ClimoBase: Lessons Learned While Rescuing Observational Data From Extinction

Katie Schmitt

@kmschmitt

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Foundations of Data Curation

- **Data Curation:**
  - The active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education
  - Curation activities and policies enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time

Source: Graduate School of Library and Information Science at University of Illinois at Urbana-Champaign

Foundations of Data Curation

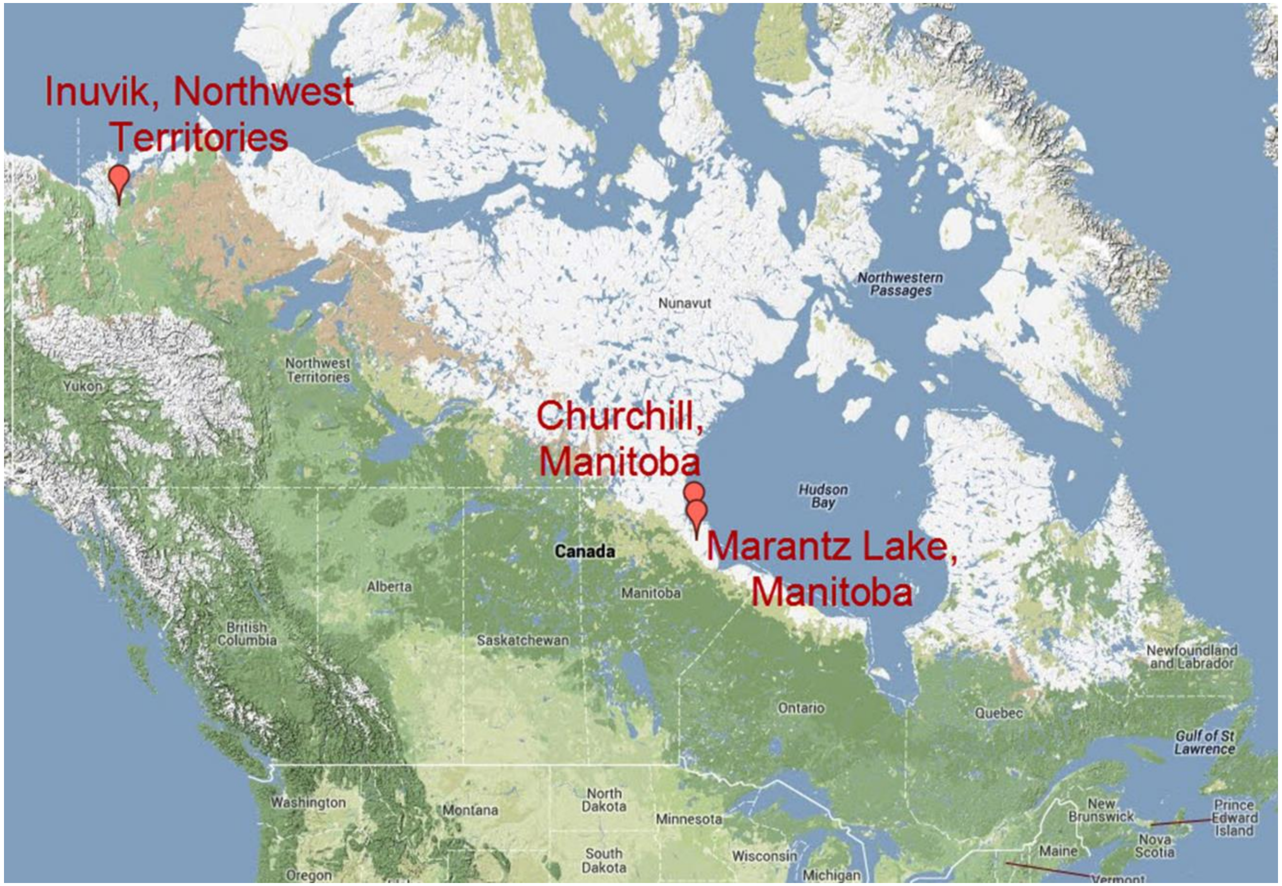
- **Data Rescue:**
  - The process of securing data at risk of being lost due to deterioration or simple obsolescence of the storage media, natural hazards, theft or vicious destruction, and ensuring that data can be easily accessed and used.

Source: World Meteorological Association
http://www.wmo.int/pages/prog/hwrp/datarescue.php

ClimoBase

- **Data collected by Dr. Wayne Rouse (PI)**
  - 1984 – 1998
  - Surface-climate studies in Northern Canada.
    - Regular observational measurements
      - Some recorded in 15 minute increments
    - Variety of terrains
      - Boulders
      - Wetlands
      - Forest, etc.

ClimoBase



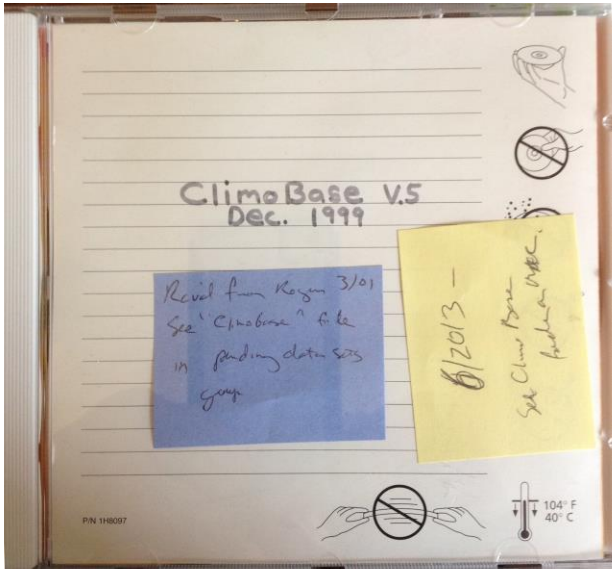
Map data ©2014 Google

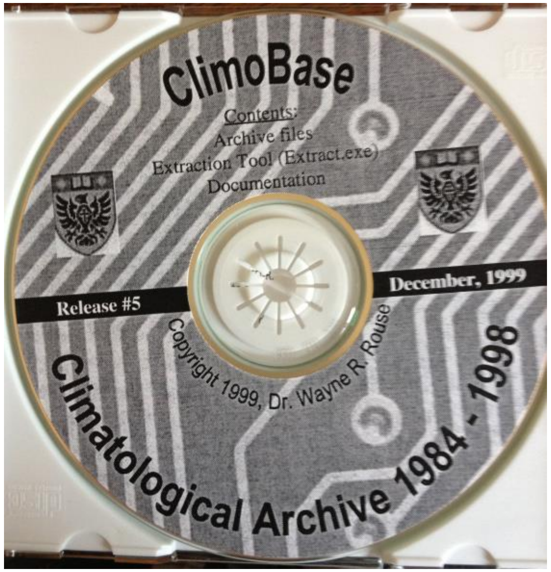
ClimoBase

- **State of the Collection**
  - 7000 files
  - Fortran extraction program
  - Minimally processed
  - Encoded file naming convention
  - Customized ‘NULL’ value
  - Fortran-delimited
  - Minimal metadata

ClimoBase

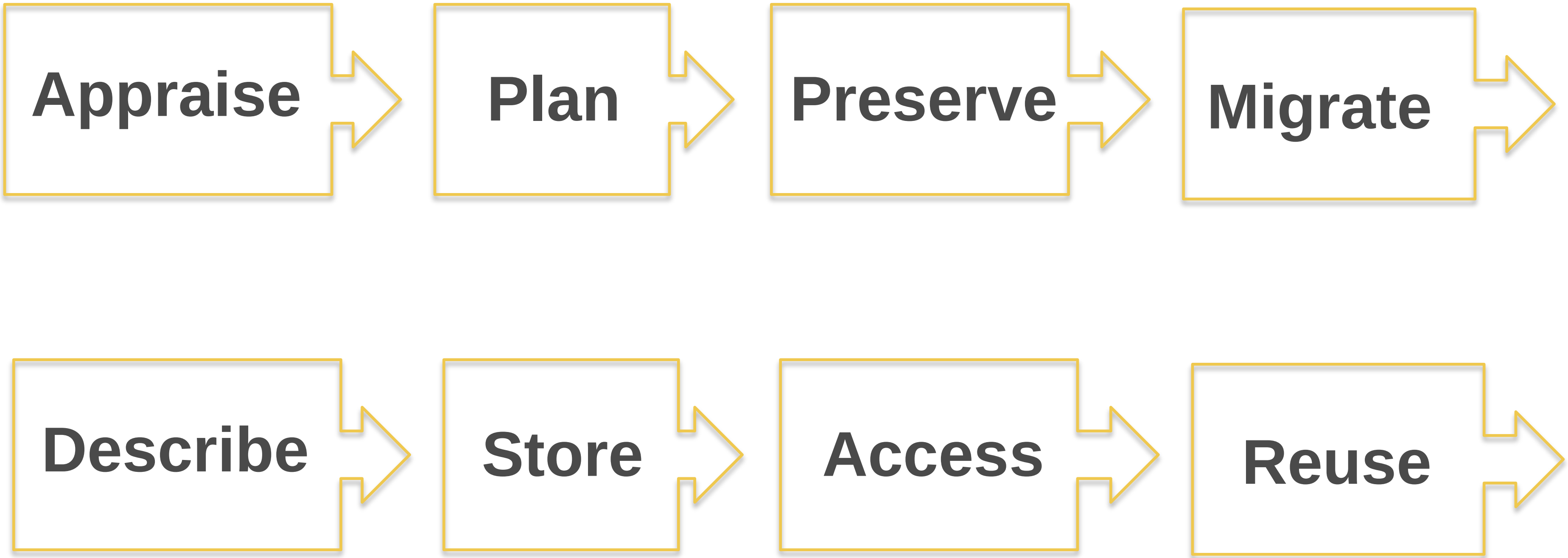
- **Supplemental files**
  - "Users' Manual" with Fortran code
  - CD labels





ClimoBase
Contents:
Archive files
Extraction Tool (Extract.exe)
Documentation
Release #5
December, 1999
Copyright 1999, Dr. Wayne R. Rouse
Climatological Archive 1984 - 1998

Data Rescue Workflow



Appraise

Plan

Preserve

Migrate

Describe

Store

Access

Reuse

Adapted from the DCC Curation Lifecycle Model
http://www.dcc.ac.uk/resources/curation-lifecycle-model

Workflow: Appraise

- **Evaluate data on potential significance**
  - Do your research and use all evidence at hand
    - Interview PI
    - Read documentation
      - Text files
      - Labels on physical storage containers
- **Cost-benefit analysis**

Workflow: Plan

- Where will the data be stored?
- What are the best file formats?
- How/where will the data be processed?
- What metadata are needed or required?

Workflow: Preserve & Migrate

- **Ensure all original data is safe and secure**
  - Use hardware to block changes to physical storage during a transfer
  - Preserve any supplemental documentation and code
- **Migrate a copy of original data**
- **Clean and validate the “new” data**

Workflow: Describe

- **Create metadata**
  - Use research from appraisal step
  - Interview PI’s and data creators/collectors
  - Track provenance
  - Examine physical materials and supplemental files
- **Create README.txt files**
- **Trust is key!**

Workflow: Store

- **Follow best practices for preservation**
  - LOCKSS: Lots Of Copies Keep Stuff Safe
    - Store three copies of data (one offsite)
  - Monitor fixity
    - Use checksums to routinely check for bit rot
  - Migrate data to new formats as needed

Workflow: Access and Reuse

- **Use metadata for discoverability**
- **Deliver copies of data to users (as needed)**
    - Use checksums to ensure fixity for each copy
- **Monitor usage statistics and citations**
- **Begin lifecycle process for versioned or manipulated data**

Lessons Learned

- **Get to know the data before you start**
  - Identify any potential issues
    - Provenance
    - Hardware/software
    - Unknown variables
    - Inconsistent NULL Values
  - Recruit subject matter experts
  - Ask questions

Lessons Learned

- **Think strategically**
  - What will you realistically be able to accomplish?
  - Will anyone use the data when you are done?
  - Budget considerations
    - Time
    - Hardware
    - Storage

Next Steps

- **NOAA @ NSIDC collection**
  - Work with NSIDC technical writers
  - Assign a unique DOI number
  - Available for download at www.nsidc.org by Fall 2014

Final Thoughts

Without quality metadata, you have nothing!

Acknowledgements

- **Florence Fetterer**
  - Project Manager and NOAA Liaison, NSIDC
- **Ruth Duerr**
  - Data Stewardship Program Manager, NSIDC
  - Adjunct Professor, GSLIS
- **Dr. Wayne Rouse**
  - Professor Emeritus at McMaster University

Questions?

Katie Schmitt

@kmschmitt

kmschmi2@illinois.edu

http://katieschmitt.me