Karlsruher Institut für Technologie

Fortschrittliche Suchmaschinentechnologie mit ElasticSearch für den openTA-Publikationsdienst

Clemens Düpmeier, Rainer Weidemann, Christian Schmitt (KIT / IAI)

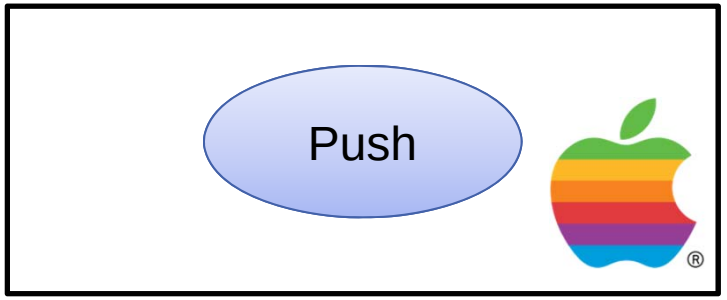
Institut für Angewandte Informatik (IAI)

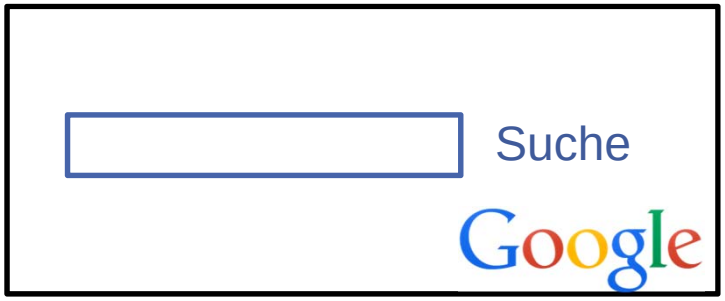
Inhalt

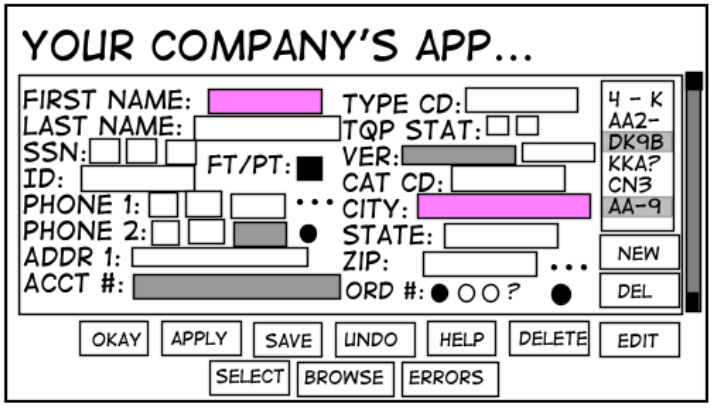
- Suchmaschinen-basierte Webanwendungen
- Strukturierte Indexserver / ElasticSearch
- Nutzung für den openTA-Publikationsdienst
- Fazit und Ausblick

Motivation – Strategien für Nutzeroberflächen

- Philosophie von Nutzeroberflächen hat sich gewandelt
- Anstatt komplexer Schnittstellen einfache Paradigmen
- Suchorientierter Ansatz eine naheliegende Variante



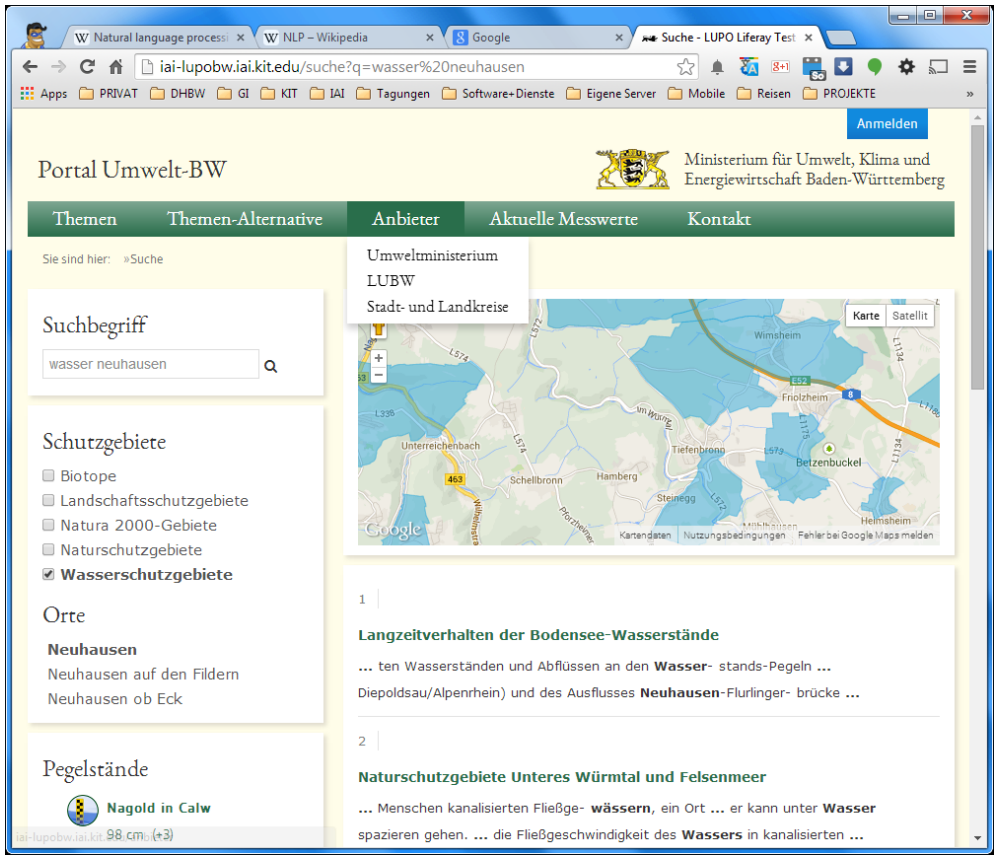




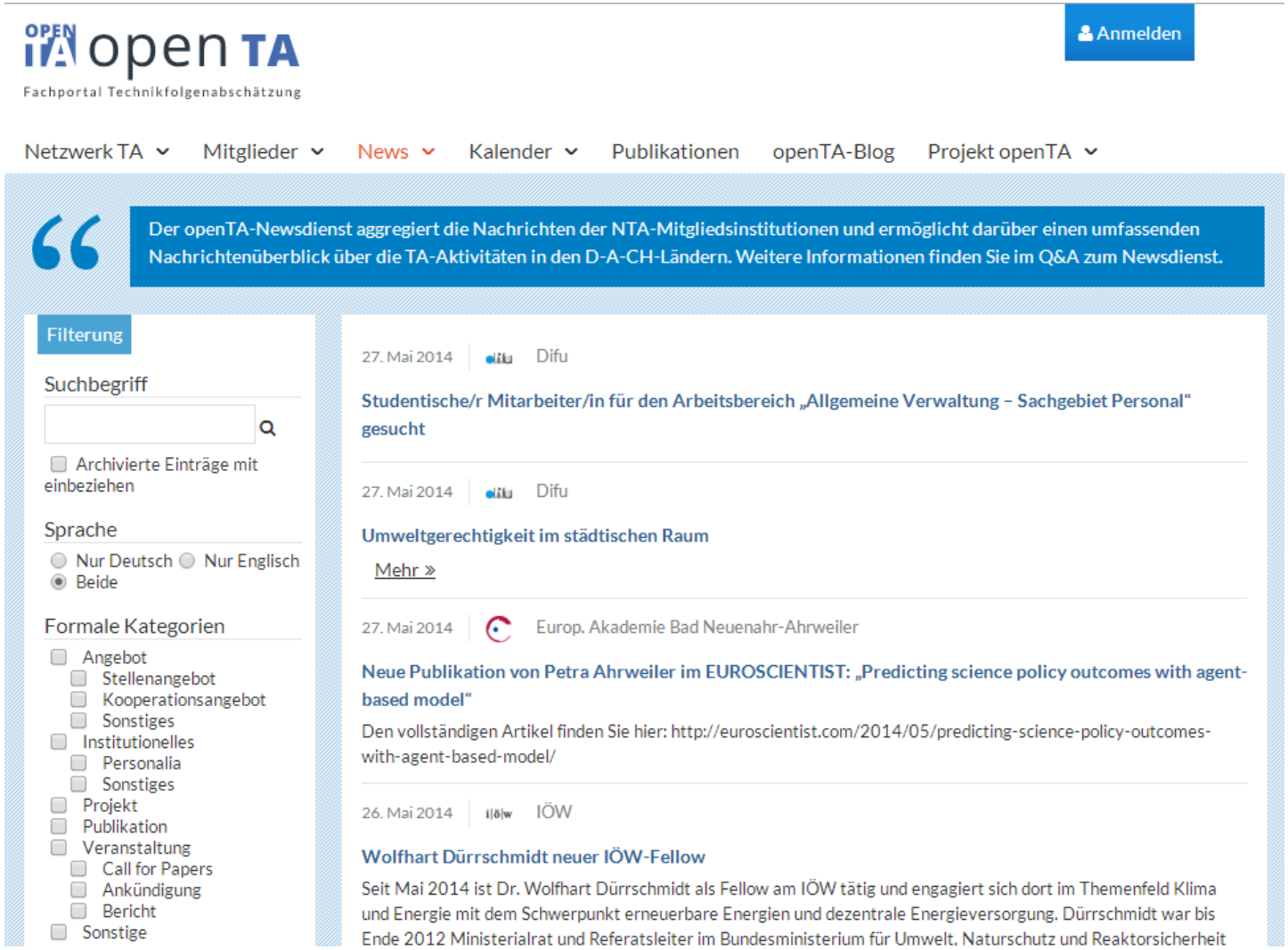
Suchmaschinen-basierte Anwendungen

- Nutzen Suchmaschinen-Technologie als Kernbestandteil
- Datengrundlage ist oft ein Mix aus
  - unstrukturierten
  - semi-strukturierten
  - strukturierten

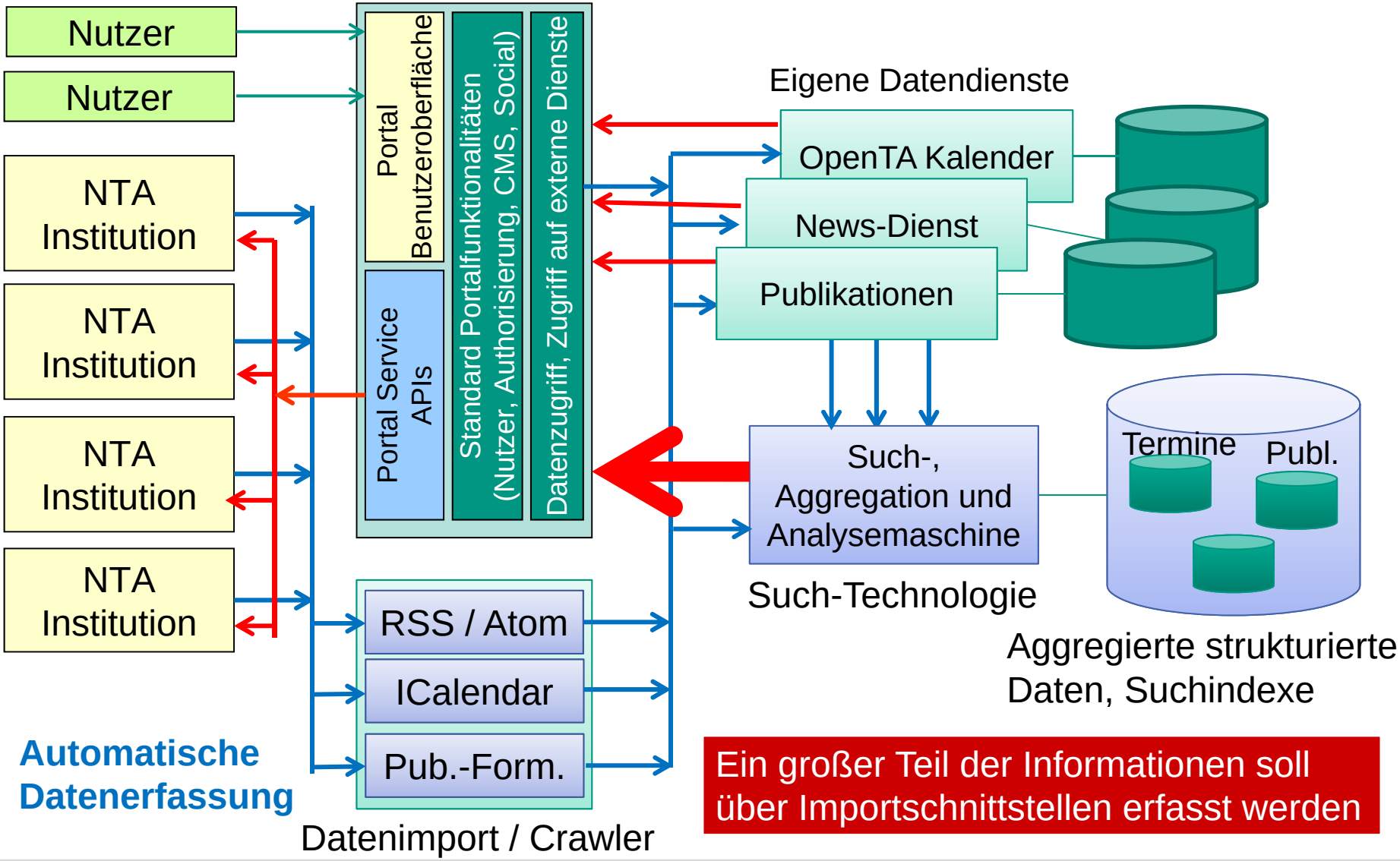
  Informationen aus verschiedenen Datenquellen / Repositorien
- Verwenden semantische Technologien zur
  - Aggregation,
  - Normalisierung und
  - Klassifikation der Informationen
- Und einen natürlich-sprachlichen Zugang zum Zugriff



openTA ebenfalls such-basiert – Bsp. News



Suchmaschinen-basierte Architektur



Welche Suchtechnologie einsetzen?

- Stand der Technik
  - Dokument-orientierte, strukturierte Indexe
  - Skalierbar über Verteilung der Indexe auf mehrere Server („Sharding“)
  - Ausgefeilte Mechanismen zur Datenanalyse, Datenaggregation und Abfrage
- Gängige Open Source Produkte sind: Solr, ElasticSearch
  - Basieren beide auf Lucine
- ElasticSearch erste Wahl für openTA
    - Neuere und modernere Architektur
    - Gegenüber dem Solr-Server werden auch hierarchische Dokumentstrukturen unterstützt
    - Übergreifende Suche über verschiedene Indexe möglich
    - Durchgehend JSON-basierte REST-Serviceschnittstellen

Einige Elemente der Suchsprache (auf Lucine basierend)

Terms	apple apple iphone
Phrases	"apple iphone"
Proximity	"apple safari"~5
Fuzzy	apple~0.8
Wildcards	app* *pp*
Boosting	apple^10 safari
Range	[2011/05/01 TO 2011/05/31] [java TO json]
Boolean	apple AND NOT iphone +apple -iphone (apple OR iphone) AND NOT review
Fields	title:iphone^15 OR body:iphone published_on:[2011/05/01 TO "2011/05/27 10:00:00"]

Indizierung bei ElasticSearch

- Strukturierte Indexe
  - Über JSON-Dokumente
  - Ohne Schema => ElasticSearch erkennt selbst Typ der Attribute
  - Kann aber auch manuell justiert werden
- Mandantenfähig => beliebig viele Indexe für unterschiedliche Mandanten möglich
- Analyse von (Text)attributen
  - Wortextraktion, Reduktion von Wörtern auf ihre Grundformen (Stemming)
  - Stoppwörter, etc.
  - Semantische Analyse (Aliasbildung, automatische Typerkennung)
  - Erweiterbar durch Plugins
- Möglichkeiten zur ID-Vergabe und Dubletten-Vermeidung

Suchfunktionalitäten

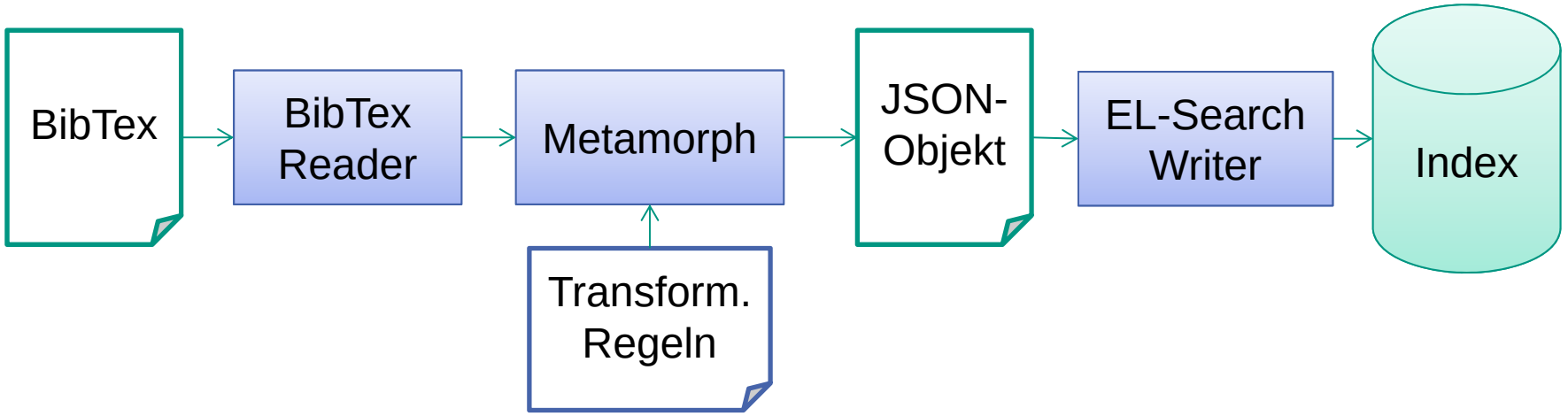
- Kombinierte Suche in den verschiedenen Attributen einer Struktur
  - Diverse Volltextsuchmöglichkeiten
    - Exakt oder mit Fehlertoleranz
      - Treffergenauigkeit einstellbar (findet Treffer auch bei Fehlern im Suchbegriff)
    - Unterstützt verschiedene Formen logischer Verknüpfung (Und / oder, …)
    - Fuzzy-Abfragen
  - Bereichsabfragen (z.B. Datumsbereiche) oder bestimmte Optionen
  - …
- Gewichtung der Treffer in verschiedenen Attributen
  - Z.B. Treffer einer Jahresangabe hat höheres Gewicht beim „Attribut Veröffentlichungsjahr“

Weitere Funktionalitäten

- Unterstützung für Mehrsprachigkeit
- Geobasierte Suchoperationen
- Unterstützung von Suchempfehlungen
- Unterstützung für Realzeitanalyse
  - Facettierung
    - Über Terme
    - Statistische Facettierung
    - Über Bereichsangaben (Klassifikation in Bereiche)
    - Filterregeln
    - Geo-Distanz
  - Statistische Berechnungen
- Percolatoren
- Aufwärmen von Indexen

BEISPIEL OPENTA PUBLIKATIONSDIENST

Überführung von Publikationsdaten in EL



- Reader für unterschiedliche Formate (transformieren in neutrales Record-basiertes Metamorph-Format
- Transformation des Formates in eigene JSON-Repräsentation über Metafacture / Metamorph (Open Source Unterprojekt von CultureGraph, DNB)
- EL-Writer schreibt die Daten in den ElasticSearch-Server

Metamorph

- Umgruppierung, Umbenennung von Attributen
- Splitten / Zusammenfügen von Attributen
- Extraktion von Teilstrings
- Transformation in andere Repräsentationen / Datentypen
- Transformation von Werten über Abbildungstabellen
- Attribute zu Records, Records zu Attribute
- Bedingte Aktionen + rekursive Bearbeitung
- Erweiterbarkeit durch eigenen Programmiercode

Bsp: Metamorph Transformationsregeln

Transformationsregeln

Splitte Authorfeld an Zeichen |

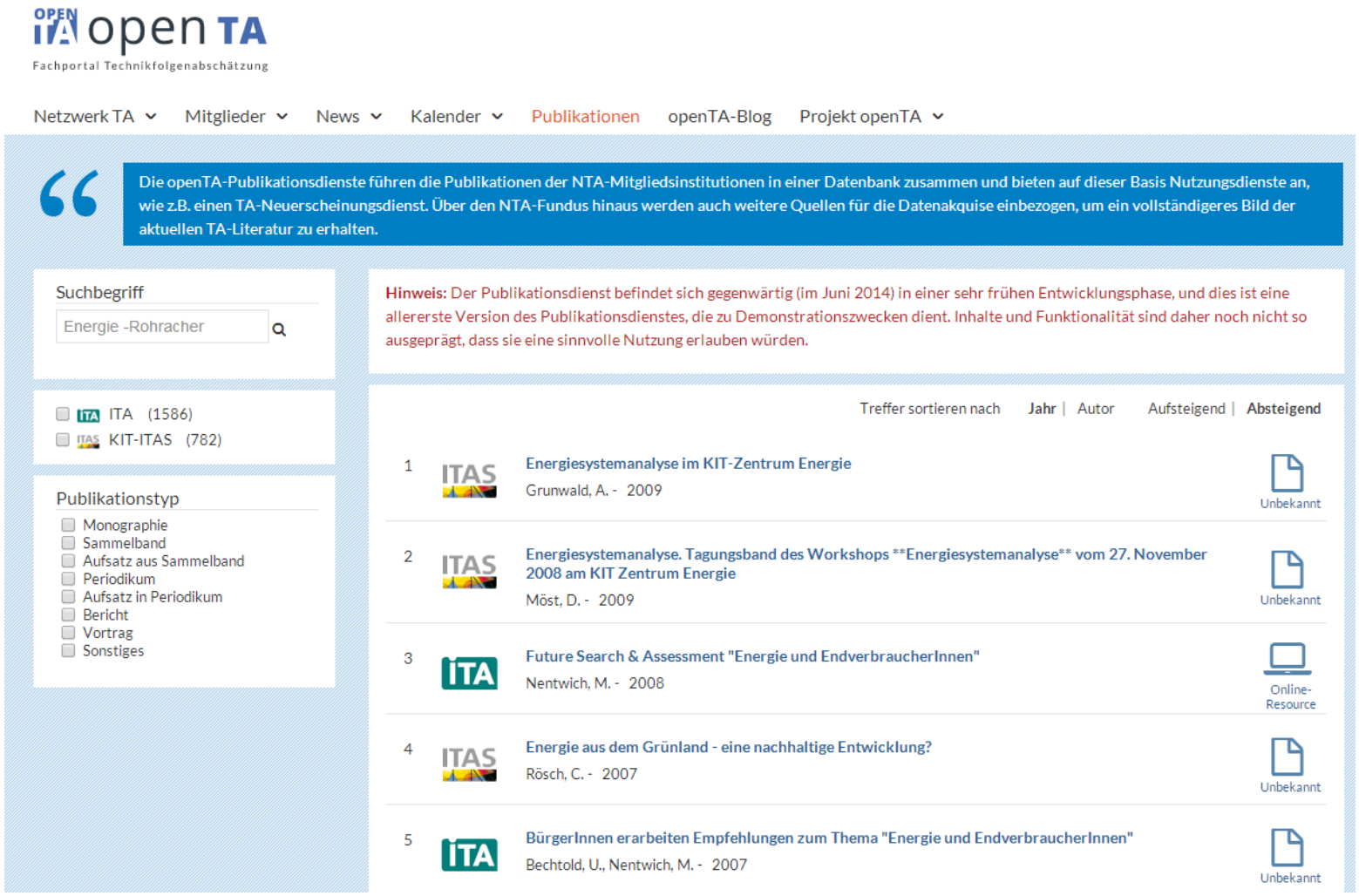
Wandle in Kleinbuchstaben und schaue Typ der Publikation in Abbildungstabelle nach

```
<?xml version="1.0" encoding="UTF-8"?>
<metamorph>
    <!-- Transformation rules -->
    <rules>
        <data name="foreignId" source="_id" />

        <data name="authors" source="author">
            <split delimiter="[|]" />
            <trim />
        </data>

        <data name="titles" source="title" />

        <data name="pubYear" source="year" />

        <data name="pubType" source="_type">
            <case to="lower" />
            <lookup in="pubtypes" />
        </data>

        <data name="contributorId" source="_contributor" />

    </rules>
</metamorph>
```

Einsatz von ElasticSearch für den openTA-Publikationsdienst

- Indizierung
  - Generierung eindeutiger Identifizierer / Dublettenkontrolle
  - Gruppierung von Dokumenten für gleiche Publikationen
  - Analyse der Attribute => Volltext-Indexbildung
- Suchfunktionalitäten
  - Textsuche mit Priorisierung der einzelnen Felder
    - Treffer in Autor-Attribut hat höheres Gewicht als Treffer im Text
    - Analog Treffer in Datumsattribut oder Titel
  - Ermittlung und Anzeige von Facettierungen zu einer Suche
    - Erscheinungsjahr
    - Top X Autoren
    - Publikationstyp

Prototyp: Suche nach „Energie –Rohracher“



Fazit und Ausblick

- Der Suchmaschinen-basierte Ansatz von openTA hat sich beim News- und Kalender-Dienst bereits bewährt
- Erster Prototyp des Publikationsdienstes ist vielversprechend
  - Import-Pipeline und Einsatz von Metafacture / Metamorph erlaubt modulare Erweiterung um neue Quellen / Formate
  - ElasticSearch-Einsatz erlaubt flexible Suche und Generierung einer ergonomischen, facettierten Suchoberfläche für den Nutzer
- Publikationsdienst wird nun in iterativen Schritten verfeinert
  - Mehr Quellen und Formate
  - Ausbau der Oberfläche und Facetten
  - Tuning der Priorisierung / Reihenfolge von Suchergebnissen

Metamorph: Publikationstypen

```
<!-- Data maps -->
<maps>

    <map name="pubtypes">
        <entry name="article" value="PERIODICAL_ARTICLE" />
        <entry name="book" value="MONOGRAPH" />
        <entry name="booklet" value="MISC" />
        <entry name="conference" value="MISC" />
        <entry name="inbook" value="COLLECTIVE_VOLUME_ARTICLE" />
        <entry name="incollection" value="COLLECTIVE_VOLUME_ARTICLE" />
        <entry name="inproceedings" value="COLLECTIVE_VOLUME_ARTICLE" />
        <entry name="manual" value="REPORT" />
        <entry name="mastersthesis" value="MISC" />
        <entry name="misc" value="MISC" />
        <entry name="phdthesis" value="MISC" />
        <entry name="proceedings" value="COLLECTIVE_VOLUME" />
        <entry name="techreport" value="REPORT" />
        <entry name="unpublished" value="MISC" />
    </map>

</maps>
```

- Metamorph erlaubt Transformation über Def.-Listen

BibTex -> JSON-Repräsentation

```
@article{ITAS-ID5582 ,
author = "Nentwich, M.| Riehm, U. ",
title = "Internationale Fachportale
für Technikfolgenabschätzung. Brauchen
wir eines oder sogar mehrere?",
journal = "Technikfolgenabschätzung -
Theorie und Praxis ",
year  = "2012",
pages = "76-80",
volume = "21",
number = "3",
evastar_pdf = "neri12a.pdf",
ISSN = "16197623",
note= "language = de"
}
```


```
{
  "id": "47e26700-90ab-38fb-9fb0-5b72fad74610",
  "foreignId": "ITAS-ID5582",
  "authors": [
    "Nentwich, M.",
    "Riehm, U."
  ],
  "editors": null,
  "pubType": "PERIODICAL_ARTICLE",
  "titles": [
    "Internationale Fachportale f\u00fcr Technikfolgenabsch\u00e4tzung.
    Brauchen wir eines oder sogar mehrere?"
  ],
  "pubYear": 2012,
  "contributorId": 12481,
  "contentType": "UNKNOWN",
  "mediaType": "UNKNOWN",
  "fullTextReference": "http:\/\/www.itas.kit.edu\/pub\/v\/2012\/neri12a.pdf",
  "valid": true,
  "created": "2014-05-28T16:23:14+0200",
  "lastModification": "2014-05-28T16:23:14+0200"
}
```


- Text-basiertes Format -> Objekt-basiertes Format
- Aufsplittung von Feldern / Umbenennung
- Normalisierung von Darstellungen (Datum, etc.)

BibTex-Format

```
@article{ITAS-ID5582 ,
author = "Nentwich, M.| Riehm, U. ",
title = "Internationale Fachportale für Technikfolgenabschätzung.
Brauchen wir eines oder sogar mehrere?",
journal = "Technikfolgenabschätzung - Theorie und Praxis ",
year  = "2012",
pages = "76-80",
volume = "21",
number = "3",
evastar_pdf = "neri12a.pdf",
ISSN = "16197623",
note= "language = de"
}
```

- Text-basiertes Format für den Transfer von Publikationsdaten zwischen Software

Internes JSON-Format

```
{
  "id": "47e26700-90ab-38fb-9fb0-5b72fad74610",
  "foreignId": "ITAS-ID5582",
  "authors": [
    "Nentwich, M.",
    "Riehm, U."
  ],
  "editors": null,
  "pubType": "PERIODICAL_ARTICLE",
  "titles": [
    "Internationale Fachportale f\u00fcr Technikfolgenabsch\u00e4tzung.
    Brauchen wir eines oder sogar mehrere?"
  ],
  "pubYear": 2012,
  "contributorId": 12481,
  "contentType": "UNKNOWN",
  "mediaType": "UNKNOWN",
  "fullTextReference": "http:\/\/www.itas.kit.edu\/pub\/v\/2012\/neri12a.pdf",
  "valid": true,
  "created": "2014-05-28T16:23:14+0200",
  "lastModification": "2014-05-28T16:23:14+0200"
}
```

- Text-basiere Repräsentation des internen JSON-Formates

openTA ebenfalls such-basiert – Bsp. Kalender