STAATS- UND UNIVERSITÄTS-
BIBLIOTHEK
HAMBURG
CARL VON OSSIETZKY

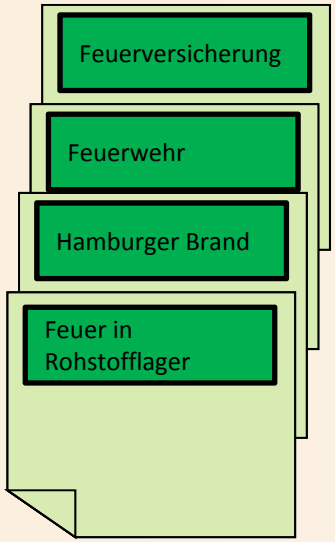
Die Gestaltung der Suche in heterogenen Datenbeständen – Optionen und Herausforderungen. Ein Werkstattbericht

Von syntaktischer zur semantischen Suche

Dr. Lothar Hotz, Arved Solth, HITeC e.V., Universität Hamburg

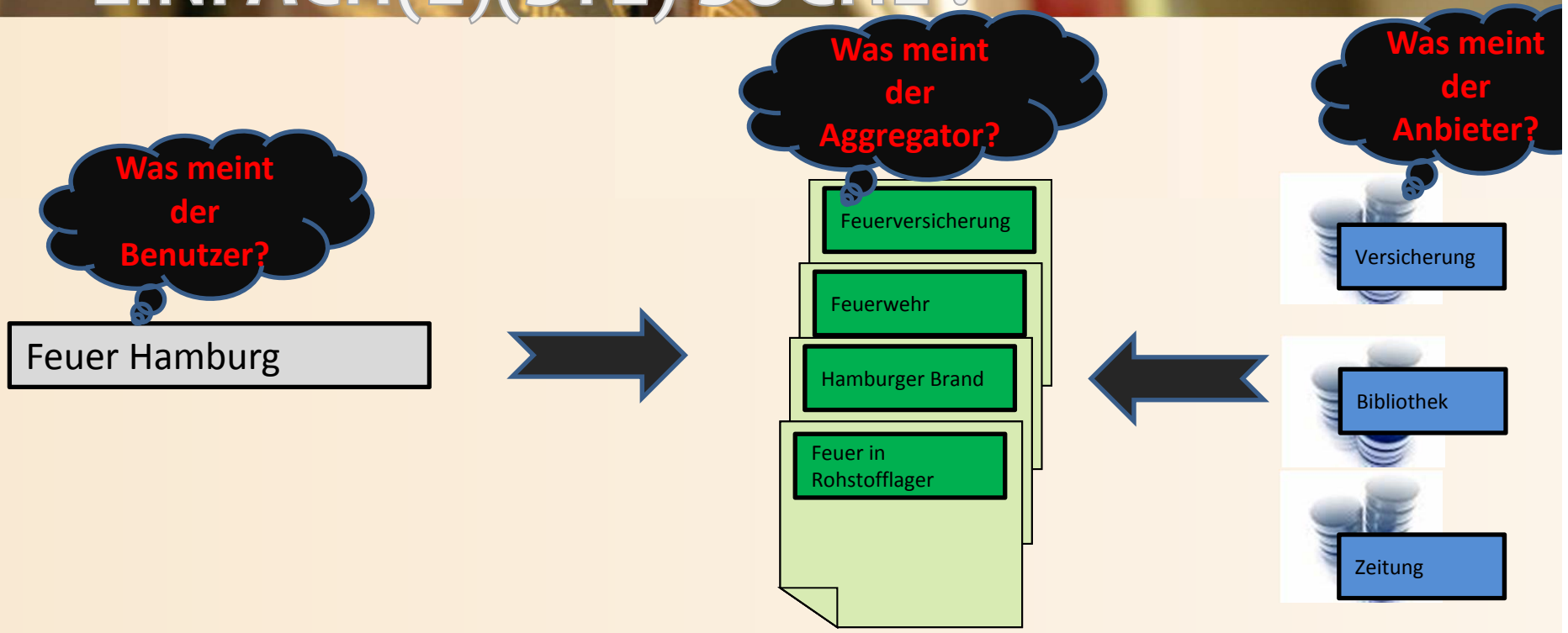
EINFACHE SUCHE

Feuer Hamburg



STAATS- UND UNIVERSITÄTS-
BIBLIOTHEK
HAMBURG
CARL VON OSSIETZKY

EINFACH(E)(STE) SUCHE ?

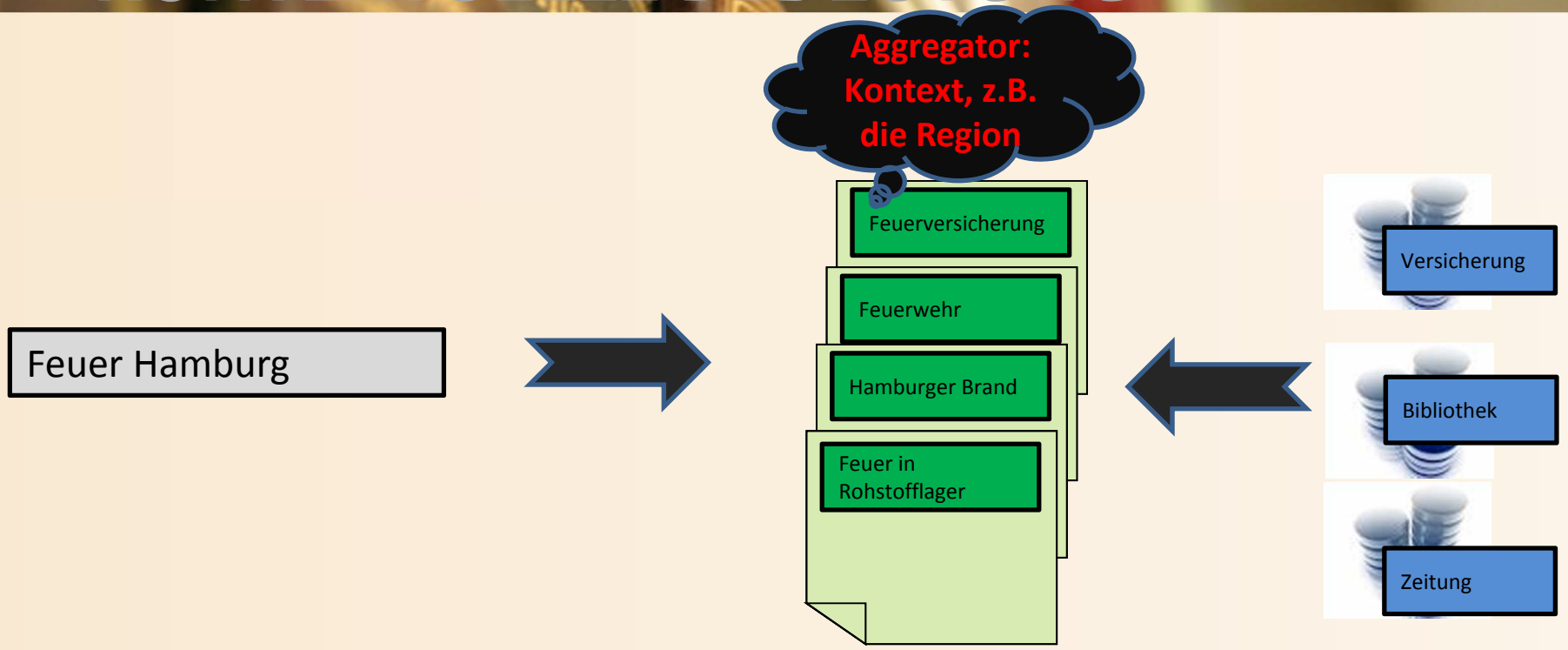


STAATS- UND UNIVERSITÄTS-
BIBLIOTHEK
HAMBURG
CARL VON OSSIETZKY

DIMENSIONEN

- Aufbereitung: Regionbezogene Filter
- Sucheingabe: Ausdruck der vom Benutzer intendierten Bedeutung
- Die Suche: Auswertung der Bedeutung
- Suchergebnis: Darstellung und Erklärung der verwendeten Bedeutung
- Verwertung: Ziel des Benutzers

KONTEXTUELLE BEDEUTUNG



Kontextuelle Bedeutung

- Kontext: **regionaler Bezug zu Hamburg**
- Auswahl entsprechender Ressourcen-Anbieter
  - ✓ **20 Anbieter und 64 Ressourcen**
- Festlegung eines Metadaten-Schemas
  - ✓ **13 Suchfelder und 26 interne Felder**
  - ❖ Standard-Metadaten-Schema
- Abbildung der Anbieter-Semantik auf Aggregator-Semantik
  - ✓ **Manuell pro Ressource**
  - ❖ Ontology-Mapping, Ontology-Alignment

Kontextuelle Bedeutung

- Abbildung und Indexierung der Informationsobjekte
  - ✓ **Automatische Metadatenabbildung**
  - ❖ Semantisches Parsing: Erkennen der kontextuellen Metadaten in Informationsobjekten

Verarbeitung von XML-Formaten

Metadaten im Original-Format (z.B. LIDO)

```
...
<lido:titleSet>
<lido:appellationValue >
Sonntagsstimmung auf der
Elbe. St. Pauli
Landungsbrücken
</lido:appellationValue>
</lido:titleSet>
</lido:titleWrap>
...
```

XSLT-Regeln

```
...
<xsl:template
match="lido:titleSet/lido:appellationValue">
    <field name="title">
      <xsl:value-of select="." />
    </field>
</xsl:template>
...
```

Metadaten im HWD-Format

```
<doc>
...
<field name="title">
Sonntagsstimmung auf der
Elbe. St. Pauli
Landungsbrücken
</field>
...
</doc>
```

Übersicht verwendeter Parser

Format	Parser/Bibliothek	Beispiel-Ressource
XML-Formate	Saxon XSL-Transformationen	Archiv des Hamburger Instituts für Sozialforschung
Webseiten (HTML)	BeautifulSoup	Beständeübersicht des Landesarchivs Schleswig-Holstein
Adobe PDF	Apache Tika	Statistische Berichte des Statistamts Nord
Comma Separated Values (CSV)	Python-CSV-Library	Stolpersteine-DB der Landeszentrale für politische Bildung

Kontextuelle Bedeutung

- Abbildung und Indexierung der Informationsobjekte
  - ✓ **Automatische Metadatenabbildung**
  - ❖ Semantisches Parsing: Erkennen der gewählten Metadaten in Informationsobjekten

- Filterung der Datenbestände
  - (Manuelle Auswahl der Informationsobjekte)
  - ✓ Modellierung der Auswahl mittels eines Thesaurus/Ontologie
    - ✓ **7485587 Hamburgensien**
  - ❖ Automatisches Lernen durch Angabe von positiven und negativen Beispielen

HWD-Indexsuche: Hamburg-Filter

Vorläufige Trefferliste:

Erich Hartmann (1886-1974). Leben und Werk eines Hamburger Malers. Mit einem Verzeichnis der Gemälde und der "Kunst am Bau"

Einfluß einer Behandlung mit dem Steroid-Implantat Synovex-S auf das natürliche Steroidhormonmuster des Muskel- und Fettgewebes von Ochsen

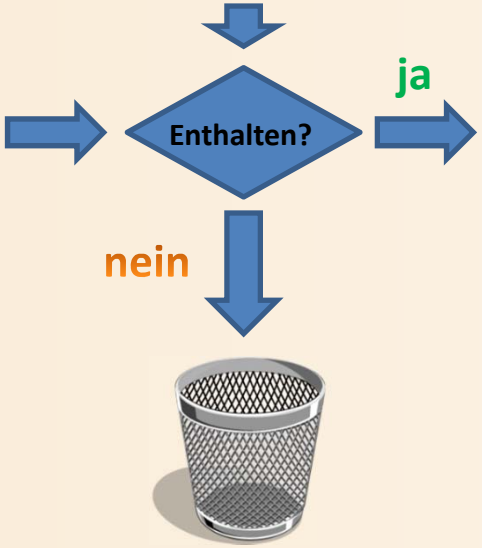
Hamburger Mietenspiegel : Texte, Erläuterungen, Arbeitshinweise / von Friedemann Sternel. XII, 96 S.

Analyse der Hamburger Rauschgifttodesfälle im Zeitraum 2004-2006 auf Hepatitis B, C, HIV und pulmonale FremdkörpergranulomeAnalysis of drug-related deaths from 2004 to 2006 for hepatitis B, C, HIV and pulmo...

Chemoenzymatische Synthese sialylierter Oligosaccharidstrukturen durch Transglycosylierung

Hamburg-Filter:

Normdaten-Wortgut (SWD, GKD, HHBiB-SW)



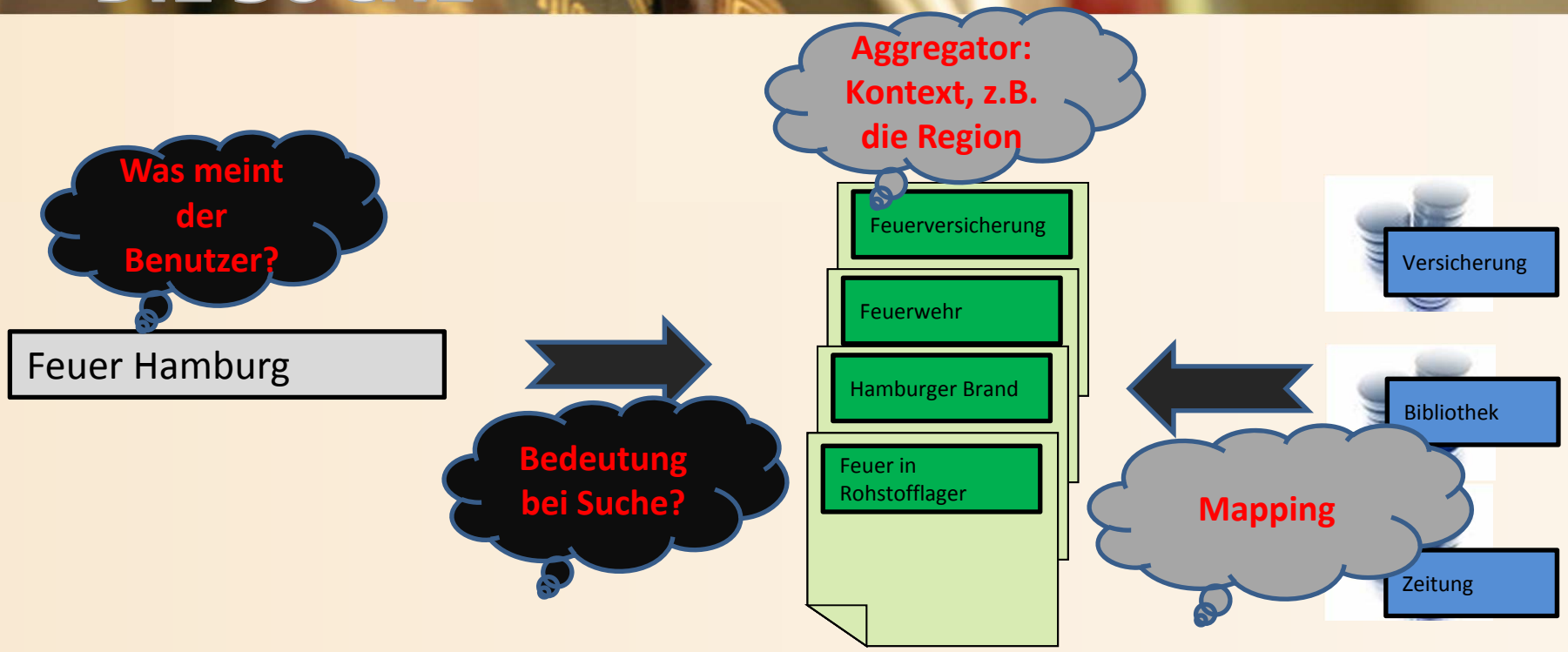
Endgültige Trefferliste:

Funktionsweise des Filters

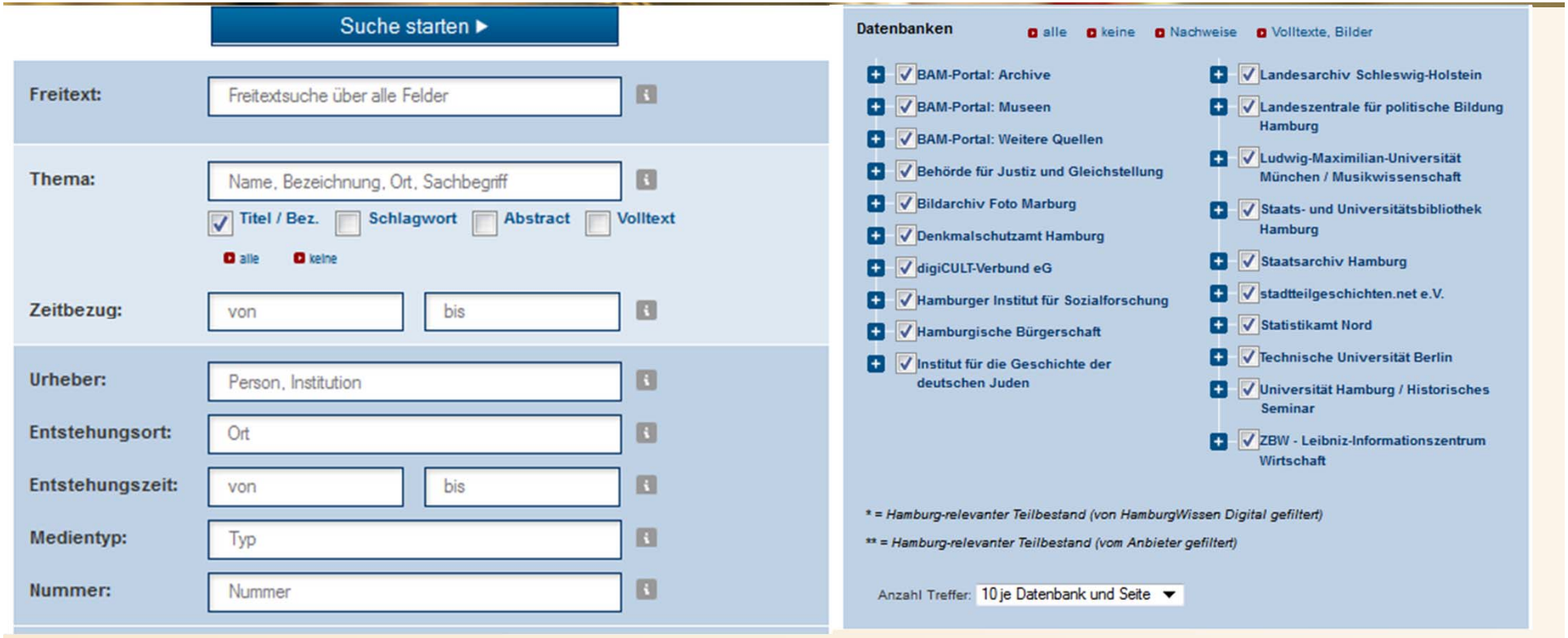
- ✓ Basis: Normdaten-Wortgut aus Schlagwortdatei (SWD), HHBIB-Schlagworte (HHBIB-SW), Gemeinsame Köperschaftsdatei (GKD), Lokalisatoren
- ✓ Kombination von ausgewählten Schlagworten mit Lokalisatoren
- ✓ Aktuell ca 7 Millionen Begriffe
- ✓ Anfrage an SOLR-Server: Hole alle Dokument, die x Hamburgensien enthalten
- ✓ Berücksichtigung eines Wortabstands bzgl. einer Hamburgensie
- ✓ Solr-Suche auf ausgewählten Feldern (Titel, Abstract, Volltext, etc.)
- ❖ Blackliste mit eindeutigen Nicht-Hamburgensien
- ✓ Laufzeit 6 bis 10 Stunden

params={fl=identifier,start=0,q=ressource:hans AND _query_:"{!edismax qf='title text author content_biography_text content_person content_place content_standard_keyword content_description content_time content_keyword ' v=$qq}",wt=javabin,qq="Hamburg Bergstedt Michael Hering Erben"~10 OR "Rahlstedt Michael Hering Erben"~10 OR "Volksdorf Michael Hering Erben"~10 OR "Barmbek German Committee for Marine Sciences and Technology"~10 OR "Hamburg Goldbeck Janssen Horst Brockstedt Hans"~10 OR "Goldbeck Hamburg Janssen Horst Brockstedt Hans"~10 OR "Treudelberg Steber Jörg"~10 OR "Mooralster Steber Jörg"~10 OR "Susebeck Steber Jörg"~10 OR "Susebek Steber Jörg"~10 OR "Alsterwanderweg Steber Jörg"~10 OR "Wandse Steber Jörg"~10 OR "Hamburg-Goldbeck Steber Jörg"~10 OR "Hamburg Goldbeck Steber Jörg"~10 OR "Goldbeck Hamburg Steber Jörg"~10 OR "Krautsand Weichmann Christian Friedrich"~10 OR "Schweinesand Weichmann Christian Friedrich"~10 OR "Pagensand Weichmann Christian Friedrich"~10 OR "Rodenbeker Quellental Weichmann Christian Friedrich"~10 OR "Klein-Flottbeck Weichmann Christian Friedrich"~10 OR "Klein-Flottbek Weichmann Christian Friedrich"~10 OR "Walddörfer Weichmann Christian Friedrich"~10 OR "Hanskalbsand Weichmann Christian Friedrich"~10 OR "Hamburg Sülldorf Imhof Hieronymus"~10 OR "Sülldorf Hamburg Imhof Hieronymus"~10 OR "Bergstedt Hamburg Musikkorps Deutschland 6"~10 OR "Bergstedt-Hamburg Musikkorps Deutschland 6"~10 OR "Hamburg-Steinbek Unbehagen Johann Andreas Christoph"~10 OR "Steinbeck-Hamburg Unbehagen Johann Andreas Christoph"~10 OR "Steinbek Hamburg Unbehagen Johann Andreas Christoph"~10 OR "Hamburg Steinbek Unbehagen Johann Andreas Christoph"~10 OR "Scharhörn Unbehagen Johann Andreas Christoph"~10 OR "Steinbek-Hamburg Unbehagen Johann Andreas Christoph"~10 OR "Schiffbek Unbehagen Johann Andreas Christoph"~10 OR "Schiffbeck Unbehagen Johann Andreas Christoph"~10 OR "Hamburg-Hoheluft Konzentrationslager Lerbeck"~10 OR "Hamburg Hoheluft Konzentrationslager Lerbeck"~10 OR "Steinbeck Hamburg Unbehagen Johann Andreas Christoph"~10 OR "Hamburg Steinbeck Unbehagen Johann Andreas Christoph"~10 OR "Fuhlsbüttel Praetzel Karl Gottlieb"~10 OR "Langenhorn Praetzel Karl Gottlieb"~10 OR "Wilhelmsburg-Hamburg Dornemann Henrich"~10 OR "Veddel Dornemann Henrich"~10 OR "Barmbek Praetzel Karl Gottlieb"~10 OR "Barmbeck Praetzel Karl Gottlieb"~10 OR "Barmbek-Süd Praetzel Karl Gottlieb"~10 OR "Barmbeck-Süd Praetzel Karl Gottlieb"~10 OR "Barmbeck-Nord Praetzel Karl Gottlieb"~10 OR "Billstedt Dornemann Henrich"~10 OR "Dulsberg Praetzel Karl Gottlieb"~10,version=2,rows=20000}

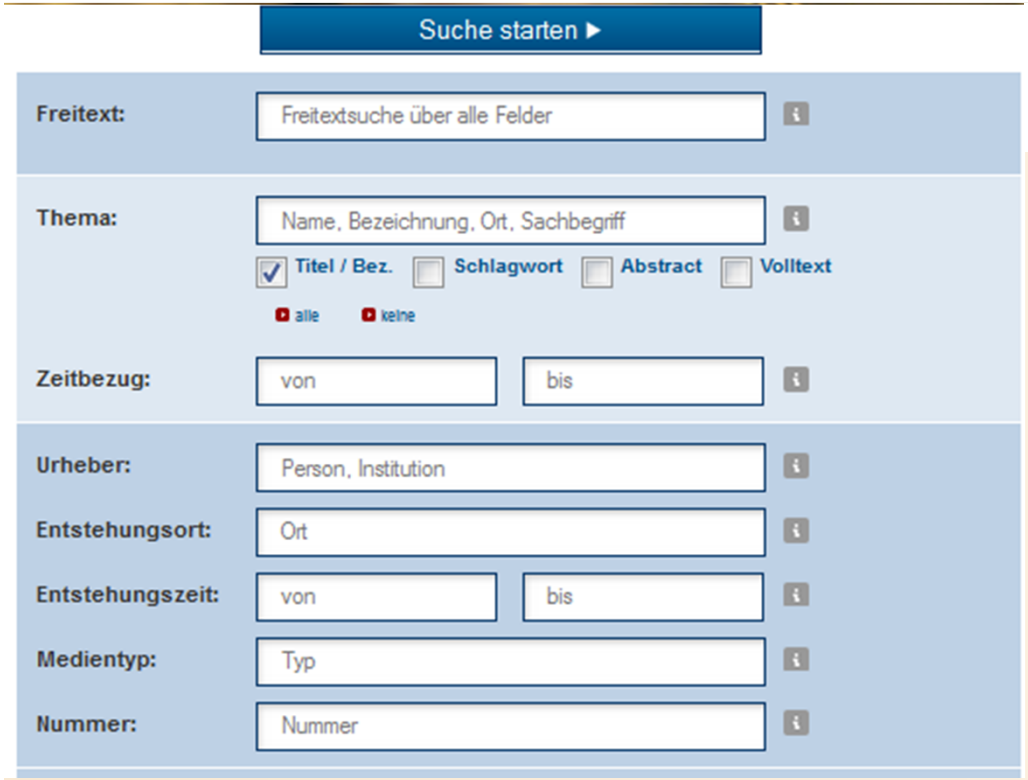
DIE SUCHE


Was meint der Benutzer?
Feuer Hamburg
Bedeutung bei Suche?
Aggregator: Kontext, z.B. die Region
Feuerversicherung
Feuerwehr
Hamburger Brand
Feuer in Rohstofflager
Mapping
Versicherung
Bibliothek
Zeitung

HWD-Metasuche – erweiterte Suchmaske

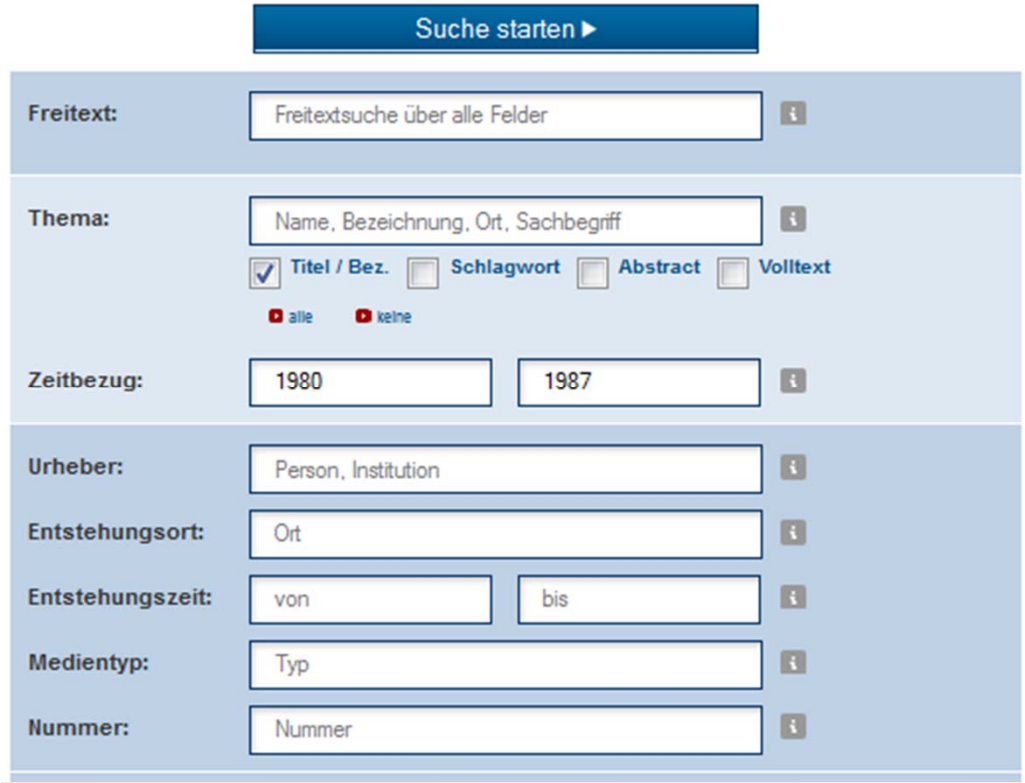


HWD-Metasuche – erweiterte Suchmaske



Ausdruckshilfe für die Darstellung der vom Benutzer intendierten Bedeutung

Felder mit Zeitbezug



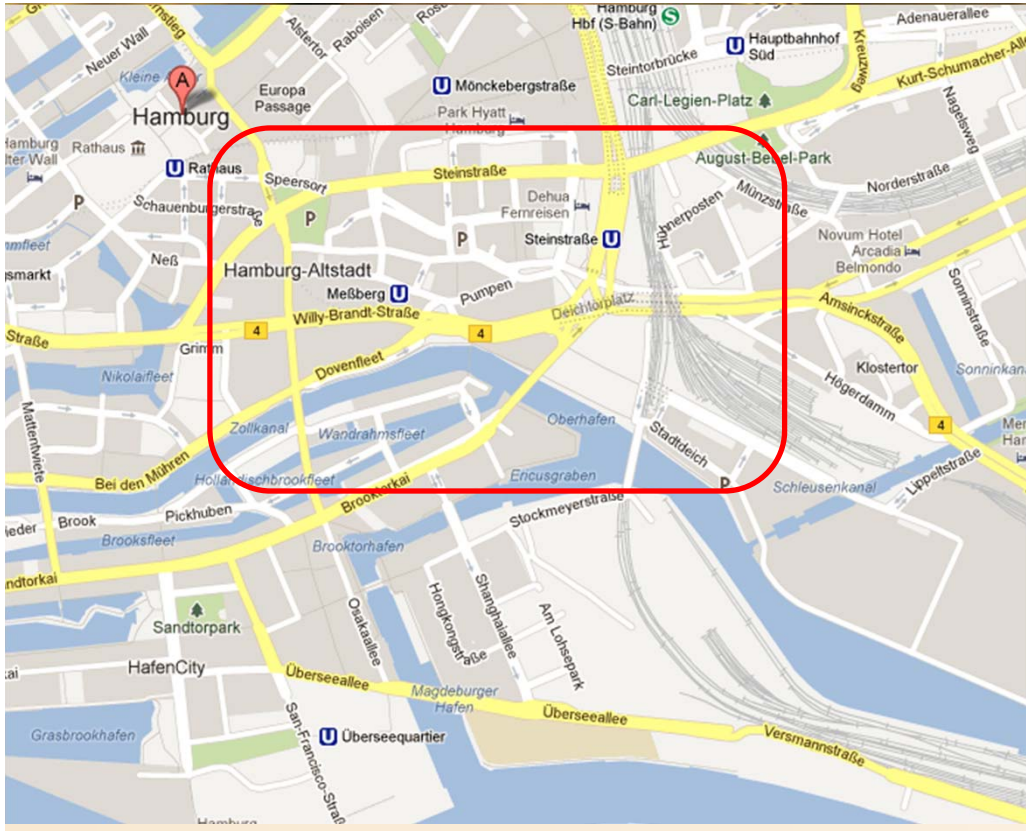
Eingabe eines Zeitbereichs: e_z
Zeitbereich des Informationsobjekts: i_z

Mögliche Interpretationen:
$i_z = e_z$: *gleiche* Zeitbereiche
$i_z \sim e_z$: *ungefähr* gleiche Zeitbereiche
$i_z \subset e_z$: Eingabebereich umfasst Informationsobjektsbereich
$e_z \subset i_z$: Eingabebereich innerhalb Informationsobjektsbereich (nur direkt umliegend(!))
$e_z \cap i_z$: Eingabebereich hat Schnitt mit Informationsobjektsbereich

Felder mit Zeitbezug

- Aufbereitung:
  - ✓ **Metadatenfelder: creation_time_from, creation_time_to, content_time_from, content_time_to**
  - ❖ Abbildung von zeitbezogenen Begriffen auf Zeitbereich
- Suche:
  - ✓ **Intervallarithmetik**
- Darstellung:
  - ❖ Zeit des Informationsobjekts mit “Im Bereich”

Felder mit Ortsbezug



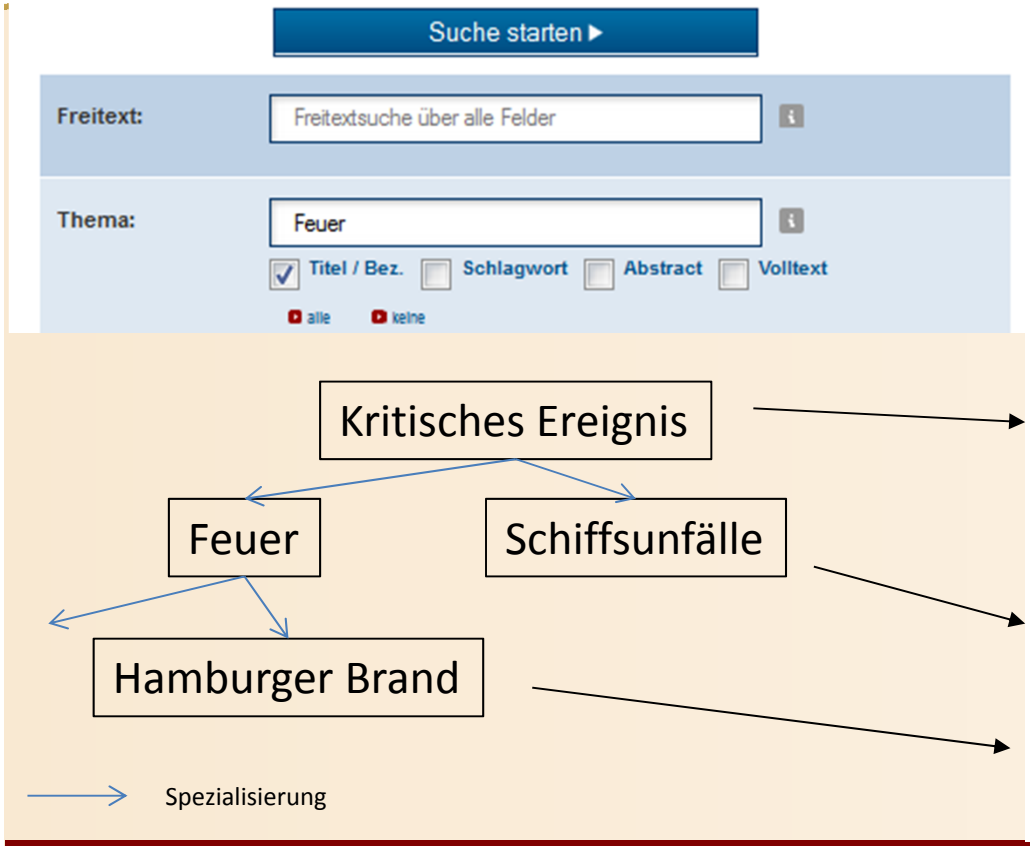
Eingabe eines Ortsbereichs: e_o
Ortsbereich des Informationsobjekts: i_o

Mögliche Interpretationen:
$i_o = e_o$: *gleiche* Ortsbereiche
$i_o \sim e_o$: *ungefähr* gleiche Ortsbereiche
$i_o \subset e_o$: Eingabebereich *umfasst* Informationsobjektsbereich
$e_o \subset i_o$: Eingabebereich *innerhalb* Informationsobjektsbereich (nur direkt umliegend(!))
$e_o \cap i_o$: Eingabebereich hat *Schnitt* mit Informationsobjektsbereich

Felder mit Ortsbezug

- Aufbereitung:
  - ❖ Metadatenfelder: content_place, content_georeference
  - ❖ Punkt, Polygon, Bounding-Box
  - ❖ Abbildung von ortsbezogenen Begriffen auf Georeferenz
- Suche:
  - ✓ **Syntaktische Suche nach Ortsbezugsbegriff**
  - ❖ Koordinatenbereichssuche
- Darstellung:
  - ❖ Kartendarstellung mit Orten der Informationsobjekte

Textuelle Felder



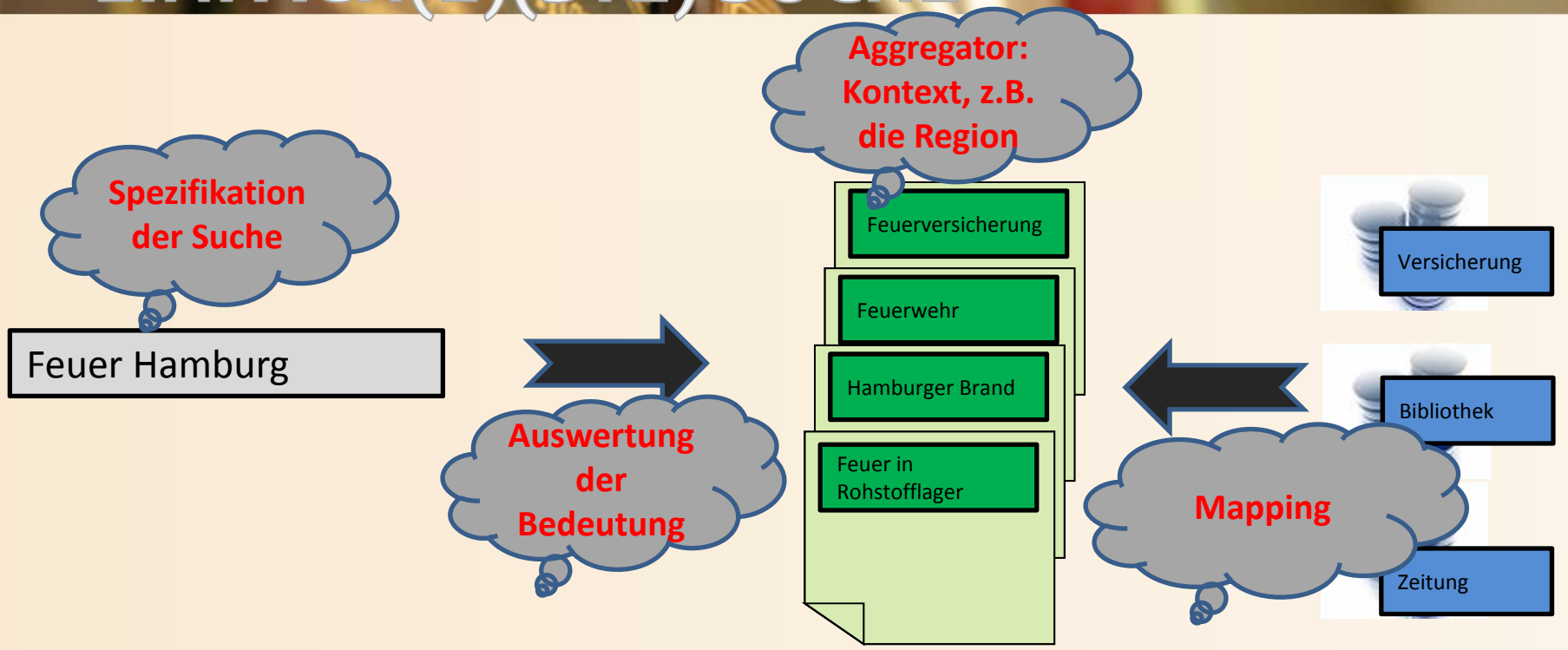
Eingabe eines Strings: e_t
Texte im Informationsobjekts: i_t

Mögliche Interpretationen:
$i_t = e_t$: ***Stringsuche,*** *Synonyme*
$i_t \sim e_t$: *Rechtschreibprüfung, etc.*
$e_t \subset i_t$: Eingabebereich *innerhalb* Informationsobjektsbereich (nur direkt umliegend(!))
$e_t \cap i_t$: Eingabebereich hat *Schnitt* mit Informationsobjektsbereich
$i_t \subset e_t$: Eingabebereich *umfasst* Informationsobjektsbereich

Textuelle Felder

- Aufbereitung:
  - ✓ Indexierung
  - ❖ Anreicherung mit Kategorien
- Suche:
  - ✓ **Syntaktische Suche**
  - ❖ Ontologische Suche
- Darstellung:
  - ✓ **Anzeige der Fundstellen des Suchworts**
  - ❖ Erklärung der ontologischen Suche

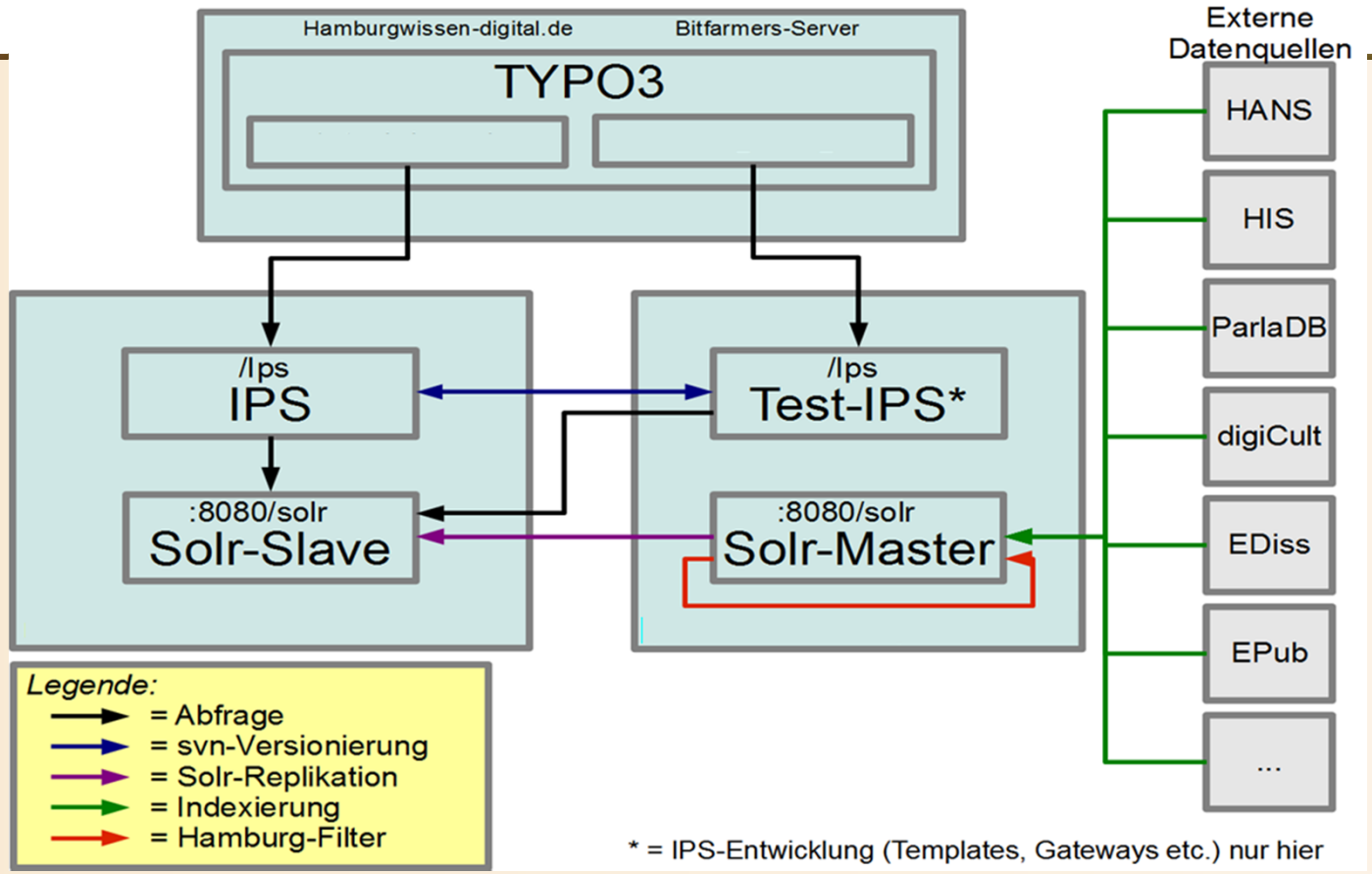
EINFACH(E)(STE) SUCHE



Technische Herausforderungen

- ✓ Trennung von Entwicklungs- und Produktivsystem
  - Entwicklung und Test von neuen Verfahren
- ✓ Trennung von Suche und Indexierung
  - Dauerhaftes Harvesting und Indexierung

Systemarchitektur



Werkzeuge

- Programmiersprachen:
  - Perl, Python, PHP, Java, Javascript, Shell, XSLT
- Infrastruktur:
  - Solr, Lucene, Tomcat, Apache, IPS, Suse, Debian, Corba, Wget, FTP, OAI-Harvester, Saxon, Mysql
- Formate:
  - Lido, XmetaDiss, proprietäre XML-Formate, HTML, museumdat, CSV

Vielen Dank für Ihre Aufmerksamkeit

Dr. Lothar Hotz

Mail: hotz@informatik.uni-hamburg.de
Tel.: 040-42883 2605

Arved Solth

Mail: solth@informatik.uni-hamburg.de
Tel.: 040-42883 2611