



APACHE CON  
**DENVER**  
WESTIN DENVER DOWNTOWN  
APRIL 7-9, 2014

# A case for polar data: integrating the NetCDF file type into Apache Tika

Annie Bryant Burgess. Postdoctoral Fellow. USC

Presented For The Apache Foundation By  
 **LINUX FOUNDATION**

Background:

PhD, Geography

Focus: Remote Sensing and  
Snow Hydrology

Digging snow...  
and writing code.



Background:

PhD, Geography

Focus: Remote Sensing and  
Snow Hydrology

Digging snow...  
and writing code.



Theory and Experimentation



Dataset Creation



Data Dissemination

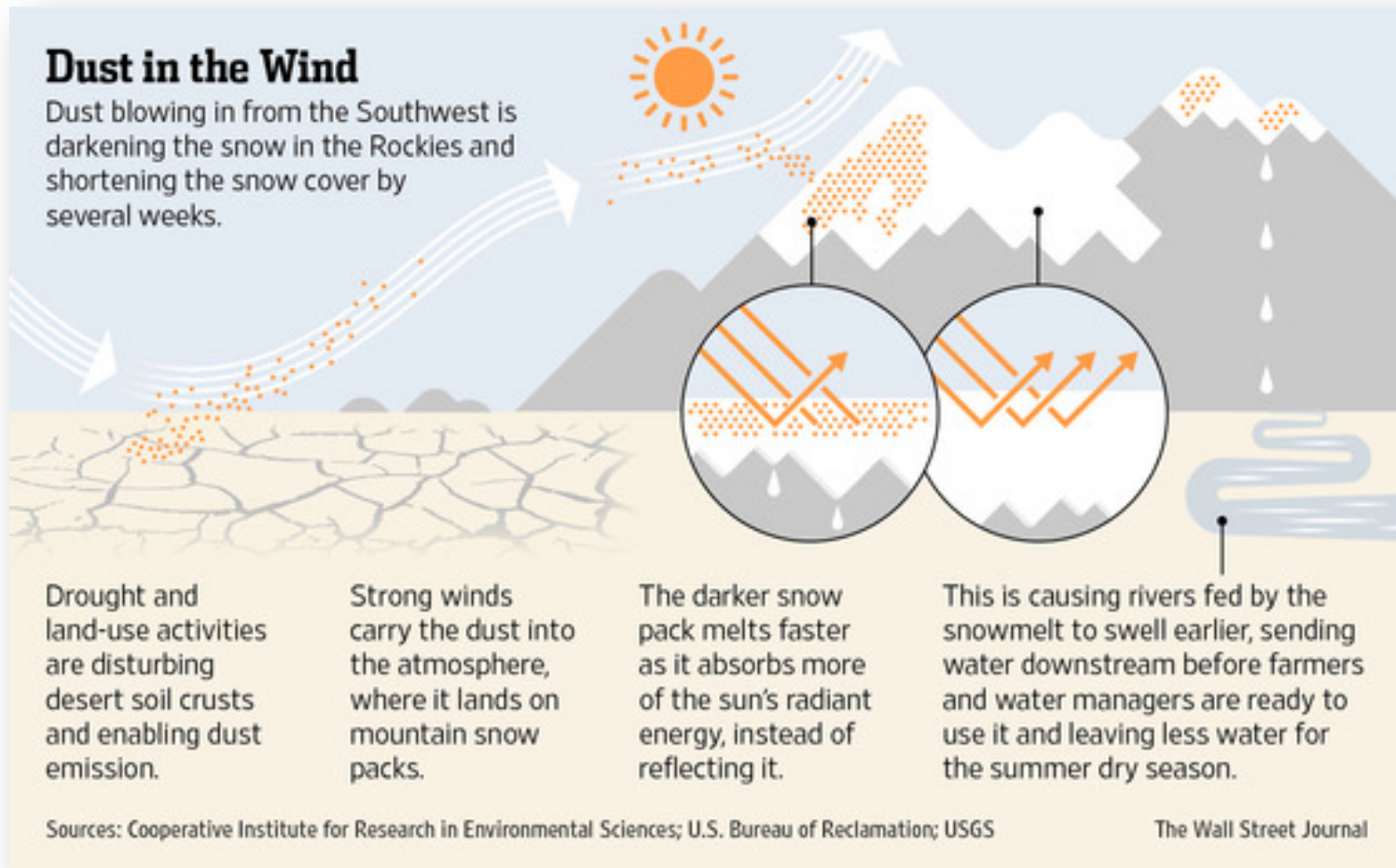


Data-Driven Science





# How much extra energy is absorbed by snow when it becomes dirty?



...and why should you care!



Photo Credit: Peter McBride

From understanding physical  
phenomenon...

to managing a data bonanza...

How one *becomes* a data  
scientist.



From understanding physical phenomenon...

to managing a data bonanza...

How one *becomes* a data scientist.

**Daily, how much additional energy is absorbed in  $W m^{-2}$  because the snow is dirty between 2000 - present?**

**12 KB of data... not quite a data scientist yet!**







From understanding physical phenomenon...

to managing a data bonanza...

How one *becomes* a data scientist.

**At ANY snow-covered area around the globe, how much more energy is absorbed by dirty snow daily?**

**> 27 TB of data... and growing**







Our data have shown...  
These results  
indicate...

Hmm... I wonder if  
those data are  
available.

We want  
your data!



Data dissemination  
through :

NASA/JPL  
[snow.jpl.nasa.gov](http://snow.jpl.nasa.gov)

**NASA** Jet Propulsion Laboratory  
California Institute of Technology

JPL HOME EARTH SOLAR SYSTEM STARS & GALAXIES SCIENCE & TECHNOLOGY  
BRING THE UNIVERSE TO YOU: JPL Email News | RSS | Mobile | Video

Home Data Publications Media People Links

## SNOW DATA SYSTEM

### Welcome to the Snow Data System

#### Data

This page provides a single point of entry into the entire repository of data provided by the Snow Data System.

#### Tips and Hints for Using Snow Data Products

The following page contains important information about understanding the data products available from this site.

#### Snow Data Product Browser

The Snow Data System product browser provides detailed information and the ability to download raw data products from the Snow Data System archive.

[Access the Data](#)

#### Snow Map (Experimental)

This experimental map overlays data from multiple remote sensing and in situ sources to provide a comprehensive picture of snow and ice properties.

[Access the Map](#)

#### Western Energy Balance of Snow (WEBS) Data

View Western Energy Balance of Snow data plotted by station and parameter, and download the raw input data directly.

[Access the Data](#)

Snow Data System Portal  
NASA Jet Propulsion Laboratory  
Built On Apache OODT  
[Privacy](#) [Image Policy](#) [Login](#)

Site Contact: Andrew F. Hart  
OODT Balance v. 0.3-SNAPSHOT

Theory and Experimentation



Dataset Creation



Data Dissemination



Data-Driven Science

Theory and Experimentation



Dataset Creation



Data Dissemination



**Data Discovery**

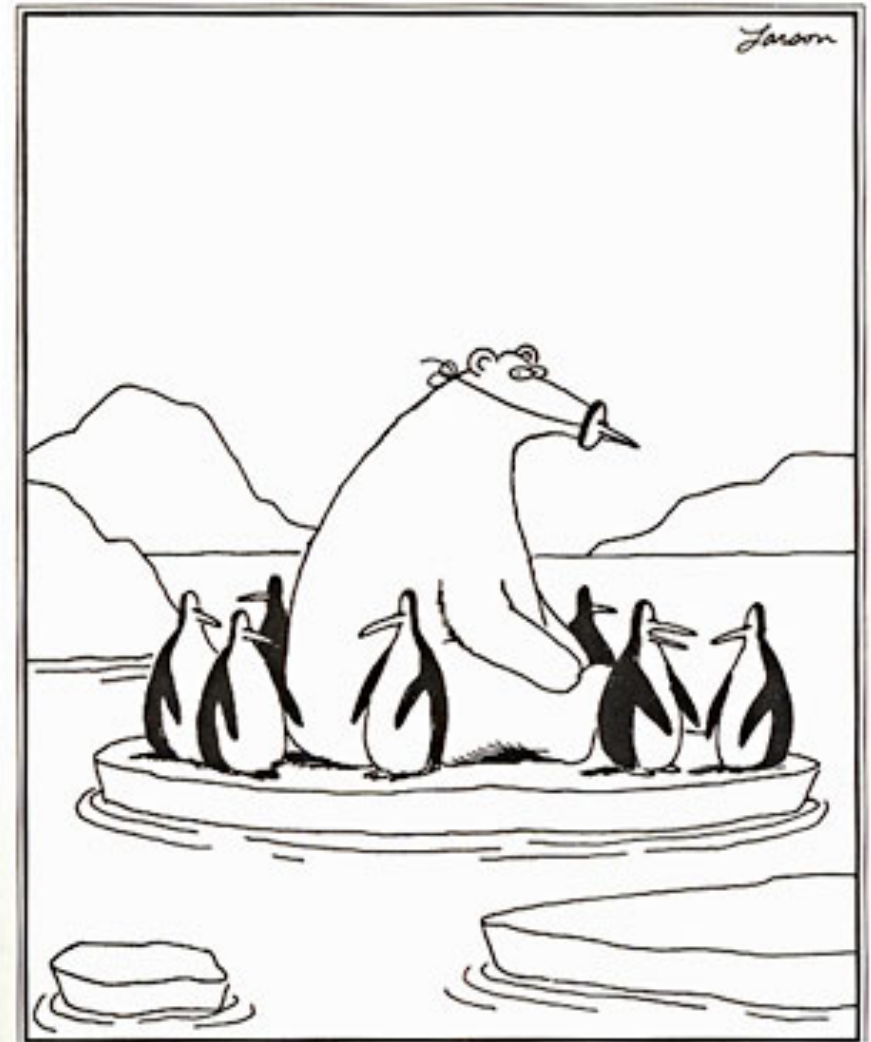


**Data-Driven Science**



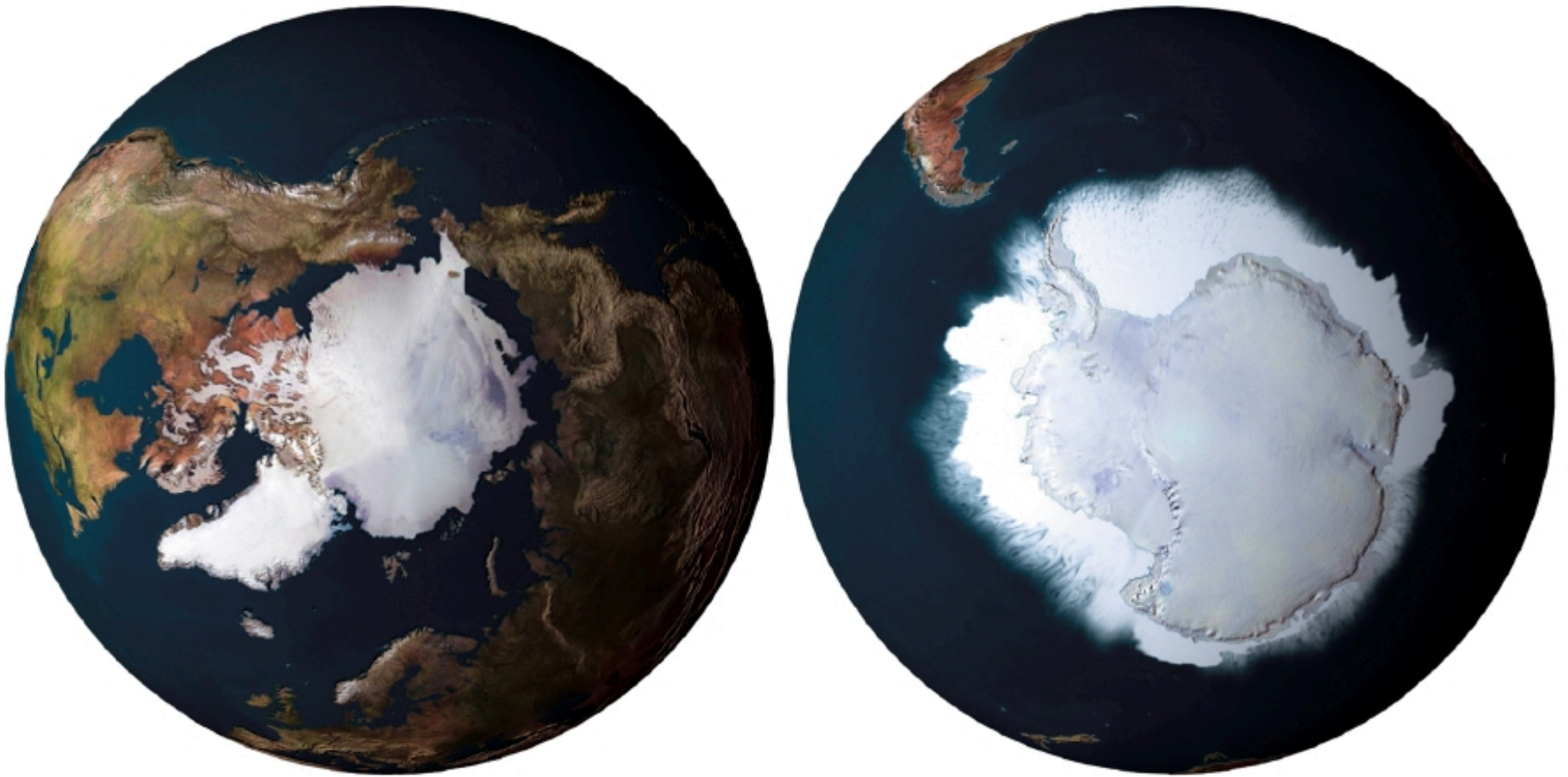
If my data are just in *another* data repository... are they truly discoverable?

# Using Apache Tika to explore Polar Data



"And now Edgar's gone. ... Something's going on around here."

# What are 'Polar' data?



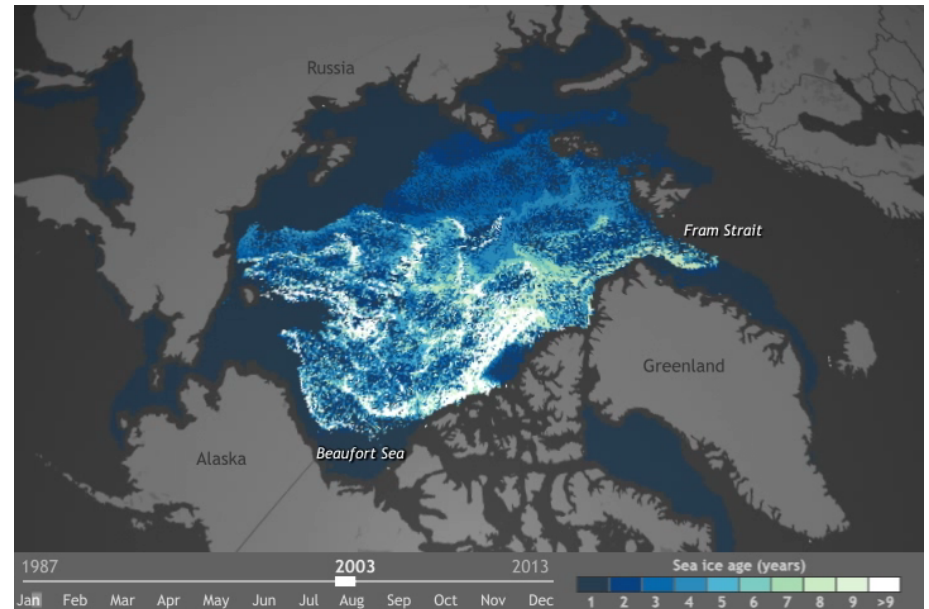
# Why is polar science so important?

**Rapid change** – The polar regions are changing faster than any other region on Earth.

**Global linkages** – The polar regions have a profound influence on the global environment, particularly on the weather and climate system.

**Human societies** – The Arctic is home to more than 4 million people. These communities face changes to their environment and the resources on which they depend.

**Sense of discovery** – The polar regions are places of wonder and, even at the beginning of the 21st century, remain largely physically and intellectually unexplored.



NSIDC/NASA



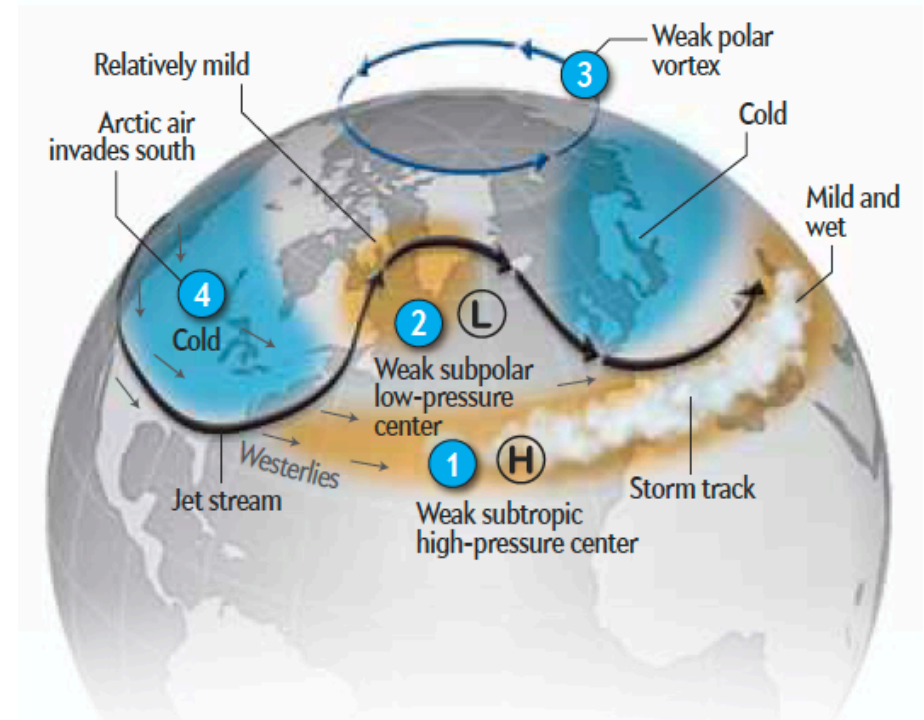
# Why is polar science so important?

**Rapid change** – The polar regions are changing faster than any other region on Earth.

**Global linkages** – The polar regions have a profound influence on the global environment, particularly on the weather and climate system.

**Human societies** – The Arctic is home to more than 4 million people. These communities face changes to their environment and the resources on which they depend.

**Sense of discovery** – The polar regions are places of wonder and, even at the beginning of the 21st century, remain largely physically and intellectually unexplored.



# Why is polar science so important?

**Rapid change** – The polar regions are changing faster than any other region on Earth.

**Global linkages** – The polar regions have a profound influence on the global environment, particularly on the weather and climate system.

**Human societies** – The Arctic is home to more than 4 million people. These communities face changes to their environment and the resources on which they depend.

**Sense of discovery** – The polar regions are places of wonder and, even at the beginning of the 21st century, remain largely physically and intellectually unexplored.



# Why is polar science so important?

**Rapid change** – The polar regions are changing faster than any other region on Earth.

**Global linkages** – The polar regions have a profound influence on the global environment, particularly on the weather and climate system.

**Human societies** – The Arctic is home to more than 4 million people. These communities face changes to their environment and the resources on which they depend.

**Sense of discovery** – The polar regions are places of wonder and, even at the beginning of the 21st century, remain largely physically and intellectually unexplored.



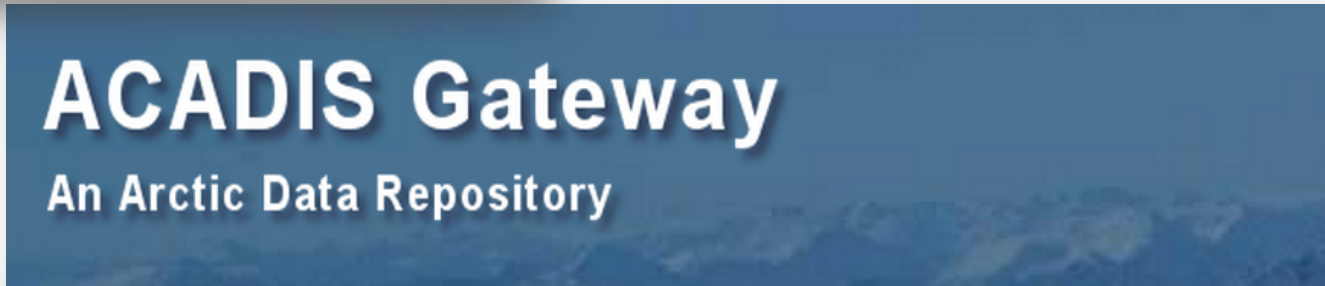


Polar scientists need provenance, content, format, and quality information to **identify** the right dataset, **evaluate uncertainty**, and **ensure the replicability** of scientific workflows. (From: *Workshop on Cyberinfrastructure for Polar Sciences*)



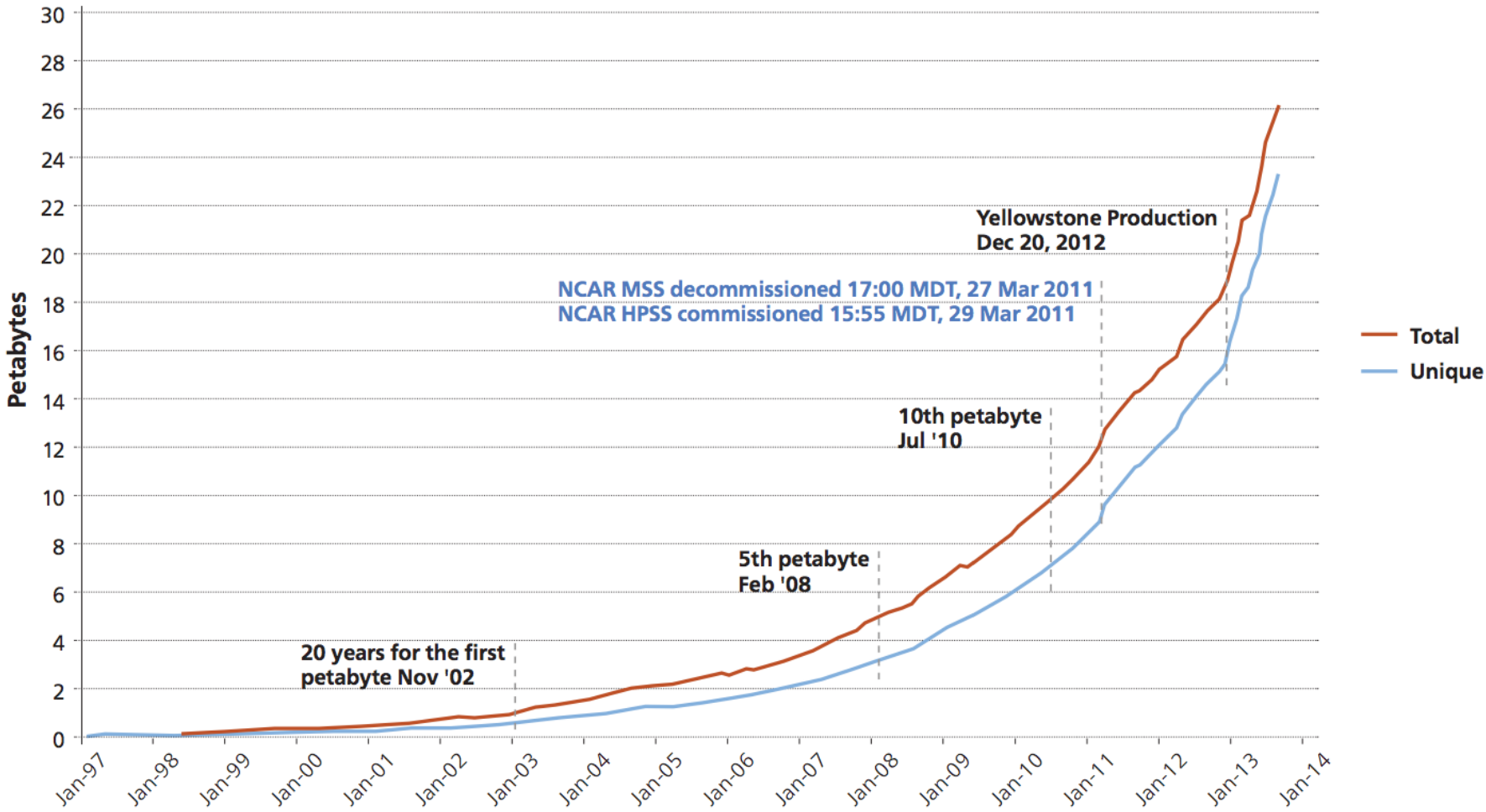


**Data Driven**      **OR**      **Theory and Experimentation**  
Must navigate:



... AND MANY MORE!

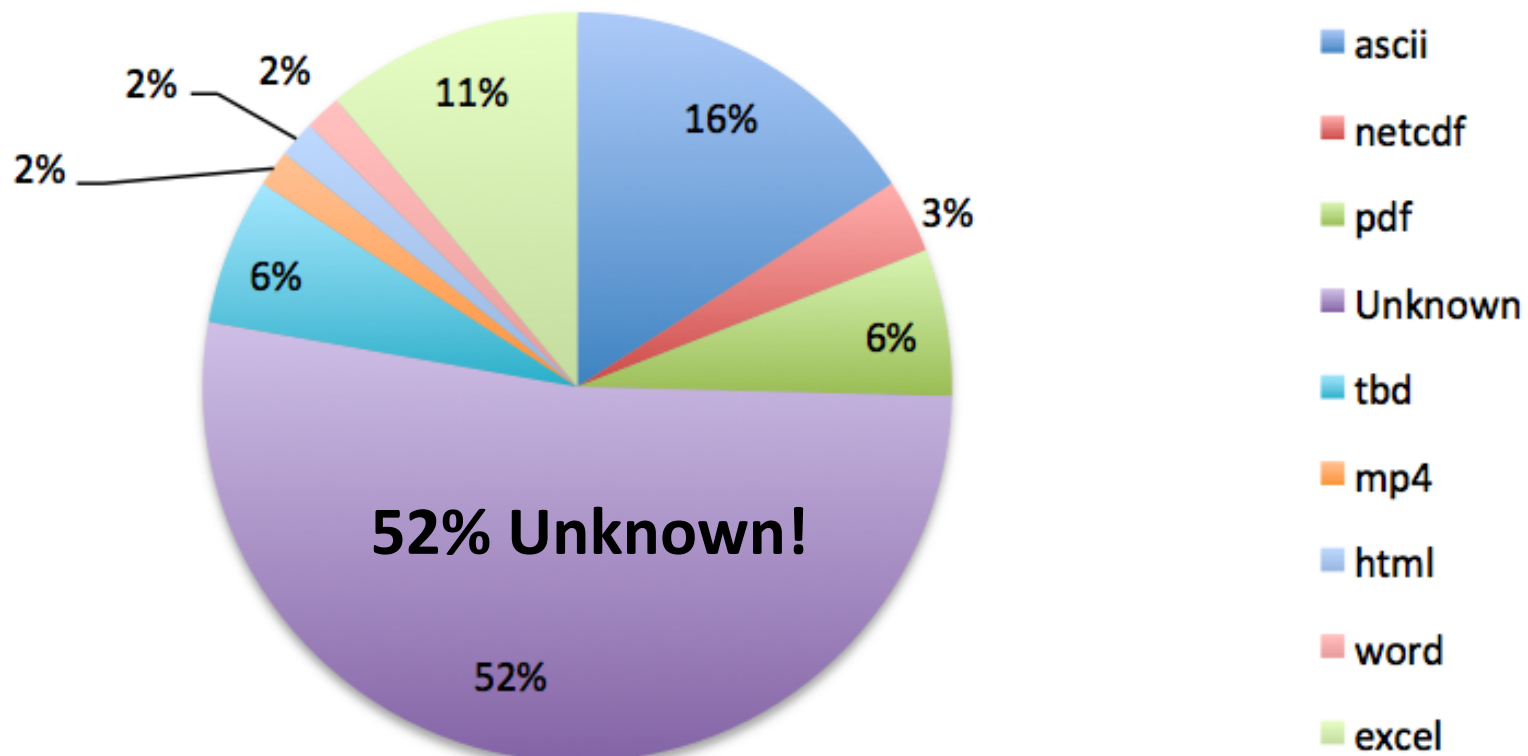
# NCAR Total Data Archive



# Apache Tika

- Content detection and analysis toolkit:
  - Automated MIME type identification and rapid parsing of text and metadata.
  - 1200 types of files including all major file types from the Internet Assigned Number Authority's MIME database.

# File types recognized by Tika in the Antarctic Master Directory



Goal: Add file formats commonly used in polar data to Tika.

Ex: NetCDF (extend), HDF (extend), IMG, GRIB,...

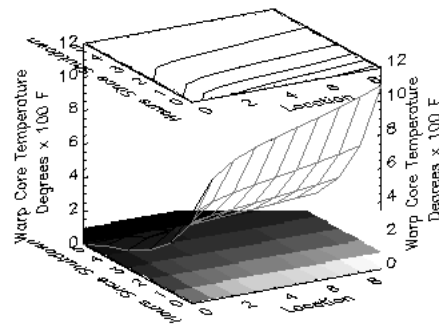
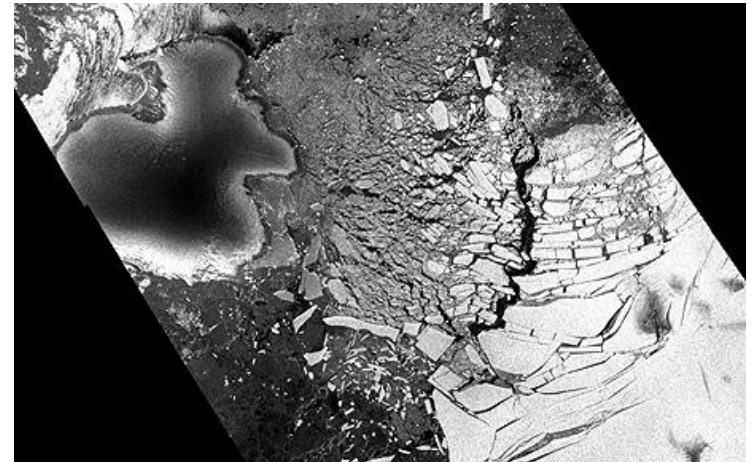


Figure 5-1: SHOWS Result of Unlimited Dimensions Example



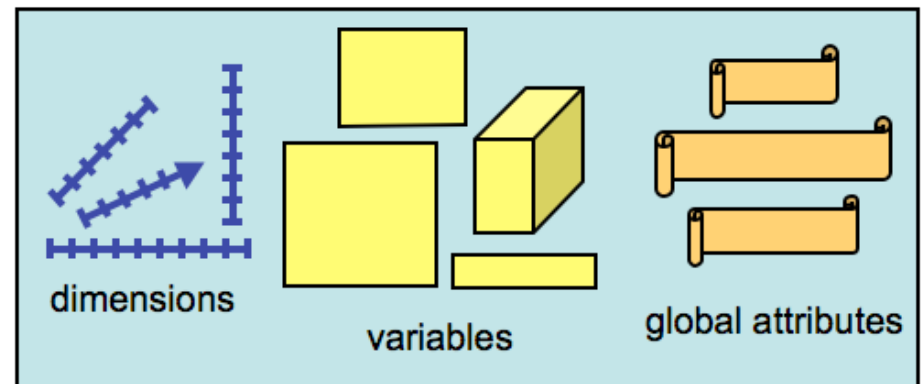
# Adding new file formats

- Augment Tika's MIME detection facilities to handle the new scientific formats.
  - Add glob patterns, regular expressions, MIME “magic” or digital signatures for files, and curation of XML root namespaces and patterns.
- Create new Tika parsers and expand existing parsers.



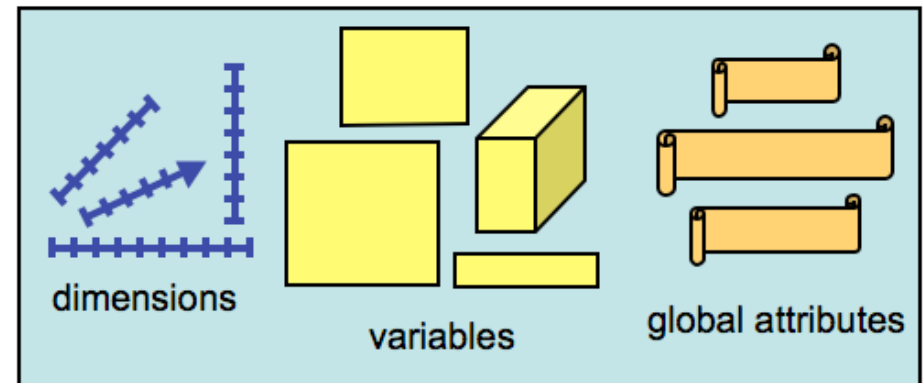
# NetCDF File Format

- Variables:
  - N-dimensional arrays of data. Variables in netCDF files can be one of six types (char, byte, short, int, float, double).
- Dimensions:
  - Describe the axes of the data arrays. A dimension has a name and a length.
- Attributes:
  - Annotate variables or files with small notes or supplementary metadata.



# NetCDF File Format

- Variables:
  - N-dimensional arrays of data. Variables in netCDF files can be one of six types (char, byte, short, int, float, double).
- Dimensions:
  - Describe the axes of the data arrays. A dimension has a name and a length.



- **Attributes:**
  - **Annotate variables or files with small notes or supplementary metadata.**

Current Tika Output

**[ab:~/tika/tika] ab% java -jar tika-app-1.5.jar --metadata /NetCDF/lsmask.nc.nc**

Content-Length: 106968

Content-Type: application/x-netcdf

Conventions: COARDS

Reference: <http://www.cdc.noaa.gov/cdc/data.unified.html>

dc:title: Unified Daily Gridded US Daily Precipitation

description: Gridded daily Precipitation

history: Thu Oct 16 16:48:20 2003: ncatted -a statistic,lsmask,c,c,Other lsmask.nc

Thu Oct 16 16:47:44 2003: ncatted -a level\_desc,lsmask,c,c,Surface lsmask.nc

Thu Oct 16 16:47:23 2003: ncatted -a dataset,lsmask,c,c,CPC .25x.25 Daily US UNIFIED  
Precipitation lsmask.nc

Thu Oct 16 16:46:18 2003: ncatted -a var\_desc,lsmask,c,c,Land-sea mask lsmask.nc

created 11/02/2000 by CAS from data obtained from NCEP

platform: Observations

resourceName: lsmask.nc.nc

title: Unified Daily Gridded US Daily Precipitation

[abryant:~/tika/tika] abryant%

**[ab:~/tika/tika] ab% java -jar tika-app-1.5.jar --text /NetCDF/lsmask.nc.nc**

[abryant:~/tika/tika] abryant%

## Tika outputs NetCDF *Attributes*:

Content-Length: 106968  
Content-Type: application/x-netcdf  
Conventions: COARDS  
Reference: <http://www.cdc.noaa.gov/cdc/data.unified.html>  
dc:title: Unified Daily Gridded US Daily Precipitation  
description: Gridded daily Precipitation  
history: Thu Oct 16 16:48:20 2003:  
dataset,ismask,c,c,CPC .25x.25 Daily US UNIFIED Precipitation Ismask.nc  
created 11/02/2000 by CAS from data obtained from NCEP  
platform: Observations  
resourceName: Ismask.nc.nc  
title: Unified Daily Gridded US Daily Precipitation

## Tika *will* output NetCDF *Dimensions* and *Variables*:

### Dimensions:

lon = 321;  
lat = 161;

### Variables:

```
float lat(lat=161);  
  :units = "degrees_north";  
  :long_name = "Latitude";  
  :actual_range = 20.0f, 60.0f; // float  
float lon(lon=321);  
  :units = "degrees_east";  
  :long_name = "Longitude";  
  :actual_range = 220.0f, 300.0f; // float  
double time(time=1);  
  :units = "hours since 1-1-1 00:00:00";  
  :long_name = "Time";  
  :actual_range = 0.0, 0.0; // double  
  :delta_t = "0000-00-01 00:00:00";  
  :avg_period = "0000-00-01 00:00:00";  
short ismask(time=1, lat=161, lon=321);  
  :long_name = "Land Sea Mask";  
  :valid_range = -1S, 1S; // short  
  :actual_range = -1.0f, 1.0f; // float  
  :add_offset = 0.0f; // float  
  :scale_factor = 1.0f; // float  
  :missing_value = 32766S; // short  
  :var_desc = "Land-sea mask";  
  :dataset = "CPC Gauge-Based Analysis of Daily Precipitation over CONUS";  
  :level_desc = "Surface";  
  :statistic = "Other";
```

Step 1: Add NetCDF --text parsing

Step 1 re-thought: **Improve Java skills!**

Step 2: `System.out.println( "Realize I've got to  
get on the Object Oriented train and OFF  
the Procedural" );`

Step 3: Start writing code and see what Tika  
can do!

Integration of scientific data types will allow more...



National Snow & Ice Data Center



British  
Antarctic Survey

NATURAL ENVIRONMENT RESEARCH COUNCIL

**ACADIS Gateway**

An Arctic Data Repository

**POLAR DATA**  
CATALOGUE



ANTARCTIC  
MASTER  
DIRECTORY

A Global Change Master Directory Portal



Integration of scientific data types will allow more...

**DATA DISCOVERY!**



National Snow & Ice Data Center



British Antarctic Survey

NATURAL ENVIRONMENT RESEARCH COUNCIL

**ACADIS Gateway**

An Arctic Data Repository

**POLAR DATA**  
CATALOGUE



ANTARCTIC  
MASTER  
DIRECTORY

A Global Change Master Directory Portal

# Closing Thoughts

- In scientific data, what should Tika see as 'text' vs. 'metadata'?
  - i.e. in NetCDF files, should *Variables* and *Dimensions* also be considered metadata?

# Questions/Acknowledgements

Thanks to:

- NSF Polar Cyberinfrastructure
- Chris Mattmann, JPL, USC
- TAC Committee for travel funding

@AB\_in\_AK  
anniebryant@gmail.com

