TREC Dynamic Domain: Polar Science

Annie Bryant Burgess[1,2,3], Chris Mattmann[1,2], Giuseppe Totaro[2]
Lewis John McGibbney[2,] Paul Ramirez[2]

[1]University of Southern California
[2]NASA Jet Propulsion Lab
[3]Earth Science Information Partners
[*]chris.a.mattmann@jpl.nasa.gov

Abstract

This paper outlines the creation of the Polar dataset within the TREC-Dynamic Domain track. The techniques used to create the Polar dataset fall into two basic categories: information extraction using Apache Tika and information retrieval using Apache Nutch. Frist, we expanded the parsing capabilities of Apache Tika, an open source framework for text and metadata extraction, to provide more searchable content within Polar data repositories. Second, we used Apache Nutch, a distributed search engine that runs on top of Apache Hadoop, to crawl three prominent Polar data repositories: the National Science Foundation Advanced Cooperative Arctic Data and Information System (ACADIS), the National Snow and Ice Data Center (NSIDC) Arctic Data Explorer (ADE), and the National Aeronautics and Space Administration Antarctic Master Directory (AMD). Because finding data is often a primary challenge in scientific discovery, the inclusion of the Polar dataset in TREC-DD helps advance science through data discovery and provides TREC-DD a new challenge in in the realm of search relevancy.

1 Introduction

1.1 Motivation

Climate change is amplified in the Polar Regions. Polar amplification is captured via space and airborne remote sensing, in-situ measurement, and climate modeling. Beyond the rich literature that documents changing Polar regions, each method of Polar-data collection produces a diverse set of data types, ranging from text-based metadata to more complex data structures (e.g. HDF, NetCDF, GRIB). There are many of these diverse data spread throughout science-specific repositories of data – we will focus on three in this paper.

The three repositories of interest to our work are (shown graphically in Figure 1 and outlined in detail in [1]): the National Science Foundation Advanced Cooperative Arctic Data and Information System (ACADIS, upper left), the National Aeronautics and Space Administration Antarctic Master Directory (AMD, upper right) and the National Snow and Ice Data Center (NSIDC) Arctic Data Explorer (ADE, bottom). These data sets represent a rich combination of web documents, PDF documents, scientific data files (HDF, NetCDF, Gridded Binary, etc.) collected over *space* (world; regional; Arctic;

Antarctic) and *time* (years-decades), with rich metadata features, and content. Access and search to this data to date has been limited to (as a baseline), scientific search of the metadata via forms; text search on the descriptions of the data; and in some (advanced) cases, geospatial search by space and time. The data also represents 10s of thousands to 100s of thousands of records, 10s-100s of Gigabytes of information, and across the three systems, which themselves are largely uncoordinated, there represents a very large chance of duplication of all types of content present in the datasets.

Figure 1: NASA AMD, NSF ACADIS and NSIDC Arctic Data Explorer.

A rough estimate is that each repository contains on order of 20,000 – 200,000 URLs; is between 10-100Gb on disk; and contains 1000-10,000 scientific records and datasets. As an idea of the richness of the types of data found in these repositories, see the pie chart in Figure 2 that compares the resultant MIME types from searching for mass balance in NASA's Antarctic Master Directory – just one of the repositories from our list.

2 Data Triage of Polar Repositories

In the Spring of 2015, we engaged in a series of crawling activities led by 200+ students in our CSCI 572 Search Engines class [2] at the University of Southern California in order to prepare a rich dataset for contribution to the TREC Dynamic Domain track. The goal of the crawling activities were to acquire as much data from ACADIS, ADE and AMD as possible, considering that the data itself was largely indicative of "dark data" and of the "dynamic domain". The data is behind web forms (typically data ordering "cart" web sites); Ajax/javascript (results pagination), and also behind heterogeneous content types. Our definition of crawl success was largely grounded in both web completeness as well as via richness of the parsed and retrieved records. Our team used Apache Nutch (http://nutch.apache.org/) and in turn the Apache Tika

(http://tika.apache.org/) frameworks for web crawling and content analysis. Our team includes the progenitors of both of these technologies.

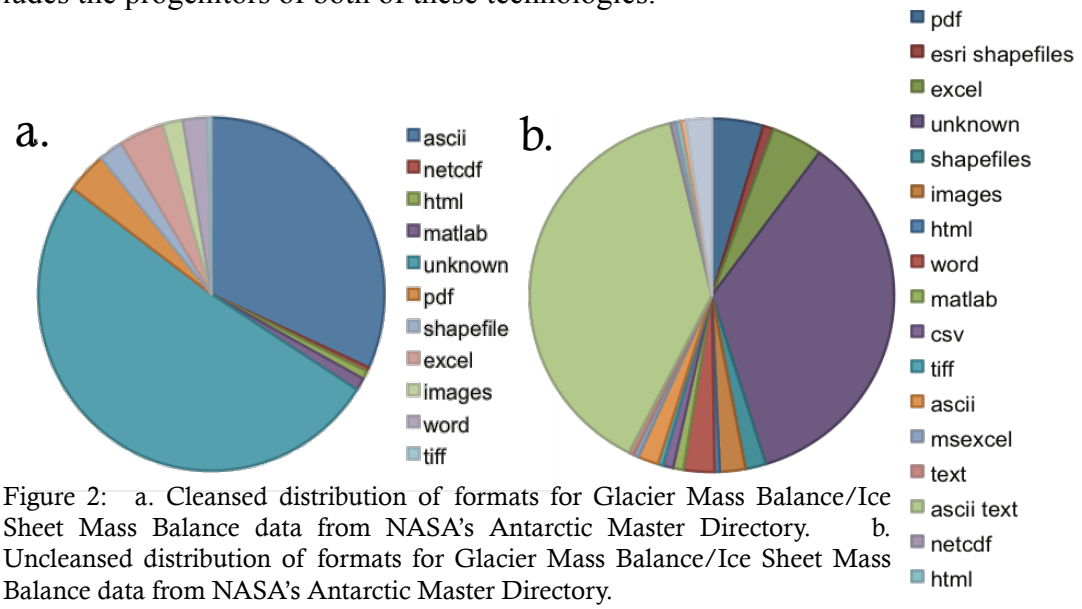


Figure 2: a. Cleansed distribution of formats for Glacier Mass Balance/Ice Sheet Mass Balance data from NASA's Antarctic Master Directory. b. Uncleansed distribution of formats for Glacier Mass Balance/Ice Sheet Mass Balance data from NASA's Antarctic Master Directory.

Our ultimate objective was to find and identify and acquire as much of the content present on these repositories as possible, while also adhering to being good crawl citizens, e.g., politeness, parallel crawling whilst traversing the expanse of the web graph and the portion of the domain and graph relevant to Polar and the Arctic sciences. Teams spent a great of effort ensuring that the content on these sites crawled was meaningful and not simply Javascript and/or CSS files that one can encounter from these archive and repository sites.

While crawling and performing the above objectives, our teams identified a range of web crawling issues and overcame them. These ranged from having to negotiate different protocols to get to the actual rich content (the HDF files, the NetCDF files, the Grib files, the Matlab files, etc.) – some are behind FTP, HTTP, HTTPS, etc; to navigation requiring web forms and having to POST to those forms to navigate to the actual content. In addition, our teams expanded Nutch to execute Javascript by creating a plugin wrapper around the Selenium (http://www.seleniumhq.org/) framework.

Our team crawled and extracted text, metadata, and language from the content and performed deduplication using an *exact* and *near duplicates* approach.

3 Web Crawl

Web crawls were focused on three polar data repositories: the National Science Foundation Advanced Cooperative Arctic Data and Information System (ACADIS), the National Snow and Ice Data Center (NSIDC) Arctic Data Explorer (ADE), and the National Aeronautics and Space Administration Antarctic Master Directory (AMD).

Each web crawl used Apache Nutch as the core framework for web crawling and Apache Tika as the main content detection and extraction framework. Nutch is a distributed search engine that runs on top of Apache Hadoop.

4 Dataset Preparation

4.1 Duplicate Records

Exact duplicate records were removed using signature based methods. Algorithms and accompanying code were developed to remove near duplicates, using jaccard similarity. However, not all teams that submitted web crawls to this dataset applied their jaccard-similarity algorithms.

4.2 Data Format

Crawled data were put into Common Crawl Format, according to TREC Dynamic Domain format, using the CommonCrawlDataDumper. The CommonCrawlDataDumper is an Apache Nutch tool that can dump Nutch segments into Common Crawl data format, mapping each crawled-by-Nutch file on a JSON-based data structure. CommonCrawlDataDumper dumps out the files and serialize them with Compact Binary Object Representation (CBOR) encoding, a data representation format used in many contexts.

Each contributed web crawl has an accompanying JSON file that lists the total records, by mimeType. A program aggregates all of the JSON files. The results of the aggregation are shown in [3] and [4] and graphically for the ACADIS website in Figure 3.

5 Dataset Characteristics

The TREC Dynamic Domain Polar dataset described in this paper is a collection of web crawls from three primary sources:

1. Dr. Chris Mattmann's crawl of ADE, performed at the Open Science Codefest and at the NSF DataViz Hackathon for Polar CyberInfrastructure
2. Dr. Mattmann's student Angela Wang, contributed 3 datasets: 2 crawls of ACADIS and one of NASA AMD.
3. Dr. Mattmann's CSCI 572 Course at USC, students submitted 13 individual crawls of NASA ACADIS, NSIDC ADE, and AMD.

The finished Polar dataset is composed of 17 distinct web crawls, containing 1,741,530 records (158 GB) across the three Polar science data repositories, which themselves are largely uncoordinated.

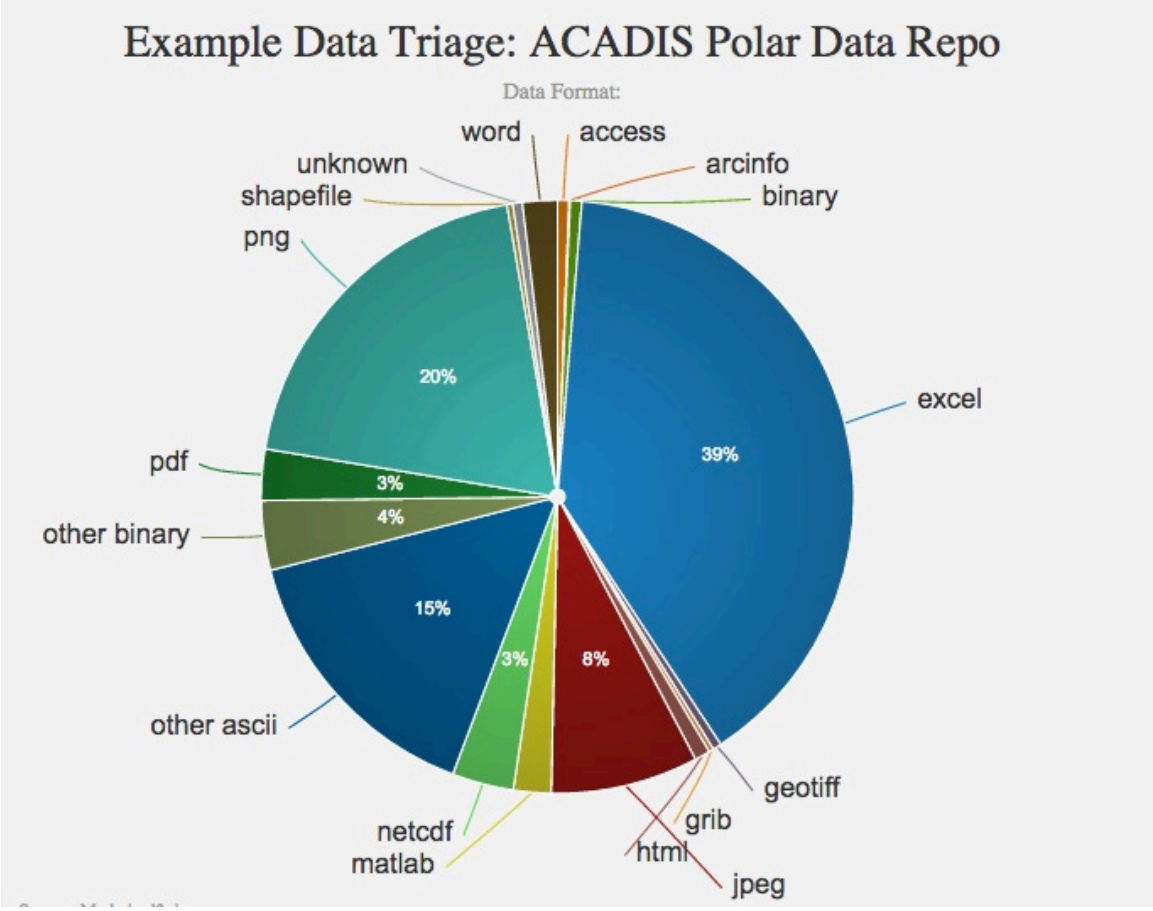


Figure 3. More precise extracted content using Apache Tika and Apache Nutch for the ACADIS Polar data repository. Our approach was able to extract 6% more of the data by correctly identifying many previously "unknown" MIME types in ACADIS.

Data types found in Polar data repositories.

	Atmosphere	Agriculture	Biological	Biosphere	Climate	Cryosphere	Human Dim.	Land Surface	Oceans	Paleoclimate	Solid Earth	Hydrosphere	%	
arcinfo							1						0.1%	
binary			4	1		2			4	1			0.6%	
geotiff				2		5	1	1					0.5%	
html	3			2	1	5	1	2	2				0.8%	
jpeg	8			21		70			2			52	8.0%	
matlab						9			29			1	2.1%	*
access			2	2	1	2	2	2	1				0.6%	
excel	55	7	10	176	11	156	4	115	82	39	7	87	39.4%	
word						11		11	3	9		2	1.9%	
netcdf	35				2	4			20			3	3.4%	*
other ascii	62	2	2	49	13	53		11	85			13	15.3%	
other binary	6	2	4	5	2	13	3	17	12	3	3	2	3.8%	
pdf	2			1		13	1	13	6	12	1	4	2.8%	
png	349						22	1	1	1			19.7%	
shapefile					1	2	1					2	0.3%	
grib	4							1					0.3%	*
unknown									10				0.5%	

Figure 4. Our approach identifies more precisely three previously unknown data types in web crawls – Matlab; NetCDF files, and Grib files. These files were acquired as a result of using Nutch Selenium integration, and Apache Tika on ACADIS, AMD and ADE.

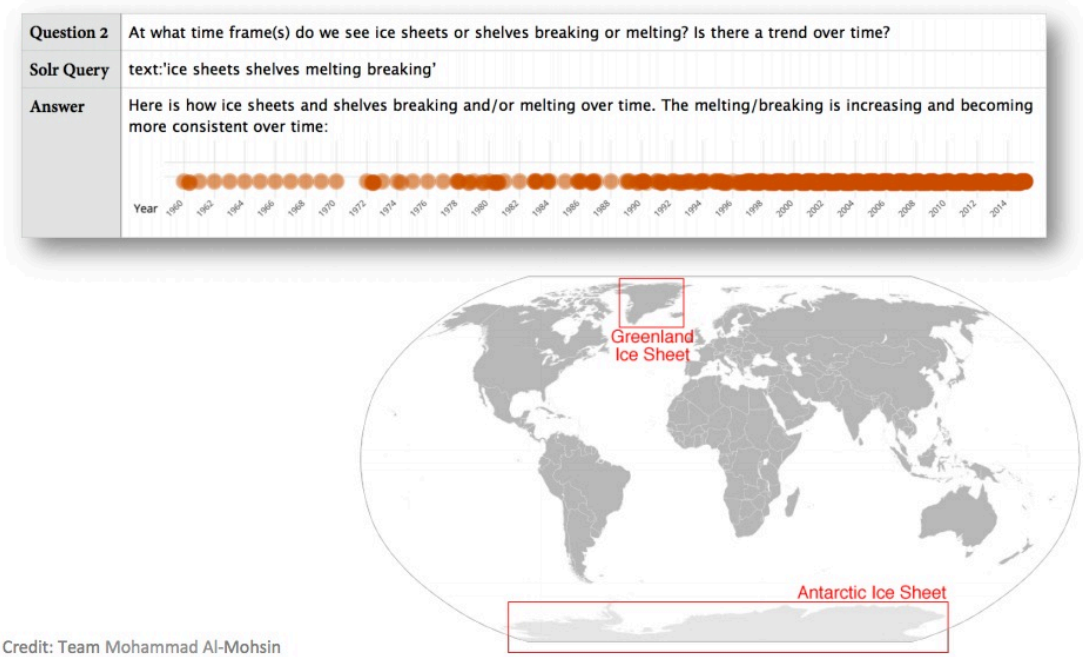
6 Querying the Data

We also explored creating ranking and querying algorithms for our crawled data. In particular, we developed and compared two sets of ranking and retrieval approaches: content-based approaches that will leveraged the indexed textual content of the TREC Dynamic Domain Polar data, which included term frequency-inverse document frequency (TF-IDF) to generate relevancy; and link-based approaches that leveraged citation relationships (graph-based) between the indexed documents and information other than the textual content of the document to perform relevancy analysis. These algorithms were focused then on allowing better answers to the below representative science queries of our Polar data:

1. What time-based trends exist for discussion of oil, iron and other natural resources in the Arctic region? Are documents and topics collocated to geographic region?
2. How many regions of interest are represented by the data you collected? Identify geographic "regions" as e.g., Circumpolar Arctic region, Antarctica and the Southern Ocean. Can you use the distribution of your documents to represent territories?
3. Can you predict areas in which there are national security interests (maritime/air/sea and land)? Which areas and why?
4. Is there a trend with respect to science data and measurements related to Climate Change? Is it time-based and/or geographic region based? What areas show a high document relevancy for sea-ice extent and decline?

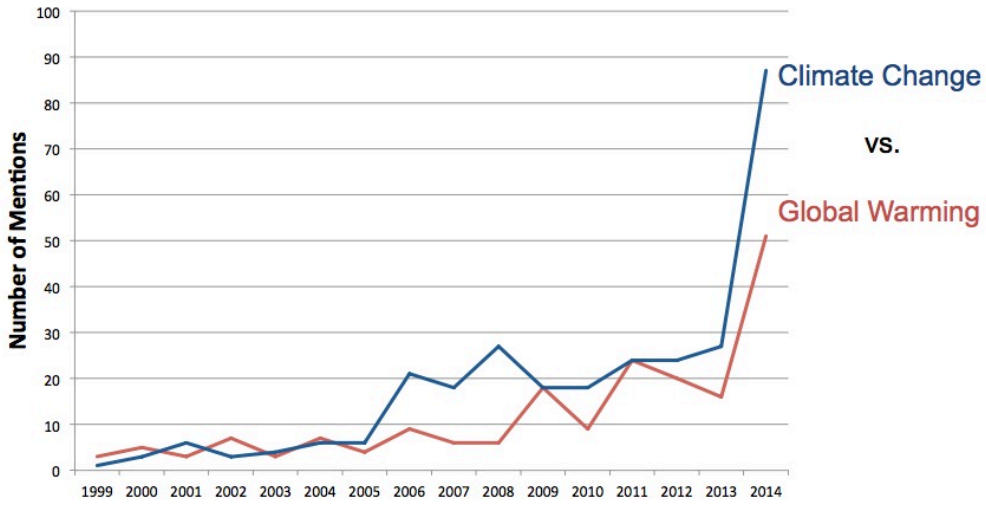
Some example answers to the above queries are demonstrated below in Figure 5.

Example 1:



Credit: Team Mohammad Al-Mohsin

Example 3:



Source: Chetan Vazirabadkar, Neeraj Joshi, Anudeep Katragadda and Niki Parmar

Figure 5. Example answers derived to Science queries of the TREC Dynamic Domain Polar data using our ranking and relevancy algorithms.

7 Acknowledgements

The TREC-DD Polar Science dataset was compiled, in part, by students in USC CS572. Individual contributors include: Lavina Advani, Mohammad Al-Mohsin, Chandrashekar Chimbili, Saurabh Gadia, Shashank Harinath, Chitra Arun Kumar, Chris Mattmann Lewis McGibbney, Indu Mohanan, Pradeep Muruganandam ,Subodh Sah, Mike Starch, Praneet Surana, Mahesh Goud Tandarpally, Giuseppe Totaro, Rishi Verma, Mengying Wang, Tianxiang Yu, and Jiaheng Zhang.

This work was partially supported by NSF Polar Cyberinfrastructure award numbers PLR-1348450 and PLR-144562. In addition the DARPA XDATA/Memex program funded a portion of the work. Effort supported in part by JPL, managed by the California Institute of Technology on behalf of NASA.

8 References

1. A. B. Burgess, C. Mattmann. Automatically Classifying and Interpreting Polar Datasets with Apache Tika. In Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, pp. 863-867, August 13-15, 2014, San Francisco, CA.
2. CSCI 572: Search Engines @ USC, http://sunset.usc.edu/classes/cs572_2015/, Accessed: October 2015.
3. A. B. Burgess. Apache Tika: Cool Insights Into Polar Data. ApacheCon NA 2015. http://events.linuxfoundation.org/sites/events/files/slides/ApacheCon2.pdf, Accessed: October 2015.
4. C. Mattmann. TREC Dynamic Domain Polar Dataset. https://github.com/chrismattmann/trec-dd-polar/, Accessed: October 2015.