Bundesamt für
Kartographie und Geodäsie

Technische Erläuterungen zur Suchfunktion des Geoportal.de

Zusammenfassung

In diesem Dokument wird die technische Funktionsweise der Suche im Geoportal.de (www.geoportal.de) beschrieben. Insbesondere wird hierzu auf die verwendeten Felder bzw. Elemente aus den Metadaten eingegangen und die Kriterien zum Ranking (Sortierung) der Suchergebnisse erläutert.

Inhaltsverzeichnis

1 Suchfunktion im Geoportal.de 4
1.1 Generierung von Stichwörtern 4
1.2 Anmerkungen zur Bedienung 4
1.3 Sortierung der Suchergebnisse 5
1.4 Technische Details 6
2 Kontakt 8

1 Suchfunktion im Geoportal.de

1.1 Generierung von Stichwörtern

Die Volltextsuche arbeitet auf den Feldern:

- uuid
- title
- keywords
- abstract
- pointOfContact (nur der 1. Eintrag)
- responsibleParty (nur der 1. Eintrag)
- topiccategory (XML-Keywords)
- isoThemen (ausgeschriebene deutsche Variante)

Aus diesen Feldern werden Stichwörter für die Suche wie folgt extrahiert:

1. Jede Folge von Zahlen und Buchstaben ist ein Stichwort. Zusammengesetzte Begriffe können dabei auch in der Mitte mit Punkt (.), Bindestrich (-), Unterstrich (_) oder Doppelpunkt (:) verbunden werden (Sonderzeichen am Anfang oder Ende gehören allerdings nicht zum Wort)
2. Zusammengesetzte Begriffe werden getrennt und in verschiedenen Formen als Stichwörter aufgenommen, z.B. ergibt „Bebauungs-Plan“ die Stichwörter
   a. Bebauungs-Plan
   b. Bebauungsplan
   c. Bebauungs
   d. Plan
3. Unspezifische Stichwörter wie „mit“ oder „von“ (sogenannte Stoppwörter) werden entfernt, die vollständige Liste findet sich hier https://github.com/apache/lucene-solr/blob/master/lucene/analysis/common/src/resources/org/apache/lucene/analysis/snowball/german_stop.txt
4. Die verbleibenden Stichwörter werden in verschiedenen Variationen berücksichtigt, z.B. Bäume + Baum.

Analog dazu werden Stichwörter aus den eingegebenen Suchbegriffen generiert. Gibt es eine Überschneidung zwischen den Stichwörtern der Metadaten und denen der Suche ergibt dies einen Suchtreffer. Dabei genügt ein Wortbestandteil, d.h. eine Suche nach „Bau“ würde auch einen Eintrag mit Stichwort „Bebauungsplan“ finden. Umgekehrt liefert eine Suche nach „Bebauungsplan“ jedoch keine Ergebnisse, die nur „Bau“ enthalten.

1.2 Anmerkungen zur Bedienung

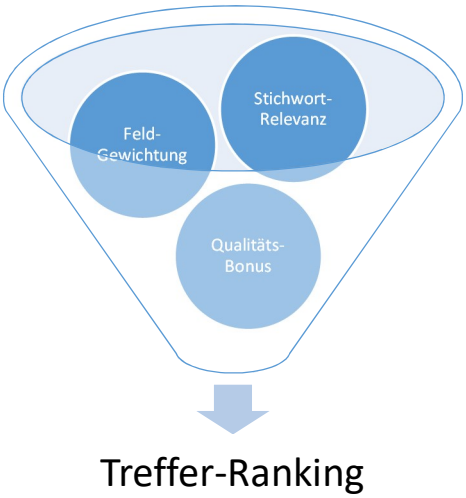
In der Suche können verschiedene Operatoren verwendet werden:

"Eine Phrase" in Anführungszeichen liefert nur Treffer, bei denen die einzelnen Wörter in genau dieser Reihenfolge vorkommen. Die oben beschriebenen Variationen der Schreibweise werden dabei nicht durchgeführt.

Der senkrechte Strich "|" kann verwendet werden um alternative Suchbegriffe anzugeben. „Umwelt Natur" liefert nur Ergebnisse in denen beide Wörter vorkommen, bei „Umwelt | Natur" hingegen genügt eines der beiden Stichwörter.

Das UUID-Feld wird beim Ranking der Ergebnisse besonders hoch gewichtet, damit die Eingabe einer UUID ins Suchfeld immer den Datensatz mit dieser UUID als ersten Treffer liefert.

1.3 Sortierung der Suchergebnisse



Zur Sortierung der Suchergebnisse werden die einzelnen Treffer bewertet. Die Bewertungsfunktion, die unter https://www.elastic.co/guide/en/elasticsearch/reference/7.13/similarity.html beschrieben ist, arbeitet nach zwei Grundsätzen:

1. Ein Treffer ist umso relevanter, je häufiger das gesuchte Stichwort in diesem Treffer vorkommt (sogenannte „term frequency")
2. Treffer zu einem bestimmten Stichwort sind umso relevanter, je seltener das darin gefundene Stichwort im Katalog insgesamt vorkommt (sogenannte „inverse document frequency"). Für die Reihenfolge der Treffer macht dies natürlich nur einen Unterschied, wenn überhaupt mehrere verschiedene Stichworte gesucht werden.

Zusammen sorgt dies dafür, dass bei einer Suche nach z.B. „Karte Bahnhöfe" solche Treffer als besonders relevant erkannt werden, die das seltene Wort „Bahnhöfe" mehrfach enthalten, während das Wort „Karte" ohnehin häufig vorkommt und daher als wenig spezifisch erkannt wird.

Die so ermittelte Relevanz wird noch mit einem Faktor zur Gewichtung multipliziert, je nachdem in welchem Metadatenfeld das Stichwort gefunden wurde:

Metadaten-Feld	**Faktor**
Titel	4
Schlüsselwörter	3
ISO-Thema, Kategorie	1

PointOfContact / ResponsibleParty	1
UUID	1000
Abstract	2

Hinzu wird ein Bonus addiert, der sich aus einer automatischen Qualitätskontrolle des Dienstes und der Metadaten ergibt:

Kriterium	**Bonus**
Titel ausgefüllt	20
Bounding Box ausgefüllt	20
Abstract ausgefüllt	20
Keywords ausgefüllt	20
PointOfContact ausgefüllt	20
Constraints ausgefüllt	20
Bild vorhanden	100
Dienst ist ansprechbar	200
Metadaten-QS-Tests ssbestanden	200

Dabei werden einige grundlegende Metadaten-Tests selbst durchgeführt, aber auch das Ergebnis der automatischen QS einbezogen.

1.4 Technische Details

Die Suche des Geoportal.de basiert auf der Open Source Software ElasticSearch (https://www.elastic.co/de/elasticsearch/). Nachfolgend werden einige technische Details exemplarisch anhand der Suchanfrage (Elasticsearch Query) „Karten Bahnhöfe SN“ erläutert:

```
{
        "query": {"bool":
                {"must": [{"simple_query_string": {
                        "query":"Karten Bahnhöfe SN",
                        "flags":"OR|PHRASE|WHITESPACE",
                        "analyzer":"default_search",
                        "quote_field_suffix":".phrase",
                        "fields":["title^4","keywords^3","topiccategory^1","abstract^2","isoThemen
                        ^1","pointOfContactName1^1","responsiblePartyName1^1","uuid^1000"],
                        "default_operator":"AND"
                }}],
                "should":[
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.title"}},
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.bbox"}},
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.abstract"}},
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.keywords"}},
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.contact"}},
```

```
                        {"rank_feature":{"boost":20,"field":"monitor_ranking.constraints"}},
                        {"rank_feature":{"boost":100,"field":"monitor_ranking.graphic"}},
                        {"rank_feature":{"boost":200,"field":"monitor_ranking.service_tests"}},
                        {"rank_feature":{"boost":200,"field":"monitor_ranking.metadata_tests"}}
                ],
                "filter":[]
        }},
        "aggs":{
                "language":{"terms":{"field":"language","size":3}},
                "service":{"terms":{"field":"service","size":2}},
                "resourcetype":{"terms":{"field":"resourcetype","size":20}},
                "keyword":{"terms":{"field":"keyword","size":2}},
                "inspireThemen":{"terms":{"field":"inspireThemen.keyword","size":100}},
                "isoThemen":{"terms":{"field":"isoThemen.keyword","size":100}},
                "datenanbieter":{"terms":{"field":"datenanbieter.keyword","size":5000}},
                "inspireumgesetzt":{"terms":{"field":"inspireumgesetzt","size":100}}
        },
        "from":0,
        "size":10,
        "track_total_hits":true
}
```

Die Stichworterzeugung liefert anschließend die Liste:

- karten, kart
- bahnhöfe, bahnhof, bahnhofe
- sachsen, sachs, SN

erzeugt mit

```
GET /metadata_all/_analyze
{
  "analyzer": "default_search",
  "text": ["Karten Bahnhöfe SN"]
}
```

Zum Vergleich würde ein Katalogeintrag mit dem Titel „Karte sächsischer Bahnhöfe“ folgende Stichwörter im Suchindex erzeugen. Die Treffer sind der Übersicht halber hervorgehoben:

karte, kar, **kart**, art, arte, rte, sächsischer, säc, säch, sächs, sächsi, sächsis, sächsisc, sächsisch, sächsische, äch, ächs, ächsi, ächsis, ächsisc, ächsisch, ächsische, ächsischer, chs, chsi, chsis, chsisc, chsisch, chsische, chsischer, hsi, hsis, hsisc, hsisch, hsische, hsischer, sis, sisc, sisch, sische, sischer, isc, isch, ische, ischer, sch, sche, scher, che, cher, her, sachsisch, sac, sach, **sachs**, sachsi, sachsis, sachsisc, ach, achs, achsi, achsis, achsisc, achsisch, sachsischer, sachsische, achsische, achsischer, **bahnhöfe**, bah, bahn, bahnh, bahnhö, bahnhöf, ahn, ahnh, ahnhö, ahnhöf, ahnhöfe, hnh, hnhö, hnhöf, hnhöfe, nhö, nhöf, nhöfe, höf, höfe, öfe, **bahnhof**, bahnho, ahnho, ahnhof, hnho, hnhof, nho, nhof, hof, bahnhofe, ahnhofe, hnhofe, nhofe, hofe, ofe

Da hier zu jedem Stichwort aus der Suchanfrage eine Entsprechung gefunden wurde, würde dieser Eintrag in den Suchergebnissen erscheinen.

2 Kontakt

Bundesamt für Kartographie und Geodäsie

Betrieb GDI-DE

Richard-Strauss-Allee 11

60598 Frankfurt am Main

E-Mail: bst@bkg.bund.de