

# Fortschrittliche Suchmaschinentechnologie mit ElasticSearch für den openTA-Publikationsdienst

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

Institut für Angewandte Informatik (IAI)

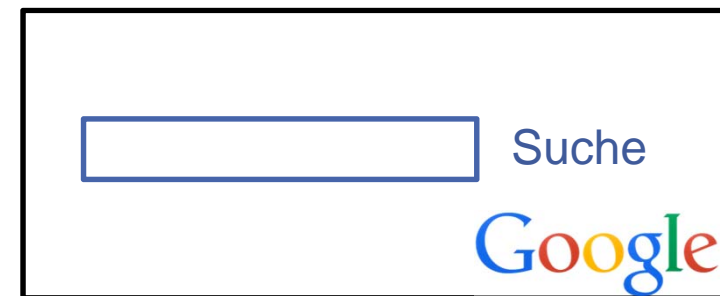
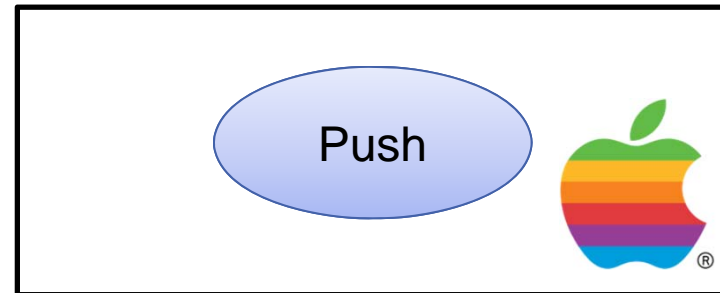


# Inhalt

- 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



YOUR COMPANY'S APP...

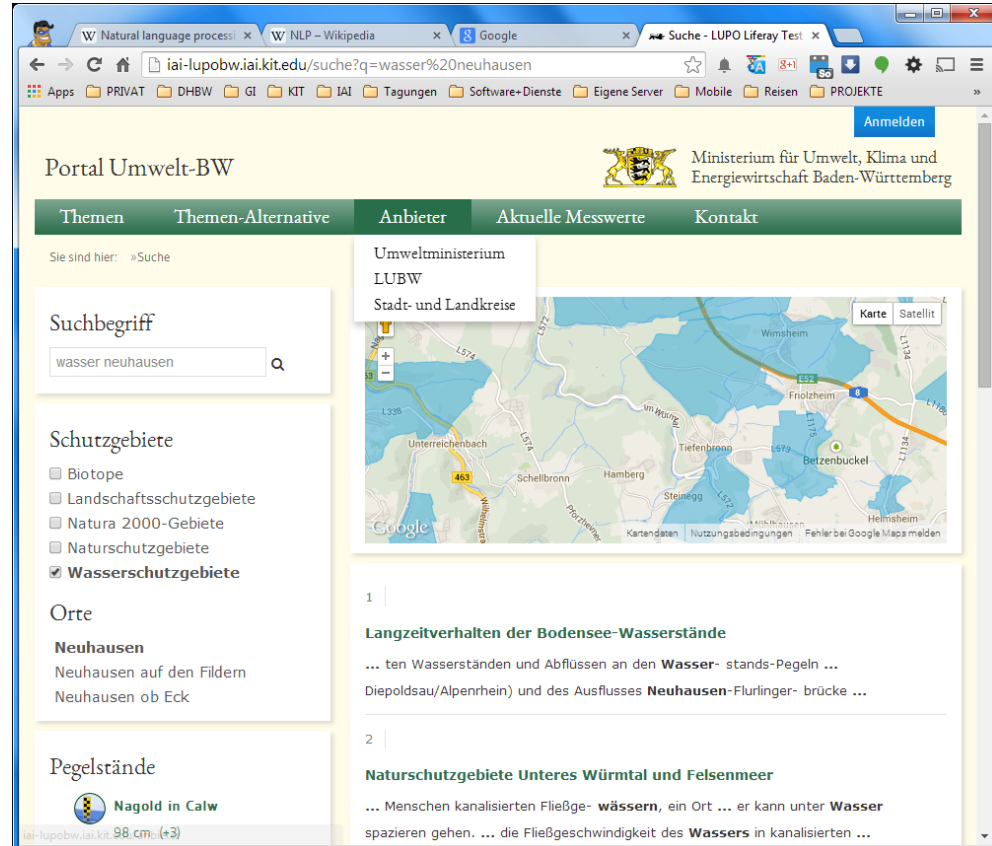
FIRST NAME: <input type="text"/>	TYPE CD: <input type="text"/>	4 - K
LAST NAME: <input type="text"/>	TQP STAT: <input type="checkbox"/>	AA2-
SSN: <input type="text"/>	FT/PT: <input type="checkbox"/>	VER: <input type="text"/>
ID: <input type="text"/>	CAT CD: <input type="text"/>	DK9B
PHONE 1: <input type="text"/>	CITY: <input type="text"/>	KKA?
PHONE 2: <input type="text"/>	STATE: <input type="text"/>	CN3
ADDR 1: <input type="text"/>	ZIP: <input type="text"/>	AA-9
ACCT #: <input type="text"/>	ORD #: <input type="text"/>	NEW

OKAY APPLY SAVE LINDO HELP DELETE EDIT

SELECT BROWSE ERRORS

# Suchmaschinen-basierte Anwendungen

- Nutzen Suchmaschinen-Technologie als Kernbestandteil
- Datengrundlage ist oft ein Mix aus
  - unstrukturierten
  - semi-strukturierten
  - strukturiertenInformationen aus verschiedenen Datenquellen / Repositorien
- Verwenden semantische Technologien zur
  - Aggregation,
  - Normalisierung und
  - Klassifikation der Informationen
- Und einen natürlich-sprachlichen Zugang zum Zugriff



The screenshot shows a web browser window displaying the 'Portal Umwelt-BW' website. The search bar contains the text 'wasser neuhausen'. The search results are organized into several sections:

- Schutzgebiete:** A list of protected areas with checkboxes. 'Wasserschutzgebiete' is checked.
- Orte:** A list of locations. 'Neuhausen' is selected, showing 'Neuhausen auf den Fildern' and 'Neuhausen ob Eck'.
- Pegelstände:** A section for water levels, with 'Nagold in Calw' selected.
- Map:** A map showing the location of Neuhausen in the region of Baden-Württemberg.
- Text Snippets:** Two snippets of text are visible:
  - 1 | **Langzeitverhalten der Bodensee-Wasserstände**  
... ten Wasserständen und Abflüssen an den **Wasser-** stands-Pegeln ...  
Diepoldsau/Alpenhein) und des Ausflusses **Neuhausen-**Flurlinger- brücke ...
  - 2 | **Naturschutzgebiete Unteres Würmtal und Felsenmeer**  
... Menschen kanalisiertem Fließge- **wässern**, ein Ort ... er kann unter **Wasser**  
spazieren gehen. ... die Fließgeschwindigkeit des **Wassers** in kanalisiertem ...

# openTA ebenfalls such-basiert – Bsp. News



Der openTA-Newsdienst aggregiert die Nachrichten der NTA-Mitgliedsinstitutionen und ermöglicht darüber einen umfassenden Nachrichtenüberblick über die TA-Aktivitäten in den D-A-CH-Ländern. Weitere Informationen finden Sie im Q&A zum Newsdienst.

## Filterung

### Suchbegriff

Archivierte Einträge mit einbeziehen

### Sprache

Nur Deutsch  Nur Englisch  
 Beide

### Formale Kategorien

- Angebot
  - Stellenangebot
  - Kooperationsangebot
  - Sonstiges
- Institutionelles
  - Personalia
  - Sonstiges
- Projekt
- Publikation
- Veranstaltung
  - Call for Papers
  - Ankündigung
  - Bericht
- Sonstige

27. Mai 2014 |  Difu

**Studentische/r Mitarbeiter/in für den Arbeitsbereich „Allgemeine Verwaltung – Sachgebiet Personal“ gesucht**

27. Mai 2014 |  Difu

**Umweltgerechtigkeit im städtischen Raum**

[Mehr »](#)

27. Mai 2014 |  Europ. Akademie Bad Neuenahr-Ahrweiler

**Neue Publikation von Petra Ahrweiler im EUROSCIENTIST: „Predicting science policy outcomes with agent-based model“**

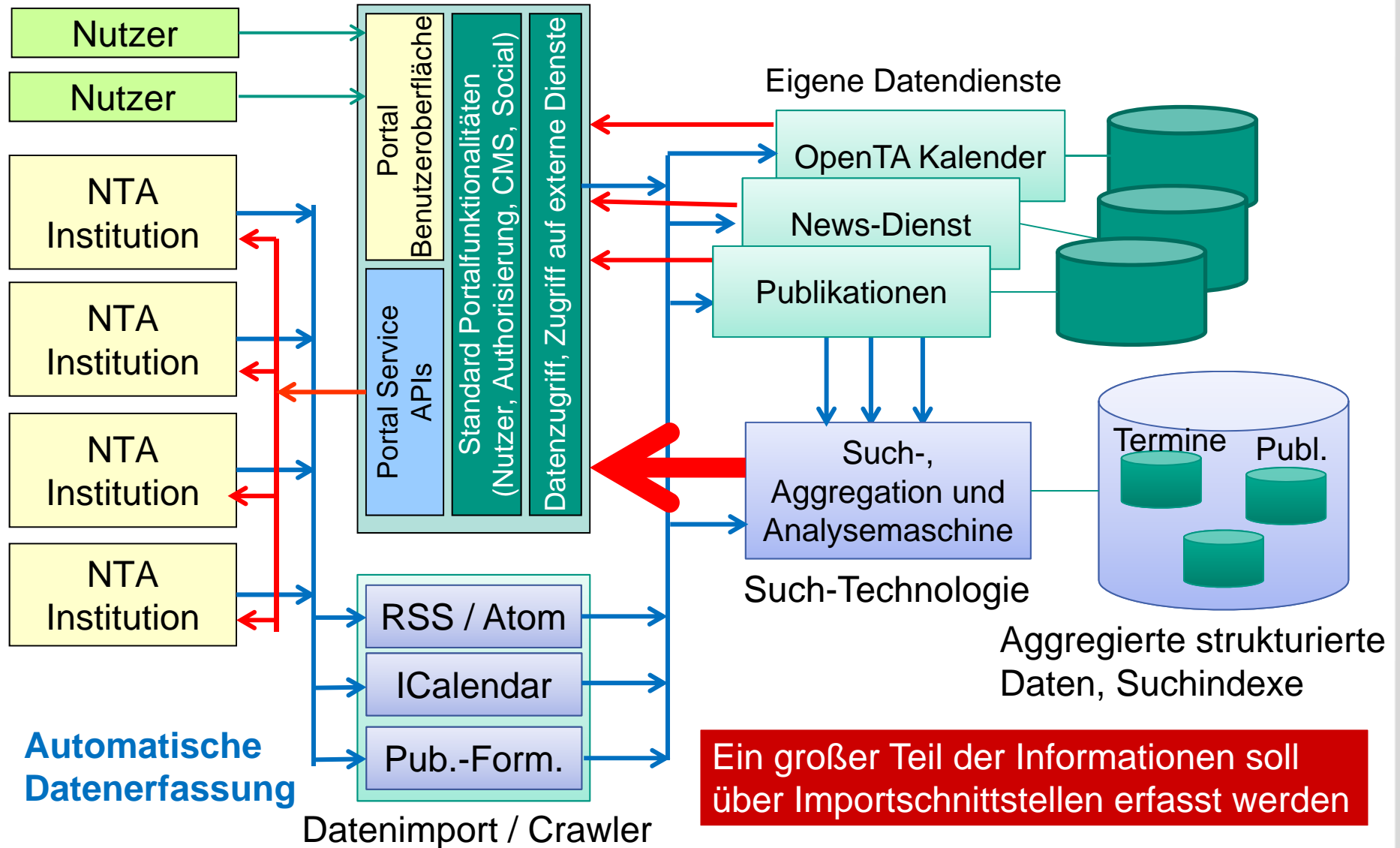
Den vollständigen Artikel finden Sie hier: <http://euroscientist.com/2014/05/predicting-science-policy-outcomes-with-agent-based-model/>

26. Mai 2014 |  IÖW

**Wolfhart Dürrschmidt neuer IÖW-Fellow**

Seit Mai 2014 ist Dr. Wolfhart Dürrschmidt als Fellow am IÖW tätig und engagiert sich dort im Themenfeld Klima und Energie mit dem Schwerpunkt erneuerbare Energien und dezentrale Energieversorgung. Dürrschmidt war bis Ende 2012 Ministerialrat und Referatsleiter im Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit

# 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

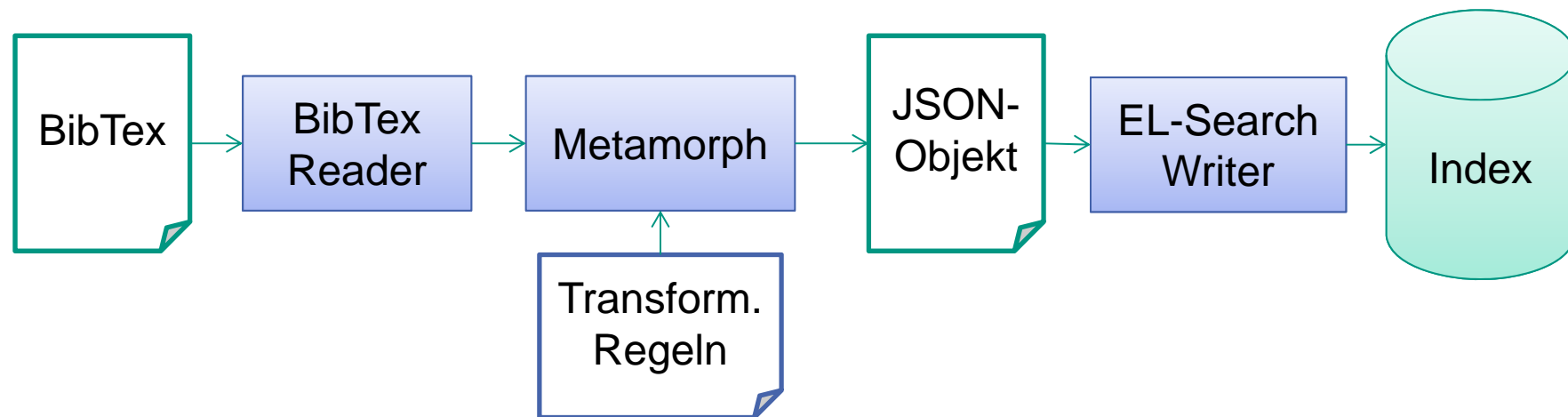
- Kombinierte Suche in den verschiedenen Attributen einer Struktur
  - Diverse Volltextsuchmöglichkeiten
    - Exakt oder mit Fehlertoleranz
      - Treffergenauigkeit einstellbar (findet Treffer auch bei Fehlern im Suchbegriff)
    - Unterstützt verschiedene Formen logischer Verknüpfung (Und / oder, ...)
    - 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 Metafactory / 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

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

Splitte Authorfeld  
an Zeichen |

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

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

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

Wandle in Kleinbuchstaben  
und schaue Typ der  
Publikation in  
Abbildungstabelle nach

```
    <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“



Die openTA-Publikationsdienste führen die Publikationen der NTA-Mitgliedsinstitutionen in einer Datenbank zusammen und bieten auf dieser Basis Nutzungsdienste an, wie z.B. einen TA-Neuerscheinungsdienst. Über den NTA-Fundus hinaus werden auch weitere Quellen für die Datenakquise einbezogen, um ein vollständigeres Bild der aktuellen TA-Literatur zu erhalten.

Suchbegriff

Energie -Rohracher











- ITA ITA (1586)
- ITAS KIT-ITAS (782)

Publikationstyp

- Monographie
- Sammelband
- Aufsatz aus Sammelband
- Periodikum
- Aufsatz in Periodikum
- Bericht
- Vortrag
- Sonstiges

**Hinweis:** Der Publikationsdienst befindet sich gegenwärtig (im Juni 2014) in einer sehr frühen Entwicklungsphase, und dies ist eine allererste Version des Publikationsdienstes, die zu Demonstrationszwecken dient. Inhalte und Funktionalität sind daher noch nicht so ausgeprägt, dass sie eine sinnvolle Nutzung erlauben würden.

Treffer sortieren nach Jahr | Autor Aufsteigend | Absteigend

- |   |   |  |  |
|---|---|--|--|
| 1 |    | Energiesystemanalyse im KIT-Zentrum Energie<br>Grunwald, A. - 2009   | <br>Unbekannt         |
| 2 |   | Energiesystemanalyse. Tagungsband des Workshops **Energiesystemanalyse** vom 27. November 2008 am KIT Zentrum Energie<br>Möst, D. - 2009 | <br>Unbekannt        |
| 3 |  | Future Search & Assessment "Energie und EndverbraucherInnen"<br>Nentwich, M. - 2008  | <br>Online-Resource |
| 4 |  | Energie aus dem Grünland - eine nachhaltige Entwicklung?<br>Rösch, C. - 2007   | <br>Unbekannt       |
| 5 |  | BürgerInnen erarbeiten Empfehlungen zum Thema "Energie und EndverbraucherInnen"<br>Bechtold, U., Nentwich, M. - 2007                     | <br>Unbekannt       |

## 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 Metafactory / 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 = "neril2a.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/neril2a.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 = "neril2a.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



Die Kalender im openTA-Portal bieten einen Überblick über die öffentlich angekündigten Veranstaltungen der NTA-Institutionen und über weitere Termine mit TA-Relevanz.

Suchbegriff

Kalender

- Verant. des NTA
- Verant. der NTA-Mitgliedsinst.
- Verant. mit TA-Relevanz
- TA-Lehrveranstaltungen

Kategorien

- Call
- Konferenz
- Lehrveranstaltung
- Vortrag
- Workshop
- Sonstiges

Calendar

1. Jun 2014

**Call: 9th International Conference on Body Area Networks**

European Alliance for Innovation

Info: <http://bodynets.org/2014>

2. Jun 2014

**Dritter openTA-Workshop**

Kontakt: [ulrich.riehm@kit.edu](mailto:ulrich.riehm@kit.edu)

Info: <http://www.openta.net/workshops>

2. Jun 2014 - 4. Jun 2014

**NTA6-TA14: Responsible Innovation. Neue Impulse für die Technikfolgenabschätzung?**

Veranstalter: ITA Wien Konferenz-Website: <http://www.oeaw.ac.at/ita/veranstaltungen/konferenzen/nta6-ta14-2014/ueberblick> Call: <http://www.oeaw.ac.at/ita/fileadmin/redaktion/Veranstaltungen/konferenzen/ta14/NTA6-TA14-CfP.pdf> #Konferenz

Mehr

Monat Woche Tag

Juni 2014

Mo	Di	Mi	Do	Fr	Sa	So
26	27	28	29	30	31	1
A. Grunwald: Ethik und...	09:30 Symposium Converging... 18:00 A. Grunwald:				Auf der Suche nach den... Call:...	Call: 9th International...