TikaGeographicInformationParser

TikaGeographicInformationParser

Currently Apache Tika lacks the required support to parse .iso19139 files that are crawled from the Acadis websites.There has been a issue that has been created by Prasanth Iyer (https://issues.apache.org/jira/browse/TIKA-1479) TIKA-1479.

The Progress is as below

1. Extract the Meta Data using Apache SIS library (Martin has been a great source of support in this regard).

2. Customize the Meta Data extracted to construct Meta Data as key multi-value.

3. The format finalized so far has been key1->[val1,val2..] , key2->[val1,val2...].

I would like suggestions on the below progress. Default Meta Data extracted from Apache SIS framework is as below

Default Meta Data

Metadata
 +-Character set…… UTF-8

 +-Contact

 ¦ +-Role…… Resource provider

 ¦ +-Party

 ¦   +-Name…… UCAR/NCAR - CISL - ACADIS

 +-Identification info

 ¦ +-Citation

 ¦ ¦ +-Title…… Barrow Atqasuk ARCSS Plant

 ¦ ¦ +-Date (1 of 2)

 ¦ ¦ ¦ +-Date…… Dec 16, 2013 12:00:00 AM

 ¦ ¦ ¦ +-Date type…… Creation

 ¦ ¦ +-Date (2 of 2)

 ¦ ¦ ¦ +-Date…… Feb 5, 2015 12:00:00 AM

 ¦ ¦ ¦ +-Date type…… Modified

 ¦ ¦ +-Cited responsible party

 ¦ ¦   +-Role…… Point of contact

 ¦ ¦   +-Party

 ¦ ¦   +-Name…… Robert Hollister

 ¦ ¦   +-Contact info

 ¦ ¦   +-Address

 ¦ ¦   +-Electronic mail address…… hollistr@gvsu.edu

 ¦ +-Abstract…… These files contain data representing the periodic plant measures of species within each plot in a text tab delimited format.
The data presented are seasonal growth of graminoids (length of leaf and length of inflorescence) and seasonal flowering of all species (number of inflorescences in flower within a plot), collected weekly during the summers of 2012-20XX for a subset of 30 grid plots at two sites (Barrow ARCSS grid and Atqasuk ARCSS grid).

 ¦ +-Status…… On going

 ¦ +-Point of contact

 ¦ ¦ +-Role…… Point of contact

 ¦ ¦ +-Party

 ¦ ¦   +-Name…… Robert Hollister

¦ ¦ +-Contact info

¦ ¦ +-Address

¦ ¦ +-Electronic mail address..... hollistr@gvsu.edu

¦ +-Resource format

¦ ¦ +-Format specification citation

¦ ¦ +-Alternate title..... Other ASCII

¦ +-Descriptive keywords (1 of 5)

¦ ¦ +-Keyword..... EARTH SCIENCE > BIOSPHERE > TERRESTRIAL ECOSYSTEMS > ALPINE/TUNDRA

¦ ¦ +-Type..... Theme

¦ ¦ +-Thesaurus name

¦ ¦ +-Title..... NASA/GCMD Earth Science Keywords

¦ ¦ +-Alternate title..... Science and Services Keywords

¦ ¦ +-Date

¦ ¦ +-Date..... May 21, 2014 12:00:00 AM

¦ ¦ +-Date type..... Revision

¦ +-Descriptive keywords (2 of 5)

¦ ¦ +-Keyword..... FIELD SURVEY

¦ ¦ +-Type..... Theme

¦ ¦ +-Thesaurus name

¦ ¦ +-Title..... ACADIS Keywords

¦ ¦ +-Alternate title..... Platforms

¦ ¦ +-Date

¦ ¦ +-Date..... Oct 7, 2014 12:00:00 AM

¦ ¦ +-Date type..... Revision

Corresponding Customized Meta Data is as below

CharacterSet-->UTF-8

ContactRole-->RESOURCE_PROVIDER

ContactPartyName-->UCAR/NCAR - CISL - ACADIS

IdentificationInfoCitationTitle-->Barrow Atqasuk ARCSS Plant

CitationDateCREATION-->Mon Dec 16 00:00:00 PST 2013

CitationDatemodified-->Thu Feb 05 00:00:00 PST 2015

[CitedResponsiblePartyRole]-->Role[POINT_OF_CONTACT]

CitedResponsiblePartyName-->Robert Hollister

CitedResponsiblePartyOrganizationName-->null

CitedResponsiblePartyPositionName-->null

CitedResponsiblePartyEMail-->hollistr@gvsu.edu

IdentificationInfoAbstract-->These files contain data representing the periodic plant measures of species within each plot in a text tab delimited format.The data presented are seasonal growth of graminoids (length of leaf and length of inflorescence) and seasonal flowering of all species (number of inflorescences in flower within a plot), collected weekly during the summers of 2012-20XX for a subset of 30 grid plots at two sites (Barrow ARCSS grid and Atqasuk ARCSS grid).

IdentificationInfoStatus-->ON_GOING

ResourceFormatSpecificationAlternativeTitle-->Other ASCII

IdentificationInfoLanguage-->English

IdentificationInfoTopicCategory-->BIOTA

DescriptiveKeyWords 1

==========

Keywords-->EARTH SCIENCE > BIOSPHERE > TERRESTRIAL ECOSYSTEMS > ALPINE/TUNDRA

KeywordsType-->THEME

ThesaurusNameTitle-->NASA/GCMD Earth Science Keywords

[ThesaurusNameAlternativeTitle]-->[Science and Services Keywords]

ThesaurusNameDateREVISION-->Wed May 21 00:00:00 PDT 2014

DescriptiveKeyWords 2

==========

Keywords-->FIELD SURVEY

KeywordsType-->THEME

ThesaurusNameTitle-->ACADIS Keywords

[ThesaurusNameAlternativeTitle]-->[Platforms]

ThesaurusNameDateREVISION-->Tue Oct 07 00:00:00 PDT 2014

I definitely feel that the Key Names could be much shorter,your suggestion would be appreciated .

Once the format would be finalized I can go ahead and start integrating the same into Apache Tika to handle .iso19139 files to make the Tika much more robust. Feel free to Comment