Programming with MTAS via its broker interface

Jan Pieter Kunst, Meertens Institute
janpieter.kunst@meertens.knaw.nl

Meertens Instituut

The broker

- The broker is a JSON-based interface to Solr; it also supports MTAS-specific queries
- It expects a POST request with a JSON structure in the body, representing a query; it returns a JSON response
- The Nederlab project uses MTAS-enhanced Solr as its backend. The end user interface (*onderzoeksportaal* = research portal) communicates with this backend via its broker interface
- What follows is a description of Solr/MTAS broker usage in the Nederlab project

Meertens Instituut

Basic example of a broker request/response

Broker URL: www.nederlab.nl/broker3/search/

request:

```
{
  "condition": {
    "type": "equals",
    "field": "NLCore_NLIdentification_nederlabID",
    "value": "81421f52-011b-11e4-b0ff-51bcbd7c379f"
  },
  "response": {
    "documents": {
      "fields": [
        "NLPerson_NLPersonName_fullName"
      ]
    }
  }
}
```

response:

```
{
  "status": "ok",
  "documents": [
    {
      "NLPerson_NLPersonName_fullName": [
        "P.J. Meertens"
      ]
    }
  ]
}
```

Meertens Instituut

Broker usage

- Query syntax and field names are documented in the web interface of the broker
- The broker web interface (especially the example section) is a major part of its usefulness

http://www.nederlab.nl/broker3/

Meertens Instituut

Broker usage: joins (1)

- By using "joins", SQL-like functionality is emulated for Solr queries. The following response (which requests the title of a publication and its authors) translates to two different Solr queries on the backend, the results of which are combined before sending the response. This is transparent to the user of the broker.

```
{
  "response": {
    "documents": {
      "fields": [
        {
          "name": "title_authorinfo",
          "join": {
            "from": "NLTitle_NLPersonRef_personID",
            "to": "NLCore_NLIdentification_nederlabID"
          },
          "fields": [
            "NLPerson_NLPersonName_preferredFullName"
          ]
        },
        "NLTitle_title"
      ]
    }
  }
}
```

Meertens Instituut

Broker usage: joins (2)

- joins can also be used in the filter/condition part of a query. This returns titles by authors with last name “Meertens”.

```
{
  "filter": {
    "condition": {
      "type": "equals",
      "field": "NLPerson_NLPersonName_lastName",
      "value": "Meertens"
    },
    "join": {
      "from": "NLCore_NLIdentification_nederlabID",
      "to": "NLTitle_NLPersonRef_personID"
    }
  }
}
```

Many other “join” possibilities exist, e.g. search for titles by authors born in some location, search for articles from some journal, etc.

Meertens Instituut

Broker usage: Lexicon service

- The broker can use external web services for query expansion. In Nederlab we use the Lexicon service of the *instituut voor de Nederlandse taal* (Institute for the Dutch language) to expand a query to also provide variants of words. This looks like this in a query:

```
"condition": {
  "type": "or",
  "list": [
    {
      "type": "cql",
      "field": "NLContent_mtas",
      "value": "[t_lc=\"koe\"]"
    },
    {
      "type": "cql",
      "field": "NLContent_mtas",
      "value": "[t_lc=$1]",
      "variables": {
        "lexicon": {
          "$1": { "word": "koe" }
        },
        "stats": {
          "key": "lexicon0",
          "type": "sum"
        }
      }
    }
  ]
}
```

The querying of the external service and feeding its results to MTAS is done transparantly by the broker.

Meertens Instituut

Broker usage: caching and collections

- The broker provides a query cache. When a query has `{ "cache": true }` its results are cached so that subsequent requests of the same query are much faster.
- The broker can save intermediate results (collections) which can be referred to and used in subsequent queries. This mechanism is not yet used in Nederlab, but we will probably do so in the future.

Meertens Instituut

Broker usage: MTAS-specific (1)

Keyword in context for found documents:

Provided for by the following in the response part of the query:


```
"mtas": {
  "kwic": [
    {
      "field": "NLContent_mtas",
      "query": {
        "type": "cql",
        "value": "[t_lc=\"koe\"]"
      },
      "output": "hit",
      "prefix": "t,lemma, pos, [... much more]",
      "number": 3,
      "start": 0,
      "left": 10,
      "right": 10,
      "key": "[t_lc=\"koe\"]0"
    }
  ]
}
```


Meertens Instituut

Broker usage: MTAS-specific (2)

Full text (annotated) for some document:

Goedkeuring van de Notulen

De Voorzitter

De Notulen	van	de	vergadering	van	gisteren	zijn rondgedeeld .
	lemma: van pos: VZ feat.vztype: init	lemma: de pos: LID feat.naamval: stan feat.npagr: rest feat.lwtype: bep	lemma: vergadering pos: N feat.ntype: soort feat.getal: ev feat.graad: basis feat.genus: zijd feat.naamval: stan		lemma: gisteren pos: BW	

Geen bezwaren ?

Cornelissen

Mevrouw de Voorzitter , collegae , gisteravond heeft een ernstig vliegtuigongeval plaastgevon
plaats waar ik woon . Bij dit ongeval zijn 32 doden en 9 gewonden te betreuren . Vele honderde
gedompeld . Namens de Commissie vervoer en toerisme en namens de Nederlandse en Belgis
mevrouw de Voorzitter , het medeleven van het Europees Parlement met de slachtoffers en de
omgekomen slachtoffers over te brengen . Ik zal vandaag met commissaris Kinnock overleggen
Commissie wordt deelgenomen in het onderzoek naar de oorzaak van het ongeval want het is

Meertens Instituut

Broker usage: MTAS-specific (2)

Provided for by the following in the response part of the query:


```
"mtas": {
   "kwic": [
     {
       "field": "NLContent_mtas",
       "query": {
         "type": "cql",
         "value": "[]"
       },
       "key": "tekst",
       "output": "token",
       "prefix": "t,lemma,pos,entity,feat.tokentype,feat.pos, [... much more]",
       "number": 1,
       "start": 0,
       "left": 0,
       "right": 500
     },
     {
       "field": "NLContent_mtas",
       "query": {
         "type": "cql",
         "value": "[]"
       },
       "key": "structuur",
       "output": "token",
       "prefix": "p,head,s",
       "number": 1,
       "start": 0,
       "left": 0,
       "right": 500
     }
   ]
 }
}
```


Meertens Instituut

Broker usage: MTAS-specific (3)

Statistics for a query result:

Statistieken

4.504 hits, gevonden in **2.965** documenten
voor CQL query: **[t_lc="koe"]**

matchende documenten

maximum aantal woorden: **309.987**
minimum aantal woorden: **2**
gemiddeld aantal woorden: **5.041,55**
totaal aantal woorden: **14.948.188**

hitstatistieken

maximum aantal hits per document: **20**
minimum aantal hits per document: **1**
gemiddeld aantal hits per document: **1,52**
totaal aantal hits: **4.504**

Meertens Instituut

Broker usage: MTAS-specific (3)

Provided for by the following in the response part of the query:


```
"mtas": {
   "stats": {
     "positions": [
       {
         "field": "NLContent_mtas",
         "key": "totaal",
         "minimum": 1,
         "type": "n,sum,mean,min,max"
       }
     ],
     "spans": [
       {
         "field": "NLContent_mtas",
         "queries": [
           {
             "type": "cql",
             "value": "[t_lc=\"koe\"]"
           }
         ],
         "key": "[t_lc=\"koe\"]",
         "minimum": 1,
         "type": "n,sum,mean,min,max"
       }
     ]
   }
 }
}
```


Meertens Instituut

Broker usage: MTAS-specific (4)

Word frequency list for a query result:

1 2 3 > >>

token		som	documenten	gem. per document	maximum
1.	de	669.294 (4,477%)	2.833	236	19.232
2.	het	360.457 (2,411%)	2.748	131	9.472
3.	van	337.379 (2,257%)	2.716	124	10.907
4.	en	333.659 (2,232%)	2.768	121	8.447
5.	een	316.328 (2,116%)	2.843	111	6.456
6.	in	234.689 (1,570%)	2.696	87	8.237
7.	ik	215.255 (1,440%)	1.649	131	6.505
8.	dat	212.208 (1,420%)	2.487	85	5.147
9.	te	157.879 (1,056%)	2.420	65	5.126
10.	op	139.896 (0,936%)	2.521	55	3.207
11.	is	138.244 (0,925%)	2.634	52	3.112

Meertens Instituut

Broker usage: MTAS-specific (4)

Provided for by the following in the response part of the query:

```
"mtas": {
  "termvector": [
    {
      "field": "NLContent_mtas",
      "key": "aantal",
      "prefix": "t_lc",
      "type": "n,sum,mean,max,median,min",
      "sort": {
        "type": "sum"
      },
      "regexp": "[a-z]+",
      "number": 100
    }
  ],
  "stats": {
    "positions": [
      {
        "field": "NLContent_mtas",
        "key": "totaal",
        "minimum": 1,
        "type": "sum"
      }
    ]
  }
}
```

Meertens Instituut

Broker usage: MTAS-specific (5)

Grouped query result (group by first word left from the hit):

4.504 hits, gevonden in **2.965** documenten
voor CQL query: **[t_lc="koe"]**

query: **[t_lc="koe"]** (1-9 van 9 items)

1.	een	**koe**	**4.494** hits (99,778%) in **2.955** documenten (99,663%)
2.	de	**koe**	**4.488** hits (99,645%) in **2.951** documenten (99,528%)
3.	heilige	**koe**	**4.488** hits (99,645%) in **2.951** documenten (99,528%)
4.	bonte	**koe**	**1** hit (0,022%) in **1** document (0,034%)
5.	gouden	**koe**	**1** hit (0,022%) in **1** document (0,034%)
6.	drinkende	**koe**	**1** hit (0,022%) in **1** document (0,034%)
7.	europese	**koe**	**1** hit (0,022%) in **1** document (0,034%)
8.	oude	**koe**	**1** hit (0,022%) in **1** document (0,034%)
9.	loslopende	**koe**	**1** hit (0,022%) in **1** document (0,034%)

Meertens Instituut

Broker usage: MTAS-specific (5)

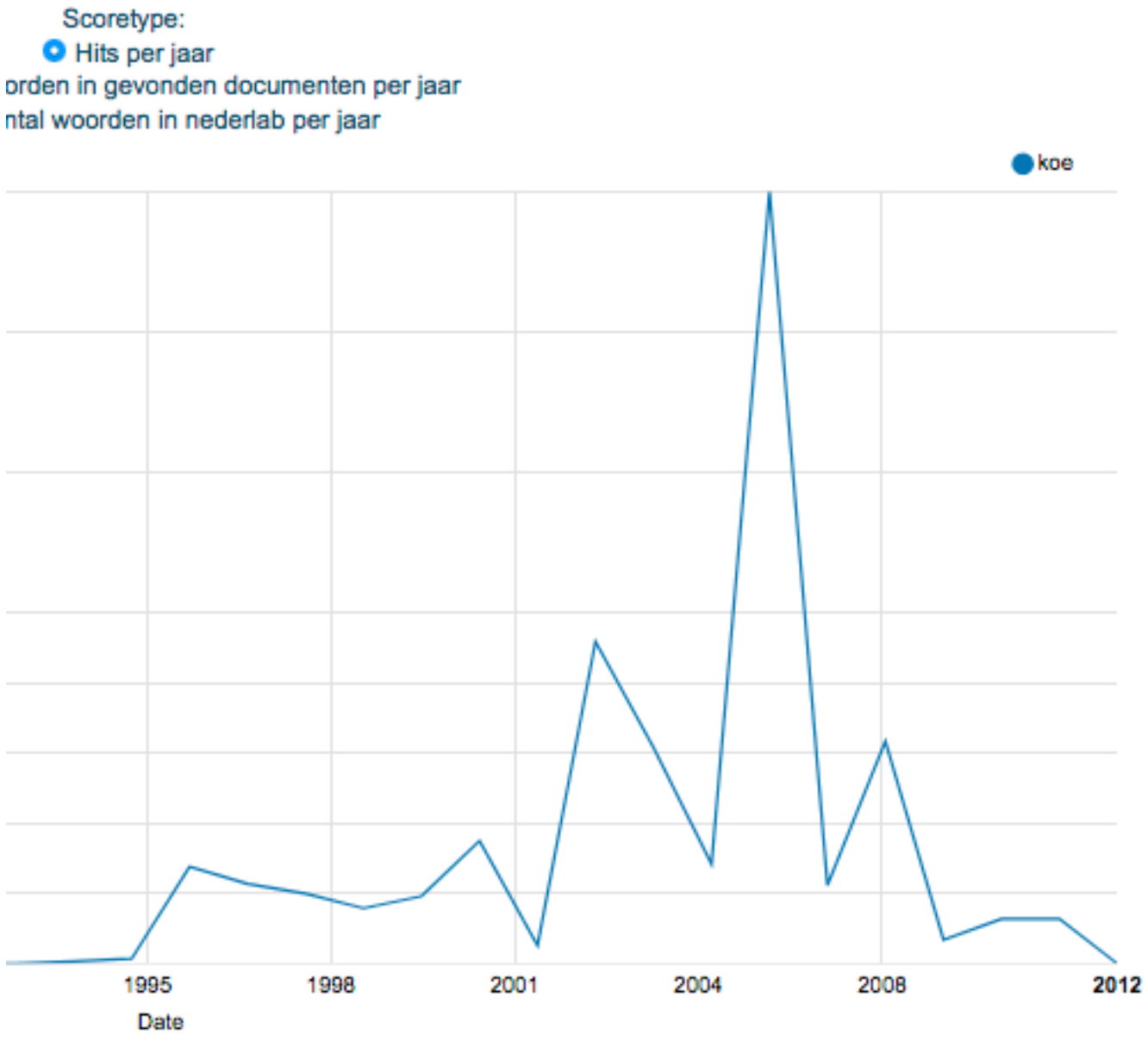
Provided for by the following in the response part of the query:


```
"mtas": {
  "group": [
    {
      "field": "NLContent_mtas",
      "query": {
        "type": "cql",
        "value": "[t_lc=\"koe\"]"
      },
      "grouping": {
        "hit": { "inside": "t_lc" },
        "left": { "0": "t_lc" }
      },
      "number": 25,
      "start": 0,
      "key": "[t_lc=\"koe\"]"
    }
  ],
  "stats": {
    "spans": [
      {
        "field": "NLContent_mtas",
        "queries": [
          {
            "type": "cql",
            "value": "[t_lc=\"koe\"]"
          }
        ],
        "key": "[t_lc=\"koe\"]",
        "minimum": 1,
        "type": "n,sum"
      }
    ]
  }
}
```


Meertens Instituut

Broker usage: MTAS-specific (6)

Visualisation: hits on a timeline



Meertens Instituut

Broker usage: MTAS-specific (6)

Provided for by the following in the response part of the query:

```
"mtas": {
  "facet": [{
      "field": "NLContent_mtas",
      "key": "nwordsmain",
      "queries": [ {
          "type": "cql",
          "value": "[]"
        }],
      "base": [{
          "field": "NLTitle_yearOfPublicationMin",
          "type": "sum",
          "sort": {
            "type": "term",
            "direction": "asc"
          }, "number": 2000 }]
    },
    {
      "field": "NLContent_mtas",
      "key": "koe",
      "queries": [{
          "type": "cql",
          "value": "[t_lc=\"koe\"]"
        }],
      "base": [{
          "field": "NLTitle_yearOfPublicationMin",
          "type": "sum,n",
          "number": 2000,
          "minimum": 1,
          "functions": [{
              "key": "nwords",
              "expression": "$n",
              "type": "sum"
            }]}
      ]}]
}
```

Meertens Instituut

Nederlab-specific practices in broker use

- Most broker access is done with Javascript in the browser
- We don't access the broker URL directly from the user's browser; we use a proxy script on our server which talks to the broker
- The main reason for this is copyright issues. If the user could access the broker directly, they could download full texts which we are not allowed to distribute. Our proxy script filters the broker output to comply with rules from our data providers, taking into account if a user is logged in or not
- Another reason to use a proxy script is to prevent cross-site origin problems in the browser: the proxy script is served from the same host as the rest of the application, while the broker might not be

Meertens Instituut

Thank you for your attention
Bedankt voor uw aandacht
Vielen Dank für Ihre Aufmerksamkeit

Meertens Instituut