TREC Dynamic Domain: Polar Science

Annie Bryant Burgess1,2,3, Chris Mattmann1,2,

… other authors

1University of Southern California

2NASA Jet Propulsion Lab

3Earth 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 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 provide TREC-DD a new challenge in in the realm of search relevancy.

# Introduction

## 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).

* Climate change, Polar amplification
* Data collection and dissemination - no more dark data

# Data Triage of Polar Repositories

To maximize extractible content from our web crawl, we first needed to expand Apache Tika to parse scientific data formats commonly found in Polar repositories.

## Tika expansion

Apache Tika is.

# 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).

## Apache Nutch

Each web crawl used [Apache Nutch](http://nutch.apache.org/) as the core framework for web crawling and [Apache Tika](http://tika.apache.org/) as the main content detection and extraction framework. Nutch is a distributed search engine that runs on top of [Apache Hadoop](http://hadoop.apache.org/).

# Dataset Preparation

## Duplicate Records

Exact duplicate records were removed using signature based methods. Algorithms and accompanying code were developed to remove near duplicates, using jaccard similarity, by graduate students in USC CS572. However, not all teams that submitted web crawls to this dataset applied their jaccard-similarity algorithms.

## Data Format

Crawled data were put into Common Crawl Format, acording to Memex 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 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, aggregate.py, aggregates all of the JSON files.

# Dataset Characteristics

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.

# 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.

# References