Data Access, Release and Distribution

Mark Schoeberl
NASA/GSFC

What do I know about this topic?

- UARS Project Scientists
  - Rolling ~ 3 year data protocol – science team only
- Aircraft mission PI (at least 4)
  - 9 mo to 1 year data protocol
- Aura – 2 phase protocol
  - Commissioning phase (~ 1 year)
  - Public release
- Data User of these and other data sets
  - Write my own codes

Views of Data: Two Extremes

"I believe to the last coil of my DNA that the experimenter should have as much time with his/her data as he/she thinks they need." – Dan Albriton, Head of NOAA Aeronomy Lab ~1988

Consequence: Vaults filled with data that have never been release because the experimenter went on to a new experiment or lost lock.

"EOS data will be released to the general public within 15 minutes of instrument turn on" – Shelby Tilford, NASA HQ Earth Sciences Director ~ 1991

What are the underlying fears?

Experimenter's viewpoint

- My reputation rests on producing good data – I need time to look at the data and make sure it is of high quality. Only I can say when the data is ready for release – it is my data!!
- If I release the data early, people will steal really exciting results and publish without my participation, while I am busy trying to improve the data.*
- I have spent years on this instrument and have

User's viewpoint

- The government funded this instrument/network – the data belongs to the community
- If you hold back the data you are impeding scientific progress
- You are using the "data is not good enough" statement to keep the data from the community …

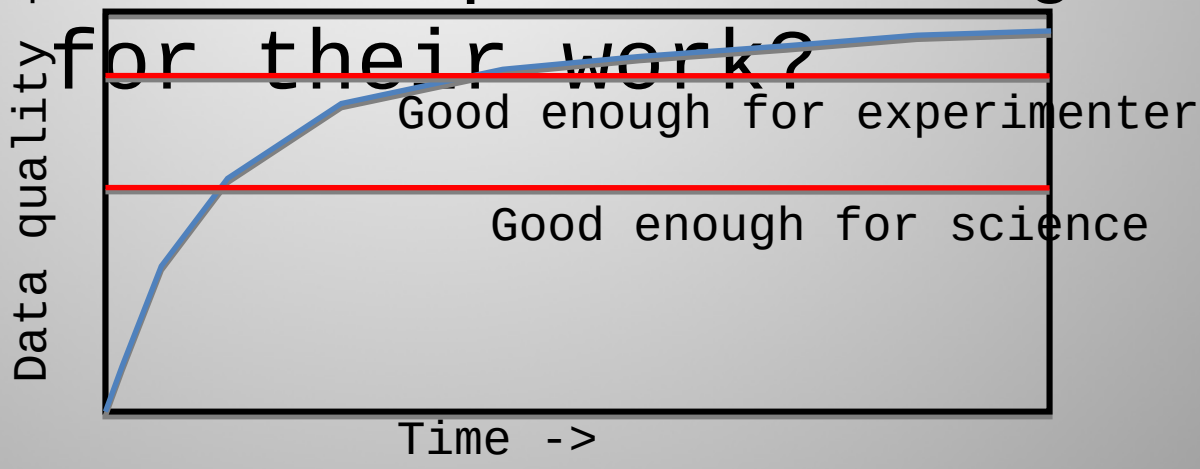
* Mostly an Urban Legend

What is the experience

- Giving the data to select users accelerate the improvement of the data. But, there needs to be enough of these users dedicated to looking at the data to find real problems.
- The experimenter has the inside track on skimming the cream from the data. Experimenters who don’t publish usually have a problem writing up there data anyway.
- There are lots of examples of people who have embarrassed themselves by not working with data provider.
  - Papers published in Science on ozone trends
- Users should recognize that the

The practical issues

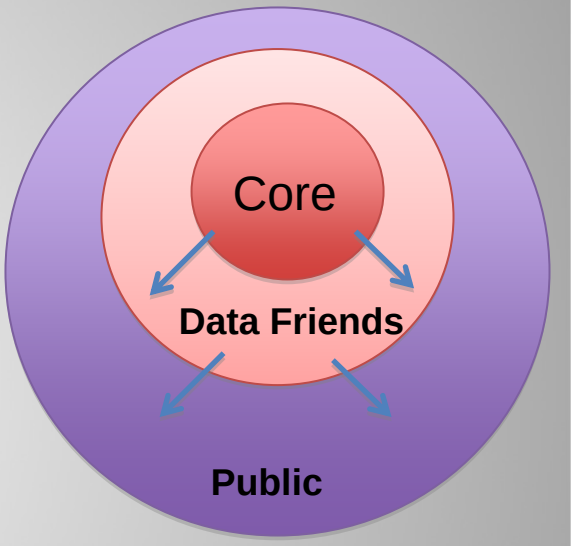
1. Data quality – When is the data good enough to release?
2. Data distribution – who should be using the early releases of the data?
3. How does the experimenter get credit for their work?



(1) When is the data good enough?

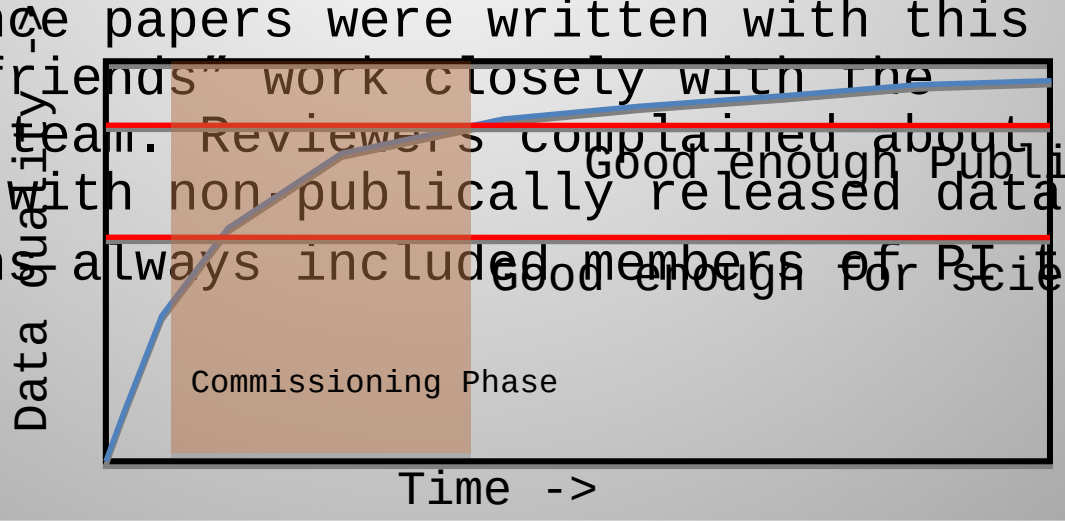
Phases of data quality

- Data is ready to be released to those providing core validation
  - Looks reasonable
  - Doesn’t violate physics or chemistry
  - Consistent with itself and internal calibration
- Data can be released to a wider circle of “data friends” – data is partially validated
  - Science papers are written
- Data is mostly validated and

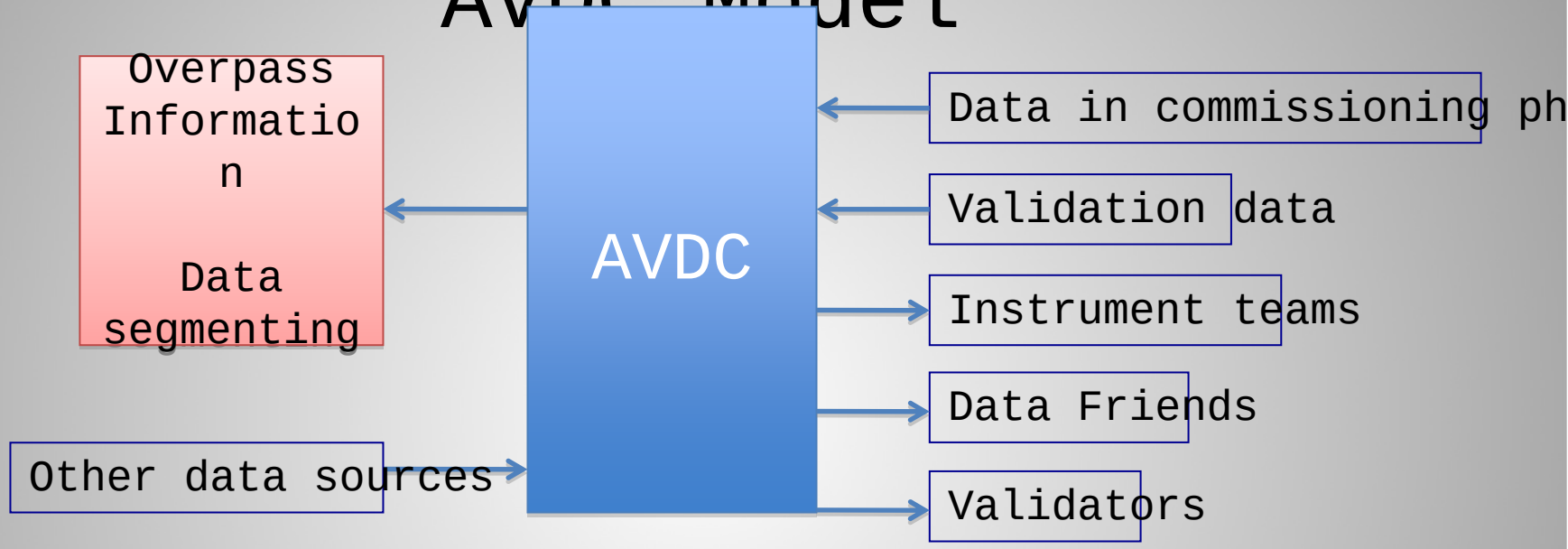


When is the data good enough – the Aura experience?

- To test satellite data algorithms, experimenters usually rely on data statistics and test retrieval points – but these aren't enough!
- Release of the data to "data friends" allowed more eyeballs on the data and quicker feedback on data problems
- With Aura the "data friends" were mostly validators and a few modelers
- Release to data friends is the "Commissioning Phase"
- Early science papers were written with this data. The "data friends" work closely with the experiment team. Reviewers complained about publishing with non-publically released data.
- Publications always included members of PI team



How do you keep the data from being released too soon: The AVDC Model



AVDC use required an “access agreement” that gave the instrument PI knowledge of who was using the data. This agreeement also gave rights to co-authorship of publications. Likewise those contributing validation data had rights of co-authorship when their data was used and had access to early releases of the satellite data (where validation was occurring). This model got us around the US

(2) When is the data ready for public release?

- Data has been validated to the extent that known problems and biases can be *clearly articulated* – the data does not have to be perfect.
- A data e-handbook is available describing the experiment data, the algorithm and including contacts for data issues and interpretation
  - A short version of this handbook is an IEEE publication (Aura, Aqua, Terra)
- Other useful things before public release
  - Provide a web site on the data
  - Make presentations on the data to encourage working with the PI team on publications and

Thoughts about the released data

- The goal is to get people to use your data!
- Metadata inclusions
  - Algorithm version and pointers documentation on changes from previous algorithms
  - Who produced the data and how to contact that person
  - Standard nomenclature “e.g. height vs altitude, N_2O vs nitrous oxide” so that data headers are machine readable
- Data
  - Standard structured files and file

Gate Keeping

Even though the data is “public” experimenters often unintentionally practice gate keeping.

- Servers hold old versions of the data – new versions are only available through the PI
- Data is in an unwieldy format
- Data provided has not been quality checked
- Documentation is incomplete or missing crucial information
- No browse tools

– Funding agencies do not want to hear complaints of “gate keeping”

– The real issues behind “gate keepers” need to be addressed – usually a resource issue

Data Types

- *Real time data* is data delivered within hours (e.g. weather data, broadcast satellite data)
  - Requires very stable algorithms
  - Requires dedicated production facilities including parallel data processing systems
- *Near-real time data* is delivered within days
  - May be operational but inherent latency in the data prevents real time delivery
  - May have a dedicated processing facility but no parallel system
- *Research data*
  - Best effort approach (no promises)
  - No dedicated production facility

Data Processing and Re-Processing

- I have never encountered a data set that wasn’t reprocessed at some point – if only once.
- Satellite data is often reprocessed many times
  - Instrument characteristics change
  - New spectroscopic information
  - Ideas on noise reduction or interfering gases included
  - Bugs in the algorithm (you can’t test everything)
- Institutions should assume that data will be reprocessed and plan enough

Giving and Getting Credit

- Experimenters often complain that they do not get enough credit
  - Aura and aircraft experiment policy
    - “If you use someone’s data in a paper you must include them as a co-author – that person can then decide if they want their name on the paper.”
    - Most people will not ask for their name to be removed and some people will actually read the paper giving

Summary

- Data Release and Distribution
  - Three stages: Core, Commissioning Phase, General release
  - Data friends and validators – get the data out to them as soon as possible
  - Commissioning phase should not be longer than a year – shorter if possible
  - Use common standards in the meta-data
  - Avoid gate keeping
- Be open about credit and publication – err on the side on inclusion