Data Discoverability Guidance

Release 0.1

Astun Technology

May 30, 2023

CONTENTS:

CHAPTER

ONE

DATA SHARING AND METADATA CREATION MADE EASY WITH GEONETWORK OPEN SOURCE

1.1 Introduction

This is a guide to extending the GeoNetwork Metadata Catalog to meet best practice guidance on storing and sharing metadata for both spatial and non-spatial datasets, and for improving data discoverability following best practice guidance from the Geospatial Commission.

The objective is to provide guidance on the configuration changes and additional schema plugins that you'll need for full spatial and non-spatial data discoverability, along with a suggested workflow for metadata creators.

It's an ongoing project by Astun Technology, and was funded in part by a grant from the Open Data Institute.

The raw files for this documentation are hosted on GitHub at https://github.com/AstunTechnology/datadiscoverabilityguidance, so if you spot a mistake then head over there and let us know. You can also suggest changes, ask questions, or even submit fixes!

1.2 Requirements

Important: To follow this guidance you will need to have an installation of GeoNetwork Open Source, with version 4.2.x. Some functionality worked differently in previous versions.

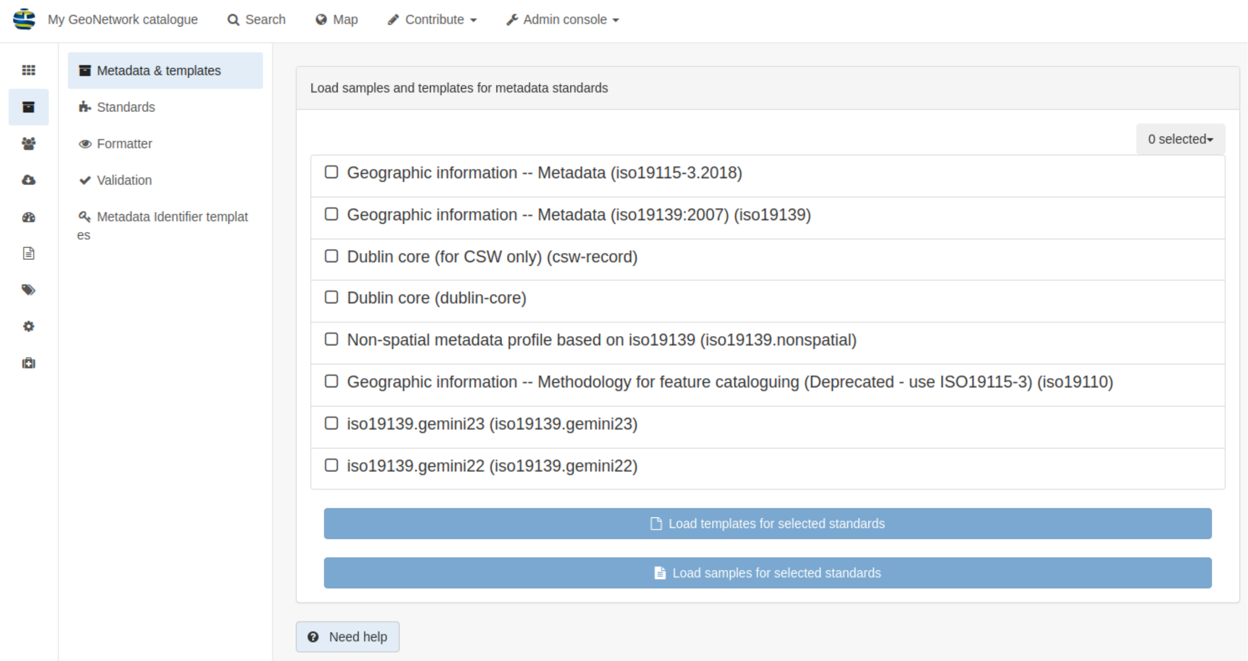
For instructions on installing GeoNetwork, and guidance on getting started, please see the official documentation.

You will also need to install the plugins for Gemini 2.3 (for spatial data), and iso19139.nonspatial (for non-spatial data).

Once you've restarted GeoNetwork, you can check that the metadata profiles have loaded correctly by logging in as an Administrator and going to the Admin console > **Metadata and templates** page.

The list should include the following additional entries (alongside the pre-loaded ones):

- Non-spatial metadata profile based on iso19139 (iso19139.nonspatial)
- iso19139.gemini23 (iso19139.gemini23)



Load samples and templates for metadata standards

□ Geographic information -- Metadata (iso19115-3.2018)
□ Geographic information -- Metadata (iso19139:2007) (iso19139)
□ Dublin core (for CSW only) (csw-record)
□ Dublin core (dublin-core)
□ Non-spatial metadata profile based on iso19139 (iso19139.nonspatial)
□ Geographic information -- Methodology for feature cataloguing (Deprecated - use ISO19115-3) (iso19110)
□ iso19139.gemini23 (iso19139.gemini23)
□ iso19139.gemini22 (iso19139.gemini22)

Click the dropdown box on the right marked **0 selected** and choose **All**, then click the buttons below the templates list to **Load templates for selected standards** and **Load samples for selected standards**. This will load all the available templates and sample records into your catalog.

1.3 Configuration

You will need to change a number of settings in the administrator panel to get best use out of GeoNetwork. Login in as an administrator, and visit Admin Console > **Settings**.

In the main **Settings** tab, we recommend making changes to the following sections. Note that there are many other options that you can also change, see the official documentation for more information.

Catalog Description

- Fill in a Catalog Name
- Fill in the Organization

Catalog Server

- Change the **Host**, **Preferred Protocol**, **Port** and **Secure Port** to match your install. For example if you access the catalog at the URL `https://mygeonetwork.com/geonetwork` then you'd set the following:
  - Host: `mygeonetwork.com`
  - Preferred Protocol: `https`
  - Port: leave blank

  - Secure Port: leave blank

- For the **Timezone**, set the most appropriate one for you. In the UK this is probably `Europe/London`

Feedback

- Change the **Email** to the address you want catalog emails to come **from**
- Fill in the address of your mailserver (the **SMTP host**, and set the rest of the options in this section as appropriate.

To test, save your changes and then click the **Test mail configuration** button. This will send an email **to** the specified address, so make sure it's one you have access to.

Warning: If the catalog server is not part of the same domain as the email address, then messages from GeoNetwork may be classified as spam.

User feedback

- Click the **Enable feedback** option to allow people to leave comments on records

Search statistics

- Click the **Enable** option to store search statistics

INSPIRE Directive configuration

- Click the **INSPIRE** option to enable the ability to display records by INSPIRE theme on the home page. Ensure you have the INSPIRE thesaurus installed (see the page on classification systems for details)
- Click the **INSPIRE search panel** option.

Metadata configuration

- Click the **Remove schema location for validation** option. This prevents validation errors from records where the schemalocation is incorrect or cannot be reached. In this case the schema files loaded on your local server are used instead.

Important: Be sure to click the blue **Save Settings** button to save your changes!

1.4 Workflow

This is a workflow for using GeoNetwork to meet the Government's guidance on sharing tabular data.

1.4.1 Choosing a data format

We recommend using CSV for your non-spatial tabular datasets to meet Government data sharing guidance but JSON may be a more suitable format if the data is more complex. See Government API guidelines for information on good practice for JSON.

If you are more accustomed to sharing data as an Excel spreadsheet, we would **strongly** recommend that you convert to CSV as above for data sharing to avoid security risks from macros, or problems arising from Excel's auto-formatting functionality.

1.4.2 Formatting your data as a CSV

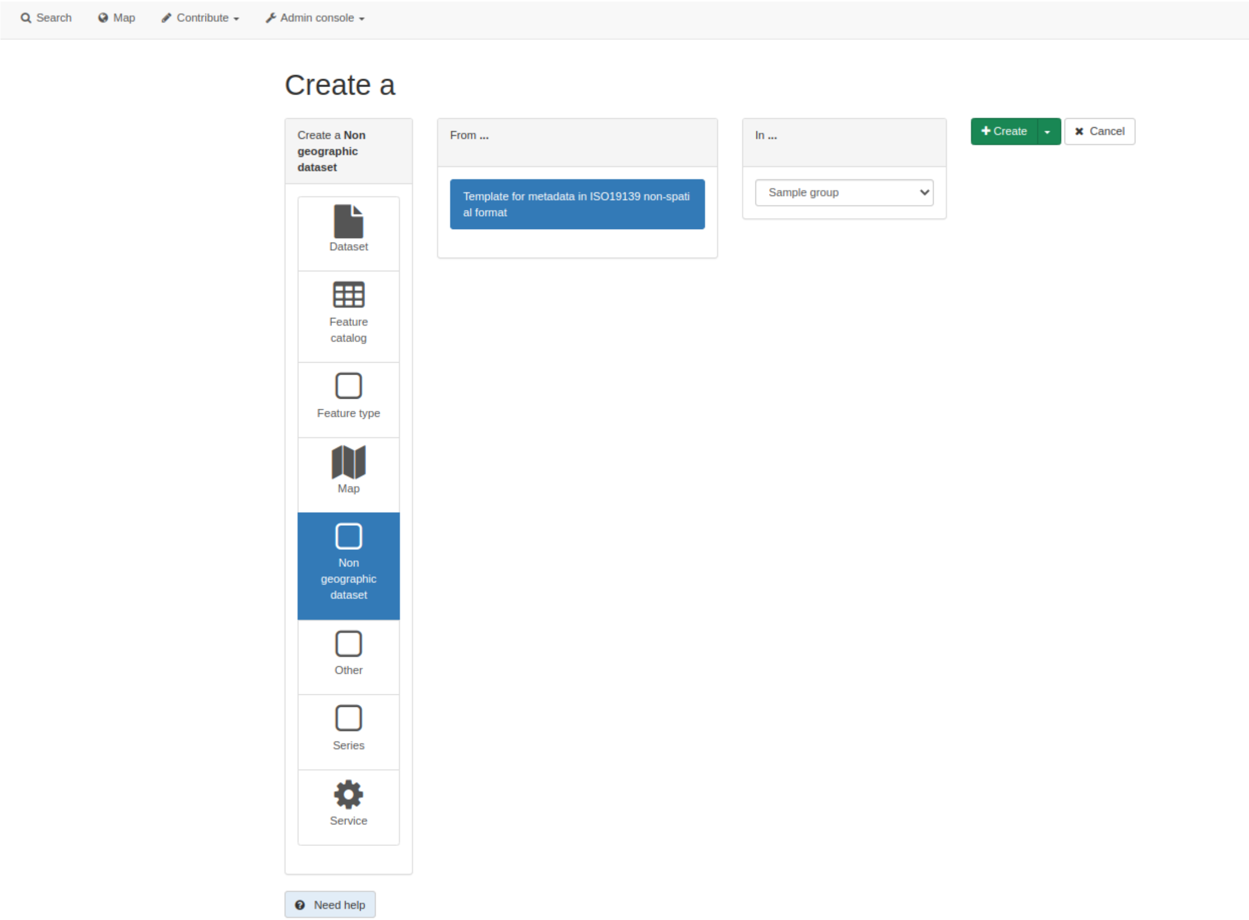
The Government's guidance on a tabular data standard recommends that you share non-spatial metadata in CSV format, meeting the following specifications:

- 0 or 1 header rows (preferrably 1)
- After the header row, each row should represent a record (eg no blank lines, totals or so on)
- Fields are separated by commas, with text optionally delimited with double quotes
- All rows have the same number of fields
- Line-breaks use windows style "\r\n"
- Use UTF8 for encoding
- No Byte Order Mark (see the link above for more information)

1.4.3 Creating a metadata record

Log into GeoNetwork as a user with Editor priviliges or higher and navigate to the **Contribute Tab**. Choose **Add a new record** and then select **Non geographic dataset** from the list on the left. Assuming you have followed the configuration instructions you should be offered the template **Template for metadata in ISO19139 non-spatial format**. Select it by clicking on it, then choose the group you wish to create the record in, and finally click the **+Create** button on the right.

Fill in all the fields shown in the default non-spatial view once the record's edititng page opens. Remember to save the changes before closing the window using the buttons at the top of the page.



From ...

Template for metadata in ISO19139 non-spatial format

In ...

Sample group

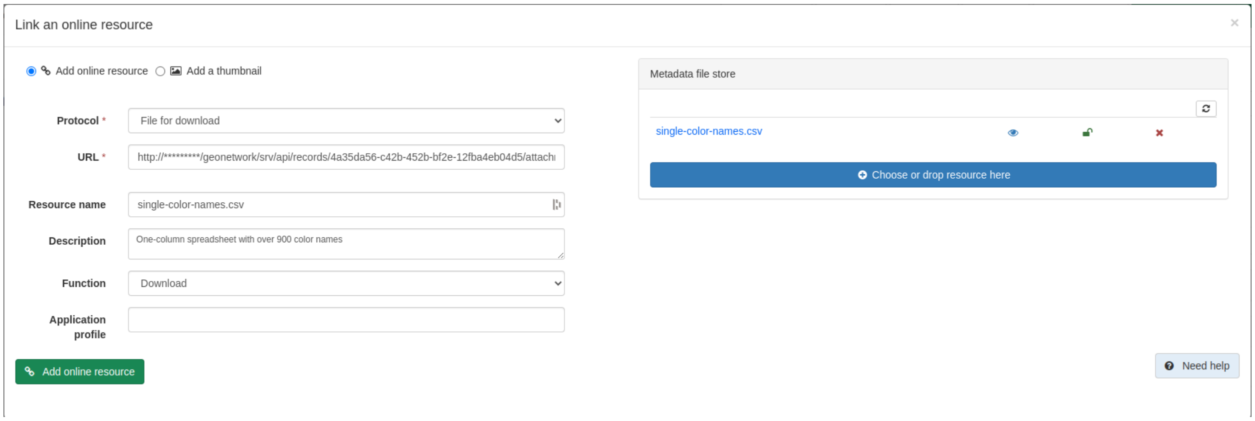
1.4.4 Uploading your dataset

In your non-spatial record editing view, use the **Associated resourcses** wizard in the top right and click **+Add**. From the list of options, choose **Link an online resource**.

In the **Metadata file store** section to the right, click the **+Choose or drop resource here** button to navigate to your CSV file. Once it is uploaded, click on its name in the list so that some of the options on the left (like the URL and Resource name) are auto-completed for you.

Fill in a Description, and choose **Download** from the list of Functions. You can leave the **Application profile** section blank.

Finally, click the **Add online resource** button.



Link an online resource

● Add online resource ○ Add a thumbnail

Protocol *: File for download

URL *: http://*********/geonetwork/srv/api/records/4a35da56-c42b-452b-bf2e-12fba4eb04d5/attachi

Resource name: single-color-names.csv

Description: One-column spreadsheet with over 900 color names

Function: Download

Application profile:

Important: GeoNetwork will check that the URL to the CSV file is reachable, and will show you an error message at the bottom if it is not. In that case, check the URL is correct.

1.4.5 Creating a Feature Catalog record from your dataset

A feature catalog describes the data model of the dataset with the list of tables, attributes, definitions, list of values, etc.

Feature catalogs can be described:

- as a document (e.g. PDF or CSV) and linked to the metadata record (see Linking a document)
- as a record and described using the ISO19115-3 standard (replacing ISO19110)

Warning: Users downloading a record will need to download the associated feature catalog (or any associated resources) separately as this is not currently downloaded at the same time by default.

Creating a link to a Feature Catalog

- Log into GeoNetwork as a user with Editor priviliges or higher and find the metadata record for your dataset
- Start an editing session of the metadata record in Default view
- On the right-hand side click the **Add** button of the **Associated resources** panel
- Choose **Link to a feature catalog** from the dropdown menu
- In the pop-up window that opens, use the search bar at the top to locate an existing feature catalog or insert a link to a remote catalog
- Click on the **Link to a feature catalog** button to link the two resources

If the steps above have been successful, you should see your linked Feature Catalog in the **Associated resources** panel on the right-hand side. Remember to save the changes before closing the window using the buttons at the top of the page.

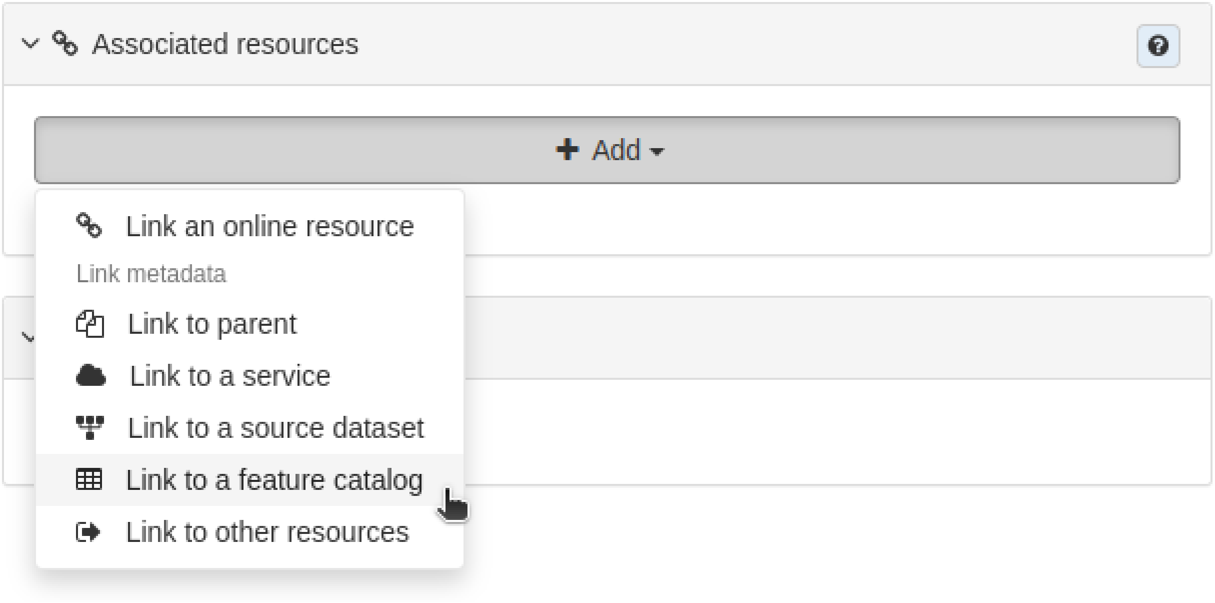
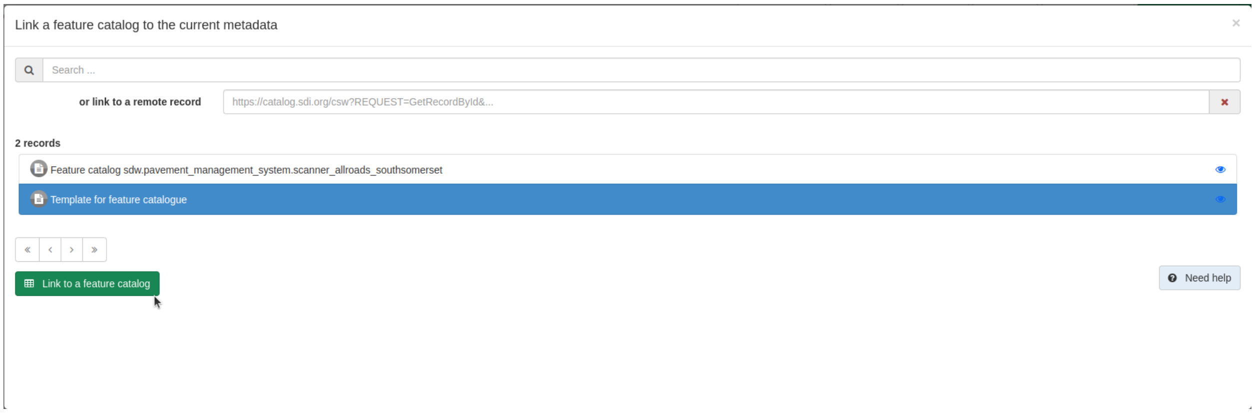


Fig. 1: Associated resources panel showing the dropdown menu options and the button to choose for linking a feature catalog


Link a feature catalog to the current metadata

Search ...

or link to a remote record: https://catalog.sdi.org/csw?REQUEST=GetRecordById&...

2 records

- Feature catalog sdw.pavement_management_system.scanner_allroads_southsomerset
- Template for feature catalogue

Fig. 2: “Link a feature catalog” pop-up window

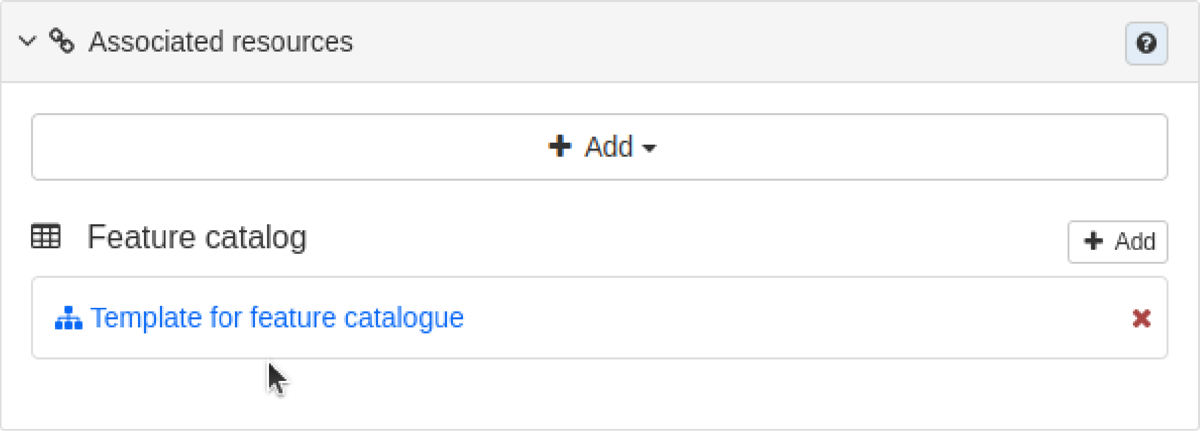


Fig. 3: Associated resources panel showing a linked feature catalog

1.5 Classification Systems

This section will explain how to add new thesauri to help with UK-specific data sharing as well as how to add custom categories to GeoNetwork.

A **thesaurus** is a **list of concepts** from a specialized field of knowledge. In a metadata catalog, concepts from a thesaurus can be assigned to a metadata record (as keywords) as a way of associating it with one or more concepts.

GeoNetwork has a concept of **categories** that can be assigned to metadata documents, but these are **not represented in the metadata**. So when the metadata is exported, the category will be lost.

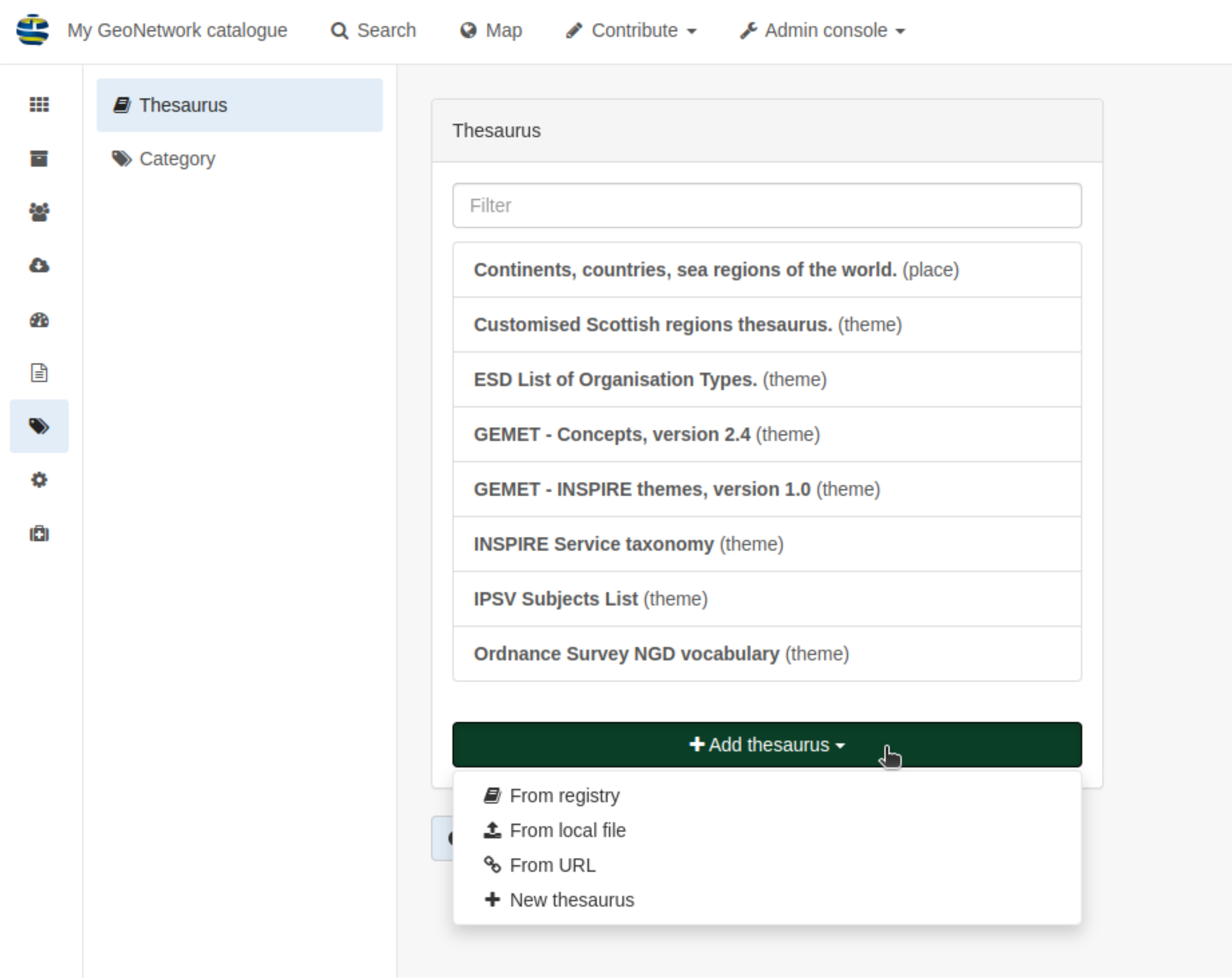
You can use these categories to separate records into groups, without changing the actual content of the metadata. Categories can also be used to filter search results, or limit the output of a custom CSW endpoint.

For further information on managing the classification systems please refer to the official GeoNetwork documentation.

1.5.1 Thesaurus

Login in as an administrator, and navigate to Admin Console > Classification systems > **Thesaurus**

Thesauri in SKOS format (XML or RDF extensions) can be managed or added here. It is also possible to interrogate the existing thesauri loaded into the catalog.



Continents, countries, sea regions of the world. (place)

Customised Scottish regions thesaurus. (theme)

ESD List of Organisation Types. (theme)

GEMET - Concepts, version 2.4 (theme)

GEMET - INSPIRE themes, version 1.0 (theme)

INSPIRE Service taxonomy (theme)

IPSV Subjects List (theme)

Ordnance Survey NGD vocabulary (theme)

It is possible to add additional thesauri by clicking the **+Add thesaurus** button. The options are as follows:

- From registry
- From local file - upload a thesaurus in SKOS format (XML or RDF extensions) from your local hard drive
- From URL - provide a link to a compatible thesaurus online
- New thesaurus - build one from scratch in Geonetwork

For more information on how to create or manage thesauri please refer to the official GeoNetwork documentation.

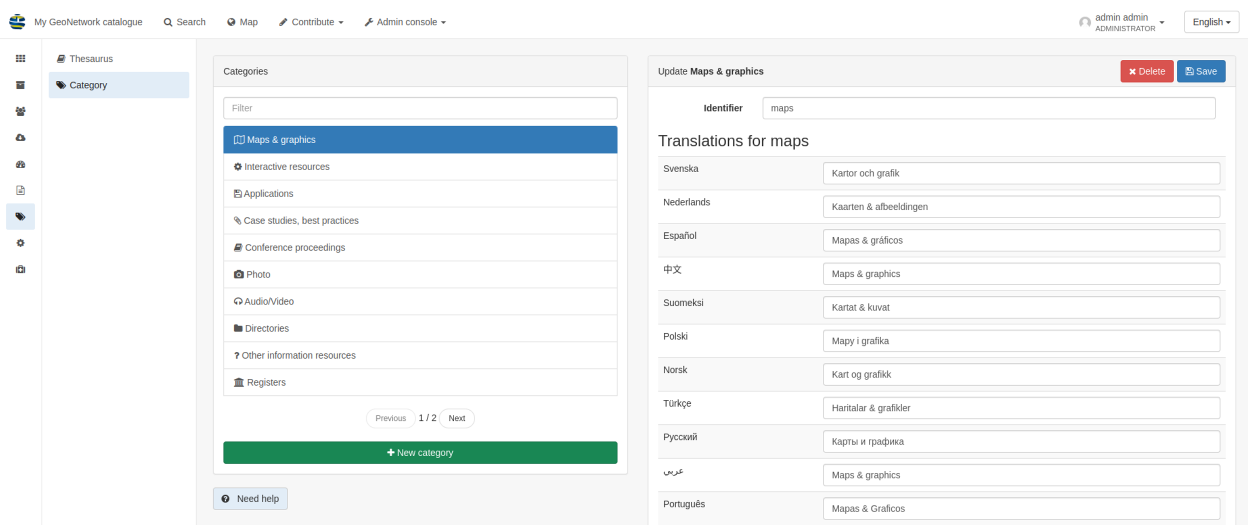
1.5.2 Category

Login in as an administrator, and visit Admin Console > Classification systems > **Category**

This page will show a list of the available categories as well as give teh option to add more by clicking on the **+ New category** button.

Each of the labels for the existing categories can be changed by clicking on the relevant category name. The label translations will open to the right. In order to change the name displayed in the catalog you will need to change the applicable label for the language(s) being used- commonly this will be English.

Important: Changing the **Identifier** value of a category will not change the wording associated with it in the catalog.



Categories

Filter

- Maps & graphics
- Interactive resources
- Applications
- Case studies, best practices
- Conference proceedings
- Photo
- Audio/Video
- Directories
- Other information resources
- Registers

Update **Maps & graphics**

Identifier maps

Translations for maps

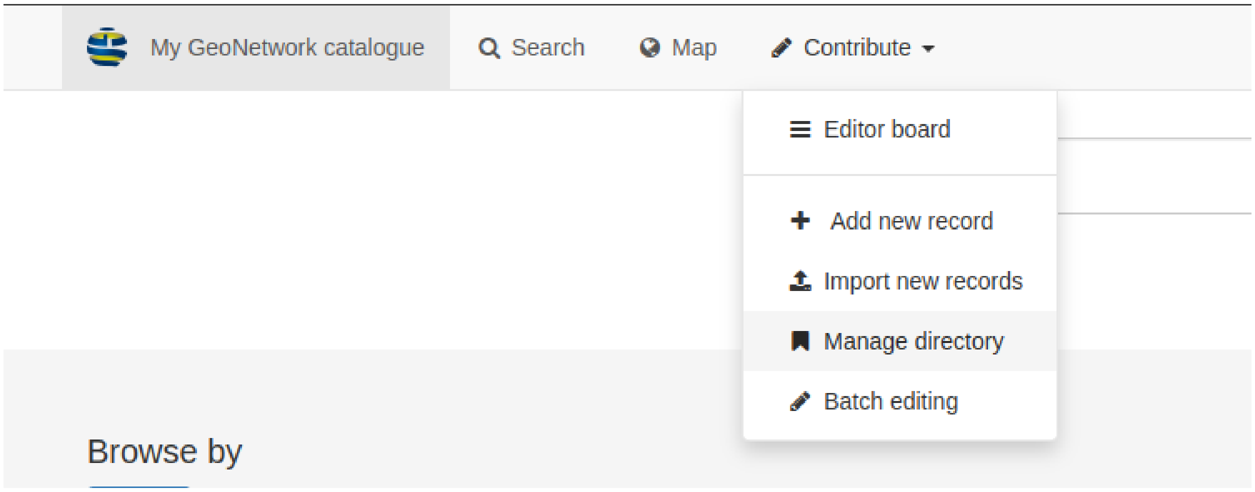
Svenska	Kartor och grafik
Nederlands	Kaarten & afbeeldingen
Español	Mapas & gráficos
中文	Maps & graphics
Suomeksi	Kartat & kuvat
Polski	Mapy i grafika
Norsk	Kart og grafikk
Türkçe	Haritalar & grafikler
Русский	Карты и графика
عربي	Maps & graphics
Português	Mapas & Graficos

1.6 Adding Snippets

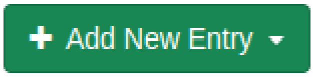
This section will explain how to add snippets from https://github.com/AstunTechnology/geonetwork-snippets to a GeoNetwork 4.2.x catalog. If you are using your own catalog, make sure that you have reviewed the instructions in GitHub to check that the additions to the **indexing templates** and the **editor layout** have been included.

1.6.1 Add Snippets to Directory

Login to the catalog as an Administrator and navigate to Contribute > **Manage directory**.



To add a snippet to the catalog from the **Manage directory** page:

- click on the  button

- select **Create an entry from scratch**
- in the textbox presented paste the XML snippet from https://github.com/AstunTechnology/geonetwork-snippets/tree/4.2.x/dataquality as needed
- click on the button

1.7 Structured Data

TODO: This section will explain about structured data embedded in both iso19139.nonspatial and iso19139.gemini23 for data-sharing and SEO

1.8 Publishing and sharing your data

TODO: This section will outline the various output formats (for download, and machine-readable)

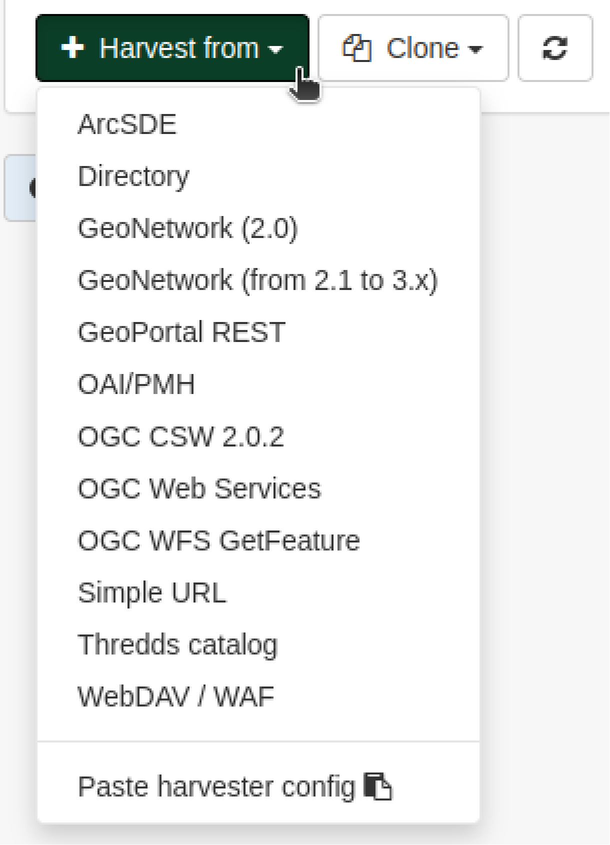
1.9 Harvesting

This section outlines the options available for harvesting spatial and non-spatial metadata from various endpoints.

To set up a new harvester in GeoNetwork 4.2.x login as an Administrator and go to Admin console > Harvesting > **Catalog harvesters**

To add a new harvester click on the **Harvest from** dropdown. The available sources are:

- ArcSDE
- Directory- this will harvest from a directory located on the same server as GeoNetwork
- GeoNetwork (2.0)
- GeoNetwork (from 2.1 to 3.x)
- GeoPortal REST
- OAI/PMH
- OGC CSW 2.0.2
- OGC Web Services
- OGC WFS GetFeature
- Simple URL
- Thredds catalog
- WebDAV/WAF



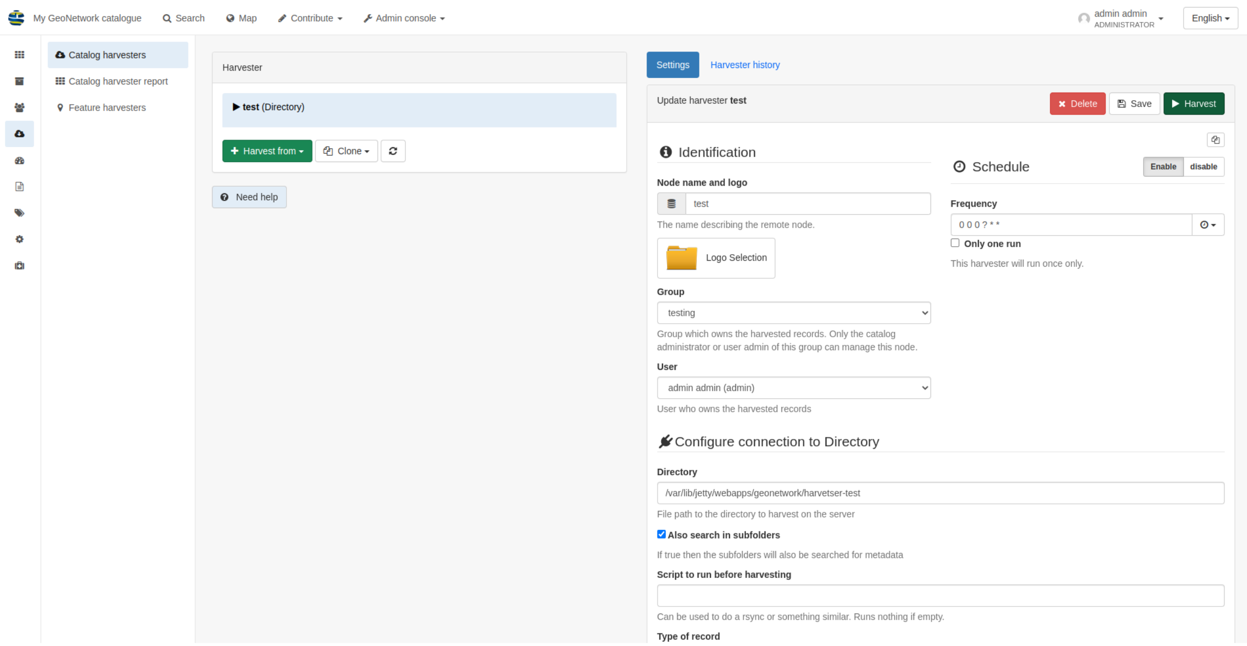
1.9.1 Directory harvesting

This option allows users to harvest records from the same server that runs GeoNetwork.

Important: If you are running GeoNetwork in a dockerised setup you will need to map the local directory to a volume and the path will become the mapped volume path. For example if the docker mapping is `./harvester-test:/var/lib/jetty/webapps/geonetwork/harvester-test`, then the path that the harvester needs to be pointed at has to be `/var/lib/jetty/webapps/geonetwork/harvester-test`

The main configuration options for a directory harvest are:

- Node name and logo- this is the name of the harvester and a logo
  - **Note:** in order to be able to associate a logo to the harvester, it needs to be pre-loaded into the catalog. This can be done in Admin console > Settings > Logo
- Group- the group which owns the harvested records
- User- a user can be picked from the list and will be the owner of the records
- Schedule- this feature can be enabled or disabled. If enabled, the user can set a recurring harvest
- Directory- this is the path to the Directory that holds the records
- Also search in subfolders- if ticked this will point the harvester to any existing subfolders too
- Action on UUID collision- this dictates what action will be taken if a UUID already exists in the catalog. This can be set to:



Update harvester **test**

Identification

Node name and logo

test

The name describing the remote node.

Group

testing

Group which owns the harvested records. Only the catalog administrator or user admin of this group can manage this node.

User

admin admin (admin)

User who owns the harvested records

Schedule

Frequency

0 0 0 ? * *

□ **Only one run**

This harvester will run once only.

Configure connection to Directory

Directory

/var/lib/jetty/webapps/geonetwork/harvetser-test

File path to the directory to harvest on the server

☑ **Also search in subfolders**

If true then the subfolders will also be searched for metadata

Script to run before harvesting

Can be used to do a rsync or something similar. Runs nothing if empty.

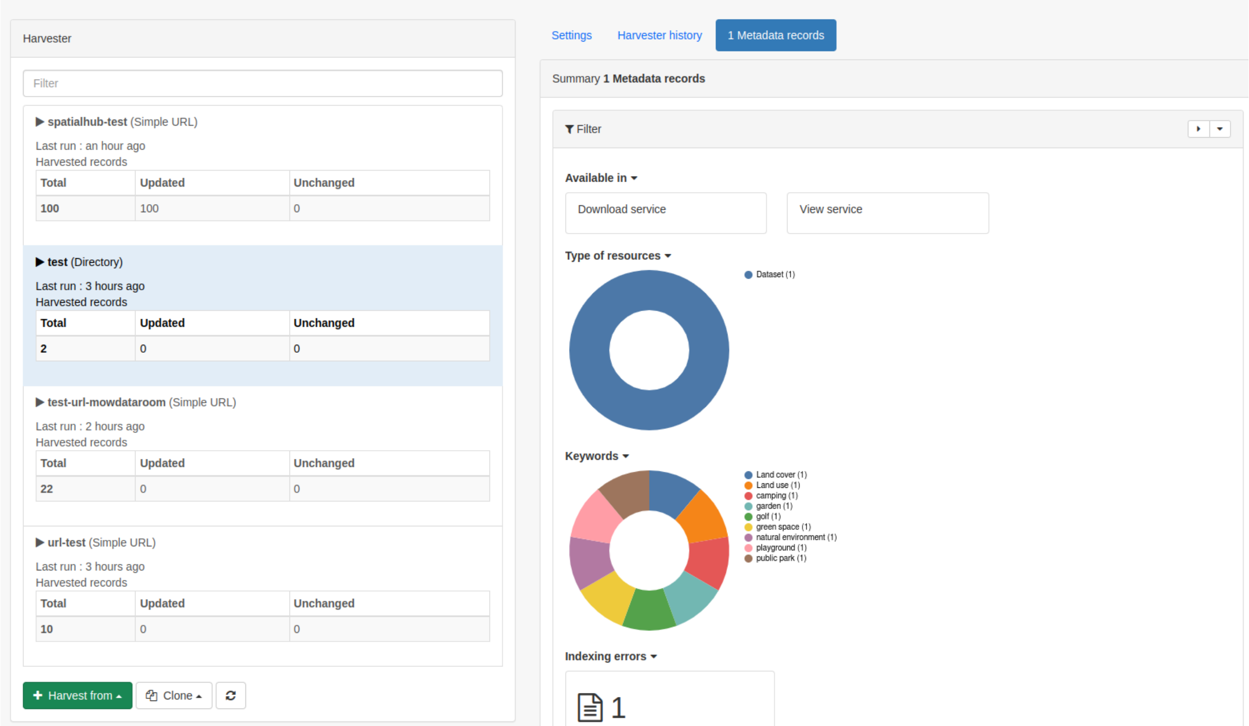
Type of record

Fig. 4: An example setup for harvesting in a dockerised setup

  - Skip record (default)
  - Overwrite record
  - Create new UUID

- Update catalog record only if file was updated (tickbox)
- Keep catalog record even if deleted at source (tickbox)
- Validate records before import- the default option is to accept all metadata without validation
- XSL transformation to apply
- Batch edits
- Category- the category to be allocated to the harvested records
- Group privileges for the harvested records

> **Warning:** This method has been tested in GeoNetwork 4.2.x and it successfuly harvested .ZIP and .XML records, however only .XML records are shown and accounted for in the **Metadata records** tab on the harvester page.



Harvester

Filter

▶ spatialhub-test (Simple URL)

Last run : an hour ago

Harvested records

Total	Updated	Unchanged
100	100	0

▶ test (Directory)

Last run : 3 hours ago

Harvested records

Total	Updated	Unchanged
2	0	0

▶ test-url-mowdataroom (Simple URL)

Last run : 2 hours ago

Harvested records

Total	Updated	Unchanged
22	0	0

▶ url-test (Simple URL)

Last run : 3 hours ago

Harvested records

Total	Updated	Unchanged
10	0	0

Summary **1 Metadata records**

Filter

Available in

Download service

View service

Type of resources

Keywords

Indexing errors

1

1.9.2 Simple URL harvesting

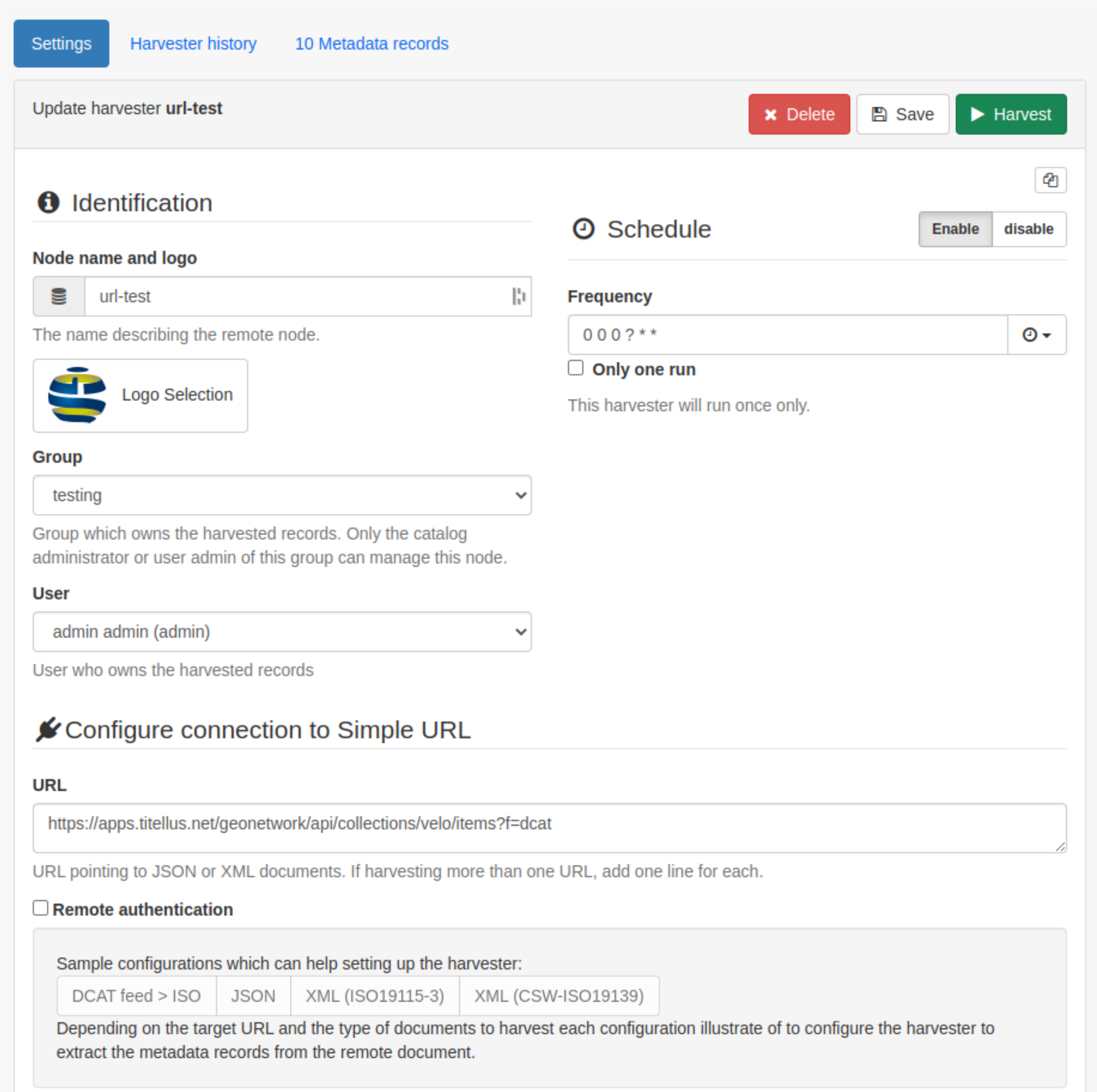
This option allows users to harvest records from various enpoints like DCAT/rdf or JSON (ESRI).

Important: You'll need to adapt the config to match the exact feed that you're trying to harvest- so manually look at it to identify the overarching dataset and identifier elements before continuing.

The main configuration options to set are:

- Node name and logo- this is the name of the harvester and a logo
  - **Note:** in order to be able to associate a logo to the harvester, it needs to be pre-loaded into the catalog. This can be done in Admin console > Settings > Logo
- Group- the group which owns the harvested records
- User- a user can be picked from the list and will be the owner of the records
- Schedule- this feature can be enabled or disabled. If enabled, the user can set a recurring harvest
- URL- path to endpoint for whole catalog (e.g. `https://apps.titellus.net/geonetwork/api/collections/velo/items?f=dcat`)
- Element to loop on- the XPath for the element that represents a dataset (e.g. `dcat:CatalogRecord`)
- Element for the UUID of each record- the element inside the dataset loop that should be used as the unique identifier (e.g. `./dct:identifier`)
- XSL transformation to apply- these are now done on a per schema basis, so find the correct file and add it as follows: `schema:{schemaname}:convert:{optional folder inside the schema's convert folder}/{filename without the xsl suffix}` (e.g. `schema:iso19115-3.2018:convert/DCAT/sparql-to-iso19115-3`)

- Batch edits
- Category- the category to be allocated to the harvested records
- Group privileges for the harvested records



Update harvester **url-test**

Identification

Node name and logo

url-test

The name describing the remote node.

Group

testing

Group which owns the harvested records. Only the catalog administrator or user admin of this group can manage this node.

User

admin admin (admin)

User who owns the harvested records

Schedule

Frequency

0 0 0 ? * *

□ **Only one run**

This harvester will run once only.

Configure connection to Simple URL

URL

https://apps.titellus.net/geonetwork/api/collections/velo/items?f=dcat

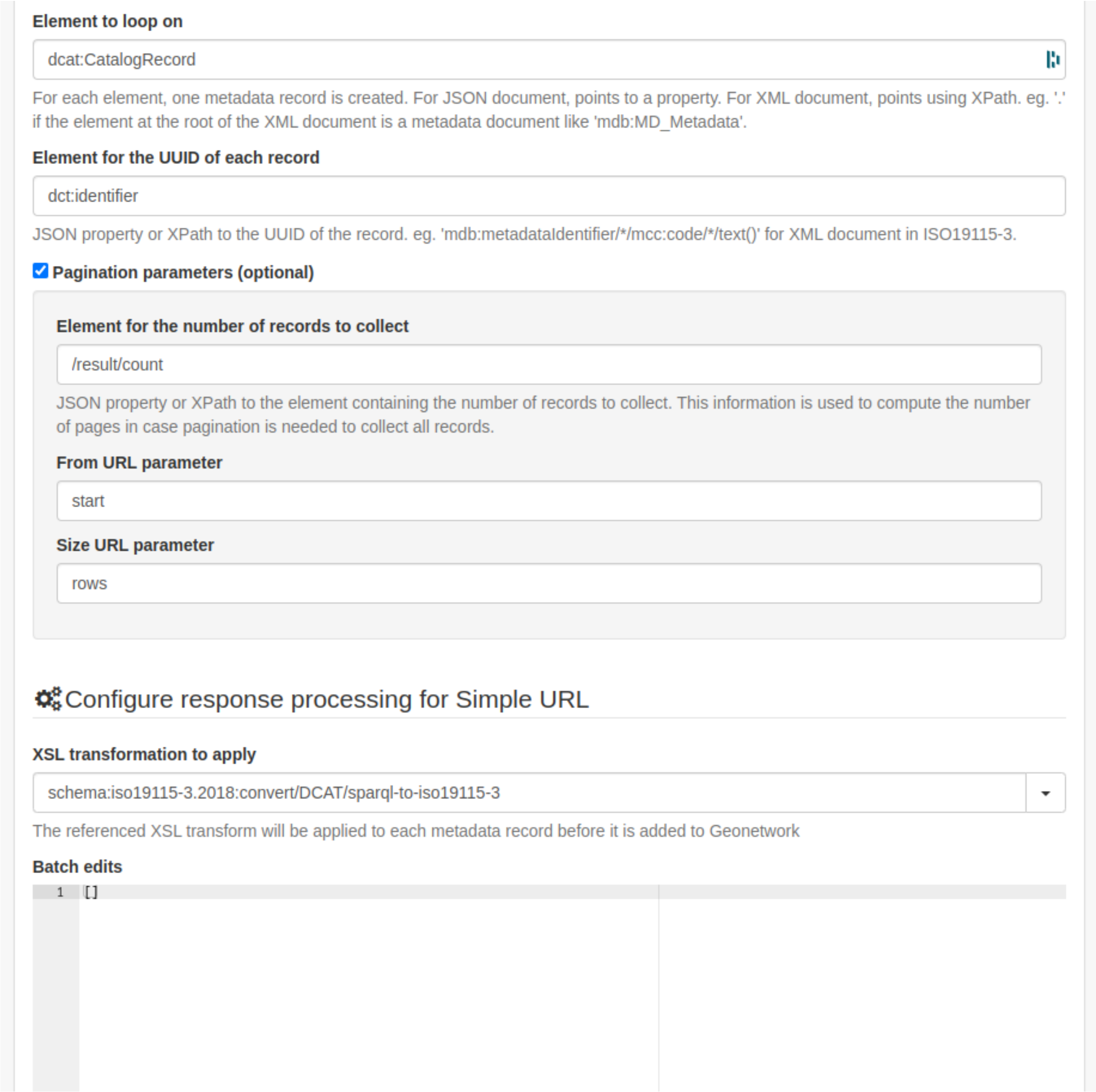
URL pointing to JSON or XML documents. If harvesting more than one URL, add one line for each.

□ **Remote authentication**

Sample configurations which can help setting up the harvester:

Depending on the target URL and the type of documents to harvest each configuration illustrate of to configure the harvester to extract the metadata records from the remote document.

Fig. 5: The top section of the configuration for an example Simple URL harvester



Element to loop on

dcat:CatalogRecord

For each element, one metadata record is created. For JSON document, points to a property. For XML document, points using XPath. eg. '.' if the element at the root of the XML document is a metadata document like 'mdb:MD_Metadata'.

Element for the UUID of each record

dct:identifier

JSON property or XPath to the UUID of the record. eg. 'mdb:metadataIdentifier/*/mcc:code/*/text()' for XML document in ISO19115-3.

☑ Pagination parameters (optional)

Element for the number of records to collect

/result/count

JSON property or XPath to the element containing the number of records to collect. This information is used to compute the number of pages in case pagination is needed to collect all records.

From URL parameter

start

Size URL parameter

rows

Configure response processing for Simple URL

XSL transformation to apply

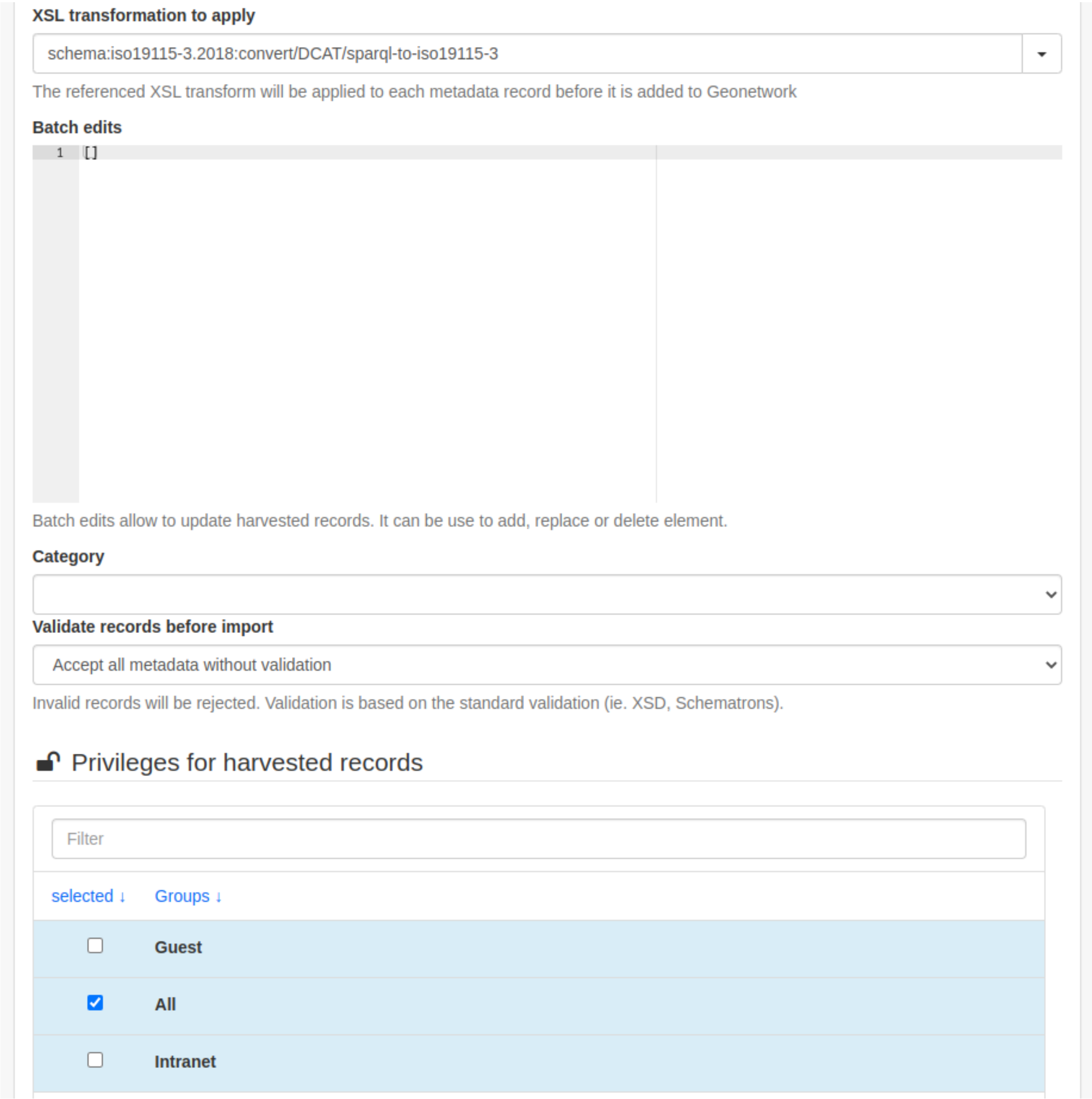
schema:iso19115-3.2018:convert/DCAT/sparql-to-iso19115-3

The referenced XSL transform will be applied to each metadata record before it is added to Geonetwork

Batch edits

```
[]
```

Fig. 6: The middle section of the configuration for an example Simple URL harvester



XSL transformation to apply

schema:iso19115-3.2018:convert/DCAT/sparql-to-iso19115-3

The referenced XSL transform will be applied to each metadata record before it is added to Geonetwork

Batch edits

```
[]
```

Batch edits allow to update harvested records. It can be use to add, replace or delete element.

Category

Validate records before import

Accept all metadata without validation

Invalid records will be rejected. Validation is based on the standard validation (ie. XSD, Schematrons).

Privileges for harvested records

Filter

selected ↓	Groups ↓
☐	**Guest**
☑	**All**
☐	**Intranet**

Fig. 7: The bottom section of the configuration for an example Simple URL harvester

1.10 Search-Engine Optimisation

TODO: This section will outline the SEO functionality built into GeoNetwork and the metadata profiles

CHAPTER TWO

INDICES AND TABLES

- genindex
- modindex
- search