen auf dem deutschsprachigen Web

Fraunhofer FOKUS

neofonie*

ParStream

temis

Technische Universität Berlin

VICO
USE DATA SMARTLY

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

MIA Konsortium

neofonie*

VICO
USE DATA SMARTLY

TU berlin

temis

Fraunhofer
FOKUS

ParStream

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

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:
Bundesministerium für Wirtschaft und Energie
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:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

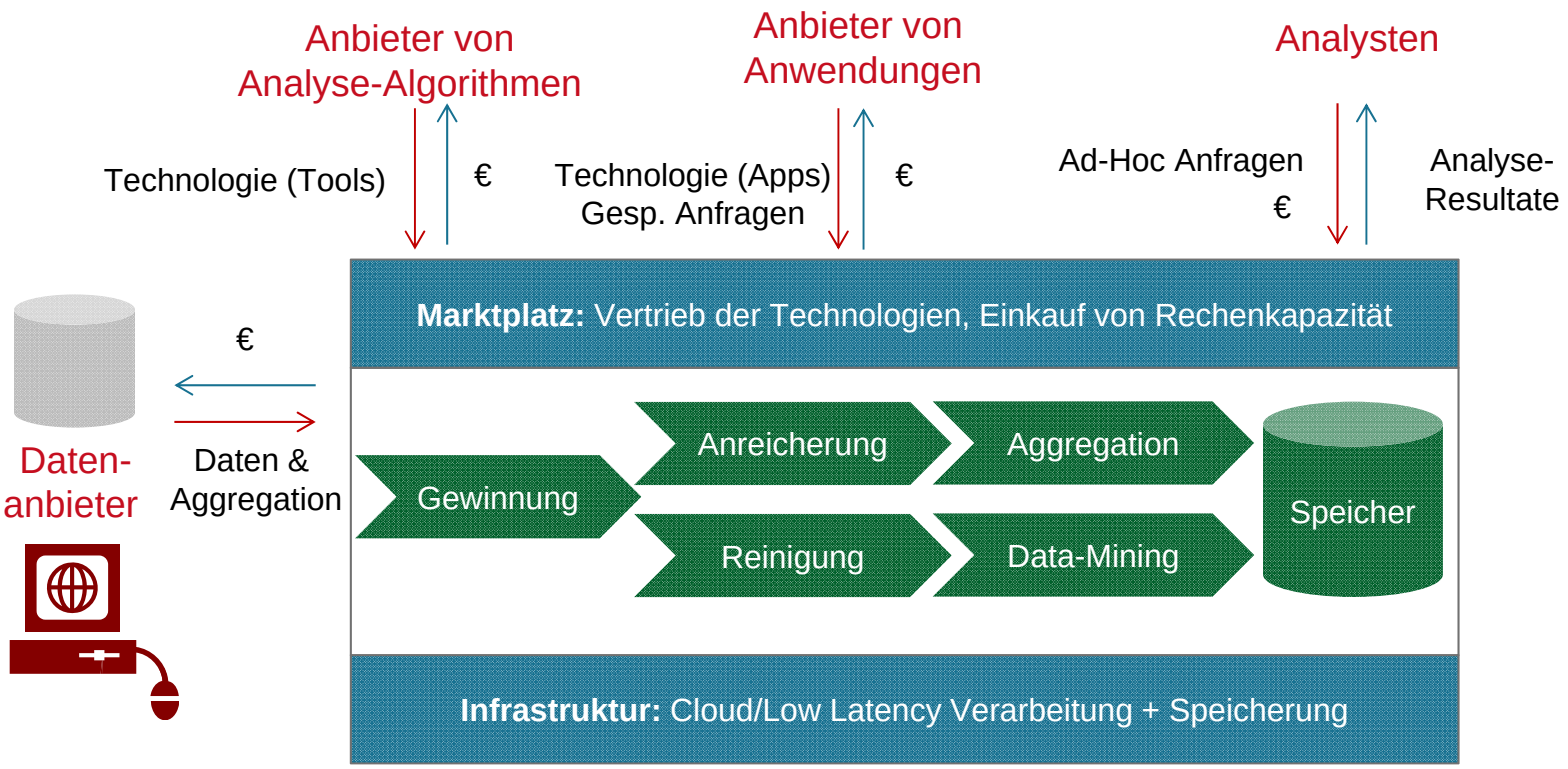
Ziel des MIA - Projektes

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:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

MIA – ein cloud-basierter Marktplatz für Informationen und Analysen



Gefördert durch:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

Datenkatalog / (Vor-)Verarbeitungsschritte

Webcrawl:
424.922.496 indizierte Seiten

Nachrichten:
115.623.652 Einträge

Blogs:
17.753.993 Einträge

Webseiten:

Klassifikation:

nach Nachrichtentopics und Erkennung medizinischer Dokumente

Nachrichten:

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

Teile der Nachrichten: Eigennamendisambiguierung

Reuters Korpus (80.000 Dokumente, EN):

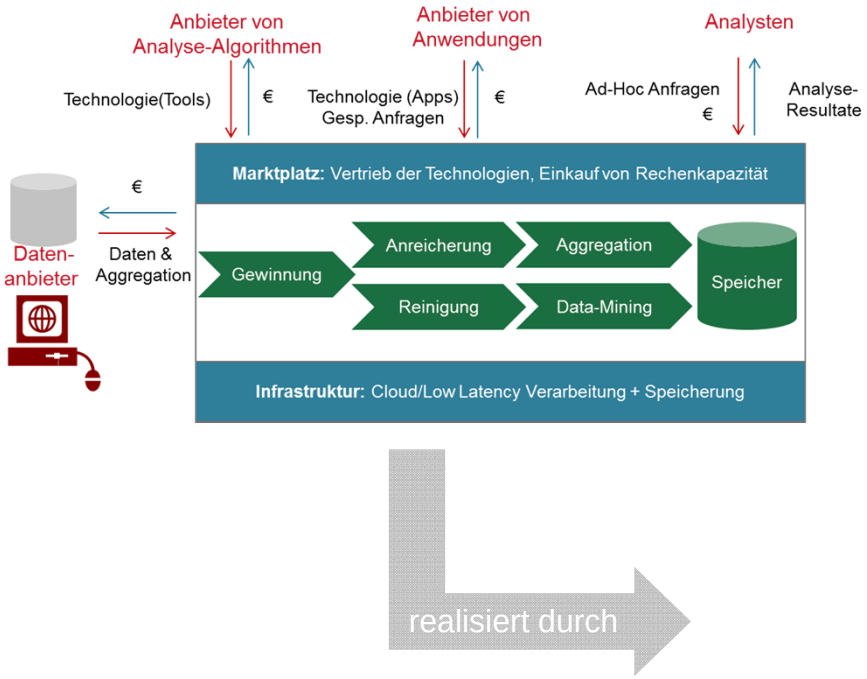
Satzerkennung, Tokenisierung, Lemmatisierung, POS-tagging, NER, Dependenzparsing, Konstituentenparsing, Relationsextraktion (Reverb)

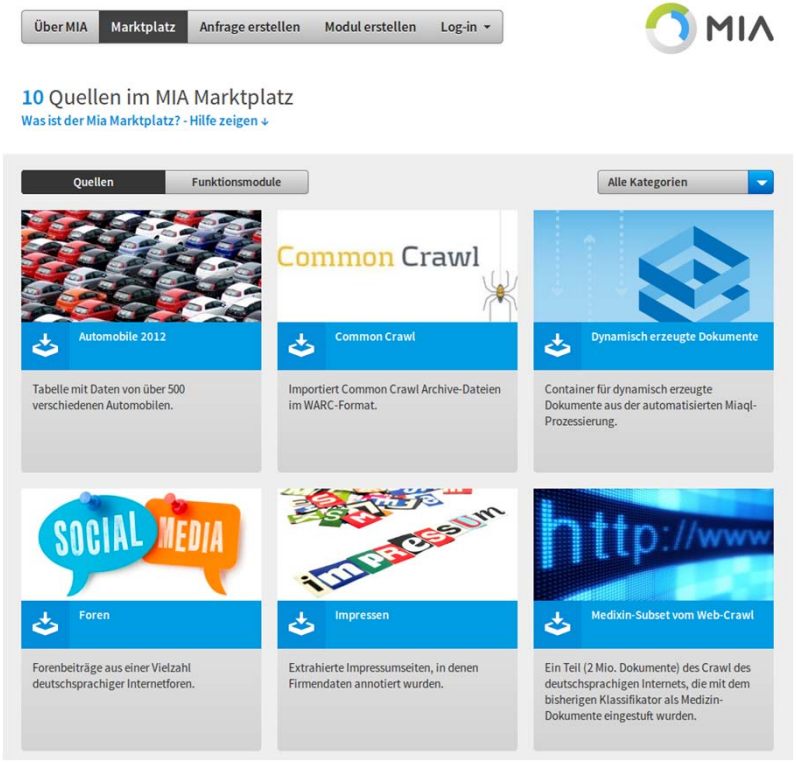
Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

MIA Analyseplattform und Marktplatz





Gefördert durch:
Bundesministerium für Wirtschaft und Energie
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 Markplatzdaten mittels ParStream.

MIA-SchulApp:

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

WATT :

Das ínteraktive 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:

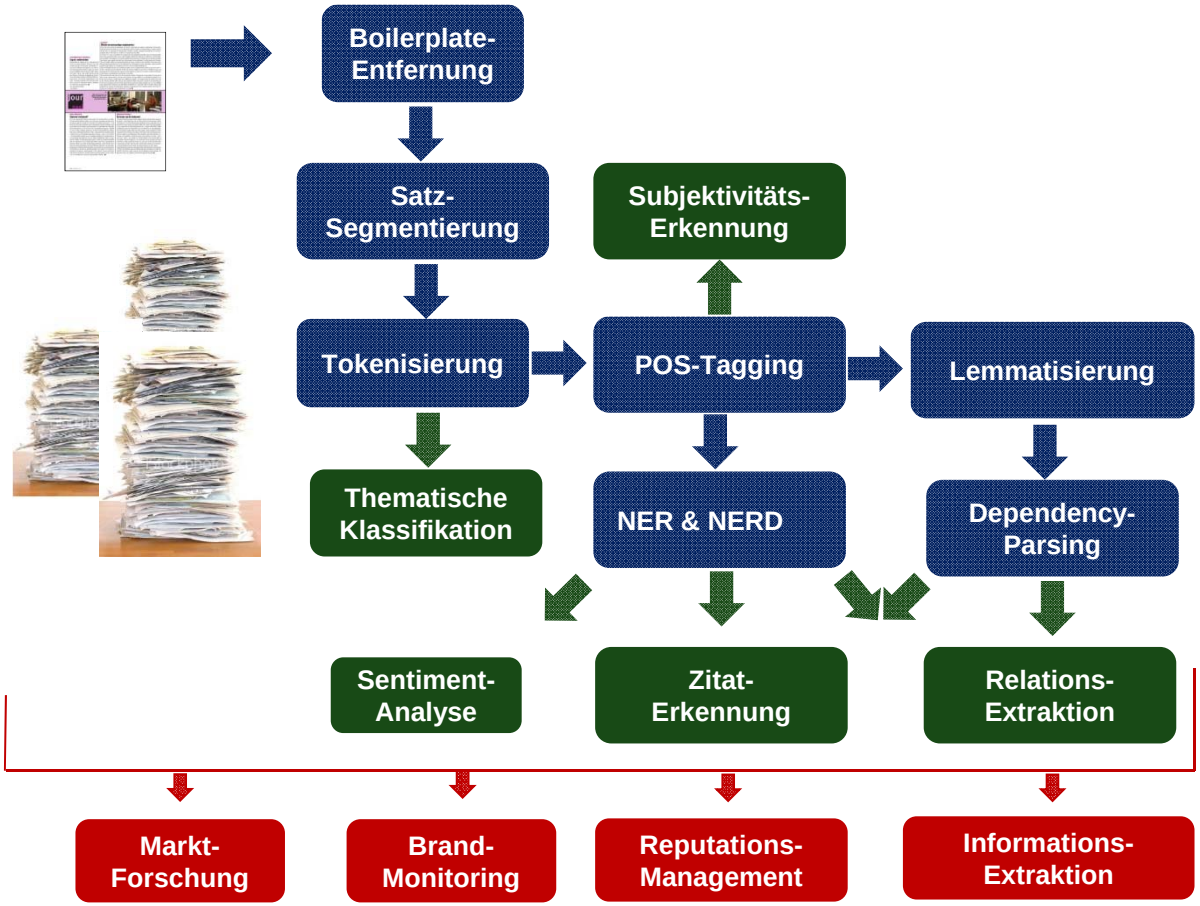
Im Rahmen von MIA wurde ebenfalls untersucht in wieweit traditionelle SQL-Datenbanken erweitert werden können, um direkt innerhalb der Datenbank Textanalyseanfragen auszuführen

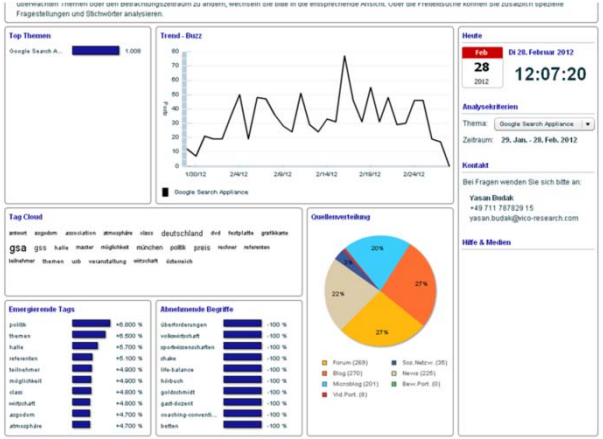
Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

Nachhaltigkeit durch Entwicklung von Tools und Showcases



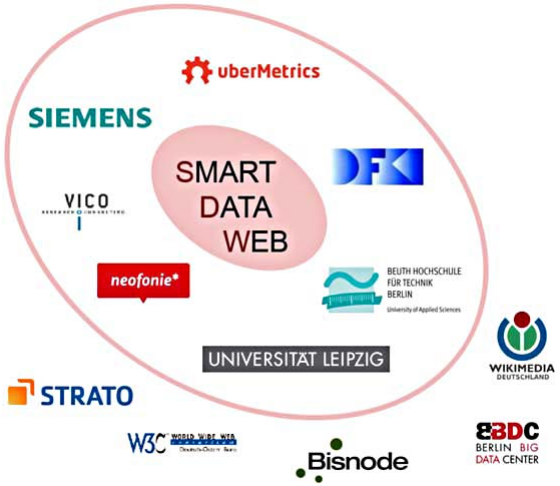


Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Wissenstransfer in Nachfolgeprojekte



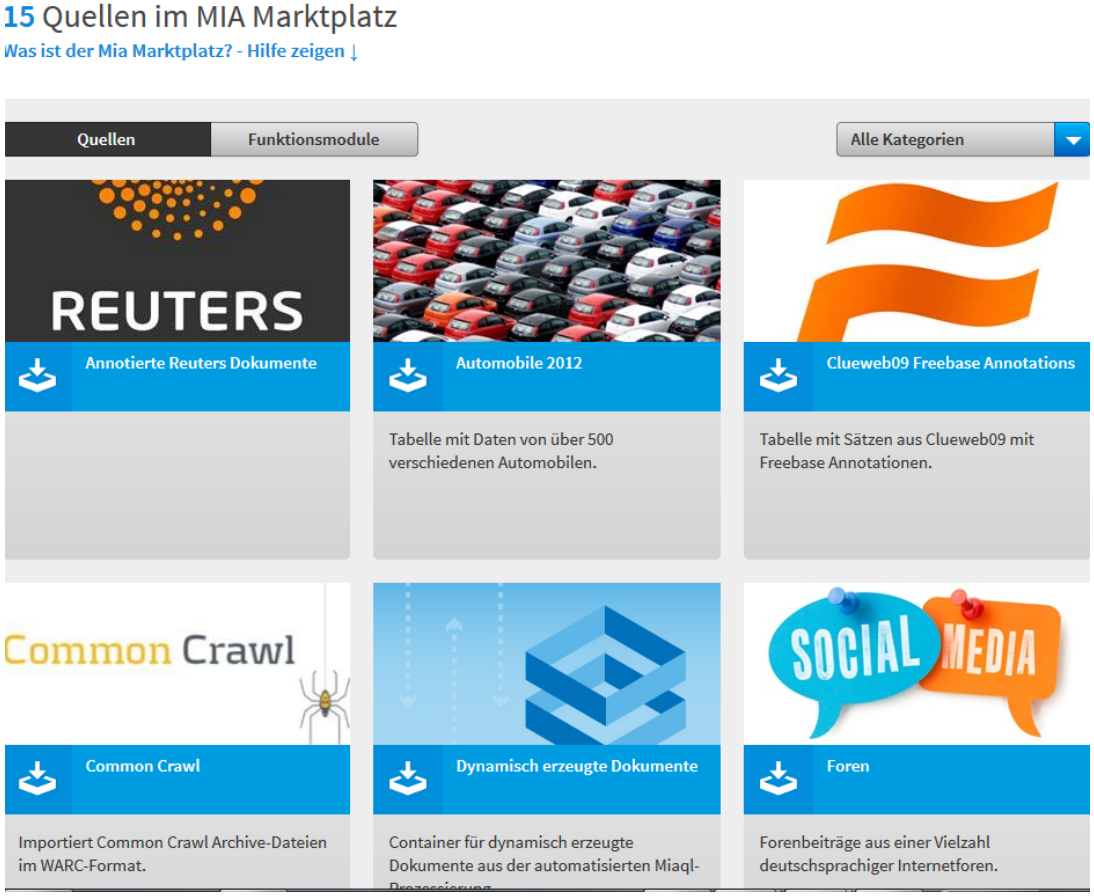
Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Nachhaltigkeit durch Realisierung eines Marktplatzes inkl. deklarativer Anfragesprache

Demo



Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

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 Hemsen, TU-Berlin, FG DIMA