GeoNetwork 4

User expectations?
It is all about search ...
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”
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?

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 configured.

Here title and keywords.
Autocomplete has unexpected results

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)
Can also suggest multiple field values (eg. title and keyword)
GN4 support phrase suggestion (ngrams in action)
GN4 suggestions are sorted with score
Give more importance to the title
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
GN4 allows configuration of scoring

Default is

```json
{
  "script_score": {
    "script": {
      "source": "_score"
    }
  }
}
```
Adjust score based on datestamp
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 independent, 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.
Spatial aggregations are also available (for points) eg. used for heatmap on WFS harvested features
Facets are still there?

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 only a fixed number of values is returned

Keep clicking “more results” until end of set
Histogram aggregations ...

- 0-10000 (81)
- 10000-20000 (188)
- 20000-40000 (49)
- 40000-50000 (2)
- 50000-100000 (23)
- 100000-250000 (9)
- 250000-500000 (2)
- 500000-650000 (1)
- 650000-1000000 (1)
- 1000000-1750000 (1)
- 1750000-* (1)
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?
<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin overview</td>
<td>History, group &amp; user statistics ...</td>
</tr>
<tr>
<td>Overview</td>
<td>Dashboard showing the main statistics about the catalog</td>
</tr>
<tr>
<td>Quality</td>
<td>Analyse catalog content, track errors and improve consistency</td>
</tr>
<tr>
<td>Search</td>
<td>This is a simple dashboard to search records</td>
</tr>
<tr>
<td>Spatial features statistics</td>
<td>Analyze the features harvested from WFS services</td>
</tr>
</tbody>
</table>
Dashboard showing the main statistics about the catalogue.

**Apply filters**
- Templates?
- Harvested?
- Record type
  - Select...

**Published in groups**
- Resource types
  - Model (9.54%)
  - Dataset (85.9%)

**Spatial representation types**
- Grid (16.32%)
- Vector (76.71%)

**Service types**
- Discovery (4.76%)
- OGC-WMTS (4.76%)
- OGC-CSW (4.76%)
- C/WMS (28.57%)
- View (42.06%)

**Resource language**
- Other (0.54%)
- Missing (0.54%)
- Ita (2.02%)
- Ger (44.44%)
- Fire (48.06%)
- Eng (44.44%)

**Data types**
- Administrative units
- Area management restriction
- Regulation rules and reporting units
- Atmospheric conditions
- Buildings
- Elevation
- Energy resources
- Geographical names
- Geology
- Habitats and biotypes
- Human health and safety
- Hydrography
- Land cover
- Land use
- Natural risk zones
- Orthophotography
- Protected sites
- Species distribution
- Transport networks
- Utility and governmental services
Admin overview

Number of records over time

Number of records per groups and users

<table>
<thead>
<tr>
<th>Group</th>
<th>Owner</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2582</td>
<td>admin admin</td>
<td>16,401</td>
</tr>
<tr>
<td>2</td>
<td>admin admin</td>
<td>12,316</td>
</tr>
<tr>
<td>1</td>
<td>admin admin</td>
<td>4</td>
</tr>
</tbody>
</table>

28,206  515

Public records - Count  Private records - Count
Quality checks

3,106 Records with no keyword - Count
0 Records with no INSPIRE theme - Count

9,511 Records with no metadata contact - Count
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

```json
"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
And what are the benefits?
Eg. Plurals in search
Analyzer per language & stemming
Indexation

Document entrée →

<table>
<thead>
<tr>
<th>Id</th>
<th>Métier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Développeuse</td>
</tr>
</tbody>
</table>

Ascii folding →

Developpeuse

developpeur

Lowercase →

developpeuse

developpeur

French stemmer →

developeu

developpeur

developeu

Recherche

<table>
<thead>
<tr>
<th>Métier</th>
</tr>
</thead>
<tbody>
<tr>
<td>développer</td>
</tr>
</tbody>
</table>

Terme recherché

developpeur

Ascii folding

developpeur

Lowercase

developpeur

French stemmer

developpeur

Index

<table>
<thead>
<tr>
<th>Clé</th>
<th>Id document</th>
</tr>
</thead>
<tbody>
<tr>
<td>développeu</td>
<td>1</td>
</tr>
</tbody>
</table>
And multilingual metadata?

In GeoNetwork 3, complex multilingual support with some drawbacks
Not yet
The idea is to use objects to store translations

(instead of one index per language in 🌍3)
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
GeoNetwork 4.0.0 OpenAPI Documentation

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://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](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 authorisation 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.