Project Proposal

Name: Mengying Wang
Instructor: Prof. Chris Mattmann
Project name: Polar CyberInfrastructure: An open source framework for metadata exploration and discovery of Polar Data
Period: 2014-08-25 to 2014-12-05

Project background:
The Arctic region is one of the Earth's last great frontiers [1]. The region is home to strong national security interests (e.g., land/air/sea and maritime), commercial interests (e.g., in oil, and in minerals and natural resources), and is heavily influenced by the Earth's changing climate including temperature increases which have led to urgent problems including glacial melting [2]. Several strategic priorities for the region are identified in the U.S. President's National Strategy for the Arctic Region document [1], including *Protect the Arctic Environment and Conserve Natural Resources*, etc.

Although myriad data sources relevant to the Arctic region are already available to increase understanding, the challenge remains sifting and sorting through those data (we refer to this process as "data triage") to properly respond to the regional priorities laid out in the U.S. national strategy. During this "data triage" process, there are a variety of data formats that a researcher must parse, analyze and interpret, not to mention the variety and quality of associated metadata available that a researcher must make sense of. For example, in the NASA's Antarctic Master Directory (AMD) [3] system, with the same Polar/Arctic related scientific parameter, of the 50 data results, 52% have an Unknown data format; similarly 6% of them are listed as TBD; the majority of the known data types (16%) are ASCII (text) format; the remaining are PDF files (11%), Excel files (6%), scientific data formats (NetCDF [13] – 3%), movie files (MP4 – 2%), Microsoft Word files (2%) and HTML files (2%).

Project description:
This project will construct and deliver a classification framework based on the open source Apache Tika technology [4] that obviates the hetereogeneity of Polar data formats, and that provides a unified mechanism for analyzing and interpreting the metadata and data once classified. Specifically, the research will comprise three steps, enumerated below.

1. **Classify and understand Polar and Arctic data and metadata.** We will deploy Apache Tika at scale using Apache OODT [5, 6], Apache Nutch [7] and Apache Solr [8]. Apache OODT is a data management and processing framework originally

developed at NASA; Nutch is a distributed search engine that runs on top of Apache Hadoop, and that indexes content metadata (including type, and language information generated by Tika). Both OODT and Nutch can index metadata in the Solr indexing engine. Once MIME types are indexed in Solr, we can easily perform MIME type classification and develop data format distributions and profiles for different representative Polar data sources, e.g., NASA AMD, PolarGrid [9], Acadis [10], etc. The metadata distribution will characterize the number and heterogeneity of Polar data formats from these representative data sources, and inform the next step.

2. **Augment Apache Tika to automatically detect the most prevalent Polar and Arctic data and metadata formats.** We will expand Tika to handle more relevant scientific data formats including OPeNDAP [11], ISO-19115 metadata [12] (through the Apache SIS project) and also FGDC. Augmenting Tika's MIME detection facilities to handle these formats will involve addition of glob patterns (file name extensions), regular expressions, MIME "magic" or digital signatures for files, and curation of XML root namespaces and patterns, especially in the case of XML-based formats.

3. **Augment Apache Tika to automatically parse the most prevalent Polar and Arctic data and metadata formats.** Parsing involves both creation of new Tika parsing facilities and the expansion of existing parsers. For example while there is existing support for parsing XML documents in Tika that can be extended for ISO-19115 and FGDC, which are XML based formats, OPeNDAP support will involve the integration and creation of new parsing facilities since OPeNDAP formats vary from DAP metadata (XML-based) to DODS (binary level) data. In addition, many sources of Polar and Arctic in-situ data are text-based, and extensions of the Text parser and the development of special Numeric oriented parsers will be necessary. In addition, we expect to construct new ContentHandler implementations that extract content and augment the extracted metadata in a streaming fashion.

My job:

My main task is first step - the classification and triage of one of the existing Polar and Arctic data sources: Acadis. As a demonstration of what I should do in this project, consider the four data format value distributions shown in Figure 2 as generated from another data source NASA AMD and the Cryosphere > Sea ice > Ice Temperature parameter page. Starting in the upper left, we see that original ("Raw") formats left uncleansed comprise a nearly even and otherwise difficult to decipher and repetitive and noisy distribution. When the formats are trimmed and lowercased ("TL") uniformly as shown in the upper right portion of Figure 2 the true distribution is less noisy; when we look for and group words similar to ASCII (e.g., "ASCII text", "ASCII text files", etc.)

we arrive at the distribution in the bottom left which generates the actual count of ASCII types in the format distribution; and finally with similar approaches applied to Microsoft Word and Excel files, we arrive at the final overall distribution shown in the bottom right comprising the actual Data format distribution of types for sea ice temperature.

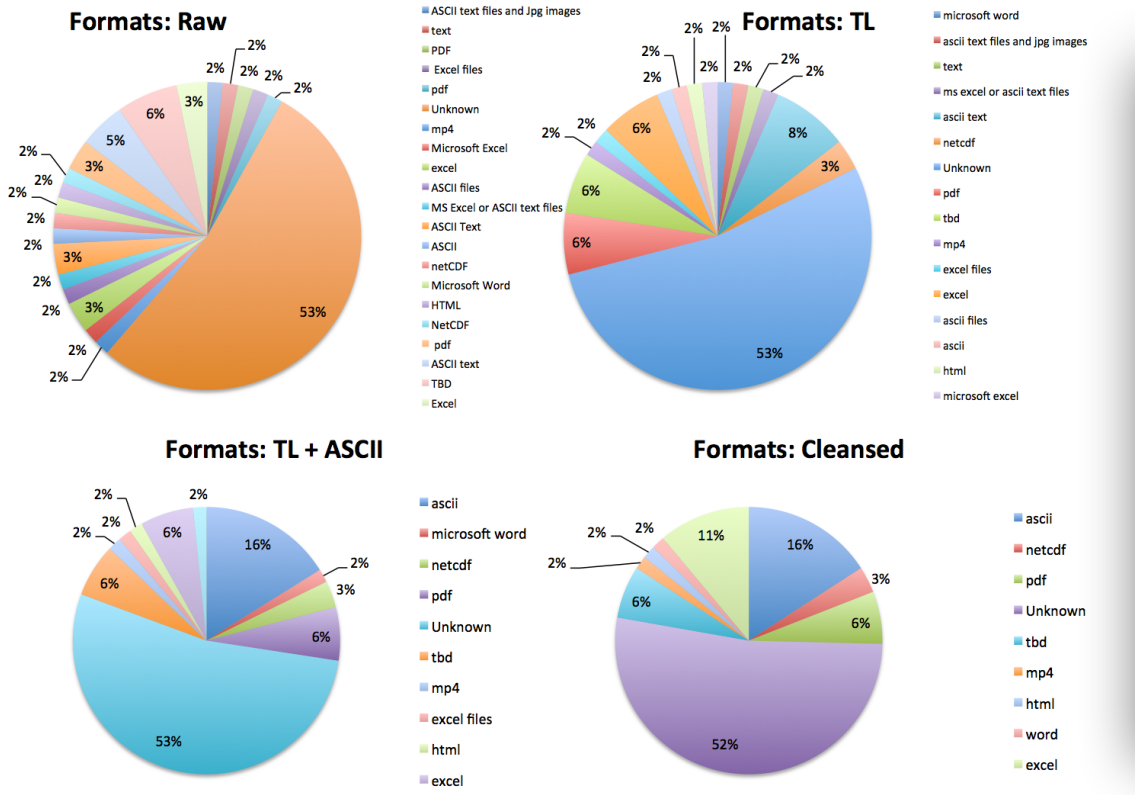

Figure 2. Distribution of Formats for Snow/Ice Temperature data from NASA's Antarctic Master Directory.

What I am going to do is something similar, but on a different data source: Acadis.

My approach:

The approach is illustrated graphically in Figure 5. Reading the diagram from the upper left, I will first acquire Acadis Cryosphere dataset descriptions via crawls of the Cryosphere dataset pages leveraging the Apache OODT Crawler technology. The Apache OODT Crawler only deals with locally available content so I will leverage Apache Nutch or the Apache OODT Push Pull framework to acquire the remote Acadis dataset descriptions in Step 1 of Figure 5. Nutch and PushPull have different strengths and weaknesses, for example Nutch automatically stores the content to the Hadoop filesystem, making it more scalable and efficient but otherwise requiring specialized filesystem APIs to retrieve; whereas PushPull stores the content to local storage as native files in a less efficient and scalable, but more easily accessible fashion. Once acquired, dataset descriptions will be ingested using the Apache OODT crawler into the Apache OODT

file management system, where associated data description files and metadata will be initially stored.

Once the dataset descriptions are stored in the File Manager, I will independently develop Tika based cleansing and triage approaches to classification based on Dataset Format as a start, and expanding to other available metadata including free-text descriptions; space and time information, and Polar and Arctic features. The Tika cleansing algorithms will then be run on the dataset descriptions at scale using the Apache OODT Workflow Manager technology shown in Step 2 of Figure 5.

In Step 3 of Figure 5, I will develop a simple ingest algorithm to ingest the Tika cleansed Dataset descriptions into Apache Solr, making the associated Polar and Arctic dataset formats, and other associated metadata available for analytics including field-based faceting, value space distribution, and visualization similar to the analysis shown in Figure 2.

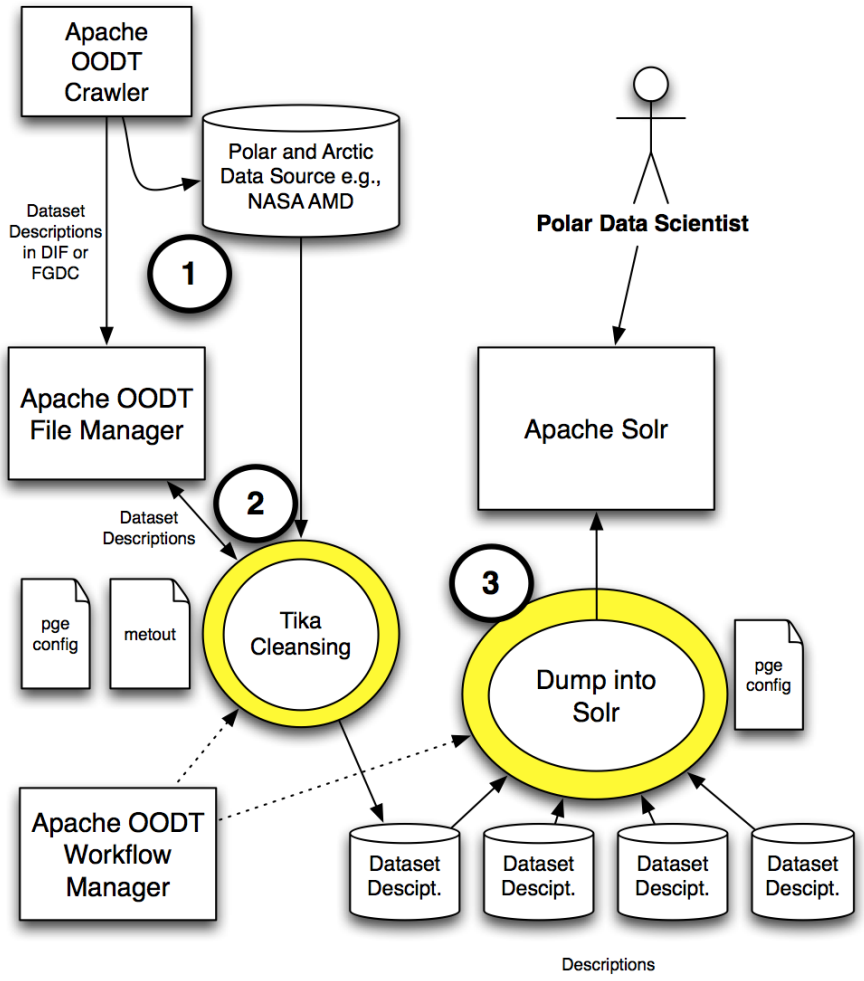

Figure 5. Overall triage and classification approach for Polar and Arctic data.

My plan:

Week 1	1. Read project proposal, understand project content and assignments. 2. Read the book *Tika in Action*, figure out what is Tika, and what Tika can do.
Week 2	1. Install Apache OODT on Mac or Windows, play with Apache OODT Crawler, Apache OODT Push Pull, Apache OODT File Manager, and Apache OODT Workflow Manager. 2. Install Apache Nutch on Mac or Windows, play with Apache Nutch. 3. Install Apache Solr on Mac or Windows, play with Apach Solr. 4. Install Apache Tika on Mac or Windows, and play with Apache Tika.
Week 3	1. Solve any possible installation problems. 2. Continue to learn Apache OODT tutorials. 3. Continue to learn Apache Nutch tutorials. 4. Continue to learn Apache Solr tutorials. 5. Continue to learn Apache Tika tutorials.
Week 4	1. Acquire the remote Acadis dataset descriptions using Apache Nutch or the Apache OODT.
Week 5	1. Acquire Acadis Cryosphere dataset descriptions in either GCMD DIF or FGDC format via crawls of the Cryosphere dataset pages leveraging

	the Apache OODT Crawler.
Week 6	1. Develop Tika based cleansing and triage approaches to Acadis classification based on Dataset Format firstly.
Week 7	1. Expand Tika based cleansing and triage approaches to Acadis classification to other available metadata including free-text descriptions; space and time information, and Polar and Arctic features.
Week 8	1. Run the Tika cleansing algorithms on the dataset descriptions at scale using the Apache OODT Workflow Manager technology.
Week 9	1. Develop a simple ingest algorithm to ingest the Tika cleansed Dataset descriptions into Apache Solr for future analytics including field-based faceting, value space distribution, and visualization.
Week 10	TBD
Week 11	TBD
Week 12	TBD
Week 13	TBD
Week 14	TBD
Week 15	TBD

References

1. U.S. National Strategy for the Arctic Region, http://www.whitehouse.gov/sites/default/files/docs/nat_arctic_strategy.pdf, May 2013

2. Overpeck, Jonathan, et al. "Arctic environmental change of the last four centuries." Science 278.5341 (1997): 1251-1256.

3. Antarctic Master Directory – NASA, http://gcmd.nasa.gov/portals/amd/, Accessed July 2013.

4. Mattmann, C., and J. Zitting. Tika in Action. Manning Publications Co., 2011.

5. Mattmann, C. A., Crichton, D. J., Medvidovic, N., and J. S. Hughes. "A software architecture-based framework for highly distributed and data intensive scientific applications." In Proceedings of the 28th international conference on Software engineering, pp. 721-730. ACM, 2006.

6. Mattmann, C. A., et al. "A reusable process control system framework for the orbiting carbon observatory and NPP Sounder PEATE missions." Space Mission Challenges for Information Technology, 2009. SMC-IT 2009. Third IEEE International Conference on. IEEE, 2009.

7. Cafarella, M. and D. Cutting. Building. "Open source search." Queue. v2 i2 (2004): 54-61.

8. Apache Solr, http://lucene.apache.org/solr/, Access July 2013.

9. Guo, Z., Raminderjeet S., and M. Pierce. "Building the polargrid portal using web 2.0 and opensocial." Proceedings of the 5th Grid Computing Environments Workshop. ACM, 2009.

10. Jodha Khalsa, S., et al. "The Advanced Cooperative Arctic Data and Information Service (ACADIS)." EGU General Assembly Conference Abstracts. Vol. 15. 2013.

11. P. Cornillon, et al., "OPeNDAP: Accessing data in a distributed, heterogeneous environment," Data Science Journal, vol. 2, pp. 164-174, 2003.

12. Karschnick, O., et al. "The UDK and ISO 19115 Standard." Proceedings of the 17th International Conference Informatics for Environmental Protection EnviroInfo. 2003.

13. R.K. Rew and G.P. Davis, "NetCDF: An Interface for Scientific Data Access," IEEE Computer Graphics and Applications, vol. 10, no. 4, 1990, pp. 76–82.