A Survey on Managing Linked Scientific Datasets

CS5352 Course Project, Spring 2021

[This project accepts one student only.]

**Contact**

Chenxu Niu (chenxu.niu@ttu.edu)

**Description**

Managing, storing, searching, and analyzing large volumes of scientific datasets are complex and difficult problems that require innovative and highly efficient data management technology. Many scientific data formats are proposed to store these data, such as HDF5[1], NetCDF[2] and FITS[4]. As the number of scientific datasets and repositories published on the web is increasing, the necessity for finding and managing relevant scientific datasets becomes crucial for scientists. There are several existing data linkage solutions, such as NetCDF LD [5], Silk[6] and LINDA[7].

The goal of this research is providing a survey of different data linkage approaches for scientific datasets, finding the advantages and disadvantages of these approaches.

1. Literature study of the related papers about managing linked scientific datasets.
2. Compare these approaches and provide pros and cons.
3. Implement one of these approaches (bonus).

**Requirements**

* Knowledge of scientific data format
* Proficiency with C programming language
* Knowledge of Linux

**Students will have a chance to learn about:**

* Scientific Data Management
* SPARQL Query Language
* Data Linkage

**References**

1. HDF5 format: <https://support.hdfgroup.org/HDF5/>
2. NetCDF format: <https://www.unidata.ucar.edu/software/netcdf/docs/>
3. NetCDF: An Introduction to NetCDF: <https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html>
4. FITS format: <https://fits.gsfc.nasa.gov/fits_documentation.html>
5. NetCDF LD: <http://docs.opengeospatial.org/DRAFTS/19-002.html>
6. Silk: <http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf>
7. LIMES: <https://dl.acm.org/doi/abs/10.1145/2396761.2398582>