Learning about Data Systems and Data Tools from Practitioner's Applications or Attributes of Usable Data Services

Richard B. Rood
rbrood@umich.edu
with thanks to
Paul N. Edwards
bunch of students and "clients"

Rood and Edwards, Climate Informatics, Earthzine, 2014

At Michigan

- UV-CDAT
  - We have several students using UV-CDAT, thanks largely to Jerry Potter.
  - They love it
  - What to make it better?
    - Stable, well documented releases, that are thoroughly tested.

Context 1

- Efforts to use climate data in adaptation planning and management
  - Students and colleagues in climate science
  - Students and scientists in, for example, ecosystem management, public health, etc.
  - Professional (urban planners, resource managers, etc.) end users
- Students in climate and information systems
  - Obtain data and apply data
  - Document barriers to data usability (self and through interviews)
  - Suggest technical and design improvements
  - Specifically evaluated ESGF and climate.data.gov

Context 2

- Climate Data and Knowledge Usability
  - Data is posed by scientists to be broadly useful
  - Practice shows that much of the data is not usable outside of a close group of scientists
  - Gap between useful and usable
- Statements like
  - 60%-80% time of project “getting data”
  - Colleague just estimated 2 years for acquiring and massaging data for Ph.D. problem
  - Many don’t actually get data and don’t want data → they want guidance placed in context of their problem.

Which Data?

- With our clients, most requested data sets
  - Station data from NCDC (National Climatic Data Center)
  - Model projections from Coupled Model Intercomparison Project (CMIP)
    - Tailored datasets derived from CMIP (this includes downscaling)
  - Local data that represent region or application
    - Lake ice, stream flow, high resolution temperature,
    - Census data, built environment, …
  - Reanalysis data and satellite data

Overarching comments

- Data availability is not a fundamental barrier; there is an abundance of available data
- However, getting the data you want is often a challenge
  - Technological
  - Information systems
  - What data to choose
- Scientists can, with effort, usually wrestle their way to data use – they know the "language."
- The situation is always changing → improving
  - The situation is far better than a decade ago

A summary of “findings” → perhaps “requirements”

- These are needed to improve usability
  - These are extracted from a body of experience
  - Some are specifically needed in ESGF and UV-CDAT

Essential, mostly missing

- Glossary
  - What do variables mean?
  - What do file names mean?
  - Jargon of field, technology,
- “Documentation”
  - User guides, quick-reference guides, example problems
  - Source, provenance → metadata
  - Strengths and weaknesses, appropriate application, experience → expert guidance

Discovery and Navigation

- Search
  - Know what you want
  - Don’t know what you want
    - How to discriminate between too many datasets
  - Controlled vocabulary
- Consistency of navigation
  - Land at a page where navigation is different
  - Lose your place, path on sites
- Get someplace and
  - Nothing there
  - What is there not what is expected

Tools and Data

- Format
  - Internal to climate (or weather) community
  - Inconsistent across individual datasets data collections, esp. campaign, research data
- Tools
  - Inconsistency go from one page to another, one site to another, and same “tool” is different
  - Many mapping tools offered, but how does one differ from another?
- Variable needed for application
  - Indices
  - Derived variables
- Basic quality control
  - Missing data
  - Wrong data

Some general recommendations

- Independent tests of usability of the data
- Focus on serving a particular community, application
  - Know your user
    - Integrate users tools rather than provide your tools to the user
  - Watch for mission creep
    - Recognize specific needs for discipline, non-discipline experts, professionals, education, public
  - Assume data to be further evaluated, tailored
- Frame data center and service as part of chain that makes end-to-end system → part of a broader community where the community can scale and extend services
  - Publish requirements, architecture, schedule, technical documentation
  - Participation in communities developing standards
  - Use of emerging services that are based on standards

End of presentation

Loading Dock Model of Data Provision

Services

- Challenges of large data sets
  - Distributed services
  - Network
  - Analysis tools
  - Visualization
    - Expert
    - Non-expert

Characteristics of usability

- For data or dataset to be usable requires:
  - Legitimacy is an attribute of objectivity, fairness, and a lack of political bias.
  - Credibility is an attribute of scientific adequacy.
  - Salience requires that information be relevant to the problem to be addressed.
- Even if a data service is easy to use, does it provide information that supports, legitimacy, credibility, and salience?

During course

- Students were asked to document barriers to data access and data use
- Specifically asked to review
  - Earth System Grid Federation (ESGF) interface to CMIP data
    - Livermore node
      - http://pcmdi9.llnl.gov/esgf-web-fe/
  - Climate.data.gov
    - https://www.data.gov/climate/

Places where we have most success getting gridded data

- For observational station data
  - National Climatic Data Center (NCDC)
- For gridded data
  - "Bureau of Reclamation"
    - http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/dcpInterface.html
  - USGS GeoDataPortal
    - http://cida.usgs.gov/gdp/
  - Reanalysis.org
    - NOAA / ESRL / PSD
    - NASA / GIOVANNI
  - Climate Wizard (Nature Conservancy)

Extend Conclusions of Rood and Edwards

- Rood and Edwards, Climate Informatics: Human Experts and the End-to-End System, Earthzine, 2014
  - Improving the usability of data systems and data services

Human experts

- Human experts are an integral part of the information system. Rather than design the human out of the information system, effort should be focused on collecting the needed human expertise and improving the efficiency of the human expert.

Evaluation

- An important part of the climate information system is the need to evaluate the suitability of data and knowledge for a particular application. Therefore, information system design needs to facilitate the evaluation step. The unmet need for evaluation stands as a barrier to delivering the most appropriate and readily usable data for particular purposes.

Audience and communication

- Information systems need to be designed with more attention and focus on classes of users with similar needs. Since the challenges of usability are more about communication than technology, data systems need to focus on effective communication.

End-to-end system

- If climate data are to be made a commodity, far more attention needs to be placed on the connective tissue of usable end-to-end data systems, such as training, example problems, glossaries, walkthroughs, and interface design. Resources to build, maintain and evolve data systems are required.

Demo November 13, 2014

- climate.data.gov
- Things are always changing, so this demo may not work.
- Task: get temperature data in my location for my problem … discovery

A Demo of Sorts (20141113)

climat.data.gov

- Google “climate.data.gov”, choose
  - https://www.data.gov/climate (note confusion of climate.data.gov and www.data.gov/climate, which some would note as “trust” or “credibility” issue)
- Click on “Data,” in Nav Bar below “Climate” (this brings up page titled “Search for a Dataset,” also tells you already that 398 datasets are found, Note diverse names of these datasets, imagine you are practitioner with specific problem in mind, how you you find? How do you choose?)
- Go to search window (“Search datasets…”), enter “temperature” (note 86 datasets found, note the word “temperature” is present in only two of the 20 resources on first page of this search (think, I searched “temperature” and I got a water level observations dataset, what is a water level dataset?, and paleoclimatology software tools, and …), but on the first page is “Climate Data Online”)
- Click “Climate Data Online” (→ next page)

A Demo of Sorts

- Click “Climate Data Online” (note you are in the catalog.data.gov, I think all resources on previous page take you to this catalog and offer a similar formated interface (good!))
- Click Data Dictionary “visit page” (you tell me as a city planner what this means, perhaps as a member of an AMS data committee, what does this mean?)
- Make your way back to Climate Data Online, probably in a tab of your browser, click Web Page “visit page” (you are now at a place that says “Climate Data Online Search”)
  - (→ next page)

A Demo of Sorts

- You are now at “Climate Data Online Search”
  - Select pull down for Weather Observation Type/Dataset,(none say temperature (your original search), some say precipitation, choose “monthly summaries”)
  - Select Data Range (leave default, assuming it is a valid entry)
  - Select Pull Down in “Search For” (select climate divisions (though not a widely known term))
    - what’s a hydrological accounting unit?, note also for “station” in “Search Guide” enter WBAN, GHCND, FAA, etc. (Meaning? I don’t know most of these)
  - Now Enter a Search Term (identifier, what’s that?, I know a location, so I enter Detroit)
  - Press search
- Result “**Sorry, but no items match the search as specified”** (Detroit is in a Climate Division, but it’s not a climate division, it does not tell me this).
- → next page

A Demo of Sorts

- I'm curious, I repeat search replacing "Climate Division" with "Station" I get results, nice map with a lot of detroits, I'm easy I click "add" next to Detroit Metropolitan Airport, MI US, screen blinks, nothing in my data cart (It's free, I get what I pay for!), I select data tools (I get the opportunity to "search across multiple datasets." I select "mapping" tools, now I'm interested, I select "marine" I sit there a while, I have a pretty map with a red grid, I have no data, I zoom in, I can't get back, I ask my graduate student how he got the temperature data and he says, after two weeks I called the front desk at NCDC and they led me to Tom Karl, and someone who works for him led me to a web page that worked.)