

GeoNetwork Remote Harvesting Capabilities

Andrea Carboni

FAO

Summary

- Harvesting overview
- Improvements over the old protocol
- Mechanism
- Harvesting types
- Metadata logos
- Tips and notes

Harvesting Overview

- It is a way to share metadata between GeoNetwork nodes
- Remote metadata are stored locally and kept in sync
- It works with all metadata types supported by GeoNetwork
- Harvesting hierarchies are allowed for `type=geonetwork`

Improvements over GN2.0

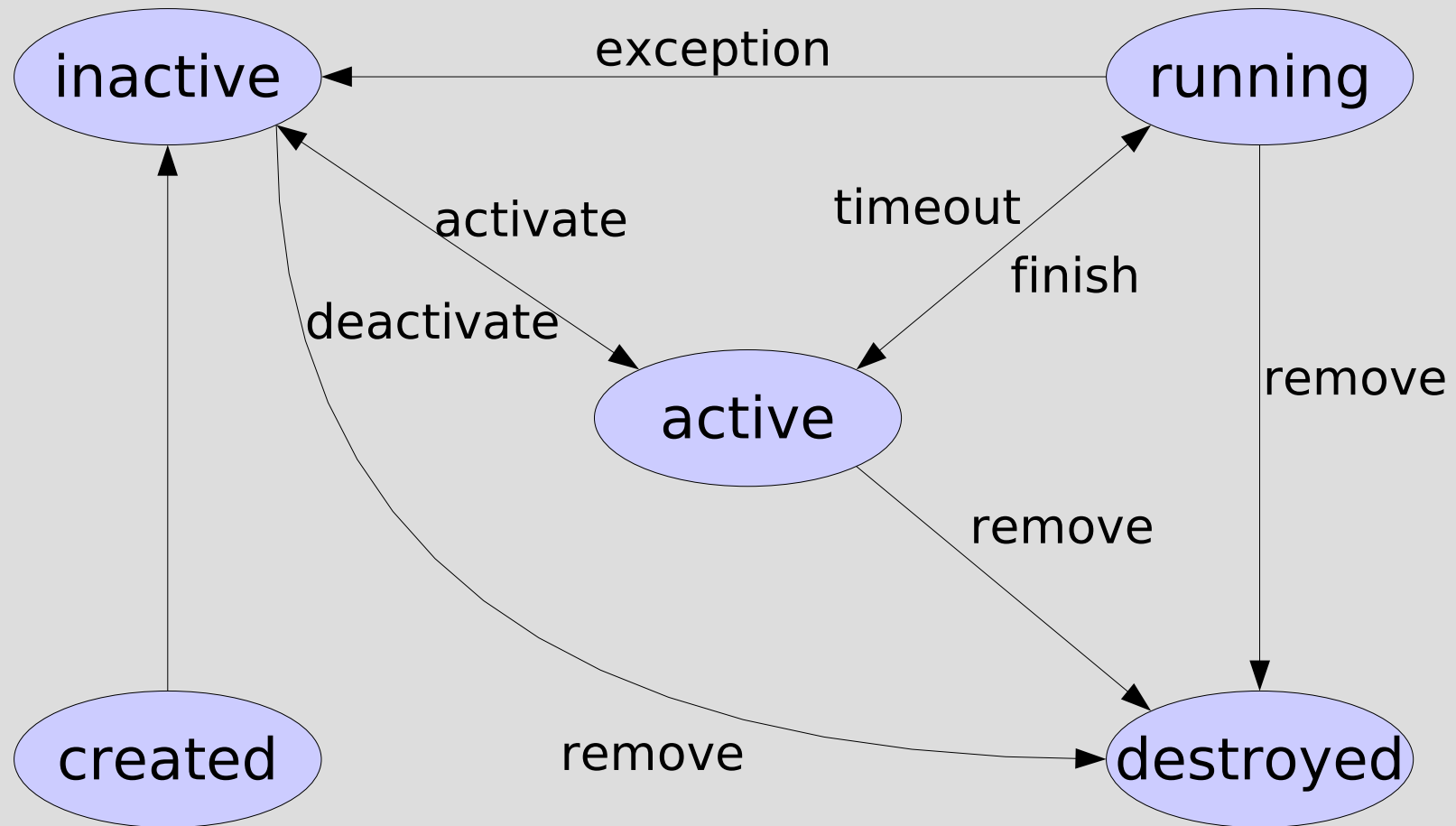
- Much more robust and reliable
- All GN's metadata schemas are supported
- Modular (easy to plug new types)
- Complete management through web interface
 - creation/change, [de]activation, run

Harvesting mechanism

Mechanism

- It uses UUIDs for metadata synchronization
- It uses metadata's last change date
 - If missing : metadata refetched every time
- This allows to:
 - Harvest a metadata, if not present locally
 - Update a local metadata, if remotely changed
 - Remove a local metadata, if remotely deleted
- The harvesting is incremental
- Privileges and categories can be assigned

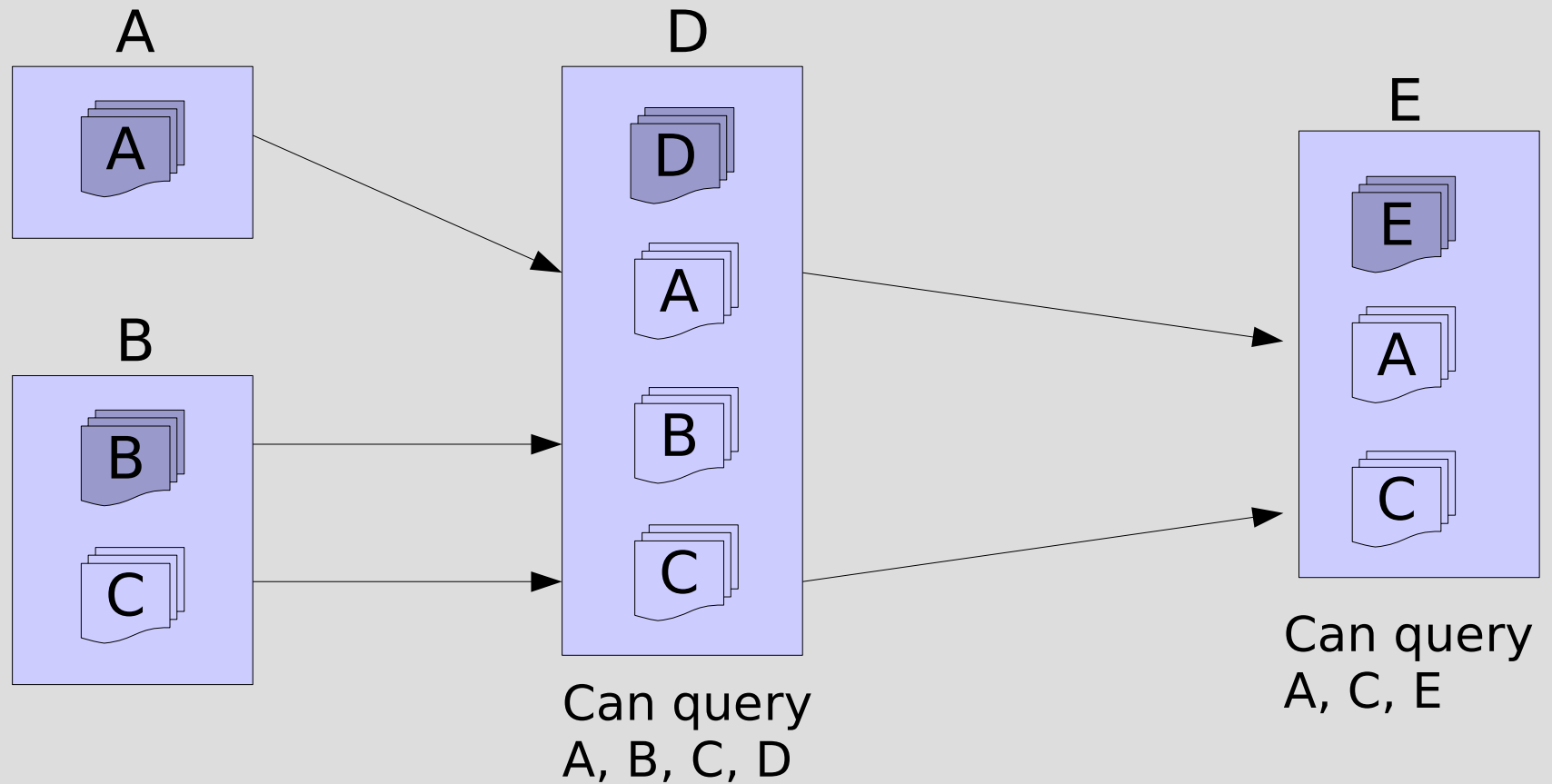
Harvesting nodes life cycle



Sources and Site IDs

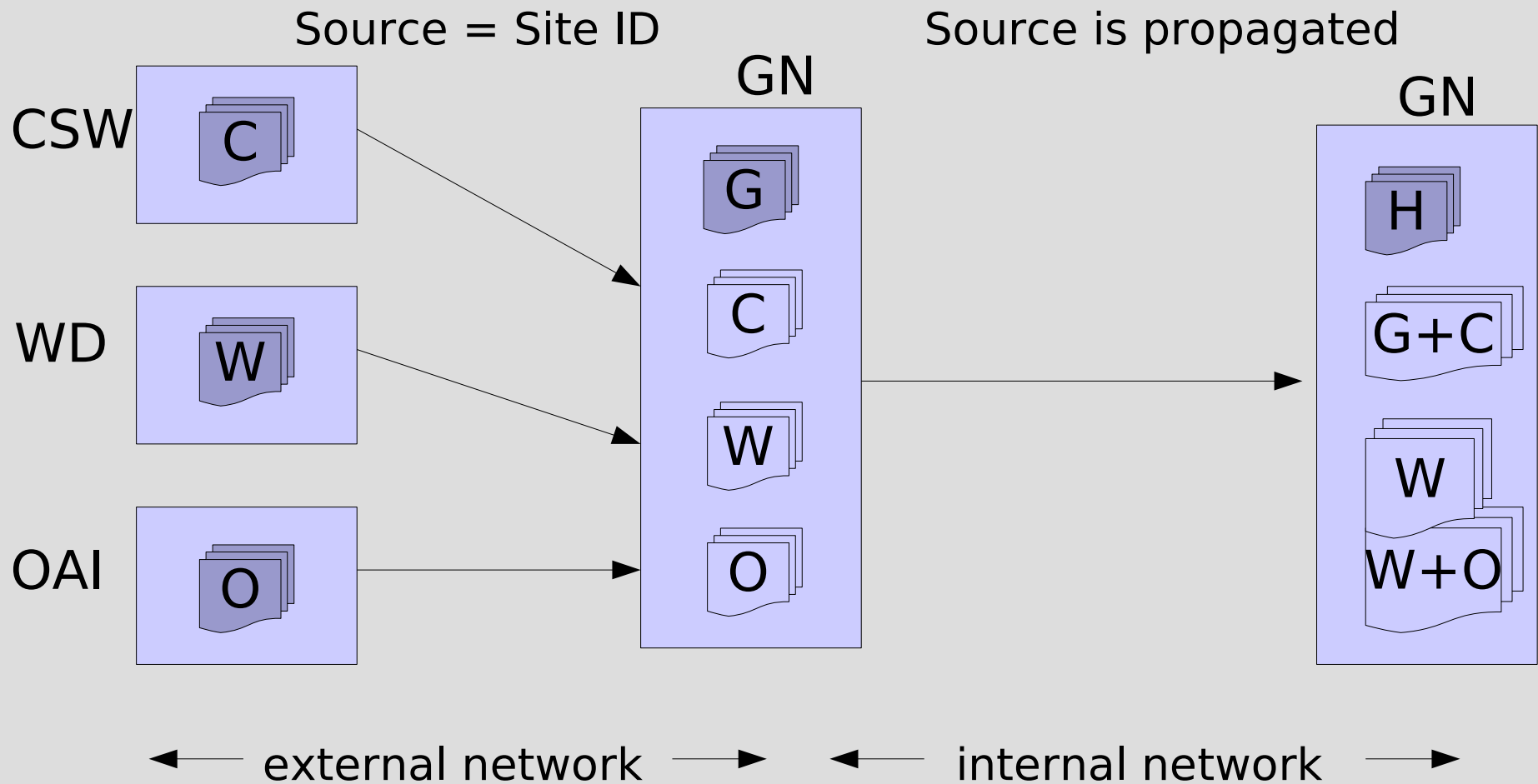
- A **source** is a metadata view on a remote server
- A **site id** identifies a GeoNetwork node
- A source *can be* equal to a site id
- Respect to harvesting protocols:
 - if internal (gn): its name comes from the remote node and can be a site id
 - if external (csw, oai, webdav): its name is set by the administrator

Old protocol case



All nodes are GeoNetwork nodes

New protocol case



Harvesting types

Supported types

- GeoNetwork 2.1
- CSW/2.0.1
- WebDAV
- OAI-PMH
- Other types (GN 2.0 & Z39.50)

GeoNetwork 2.1 harvesting type

Overview

- Native (all information is harvested)
 - Uses XML over HTTP to call remote services
 - Uses MEF format to export metadata
 - Privileges can be imported through policies
- General or source based search
- Allows hierarchical harvesting
- Allows harvested metadata to be rated

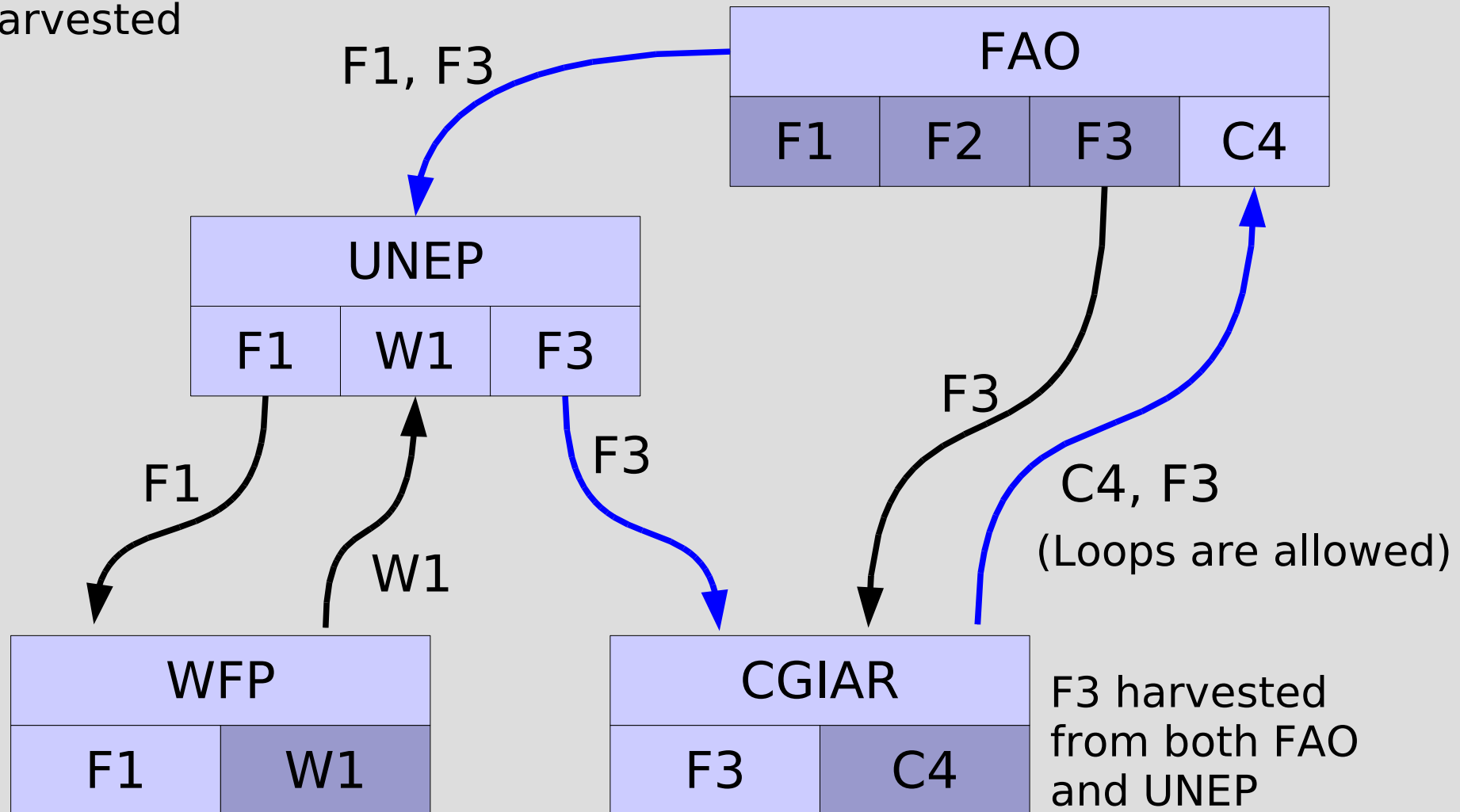
Protocol

- Login on remote node (opt)(xml.user.login)
- Info retrieval (sources, groups) (xml.info)
- Search and merge results (xml.search)
- Alignment
 - Old metadata removal based on UUID
 - Insert/update of metadata and thumbnails (mef.export)
 - Update of assigned categories and group privil
- Update source names and logo retrieval

An harvesting hierarchy

Local

Harvested



Catalogue Services for the Web harvesting type

Overview

- CSW is a specification for web services made by Open Geospatial Consortium
- Implementation does not use harvesting services
- Main queryables can be searched : title, abstract, subject, any
- <http://www.opengeospatial.org/standards/cat>

Protocol

- Call GetCapabilities
 - Retrieve GetRecords and GetRecordById oper
- Search and merge results (GetRecords)
 - Autodetect for GET/CQL and POST/Filters supp
 - Several calls to scroll the result set

Protocol (2)

- Alignment
 - Old metadata removal based on UUID
 - Insert/update of metadata based on change date (GetRecordById)
 - Autodetect for GET or POST bindings
 - Autodetect metadata schema
 - Update of categories and privileges

Limitations

- If dct:modified is missing in a metadata, that metadata is always harvested
- Based on ambiguous 2.0.1 spec: probable communication problems

Web DAV

harvesting type

Overview

- Extensions to HTTP protocol to manage files on remote servers
- Implementation uses Jakarta Slide library
- Retrieves all ".xml" files
- It is not a *Web Accessible Folder*
- <http://www.webdav.org>

Protocol

- Scanning (recursive) of all files
- Alignment
 - Old metadata removal based on URI
 - Insert/update of metadata based on HTTP and change date
 - Autodetect metadata schema
 - Update of categories and privileges

Limitations

- There are some bugs on the library (it is 3 years old)
- The same metadata could be harvested many times creating duplicates

OAI-PMH

harvesting type

Overview

- Protocol designed to harvest metadata (NASA, Cornell University and others)
- Implementation uses our own library
- Search by: from, until, set, prefix
- <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- GeoNetwork can act as an OAIPMH server

Protocol

- Search and merge results (ListIdentifiers)
- Alignment
 - Old metadata removal based on remote id
 - Insert/update of metadata based on change date (GetRecord)
 - Autodetect metadata schema
 - Conversion of oai_dc into dublin core
 - Response validated against oai schema
 - Update of categories and privileges

Limitations

- The remote id must be a UUID

Other harvesting types

Other harvesting types

- GeoNetwork 2.0
 - deprecated, not reliable
 - harvested metadata loose their site id
- Z39.50
 - Under development

Logos, tips and notes

Metadata logos

- Logos are located in `<gn>/images/logos` and their names are UUIDs
- Beside the local one, there is one logo for each harvesting node. Images are taken from `<gn>/images/harvesting`
- Logos are shown in search results page:
 - local metadata: site logo
 - harvested from GN : logos are harvested
 - harvested from GN20 : fixed image
 - generic harvesting : set by user interface

Tips

- Use a different timeout in each harvesting node. This distributes the harvesting load over time

Notes

- When a node is changed, the timeout is restarted
- On errors, the node is deactivated
- Harvesting results are not stored. They are lost if the server reboots.
- Harvested metadata are readonly. When the node is deleted they are removed.
- Settable privileges: view, dynamic, featured

That's all