Technical Requirements

Tier 3 Excel publication to OGC services upon harvest of datasets into NGDS CKAN

Stephen M Richard and Christy Caudill, US Geoscience Information Network

Jon Weers, National Renewable Energy Lab

*revised 2015-07-01*

Objective: When metadata harvesting from Geothermal Data Repository (GDR) or any source to the NGDS aggregator or an NGDS GINstack node, add process to determine if a data is available in a Tier 3 content model Excel file. If so, convert the Excel file to a publishable format (CSV) then publish dataset automatically to schema-compliant OGC web services.

Precondition: ISO19139 metadata conforming to encoding requirements specified here is available in a CSW or WAF for harvesting into NGDS. To be eligible for service deployment, datasets submitted through GDR must be accessible online, in a content model in the approved Excel workbook format (from <http://schemas.usgin.org/models/>), or the Data tab of such workbook in CSV format. Metadata records for these datasets must indicate: 1) USGIN content model type keyword; 2) URL to the dataset; 3) content model URI for calling usginmodels API; and 4) the name of the file, prefixed with “NGDS Tier 3 Data,”. See these components in Figure 1; example text is in black, bold text.

Requirements: Tier 3 dataset has metadata conforming to the description in the GDR requirements section (below), and is accessible online (HTTP GET) in an approved content model. As described in the NGDS requirements, NGDS will add functionality to convert the data tab of the Excel-formatted content model to CSV format. If the CSV format of the dataset file validates against the USGIN validation service (<http://schemas.usgin.org/validate/cm>), it will be deployed as OGC WMS and WFS services during harvesting. If the file does not validate against the specified USGIN schema, it will still be harvested in as the original file but not published to web services.

### 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 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 identified as an Excel format content model dataset, read the About tab of the content model to retrieve: 1) typeLayerName (as from the given data type schema, this cell has a DefinedName that is the same as the worksheet tab containing the data for this feature type); 2) usginKeyword; 3) dataTab (name of the tab of the Excel workbook containing the data to be transformed to CSV).
* Convert the data tab of the Excel file to a CSV file.
  + Type and version of Excel files to convert can be narrowed down through: 1) Use of Microsoft Excel 97-2003 Worksheet format, as is the specified format at <http://schemas.usgin.org/models/>; 2) Use of specific Excel parsing libraries, such as:
    - PHPExcel: <https://phpexcel.codeplex.com/> (PHP; GNU license)
    - Roo: <https://github.com/roo-rb/roo> (Ruby gem; MIT license)
    - Python-Excel org: <http://www.python-excel.org/>
      * XLRD is probably the best: <https://github.com/python-excel/xlrd>
* Use this CSV file, or if the resource is already accessible in a Tier 3 CSV file as indicated in the metadata, run USGIN validator on the CSV.
* 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. Content model keywords and URI quick look-up can be found at <https://github.com/ngds/documents/wiki/usgincm:-Content-Model-Keywords-List>.
2. Metadata harvested from GDR **must** be in ISO 19139 XML format, and must include a *distributionInfo//distributor* element that matches the example in Figure 1 below. The required content includes four elements:
   1. gmd:keyword// gco:CharacterString = ‘*USGIN content model keyword*’
   2. gmd:distributorTransferOptions//gmd:linkage/gco:URL = ‘*url to download the csv file*’
   3. gmd:distributorTransferOptions//gmd:applicationProfile/gco:Character­String = ‘*USGIN content model URI*’
   4. gmd:distributorFormat//gmd:name/gco:CharacterString = ‘*NGDS Tier 3 Data, csv format: FileName.csv* or *FileName.xls*’

### 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.

<!—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>