ScientificDataFormatTextRepresentation

Scientific Data Text Representation

Tika now parses scientific data files such as HDF, NetCDF, and MatLab. Currently, the parsed output for these files is metadata. While there is a 'text/ -t' output for NetCDF files, it includes variable names and dimensions, which is still in its essence, metadata. Enabling the *data* themselves to be parsed would enable a new paradigm in search.

NetCDF Example

A netCDF classic or 64-bit offset dataset is stored as a single file comprising two parts:

- a header, containing all the information about dimensions, attributes, and variables except for the variable data;
- a data part, comprising fixed-size data, containing the data for variables that don't have an unlimited dimension; and variable-size data, containing the data for variables that have an unlimited dimension.

The NetCDFparser currently extracts the "header part".

- -text extracts file Dimensions and Variables
- -metadata extracts Global Attributes

We want the option to extract the "data part" of NetCDF files.

JIRA Issue: TIKA-1577

Lets use the NetCDF test file for our dev testing: tika/tika-parsers/src/test/resources/test-documents/sresa1b_ncar_ccsm3_0_run1_200001.nc

Suggestions include:

Use nested tables in HTML5:

http://stackoverflow.com/questions/10297874/html5-validity-of-nested-tables