Inhaltliche Erschließung neuer Fachdomänen
auf der Grundlage von Termextraktionsverfahren

DTT-Symposion 2018
22. bis 24. März, Mannheim

Axel Theofilidis
axel.theofilidis@iailc.de
IAI Linguistic Content AG, Saarbrücken

Was die Zukunft bringt

Allenthalben entstehen neue Technologien …

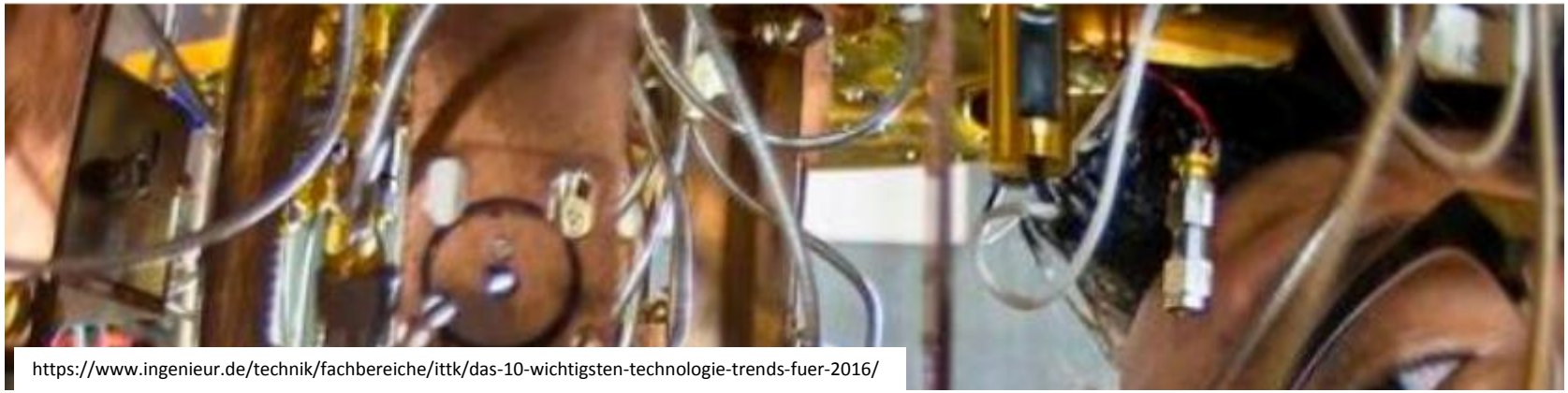
© 2018 IAI Linguistic Content AG

Was die Zukunft bringt

TECHNOLOGIE-RANKING 20.12.2017, 10:45 Uhr

Das sind die 10 wichtigsten Technologietrends für 2018

Was die Zukunft bringt, wollen wir alle gerne wissen. Für Unternehmen ist das sogar überlebenswichtig. Deshalb hat das Marktforschungsinstitut Gartner die zehn wichtigsten Technologietrends des kommenden Jahres eruiert. Sie werden unsere Zukunft (mit-)formen.



https://www.ingenieur.de/technik/fachbereiche/ittk/das-10-wichtigsten-technologie-trends-fuer-2016/

Was die Zukunft bringt

Neue Technologien prägen neue Begriffssysteme
mit ihren jeweils spezifischen Termen.

⇒ Für Technologieunternehmen gilt es, kontinuierlich Terminologiebestände zu aktualisieren und zu konsolidieren:

→ neu aufkommende Terminologie erfassen

→ neue Terme in Bestandsterminologie einbinden

→ terminologischem Wildwuchs vorbeugen

© 2018 IAI Linguistic Content AG

Was die Zukunft bringt

Neue Technologien prägen neue Begriffssysteme
mit ihren jeweils spezifischen Termen.

⇒ Für die Informationsgesellschaft insgesamt gilt es,
den Zugang zu neuen Technologien zu sichern:

→ intelligente Informationsaufbereitung etablieren

→ effektive Strategien der Informationssuche unterstützen

© 2018 IAI Linguistic Content AG

Methoden der inhaltlichen Erschließung …

Übersicht

- Ausgangspunkt Termextraktion
- Automatische Hierarchisierung von Termen
- Automatische Verschlagwortung von Texten
- Termwolken & Textwortwolken
- Wortwolkenbasierte Suche
- Zusammenfassung

© 2018 IAI Linguistic Content AG

Ausgangspunkt Termextraktion

- Das Wissen einer Fachdomäne manifestiert sich in ihren Texten.
- Wesentlicher Informationsträger in den Texten einer Fachdomäne sind deren spezifische Terme.

⇒ Termextraktion als grundlegendes Verfahren für die inhaltliche Erschließung einer Fachdomäne.

© 2018 IAI Linguistic Content AG

Ausgangspunkt Termextraktion

Ausgehend von einem umfassenden fachspezifischen Textkorpus:

→ Welche Terme treten erstmals mit signifikanter Häufigkeit auf?

→ In was für Beziehungen stehen neu auftretende Terme zu den bereits etablierten Termen des erweiterten Fachbereichs?

⇒ Neu auftretende Terme müssen sich in eine gegebene Bestandterminologie begrifflich einfügen!

© 2018 IAI Linguistic Content AG

Einbeziehung einer Bestandsterminologie

Im Zuge der Termextraktion ist ein Abgleich mit einer gegebenen Bestandsterminologie durchzuführen:

→ Bereits erfasste Terme werden erkannt und gesondert ausgewiesen, darunter auch …

- **nicht erlaubte Terme** (Negativterme) sowie …
- **sprachliche Varianten** von Vorzugstermen!

© 2018 IAI Linguistic Content AG

Einbeziehung einer Bestandsterminologie

Grundform	Originalwort	Wortart	Genus	Kriterium
Kühleranschluss	Kühleranschluss	noun	m	termcand
Kühlermodullager	Kühlermodullager Kühlermodullagern	noun	n	kterm
Kühler-Niedertemperatur	Kühler-Niedertemperatur	noun	f	termcand
Kühlerpin	Kühlerpin Kühlerpins	noun	m	DEPR
Kühlerschutzblech	Kühlerschutzblech	noun	n	termcand
Kühlerverschraubung	Kühlerverschraubung	noun	f	termcand
Kühlmittel	Kühlmittel Kühlmittels	noun	n	kterm
Kühlmittel-Ablassschraube Kühlmittelablassschraube	Kühlmittel-Ablassschraube Kühlmittelablassschraube	noun	f	VAR kterm
Kühlmittelausgleichbehälter	Kühlmittelausgleichbehälter	noun	m	termcand

⇒ Die effektive Erkennung von Termvarianten legt ein linguistisch fundiertes Termextraktionsverfahren nahe!

© 2018 IAI Linguistic Content AG

Linguistisch fundierte Termextraktion

Morphologische und grammatische Analyse des Textkorpus:

→ Zuweisung allgemeiner linguistischer Merkmale (z.B. Wortart, Genus, Numerus, Semantikklasse)

→ Erkennung von Wortstrukturen

→ Rückführung von Wortbestandteilen auf ihre Wortwurzel

Emissionsreduktion

{lu=**emissionsreduktion**,c=**noun**,g=**f**,nb=**sg**,
s=**massnahme**,cs=**n#n**,ls=**emittieren#reduzieren**}

© 2018 IAI Linguistic Content AG

Linguistisch fundierte Termextraktion

Morphologische und grammatische Analyse des Textkorpus:

→ Erkennung grammatischer Strukturen

Aktuelle Forschungsergebnisse zeigen, dass bei den bestehenden Motorenkonzepten noch Raum für Emissionsreduktion existiert.

```
NP [ Aktuelle<adj> Forschungsergebnisse<n> ]
zeigen<fiv> , dass<subj>
PP [ bei<p>
   NP [ den<art> bestehenden<adj> Motorenkonzepten<n> ]]
noch<part> Raum<n>
PP [ für<p>
   NP [ Emissionsreduktion<n> ]]
existiert<fiv> .
```

© 2018 IAI Linguistic Content AG

Linguistisch fundierte Termextraktion

Emissionsreduzierung

Emissionsreduzierung > Reduzieren

Term: Emissionsreduzierung
Term-ID: 30696
Anzahl: 38

So wurde der Term gefunden:

Anzahl	Match-Typ	Text	ls
26	kterm	Emissionsreduzierung	emittieren#reduzieren
5	VAR	Reduktion der Emissionen	reduzieren d_art emittieren
3	VAR	Reduktion von Emissionen	reduzieren von emittieren
2	VAR	Emissionsverminderung	emittieren#ver$mindern
1	VAR	Emissionseinsparungen	emittieren#ein_$sparen
1	VAR	Verminderung der Emissionen	ver$mindern d_art emittieren

Variantenerkennung durch

→ Abgleich der Wortwurzeln

→ Synonymisierung

→ Transformation

© 2018 IAI Linguistic Content AG

Automatische Termhierarchisierung

Das Wissen über die sprachliche Struktur von Termen ermöglicht zugleich die automatische Hierarchisierung der Terme.

→ Durch den sukzessiven Abzug von Bestimmungswörtern werden zu jedem Term Hyperonyme aufgezeigt.

Brennstoffzellen-Elektroantriebsstrang

Brennstoffzellen-Elektroantriebsstrang > Elektroantriebsstrang > Antriebsstrang > Strang

© 2018 IAI Linguistic Content AG

Automatische Termhierarchisierung

Das Wissen über die sprachliche Struktur von Termen ermöglicht zugleich die automatische Hierarchisierung der Terme.

→ Termkandidaten und Bestandsterme werden durch unterschiedliche Formatierung kenntlich gemacht.

Brennstoffzellen-Elektroantriebsstrang

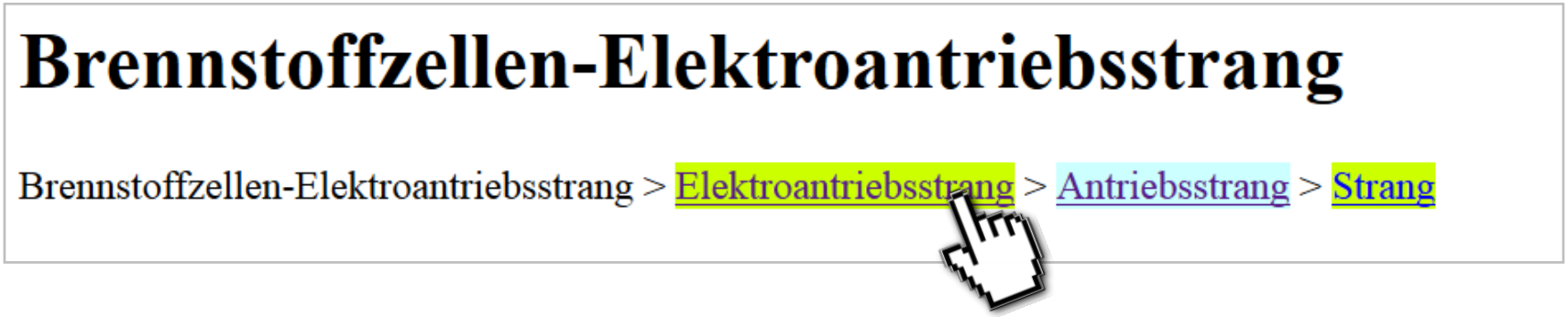
Brennstoffzellen-Elektroantriebsstrang > Elektroantriebsstrang > Antriebsstrang > Strang

© 2018 IAI Linguistic Content AG

Automatische Termhierarchisierung

Das Wissen über die sprachliche Struktur von Termen ermöglicht zugleich die automatische Hierarchisierung der Terme.

→ Durch Verlinkung der Terme kann entlang der Begriffshierarchien navigiert werden.



© 2018 IAI Linguistic Content AG

Automatische Termhierarchisierung

Das Wissen über die sprachliche Struktur von Termen ermöglicht zugleich die automatische Hierarchisierung der Terme.

→ Durch Verlinkung der Terme kann entlang der Begriffshierarchien navigiert werden.

© 2018 IAI Linguistic Content AG

… und andere sprachliche Ähnlichkeiten

Antriebsstrang

Antriebsstrang > Strang

Ähnlichkeiten rechts

Termkandidaten

2	Abgasstrang
3	alternativer Antriebsstrang
1	aufgebauter Antriebsstrang
5	autarker Hybrid-Antriebsstrang
1	automobiler Antriebsstrang
1	automobiler Triebstrang
2	Basisantriebsstrang

Neben hierarchischen Begriffsrelationen werden vielfältige Ähnlichkeiten bezüglich der Grund- und Bestimmungswörter von Termen aufgezeigt.

© 2018 IAI Linguistic Content AG

… und andere sprachliche Ähnlichkeiten

Die wissensbasierte Termextraktion unterstützt durch die Explikation sprachlicher Beziehungen:

→ Entscheidungsprozesse beim Terminologieaufbau

→ Festlegung von Begriffsrelationen innerhalb der Terminologie

⇒ Ergänzt um statistische Methoden ist die wissensbasierte Termextraktion zugleich Grundlage für die **automatische Verschlagwortung** von Texten.

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

Die Verschlagwortung von Texten zielt auf die optimale Repräsentation von Textinhalten.

Dies ist Grundvoraussetzung für die effektive Informationssuche in Textbeständen.

→ Welches sind die für einen Text relevantesten Terme?

⇒ Zuweisung von Termen als gewichtete Schlagwörter

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

Bestimmung der informationellen Relevanz von Termen auf Grundlage des **tf-idf-Maßes**:

- **tf** (Termfrequenz): Terme, die in einem Dokument oft vorkommen, sind wichtiger als weniger oft vorkommende Terme.
- **idf** (inverse Dokumentfrequenz): Terme, die in wenigen Dokumenten oft vorkommen, sind wichtiger als Terme, die in sehr vielen Dokumenten jeweils selten vorkommen.

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

Hinzuziehung weiterer Faktoren zur Bestimmung der Termrelevanz:

→ Berücksichtigung von Strukturelementen und Textposition
`Titel`, `Überschrift`, `Absatz`, …

⇒ Terme in Titeln oder Überschriften werden stärker gewichtet.

→ Häufigkeitsverteilung semantischer Klassen in einem Text
AGENT, DING, INSTRUMENT, MATERIAL, PROZESS, …

⇒ Terme in häufig vorkommenden semantischen Klassen werden stärker gewichtet.

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

Wichtig für die Bestimmung der Termrelevanz und
für die entsprechende Zuordnung von Schlagwörtern:

→ Zusammenführung von Termvarianten

Emissionsreduzierung
Reduktion der Emissionen
Reduktion von Emissionen
Emissionsverminderung
Emissionseinsparungen
Verminderung der Emission
…

} **Emissionsreduzierung**

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

Wichtig für die Bestimmung der Termrelevanz und
für die entsprechende Zuordnung von Schlagwörtern:

→ Erkennung von Termen in Wortzusammensetzungen

⇒ Unterschiedliche Gewichtung, je nachdem ob

- **allein stehend**
- **Grundwort** einer Wortzusammensetzung
- **Bestimmungswort** einer Wortzusammensetzung

© 2018 IAI Linguistic Content AG

Automatische Verschlagwortung

<h1>**Leichtbau** bei **Elektrofahrzeugen**</h1>
Hohe Kosten sowie eine eingeschränkte Reichweite sind die zentralen Hürden für einen breiten Markteintritt von **Elektrofahrzeugen**. Ein entscheidender Stellhebel zur Reduktion der zu installierenden Batteriekapazität ist die Senkung des Energieverbrauchs des Fahrzeugs. **Leichtbau** wird hier als ein Schlüssel zum wirtschaftlichen Erfolg des **Elektrofahrzeugs** angesehen. Um die Bedeutung von **Leichtbau**maßnahmen in **Elektrofahrzeugen** aufzuzeigen, werden die bisherigen **Leichtbau**-maßnahmen in konventionellen Fahrzeugen beschrieben. Nach der Quantifizierung des Kostenpotenzials von **Leichtbau**maßnahmen in **Elektrofahrzeugen** erfolgt auf Basis der geänderten Gewichtsverteilung der Baugruppen und Komponenten im **Elektrofahrzeug** die Beschreibung von sinnvollen **Leichtbau**maßnahmen. Hierzu wird sowohl auf bestehende **Leichtbau**konzepte eingegangen, deren Anwendung sich im **Elektrofahrzeug** lohnen könnte, als auch auf **Leichtbau**potenziale, die bis dato keine Anwendung in konventionellen Fahrzeugen gefunden haben.

Descriptors: **Elektrofahrzeug [100]**; **Leichtbau [40]**; Energieverbrauch [28]; hohe Kosten [18]; Batteriekapazität [3]

Free Descriptors: Leichtbaumaßnahme [100]; Leichtbaukonzept [20]; Stellhebel [14]; Leichtbaupotenzial [12]; eingeschränkte Reichweite [6]; zentrale Hürde [6]; breiter Markteintritt [6]; Kostenpotenzial [4]

Hyperonyms: Fahrzeug [100]; Bauweise [40]; Verbrauch [28]; Kosten [18]

© 2018 IAI Linguistic Content AG

Dokumentähnlichkeiten

Ergänzt um weitere statistische Methoden ist die automatische Verschlagwortung von Texten zugleich Grundlage für die Ermittlung von **Dokumentähnlichkeiten.**

→ Abgleich des Textindexats mit den Indexaten aller anderen Texte unter Einrechnung der Termrelevanzen

⇒ Welche ähnlichen Texte gibt es zu einem gegebenen Text?

⇒ Weitere **Termähnlichkeiten** werden dagegen durch **Wortwolken** aufgezeigt.

© 2018 IAI Linguistic Content AG

Termwolken

Wortwolken sind Gruppen semantisch ähnlicher Wörter.

Wortwolken werden auf Basis der Kookkurrenz von Wörtern in umfangreichen Textkorpora ermittelt.

→ Berechnung von **Termwolken** im Zuge einer Termextraktion

⇒ Mit welchen anderen Termen tritt ein Term im Textbestand besonders häufig zusammen auf?

© 2018 IAI Linguistic Content AG

Termwolken

Im Zuge einer Termextraktion ermittelte Termwolken unterstützen:

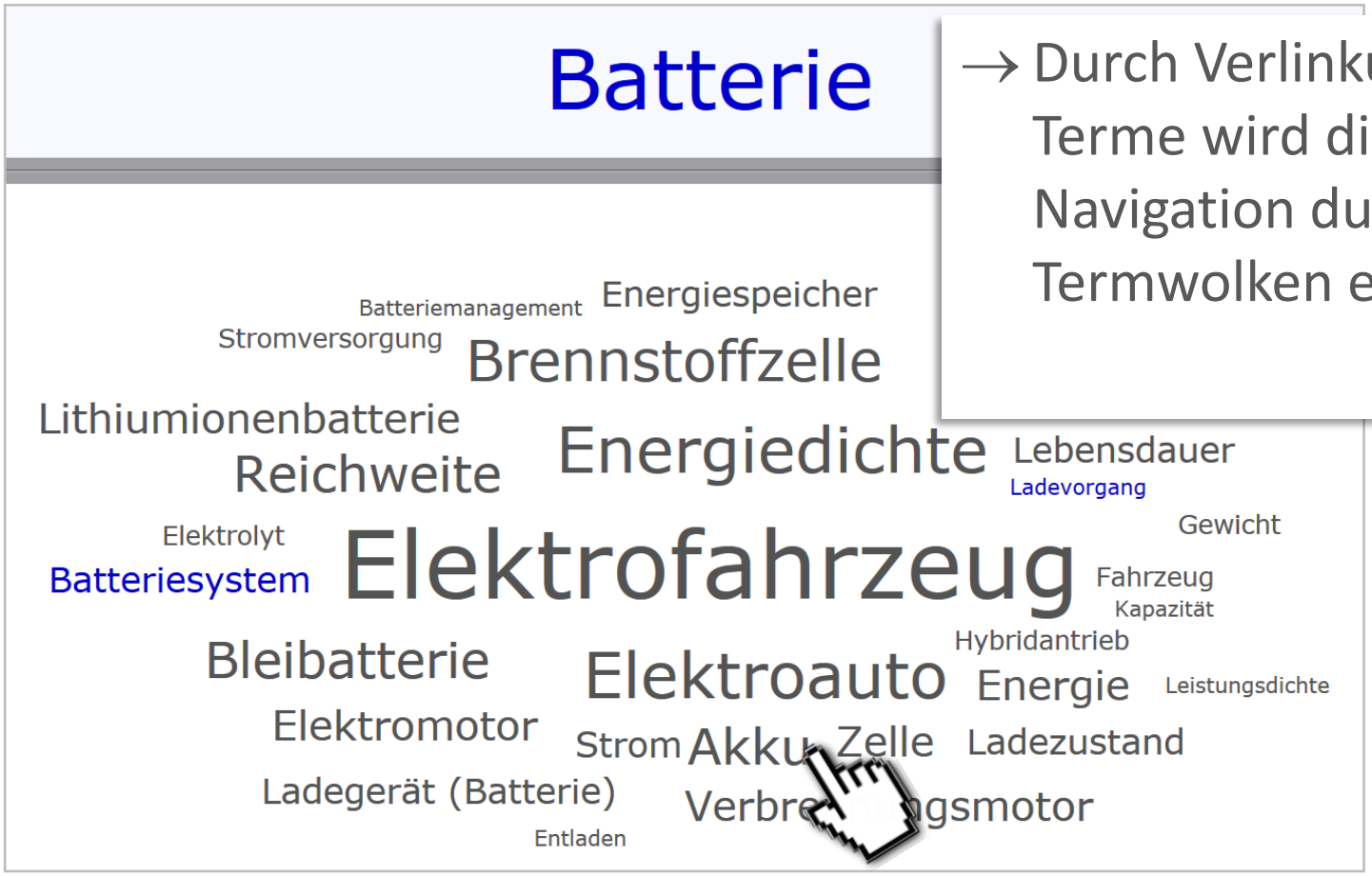
→ den Terminologieaufbau:
durch Aufzeigen weiterer semantischer Ähnlichkeiten

→ die Verschlagwortung von Texten:
durch Vereindeutigung von Homonymen

→ das Auffinden von Inhalten in Textbeständen:
durch Bereitstellen von relevanten Suchhinweisen

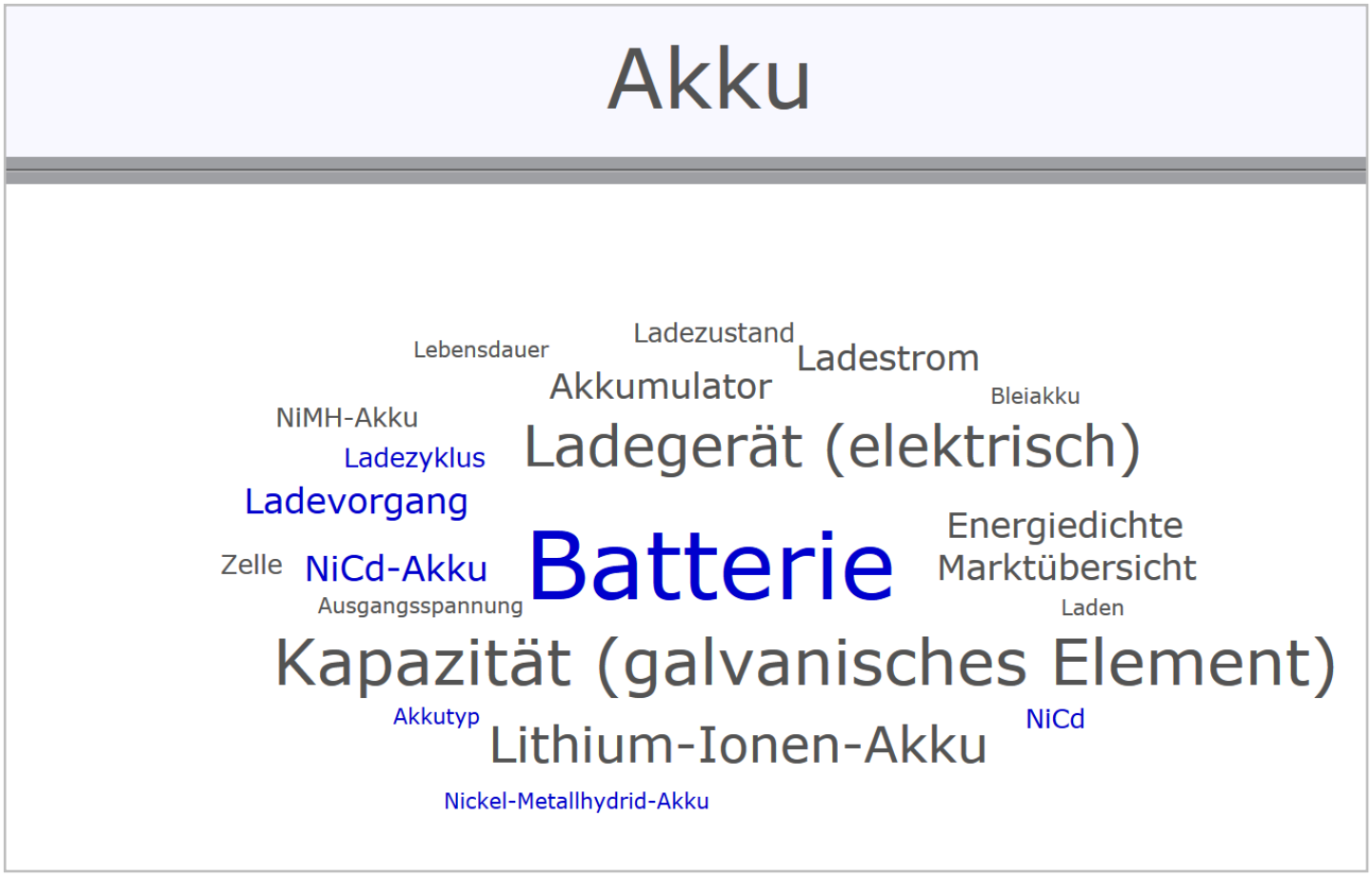
© 2018 IAI Linguistic Content AG

Termwolken



→ Durch Verlinkung der Terme wird die Navigation durch Termwolken ermöglicht.

Termwolken

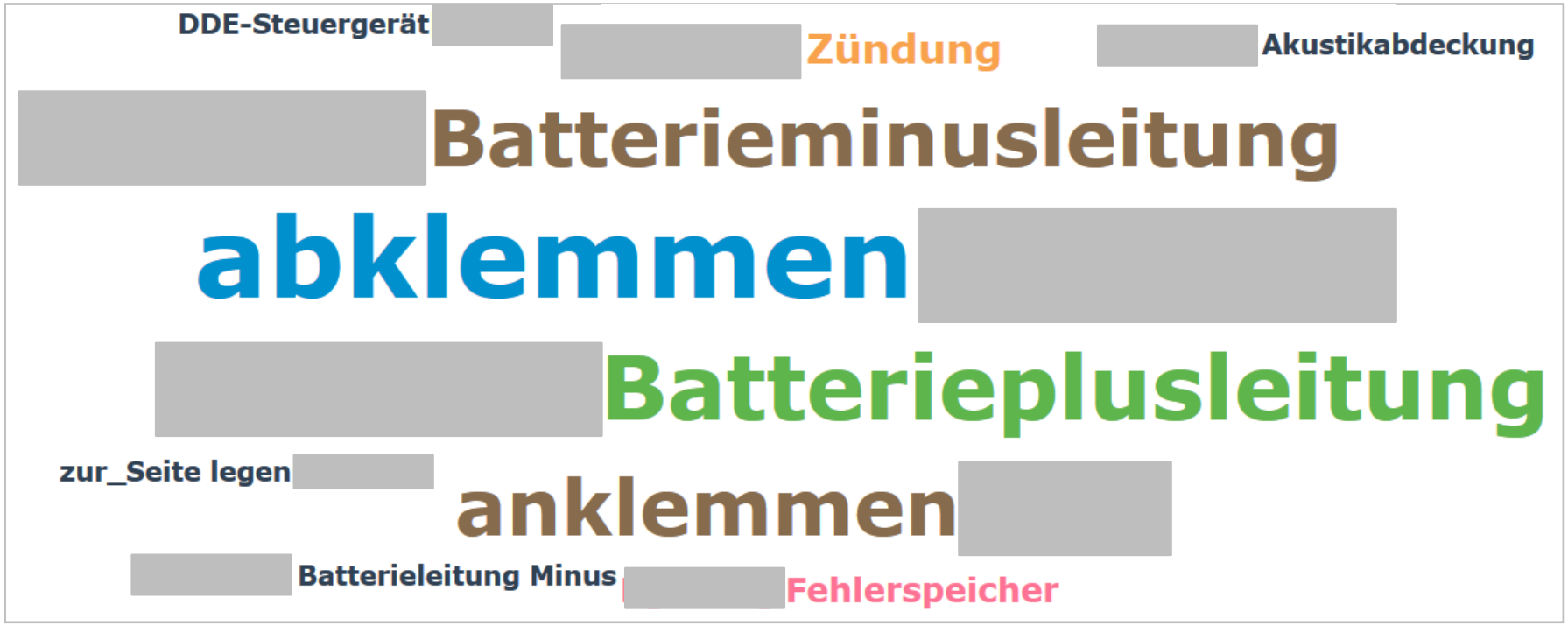


© 2018 IAI Linguistic Content AG

… auch für Verben / Prozesse

© 2018 IAI Linguistic Content AG

… und ihre Mitspieler



© 2018 IAI Linguistic Content AG

Von Termwolken zu Textwortwolken

Ausgehend von Termwolken lässt sich für jeden Text die spezifische **Textwortwolke** ermitteln.

→ Zu jedem Text sind die relevantesten Terme bekannt.

→ Zu jedem Term ist die korpusspezifische Termwolke bekannt.

⇒ Durch Aufsummierung der Termwolken der prominentesten Terme eines Texts ergibt sich die spezifische Textwortwolke.

© 2018 IAI Linguistic Content AG

Textwortwolken

Nächste Steckdose auf See fertig

ABB hat kürzlich die mit 200 km längste Kabelstrecke installiert, die bislang für eine Offshore-Windenergieanlage (WEA) verlegt wurde. Die Konverterstation dazu befindet sich rund 100 km von Borkum, auf dem Festland erfolgt die Einspeisung in das 380 kV-Netz bei Papenburg. Die Konverterstation soll 2010 400 MW übertragen, und zwar erstmals in HGÜ-Technik (Hochspar Leitungsverluste auszeichnet. Angesichts des wachsenden Bedarfs werden außerdem derzeit Untersuchungen durchgeführt, Kabeltras sondern in den Flussmündungen der großen Nordseeströme Elbe,

→ Nur die unterstrichenen Terme kommen im Text überhaupt vor.

→ Alle anderen Terme wurden „eingeschleppt“.

Von Termwolken zu Textwortwolken

Textwortwolken zeigen den weiteren informationellen Kontext eines Texts auf.

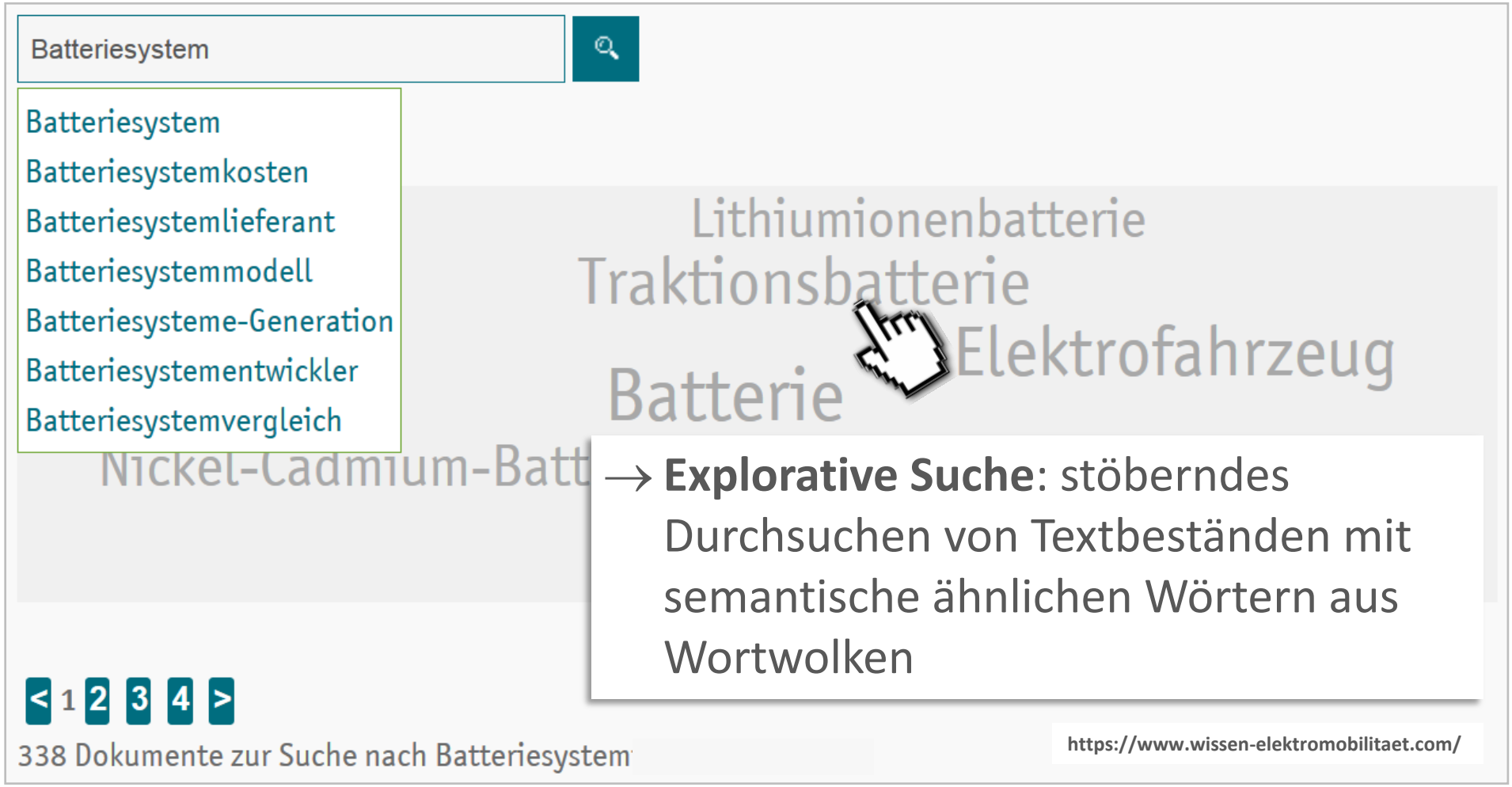
→ Da die aufsummierten Termwolken über den gesamten Textbestand ermittelt wurden, …

→ schleppen sie Terme ein, die dem betreffenden Term ähnlicher sind als alle anderen im Text vorkommenden Terme.

⇒ Textwortwolken wie Termwolken unterstützen so die **effektive Suche** in Textbeständen.

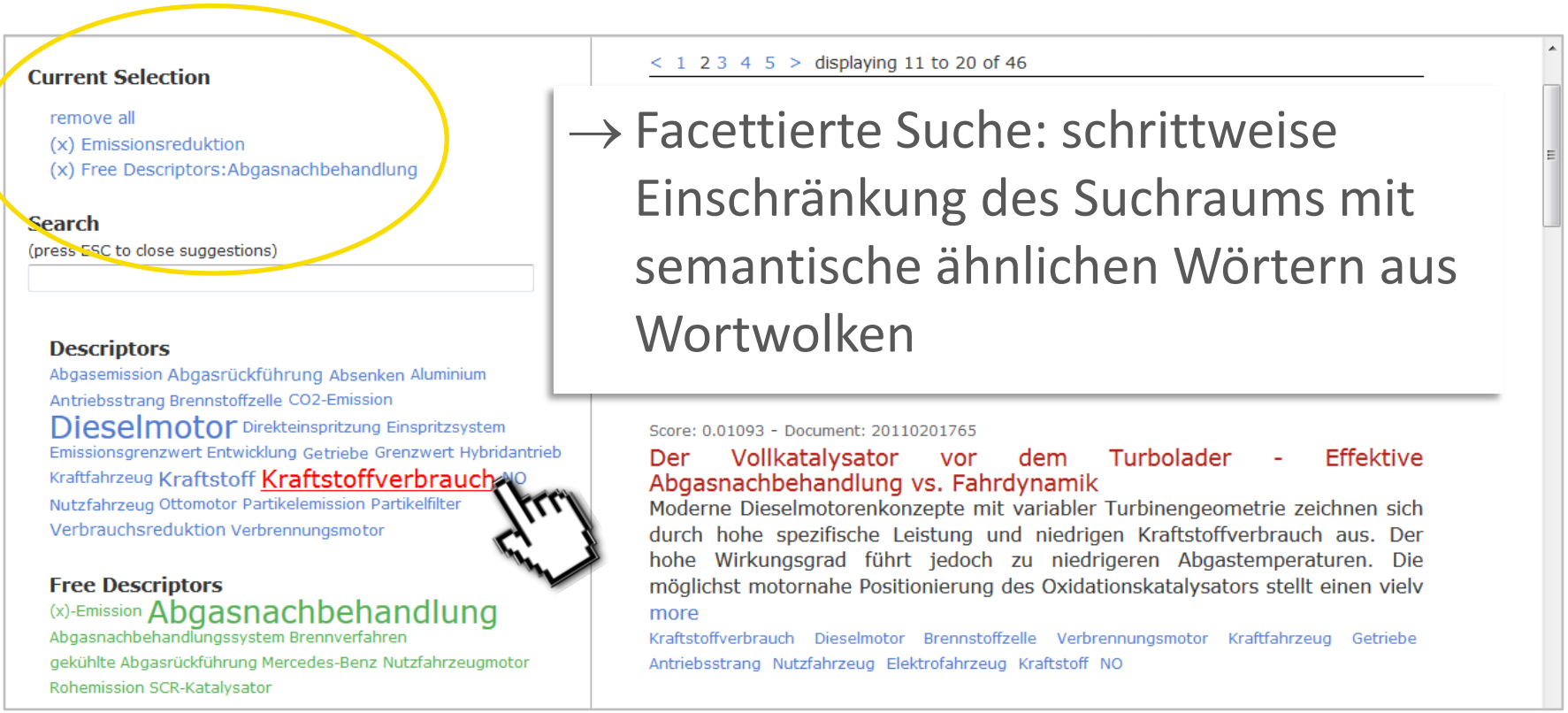
© 2018 IAI Linguistic Content AG

Effektive Suche - wortwolkenbasiert



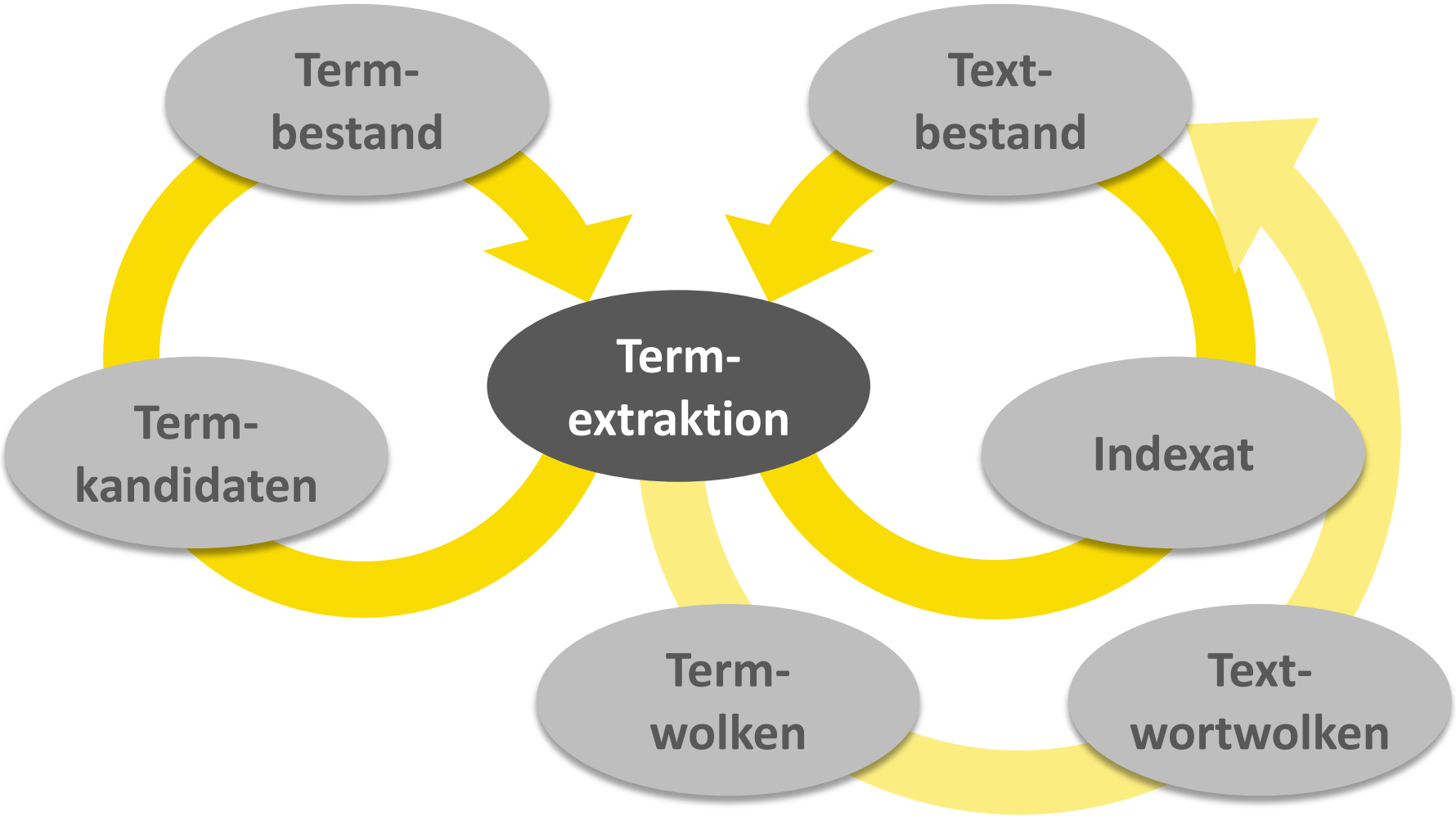
→ **Explorative Suche**: stöberndes Durchsuchen von Textbeständen mit semantische ähnlichen Wörtern aus Wortwolken

Effektive Suche - wortwolkenbasiert



→ Facettierte Suche: schrittweise Einschränkung des Suchraums mit semantische ähnlichen Wörtern aus Wortwolken

Zusammenfassung



© 2018 IAI Linguistic Content AG

Vielen Dank für Ihre Aufmerksamkeit!

© 2018 IAI Linguistic Content AG