Technical Requirements

Harvest Tier 3 datasets into an aggregator node based on metadata content

Stephen M Richard and Christy Caudill, US Geoscience Information Network

Jon Weers, National Renewable Energy Lab

Objective: When harvesting from a source, add process to determine if a data is available in a Tier 3 CSV-format file. If so, harvest the data into the harvesting CKAN node and publish as NGDS services.

Precondition: ISO19139 metadata conforming to encoding requirements specified here is available in a CSW or WAF for harvesting into and NGDS aggregator. To be eligible for aggregator service deployment, datasets submitted through GDR must be accessible online in a CSV format. Metadata records for these datasets must indicate which tier 3 data type (content model) for the dataset and new web service (in TransferOptions, see below).

Requirements: Tier 3 dataset has metadata conforming to the description in the GDR requirements section (below), and is accessible online (HTTP GET) in a CSV format. If the file validates against the USGIN validation service (<http://schemas.usgin.org/validate/cm>), it will be deployed as OGC WMS and WFS services from the aggregator node.

## Geothermal Data Repository (GDR) - NGDS Portal Integration

### NGDS Requirements:

Implement function for NGDS Portal CKAN build to enhance the metadata harvest process as follows:

* When a harvested record is imported, scan the metadata to determine if the described resource is available from the publisher as a csv file conforming to one of the USGIN content model schemas (Tier 3 data). A current list of content models and their identifiers can be obtained at <http://schemas.usgin.org/contentmodels.xml> (also available as .json or .html). Convention for how this will be indicated in the metadata is discussed below (Figure 1).
* If the resource is accessible in a Tier 3 .csv file, get the file and run validator
* If it passes validation, import the csv file to the node, and deploy OGC services to Geoserver. This will be exactly the same process that runs when someone uploads a Tier 3 dataset to an NGDS CKAN node using the UI.
* If service deploys successfully, update the distribution elements in the metadata in the NGDS portal catalog to include the new OGC service distributions.
  + Maintain link to original file on GDR
* Add UI control to harvest configuration to determine if this process will be run on records from that source. This will allow the user to indicate that this process is necessary for the given harvest.
* Code to be developed in a new repository in the NGDS Organization GitHub.
* Code should include unit tests to validate functionality.

### GDR Requirements:

1. GDR submission process will need to identify the tier 3 content model used.
   1. USGIN suggests pulling this list dynamically from the content model register (<http://schemas.usgin.org/contentmodels.xml> or http://schemas.usgin.org/contentmodels.json).
2. Metadata harvested from GDR **must** be in ISO 19139 XML format, and must include a *distributionInfo//distributor* element that matchs the example in Figure 1. below. Figure 2 is an example of the recommended metadata content in the *distributor* element. The required content includes 3 elements:
   1. gmd:distributorFormat//gmd:name/gco:CharacterString = csv
   2. gmd:distributorTransferOptions//gmd:applicationProfile/gco:Character­String = ‘*uri for a known Tier 3 content model*’
   3. gmd:distributorTransferOptions//gmd:linkage/gco:URL = ‘*url to download the csv file*’

### Figure 1. Mandatory content in ISO gmd:MD\_Metadata record for NGDS deployment of Tier 3 services from GDR data. The onlineFunctionCode element is required for XML schema validation. [Note—these are updated in version 2 (this version) reflecting testing during implementation].

<!—in the keywords section of gmd:MD\_DataIdentification, a usgincm: namespace keyword identifying the content model; these keywords are summarized at <https://github.com/ngds/documents/wiki/usgincm:-Content-Model-Keywords-List>; keywords are not case sensitive -->

<gmd:descriptiveKeywords>  
 <gmd:MD\_Keywords>  
 <gmd:keyword>  
 <gco:CharacterString>usgincm:well log observation</gco:CharacterString>  
 </gmd:keyword>  
 <gmd:type>  
 <gmd:MD\_KeywordTypeCode codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO\_19139\_Schemas/resources/Codelist/gmxCodelists.xml#MD\_KeywordTypeCode" codeListValue="theme">theme</gmd:MD\_KeywordTypeCode>  
 </gmd:type>  
 </gmd:MD\_Keywords>  
 </gmd:descriptiveKeywords>

…

<!-- The DigitalTransferOptions element may be in a gmd:transferOptions or gmd:distributorTransferOptions property -->  
 <gmd:MD\_DigitalTransferOptions>  
 <gmd:onLine>  
 <gmd:CI\_OnlineResource>  
 <gmd:linkage>  
 <gmd:URL>**http://url to get csv file**</gmd:URL>  
 </gmd:linkage>

<gmd:applicationProfile>

<!-- uri for the content model from <http://schemas.usgin.org/contentmodels.json>; MUST identify the version -->  
 <gco:CharacterString>

**http://stategeothermaldata.org/uri-gin/aasg/xmlschema/welllog/0.8**</gco:CharacterString>  
 </gmd:applicationProfile>

<!—Client applications MUST identify the proper distribution by searching for “NGDS Tier 3 Data, csv format” VERBATIM, in the online resource name element. -->  
 <gmd:name>  
 <gco:CharacterString>**NGDS Tier 3 Data, csv format**:nmwelllog.csv</gco:CharacterString>  
 </gmd:name>  
 <gmd:function>  
 <gmd:CI\_OnLineFunctionCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI\_OnlineFunctionCode" codeListValue="download">download</gmd:CI\_OnLineFunctionCode>  
 </gmd:function>  
 </gmd:CI\_OnlineResource>  
 </gmd:onLine>  
 </gmd:MD\_DigitalTransferOptions>

### Figure 2. Recommended content in ISO gmd:MD\_Metadata record for Tier 3 datasets from GDR data:

<gmd:distributionInfo>  
 <gmd:MD\_Distribution>

<gmd:distributor>  
 <gmd:MD\_Distributor>  
 <gmd:distributorContact>  
 <gmd:CI\_ResponsibleParty>  
 <gmd:organisationName>  
 <gco:CharacterString>Geothermal Data Repository (GDR), National Renewable Energy Laboratory (NREL)</gco:CharacterString>  
 </gmd:organisationName>  
 <gmd:contactInfo>  
 <gmd:CI\_Contact>  
 <gmd:address>  
 <gmd:CI\_Address>  
 <gmd:electronicMailAddress>  
 <gco:CharacterString>e-mail for GDR curator at NREL</gco:CharacterString>  
 </gmd:electronicMailAddress>  
 </gmd:CI\_Address>  
 </gmd:address>  
 </gmd:CI\_Contact>  
 </gmd:contactInfo>  
 <gmd:role>  
 <gmd:CI\_RoleCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI\_RoleCode" codeListValue="distributor"></gmd:CI\_RoleCode>  
 </gmd:role>  
 </gmd:CI\_ResponsibleParty>  
 </gmd:distributorContact>  
 <gmd:distributorFormat>  
 <gmd:MD\_Format>  
 <gmd:name> <!—explicit indication that a CSV distribution is available -->  
 <gco:CharacterString>csv</gco:CharacterString>  
 </gmd:name>  
 </gmd:MD\_Format>  
 </gmd:distributorFormat>  
 <gmd:distributorTransferOptions>  
 <gmd:MD\_DigitalTransferOptions>

<gmd:onLine>  
 <gmd:onLine>  
 <gmd:linkage>  
 <gmd:URL>**http://url to get csv file**</gmd:URL>  
 </gmd:linkage>

<gmd:applicationProfile>

<!-- uri for the content model from <http://schemas.usgin.org/contentmodels.json>; should identify the version! -->  
 <gco:CharacterString>

**http://stategeothermaldata.org/uri-gin/aasg/xmlschema/welllog/0.8**</gco:CharacterString>  
 </gmd:applicationProfile>

<!-- Client applications MUST identify the proper distribution by searching for “NGDS Tier 3 Data, csv format” VERBATIM, in the online resource name element. -->  
 <gmd:name>  
 <gco:CharacterString>**NGDS Tier 3 Data, csv format**:nmwelllog.csv</gco:CharacterString>  
 </gmd:name>

<gmd:description>  
 <gco:CharacterString>NGDS Tier 3 dataset in comma-delimited table format</gco:CharacterString>  
 </gmd:description>  
 <gmd:function>  
 <gmd:CI\_OnLineFunctionCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI\_OnlineFunctionCode" codeListValue="download">download</gmd:CI\_OnLineFunctionCode>  
 </gmd:function>  
 </gmd:CI\_OnlineResource>  
 </gmd:onLine>  
 </gmd:MD\_DigitalTransferOptions>  
 </gmd:distributorTransferOptions>  
 </gmd:MD\_Distributor>  
 </gmd:distributor>  
 </gmd:MD\_Distribution>  
 </gmd:distributionInfo>