Excel file shape configuration

Shape of the excel file is set at the beginning of the script for the 4 sheets of the excel file :

- *MD Fields* (column and row start)
- *Help* (associate columns with their names)
- *MD generic* (row start and associate columns with their names)
- *MD Thesaurus* (col start and associate columns with their names)

Be careful always to keep this values consistent with excel file shape, particularly for the delta between *MD Fields* columns and associated *Help* rows.

The addition of a column in *MD Field* or a row in *Help* does not impact this shape configuration.

Read and parse excel file with xlrd

Get sheets and check excel file shape :

- Do *MD Fields* ID match with *Help* ID ?
- Are mandatory fields filled ?

If an error is identified the script stops and an appropriate error

message is displayed.

Parse the template as an XML tree with etree

Add metadata elements

An element is added only when its value is not null.

1. Generic metadata (*MD generic* sheet)

Each row of *MD generic* sheet stands for a metadata element. The element's value is added at the location specified by the xpath. A codelist and an attribute can also be added.

For some elements - identified thanks to their tag name - some other tags are added :

- **Resource locator url**

  addition of online resource protocol *WWW:LINK-1.0-http–link*

2. Non generic metadata (*MD Fields* and *Help* sheets)

An XML metadata file is generated for each row of MD Field sheet.

For each row :

- *MD Fields* columns are parsed one by one to get the values ;
- linked *Help* rows are parsed to get :
    - xpath
    - attribute (is the metadata element an attribute)
    - thesaurus name
    - multivalue (do the tag need to be added only once or several times)
    - codelist value
    - type (Date or Keyword)
    - keyword attribute ID

Regular case is to add the metadata value (tag or attribute) at the location specified by the xpath.

Other elements can be added according to which *Help* cells are filled :

- attribute id for Free Keywwords ;
- codeList and codeListValue ;
- thesaurus name, date, dateType and codeList (information available in *MD Thesaurus* sheet) ;
- Date elements : dateType and codelist ;
- Keyword elements : KeywordType and codeList.

Some values are kept in memory such as :

- URN ;

- resource Title ;
- file name pattern.

Some elements need a specific processing :

- **URN** [identified with its xpath]

  - *MD Field* Unique identifier value is replaced by the URN - the concatenation of *MD generic* UID and *MD Fields* UID
  - 2 *MD generic* tags are replaced by the concatenation of their value and the URN :
    - location (address) for on-line access
    - permanent link
  - If *MD generic* field *OpenWIS only: DCPC local data source* is not null, two online resources are added (Suscribe on DCPC and Request on DCPC)

- **GTS Priority** [identified with its xpath]
  Value is replaced by *"GTSPriorityN"* with *N* the 9th character of *MD Field* cell value

- **File name** [identified with its xpath]
  GFNC file name and associated information are added :

  - file description (dynamic value : Resource title)
  - file type (static value)
  - file format name (static value)
  - file format version (static value)

- **Free links** [identified with its xpath]

  Several links can be added. 3 tags are added for each link :

    - name (optional, URL value if not specified)

    - protocol (static value)

    - URL (dynamic value)

- **Resource Format** [identified with its xpath]

  Several resource formats can be added. 3 tags are added for each resource format (dynamic values) :

    - name

    - version

    - specification (optional)

- **Tag or group of tags added several times** [Mutlivalue set to Yes]

  Add a group of tags as many times as the number of comma-separated values. First multivalued tag to add is suffixed by *"[]"*.

Warn messages

Warn message can be displayed for each metadata in case where :

- there is a field value but no xpath ;
- a xpath is incorrect.