Provenance of Earth Science Datasets – How Deep Should One Go?

Gerald Manion\(^1\) geraldjohn.m.manion@jpl.nasa.gov, Hampapuram Ramapriyan\(^2\) Hampapuram.Ramapriyan@ssaihq.com, Steve Aulenbach\(^3\) saulenbach@usgs.gov, Brian Duggan\(^4\) brian@promptworks.com, Justin Goldstein\(^5\) goldstein@usgcrp.gov, Hook Hua\(^6\) hook.hua@jpl.nasa.gov, Dexter Tan\(^7\) dexter.c.tan@jpl.nasa.gov, Curt Tilmes\(^7\) curt.tilmes@nasa.gov, Brian Wilson\(^1\) bdwilson@jpl.nasa.gov, Robert Wolfe\(^8\) reiwolfe@usgcrp.gov, Stephan Zednik\(^9\) zednis2@rpi.edu
\(^1\)Jet Propulsion Laboratory, California Institute of Technology, \(^2\) Science Systems and Applications, Inc., \(^3\)US Geological Survey, \(^4\)PromptWorks, \(^5\)US Global Change Research Program, \(^6\)Raytheon Company, Pasadena, \(^7\)NASA Goddard Space Flight Center, \(^8\)Rensselaer Polytechnic Institute

Why Provenance?

- Transparency and reproducibility essential for scientific credibility
- Answers to scientific questions can have global socio-economic impact and are viewed critically with “healthy skepticism”
  - Emphasizes the need for reproducibility of influential scientific results. In this context, influential implies that the information has a “clear and substantial impact on important public policies or important private sector decisions”.
  - “If an agency is responsible for disseminating influential scientific, financial, or statistical information, agency guidelines shall include a high degree of transparency about data and methods to facilitate the reproducibility of such information by qualified third parties.”

Traceability – Prerequisite for Reproducibility

- From conclusion in a paper to objects, their provenance and context leading to the conclusion
- Provenance (lineage) tracing tools help – generally can express trace-back as (Entities (inputs & outputs), Activities, Agents)

Global Change Information System

- Established by [US Global Change Research Program (USGCRP)](http://www.usgcrp.gov) “to better coordinate and integrate the use of federal information products on changes in the global environment and the implications of those changes for society”.
- Open-source, web-based resource for “traceable, sound global change data, information, and products.”

The Third National Climate Assessment (NCA3)

- Summarizes impacts of climate change on the United States, now and in the future.
- Produced by >300 experts guided by a 60-member Federal Advisory Committee
- Extensively reviewed by the public and experts, including federal agencies and a panel of the National Academy of Sciences.

GCIS and NCA3

- GCIS provides capabilities for interactive exploration of NCA3
- NCA3 - “featured report” in GCIS
- Report has chapters, figures, tables, findings (key messages), and references
- Generally figures, tables and references support findings
- Datasets and/or images are used in figures
- Important to trace back to sources of all items supporting a key message
  - e.g. finding → figure → image → dataset → algorithm → input data → instrument → satellite; algorithm → ATBD → Reference; dataset → archive; dataset → metadata/documentation)

Key message: “Global sea level has risen by about 8 inches since reliable record keeping began in 1880. It is projected to rise another 1 to 4 feet by 2100”

- Inputs: four inputs for four sections of graph - Proxy Records, Tide Gauge Data, Satellite-Derived, Scenarios
- Method
  - Part 1: Plotted using Matlab code by first converting 51 units into feet (Trace back includes code)
  - Part 2: For satellite data (TOPEX, Jason-1, Jason-2), Time periods covered by each is given in table at [http://sealevel.colorado.edu/content/data-processing-methods](http://sealevel.colorado.edu/content/data-processing-methods). Processing method is described in detail in: DOI 10.1080/01490419.2010.491031
- Comment: Detailed trace back needed since data are used to draw quantitative conclusions

Key message: “Infrastructure will be increasingly compromised by climate-related hazards, including sea level rise, coastal flooding, and intense precipitation events”

- Inputs: Two granules of MODIS Aqua dataset MYD021KM
- Method: Input granules were mosaicked and subset to extract desired area to display region covered by Hurricane Irene
- Comment: Simple trace to identify granules used is enough - figure only illustrates extent of precipitation event; other sources are used to show quantitative consequences

Lessons Learned

- Trace back difficult after report delivery
  - Difficult to contact authors and follow-up to get complete provenance information
  - Follow-up attempts could be misinterpreted as questioning their research
  - Instructions and templates to record provenance should be provided to authors before report generation
  - Data and images used for the report should be held in a long-lived, user-accessible repositories
  - Generally information in the form of inputs, outputs, agents and activities (descriptive and/or mathematical) is useful for each dataset used, images or figures generated and key messages
  - Capturing the extra contextual knowledge around the findings could shed light into why certain decisions were made during the findings.
  - Depth of trace back should be commensurate with information needed to reproduce results

Figures in NCA3 using NASA Datasets

- Figure 2.26: Past and Projected Changes in Global Sea Level Rise
- Figure 9.3: Wildfire Smoke has Widespread Health Effects
- Figure 16.3: Flooding and Hurricane Irene
- Figure 16.5: Urban Heat Island
- Figure 33.14: Warming-trend-and-effects-of-el-Niño-la-Niña
- Figure 34.16: Observed Change in Global Average Temperature

Tracing Back – Typical Issues

- Missing source images – Available when the report was written; author downloaded and used it, but image not in long-term archive
- Author unavailable for answering questions
- Description of how a figure was created is not available (even when source of data has been acknowledged)
- Completeness of available metadata - Original spatial or temporal extent not available
- Handling discovered errors – cannot fix report but can note errata

IN21C-1700