Provenance, Planning and Production

Basic Facts
- Data come in Time Series of Files
- Data produced in discrete time series of jobs

Definitions
- Files belong to Data Products:
  - Same Data Time Intervals in each file, S
  - Same Parameters in each file
- Data Sets are Subsets of Data Products:
  - Data sources determine start and end times
  - Number of possible files in Data Set = End - Start
- Data Set Versions are Subsets of Data Data:
  - Error structure of data in a Data Set version should be homogeneous

A Climate Data Record is a Single Data Set Version

Derivation
- Jobs also come in discrete time series
- There are 4 relationships between files and jobs:
  1. File is produced by a job
  2. File is used as input by a job
  3. Job uses a file as input
  4. Job produces a file as output
- Files and Jobs form a graph (or network):
  1. Files and Jobs form Vertices (nodes)
  2. Relationships produce Edges (arcs)

Common operations correspond to graph manipulations

Implications
- Error Sources Changes:
  1. Input Data
  2. Source Code
  3. Parameters
  4. Connectivity

Derived Collection Organization:
- Archive
- Data Producer
- Data Product
- Data Set
- Data Set Version (Climate Data Record)
  - File

Provenance Tracking Graph

Production Planning Graph

Bruce R. Barkstrom
NOAA’s National Climatic Data Center, Asheville, NC