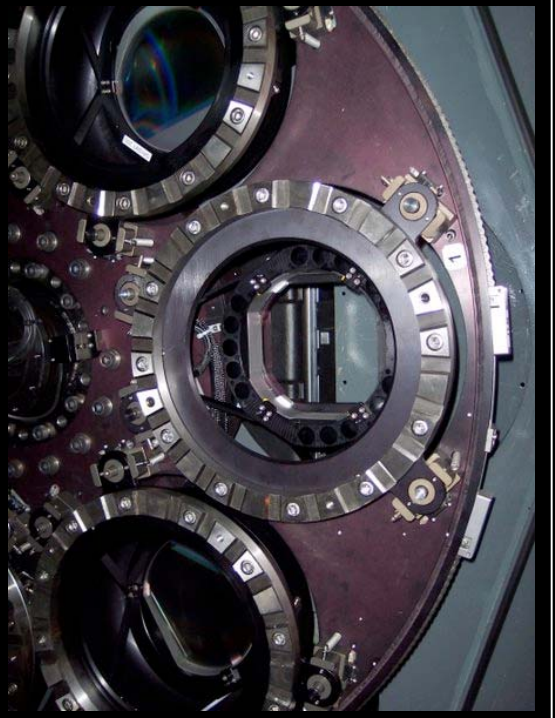
# Archiving PRIMUS at New York University:

## Design Methodology

Gretchen Gano
Brian Hoffman
Alex Tzanov



# Overview

- PRIMUS project context & library data curation team
- Data size and scope
- Core classes of content data
- Non-content data
- Design deliverables
- Design decisions
  - Evaluation criteria
- Concluding themes

# National/International developments

- NSF office of cyberinfrastructure
- Select reports
  - National Science Board report "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century " (2005)
    - Particuarly recommendations for roles of data managers
  - NSF Workshop on Challenges of Scientific Workflows (2006)
  - NSF Mellon workshop on interoperability of workflows (2007)

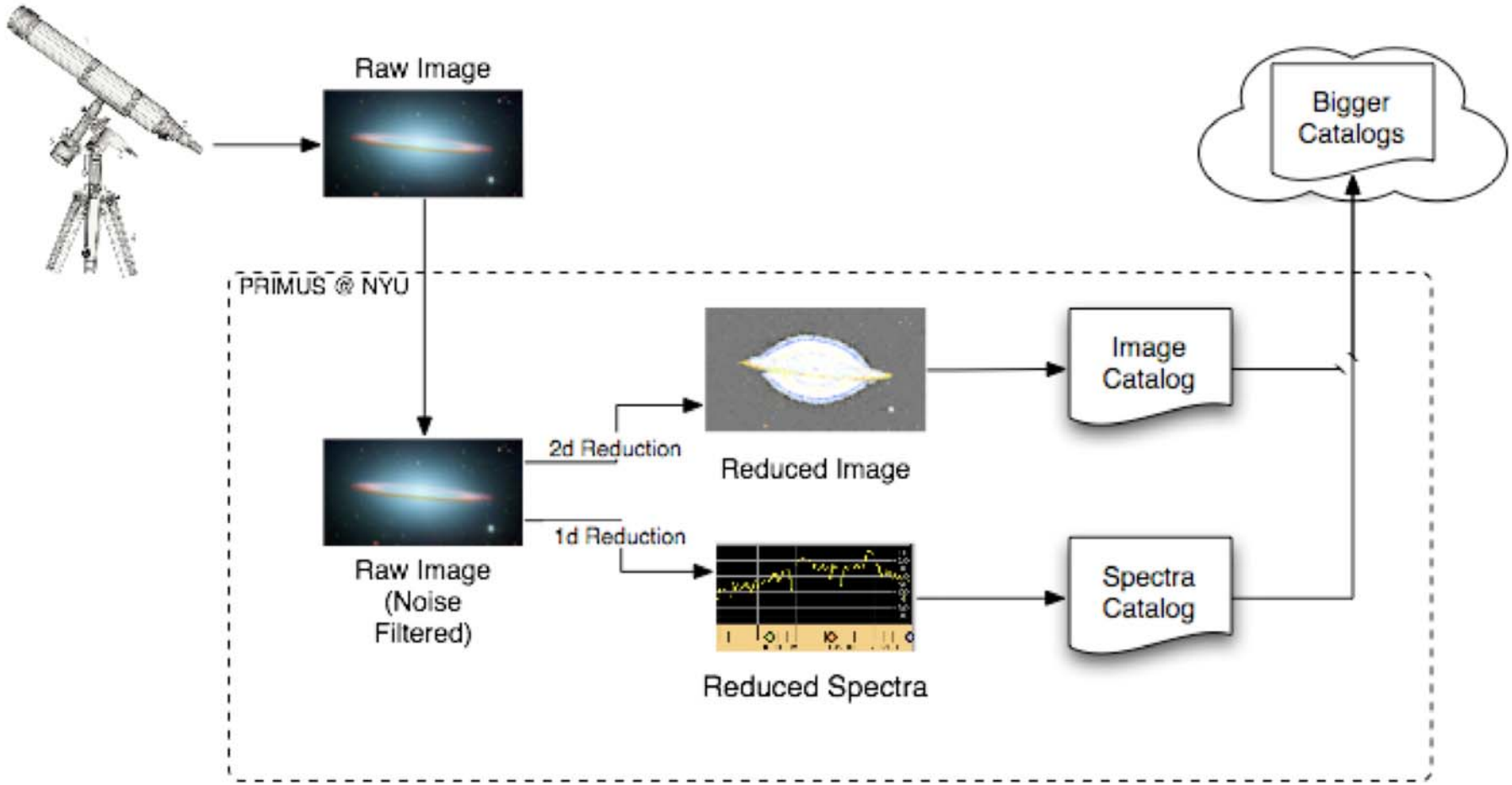
# The project

- Emerged from general discussions with NYU faculty about data management needs
- Pilot astronomy archiving project with the Center for Cosmology and Particle Physics in the NYU Department of Physics
- Data curation team composition from diverse library departments, including HPC representative

# Intro to PRIMUS

- PRIsm MUlti-object Survey (PRIMUS) is a cosmology research project to advance the study of galaxy evolution, quasars, clusters, and the large-scale structure of the universe
  - Similar to the Sloan Digital Sky Survey
- Wide-field redshift survey
  - Redshift occurs when a light source moves away from an observer, corresponding to the Doppler shift that changes the perceived frequency of sound waves.

# Research workflow & data products



# Flexible Image Transport System (FITS)

- All data in the research workflow is wrapped in FITS, **a data format designed to provide a means for convenient exchange of astronomical data between installations whose standard internal formats and hardware differ**
- Nonstandard metadata practices

# Data Size and Scope

- Size -
  - Pilot Project - “Trivial” (< 300 Gigs)
  - Full Commitment -10s of Terabytes
- Scope (Data Types)
  - “Content Data” Vs. Non-Content Data
  - Levels of Service



8

# Design Deliverables

- 1. SIP / AIP / DIP
- 2. Service Levels
- 3. Domain Model, extensible to general research data model/ontology
- 4. Data Scale Analysis



9

# 1. SIP / AIP / DIP

- SIP / AIP / DIP contents determined by:
  - data use / dissemination scenario analysis
    - what data do we need to support scenarios?
  - preservation requirements
    - what data do we need to support preservation?
  - selected Service Level Options

# PRIMUS **Data** SIPs

- Data SIP (Raw/Processed/Supporting)
  - PRIMUS has a well defined data model
    - http://howdy.physics.nyu.edu/index.php/PRIMUS_Data_Model
    - construct **data SIP** by leveraging existing PRIMUS structure
    - need to perform gap analysis
      - are additional data required to support use / dissemination scenarios?
      - are additional data required to support preservation?
    - some supporting data already in PRIMUS data model

# PRIMUS **Data** AIP

- Data AIP (Raw/Processed/Supporting)
  - METS for structural metadata
    - references to associated “Supporting Data” objects
  - Persistent Identifiers (PID):
    - Determine best way to use PIDs for PRIMUS data:
      - one PID per observation session?
        - observation session object can contain all raw, processed, and supporting data related to one observation session
      - one PID per raw image with links to supporting data?
        - raw image analogous to “master”
        - processed data analogous to “derivatives”
        - image object contains references to “Supporting Data”, “Tool”, and “Instrument” objects
      - others?

# PRIMUS **Data** DIP

- Data DIPs
  - what are the dissemination scenarios?
    - web interface
    - publication to external catalogs
    - others?
  - are intermediaries required to facilitate access?
    - data extracted from FITS headers?
  - the way we need to disseminate content determines some of the data required in the SIPs

# 2. Service Levels

- Data flow determines Service Level Options:
  - investigators choose what to preserve
  - e.g.,

| | **Processed** Data | **Raw** Data | **Supporting** Data | **Tool** Source / Docs | **Instrument** Data |
|---|---|---|---|---|---|
| **PI 1** | **preserve** | **no** | **no** | **no** | **no** |
| **PI 2** | **preserve** | **preserve** | **preserve** | **no** | **no** |
| **PI 3** | **preserve** | **preserve** | **preserve** | **preserve** | **no** |
| **PI 4** | **no** | **preserve** | **preserve** | **no** | **preserve** |
| **...** | **...** | **...** | **...** | **...** | **...** |

  - Note: some options **require researchers** to use best practices (e.g., software releases, documentation)

# 3. Domain Model

- Create abstract classes to contain PRIMUS data types
- Classes should scale to other scientific domains
- Implications for Access and Data Sharing

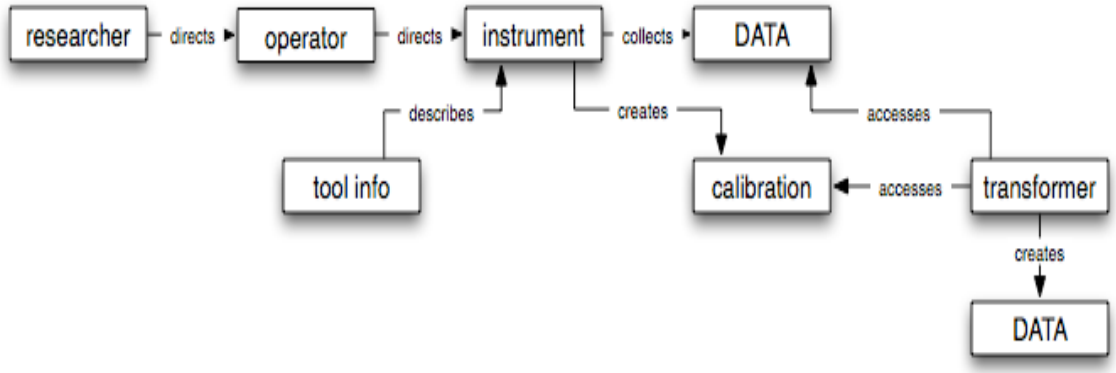


15

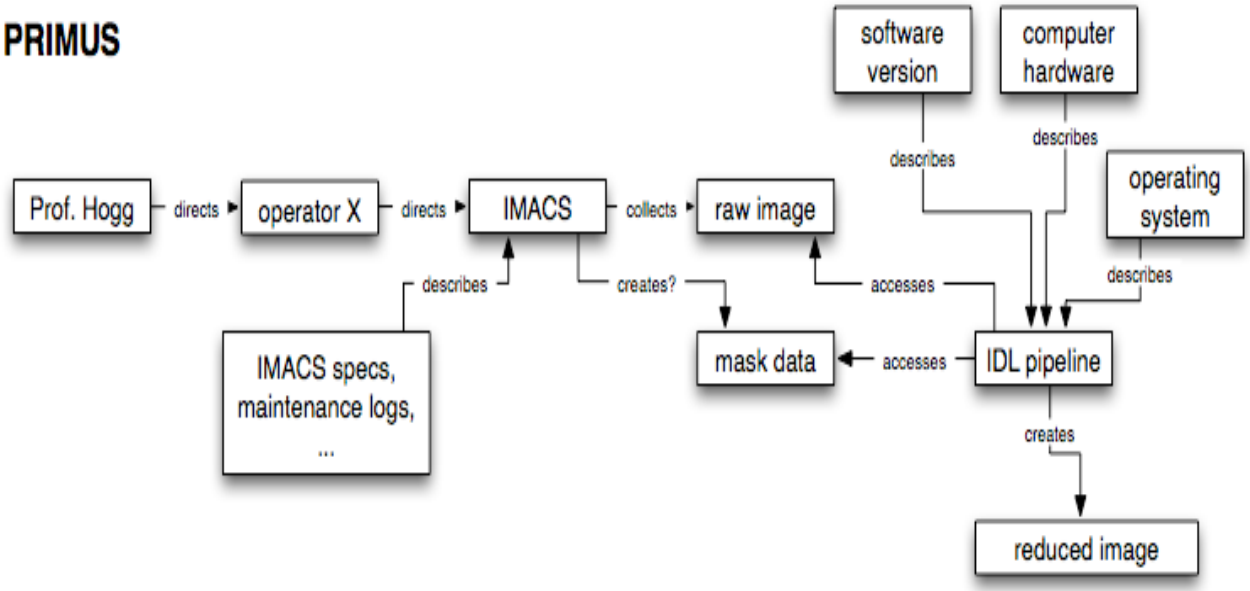
# Proposed model properties

- Holds sufficient and required information
- Must be able to accommodate other data types in the future i.e. different data types

## General




## PRIMUS




extend for image catalog updates, reduced spectra, etc.
value add: create tool/webapp that allows researchers to navigate ontology.

# Object Types

- Use **data flow** to identify **object types**:

  **data** type

  **supporting data** type

  **software tool** type

  **instrument** type

# Data and uncertainties

- **Data depend on:**
  - used different observables/experiments
  - type of instrument applied
  - the way the particular instrument is calibrated
- **Various coordinate systems and transformations are used**
  - Cartesian coordinates
  - Galactic coordinates
  - Elliptic coordinates … etc

# Existence of metadata

- Some partial information about instrument may exist – FITS BUNIT field
- Algorithms are not preserved
- Spectrum is a function of spectral coordinates corrected for systematic errors
- Spectral survey consist of bunch of spectral datasets which map given spectrum back to particular instrumentClass coordinates

# 4. Data Scale

- How much observation data is being created at NYU?
- How fast?
- What kinds of hardware configurations might long-term curation require?
- How much would it cost?



21

# Design decisions

# PRIMUS Scope?

- What is in scope for PRIMUS pilot project?
  - Certainly data, but all data?
    - processed : yes
    - raw : ???
    - supporting : yes?
  - Do we want to investigate Tool and Instrument data preservation using PRIMUS?
    - Tool SIP / AIP / DIP ?
      - is tool preservation useful for the researchers?
    - Instrument SIP / AIP / DIP ?
      - is instrument data preservation useful for the researchers?
      - the change history of the instrument?

# Evaluation of design deliverables

- For each deliverable
  - Assessment of infrastructure, tools, and policy needed to accomplish (which can use existing resources)
- develop a detailed cost analysis
- summary recommendation on the generalizability of the recommended workflow and ingest pipeline to other data curation projects at NYU.

# Concluding themes

- Data management will begin and run concurrently with the research activity, tracking the data lifecycle
- Data management tools must interoperate with the going research environment and workflows
- Libraries will need standard, extensible data management tools -- there will be limited opportunities to capture research funding beyond initial implementation

# Concluding themes, cont.

- Data management teams will need domain specialists.
  - Digital libraries could take on task of training data managers as a joint research mission - could a subset of DLF partners take on this role?
- What is data archiving when going practice is a distributed, peer to peer environment? DIPs will be important.
- Data management will make possible new web-based analysis tools as well as support preservation

# Contacts

- NYU Division of Libraries
  - Gretchen Gano, gretchen.gano@nyu.edu
  - Brian Hoffman, brianjhoffman@nyu.edu