Automation at your Fingertips

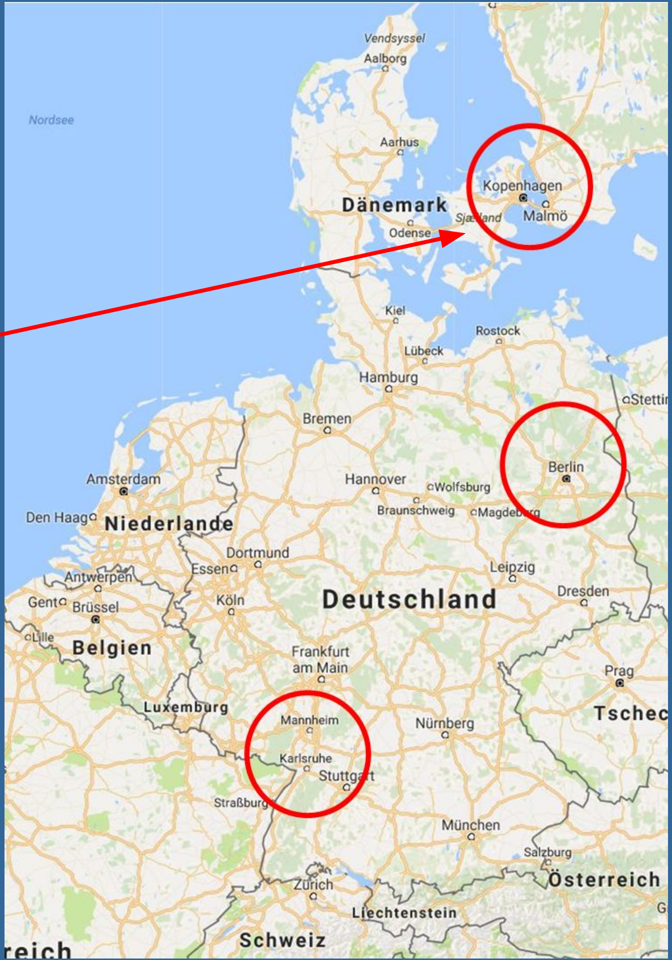
Metadata-based autocompletion for Primo (and possibly others)

Felix Ostrowski (graphthinking GmbH, Berlin)
Uwe Dierolf (KIT Library, Karlsruhe)

Introduction

- In the beginning there was EXIT !

ELAG 2016 | JUNE 6-9 IN COPENHAGEN



Felix Ostrowski

Uwe

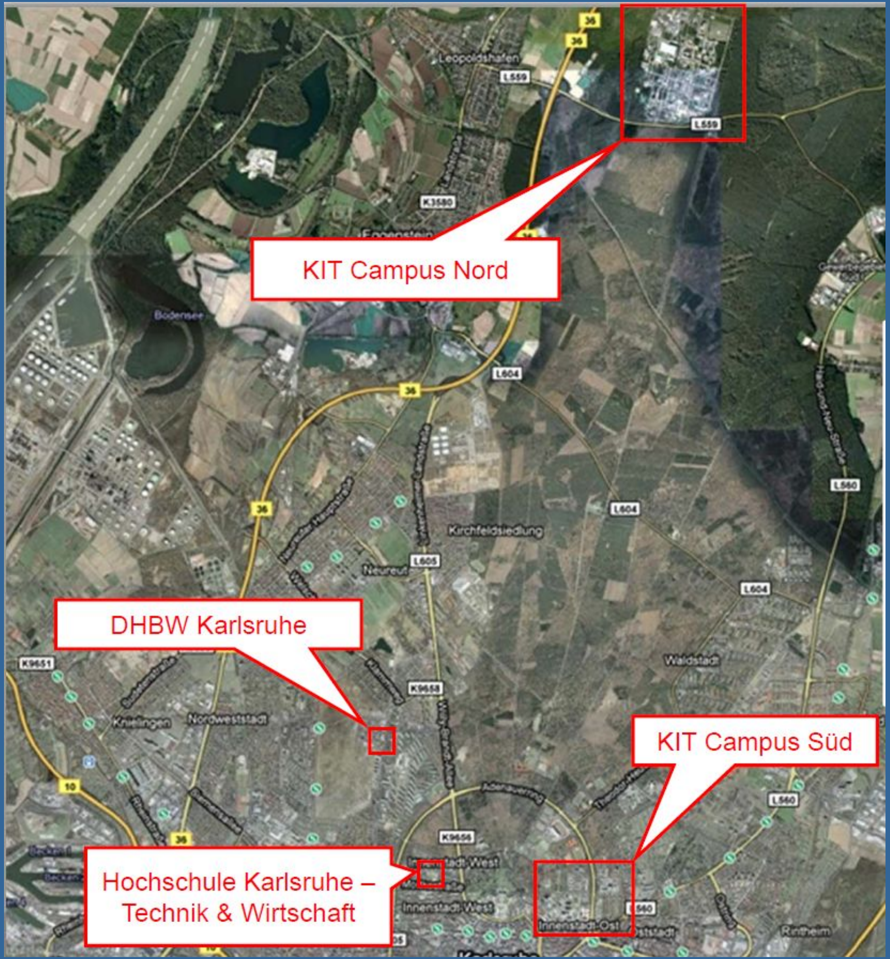
Royal Library sunbeds - a project idea was born

KIT Library

- KIT = Karlsruhe Institute of Technology
- 24/7 since April 2006
- 26.000 students
- 10.000 staff and researchers
- 5000 visitors per day

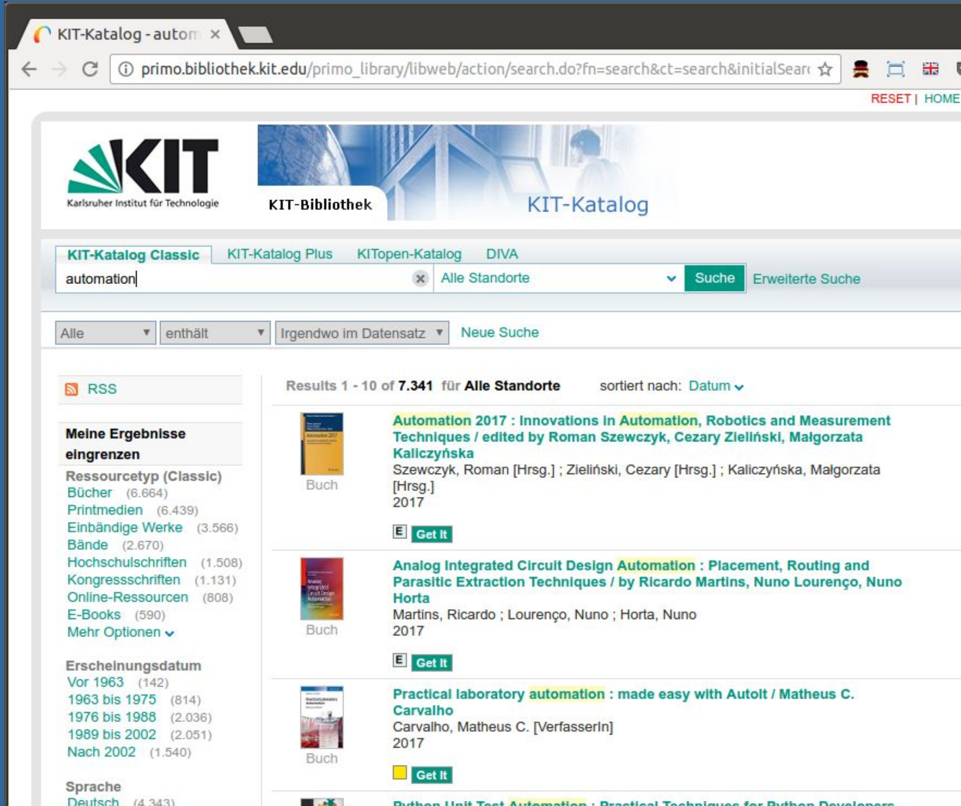
KIT Library as a service provider

- Hochschule Karlsruhe
  - university of applied sciences
  - 8500 students
  - 1000 staff & researchers
- DHBW Karlsruhe
  - Baden-Württemberg Cooperative State University
  - 3500 students
  - 700 assistant professors
  - 80 full-time professors



The KIT catalog

- is based on Ex Libris' resource discovery system Primo
- running at KIT since april 2012
- includes titles from different scopes
  - 3 universities
  - institutional repository KITopen
- offers features such as faceting
- but lacks an useable autocomplete feature



KIT-Katalog - autom ×

primo.bibliothek.kit.edu/primo_library/libweb/action/search.do?fn=search&ct=search&initialSear

RESET | HOME

Karlsruher Institut für Technologie

KIT-Bibliothek

KIT-Katalog

KIT-Katalog Classic | KIT-Katalog Plus | KITopen-Katalog | DIVA

automation

Alle Standorte

Suche

Erweiterte Suche

Alle ▾ | enthält ▾ | Irgendwo im Datensatz ▾

Neue Suche

RSS

Meine Ergebnisse eingrenzen

Ressourcetyp (Classic)

Bücher	(6.664)
Printmedien	(6.439)
Einbändige Werke	(3.566)
Bände	(2.670)
Hochschulschriften	(1.508)
Kongressschriften	(1.131)
Online-Ressourcen	(808)
E-Books	(590)

Mehr Optionen

Erscheinungsdatum

Vor 1963	(142)
1963 bis 1975	(814)
1976 bis 1988	(2.036)
1989 bis 2002	(2.051)
Nach 2002	(1.540)

Sprache

Deutsch	(4.343)

Results 1 - 10 of **7.341** für **Alle Standorte**   sortiert nach: Datum

Automation 2017 : Innovations in Automation, Robotics and Measurement Techniques / edited by Roman Szewczyk, Cezary Zieliński, Małgorzata Kaliczyńska
Szewczyk, Roman [Hrsg.] ; Zieliński, Cezary [Hrsg.] ; Kaliczyńska, Małgorzata [Hrsg.]
2017
Buch
E Get It

Analog Integrated Circuit Design Automation : Placement, Routing and Parasitic Extraction Techniques / by Ricardo Martins, Nuno Lourenço, Nuno Horta
Martins, Ricardo ; Lourenço, Nuno ; Horta, Nuno
2017
Buch
E Get It

Practical laboratory automation : made easy with AutoIt / Matheus C. Carvalho
Carvalho, Matheus C. [VerfasserIn]
2017
Buch
Get It

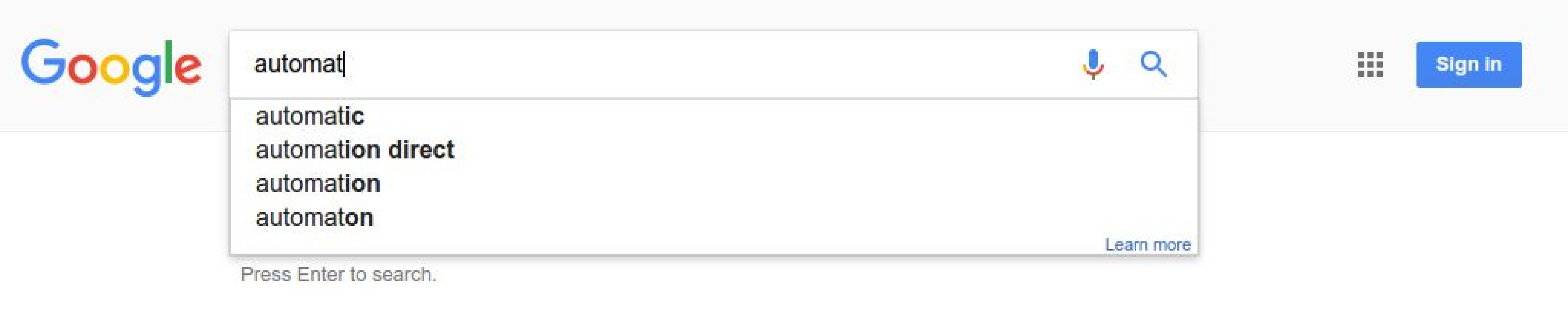
Python Unit Test Automation : Practical Techniques for Python Developers

Requirements for a self-made autocompletion

- Autocomplete current word
- Suggest next word
- Correct spelling errors
- Avoid zero-hit suggestions
- Consider Primo scopes and facets
- Offer advanced mode for librarians and power-users
- Integration of this service should be possible also in the libraries website

Google-like Autocomplete

- Suggestions are based on previous queries
- This works well with large scale usage and index size
- Avoids zero hit queries



automat

automatic
automation direct
automation
automaton

Press Enter to search.

Problems in Library Land

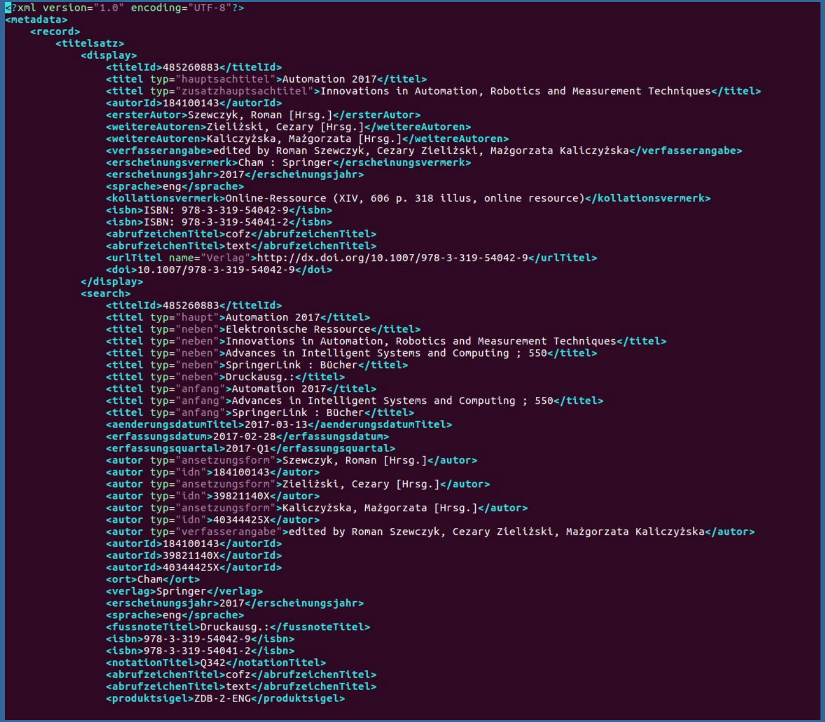
- Usage of systems like Primo typically experience much less usage
- Deriving sensible suggestions from user input is almost impossible
- Thus Ex Libris' autocompletion is based on global user input
- At least in case of KIT catalog the results are mediocre
    - far too many suggestions lead to zero hit queries
    - and many words present in the index are not suggested at all

Metadata to the rescue!

While we don’t have enough user input to generate sensible word suggestions,

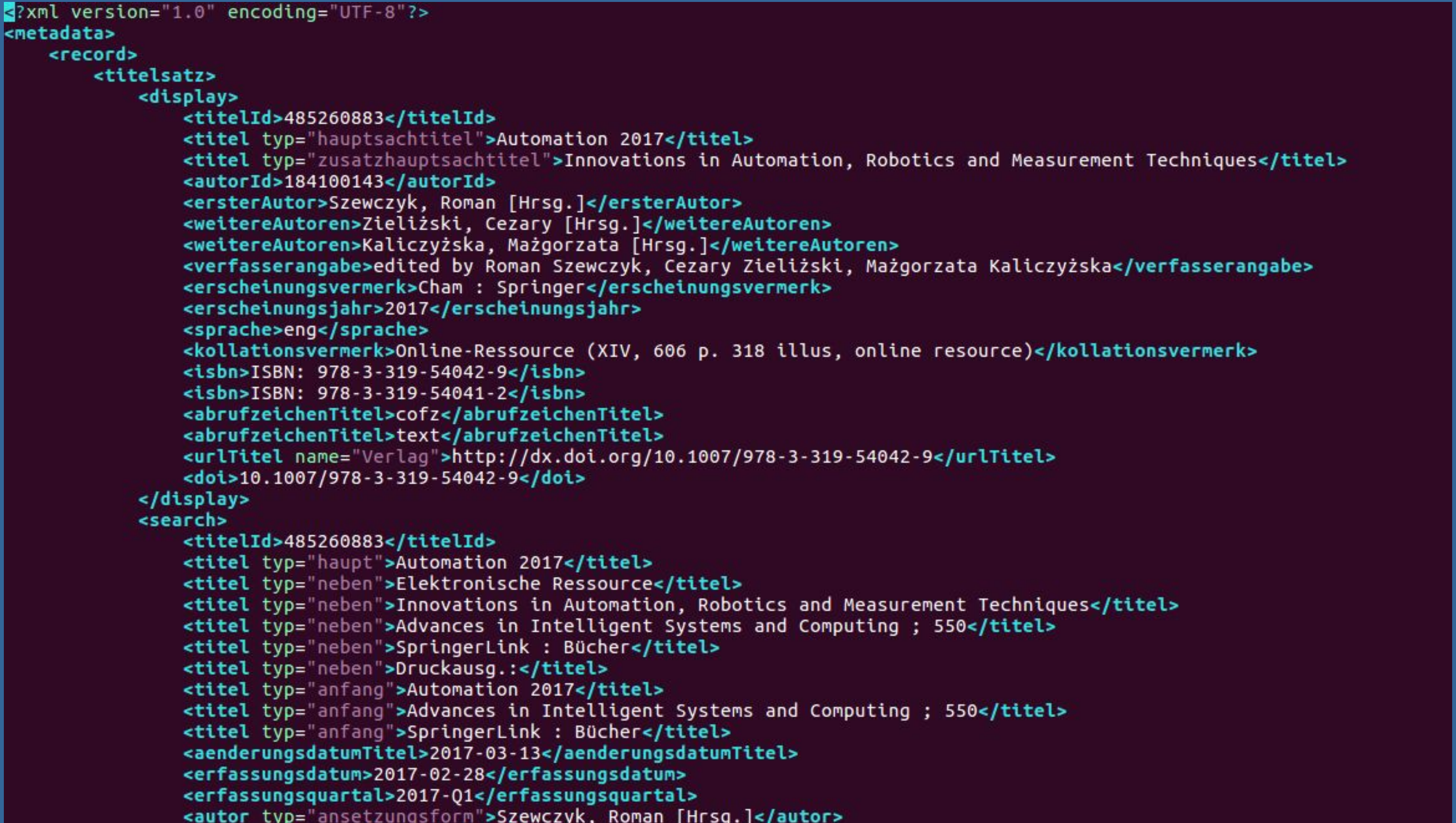
But we do have another source: our beloved metadata!

If it contains enough textual information to be used for search, can’t it be used for word completion and suggestion as well?



```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
    <record>
        <titelsatz>
            <display>
                <titelId>485260883</titelId>
                <titel typ="hauptsachtitel">Automation 2017</titel>
                <titel typ="zusatzhauptsachtitel">Innovations in Automation, Robotics and Measurement Techniques</titel>
                <autorId>184100143</autorId>
                <ersterAutor>Szewczyk, Roman [Hrsg.]</ersterAutor>
                <weitereAutoren>Zieliżski, Cezary [Hrsg.]</weitereAutoren>
                <weitereAutoren>Kaliczyżska, Mażgorzata [Hrsg.]</weitereAutoren>
                <verfasserangabe>edited by Roman Szewczyk, Cezary Zieliżski, Mażgorzata Kaliczyżska</verfasserangabe>
                <erscheinungsvermerk>Cham : Springer</erscheinungsvermerk>
                <erscheinungsjahr>2017</erscheinungsjahr>
                <sprache>eng</sprache>
                <kollationsvermerk>Online-Ressource (XIV, 606 p. 318 illus, online resource)</kollationsvermerk>
                <isbn>ISBN: 978-3-319-54042-9</isbn>
                <isbn>ISBN: 978-3-319-54041-2</isbn>
                <abrufzeichenTitel>cofz</abrufzeichenTitel>
                <abrufzeichenTitel>text</abrufzeichenTitel>
                <urlTitel name="Verlag">http://dx.doi.org/10.1007/978-3-319-54042-9</urlTitel>
                <doi>10.1007/978-3-319-54042-9</doi>
            </display>
            <search>
                <titelId>485260883</titelId>
                <titel typ="haupt">Automation 2017</titel>
                <titel typ="neben">Elektronische Ressource</titel>
                <titel typ="neben">Innovations in Automation, Robotics and Measurement Techniques</titel>
                <titel typ="neben">Advances in Intelligent Systems and Computing ; 550</titel>
                <titel typ="neben">SpringerLink : Bücher</titel>
                <titel typ="neben">Druckausg.:</titel>
                <titel typ="anfang">Automation 2017</titel>
                <titel typ="anfang">Advances in Intelligent Systems and Computing ; 550</titel>
                <titel typ="anfang">SpringerLink : Bücher</titel>
                <aenderungsdatumTitel>2017-03-13</aenderungsdatumTitel>
                <erfassungsdatum>2017-02-28</erfassungsdatum>
                <erfassungsquartal>2017-Q1</erfassungsquartal>
                <autor typ="ansetzungsform">Szewczyk, Roman [Hrsg.]</autor>
                <autor typ="idn">184100143</autor>
                <autor typ="ansetzungsform">Zieliżski, Cezary [Hrsg.]</autor>
                <autor typ="idn">39821140X</autor>
                <autor typ="ansetzungsform">Kaliczyżska, Mażgorzata [Hrsg.]</autor>
                <autor typ="idn">40344425X</autor>
                <autor typ="verfasserangabe">edited by Roman Szewczyk, Cezary Zieliżski, Mażgorzata Kaliczyżska</autor>
                <autorId>184100143</autorId>
                <autorId>39821140X</autorId>
                <autorId>40344425X</autorId>
                <ort>Cham</ort>
                <verlag>Springer</verlag>
                <erscheinungsjahr>2017</erscheinungsjahr>
                <sprache>eng</sprache>
                <fussnoteTitel>Druckausg.:</fussnoteTitel>
                <isbn>978-3-319-54042-9</isbn>
                <isbn>978-3-319-54041-2</isbn>
                <notationTitel>Q342</notationTitel>
                <abrufzeichenTitel>cofz</abrufzeichenTitel>
                <abrufzeichenTitel>text</abrufzeichenTitel>
                <produktsigel>ZDB-2-ENG</produktsigel>
```

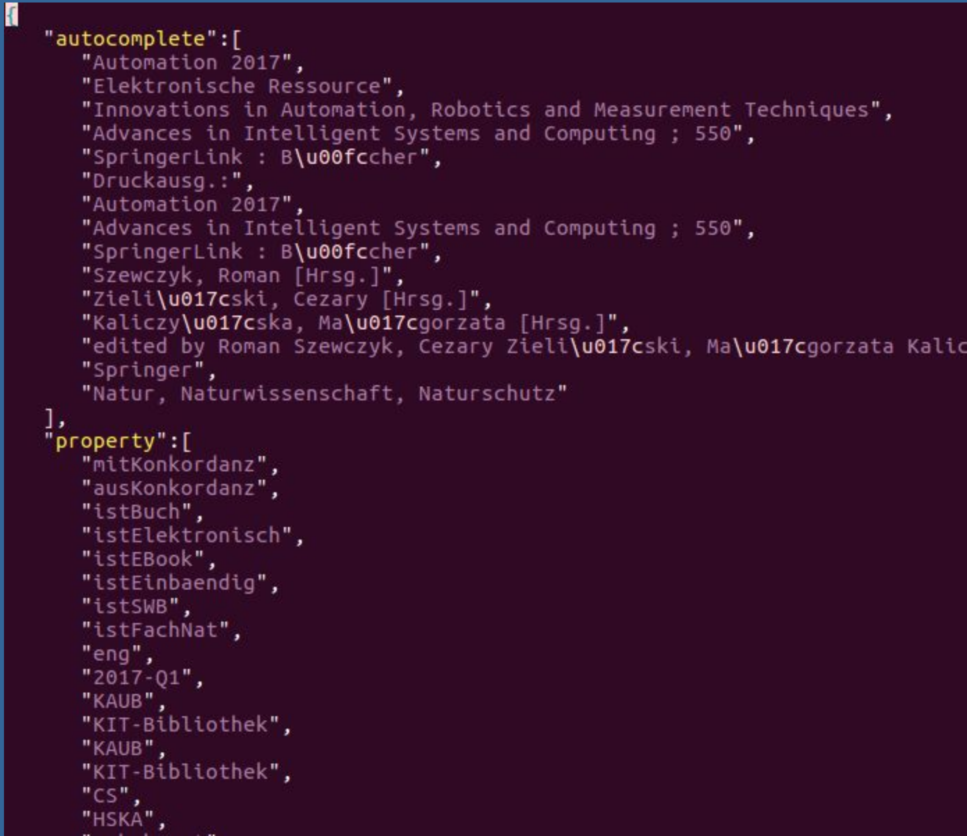
Metadata to the rescue!



```xml
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
    <record>
        <titelsatz>
            <display>
                <titelId>485260883</titelId>
                <titel typ="hauptsachtitel">Automation 2017</titel>
                <titel typ="zusatzhauptsachtitel">Innovations in Automation, Robotics and Measurement Techniques</titel>
                <autorId>184100143</autorId>
                <ersterAutor>Szewczyk, Roman [Hrsg.]</ersterAutor>
                <weitereAutoren>Zieliżski, Cezary [Hrsg.]</weitereAutoren>
                <weitereAutoren>Kaliczyżska, Mażgorzata [Hrsg.]</weitereAutoren>
                <verfasserangabe>edited by Roman Szewczyk, Cezary Zieliżski, Mażgorzata Kaliczyżska</verfasserangabe>
                <erscheinungsvermerk>Cham : Springer</erscheinungsvermerk>
                <erscheinungsjahr>2017</erscheinungsjahr>
                <sprache>eng</sprache>
                <kollationsvermerk>Online-Ressource (XIV, 606 p. 318 illus, online resource)</kollationsvermerk>
                <isbn>ISBN: 978-3-319-54042-9</isbn>
                <isbn>ISBN: 978-3-319-54041-2</isbn>
                <abrufzeichenTitel>cofz</abrufzeichenTitel>
                <abrufzeichenTitel>text</abrufzeichenTitel>
                <urlTitel name="Verlag">http://dx.doi.org/10.1007/978-3-319-54042-9</urlTitel>
                <doi>10.1007/978-3-319-54042-9</doi>
            </display>
            <search>
                <titelId>485260883</titelId>
                <titel typ="haupt">Automation 2017</titel>
                <titel typ="neben">Elektronische Ressource</titel>
                <titel typ="neben">Innovations in Automation, Robotics and Measurement Techniques</titel>
                <titel typ="neben">Advances in Intelligent Systems and Computing ; 550</titel>
                <titel typ="neben">SpringerLink : Bücher</titel>
                <titel typ="neben">Druckausg.:</titel>
                <titel typ="anfang">Automation 2017</titel>
                <titel typ="anfang">Advances in Intelligent Systems and Computing ; 550</titel>
                <titel typ="anfang">SpringerLink : Bücher</titel>
                <aenderungsdatumTitel>2017-03-13</aenderungsdatumTitel>
                <erfassungsdatum>2017-02-28</erfassungsdatum>
                <erfassungsquartal>2017-Q1</erfassungsquartal>
                <autor typ="ansetzungsform">Szewczyk, Roman [Hrsg.]</autor>
```

Main idea - deconstructing highly structured metadata

- In the end, we simply need a list of possible words
- So we throw all metadata fields that contain sensible words for suggestions into a single field
    - “autocomplete”
- And another field for internal usage such as faceting and scoping into another one
    - “property”



```
{
   "autocomplete":[
      "Automation 2017",
      "Elektronische Ressource",
      "Innovations in Automation, Robotics and Measurement Techniques",
      "Advances in Intelligent Systems and Computing ; 550",
      "SpringerLink : B\u00fccher",
      "Druckausg.:",
      "Automation 2017",
      "Advances in Intelligent Systems and Computing ; 550",
      "SpringerLink : B\u00fccher",
      "Szewczyk, Roman [Hrsg.]",
      "Zieli\u017cski, Cezary [Hrsg.]",
      "Kaliczy\u017cska, Ma\u017cgorzata [Hrsg.]",
      "edited by Roman Szewczyk, Cezary Zieli\u017cski, Ma\u017cgorzata Kalic
      "Springer",
      "Natur, Naturwissenschaft, Naturschutz"
   ],
   "property":[
      "mitKonkordanz",
      "ausKonkordanz",
      "istBuch",
      "istElektronisch",
      "istEBook",
      "istEinbaendig",
      "istSWB",
      "istFachNat",
      "eng",
      "2017-Q1",
      "KAUB",
      "KIT-Bibliothek",
      "KAUB",
      "KIT-Bibliothek",
      "CS",
      "HSKA",
```

Main idea - further details

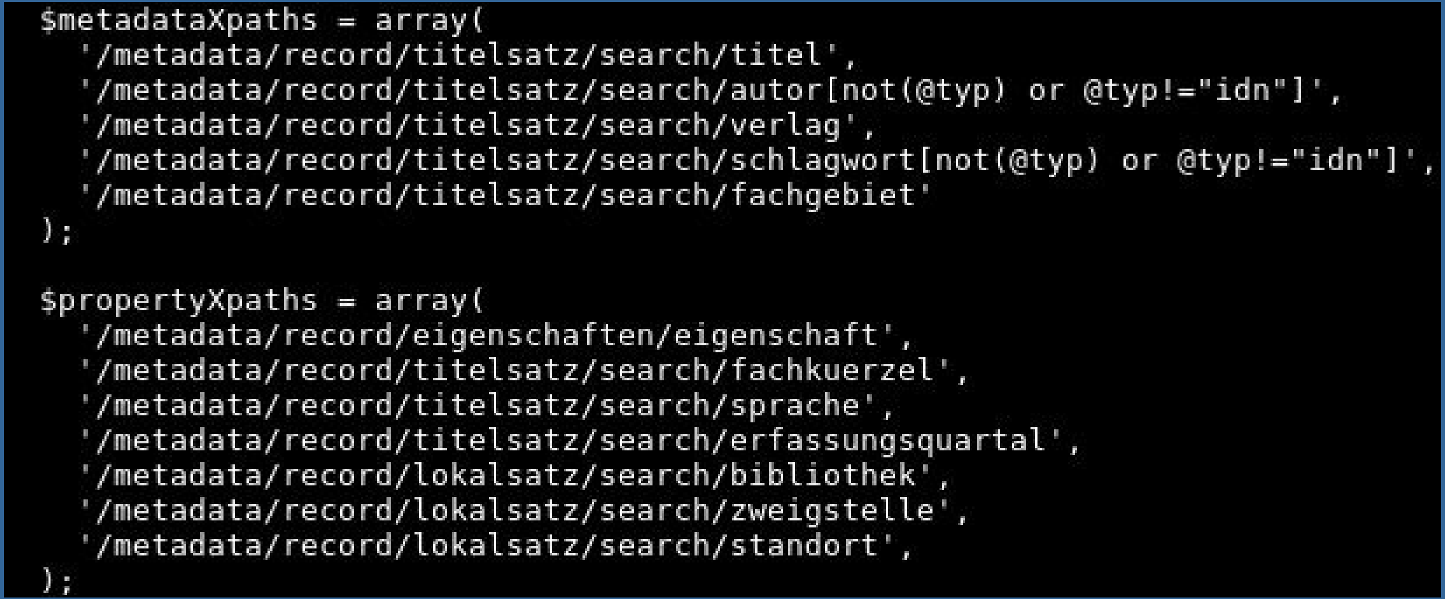
- Extract all words from our metadata
- Create an index with only 2 fields
    - autocomplete & property
    - the transformation is configurable using XPath expressions
- “autocomplete”-field
    - correct typing mistakes
    - create auto-completion for the current word
    - create auto-suggestion for next word
- “property”-field as a helper
    - scopes - restrict the results in all 3 use cases to relevant parts of the whole index (like a view)
    - facets - contains all internally used facet values

How to implement your own autocomplete feature?

- transform your data
  - use the language of your choice
- create an index using a search engine technology
  - Elasticsearch
- implement the autocompletion queries
  - JSON
- incept your user interface code
  - Javascript

Transforming XML to JSON

- xml2json.php
- Xpath to extract data



```
$metadataXpaths = array(
   '/metadata/record/titelsatz/search/titel',
   '/metadata/record/titelsatz/search/autor[not(@typ) or @typ!="idn"]',
   '/metadata/record/titelsatz/search/verlag',
   '/metadata/record/titelsatz/search/schlagwort[not(@typ) or @typ!="idn"]',
   '/metadata/record/titelsatz/search/fachgebiet'
);

$propertyXpaths = array(
   '/metadata/record/eigenschaften/eigenschaft',
   '/metadata/record/titelsatz/search/fachkuerzel',
   '/metadata/record/titelsatz/search/sprache',
   '/metadata/record/titelsatz/search/erfassungsquartal',
   '/metadata/record/lokalsatz/search/bibliothek',
   '/metadata/record/lokalsatz/search/zweigstelle',
   '/metadata/record/lokalsatz/search/standort',
);
```

Elasticsearch

- Lucene based full-text search
- Near real-time
- Schema-less
- Open source
- JSON over HTTP

Index configuration

- Use the Elasticsearch analyzer
  - Tokenize
  - Lowercase & more normalization
  - Remove stopwords
- analyze “autocomplete” field
- do not analyze “property” field

```
$ curl -s -XGET 'localhost:9200/autocomplete/_analyze' -d '
{
  "analyzer" : "autocompletion",
  "text" : "Innovations in Automation, Robotics and
Measurement Techniques"
}' | json_pp
{
   "tokens" : [
      {
         "token" : "innovations"
      },
      {
         "token" : "automation"
      },
      {
         "token" : "robotics"
      },
      {
         "token" : "measurement"
      },
      {
         "token" : "techniques"
      }
   ]
}
```

Implement the autocompletion queries with elasticsearch

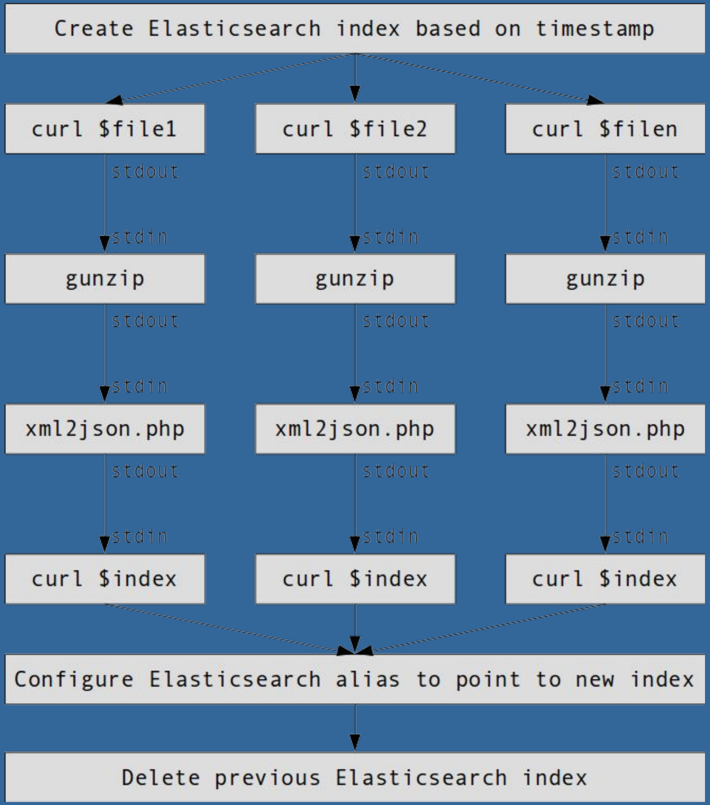
- Term suggester for spelling corrections
- Term aggregations to autocomplete current
- Term aggregations to suggest next word
- Filter aggregations queries to avoid zero-hit suggestions
- Exact match queries to limit to scopes and facets
- All tied together in a single query issued by a simple Javascript plugin

Staying up-to-date - automated ingest process

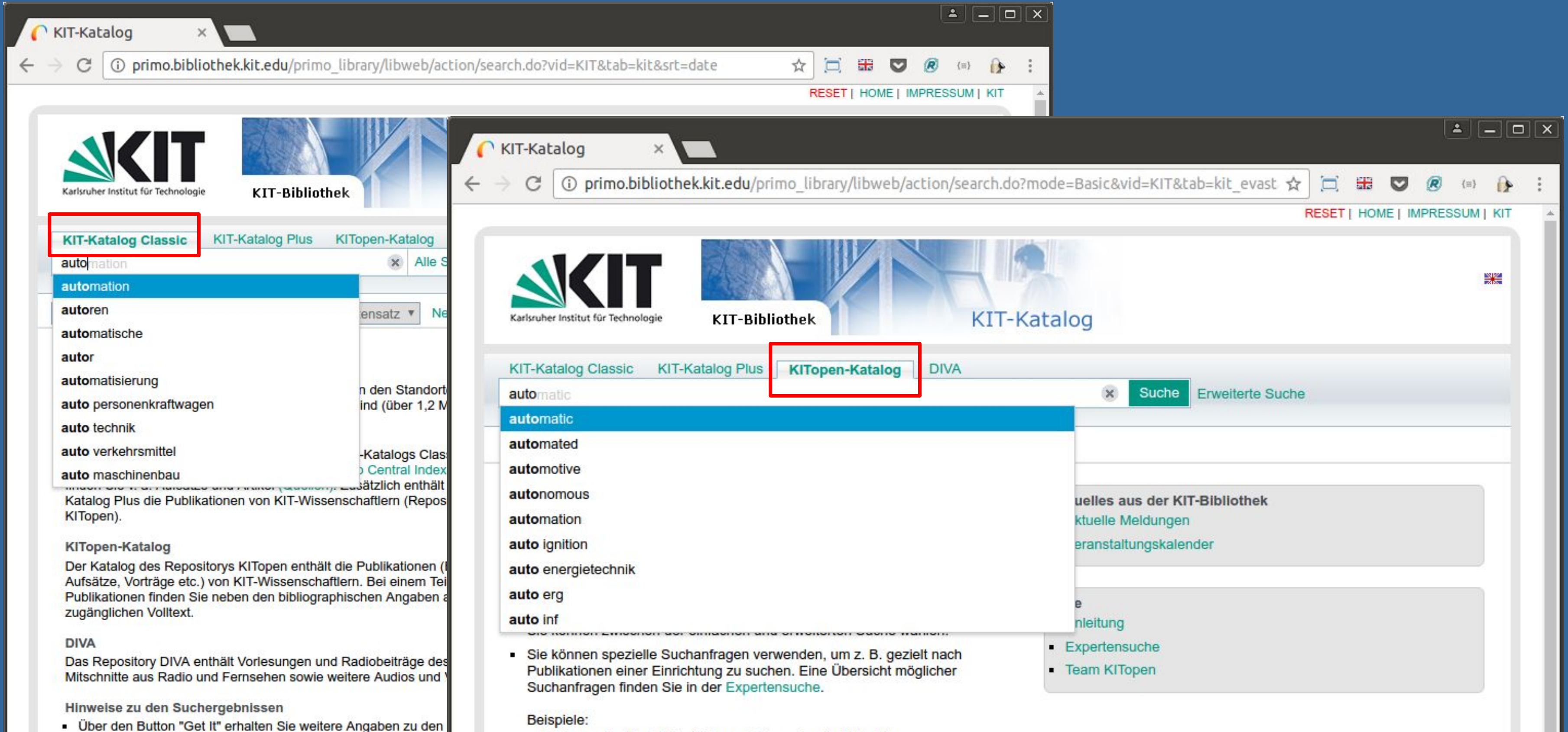
- RDS workflow has been adapted by IT team
- KIT XML for all records is built every weekend
- ingest script works on elasticsearch server
  - harvests the KIT XML from RDS server via HTTP
  - transforms XML into JSON
  - runs for both Primo servers (test & production)

Stream processing pipeline

- Unzip compressed XML metadata to STDOUT
- Stream process XML from STDIN using XPaths, map XML to two JSON fields in PHP:
    - autocomplete
    - property
- Output JSON format suitable for Elasticsearch Bulk API to STDOUT
- Pipe directly via cURL to Elasticsearch Index
- 1.5 Mio records indexed in ~8min on a virtual Dual Core with 8 GB RAM running both the conversion and Elasticsearch



Automation at your fingertip - scopes



auto|mation

n den Standort
ind (über 1,2 M

-Katalogs Clas
Central Index
finden Sie v. a. Aufsätze und Artikel (Quellen). Zusätzlich enthält Katalog Plus die Publikationen von KIT-Wissenschaftlern (Repos KITopen).

KITopen-Katalog

Der Katalog des Repositorys KITopen enthält die Publikationen (Aufsätze, Vorträge etc.) von KIT-Wissenschaftlern. Bei einem Tei Publikationen finden Sie neben den bibliographischen Angaben a zugänglichen Volltext.

DIVA

Das Repository DIVA enthält Vorlesungen und Radiobeiträge des Mitschnitte aus Radio und Fernsehen sowie weitere Audios und

Hinweise zu den Suchergebnissen

- Über den Button "Get It" erhalten Sie weitere Angaben zu den

auto|matic

uelles aus der KIT-Bibliothek
ktuelle Meldungen
eranstaltungskalender

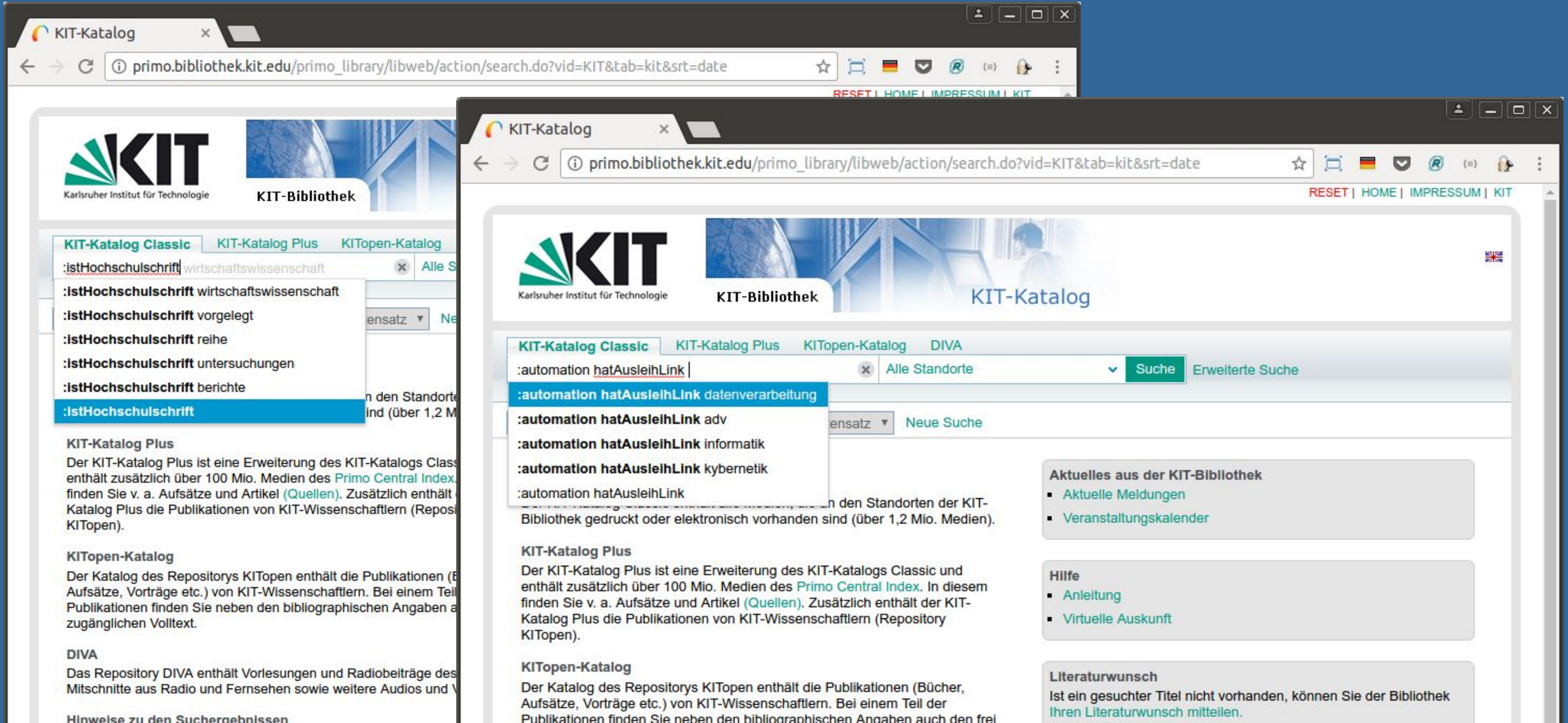
e
nleitung
- Expertensuche
- Team KITopen

Sie können zwischen der einfachen und erweiterten Suche wählen.

- Sie können spezielle Suchanfragen verwenden, um z. B. gezielt nach Publikationen einer Einrichtung zu suchen. Eine Übersicht möglicher Suchanfragen finden Sie in der Expertensuche.

Beispiele:

Poweruser mode with “:” as prefix



KIT-Katalog

primo.bibliothek.kit.edu/primo_library/libweb/action/search.do?vid=KIT&tab=kit&srt=date

RESET | HOME | IMPRESSUM | KIT

Karlsruher Institut für Technologie

KIT-Bibliothek

KIT-Katalog Classic | KIT-Katalog Plus | KITopen-Katalog

:istHochschulschrift wirtschaftswissenschaft

- **:istHochschulschrift** wirtschaftswissenschaft
- **:istHochschulschrift** vorgelegt
- **:istHochschulschrift** reihe
- **:istHochschulschrift** untersuchungen
- **:istHochschulschrift** berichte
- **:istHochschulschrift**

n den Standorte
ind (über 1,2 M

KIT-Katalog Plus

Der KIT-Katalog Plus ist eine Erweiterung des KIT-Katalogs Class
enthält zusätzlich über 100 Mio. Medien des Primo Central Index.
finden Sie v. a. Aufsätze und Artikel (Quellen). Zusätzlich enthält
Katalog Plus die Publikationen von KIT-Wissenschaftlern (Reposi
KITopen).

KITopen-Katalog

Der Katalog des Repositorys KITopen enthält die Publikationen (E
Aufsätze, Vorträge etc.) von KIT-Wissenschaftlern. Bei einem Tei
Publikationen finden Sie neben den bibliographischen Angaben a
zugänglichen Volltext.

DIVA

Das Repository DIVA enthält Vorlesungen und Radiobeiträge des
Mitschnitte aus Radio und Fernsehen sowie weitere Audios und V

Hinweise zu den Suchergebnissen

KIT-Katalog

primo.bibliothek.kit.edu/primo_library/libweb/action/search.do?vid=KIT&tab=kit&srt=date

RESET | HOME | IMPRESSUM | KIT

Karlsruher Institut für Technologie

KIT-Bibliothek

KIT-Katalog

KIT-Katalog Classic | KIT-Katalog Plus | KITopen-Katalog | DIVA

:automation hatAusleihLink

Alle Standorte | Suche | Erweiterte Suche

- **:automation hatAusleihLink** datenverarbeitung
- **:automation hatAusleihLink** adv
- **:automation hatAusleihLink** informatik
- **:automation hatAusleihLink** kybernetik
- :automation hatAusleihLink

ensatz ▼ Neue Suche

n den Standorten der KIT-Bibliothek gedruckt oder elektronisch vorhanden sind (über 1,2 Mio. Medien).

KIT-Katalog Plus

Der KIT-Katalog Plus ist eine Erweiterung des KIT-Katalogs Classic und enthält zusätzlich über 100 Mio. Medien des Primo Central Index. In diesem finden Sie v. a. Aufsätze und Artikel (Quellen). Zusätzlich enthält der KIT-Katalog Plus die Publikationen von KIT-Wissenschaftlern (Repository KITopen).

KITopen-Katalog

Der Katalog des Repositorys KITopen enthält die Publikationen (Bücher, Aufsätze, Vorträge etc.) von KIT-Wissenschaftlern. Bei einem Teil der Publikationen finden Sie neben den bibliographischen Angaben auch den frei

Aktuelles aus der KIT-Bibliothek

- Aktuelle Meldungen
- Veranstaltungskalender

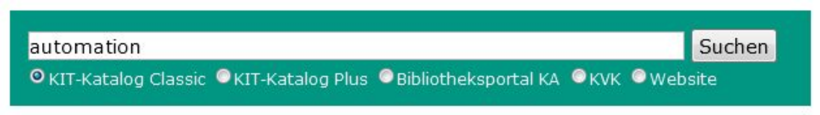
Hilfe

- Anleitung
- Virtuelle Auskunft

Literaturwunsch

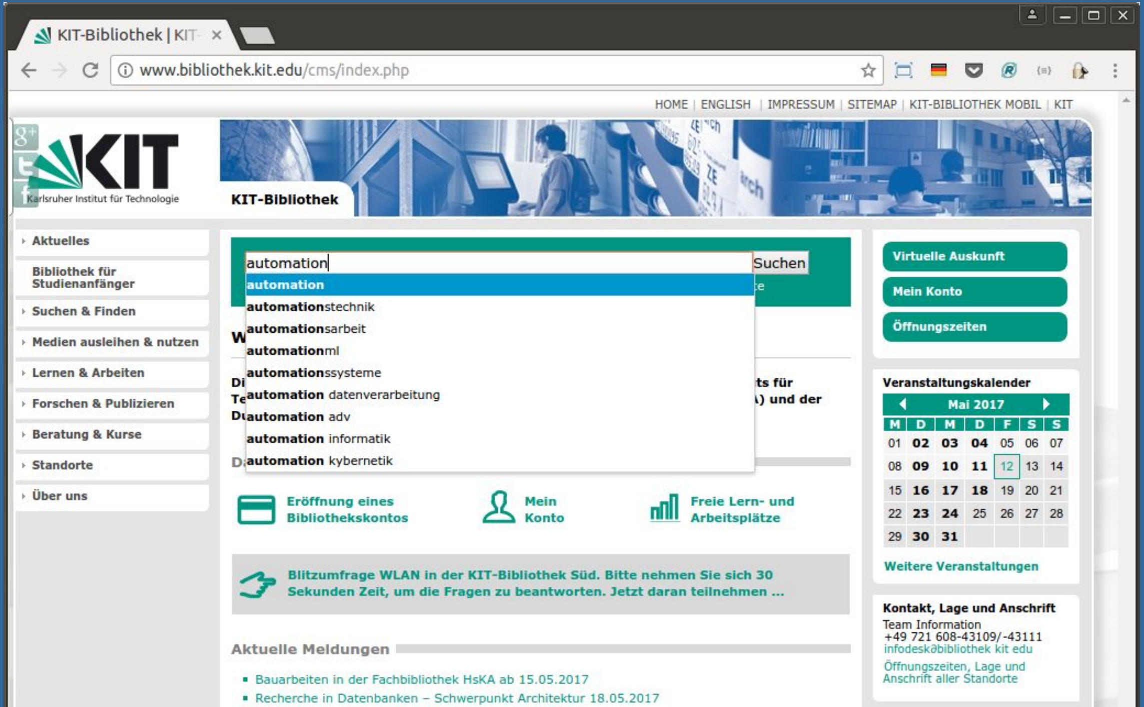
Ist ein gesuchter Titel nicht vorhanden, können Sie der Bibliothek Ihren Literaturwunsch mitteilen.

Beyond Primo - autocompletion everywhere



automation Suchen

● KIT-Katalog Classic ○ KIT-Katalog Plus ○ Bibliotheksportal KA ○ KVK ○ Website



KIT-Bibliothek | KIT-

www.bibliothek.kit.edu/cms/index.php

HOME | ENGLISH | IMPRESSUM | SITEMAP | KIT-BIBLIOTHEK MOBIL | KIT

KIT Karlsruher Institut für Technologie

KIT-Bibliothek

- Aktuelles
- Bibliothek für Studienanfänger
- Suchen & Finden
- Medien ausleihen & nutzen
- Lernen & Arbeiten
- Forschen & Publizieren
- Beratung & Kurse
- Standorte
- Über uns

automation Suchen

automation
automationstechnik
automationsarbeit
automationml
automationssysteme
automation datenverarbeitung
automation adv
automation informatik
automation kybernetik

ts für
) und der

Eröffnung eines Bibliothekskontos

Mein Konto

Freie Lern- und Arbeitsplätze

Blitzumfrage WLAN in der KIT-Bibliothek Süd. Bitte nehmen Sie sich 30 Sekunden Zeit, um die Fragen zu beantworten. Jetzt daran teilnehmen ...

Aktuelle Meldungen

- Bauarbeiten in der Fachbibliothek HsKA ab 15.05.2017
- Recherche in Datenbanken – Schwerpunkt Architektur 18.05.2017

Virtuelle Auskunft

Mein Konto

Öffnungszeiten

Veranstaltungskalender

Mai 2017

M	D	M	D	F	S	S
01	02	03	04	05	06	07
08	09	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Weitere Veranstaltungen

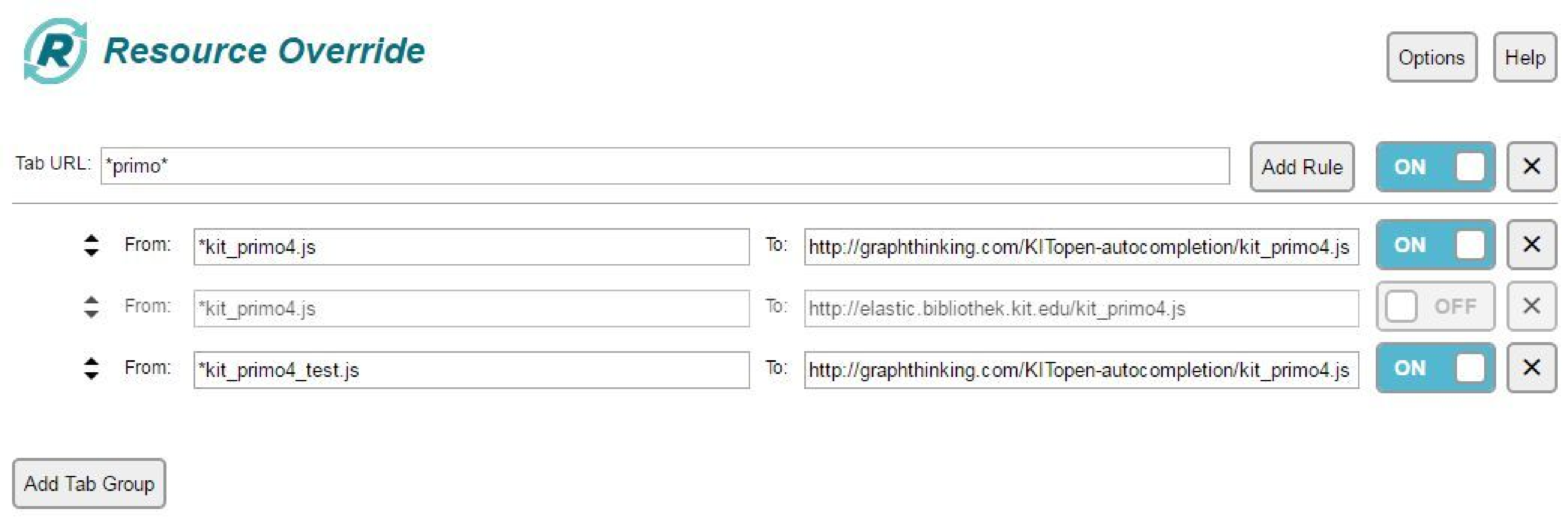
Kontakt, Lage und Anschrift

Team Information
+49 721 608-43109/-43111
infodesk∂bibliothek kit edu
Öffnungszeiten, Lage und Anschrift aller Standorte

How to manage a remote project?

- Distance between Karlsruhe and Berlin are in about 6 hours by train
- Which toolset did we use?
  - telephone and teamviewer
    - to communicate and see what you are talking about
  - github
    - to manage source code and issues
  - work in Berlin as if you were located in Karlsruhe
    - use tunneling (SOCKS)
  - incept the live Primo service running in Karlsruhe from Berlin
    - "Resource Override" plugin for Google Chrome

Resource Override for Google Chrome



Tab URL: *primo* | Add Rule | ON | ✕

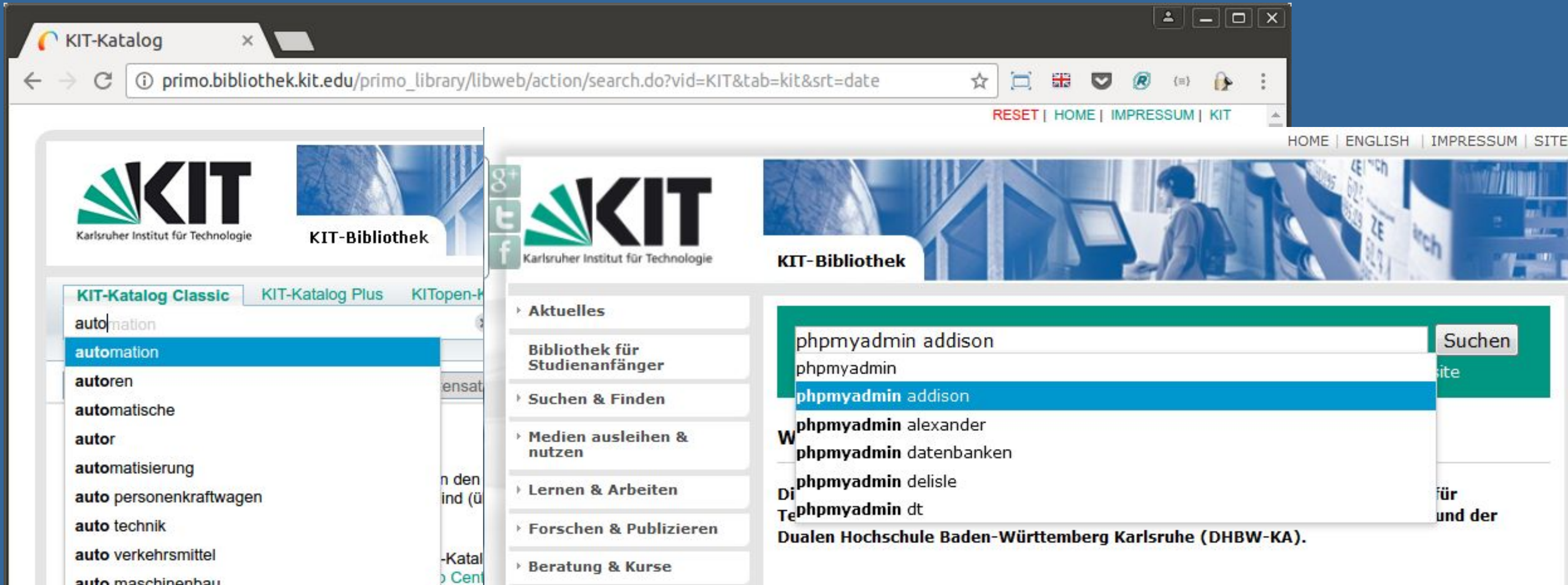
From: *kit_primo4.js To: http://graphthinking.com/KITopen-autocompletion/kit_primo4.js ON ✕

From: *kit_primo4.js To: http://elastic.bibliothek.kit.edu/kit_primo4.js OFF ✕

From: *kit_primo4_test.js To: http://graphthinking.com/KITopen-autocompletion/kit_primo4.js ON ✕

Results

- autocompletion within Primo
- autocompletion on every page of the KIT Library Website having a searchslot



primo.bibliothek.kit.edu/primo_library/libweb/action/search.do?vid=KIT&tab=kit&srt=date

auto|mation

phpmyadmin addison

More Results

- suggestions depend on the "context"
  - works within the live production system but also in the testing environment
  - Primo scopes are supported
    - location specific search restrictions (KIT, HSKA, DHBW)
    - data specific search restrictions (KITopen repository)
  - many Primo facettes are supported
- advanced mode for power users offers more suggestions
- solution is adaptable to other environments
  - blogs, CMS, database

Questions?

Here and now or any time to
uwe.dierolf@kit.edu
and
felix.ostrowski@graphthinking.com