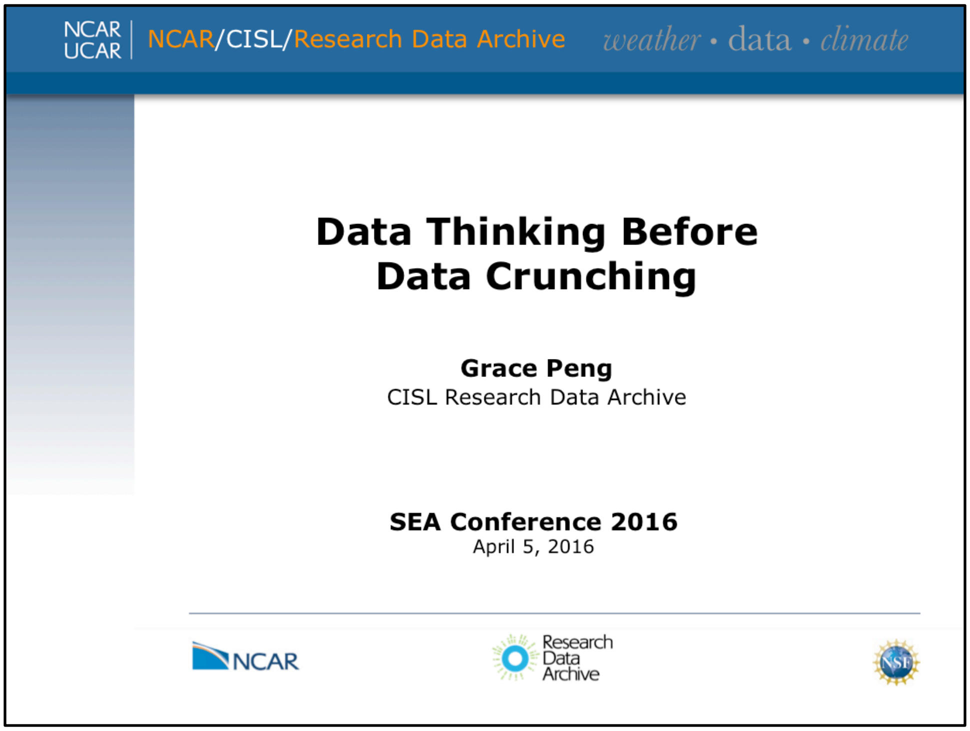


NCAR UCAR | NCAR/CISL/Research Data Archive *weather* • data • *climate*

Data Thinking Before Data Crunching

Grace Peng
CISL Research Data Archive

SEA Conference 2016
April 5, 2016

NCAR
Research Data Archive

I'm Grace Peng, one of the data specialists at the NCAR Research Data Archive aka as the RDA.

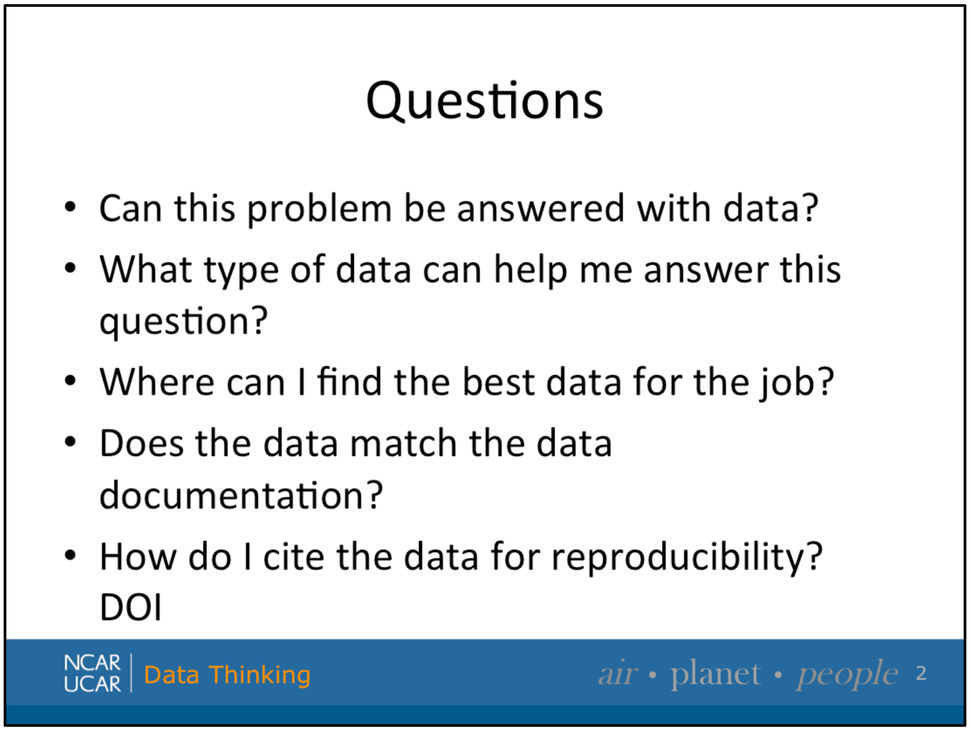
Many within NCAR know us as the people who stage data on /glade for you. Some people in this room may have never heard of us. For more than 40 years, the RDA has been collecting and disseminating weather and climate data to the research community. We host a growing collection of over 600 datasets.

Data specialists are part data curators, part data engineers, part software developers, and part subject matter experts. Among our data specialists, we have experts in meteorology, oceanography, mathematics, physics, chemistry and engineering.

We're witnessing a large increase in the use of wx and climate data in the commercial, government and educational sector. Our growing user base includes more non-experts so we find ourselves becoming educators about wx and climate data as well.

Data remains alive as long as people are using it. Working with students helps us learn what we need to do to today's data to make it useful for future users.

By the people come to me with their questions, they have usually begun their



Questions

- Can this problem be answered with data?
- What type of data can help me answer this question?
- Where can I find the best data for the job?
- Does the data match the data documentation?
- How do I cite the data for reproducibility? DOI

NCAR UCAR | Data Thinking

air • planet • people 2

I'll try to get through these five fundamental questions in 25 minutes. If I don't, please come to the Better than Free workshop tomorrow.

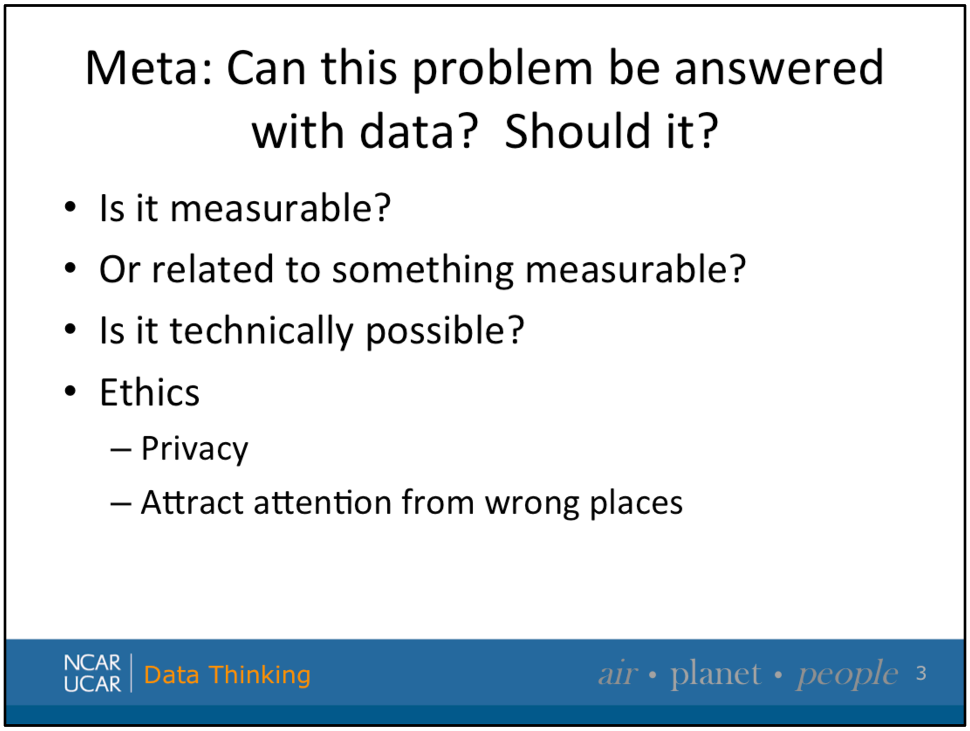
The first Q, Can this… sounds obvious, but it's actually difficult. This is the part where you make sure you have a science project and not an Op-Ed piece.

The second (and first) Q together helps you formulate your research plan.

Data discovery, finding the best data for your project, is a difficult technical problem. Data search engines founded on metadata can help you.

If the data doesn't match what you expect (from the documentation), you could be experiencing technical difficulties reading the data in correctly, or you may have discovered an error in the data.

Science should be reproducible. You've heard other talks this morning about data curation and digital object identifiers for data sets. I'll only need to cover the mechanics of how to cite data so that others understand what you did.



Meta: Can this problem be answered with data? Should it?

- Is it measurable?
- Or related to something measurable?
- Is it technically possible?
- Ethics
  - Privacy
  - Attract attention from wrong places

NCAR UCAR | Data Thinking air • planet • people 3

Can this problem be answered with data? is a meta question that you need to think through before starting on any endeavor.

Is it measurable?

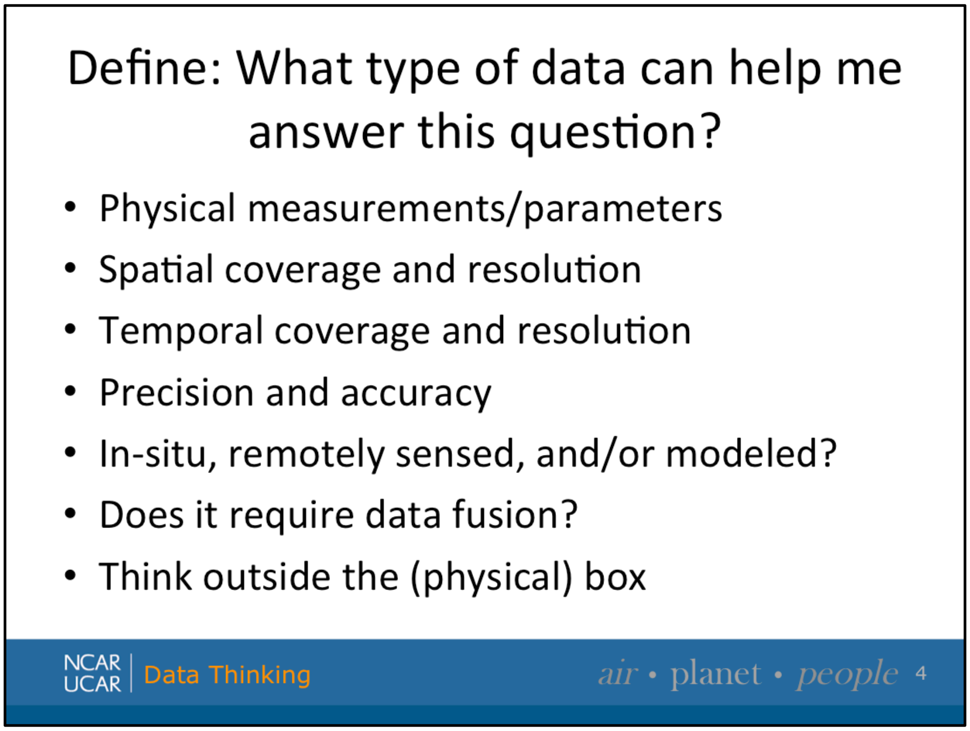
If not directly measureable, is it related to something measurable?
Has the data been collected already? May have to wait until data exists.

Should this problem be tackled?

Don't collect sensitive information if you can't safeguard it.

We think of weather as benign data. But, as weather apps turn people into mobile weather stations, what are the ethical and safety considerations? If we know where someone lives (their night-time closest cell tower), and also where they work or go to school during the day, we can narrow it down to a few people. If that information ends up in the hands of advertisers, it is annoying and intrusive. If it ends up in the hands of a violent ex or a repressive government, it could be deadly.

In my own work, I help researchers obtain data for pollution studies in countries where questioning the environmental costs of development has landed people in jail. What if their government demands I hand over their data request records?



Define: What type of data can help me answer this question?

- Physical measurements/parameters
- Spatial coverage and resolution
- Temporal coverage and resolution
- Precision and accuracy
- In-situ, remotely sensed, and/or modeled?
- Does it require data fusion?
- Think outside the (physical) box

NCAR UCAR | Data Thinking air • planet • people 4

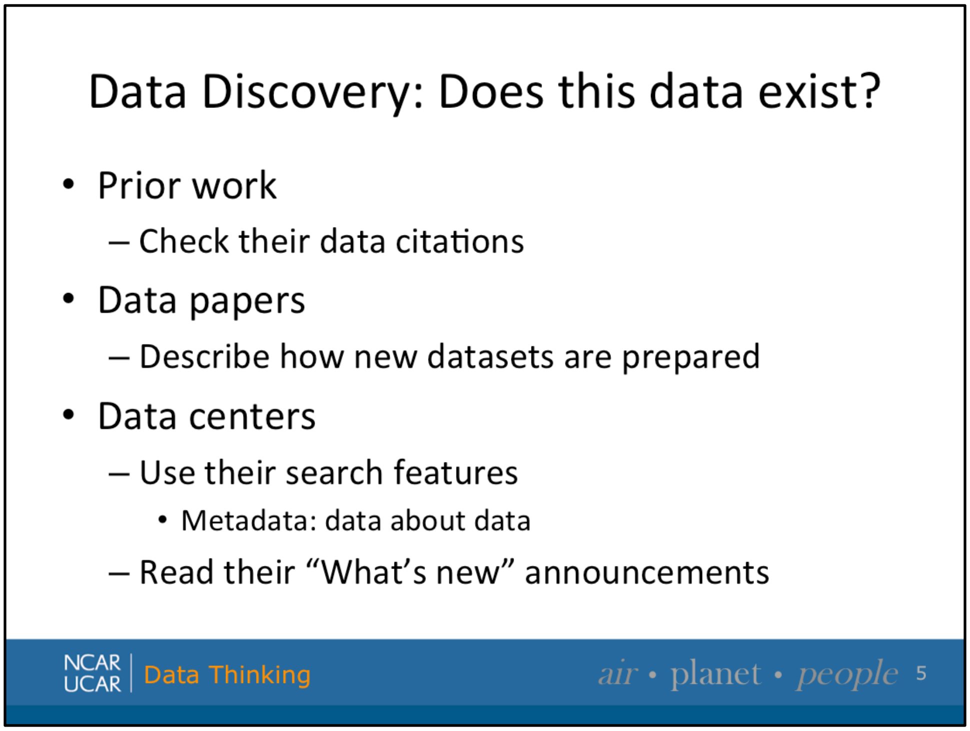
Data availability informs our approaches to research problems.

What kind of spatio-temporal coverage and resolution do you need?
Precision and accuracy (P&A) of measurements?

We all want the “best” data, with the most P&A. But, work through your error budget to find the minimum acceptable P&A. This will open up more data possibilities for your research.

Does it require data fusion; do you need to combine data from multiple sources? This is usually the case or else it wouldn’t be research.

New data enables new inquiries. But, do not overlook the value of looking at old data in new ways. My favorite example is the paper that examined the prevalence of cirrus clouds in the SLC area before and after it became a Delta Airlines hub airport. They correlated jet fuel tax receipts and cirrus prevalence and made a very convincing case that jet contrails are the cause of the increasing prevalence of cirrus clouds near SLC.



Data Discovery: Does this data exist?

- Prior work
  - Check their data citations
- Data papers
  - Describe how new datasets are prepared
- Data centers
  - Use their search features
    - Metadata: data about data
  - Read their "What's new" announcements

NCAR UCAR | Data Thinking air • planet • people 5

The practice of Data discovery is still in the juvenile phase. There isn't one killer app or dominant search engine that can search all data repositories yet.

This is a mixed bag with upsides. Consider the time spent in data discovery a learning opportunity.

Read research related to your project. What kinds of data do they use and where did they obtain it?

Data can be cited directly, as an entity in its own right, often with a DOI. This is the modern, preferred method.

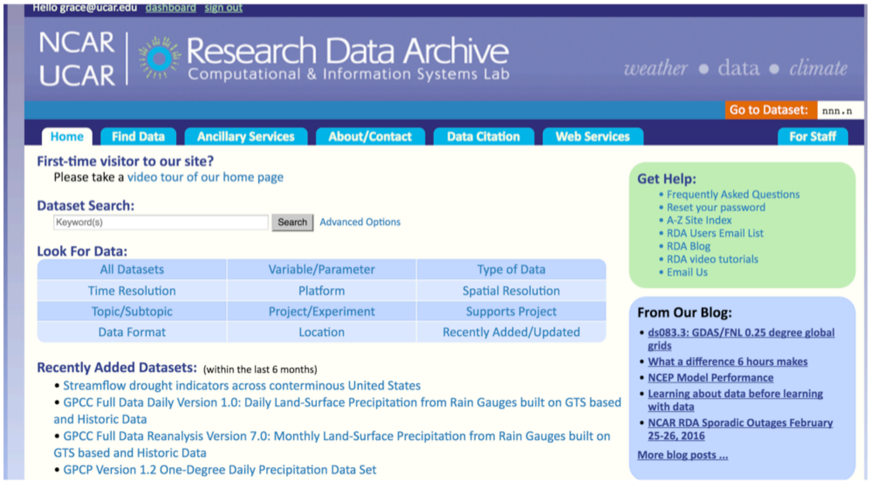
Aside:
[DOI stands for Digital Object Identifiers are one to one mappings of data sets (DS) to DOI number (unlike DS names). Some places give new DOIs to new versions of a DS, while others give new DOIs for major new versions of DSs.

However, the practice of citing data papers, papers describing the preparation of a data set, is also in widespread use and has been used for much longer.

Before the advent of data papers, people would cite the first science paper that used

RDA Search Matrix

rda.ucar.edu



First-time visitor to our site?
Please take a video tour of our home page

Dataset Search:

Look For Data:

All Datasets	Variable/Parameter	Type of Data
Time Resolution	Platform	Spatial Resolution
Topic/Subtopic	Project/Experiment	Supports Project
Data Format	Location	Recently Added/Updated

Recently Added Datasets: (within the last 6 months)

- Streamflow drought indicators across conterminous United States
- GPCC Full Data Daily Version 1.0: Daily Land-Surface Precipitation from Rain Gauges built on GTS based and Historic Data
- GPCC Full Data Reanalysis Version 7.0: Monthly Land-Surface Precipitation from Rain Gauges built on GTS based and Historic Data
- GPCP Version 1.2 One-Degree Daily Precipitation Data Set

Get Help:

- Frequently Asked Questions
- Reset your password
- A-Z Site Index
- RDA Users Email List
- RDA Blog
- RDA video tutorials
- Email Us

From Our Blog:

- ds083.3: GDAS/FNL 0.25 degree global grids
- What a difference 6 hours makes
- NCEP Model Performance
- Learning about data before learning with data
- NCAR RDA Sporadic Outages February 25-26, 2016

More blog posts ...

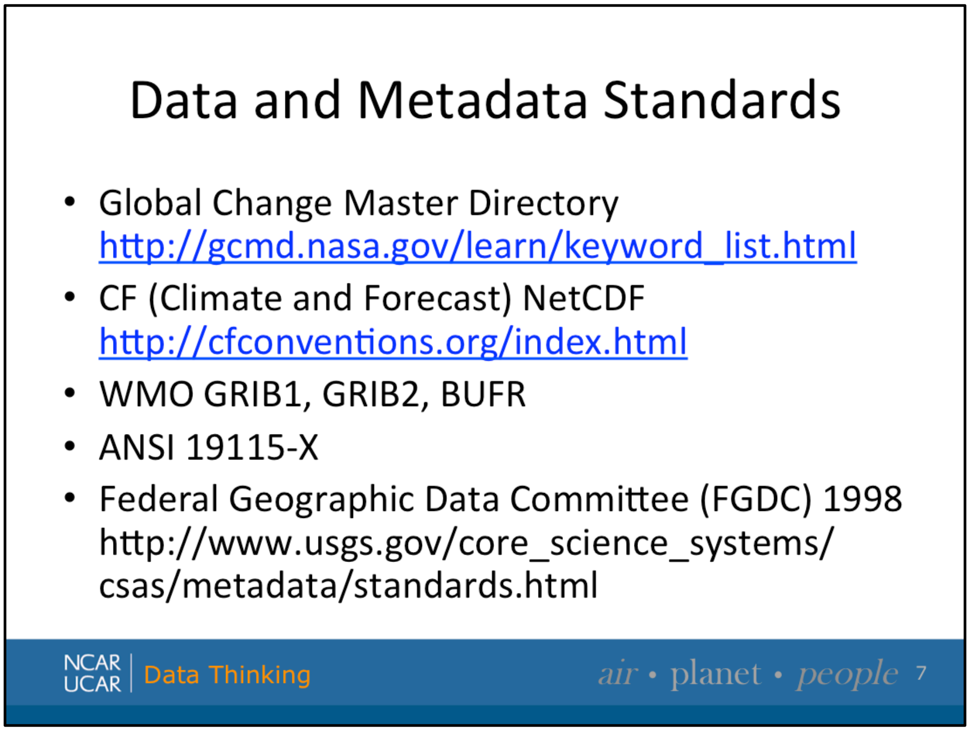
NCAR UCAR | Data Thinking air • planet • people 6

As an example, take a look at the RDA front door, rda.ucar.edu

You can do a free text search for data at the top.
Or, you can perform a faceted search, successively narrowing selections based upon metadata criteria.
These are just a few of the many options enabled at the RDA.

Check out our new datasets and other announcements on our blog.

Take a video tour of our home page to learn how to use the myriad features of our site.



Data and Metadata Standards

- Global Change Master Directory http://gcmd.nasa.gov/learn/keyword_list.html
- CF (Climate and Forecast) NetCDF http://cfconventions.org/index.html
- WMO GRIB1, GRIB2, BUFR
- ANSI 19115-X
- Federal Geographic Data Committee (FGDC) 1998 http://www.usgs.gov/core_science_systems/csas/metadata/standards.html

NCAR UCAR | Data Thinking air • planet • people 7

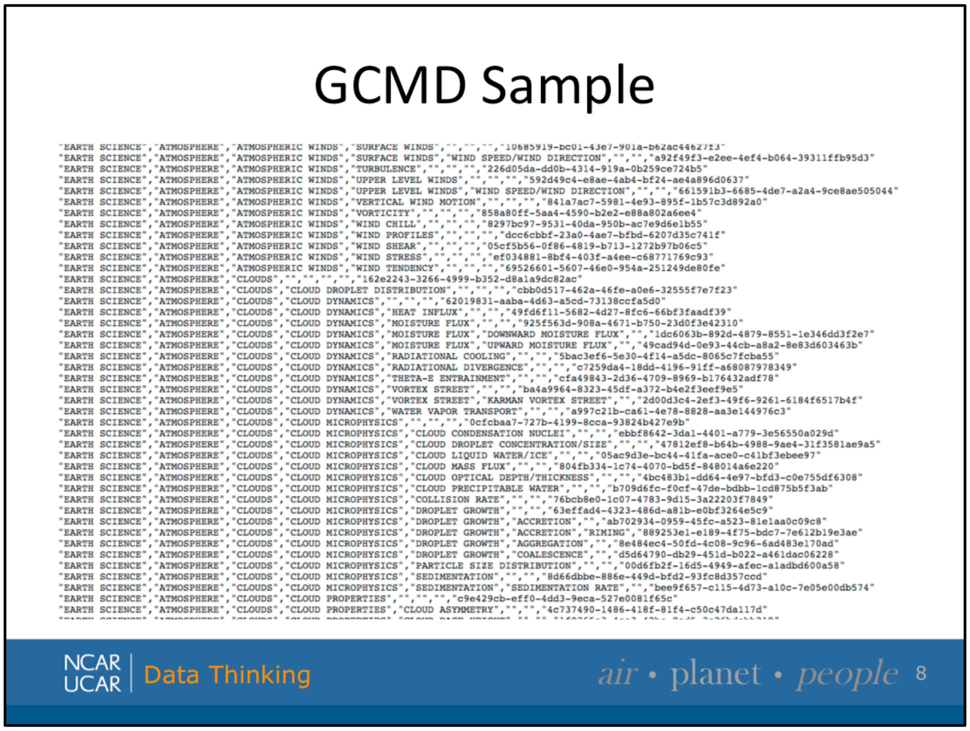
The RDA's search features are powered by standardized metadata, most commonly Global Change Master Directory keywords.

NetCDF is closely associated with CF standards and NCAR. It is self-describing and self-contained.

There are several other standards in wide use. Some older standards such as FGDC may be superceded by newer ones.

However, some old standards such as WMO and GRIB remain in daily use worldwide. GRIB and BUFR are self-describing, but not self-contained. They require external table to unpack the binary data correctly. If you apply the wrong table, you read the values incorrectly. GRIB and BUFR are widely used in operational transmission, but must be careful if used later. If you are unsure about the correct GRIB table to apply, ASK!

A data file can be mapped to more than one data standard for interoperability. In fact, older but still useful data sets should be mapped to new metadata standards to maintain their discoverability by both humans and machines.



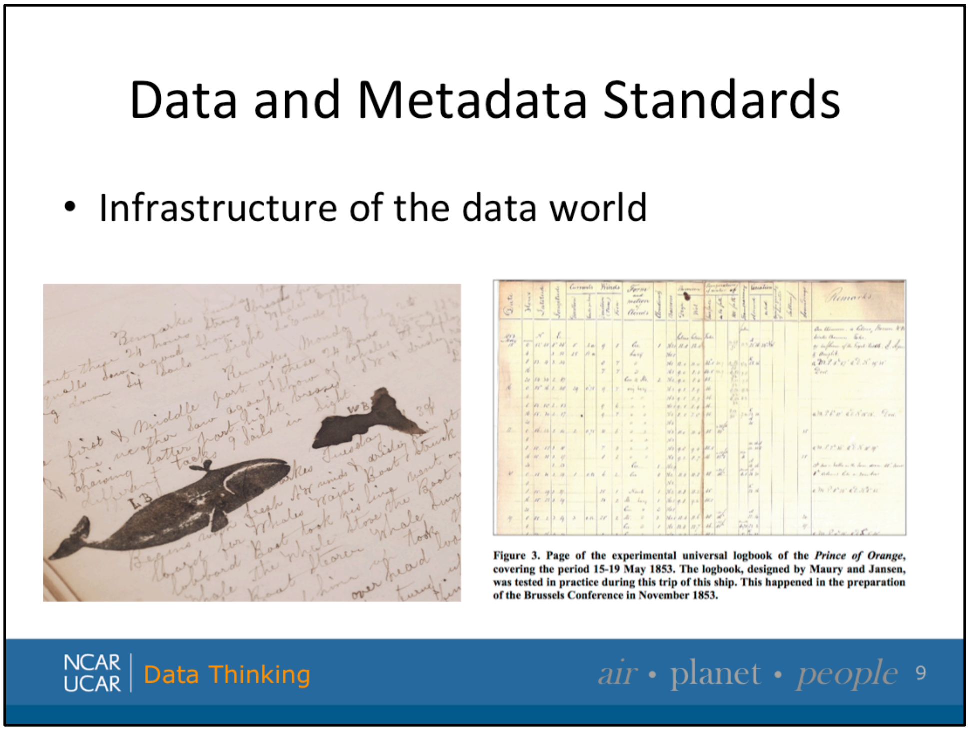
GCMD Sample

```
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","SURFACE WINDS","","","","10685919-bc01-43e7-901a-b62ac44627f3"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","SURFACE WINDS","WIND SPEED/WIND DIRECTION","","","a92f49f3-e2ee-4ef4-b064-39311ffb95d3"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","TURBULENCE","","","","226d05da-dd0b-4314-919a-0b259ce724b5"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","UPPER LEVEL WINDS","","","","592d49c4-e8ae-4ab4-bf24-ae4a896d0637"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","UPPER LEVEL WINDS","WIND SPEED/WIND DIRECTION","","","661591b3-6685-4de7-a2a4-9ce8ae505044"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","VERTICAL WIND MOTION","","","","841a7ac7-5981-4e93-895f-1b57c3d892a0"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","VORTICITY","","","","858a80ff-5aa4-4590-b2e2-e88a802a6ee4"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","WIND CHILL","","","","8297bc97-9531-40da-950b-ac7e9d6e1b55"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","WIND PROFILES","","","","dcc6cbbf-23a0-4ae7-bfbd-6207d35c741f"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","WIND SHEAR","","","","05cf5b56-0f86-4819-b713-1272b97b06c5"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","WIND STRESS","","","","ef034881-8bf4-403f-a4ee-c68771769c93"
"EARTH SCIENCE","ATMOSPHERE","ATMOSPHERIC WINDS","WIND TENDENCY","","","","69526601-5607-46e0-954a-251249de80fe"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","","","","","162e2243-3266-4999-b352-d8a1a9dc82ac"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DROPLET DISTRIBUTION","","","","cbb0d517-462a-46fe-a0e6-32555f7e7f23"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","","","","62019831-aaba-4d63-a5cd-73138ccfa5d0"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","HEAT INFLUX","","","49fd6f11-5682-4d27-8fc6-66bf3faadf39"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","MOISTURE FLUX","","","925f563d-908a-4671-b750-23d0f3e42310"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","MOISTURE FLUX","DOWNWARD MOISTURE FLUX","","1dc6063b-892d-4879-8551-1e346dd3f2e7"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","MOISTURE FLUX","UPWARD MOISTURE FLUX","","49cad94d-0e93-44cb-a8a2-8e83d603463b"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","RADIATIONAL COOLING","","","5bac3ef6-5e30-4f14-a5dc-8065c7fcba55"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","RADIATIONAL DIVERGENCE","","","c7259da4-18dd-4196-91ff-a68087978349"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","THETA-E ENTRAINMENT","","","cfa49843-2d36-4709-8969-b176432adf78"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","VORTEX STREET","","","ba4a9964-8323-45df-a372-b4e2f3eef9e5"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","VORTEX STREET","KARMAN VORTEX STREET","","2d00d3c4-2ef3-49f6-9261-6184f6517b4f"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD DYNAMICS","WATER VAPOR TRANSPORT","","","a997c21b-ca61-4e78-8828-aa3e144976c3"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","","","","0cfcbaa7-727b-4199-8cca-93824b427e9b"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD CONDENSATION NUCLEI","","","ebbf8642-3da1-4401-a779-3e56550a029d"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD DROPLET CONCENTRATION/SIZE","","","47812ef8-b64b-4988-9ae4-31f3581ae9a5"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD LIQUID WATER/ICE","","","05ac9d3e-bc44-41fa-ace0-c41bf3ebee97"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD MASS FLUX","","","804fb334-1c74-4070-bd5f-848014a6e220"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD OPTICAL DEPTH/THICKNESS","","","4bc483b1-dd64-4e97-bfd3-c0e755df6308"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","CLOUD PRECIPITABLE WATER","","","b709d6fc-f0cf-47de-bdbb-1cd875b5f3ab"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","COLLISION RATE","","","76bcb8e0-1c07-4783-9d15-3a22203f7849"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","DROPLET GROWTH","","","63effad4-4323-486d-a81b-e0bf3264e5c9"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","DROPLET GROWTH","ACCRETION","","ab702934-0959-45fc-a523-81e1aa0c09c8"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","DROPLET GROWTH","ACCRETION","RIMING","889253e1-e189-4f75-bdc7-7e612b19e3ae"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","DROPLET GROWTH","AGGREGATION","","8e484ec4-50fd-4c08-9c96-6ad483e170ad"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","DROPLET GROWTH","COALESCENCE","","d5d64790-db29-451d-b022-a461dac06228"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","PARTICLE SIZE DISTRIBUTION","","","00d6fb2f-16d5-4949-afec-a1adbd600a58"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","SEDIMENTATION","","","8d66dbbe-886e-449d-bfd2-93fc8d357ccd"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD MICROPHYSICS","SEDIMENTATION","SEDIMENTATION RATE","","bee9f657-c115-4d73-a10c-7e05e00db574"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD PROPERTIES","","","","c9e429cb-eff0-4dd3-9eca-527e0081f65c"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD PROPERTIES","CLOUD ASYMMETRY","","","4c737490-1486-418f-81f4-c50c47da117d"
"EARTH SCIENCE","ATMOSPHERE","CLOUDS","CLOUD PROPERTIES","CLOUD BASE HEIGHT","",""
```

NCAR UCAR | Data Thinking *air* • planet • *people* 8

GCMD keywords map out explicitly the contents of data files. This should eliminate guesswork, which can lead to misinterpreting data.
This looks excessively detailed, but this is how machine-to-machine communication works.
This also enables distributed search, where one portal can crawl other data sites for you.

Data producers can obtain help from data curators/metadata specialists to map out their data.



Data and Metadata Standards

- Infrastructure of the data world

Figure 3. Page of the experimental universal logbook of the *Prince of Orange*, covering the period 15-19 May 1853. The logbook, designed by Maury and Jansen, was tested in practice during this trip of this ship. This happened in the preparation of the Brussels Conference in November 1853.

NCAR UCAR | Data Thinking

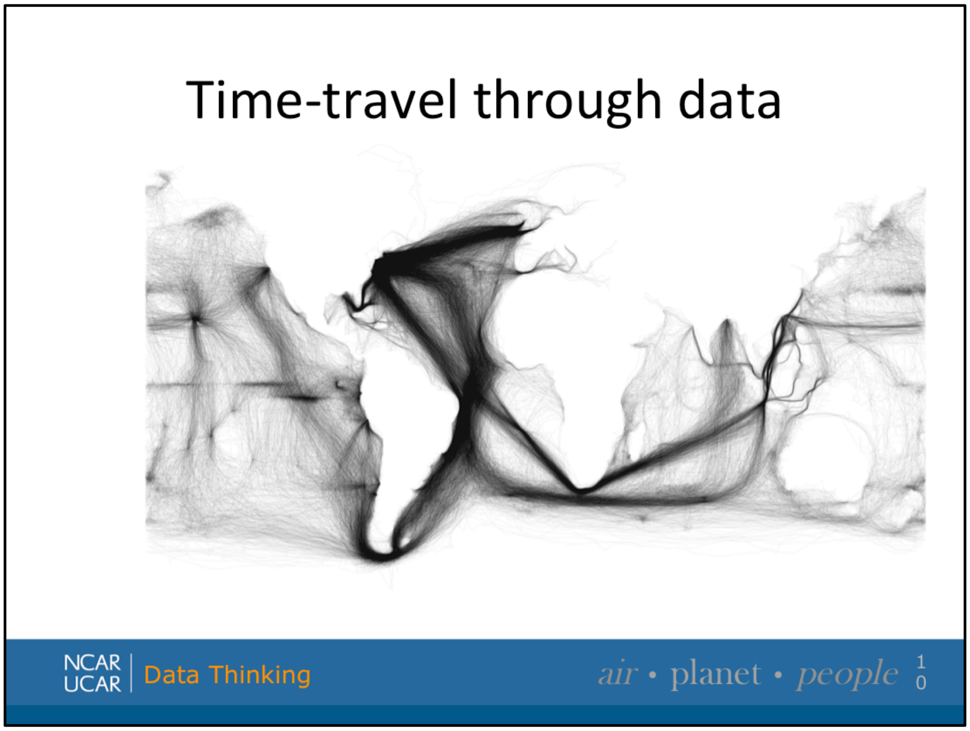
air • planet • people 9

Standardization of data is not a new idea. Once, whaling ship logbooks looked like the one at left. In 1853, Maury and Jansen created and tested an universal logbook format that is similar to what's used today. Then they organized an international conference to promote the use of their universal logbook format.

Seafarers could still maintain idiosyncratic logbooks like those at left, but they were required to also log in their information in the universal logbook with standardized units.

Bauman Rare Books http://sappingattention.blogspot.com/2012/11/reading-digital-sources-case-study-in.html

The Brussels Conference and its legacy ftp://ftp.wmo.int/Documents/PublicWeb/amp/mmop/documents/JCOMM-TR/J-TR-27-BRU150-Proceedings/DOCUMENTS_JCOMM_27/Session_4/4_2_Konnen.pdf



Time-travel through data

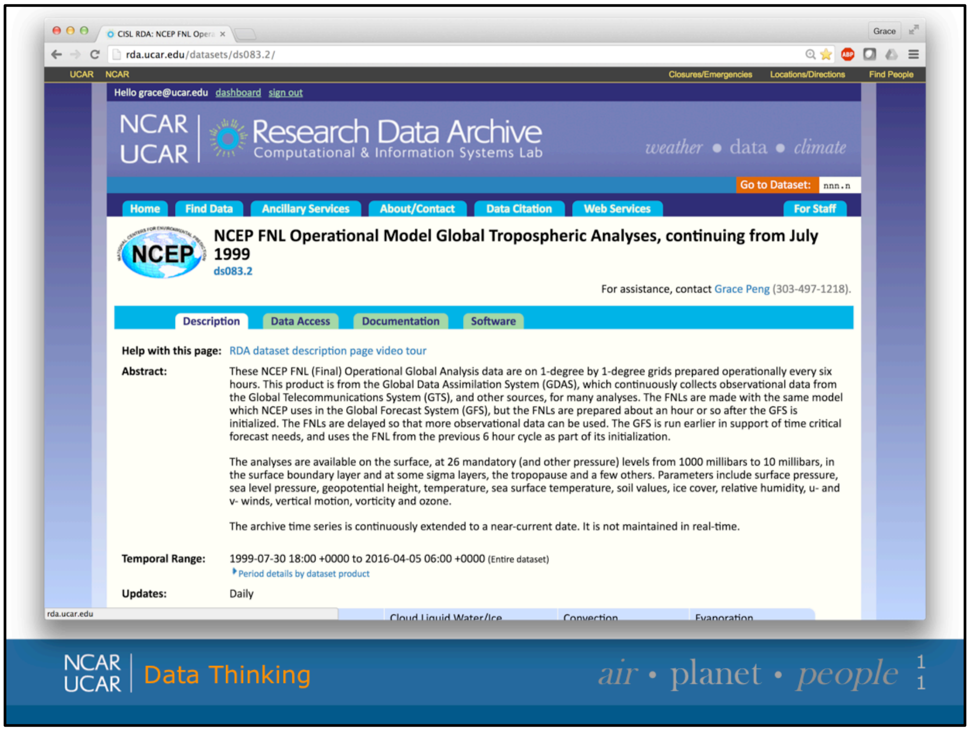
NCAR UCAR | Data Thinking

air • planet • people 10

This data standardization enables us to create maps like this.

We know global shipping routes and traffic over time from these logbooks. We agreed to common units of measurement (latitude and longitude, Greenwich time, Celsius scale, frequency of recorded measurements) http://sappingattention.blogspot.com/2012/11/reading-digital-sources-case-study-in.html

The data of the past is available to us today because of preservation efforts of data compilers and curators of yesterday. It's understandable to us only because they handed down not just the data, but the knowledge about the data in the form of metadata and adherence to well-described data conventions. These are known as data and metadata standards. The corollary is that our adherence to data standards today will enable your data of today to time travel and aid future scientists.



NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999

ds083.2

For assistance, contact Grace Peng (303-497-1218).

Help with this page: RDA dataset description page video tour

Abstract: These NCEP FNL (Final) Operational Global Analysis data are on 1-degree by 1-degree grids prepared operationally every six hours. This product is from the Global Data Assimilation System (GDAS), which continuously collects observational data from the Global Telecommunications System (GTS), and other sources, for many analyses. The FNLs are made with the same model which NCEP uses in the Global Forecast System (GFS), but the FNLs are prepared about an hour or so after the GFS is initialized. The FNLs are delayed so that more observational data can be used. The GFS is run earlier in support of time critical forecast needs, and uses the FNL from the previous 6 hour cycle as part of its initialization.

The analyses are available on the surface, at 26 mandatory (and other pressure) levels from 1000 millibars to 10 millibars, in the surface boundary layer and at some sigma layers, the tropopause and a few others. Parameters include surface pressure, sea level pressure, geopotential height, temperature, sea surface temperature, soil values, ice cover, relative humidity, u- and v- winds, vertical motion, vorticity and ozone.

The archive time series is continuously extended to a near-current date. It is not maintained in real-time.

Temporal Range: 1999-07-30 18:00 +0000 to 2016-04-05 06:00 +0000 (Entire dataset)

Period details by dataset product

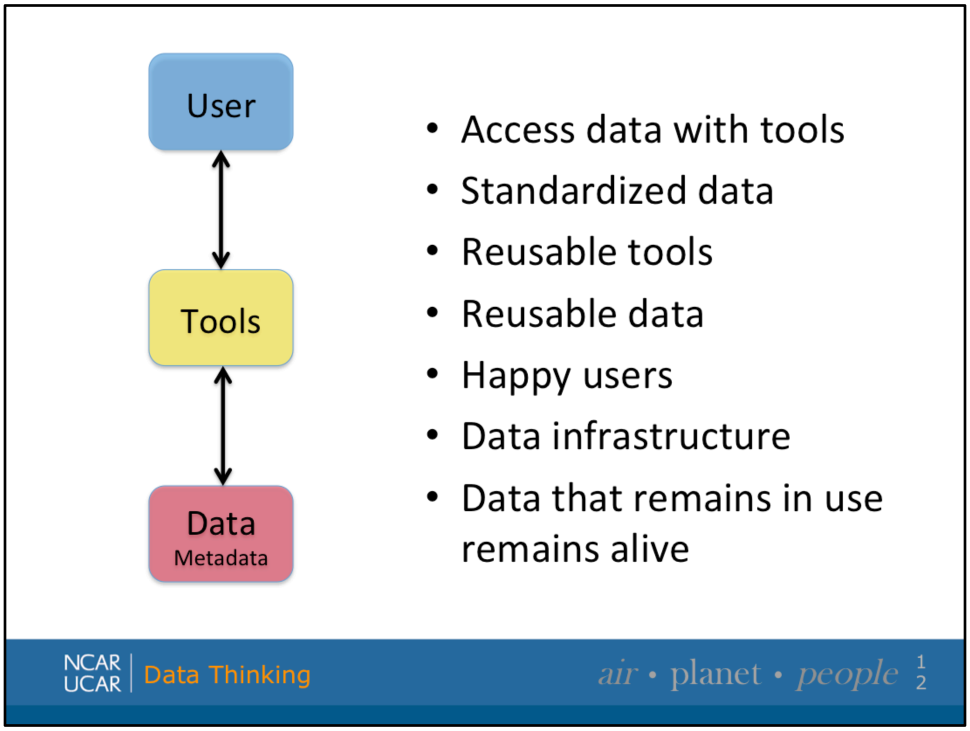
Updates: Daily

This is currently the RDA's most popular dataset; it is used to initialize Numerical Weather Prediction (NWP) models such as WRF.

If we archive a data set, we also archive documentation about how it was prepared and software to read it.
If the data set is archived in one of the major standards, such as NetCDF or GRIB, we don't need to say much more about it.

But, if it is in a more idiosyncratic format, we will provide specialized readers and other help.

Note that each data set is assigned to a data specialist whose contact information is at the top right. Data specialists serve as a conduit between data users and data producers.



User

Tools

Data
Metadata

- Access data with tools
- Standardized data
- Reusable tools
- Reusable data
- Happy users
- Data infrastructure
- Data that remains in use remains alive

NCAR UCAR | Data Thinking

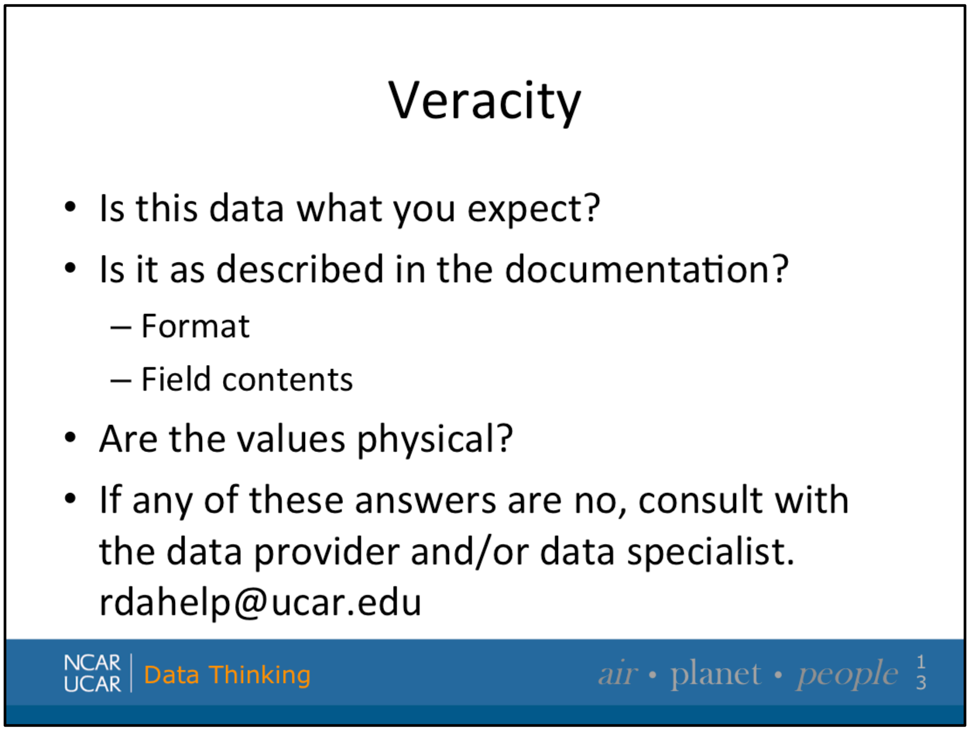
air • planet • people 12

Users access data through tools. One-off (ad-hoc) data requires one-off tools.

Data and metadata standardization allows us to create reusable tools.

This ecosystem of data standards, tools, search and delivery tools makes up our shared data infrastructure. When infrastructure works smoothly, the magic becomes mundane and invisible.

Users are not charged the full cost of developing or maintaining infrastructure. But please keep in mind that infrastructure is not cheap and delaying maintenance can cause catastrophic failures in the future.



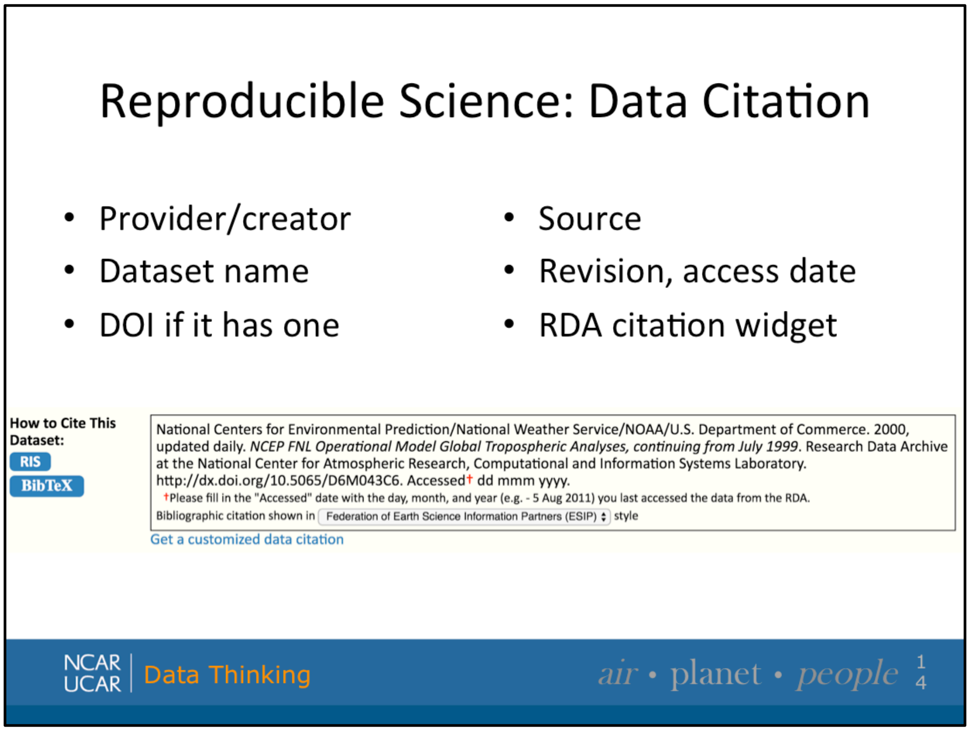
Veracity

- Is this data what you expect?
- Is it as described in the documentation?
  - Format
  - Field contents
- Are the values physical?
- If any of these answers are no, consult with the data provider and/or data specialist. rdahelp@ucar.edu

NCAR UCAR | Data Thinking *air* • planet • *people* 13

Speaking of magic. When you open up those data files, take some time to interrogate its contents. If you have time, explore beyond the values that you originally wanted to use to see if something else might open up new possibilities for your research.

If you don't get what you expect to see, either in the file contents or values that don't make physical sense, stop. Consult with the data provider. If you got the data from us, send a message to rdahelp@ucar.edu and cc it to the data specialist for that data set.



Reproducible Science: Data Citation

- Provider/creator
- Dataset name
- DOI if it has one
- Source
- Revision, access date
- RDA citation widget

How to Cite This Dataset:

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce. 2000, updated daily. *NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://dx.doi.org/10.5065/D6M043C6. Accessed† dd mmm yyyy.

†Please fill in the "Accessed" date with the day, month, and year (e.g. - 5 Aug 2011) you last accessed the data from the RDA.

Bibliographic citation shown in Federation of Earth Science Information Partners (ESIP) ⬍ style

Get a customized data citation

NCAR UCAR | Data Thinking

air • planet • people 14

For the sake of reproducible science, use sound data citation practices.   Different journals have different data citation styles. They share common required elements.

Who created the dataset?

How is it called?. Give the DOI if it has one.

Where did you get it? Data can be replicated in multiple places that serve different subsets of the data. Citing the data source is important.

Data is sometimes reprocessed or corrected. Put down the version number and data access date.

If you make a data request through the RDA web interface, the request is automatically logged in our database. Should anything happen, we can regenerate the data request for you or anyone else who wants to repeat your work. This can save you storage space as you need only save locally what you are currently using.

Our database records also allow us to make custom data citations for you with this widget found on all of our data set home pages. The widget generates citations in AMS, AGU and several other popular styles. You can also create RIS and BibTeX

Questions? rdahelp@ucar.edu

http://ncarrda.blogspot.com/p/recommended-data-reading.html

NCAR UCAR | Data Thinking air • planet • people 15

If you are interested in thinking meta about data, I put up a list of my favorite data books. We also maintain a blog where we make announcements, give tips and post tutorials.

Thank-you.