Building a Better Search and Discovery Platform Using Elasticsearch

EDDI 2016-12-06, Cologne

Stefan Jakobsson  stefan.jakobsson@snd.gu.se
Olof Olsson  olof.olsson@snd.gu.se
Johan Fihn  johan.fihn@snd.gu.se
Pablo Millet  pablo.millet@snd.gu.se
About the presenter

Stefan Jakobsson
Systems Developer at Swedish National Data Service (SND)

SND
- Infrastructure for Swedish research within the humanities, social sciences and health sciences
- Helps researchers gain access to existing research data within and outside of Sweden (via online catalogue)
- Provides support and guidance to researchers throughout the entire research process
Search before Elasticsearch at SND

- Only text search, no filters
- List functions
- No support for geographical search
Search before Elasticsearch at SND

- Only textual listing of results
- No further filtering possible
Why rebuild the search platform?

- Implement facet filters
- Support geographical searches and visualization
- Support timeline searches and visualization
- Simplify the online catalogue technically
  - Remove a lot of database calls to internal systems
  - Simplified database model optimized for search, discovery and generating landing pages
New search: Start page

**The Northern Sweden MONICA Study**

The Northern Sweden MONICA Study started with the WHO MONICA Study in 1985. The purpose was to monitor trends in mortality and morbidity in cardiovascular disease and relate them to trends in risk factors. Since 1985...

- Stefan Söderberg, Umeå University, Department of Public Health and Clinical Medicine
- Ann-Sofie Forslund, Norrbotten County Council, Department of Research

**NSHDS-VIP**

The cohort is population-based and consists of blood and data from primarily 40, 50 and 60 year olds, taken every year in this age group in connection with the Västerbotten health surveys from 1985 - present. The data...

- Göran Hallmans, Umeå University, Department of Biobank Research

**Northern Sweden Diet Database (NSDD)**

NSDD is a database with harmonised data from two population based projects “Västerbotten Intervention Project”
New search: Filtering example, by keyword

GIS-material for the archaeological project: Settlement near the medieval village site of Kallerstad
Unavailable
- Swedish National Heritage Board, UV Öst

GIS-material for the archaeological project: Södra Änggård
The information in the abstract is translated from the archaeological report: The Swedish National Heritage Board's Contract Archaeology Service, UV Öst, carried out a field evaluation in August 2002, and a final exca...
- Swedish National Heritage Board, UV Öst
New search: Geographic visualization and filtering

Bounding box is defined by panning/zooming the map and is used as filter in query

Displays heat map of results if results > 100
New search: Geographic visualization and filtering, zoomed

Displays pins of individual results if results < 100
New landing pages: Overview
/catalogue/study/SND0805

Principal investigator(s):
Stefan Svallfors - Umeå University, Department of Sociology
Jonas Edlund - Umeå University, Department of Sociology

Abstract:
This is the Swedish part of the 2003 'International Social Survey Program' (ISSP), and it is the second time ISSP focuses on national identity. The questions on national consciousness and national identity include: identification with the town, the city, the region, the nation and with the respective continent; most important characteristics for national identity; identification with one's own nation and national pride (scale); perceived pride in the democracy of the country, the political influence of the country in the world, the economic achievement, the social security system, the scientific achievements, the achievements in sports, the achievements in arts or literature, the armed forces, the history and equal rights of all social groups in society; preference for protective duty to support the national economy; attitude to the right of international institutions to enforce solutions to be accepted nationally; attitude to enforcing national interests regardless of evoking conflicts with other countries; rejection of acquisition of land by foreigners in one's country; preference for national

... more

Keywords:
trade policy, international relations, immigrants, community identification, minorities, national cultures, national identity

Availability status:
1b - Freely available via ordering or Open data 1

Homepage:
The Swedish ISSP surveys at Umeå University
New landing pages: DOI resolved to specified dataset version
/catalogue/study/SND0805/001/1.1

ISSP 2003 - National identity II: Sweden

Citation:

Kind of data:
Survey data

Collection events:
2003-02 - 2003-04 (Fixed form self-administered questionnaire: paper)

Variables:
114

Number of individuals/objects:
1186

Response rate:
61.9%

Languages:
Data: Swedish
Frågeformulär: Swedish

You have requested the latest version of this dataset

Version 1.0:
2005-09-23 doi:10.5878/002399

Version 1.1:
2009-08-28 doi:10.5878/002400

Version Original:
2004-02-12
Technology overview

Internal metadata database (SQL)

Elasticsearch

SND web Drupal

Clients & API

Public REST API (HTTP)

JSON documents via REST API (HTTP)

Harvested metadata

metadata export

ingest of metadata from other organizations

ingest of metadata from other organizations

metadata export
What is Elasticsearch?

- Distributed search and analytics engine
- Supports full text search, geographical search and more
- Document-oriented, uses schema-free JSON documents
- Scalable, built from the ground up to be distributed
- Open source
- Based on Apache Lucene (text search engine)
- Cross platform (written in Java)
Why Elasticsearch?

- Prior experience, we know that it works well for our use case
- Built on Lucene which is one of the best full text search engines available
  - Lucene is very complex, Elasticsearch hides the complexity behind a coherent interface
  - Adds features on top of Lucene, for example geographical support
- Scalable if needed: Elasticsearch is distributed by default and built for large scale scenarios
- Elasticsearch is a popular choice with a lot of traction, which means there is a large community and lots of help/tips online
Mappings in Elasticsearch

- Defines how a document and its fields are stored and indexed
- Fields can be mapped to different data types e.g. full text, integer, date, geographical etc.
- Boost: tailor a field’s impact on search score

Full text mapping
- Uses analyzer to tokenize, remove stop words, stem into root forms etc.
- Many analyzers tailored for specific languages available

Geographical mapping
- Geo points
- Geo shapes (e.g. polygons)
Mappings, examples

```
"title": {
    "type": "text",
    "analyzer": "english",
    "boost": 2
}
```
- Type “text” means full text indexed
- Uses “english” analyzer
- Boosted to twice the default weight

```
"identifier": {
    "type": "keyword"
}
```
- Type “keyword” is for content that is not analyzed
- Typically for fields that are used for filtering, sorting and aggregations

```
"polygon": {
    "type": "geo_shape",
    "precision": "10m"
}
```
- Type “geo_shape” for arbitrary geo shapes e.g. polygons
- Choice of precision is important, big impact on index size
SND has developed an open source example mapping for studies and datasets (available on Bitbucket, link on last page of presentation)

Many fields map to DDI equivalents, examples below

<table>
<thead>
<tr>
<th>DDI</th>
<th>Elasticsearch mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>StudyUnit/KindOfData</td>
<td>&quot;kindofdata&quot;: {</td>
</tr>
<tr>
<td></td>
<td>&quot;type&quot;: &quot;text&quot;,</td>
</tr>
<tr>
<td></td>
<td>&quot;analyzer&quot;: &quot;english&quot;,</td>
</tr>
<tr>
<td></td>
<td>&quot;fields&quot;: {</td>
</tr>
<tr>
<td></td>
<td>&quot;raw&quot;: {</td>
</tr>
<tr>
<td></td>
<td>&quot;type&quot;: &quot;keyword&quot;</td>
</tr>
<tr>
<td></td>
<td>}</td>
</tr>
<tr>
<td></td>
<td>}</td>
</tr>
<tr>
<td>StudyUnit/Coverage/SpatialCoverage/BoundingBox</td>
<td>&quot;boundingbox&quot;: {</td>
</tr>
<tr>
<td></td>
<td>&quot;type&quot;: &quot;geo_shape&quot;,</td>
</tr>
<tr>
<td></td>
<td>&quot;precision&quot;: &quot;25m&quot;</td>
</tr>
<tr>
<td></td>
<td>}</td>
</tr>
</tbody>
</table>
Simple full text search query

GET /sndcatalogue-sv/_search

```
{
    "query": {
        "simple_query_string": {
            "query": "swedish"
        }
    }
}
```

full text search string matching indexed fields

Each hit is scored
Filter by term example

[Image of a filter interface with the term "mass media" selected]
Filter by term example query

```json
{
    "query": {
        "bool": {
            "must": {
                "match_all": {}}
            },
        "filter": [
            {
                "term": {
                    "keyword.value.raw": "mass media"
                }
            }
        ]
    }
}
```

default match all; no search string given

filter must match, multiple filters are allowed
Filter by term example query, combined with text search

```json
{
    "query": {
        "bool": {
            "must": {
                "simple_query_string": {
                    "query": "swedish"
                }
            },
            "filter": [
                {
                    "term": {
                        "keyword.value.raw": "mass media"
                    }
                }
            ]
        }
    }
}
```

- Full text search string matching indexed fields
- Filter must match, multiple filters are allowed
Aggregations

- Provides aggregated data based on a search query
- Aggregations can be composed to build complex summaries

**Used aggregations at SND**

- Term aggregations
  - For displaying facet filters (available terms and frequency)
- Geohash grid aggregations
  - For displaying geographical heat map of search results
- Geo bounds aggregations
  - For finding bounding box of search results

Lots more available...
Terms aggregation example

How is “keyword” facet terms and frequency calculated?
Terms aggregation example

GET /sndcatalogue-en/_search

```json
{
    "query": {
        "bool": {
            "must": {
                "match_all": {}
            }
        }
    },
    "aggregations": {
        "keyword": {
            "terms": {
                "field": "keyword.value.raw"
            }
        }
    }
}
```

Aggregated into buckets with count for each

```json
{
    "took": 1109,
    "timed_out": false,
    "_shards": {
        "total": 5,
        "successful": 5,
        "failed": 0
    },
    "hits": {
        "aggregations": {
            "keyword": {
                "doc_count_error_upper_bound": 15,
                "sum_other_doc_count": 1024,
                "buckets": [
                    {
                        "key": "mass media",
                        "doc_count": 73
                    },
                    {
                        "key": "television news",
                        "doc_count": 30
                    },
                    {
                        "key": "elections",
                        "doc_count": 28
                    }
                ]
            }
        }
    }
}
```
Future plans

- Timeline visualization and filtering
- Question and variable search
- More like this (on study landing page)
  - Thematically (e.g. based on subjects, keywords, title)
  - Geographically
Links

Public repository containing mappings used at SND, 41 example documents, example queries etc [https://bitbucket.org/swedishnationaldataservice/elasticsearch-mappings](https://bitbucket.org/swedishnationaldataservice/elasticsearch-mappings)

You can use docker to run the examples and import the example data, script and instructions in the bitbucket repository

Questions, collaborations? Contact us!
Stefan Jakobsson  stefan.jakobsson@snd.gu.se
Olof Olsson  olof.olsson@snd.gu.se
Johan Fihn  johan.fihn@snd.gu.se
Pablo Millet  pablo.millet@snd.gu.se