Aufbau einer Wissensdatenbank mit einem Terminologieverwaltungssystem: Ein Erfahrungsbericht

Dr. François Massion,

D. O. G. Dokumentation ohne Grenzen GmbH

(francois.massion@dog-gmbh.de)

Motivation

- D.O.G. GmbH: bietet Übersetzungen und Softwareprodukte.
- Viele Projekte sind komplex und benötigen Wissen:
  - **Sprachliche** Aspekte: "*Gilt in Singapur britisches oder amerikanisches Englisch*"?
  - **Fachliche** Aspekte: "*Wo finde ich die Definition von einem 'Abstreifer'*"?
  - **Technische** Aspekte: "*Wie gehen wir mit JSON-Inhalten in CDATA-Blöcken um*"?
  - **Organisatorische** Aspekte: Unterschied zwischen *Revisor und Lektor*?
- Dazu kommen viele Wissensfragen aus dem Arbeitsalltag.

Motivation

- Das Wissen ist **in den Köpfen**.

> *... Das Wissen einer Organisation ist zu komplex als dass eine einzelne Person alles überblicken könnte.*

- Gerade wenn man **unter Zeitdruck** steht, taucht eine schwierige Frage auf, die man recherchieren muss.
- Zeit**aufwand** für:
  - Einarbeitung neuer Mitarbeiter
  - Reparatur von Fehlern (nicht gewusst wie…)
- Wir wollten Wissen sammeln und vom **Schwarmwissen** profitieren.

Unsere Datenbank und der Weg dahin

- Aktuell: 811 Begriffe – 988 deutsche Benennungen – 534 Relationen

Anzahl Sprachen: 3 (0 davon ganz ohne Benennungen)
Anzahl der Begriffe: 811
Anzahl der Benennungen: 1110
Anzahl der Werte: 2750

Sprache	Begriffe	Benennungen	Deckungsgrad
Deutsch	811	988	100,00%
Englisch	88	120	10,85%
Französisch	2	2	0,25%

Anzahl der Relationen (Subjektebene -- Objektebene)

Relation	C -- C	C -- T	T -- C	T -- T	total
Hat	23	0	0	0	23
Teil_von	82	0	0	0	82
Ist_ein	164	0	0	0	164
Beeinflusst	126	0	0	0	126
Benutzt	52	0	0	0	52
Lösung_für	46	0	0	0	46
Relation_mit	13	0	0	0	13
Wird_angeboten_mit	23	0	0	0	23
Voraussetzung_für	5	0	0	0	5

C = concept level
T = term level

Was ist eigentlich eine Wissensdatenbank?

- Knowledgebase, Wiki, Ontologie ….Keine standardisierte Definition:

Was ist eine Wissensdatenbank?

*… Eine **Wissensdatenbank** oder **Wissensbasis** (englisch knowledge base) ist eine spezielle Datenbank für das Hinterlegen von Wissen.*

*Wissensdatenbank ist dabei im deutschen Sprachgebrauch **ein nicht definierter Begriff**, der meistens im Zusammenhang mit Wissensmanagement verwendet wird und eine Sammlung expliziten Wissens in meist schriftlicher Form darstellt.*

Wikipedia – Zugriff 13.10.2021

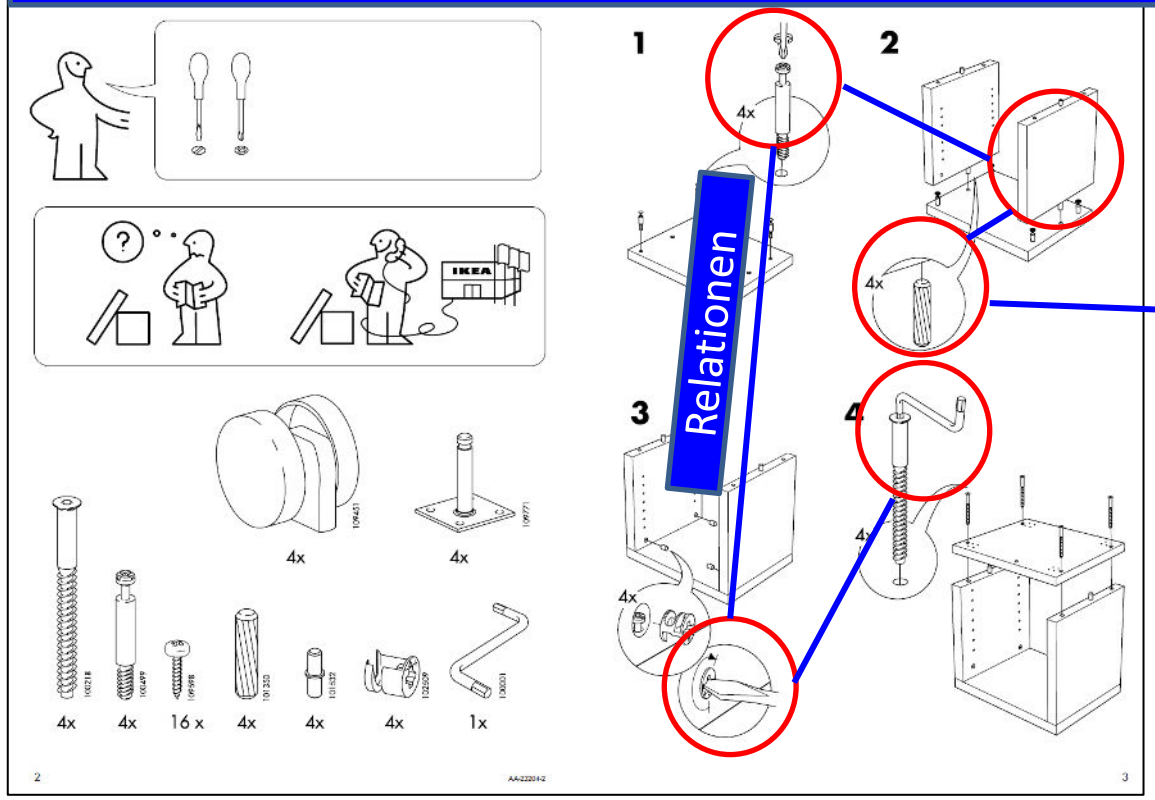
- Datenbank mit Sammlung von Texten
- Aber auch: Ontologie, ontologieartige Begriffssammlung

Begriffsklärung

- Einige Termini, wie sie in dieser Präsentation verwendet werden:
  - **Begriff**: Abstrakte Denkeinheit (z. B. was man sich unter *Software* vorstellt)
  - **Benennung**: Wort oder Wortgruppe, die einen Begriff bezeichnen
  - **Wissenseinheit**: Gruppe von Begriffen, die durch Relationen verbunden sind und einen semantischen Zusammenhang haben.
  - **Concept Map**: Grafische Darstellung einer Wissenseinheit

Unsere Lösung: Zwischen IKEA und Wikipedia

Begriffskontext + Begriffsinformationen



Dübel Begriff 15

Definition	beschichteter, glatter Stahlstab, der aneinander angrenzende Platten einer Fahrbahnbefestigung aus Beton an der Fuge miteinander verbindet, um die Lastübertragung zu verbessern und Stufenbildung zu verhindern
Quelle	DIN EN 13877-1:2013-06 Fahrbahnbefestigungen aus Beton - Teil 1: Baustoffe; Deutsche Fassung EN 13877-1:2013
Bild	4x
Kommentar	Dübel können aus verschiedenen Materialien sein: Holz, Kunststoff, Stahl, je nachdem, wozu sie eingesetzt werden. In der Möbelindusttrie sind die Dübel oft aus Holz.
Weitere Infos	308521746.1990425.PDF

Deutsch

Dübel

Status	Freigegeben
Verwendung	Erlaubt

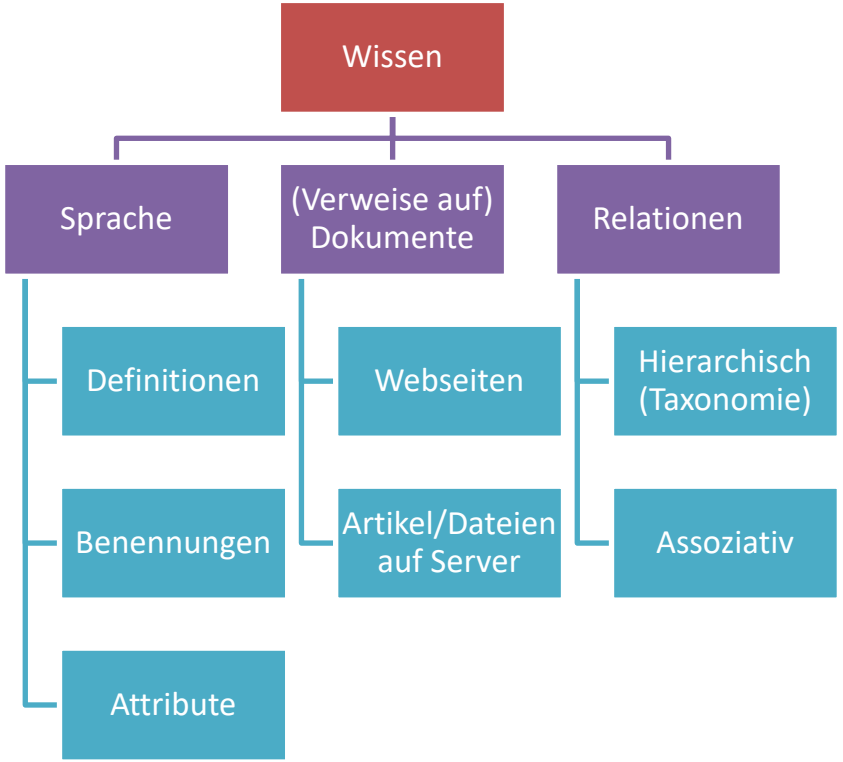
Historie

Englisch

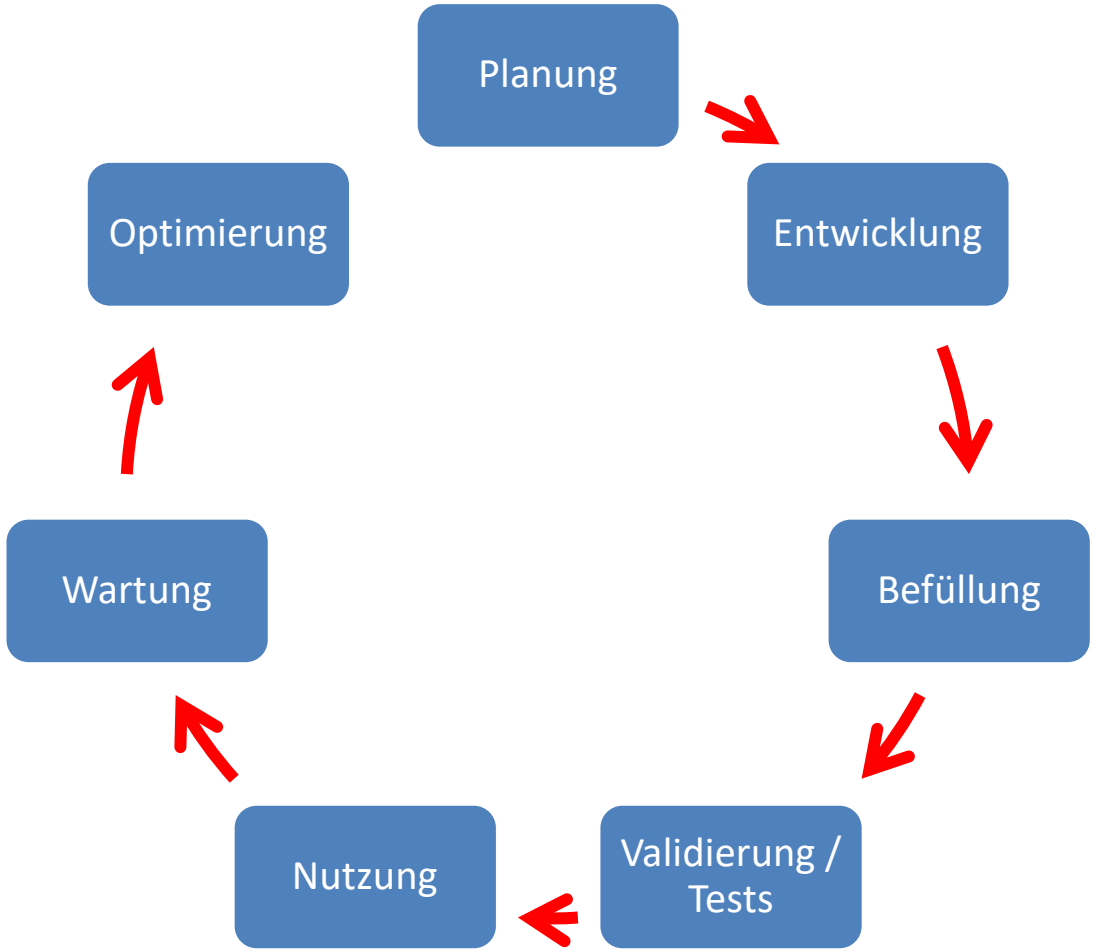
dowel

Der (grobe) Plan

- Unsere intelligente Terminologiedatenbank verwenden, um 3 **Komponenten eines Wissenssystems** zu vereinen:



Die Arbeitsschritte



Planung des Wissensaufbaus

- Wissensbedarf und Nutzen ermitteln:
  - **Was** möchten wir wissen: **Themen**
  - **Für wen** ist das Wissen bestimmt: **Zielgruppen**
  - **Was wird mit** dem Wissen **getan**: **Einsatzszenarien**
- Daraus: Quellen für die Terminologie- und Relationsextraktion
- **Relationstypen** festlegen, welche Relation für welche Aufgabe
- Validieren anhand konkreter Situationen

Schritt 1: Bedarf und Zielgruppen

- Bei D.O.G. arbeiten verschiedene Mitarbeitergruppen, die ihre eigenen Aufgaben haben:

Zielgruppe	Aufgabe
Projektmanager	Organisation von Kundenprojekten
EDV-Mitarbeiter	technische Vorbereitung und Nachbearbeitung der Projekte
Übersetzer, Terminologen und Revisoren	Sprachliche und liguistische Aufgaben
Vertriebs- und Marketing	Angebote, Beratung, Kommunikation
Verwaltung, Buchhaltung	kaufmännische Abwicklung, Compliance
Programmiererteam	Entwicklung intelligenter Lösungen von der Qualitätssicherung bis zur maschinellen Übersetzung.

Schritt 2: Auswahl der Texte/Quellen

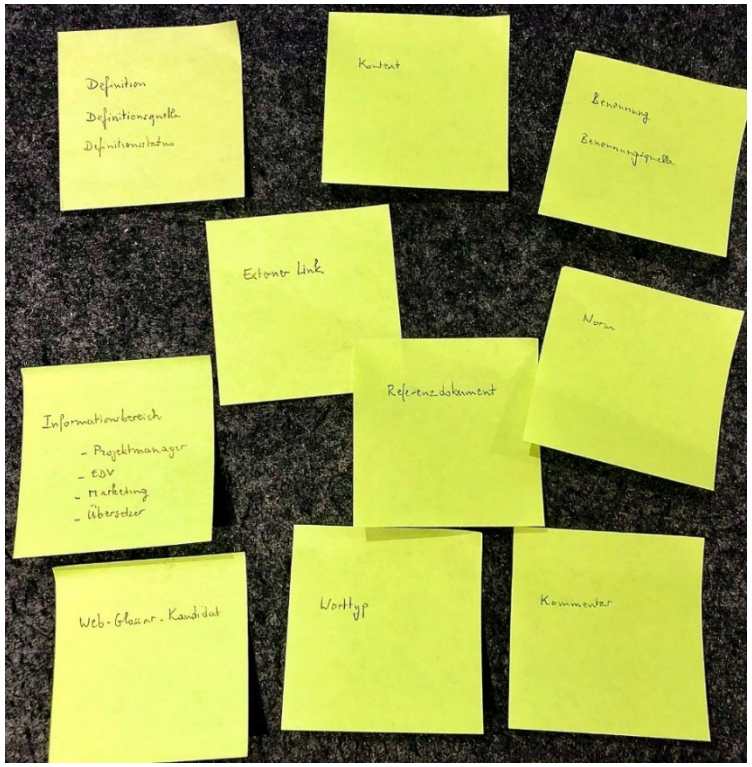
- Normen (ISO 17100)
- Firmeninternes Qualitätshandbuch
- Branchenpublikationen, Webseiten, Blogs (Tekom, Wettbewerber, Hochschulen, Konferenzprogramme usw..)
- Bücher, Fachartikel
- Kundenanfragen und Projektdaten
- Entwicklerdokumentation
- …

Schritt 3: Terminologie extrahieren

- Welche Termini: Terminologie unseres Geschäfts sammeln
  - Deutsch, teilweise Englisch
  - Begriffsorientiert, Begriffe mit Synonymen wie (*Ausgangstext* und *Quelltext*)
  - Definitionen, Bilder und Attribute
- Verschiedene **Extraktionsmethoden**:
  1. Manuelle Markierungen
  2. Statistische Verfahren: Kookkurrenz-Analyse, Häufigkeit
  3. Linguistische Verfahren: Nach grammatikalischen Kategorien (POS)
  4. Machine Learning: Word embeddings

Schritt 4: Wissensgerechtes Datenmodell

- **Informationsmodell**
  - Attribute und Felder auf Begriffs-, Sprach- und Benennungsebene
  - Klassifikationsfelder (Basis für **Taxonomien**):
    - für welche **Nutzergruppen** ist der Eintrag relevant?
    - Um welches **Sachgebiet** (Technik, Programmieren, Vertrieb, ...) handelt es sich?
  - Verweis auf externe oder interne Dokumente/Quellen
- Basis für bedarfsgerechtes Filtern von Wissen

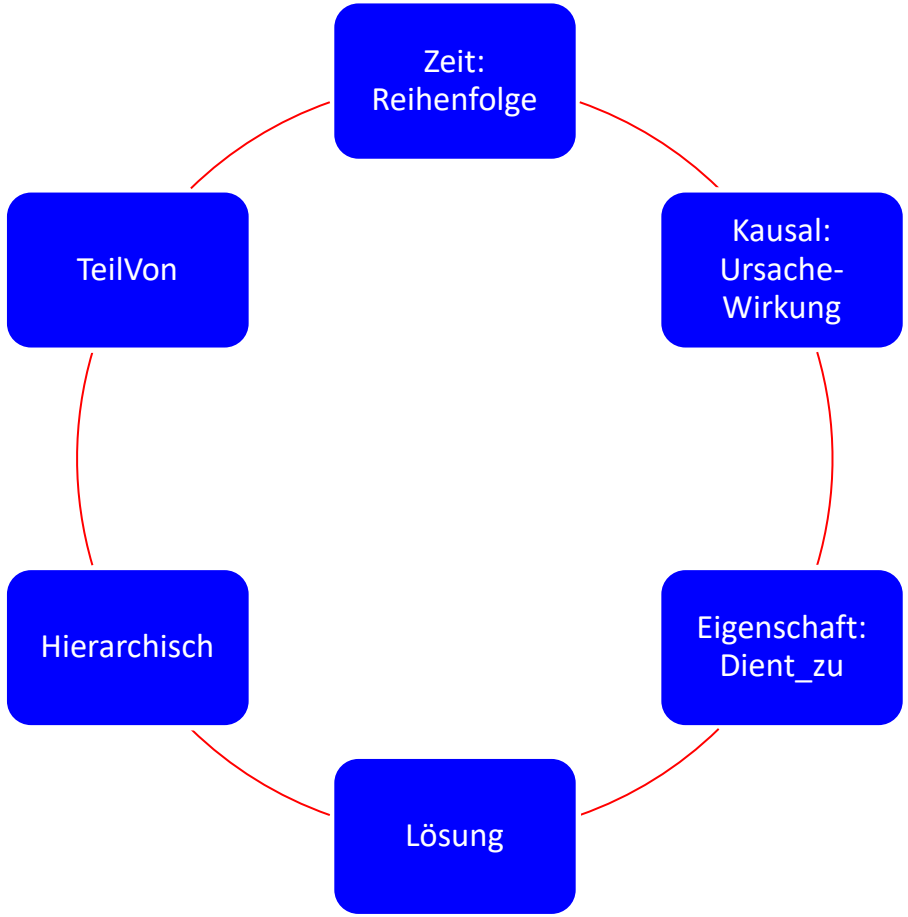


Schritt 5: Festlegen der benötigten Relationen

- **Nutzergruppenspezifisch** arbeiten
  - Welcher Informationsbedarf je Nutzergruppe?
  - Welche Hilfestellung bringt die Relation?
- Beispiel: EDV-Abteilung (technische Projektvorbereitung):
  - Mögliche Problemursachen
  - Mögliche Voraussetzungen für Bearbeitungsschritte
- Beispiel: Marketing:
  - Welche Produkte/Leistungen von DOG werden zusammen angeboten
  - Bessere SEO - Wörter finden

Liste der festgelegten Relationen

- Beeinflusst
- Benutzt
- Hat
- Ist_ein
- Lösung_für
- Relation_mit
- Teil_von
- Voraussetzug_für
- Wird_angeboten_mit



Schritt 6: Relationen extrahieren

- **Relationsmuster**: Triple S – P – O

Subjekt	Prädikat	Objekt
Modul	**Teil_von**	**CMS**

- Basis für Ontologien und Wissenssysteme
- Basis von RDF- und OWL-Format
- Basis für SPARQL-Abfragen

Relationen suchen: Wie gehen wir vor?

1. **manuelle** Relationsextraktion:
   - Nach Bedarf, um einzelne Themen/Fragen zu modellieren
   - Vorteil: Qualität der Ergebnisse
2. **(halb-)automatisierte** Relationsextraktion:
   a) Zuerst assoziierte **Begriffe** identifizieren
   b) Dann **Relationen** ermitteln

Regelbasierte Verfahren

- Verfahren aus dem Jahr 1992, nach M. Hearst *)
- Schablone (Relationsmuster) werden definiert
- Für jede Relation ein Muster
- Vorteil:
  - kein Training erforderlich
  - höhere **Präzision** (richtige Treffer), aber **Recall** (gefundene Relationen) geringer.

*) Quellenangabe: Marti Hearst, der 1992 einen Beitrag dazu geschrieben hat: "Automatic Acquisition of Hyponyms from Large Text Corpora": https://people.ischool.berkeley.edu/~hearst/papers/coling92.pdf.

Relationsmuster

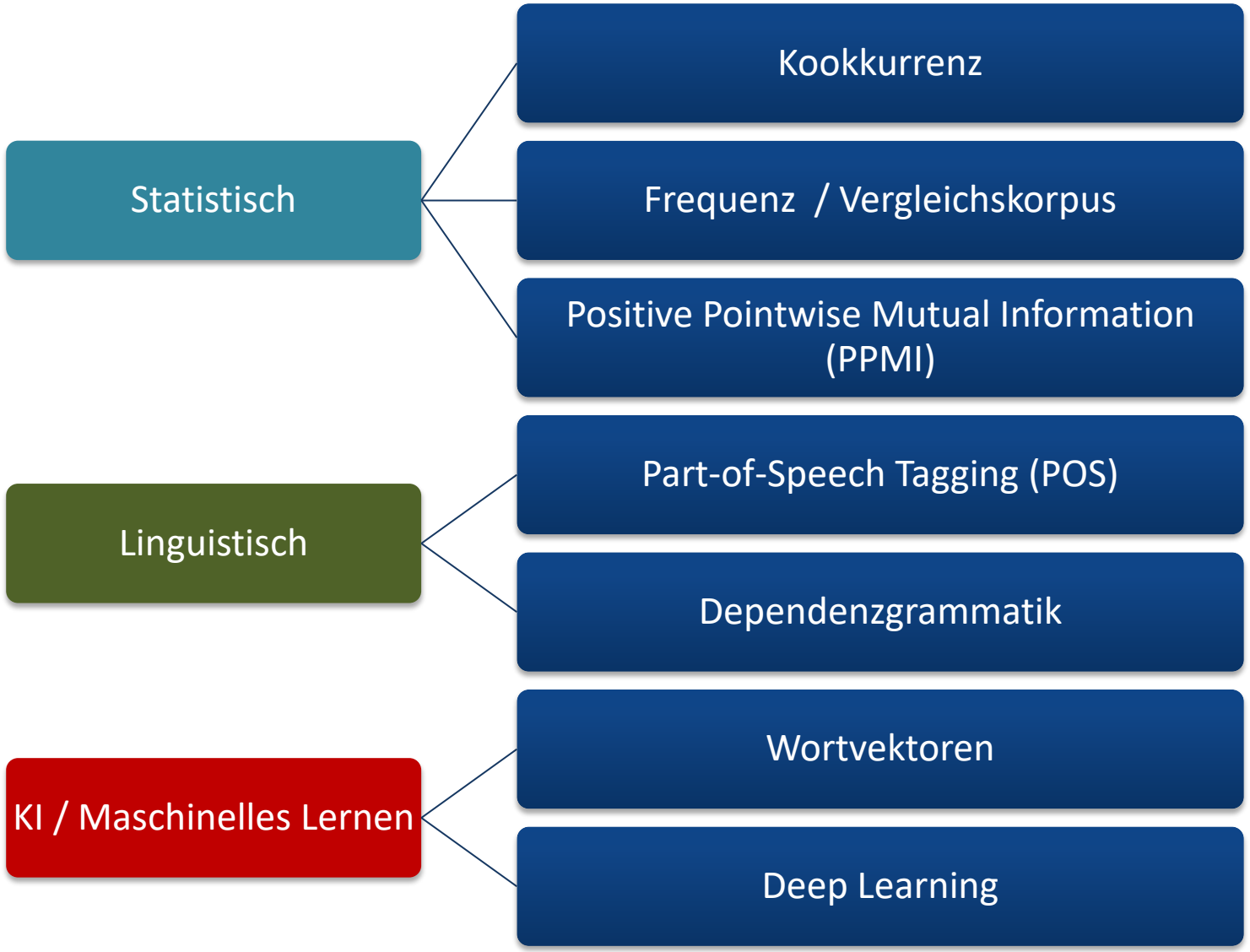
- Beispiel: Muster für die Extraktion einer **Ist_ein** Relation

Muster	Beispiel
X (…) **wie** (…) Y	Zu beachten sind spezielle Normen, wie die internationale IEC 82079-1:2012 …
X (…) **ist ein** (…) Y	Die tekom ist ein Verband zum Anfassen.

- Beispiel: Muster für die Extraktion einer **Dient_zu** Relation

Muster	Beispiel
X (…) **verwendet** (…) Y	Marketing-Cookies werden verwendet, um personalisierte Werbung anzuzeigen.
X (…) **mittels** (…) Y	Beschwerden können Sie mittels des Formulars an die Zertifizierungsstelle richten

Automatisierung: Methoden und Möglichkeiten



Suche nach Wörtern für Relationsmuster

- Das **Bootstrapping-Verfahren**:
  - Zuerst einige **Beispiele von Wortpaaren nehmen**, die eine bekannte Relation **R** ("**seed**") haben
  - Dann Satzbeispiele sammeln, die diese Paare enthalten
  - Den Kontext untersuchen (d. h. die Wörter in der Nähe des Paares)
  - Daraus **Muster für typische Beziehungen** ableiten

Suche nach Begriffspaaren (1)

- Statistisch

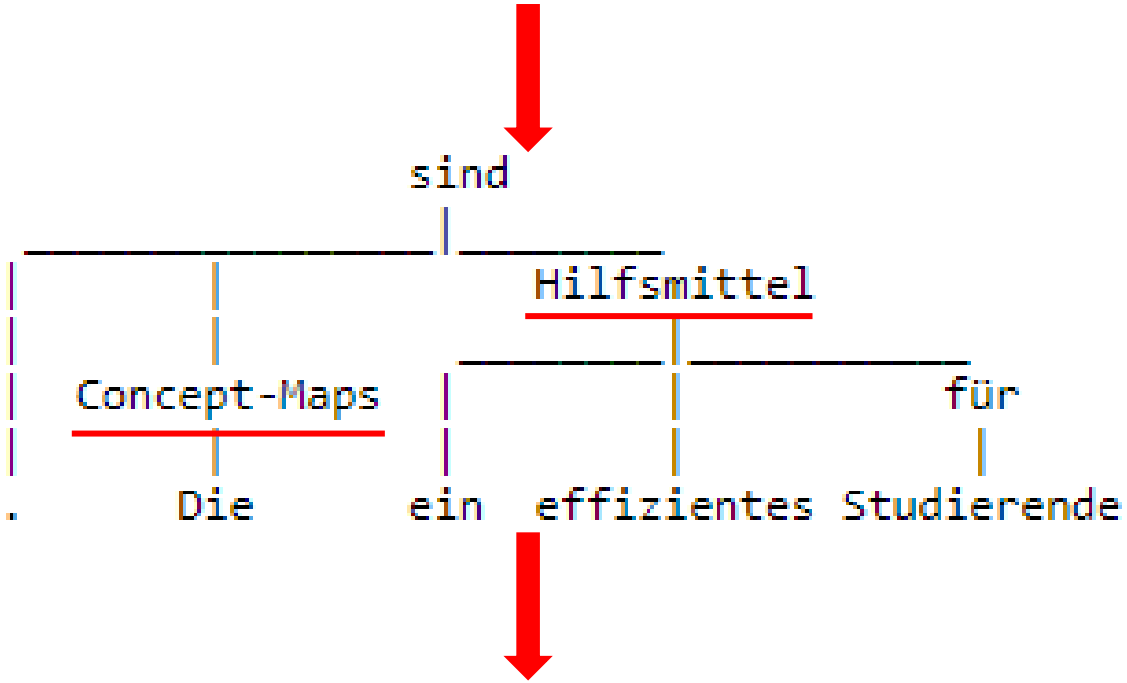
```
TermBase» TBX: 3, ISO: 3¶
Terminologie-Management»Vektor: 6¶
Terminologiearbeit» Beziehung: 5¶
Terminologiemanagement» tekom: 6¶
Terminologiesuche»Terminologie: 3, Vektor: 3, Suche: 3¶
```

TermBase *und* ***TBX*** *erscheinen 3 Mal zusammen*

Suche nach Begriffspaaren (2)

- Linguistisch (hier Dependenz)

"Die Concept-Maps sind ein effizientes Hilfsmittel für Studierende."



{Concept Map, Hilfsmittel}

Suche nach Begriffspaaren (3)

- Assoziierte Begriffe:
  - Kommen **im gleichen Satz** vor (→ Statistik)
  - Kommen **im gleichen Kontext** vor, müssen aber nicht im gleichen Satz erscheinen (→ Machinelles Lernen)
- Maschinelles Lernen (Word embeddings)

```
In [19]:
1 term = "satzstudios"
2 model.wv.similar_by_word(term,topn=20, restrict_vocab=None)

Out[19]: [('reguläre', 0.7579091787338257),
 ('benennungen', 0.7514564990997314),
 ('speicherort', 0.7497266530990601),
 ('vorschüsse', 0.6772931814193726),
 ('dtp-programm', 0.6578013896942139),
 ('synonym', 0.6426221132278442),
 ('rechtschreib-', 0.634267270565033),
 ('referenzdateien', 0.6088831424713135),
 ('textverarbeitungssystem', 0.6010711789131165),
 ('virenprüfung', 0.5745443105697632),
 ('warnung', 0.5681763887405396),
 ('eingriff', 0.5408499240875244),
 ('auftragsunterlagen', 0.5369100570678711),
 ('schreibanweisungen', 0.5357495546340942),
 ('makros', 0.5197840332984924),
 ('lokalisierte', 0.5109772086143494),
 ('skripten', 0.5085192322731018),
 ('honorarabrechnung', 0.5056849122047424),
 ('projektablauf', 0.4978453814983368),
 ('mahnwesen', 0.48487651348114014)]
```

Anwendungsbeispiel: Terminologie wird nicht erkannt

- Wie modellieren wir dieses Problem?

3 nicht erkannte Fehler!

Mit elektrischem Handrad, Tasten für Achs-Richtung, Vorschub, Eilgang, Not-Aus- und Zustimmtaste, sowie drei belegbaren Funktionstasten.

With electrical handwheel, keys for ~~**axial direction**~~ axis direction, feed, rapid traverse, ~~**Emergency button**~~ Emergency Stop button and acceptance key as well as three assignable ~~**functional keys**~~ function keys.

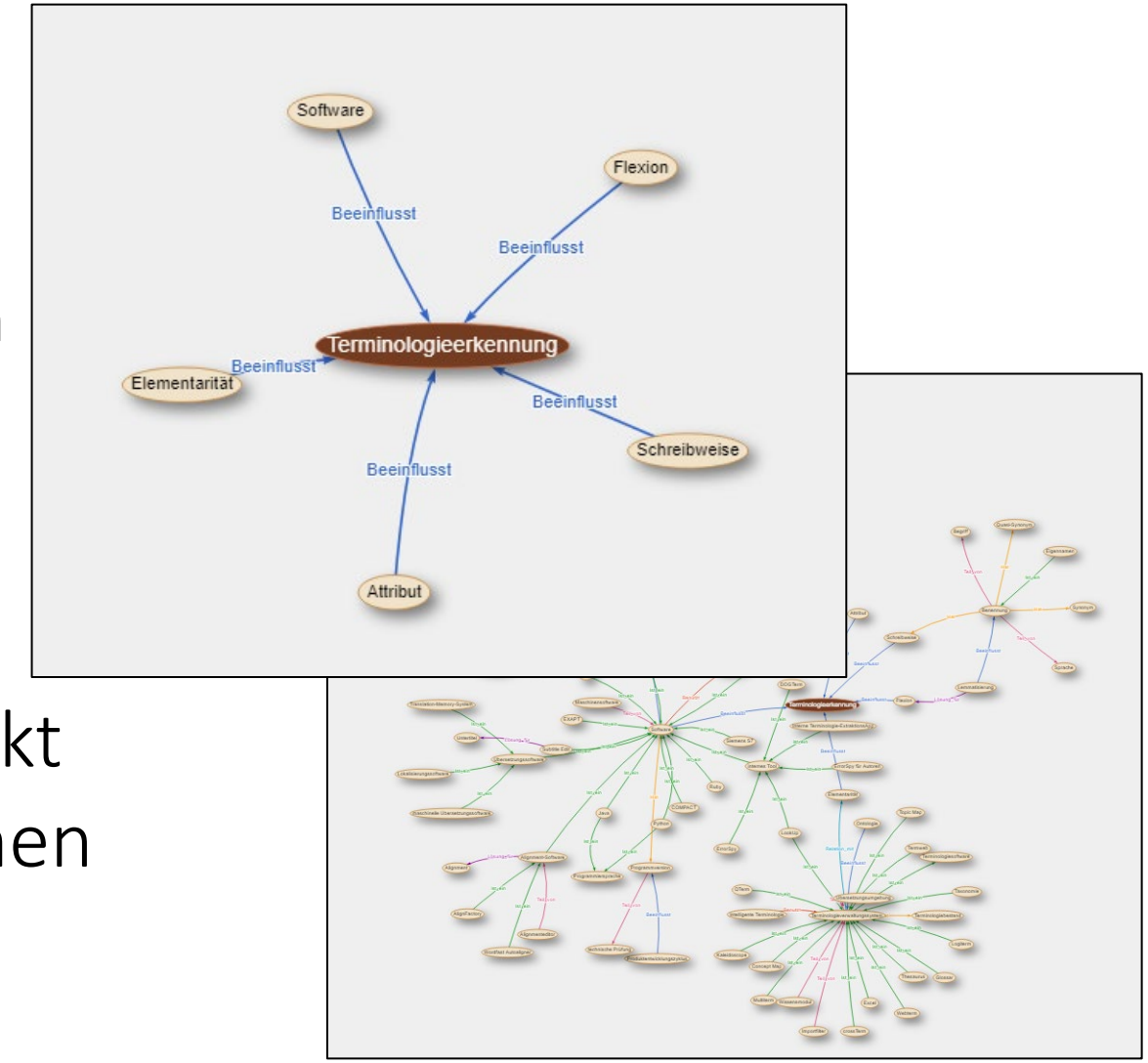
- Vorhanden und erkannt
- Vorhanden und nicht erkannt
- Nicht in der Terminologie erfasst

Eintrag in Terminologiedatenbank:

- Not-Aus-Taste
- Achsrichtung *(Bindestrich bitte prüfen)*

Ergebnis

- Benötigte **Begriffe**:
  - Einflussfaktoren für die Nicht-Erkennung einer Benennung durch eine Qualitätssicherungssoftware
- **Relationen**: Beeinflusst, Bestandsbeziehungen
- **Concept Map** ist nur als Startpunkt für weitere Recherche zu verstehen



Einige Herausforderungen

- **Richtung** der Relation ermitteln

Automatische Erkennung von Subjekt (blau) und Objekt (rot)?

Die ANSI-Norm ***verwendet*** *das Munsell-Farbsystem.*

Das RGB- oder das CMYK-Farbmodell wird in den europäischen Richtlinien ***verwendet****.*

- **Welche Wörter** stehen in Relation?

Automatische Erkennung verbundener Begriffe

Relation ***Dient_zu*** mit Muster: X (. . .) **mittels** (. . .) Y

Beschwerde [1] *können Sie* ***mittels*** *des Formulars* [2] *an die Zertifizierungsstelle* [3] *richten.*

Test und Validierung der Concept Maps

- Unterschiedliche Autoren und Aufgaben führen zu **Inkonsistenzen**, **Teilansichten** und **Momentaufnahmen**
- Typische Problemfälle:
  - In vergleichbaren Situationen werden **unterschiedliche Relationen** eingesetzt
  - **Neue Begriffe** fehlen bei einer Wissenseinheit
  - Relation **nicht präzise** genug (X *Relation_mit* Y) oder passt zur Fragestellung nicht.
  - **Startpunkt der Relation** bei Begriffshierarchien (Hat X oder der Oberbegriff von X eine Relation zu Y?)

Arbeit mit der Datenbank

- Die Wissensdatenbank wird unterschiedlich verwendet:

1. Für die **klassische Begriffssuche** (*Was bedeutet ….?*)
2. Für die **Lösung einzelner Aufgaben** (*Was verursacht XYZ ….?*)
3. Für die **thematische Aufbereitung einer Aufgabe**, z.B. komplexes Projekt: Themen können in einem Dokument in unterschiedlichen Farben markiert werden.

Erkenntnisse in Bezug auf Terminologie

- Bedarf nach neuen Begriffstypen:
  - **Strukturierungsbegriffe**: Software, Norm, Kostenkalkulation
    → Für Taxonomien, Hierarchien
  - **Eigennamen**: tekom, BDÜ, C++, msg (als Dateiendung)
  - **Firmeninterne Begriffe**: DOG-Termin (als Ergänzung zu Kundentermin)
- Schritt in Richtung Ontologie-Klassen

Arbeiten mit Relationen

- Aufgrund der **Vielfältigkeit der Themen** und Wissenseinheiten kann die Datenbank schnell **unübersichtlich** werden.
- Ausgehend von **einem** Begriff können **mehrere Wissenseinheiten** modelliert werden.
- Daher sind **Filtermöglichkeiten** sehr nützlich:
  - Filtern nach Begriffsattributen (Taxonomien). Z. B. Thema Vertrieb
  - Filtern nach Relationstyp

Was haben wir gelernt?

- **Planung** ist die halbe Miete.
- Größte Herausforderung = **Wissensaufbau standardisieren**:
  - Relationen beeinflusst durch persönliche Erfahrung + Vorwissen
  - Relationen beeinflusst durch Aufgabe/Fragestellung
- **Notwendigkeit**, **regelmäßig** Terminologie und Wissenseinheiten **zu aktualisieren**
- **Schulung**sbedarf
- Viel **Überzeugungsarbeit** intern erforderlich

Ausblick und Challenges

- Die Datenbank bringt bereits eine Erleichterung und Vorteile im Alltag, aber…
- Die **Motivation** der Mitarbeiter, die Datenbank systematisch zu nutzen und mitaufzubauen bleibt eine Herausforderung.
- Ausblick: Entwicklung in Richtung **ontologiefähiges System**.
- Das bedeutet:
  - Funktionen im TVS, die den Umgang mit Wissen unterstützen.
  - Relationen mit Attributen anreichern, u.a. um **SPARQL-Abfragen** und **Reasoning** (Schlussfolgern) zu ermöglichen.
  - Austauschformat auf **RDF** / OWL-Basis

Vielen Dank für Ihre Aufmerksamkeit!

Kontaktperson:

Dr. Francois Massion
E-Mail: francois.massion@dog- gmbh.de
Telefon: +49 (0)7152/35411-10

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback über den QR-Code