About keyword parsing

Let’s look at one record: NODC\ [000294EA-2BC7-44A6-8C72-DBD4D14BFF2D.xml](http://hydro10.sdsc.edu/metadata/NODC/000294EA-2BC7-44A6-8C72-DBD4D14BFF2D.xml) (2nd in the NODC folder)

It has several groups of keywords, organized by different thesauri, and described using Anchor (so that to allow xlink:href pointing to a definition.

Some of these thesauri we would like to have in all records.

These are

“Parameter thesaurus” (in NODC called NODC DATA TYPES THESAURUS)

<gmd:descriptiveKeywords>

<gmd:MD\_Keywords>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/10"xlink:actuate="onRequest">AMMONIA</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/223"xlink:actuate="onRequest">NITRATE</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/237"xlink:actuate="onRequest">OXYGEN</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/457"xlink:actuate="onRequest">PHAEOPHYTIN</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/263"xlink:actuate="onRequest">PHOSPHATE</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/454"xlink:actuate="onRequest">PIGMENTS</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/294" xlink:actuate="onRequest">PRIMARY PRODUCTIVITY</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/440"xlink:actuate="onRequest">ZOOPLANKTON</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/datatype/details/441" xlink:actuate="onRequest">ZOOPLANKTON BIOMASS</gmx:Anchor>

</gmd:keyword>

<gmd:type>

<gmd:MD\_KeywordTypeCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD\_KeywordTypeCode"codeListValue="theme" codeSpace="005">theme</gmd:MD\_KeywordTypeCode>

</gmd:type>

<gmd:thesaurusName>

<gmd:CI\_Citation>

<gmd:title>

<gco:CharacterString>NODC DATA TYPES THESAURUS</gco:CharacterString>

</gmd:title>

<gmd:date gco:nilReason="inapplicable"/>

</gmd:CI\_Citation>

</gmd:thesaurusName>

</gmd:MD\_Keywords>

</gmd:descriptiveKeywords>

Also, “Instrument type” (>NODC INSTRUMENT TYPES THESAURUS)

<gmd:descriptiveKeywords>

<gmd:MD\_Keywords>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/insttype/details/3"xlink:actuate="onRequest">bottle</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/insttype/details/13"xlink:actuate="onRequest">fluorometer</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/insttype/details/31" xlink:actuate="onRequest">net - plankton net</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/insttype/details/68" xlink:actuate="onRequest">net - zooplankton net</gmx:Anchor>

</gmd:keyword>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.nodc.noaa.gov/cgi-bin/OAS/prd/insttype/details/27" xlink:actuate="onRequest">oxygen meter</gmx:Anchor>

</gmd:keyword>

<gmd:type>

<gmd:MD\_KeywordTypeCodecodeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schema/resources/Codelist/gmxCodelists.xml#MD\_KeywordTypeCode"codeListValue="instrument">instrument</gmd:MD\_KeywordTypeCode>

</gmd:type>

<gmd:thesaurusName>

<gmd:CI\_Citation>

<gmd:title>

<gco:CharacterString>NODC INSTRUMENT TYPES THESAURUS</gco:CharacterString>

</gmd:title>

<gmd:date gco:nilReason="inapplicable"/>

</gmd:CI\_Citation>

</gmd:thesaurusName>

</gmd:MD\_Keywords>

</gmd:descriptiveKeywords>

Additional thesauri are: NODC PLATFORM NAMES THESAURUS, NODC COLLECTING INSTITUTION NAMES THESAURUS, NODC SUBMITTING INSTITUTION NAMES THESAURUS, NODC SEA AREA NAMES THESAURUS.

If we follow this model, we need to have cross-domain thesauri, refer to them from keywords we generate, and create respective gmd:descriptiveKeywords sections.

The issues are:

1. Creating cross-domain thesauri from those vocabularies we find?
   1. Use existing cross-domain vocabularies, eg SWEET
   2. Take a few generic vocabularies, for main parameters (chemical, biological, physical…), instruments, mediums, processing levels, supported protocols, information models, authors, organizations. In other words, each dataset will be characterized by the following facets:
      1. Parameters measured (extracted from the title and record)
      2. Instrument and/or platform (extracted from the title and record)
      3. Processing level (by default, not specified)
      4. Medium (stream, sediment, rock… should use some heuristics)
      5. Authors (from the record) give after initial search
      6. Organization (from the record)
      7. Supported access protocols (from the record)
      8. Supported information model (from the record, and heuristics. Instead, could be file type)
      9. What vocabulary or vocabularies conforms with
      10. In what repository resides
      11. What catalog the metadata record is harvested from
      12. Domain (as given in <http://abstractsearch.agu.org/keywords>)
          1. As rdf: [http://abstractsearch.agu.org/keywords.rdf](http://abstractsearch.agu.org/keywords/1800.rdf)
      13. More detailed index terms as in <http://abstractsearch.agu.org/keywords/1800.html>
          1. As rdf: <http://abstractsearch.agu.org/keywords/1800.rdf>
          2. As json: <http://abstractsearch.agu.org/keywords/1800.json>

See High level term numbers below

* + 1. Spatial extent (from the record, or
    2. Links to other resources

The above is focused on data; software may be different.

1. For each harvesting endpoint, create mappings that define how the above information is extracted and mapped to ISO
2. Based on the mapping file, each parser would then create a respective keyword record and insert it in the file, e.g.

<gmd:descriptiveKeywords>

<gmd:MD\_Keywords>

<gmd:keyword xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:xlink="http://www.w3.org/1999/xlink">

<gmx:Anchor xlink:href="http://www.cinergi.earthcube.org/thesauri/domain/1>4800 OCEANOGRAPHY: BIOLOGICAL AND CHEMICAL</gmx:Anchor>

</gmd:keyword>

<gmd:type>

<gmd:MD\_KeywordTypeCodecodeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schema/resources/Codelist/gmxCodelists.xml#MD\_KeywordTypeCode"codeListValue="instrument">instrument</gmd:MD\_KeywordTypeCode>

#

# Basically, a pointer to a vocabulary maintained at CINERGI or elsewhere.

#

</gmd:type>

<gmd:thesaurusName>

<gmd:CI\_Citation>

<gmd:title>

<gco:CharacterString>CINERGI DOMAIN VOCABULARY BASED ON AGU LIST at http://abstractsearch.agu.org/keywords</gco:CharacterString>

</gmd:title>

<gmd:date gco:nilReason="inapplicable"/>

</gmd:CI\_Citation>

</gmd:thesaurusName>

</gmd:MD\_Keywords>

</gmd:descriptiveKeywords>

High level term numbers

1. [[0300] ATMOSPHERIC COMPOSITION AND STRUCTURE](http://abstractsearch.agu.org/keywords/0300)
2. [[0400] BIOGEOSCIENCES](http://abstractsearch.agu.org/keywords/0400)
3. [[0500] COMPUTATIONAL GEOPHYSICS](http://abstractsearch.agu.org/keywords/0500)
4. [[0600] ELECTROMAGNETICS](http://abstractsearch.agu.org/keywords/0600)
5. [[0700] CRYOSPHERE](http://abstractsearch.agu.org/keywords/0700)
6. [[0800] EDUCATION](http://abstractsearch.agu.org/keywords/0800)
7. [[0900] EXPLORATION GEOPHYSICS](http://abstractsearch.agu.org/keywords/0900)
8. [[1000] GEOCHEMISTRY](http://abstractsearch.agu.org/keywords/1000)
9. [[1100] GEOCHRONOLOGY](http://abstractsearch.agu.org/keywords/1100)
10. [[1200] GEODESY AND GRAVITY](http://abstractsearch.agu.org/keywords/1200)
11. [[1500] GEOMAGNETISM AND PALEOMAGNETISM](http://abstractsearch.agu.org/keywords/1500)
12. [[1600] GLOBAL CHANGE](http://abstractsearch.agu.org/keywords/1600)
13. [[1700] HISTORY OF GEOPHYSICS](http://abstractsearch.agu.org/keywords/1700)
14. [[1800] HYDROLOGY](http://abstractsearch.agu.org/keywords/1800)
15. [[1900] INFORMATICS](http://abstractsearch.agu.org/keywords/1900)
16. [[2100] INTERPLANETARY PHYSICS](http://abstractsearch.agu.org/keywords/2100)
17. [[2400] IONOSPHERE](http://abstractsearch.agu.org/keywords/2400)
18. [[2700] MAGNETOSPHERIC PHYSICS](http://abstractsearch.agu.org/keywords/2700)
19. [[3000] MARINE GEOLOGY AND GEOPHYSICS](http://abstractsearch.agu.org/keywords/3000)
20. [[3200] MATHEMATICAL GEOPHYSICS](http://abstractsearch.agu.org/keywords/3200)
21. [[3300] ATMOSPHERIC PROCESSES](http://abstractsearch.agu.org/keywords/3300)
22. [[3600] MINERALOGY AND PETROLOGY](http://abstractsearch.agu.org/keywords/3600)
23. [[3900] MINERAL PHYSICS](http://abstractsearch.agu.org/keywords/3900)
24. [[4200] OCEANOGRAPHY: GENERAL](http://abstractsearch.agu.org/keywords/4200)
25. [[4300] NATURAL HAZARDS](http://abstractsearch.agu.org/keywords/4300)
26. [[4400] NONLINEAR GEOPHYSICS](http://abstractsearch.agu.org/keywords/4400)
27. [[4500] OCEANOGRAPHY: PHYSICAL](http://abstractsearch.agu.org/keywords/4500)
28. [[4800] OCEANOGRAPHY: BIOLOGICAL AND CHEMICAL](http://abstractsearch.agu.org/keywords/4800)
29. [[4900] PALEOCEANOGRAPHY](http://abstractsearch.agu.org/keywords/4900)
30. [[5100] PHYSICAL PROPERTIES OF ROCKS](http://abstractsearch.agu.org/keywords/5100)
31. [[5200] PLANETARY SCIENCES: ASTROBIOLOGY](http://abstractsearch.agu.org/keywords/5200)
32. [[5400] PLANETARY SCIENCES: SOLID SURFACE PLANETS](http://abstractsearch.agu.org/keywords/5400)
33. [[5700] PLANETARY SCIENCES: FLUID PLANETS](http://abstractsearch.agu.org/keywords/5700)
34. [[6000] PLANETARY SCIENCES: COMETS AND SMALL BODIES](http://abstractsearch.agu.org/keywords/6000)
35. [[6200] PLANETARY SCIENCES: SOLAR SYSTEM OBJECTS](http://abstractsearch.agu.org/keywords/6200)
36. [[6300] POLICY SCIENCES](http://abstractsearch.agu.org/keywords/6300)
37. [[6600] PUBLIC ISSUES](http://abstractsearch.agu.org/keywords/6600)
38. [[6900] RADIO SCIENCE](http://abstractsearch.agu.org/keywords/6900)
39. [[7200] SEISMOLOGY](http://abstractsearch.agu.org/keywords/7200)
40. [[7500] SOLAR PHYSICS, ASTROPHYSICS, AND ASTRONOMY](http://abstractsearch.agu.org/keywords/7500)
41. [[7800] SPACE PLASMA PHYSICS](http://abstractsearch.agu.org/keywords/7800)
42. [[7900] SPACE WEATHER](http://abstractsearch.agu.org/keywords/7900)
43. [[8000] STRUCTURAL GEOLOGY](http://abstractsearch.agu.org/keywords/8000)
44. [[8100] TECTONOPHYSICS](http://abstractsearch.agu.org/keywords/8100)
45. [[8400] VOLCANOLOGY](http://abstractsearch.agu.org/keywords/8400)
46. [[9300] GEOGRAPHIC LOCATION](http://abstractsearch.agu.org/keywords/9300)
47. [[9600] INFORMATION RELATED TO GEOLOGIC TIME](http://abstractsearch.agu.org/keywords/9600)
48. [[9800] GENERAL OR MISCELLANEOUS](http://abstractsearch.agu.org/keywords/9800)
49. [[9900] CORRECTIONS](http://abstractsearch.agu.org/keywords/9900)