Data Products: Whither ELTs?

David Silva (NOAO)

Science with Giant Telescopes

2008 June

Initial thoughts

- Primary data product: refereed publication
- Secondary data products (illustrative hierarchy)
  - Well-documented raw data associated with appropriate calibration data and data processing software
  - Calibrated observations with complete information (origin, processing history, quality assessment)
  - Multi-parameter object catalogs
- Hypothesis: easy-to-find, easy-to-use, secondary data products maximize number of refereed papers per unit observation
  - See: Great Observatories science archives, 2MASS, SDSS
- Hypothesis: must capture raw data stream but much less useful

Production requirements

- Producing a good (useful) data product requires…
  - Expert knowledge of telescope/instrument system
  - Sufficient calibration data
  - Adequate, flexible data processing software
  - Comprehensive data quality control (assessment) process
  - Complete information capture (meta-data), from start to end
  - Timely processing and release

- Automatic production requires in addition…
  - Structured, documented observing process
  - Automatic data organization, post-execution
  - Sufficient compute hardware for timely reduction

Data processing complexity

- Imaging
- Long-slit spectroscopy (point source)
- Long-slit spectroscopy (extended source)
- Multi-object spectroscopy (fiber)
- Multi-object spectroscopy (slits)
- Integral-field spectroscopy

- Add: weather
- Add: wavelength-dependent atmospheric absorption, emission
- Add: time-variable PSF
- Add: AO
- Add: data volume (number of bytes per unit time)

Where Are We?

- Obviously, most observations can be processed into useful secondary data products
  - See ApJ, AJ, AA, MNRAS, etc…
- Most current general purpose, ground-based observatories have been unwilling or unable to spend enough money to produce and archive good secondary data products, independently of initial users
  - Exceptions exist, mostly imaging or single object spectroscopy
- Most users of these facilities think in terms of individual observing runs and projects, not long-term legacy/reusability of secondary data products
  - Observing process not necessarily well-structured for automatic data processing

Where Do We Want To Go?
Discussion points

- Is our goal to provide ELT time or produce secondary data products for possible reuse?

- Semi-ignore small, self-contained data sets (<= 2 nights)?
  - Secondary data products only guaranteed to be useful to that project team
  - Does this “throw away” 70%++ of ELT observations?

- Process all other datasets?
  - Find right division of labor
    - Project team, observatory, science/data center
  - How much centralized coordination is required?
  - Who is responsible for quality of final data product in archive?