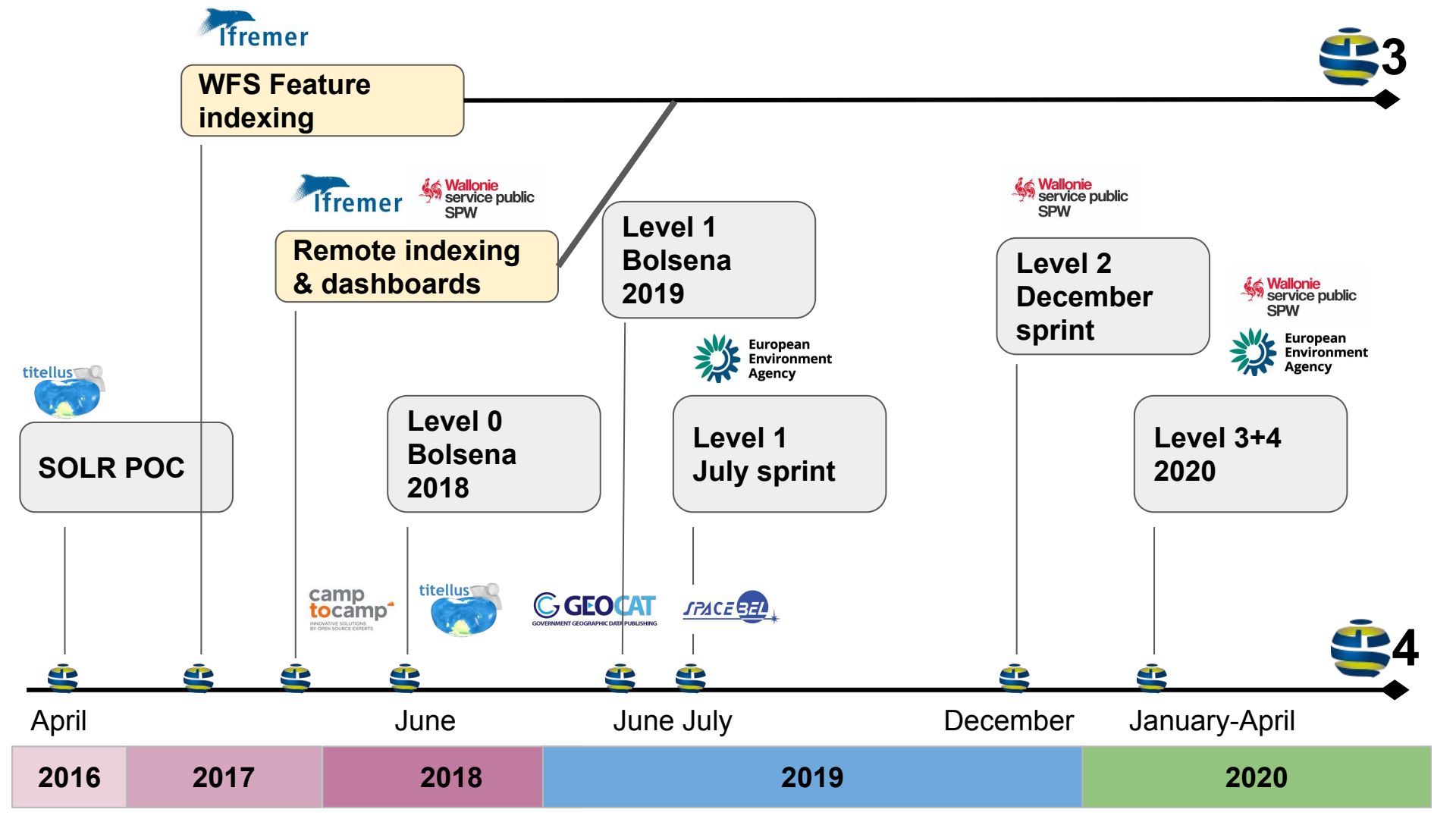
GeoNetwork 4

User expectations?
It is all about search ...

camptocamp
INNOVATIVE SOLUTIONS
BY OPEN SOURCE EXPERTS

GEOCAT
GOVERNMENT GEOGRAPHIC DATA PUBLISHING

titellus


Ifremer
WFS Feature indexing
3
Ifremer
Wallonie service public SPW
Remote indexing & dashboards
Level 1 Bolsena 2019
Wallonie service public SPW
Level 2 December sprint
Wallonie service public SPW
European Environment Agency
European Environment Agency
titellus
SOLR POC
Level 0 Bolsena 2018
Level 1 July sprint
Level 3+4 2020
camp tocamp
titellus
GEOCAT
SPACEBEL
4
April
June
June July
December
January-April
2016
2017
2018
2019
2020

GeoNetwork 4

= GeoNetwork 3.10.3 (release planned tomorrow)

+ Elasticsearch (instead of Lucene)

- Some features are not available

User interface is the same

Database is the same

Is search faster?

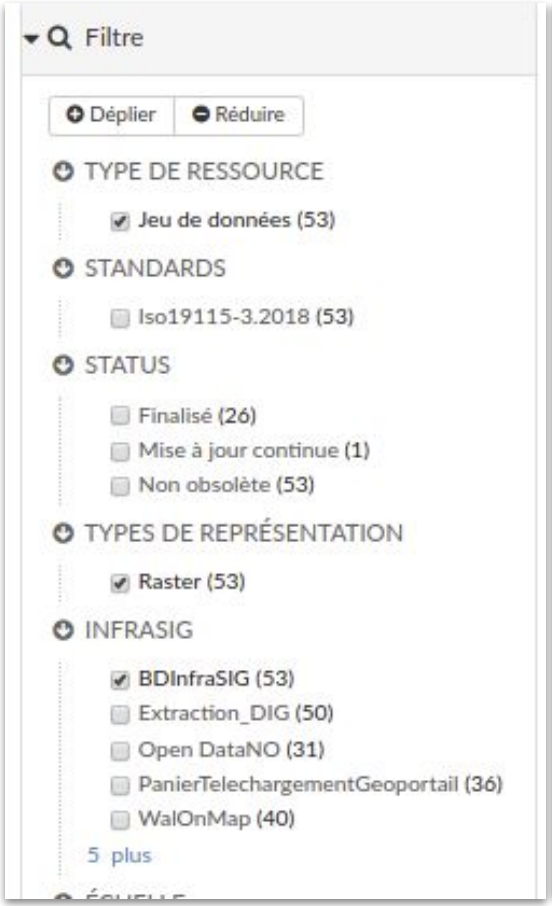
Simple scenario

Load search page

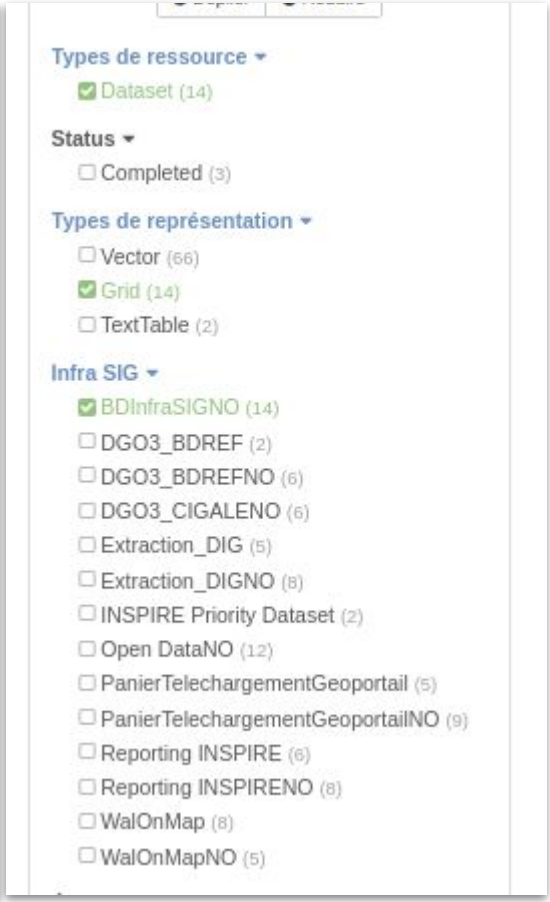
3 clicks on facets

- “Datasets”
- Then “BDinfraSIG”
- Then “Grid”

3



4



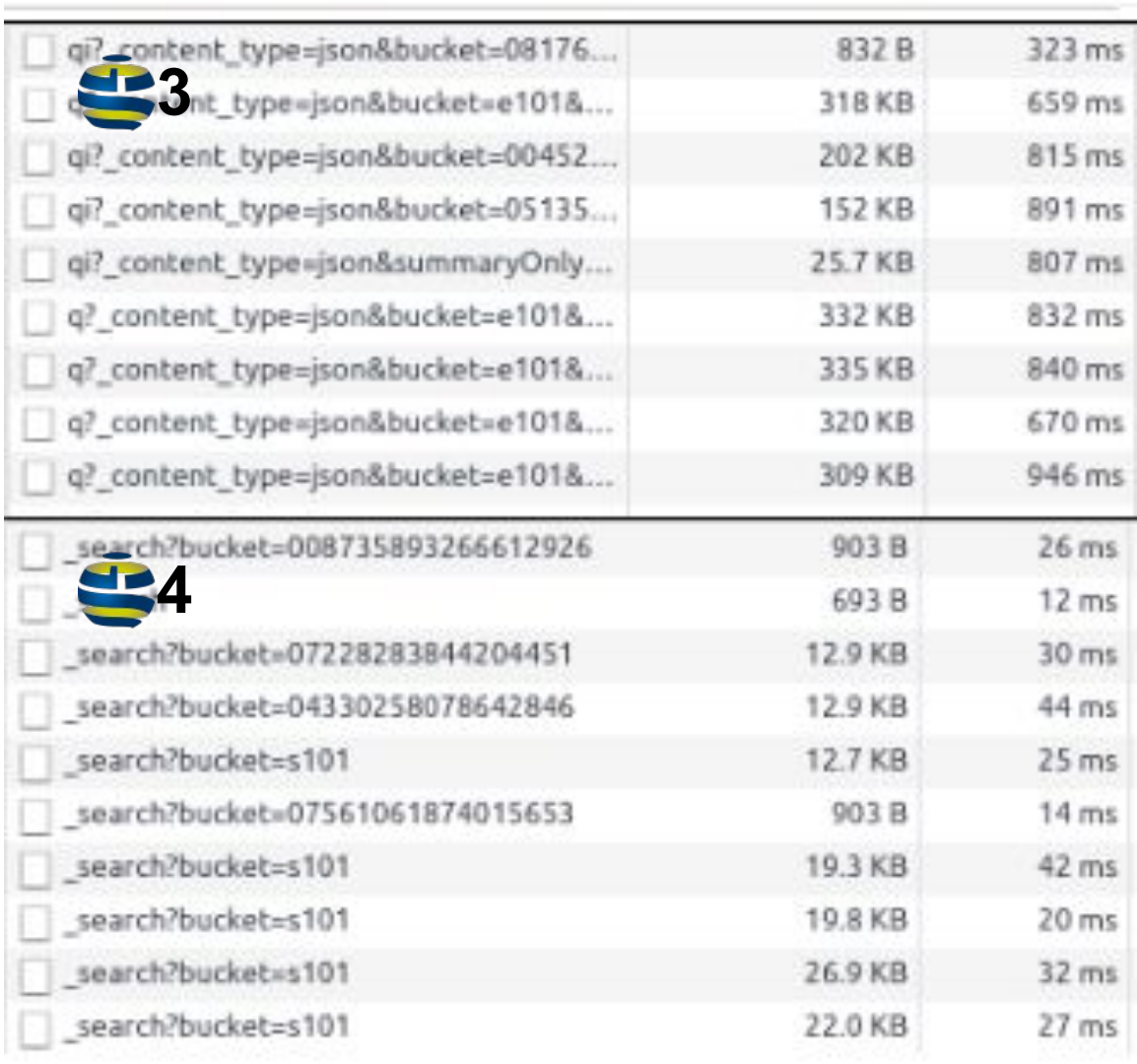
Search sequence =

Search is

minimum **10** time faster

Response size is

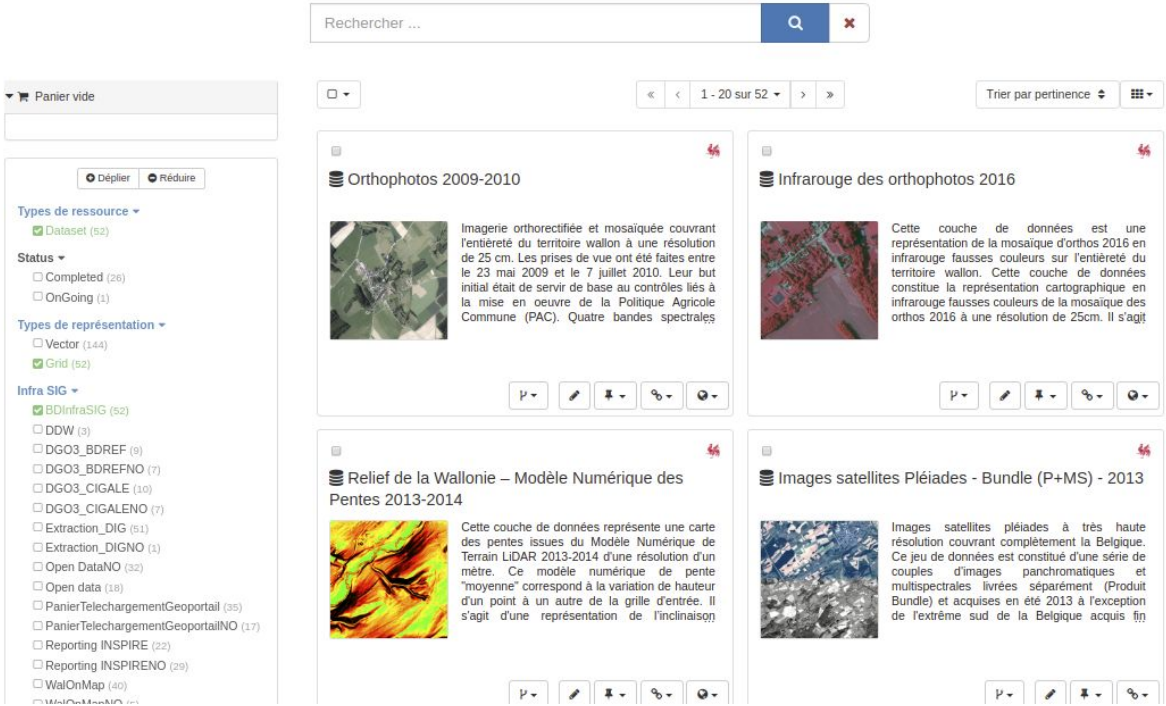
10 time smaller



Search is fast … but rendering is still “slow”

- Related records
- Angular one time binding
- Watchers
- Facet tree

More about that in
the UX/UI session



Why is the response size smaller?

3

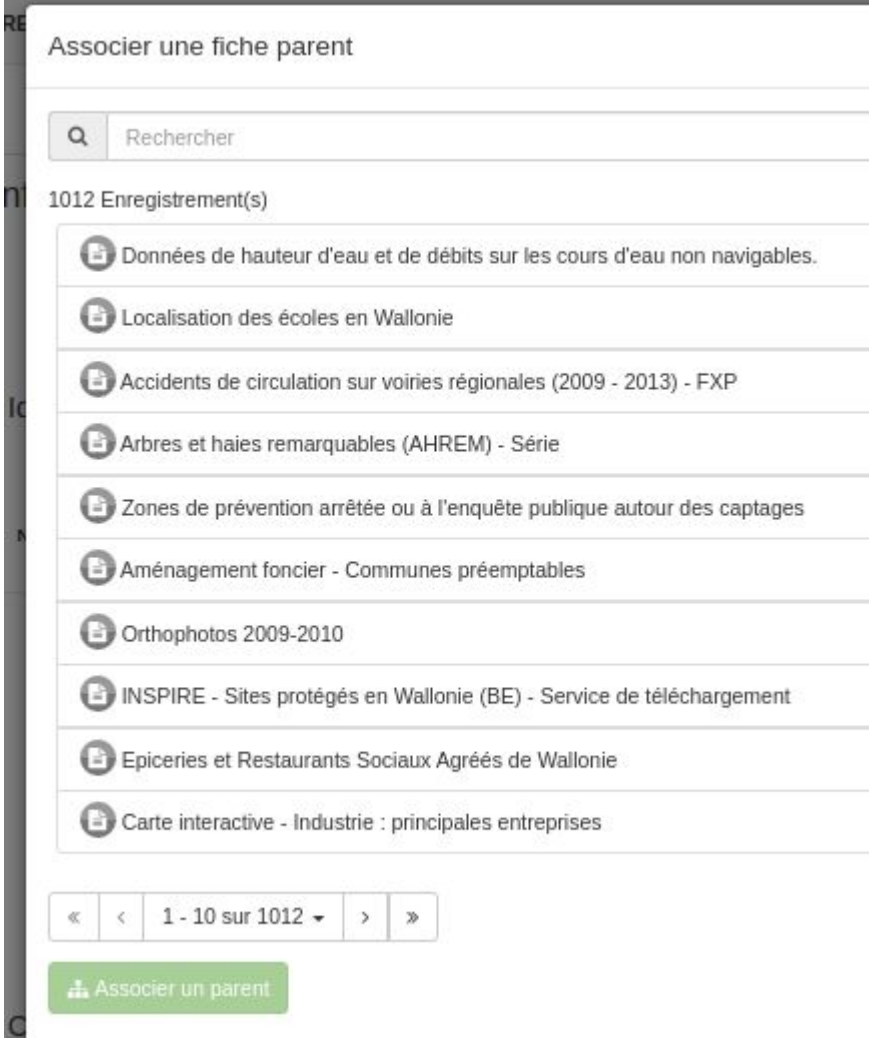
In GeoNetwork 3, search response size is always the same while searching with the main search page or searching a contact in the directory in the editor …

Only option to limit the response is to define: facet or facet & results.

JSON format is mapped from legacy XML responses and is verbose.

In Elasticsearch API the request defined which elements are returned in the response

Eg. we only need the title for making a list of record when picking up a parent record in the associated panel



Display similar records?

In GeoNetwork 3, users are encoding lots of relations eg. dataset/service, parents, sources

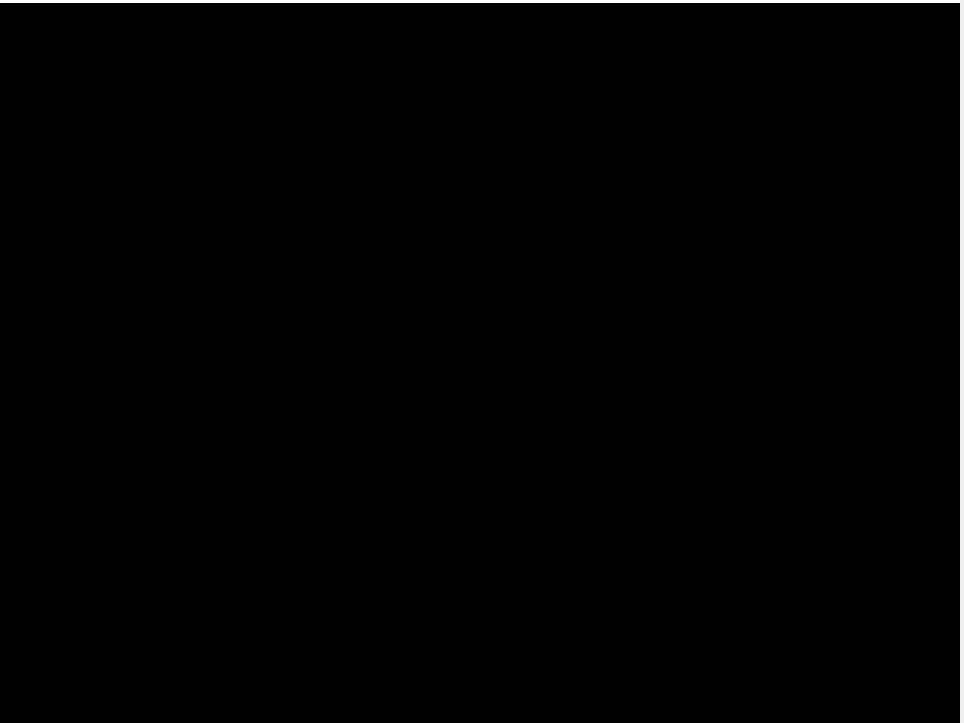
It would be interesting to also display similar records based on thematic, geographical and temporal similarities

More like this

The proximity is computed on the field you configure

Here title and keywords.

Autocomplete has unexpected results



3

In GeoNetwork 3, suggestions search on field value and if none found, search for title.

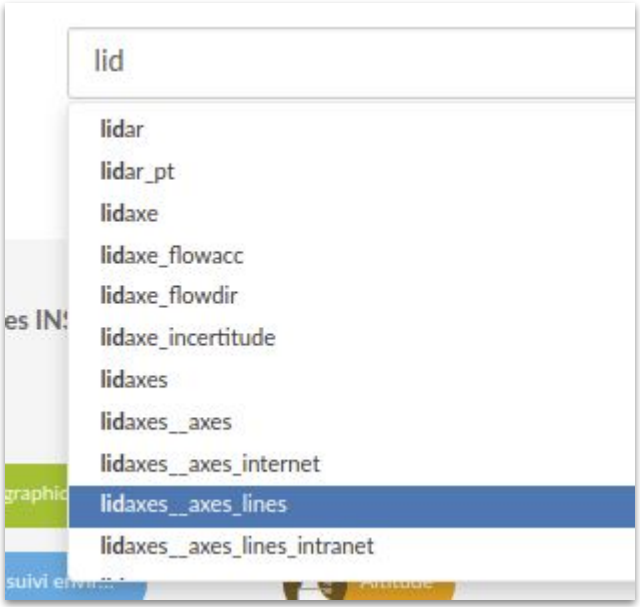
Suggestion does not take privileges into account.

Suggestion on more than one work does not work.

GN4 suggest a field (eg. title) searching on all content

(taking privileges into consideration)

3



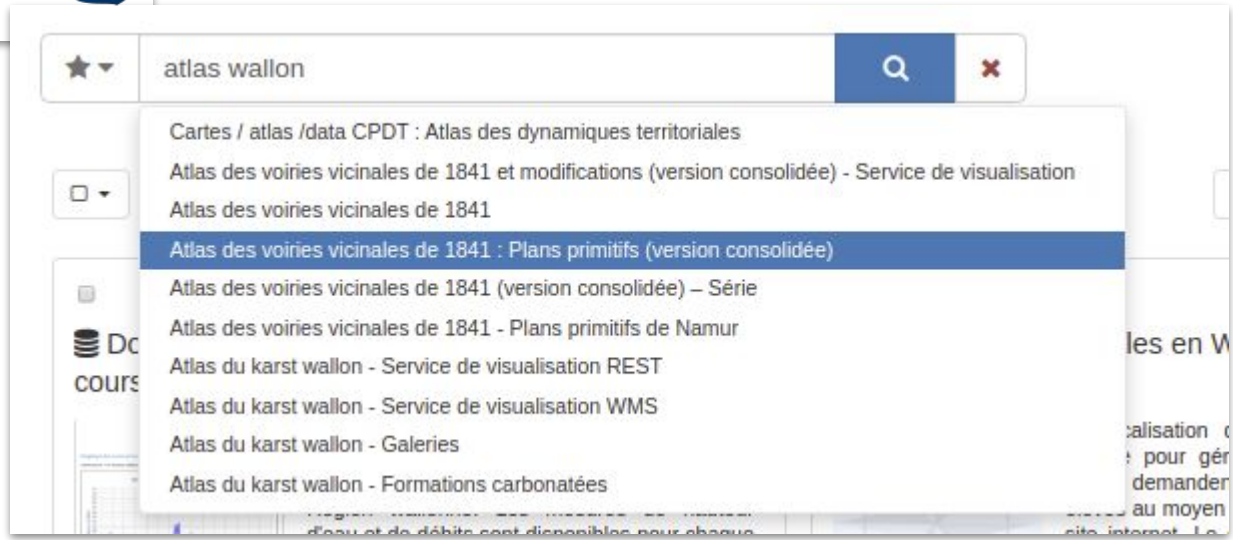
4

Can also suggest multiple field values (eg. title and keyword)

GN4 support phrase suggestion (ngrams in action)

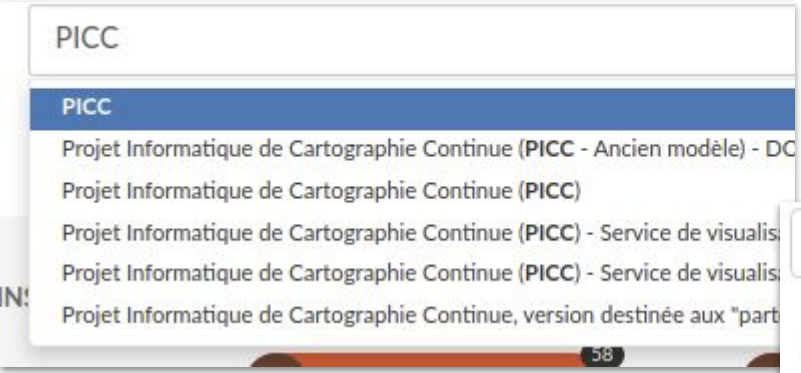
3

4

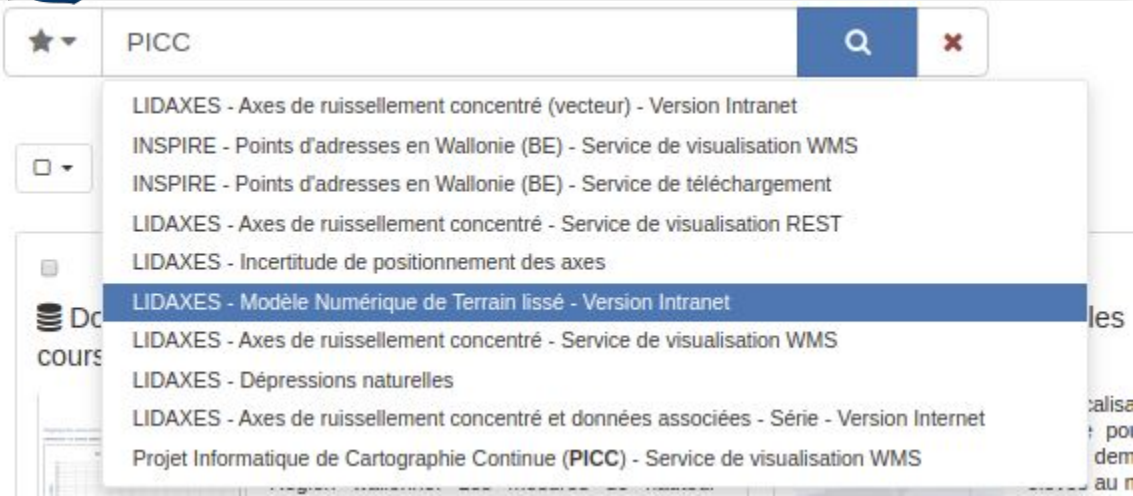


GN4 suggestions are sorted with score

3



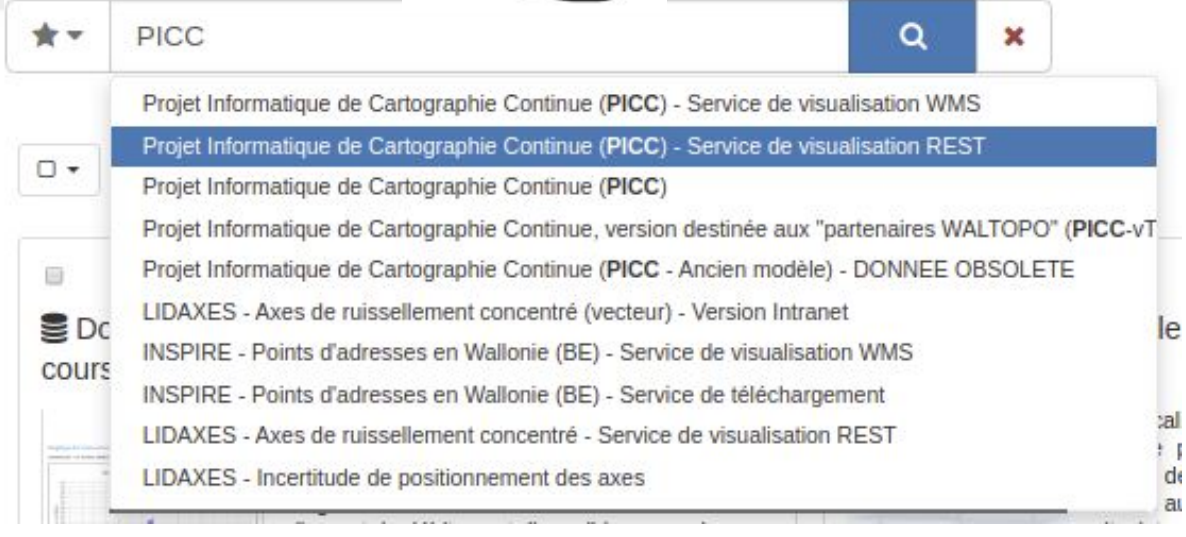
4



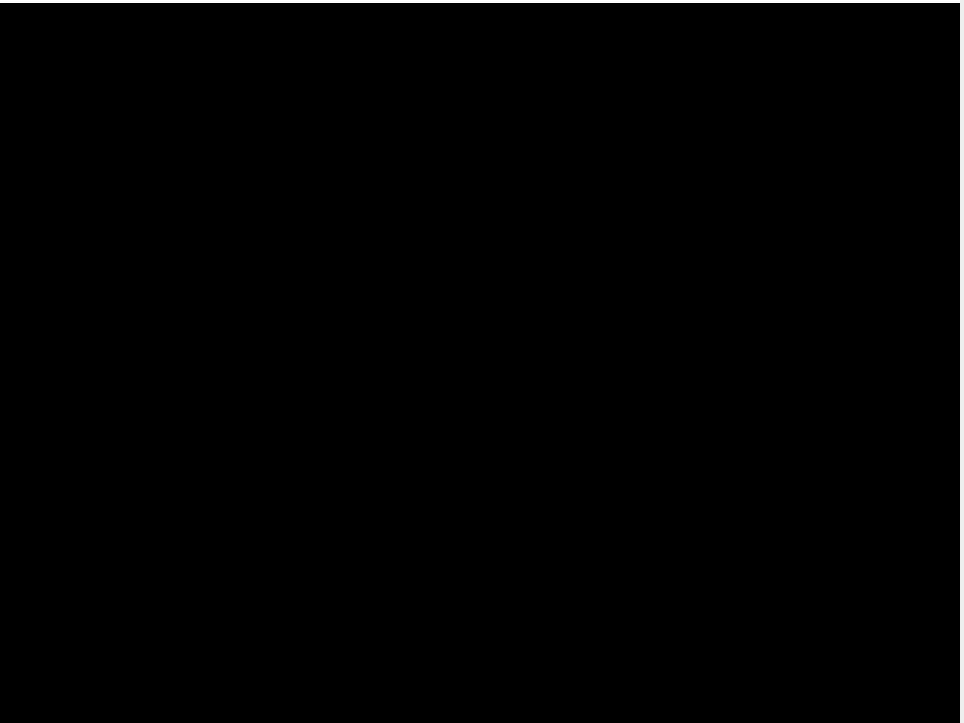
Give more importance to the title

ui-autocompleteConfig

```
{
  "query": {
    "multi_match": {
      "query": "",
      "type": "bool_prefix",
      "fields": [
        "resourceTitle",
        "anytext",
        "anytext._2gram",
        "anytext._3gram"
      ]
    }
  },
  "_source": [
    "resourceTitle"
  ]
}
```



Match on title higher ranking?



3

In GeoNetwork 3, score is mainly computed on term frequency between searched terms and document terms ...

Relevance can be adjusted based on 2 things

- Fields you are searching on
- Scoring

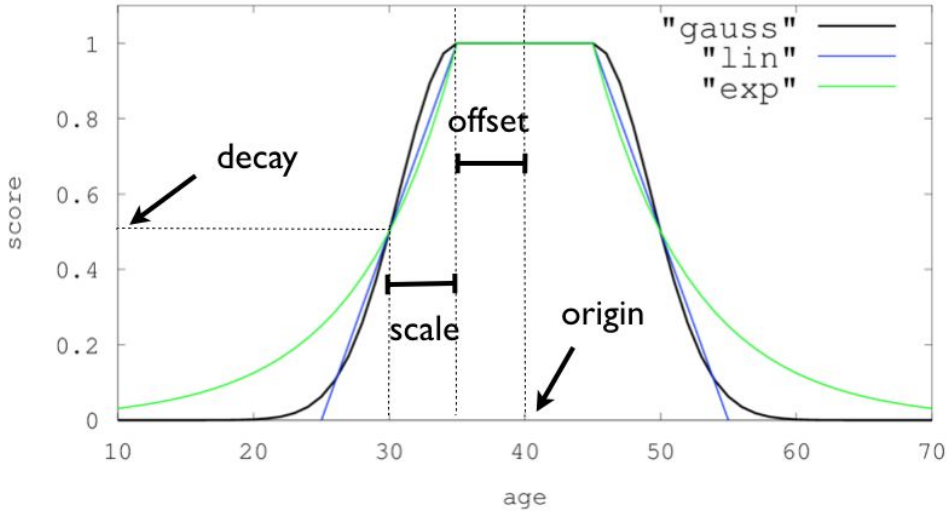
ui-autocompleteConfig

```
{
  "query": {
    "multi_match": {
      "query": "",
      "type": "bool_prefix",
      "fields": [
        "resourceTitle",
        "anytext",
        "anytext._2gram",
        "anytext._3gram"
      ]
```

GN4 allows configuration of scoring

Default is

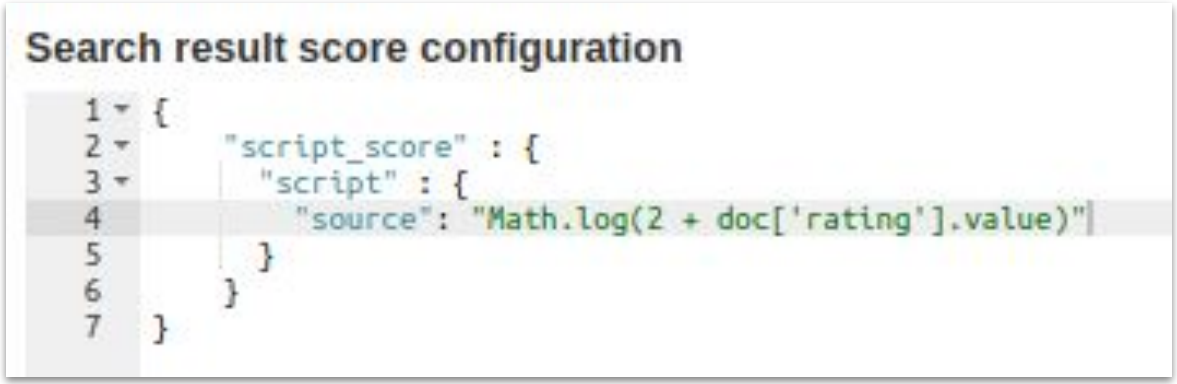
Adjust score based on datestamp



Search result score configuration

```
{
    "gauss": {
      "dateStamp": {
        "scale": "200d"
      }
    }
}
```

Score based on a function? Eg. promote those with higher rating?



And much more see API

Ideas : decrease score for obsolete records, for records member of a serie, ...

Spatial search works ?

GeoNetwork 3 has a post process to manage spatial filters. It searches based on criteria and then applies with AND the spatial filter.

Spatial index is independant, managed by GeoTools

Embedded spatial search

Elasticsearch introduces spatial index and search capabilities

Introduces capability to combine spatial and attribute filters (required by OGC compliance)

Polygons are not indexed yet in GN4.

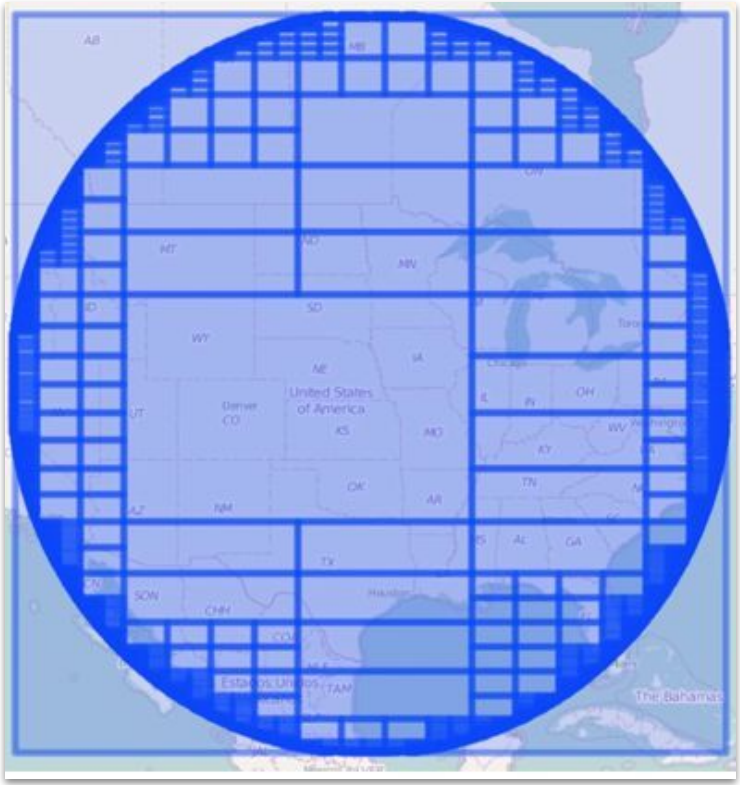
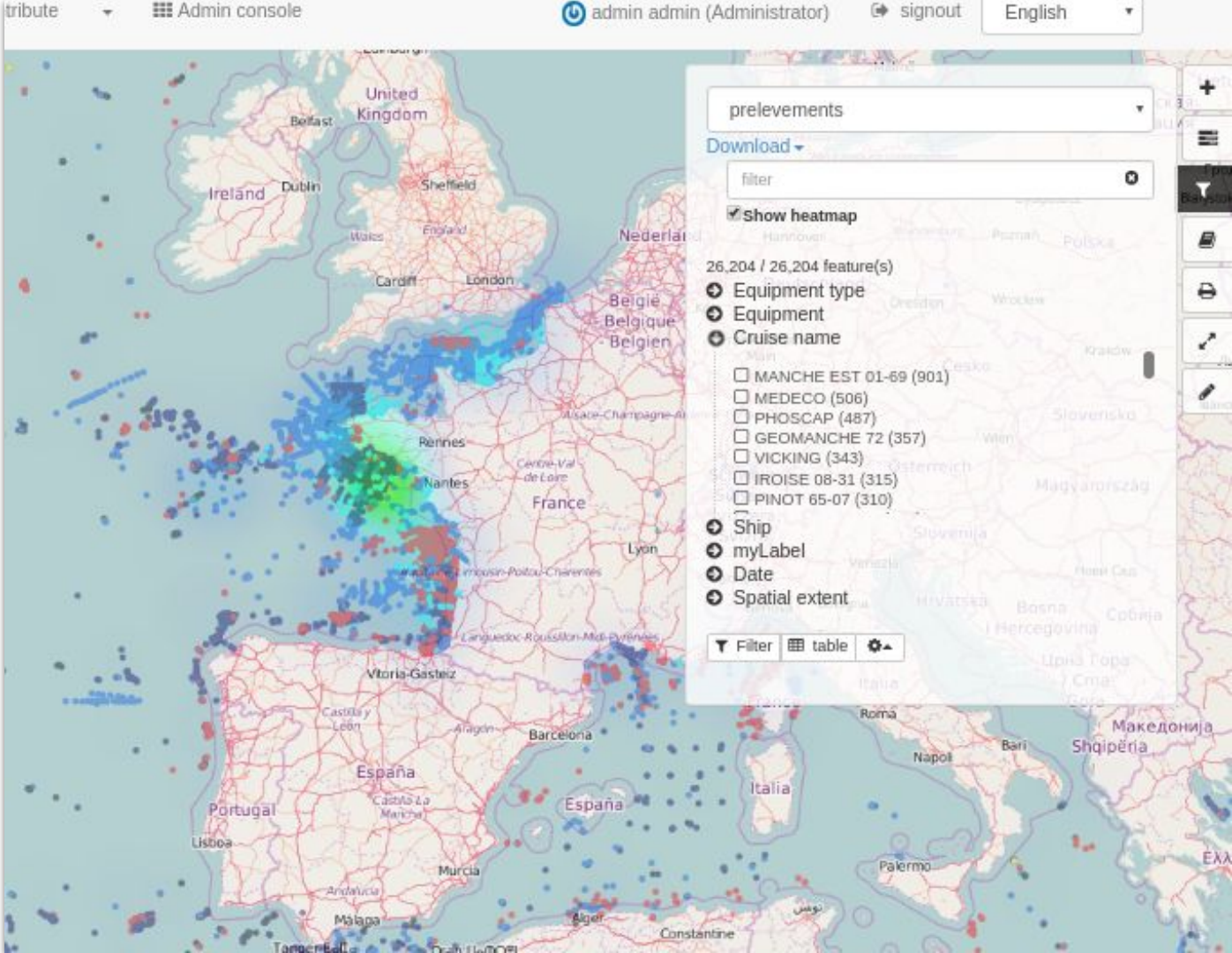


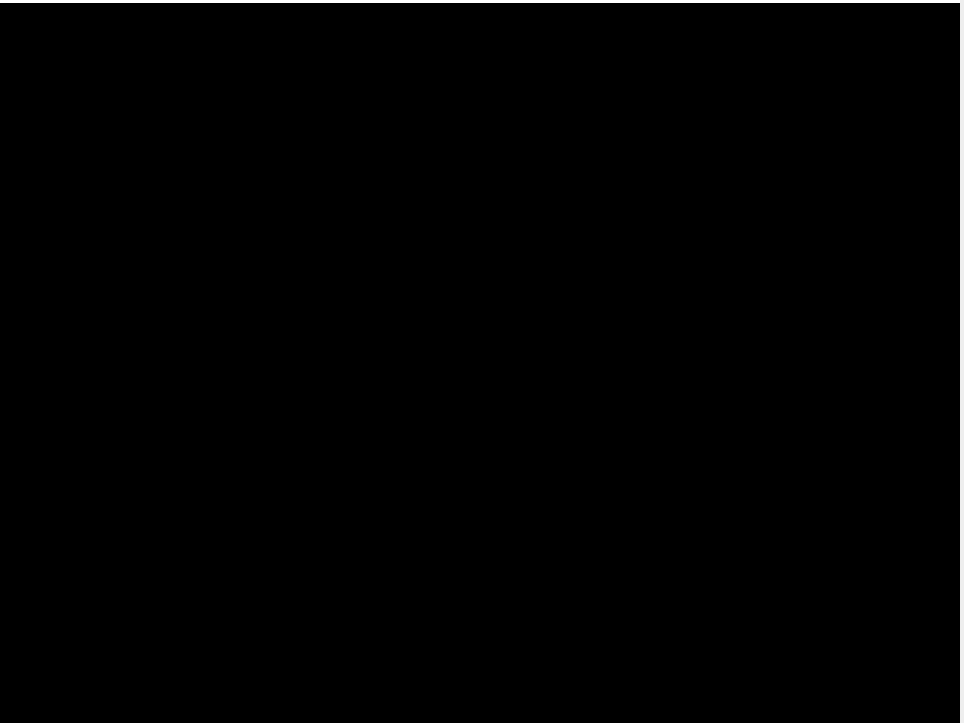
Image: elasti.co

Spatial aggregations are also available (for points)

eg. used for heatmap on WFS harvested features



Facets are still there?



3

In GeoNetwork 3, different types of facets are available : simple terms list, tree view eg. GEMET.

Only predefined set of facets are available - one for search, one for editor board.

Facets are thematic aggregations of search results, to further filter the search.

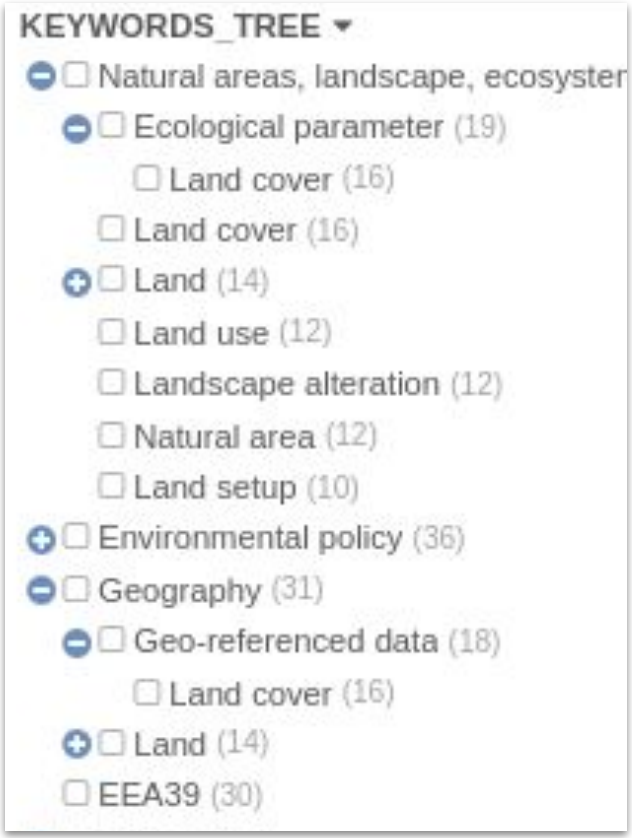
Aggregations with OR

Select multiple facet-values to be included in results.

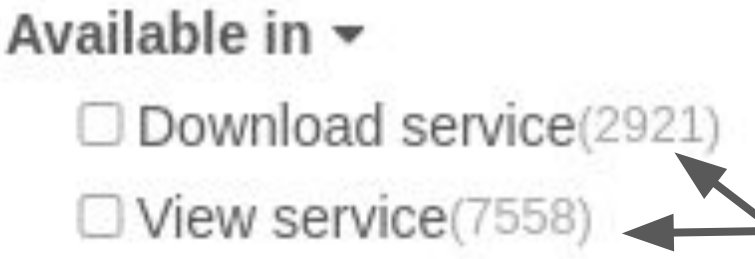
Exclude a value

Click “+” to add, click “-” to “exclude”

Tree based on hierarchy or by nesting fields



Flexibility with filter aggregations



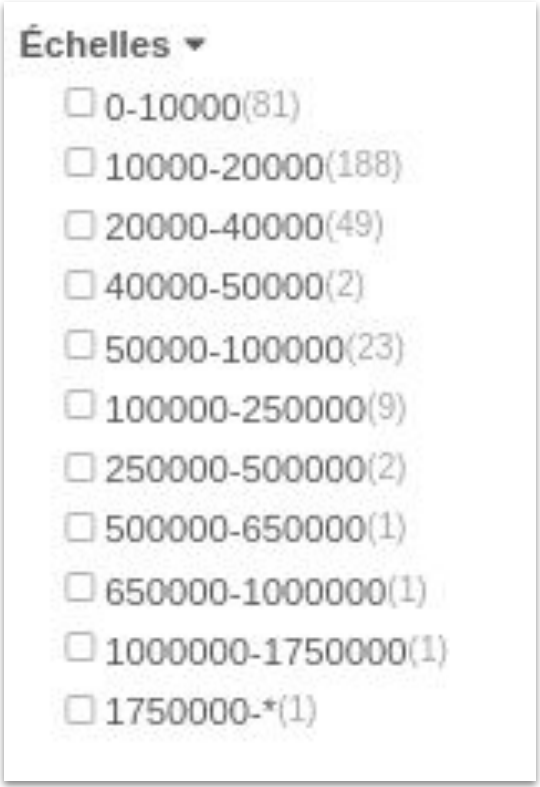

```
GET /gn-records/_search
{
  "aggs": {
    "messages" : {
      "filters" : {
        "filters" : {
          "availableInViewService" :  {
              "query_string": {
                "query": "+linkProtocol:/OGC:WMS.*/"
              }
            },
          "availableInDownloadService" :  {
              "query_string": {
                "query": "+linkProtocol:/OGC:WFS.*/"
              }
            }
        }
      }
    }},
  "query": {
  "match_all": {}
}}
```

Aggregations with paging

In 3 only a fixed number of values is returned

Keep clicking “more results” until end of set

Histogram aggregations …



All is configurable in the web admin

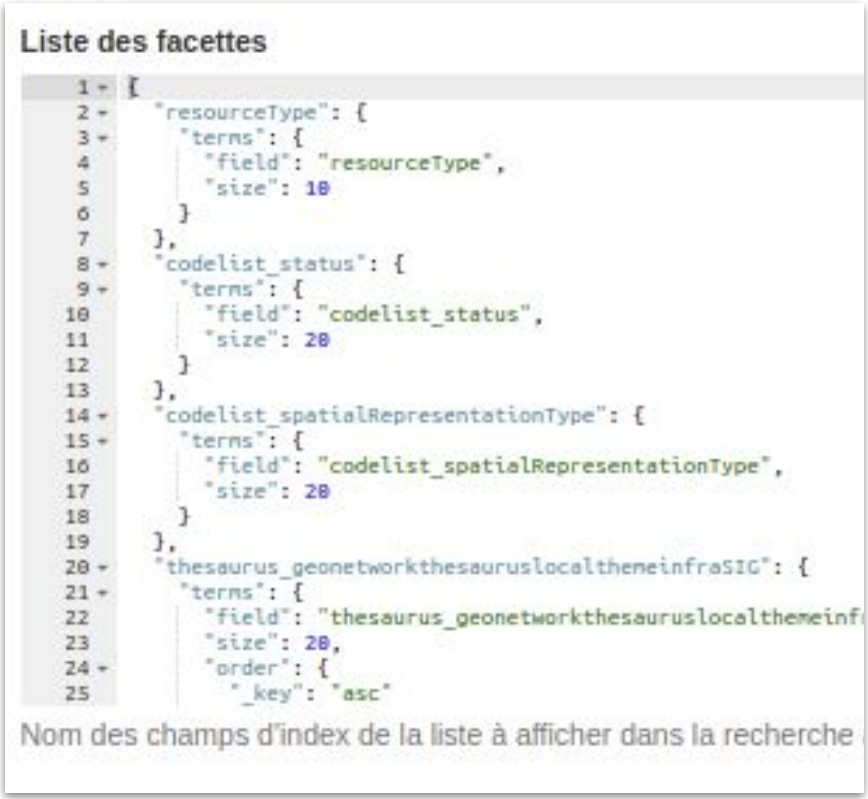
Configuration is based on Elasticsearch API

E.g.

- Facet list
- Autocomplete mode

Needs alternative “dummy mode”?

Needs validation option & preview



Other usages of facets?

Dashboards

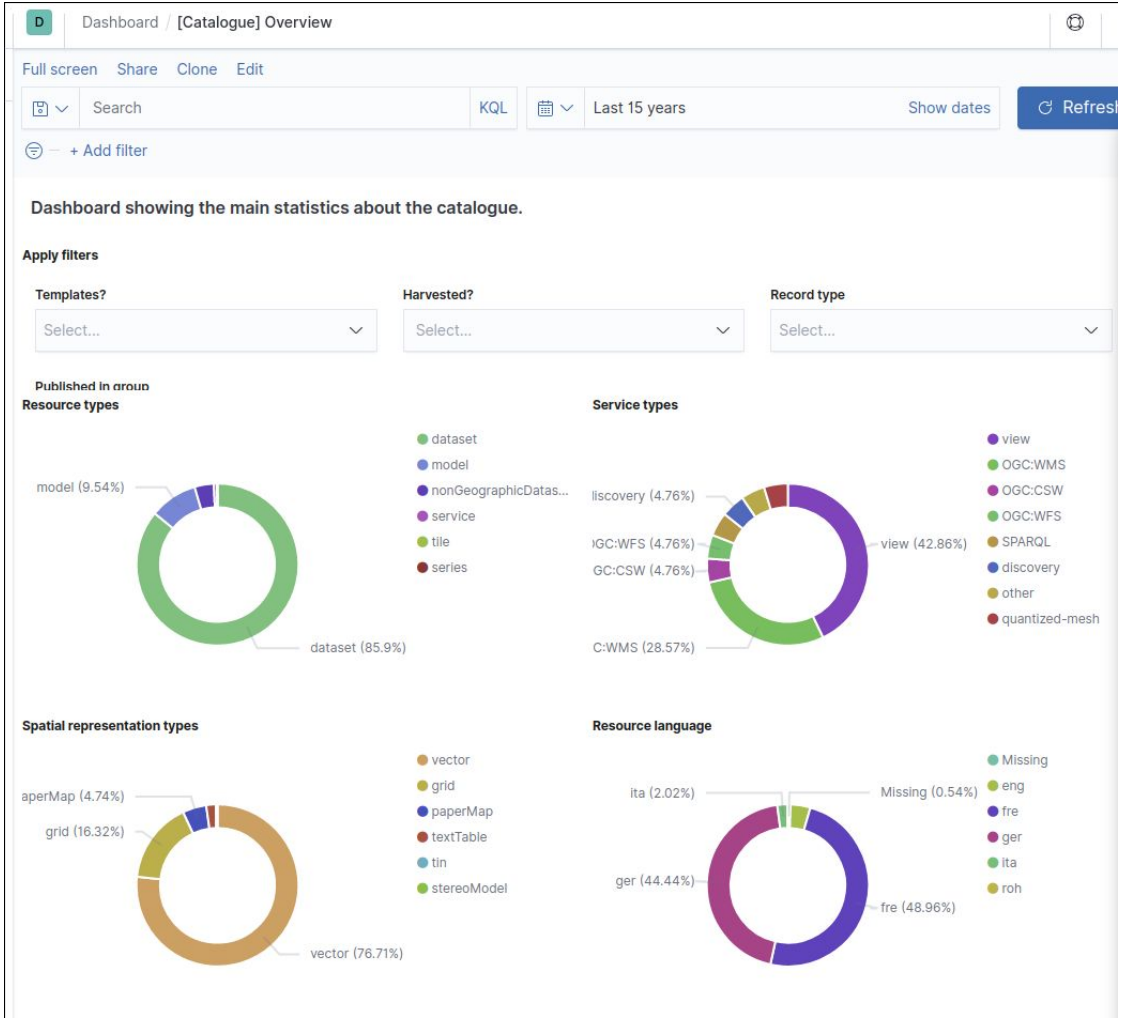
Create dashboard

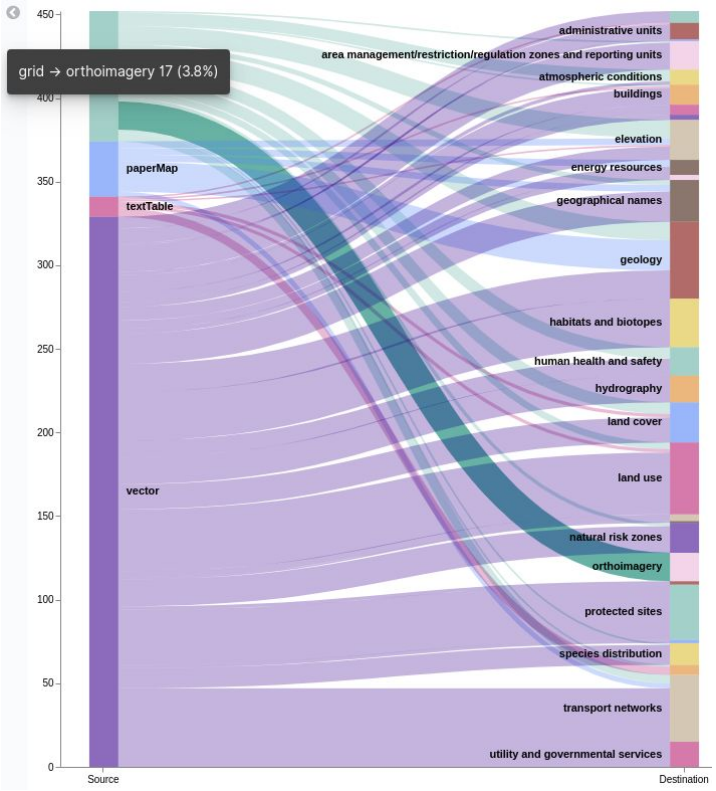
Search...

Title	Description	Actions
Admin overview	History, group & user statistics ...	
Overview	Dashboard showing the main statistics about the catalog	
Quality	Analyse catalog content, track errors and improve consistency	
Search	This is a simple dashboard to search records	
Spatial features statistics	Analyze the features harvested from WFS services	

Rows per page: 20

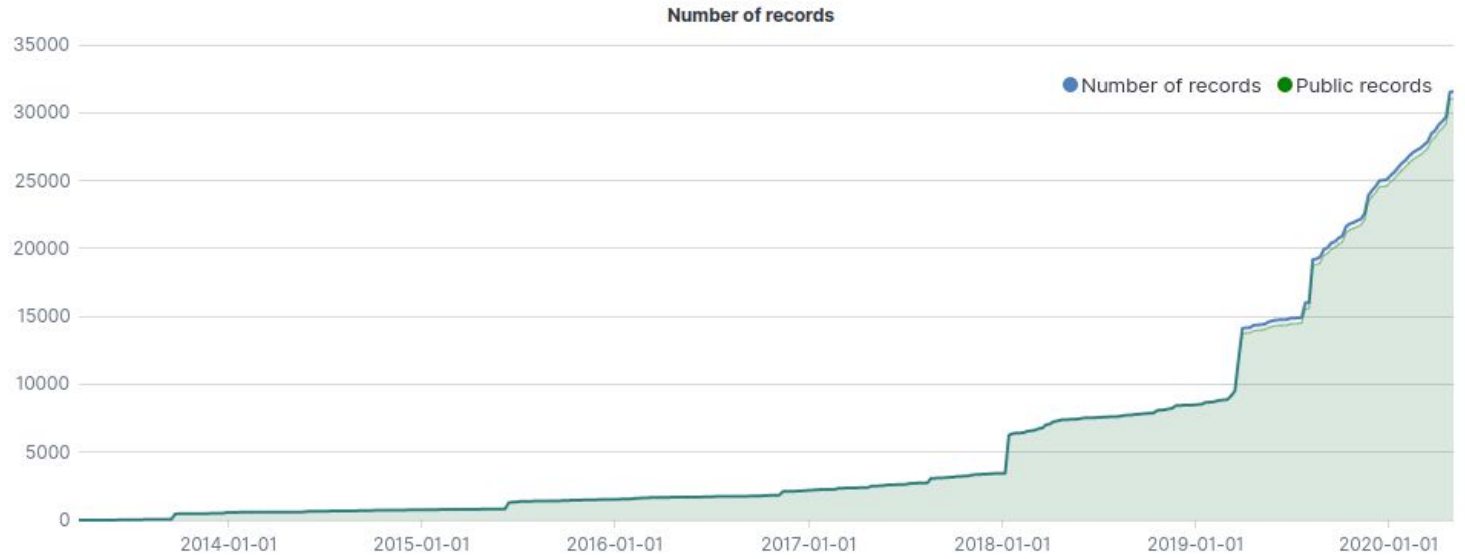
1


D
Dashboard / [Catalogue] Overview
Full screen
Share
Clone
Edit
Search
KQL
Last 15 years
Show dates
Refresh
+ Add filter
Dashboard showing the main statistics about the catalogue.
Apply filters
Templates?
Select...
Harvested?
Select...
Record type
Select...
Published in group
Resource types
model (9.54%)
dataset (85.9%)
dataset
model
nonGeographicDatas...
service
tile
series
Service types
liscovery (4.76%)
GC:WFS (4.76%)
GC:CSW (4.76%)
C:WMS (28.57%)
view (42.86%)
view
OGC:WMS
OGC:CSW
OGC:WFS
SPARQL
discovery
other
quantized-mesh
Spatial representation types
aperMap (4.74%)
grid (16.32%)
vector (76.71%)
vector
grid
paperMap
textTable
tin
stereoModel
Resource language
ita (2.02%)
ger (44.44%)
Missing (0.54%)
fre (48.96%)
Missing
eng
fre
ger
ita
roh


grid → orthoimagery 17 (3.8%)
450
400
350
300
250
200
150
100
50
0
Source
Destination
paperMap
textTable
vector
administrative units
area management/restriction/regulation zones and reporting units
atmospheric conditions
buildings
elevation
energy resources
geographical names
geology
habitats and biotopes
human health and safety
hydrography
land cover
land use
natural risk zones
orthoimagery
protected sites
species distribution
transport networks
utility and governmental services

Admin overview

Number of records over time



Number of records per groups and users

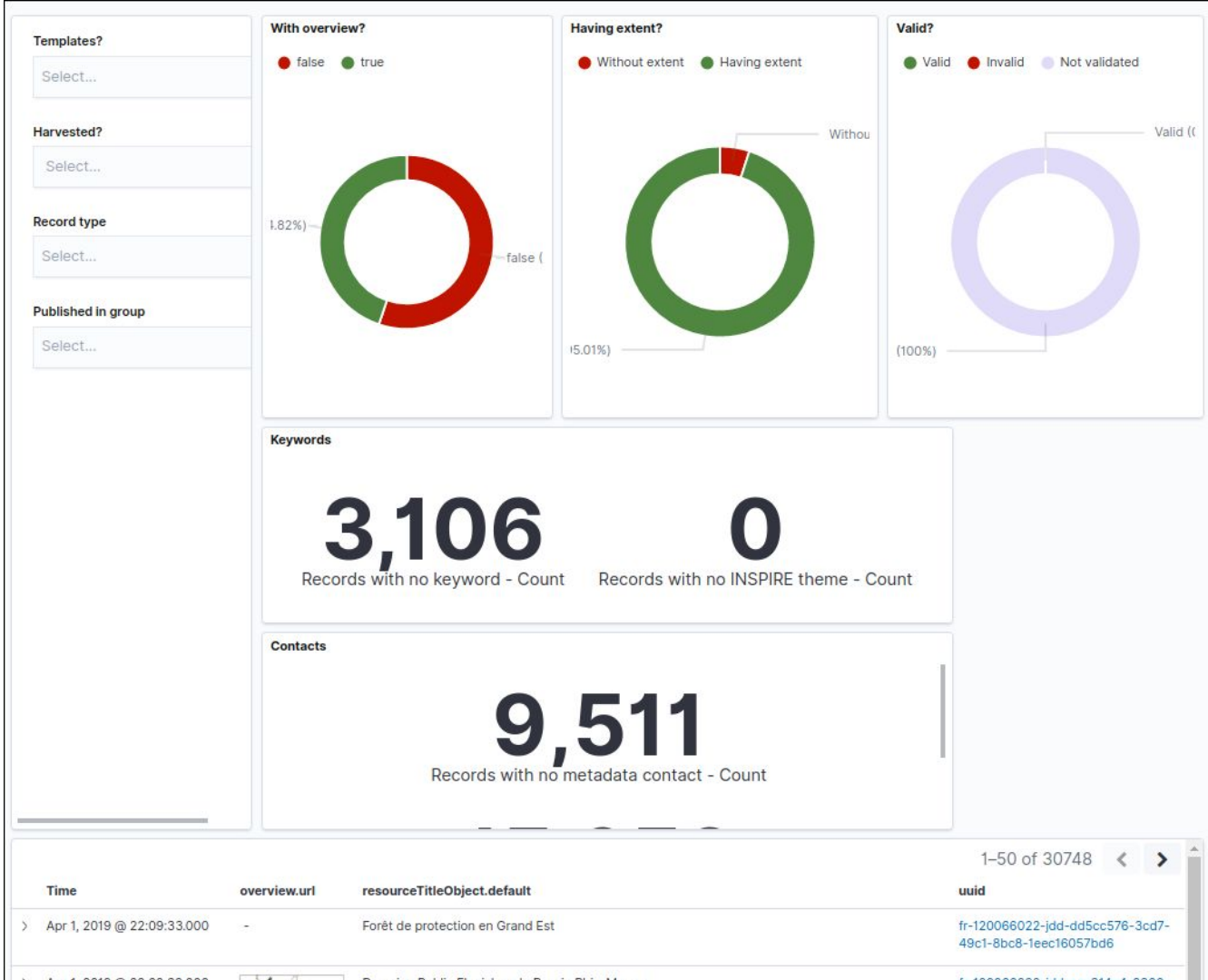
Group	Owner	Count
2582	admin admin	16,401
2	admin admin	12,316
1	admin admin	4

Export: Raw Formatted

28,206
Public records - Count

515
Private records - Count

Quality checks



What about synonyms?

In GeoNetwork 3, no simple support for synonyms exists

Synonyms are typically defined in thesauri, a keyword has same-as, narrower or broader relations to other concepts.

A query can use these relations to locate a record by matching a synonym of the search term.

Configure index to expand synonyms

```
"french_synonym": {
  "type": "synonym",
  "ignore_case": true,
  "expand": true,
  "synonyms": [
    "sig, systeme d'information geographique, ids, gis, s
    "shp, shapefile",
    "raster, image",
    "salade, laitue"
  ]
},
```

More to experiment

Link to synonyms from thesauri

Synonyms at index time or at query time

Use the Elasticsearch synonyms API

Is the search engine aware of language specific aspects?

In GeoNetwork 3, Lucene configuration can load different Analyzers

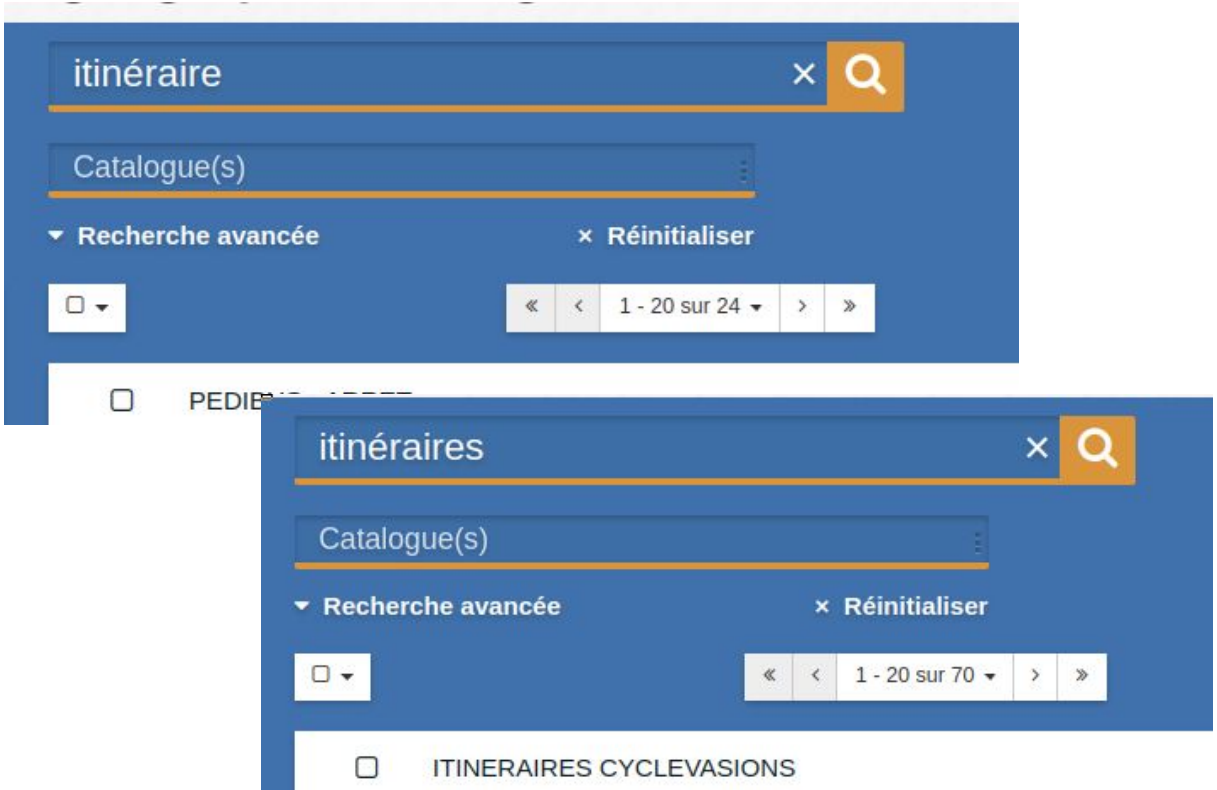
Yes. Ongoing experiment for French in Wallonia

```
"filter": {
  "shingle": {
    "type": "shingle",
    "min_shingle_size": 2,
    "max_shingle_size": 3
  },
  "french_elision": {
    "type": "elision",
    "articles_case": true,
    "articles": ["l", "m", "t", "qu", "n", "s", "j", "d", "c", "jus
  },
  "french_synonym": {
    "type": "synonym",
    "ignore_case": true,
    "expand": true,
    "synonyms": [
      "sig, systeme d'information geographique, ids, gis, sdi",
      "shp, shapefile",
      "raster, image"
    ]
  },
  "french_stemmer": {
    "type": "stemmer",
    "language": "light_french"
```

```
"analyzer": {
  "french_heavy": {
    "tokenizer": "icu_tokenizer",
    "filter": [
      "french_elision",
      "icu_folding",
      "french_synonym",
      "french_stemmer"
    ]
  },
  "french_light": {
    "tokenizer": "icu_tokenizer",
    "filter": [
      "french_elision",
      "icu_folding"
    ]
  },
```

And what are the benefits?

Eg. Plurals in search
Analyzer per language & stemming



Indexation

Document entrée →

Id	Métier
1	Développeuse

Ascii folding → Developpeuse

Lowercase → developpeuse

French stemmer → developeu

Recherche

Métier
développeur

← Terme recherché

developpeur ← Ascii folding

developpeur ← Lowercase

developeu ← French stemmer

Index

Clé	Id document
developeu	1

And multilingual metadata?

In GeoNetwork 3, complex multilingual support with some drawbacks

Not yet

But

The idea is to use objects to store translations

(instead of one index per language in 

)

Potential Multilingual features

- Display the metadata element in the UI language
- Match keywords from thesauri with their translation in the UI language
- Higher score for matches on elements in the UI language
- Machine translation of metadata content to the UI language

So there is some more work needed?

So there is some more work needed?

Yes. That’s why we are here ;)

What are the next steps?

Install it, test & report issues

=

DB

\+ webapp

\+ Elasticsearch

\+ (optional) Kibana

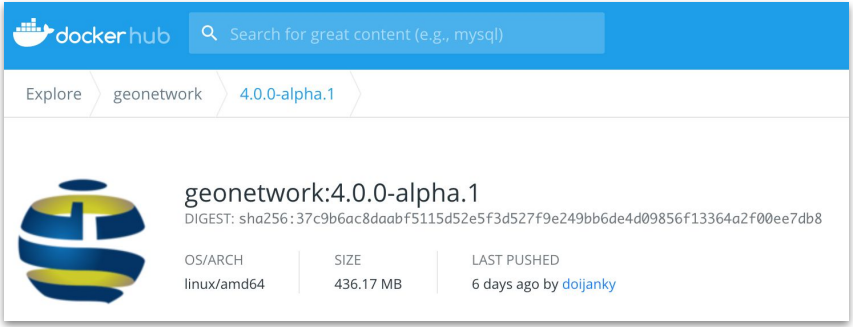
Start it with docker, test & report issues

Run Elasticsearch

docker run -p 8080:8080 Geonetwork:4.0.0-alpha

And open http://localhost:8080/geonetwork/

Or deploy ES+GN+PG+kibana using docker-compose.yml



Some features are not there yet?

Yes,

some even completely removed!

Deprecated features removed

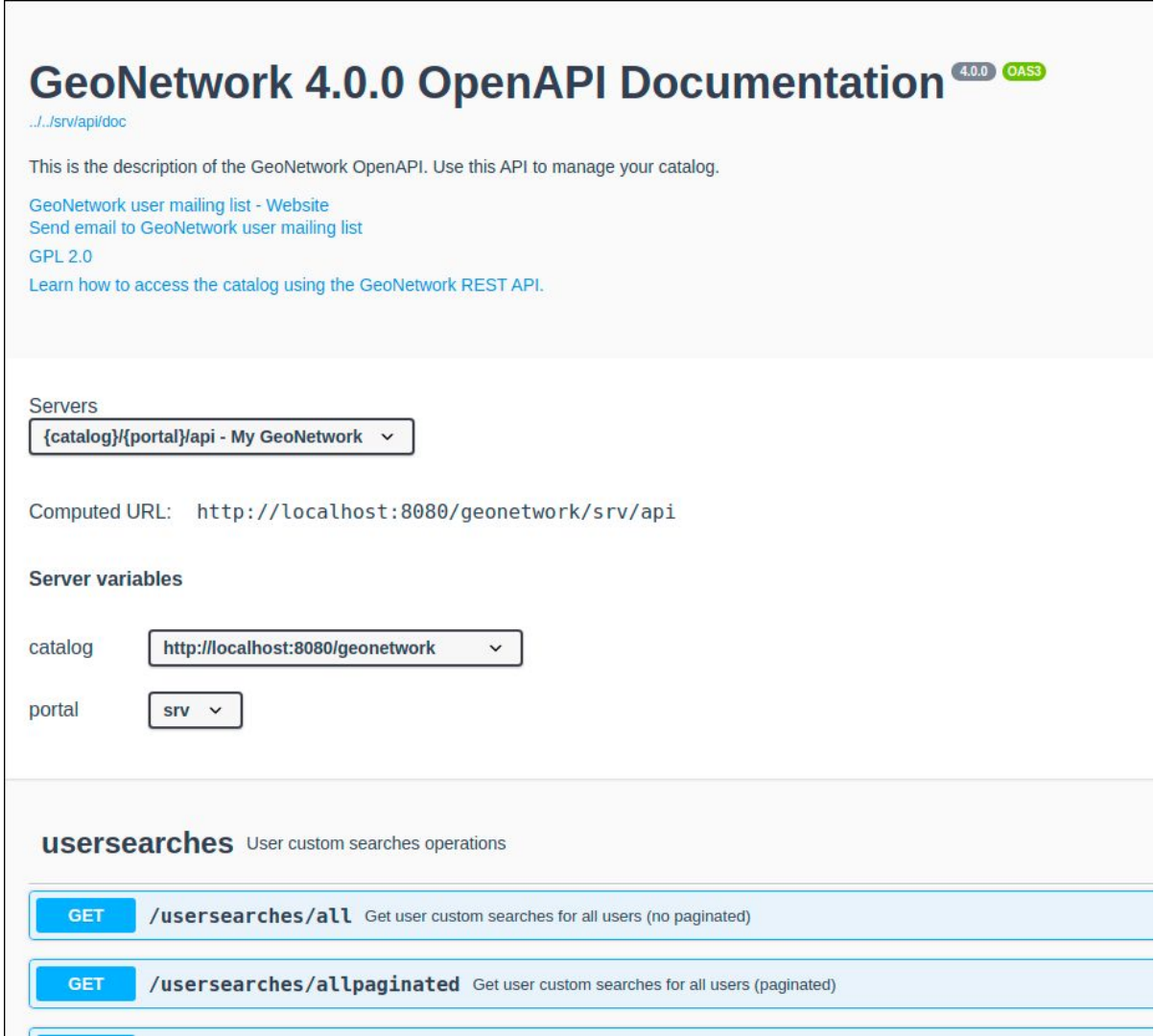
The new generation of GeoNetwork removed some of the previously deprecated functionalities.

Impact on User Interface is minimal, it mostly relates to unavailability/move of deprecated service endpoints.

q search service is removed. Use /api/search

- CSW / Virtual CSW is replaced by sub-portal (see Portal configuration)
- CSW / results_with_summary custom GeoNetwork output schema is removed. To retrieve facets use the main search API.
- Deprecated Jeeves services removed (ie. services not used by Angular application like harvester config). Use the swagger API instead (see GeoNetwork API).
- Metadata notifier manager (can be replaced by event system)
- Overrides mechanism of configuration file

OpenAPI

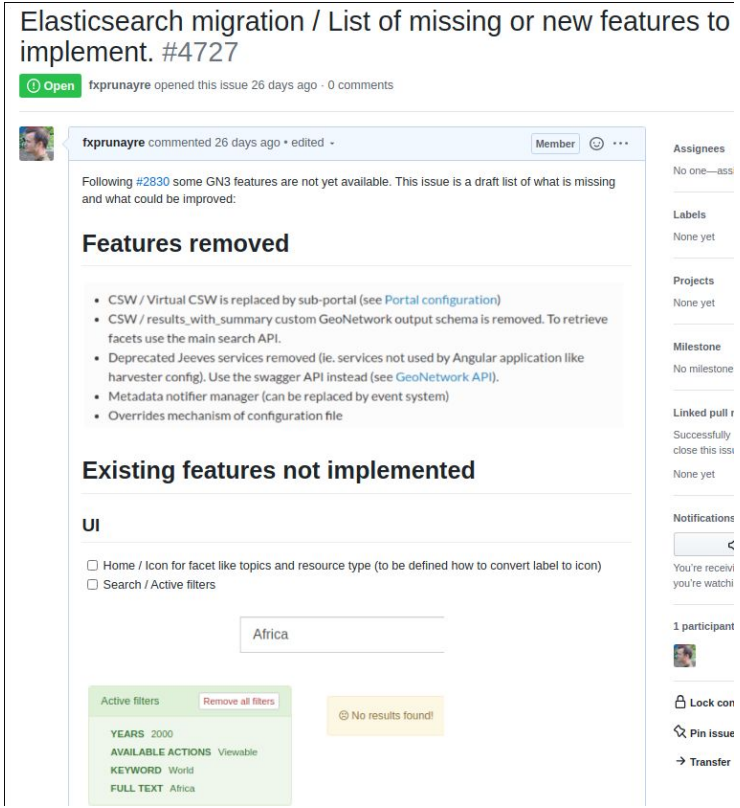

GeoNetwork 4.0.0 OpenAPI Documentation
4.0.0
OAS3
../../srv/api/doc
This is the description of the GeoNetwork OpenAPI. Use this API to manage your catalog.
GeoNetwork user mailing list - Website
Send email to GeoNetwork user mailing list
GPL 2.0
Learn how to access the catalog using the GeoNetwork REST API.
Servers
{catalog}/{portal}/api - My GeoNetwork
Computed URL: http://localhost:8080/geonetwork/srv/api
Server variables
catalog
http://localhost:8080/geonetwork
portal
srv
usersearches
User custom searches operations
GET
/usersearches/all
Get user custom searches for all users (no paginated)
GET
/usersearches/allpaginated
Get user custom searches for all users (paginated)

So what is still missing?

Which features to add before 4.0.0 release?

See

- https://github.com/geonetwork/core-geonetwork/issues/4727
- https://geonetwork-opensource.org/manuals/4.0.x/en/overview/change-log/version-4.0.0-alpha.1.html#what-is-next



Mid-day survey to collect your interest

What MUST be in 4.0.0?

Thank you

And we have demo that you can play with https://apps.titellus.net/geonetwork/

Scaling and Sharding

- Elastic is a separate process, which by design has scaling and sharding options.
- Any number of GeoNetwork instances can run on top of an Elastic Index.
- Elastic can scale by replicating the index over multiple nodes and load balance the traffic to each of the nodes, while synchronising the content.
- Elastic can split (shard) the index over multiple nodes, a search query is fired to each of the nodes. The partial result sets are then merged.

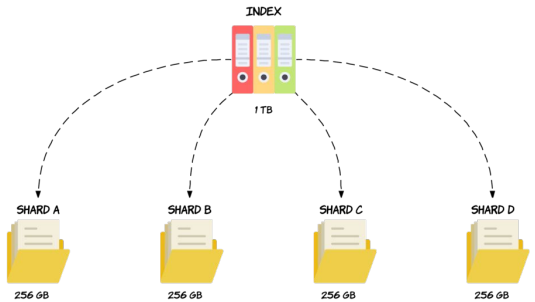


Image: codingexplained.com

OGC API Records

Standard currently under design at OGC as an alternative to CSW

- Uses similar technologies as current GeoNetwork API (Open API, JSON, content negotiation)
- Facets will be an extension, we should contribute GeoNetwork experience on this topic

Authorisation

Elastic provides an authorisation mechanism similar to GN, with roles in groups. GeoNetwork could adopt this authorisation model.

This allows third party tooling (such as kibana) to engage with the index directly, while benefiting from the authorisation provided by elastic.

Do we still need a database?

In GN<4 the most important use of the database is to provide a backup for metadata (if the lucene index is corrupted, it can be recompiled from the database), facilitate autorisation and manage spatial filters.

In GN4

- Index can have its own backup mechanism.
- Spatial filters are managed by index
- Authorisation aspects should be included in (or delegated to) index

Other aspects which are stored in the database can be stored in the index, such as user feedback, search statistics. Or in config files, e.g. settings, translations.