Architectural considerations for AstroGRID

David Giaretta

Outline

- Look at various types of probable requirements
  - Users
  - Archives
  - AstroGRID administration
- "Standard" GRID capabilities
- Extra capabilities needed
- Data Model(s) and Data Access
- Metadata
- Techniques

Probable requirements - Users/Clients

- Able to do important science
- easy to join as a user
  - And also be able to use other "GRID" facilities if required i.e. be a member of several GRIDS e.g computational as well as data GRIDs
- well understood GRID access rights:
  - small quick things → free
  - larger things → by grant application (like PATT)
    - need to be able to control "expenditure" of allocation

.... User req ...continued

- transparent access to information
  - but able to see details if required
  - including access to non UK facilities
- friendly GUIs for standard applications
  - Database queries
  - Data analyses
- "easy" API to write own applications, and easy scripting
- easy to understand results
  - NB may be unfamiliar datasets
- trustworthiness of the data?

Probable requirements - Archives

- easy to "join" the Astro-GRID club of archives
- easy to publish
  - interfaces and other access to the archive's databases, other data, metadata, tools
    - but not too prescriptive
- maintain control of allocation and usage of resources
- maintain control of access to data
- tools/procedures for data curation

Probable Astro-GRID administrators' requ.

- Monitor performance and identify bottle-necks
- Control response times
- Able to dedicate resources as required to do big science
- Control resource use
  - UK
    - Short queries – free?
    - Long queries – allocation of time needed?
    - Background or Serendipitous queries
  - Non-UK – lower priority than UK?

Admin. Req. ...continued

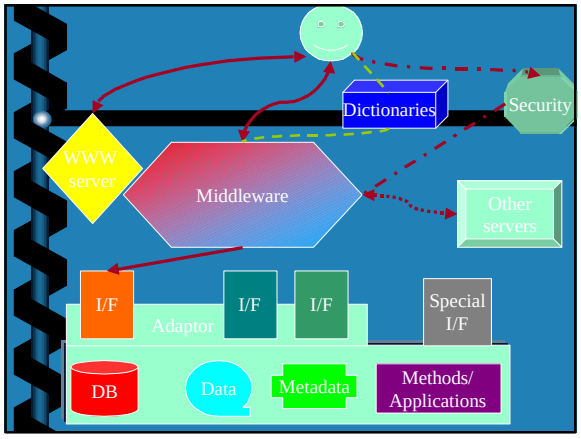
- Produce statistics of use
- Predict growth of use
- Troubleshoot
- Expand the membership of archives
  - Including foreign archives

"Standard" GRID capabilities

- Distributed security and authentication
- Generic information discovery tools
- Generic data transport and handling
  - Storage transparency, name and location transparency, collection management, replica management, etc.
  - Technologies: SRB, DPSS, MCAT, GridFTP, XML-based, etc.
- Request planning and resource scheduling / optimization
- Distributed execution management
- Wide-area event service

Grid extensions required

- Astronomy metadata standards
- Request translation
- Data access layer: data models, procedures, access protocols
- Simple archive interface for publishing data
- Distributed data mining tools
- Distributed data analysis tools
- Visualization tools for multi-parametric data

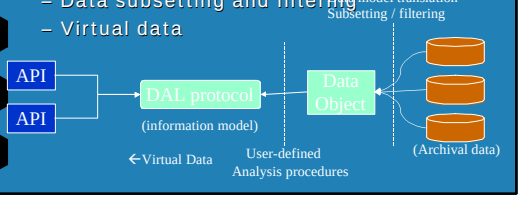


Data Model

- Astronomical
  - Image, spectrum
  - Coordinate systems
  - Quality, errors
- STP
  - Tabular
  - Timestamp index on rows

Data Access Layer

- Data access
  - Dataset file management
    - collection management, replica management, caching
  - Data model translation
    - provides data format transparency
    - example classes: images, spectra, event data
  - Data subsetting and filtering
  - Virtual data



Data Access Layer

- Computational services
  - Server-side computation is critical to maximize network performance, distribute computation for large queries
  - Standard subsetting/filtering methods for each type of data
  - Global catalog of object-specific analysis procedures
  - User downloadable functions
  - Analysis of image subsets
  - Dynamic extension of queries
  - Custom server-side analysis functions

Metadata

- Schema:
  - e.g. database schema or document DTD
- Navigational:
  - e.g. access such as URL
- Associative:
  - descriptive:
  - restrictive: e.g. user access rights
  - supportive: e.g. dictionaries, thesauri, ontologies

Software suites

- Starlink
- IRAF
- IDL
- Others?

Techniques

- Local applications:
  - Replace data access libraries
  - Add data location GUI
  - Add authentication hooks
- Remote Applications
  - application wrappers/ scripts
  - send scripts to applications sitting on server already
    - client-server
      - CORBA
      - SOAP
  - pipeline systems (ORAC-DR)
- Send applications to server
  - Globus toolkit

Info Discovery:Probable Requirements

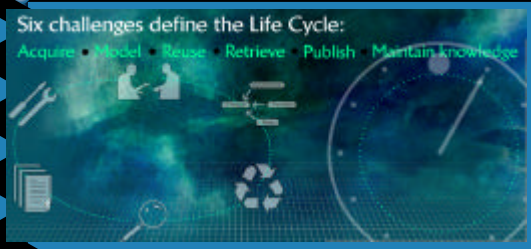
- Let the system filter out unwanted info
- Be able to locate information relevant to a more or less natural language question
  - Be able to handle scientific terms in various disciplines
- Be notified when something happens e.g. when specified object is observed or when some specified threshold passed

Caching

- Caching of catalog metadata
  - Allows information discovery
  - Permits correlations to be performed locally
  - Dataset replication permits efficient pixel level computation
- Caching of Catalogues
  - Allows efficient joins/correlations
  - Allows fail-over option
  - Improves response
- Caching Images/Spectra
  - Allow failover
    - Avoids delays on restore in case of catastrophe at one site

END

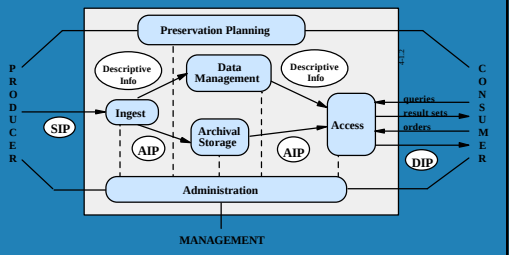
Challenge: Supporting the Knowledge Life Cycle



Knowledge Acquisition Technologies

- Protocol Analysis
- Process Mapping
- Laddering
- Repertory Grids
- Machine Induction
- Neural Networks.....

ISO Archive Reference Model



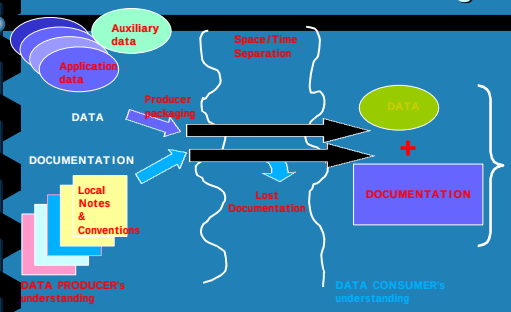
Mandatory Requirements

- *Negotiate and accept appropriate information from information Producers.*
- *Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.*
- *Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and therefore should be able to understand the information provided.*

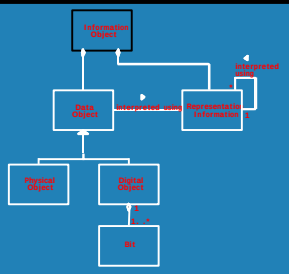
.....Mandatory Requirements

- *Ensure the information to be preserved is independently understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.*
- *Follow documented policies and procedures which ensures the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.*
- *Make the preserved information available to the Designated Community.*

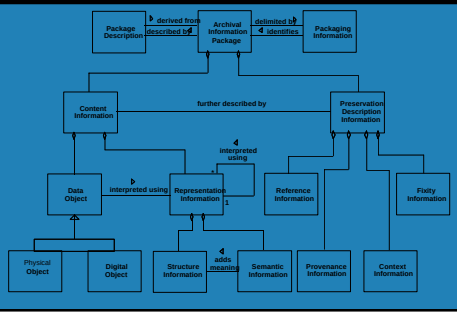
Problems in Automated Information Interchange



Information Object



AIP detailed view



Additional items

- Reports on computing and storage architectures
- Evaluation and possible inclusion of Agent technology
  - "encapsulated computer system, situated in some environment, and capable of flexible autonomous action in that environment in order to meet its design objectives"
- Model data flows and network response
- Naming schemes
- Ontologies (world view with respect to a given domain – a shared understanding)
  - Various domains with common engine as far as possible

The STP Specific Problems

- STP datasets are hosted on a variety of
  - computers using
  - many different operating systems, and
  - are held in many different formats
    - each dataset usually has its own retrieval method and
    - requires specially written software to be used for access and scientific manipulation.
- Most is conceptually TABULAR data
- Data at the archives should remain in its original format if possible
  - could put all data into e.g. ORACLE but there have been problems in the past doing this
  - lots of work

Possible Implementation steps

- LDAP
- Agents

Agent

"encapsulated computer system, situated in some environment, and capable of flexible autonomous action in that environment in order to meet its design objectives" (Wooldridge)

- control over internal state and over own behaviour
- experiences environment through sensors and acts through effectors
- reactive: respond in timely fashion to environmental change
- proactive: act in anticipation of future goals

31

Agents in the Grid

- Agents act on behalf of service owner
  - Managing access to services
  - Ensuring agreed contracts are fulfilled
    - Scheduling local activities according to available resources
    - Ensuring results are delivered
- Agents act on behalf of service consumer
  - Locating appropriate services
  - Receiving and presenting results

...caching

- Hierarchy of distributed queries is also possible
  - Supports large adhoc queries
  - Requires high network bandwidth between tier 1-2 nodes
  - Server side computation can be distributed to improve performance
  - Dataset replication and server-side procedures provide strategies to reduce network performance requirements