The Global Virtual Observatory:

How do we get there from here?

David Schade

National Research Council Canada

Canadian Astronomy Data Centre

Random thoughts about the GVO

NRC · CNRC

Why does the GVO not yet exist?

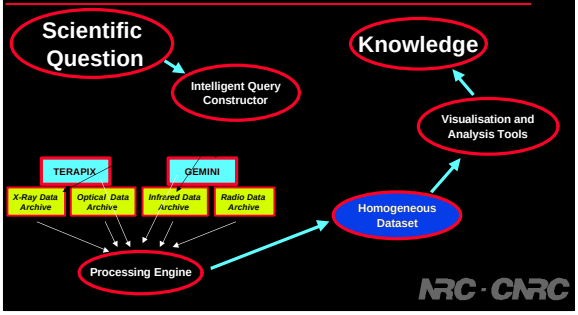
Technology is way ahead of our ability to use it to scientifically exploit astronomical observations.

Our archives could have much better linkages than they do right now.

Why are they in the state they are in?

NRC · CNRC

Data Mining (the GVO)



NRC · CNRC

Vision

Many of us now share a clear vision of what the GVO should be.

But very little time has been spent assessing where we are and how existing facilities (archives, data centres, projects) and skilled staff can be made to evolve into the GVO.

[It is curious that the GVO (in the US and elsewhere?) is driven by forces external to the data centres.]

** (Are the data centres losing their status as the visionaries of the astronomy/information-technology dynamic duo).*

NRC · CNRC

Vision

The most serious need for vision NOW is the need to see the *path* from present reality to the global virtual observatory.

NRC · CNRC

1. What is the Global Virtual Observatory?

Is the GVO about catalogues? Is the GVO about data?

The GVO is about catalogues and data and everything in between.

Every interesting query begins with:

" Give me a fair sample of … "

The GVO is about propagating the selection functions of all of the observational material through to the catalogue and the user.

NRC · CNRC

1. What is the Global Virtual Observatory?

The GVO is about working with catalogues but having the ability to drill down to observations when necessary.

NRC · CNRC

1b. What the GVO is NOT

1) Imagine that we had a network of infinite bandwidth

2) Imagine we had the equivalent of the local capabilities of all data centres in the world at your fingertips

3) Imagine we had infinite CPU cycles

4) Imagine we had infinite mass storage

We would have nothing remotely like the GVO

The GVO is not a technical capability

NRC · CNRC

2. Take Inventory: Where are we?

What facilities and content to we have in place now that form a foundation for the IVO ?

Canada: CADC (HST, CFHT, JCMT, GEMINI)

Europe: CDS Strasbourg, ESO, ST-ECF

UK: Cambridge Survey Unit, APM, COSMOS, La Palma archive, Leicester: X-rays

US: STScI MAST, HEASARC, other NASA facilities, Sloan

These facilities form an infrastructure and a skill base that form some part of the foundation for a GVO.

NRC · CNRC

Motivation for existing data centres to develop the Global Virtual Observatory

GVO is inevitable.

If existing data centres continue to limit their role to being providers of raw content, then the developers of the GVO will simply suck content from them and the data centres will lose their identities and their justification for funding.

NRC · CNRC

Do we have content worth connecting to the GVO?

• We must start acquiring data in a manner which is consistent with its function as content for the GVO

E.g., scientific queries will only be successful if data are indexed scientifically and if their selection function is known.

• Surveys provide good content. Our selection of programmes which are allocated time should consider GVO value.

• There is a huge scientific job of processing data from raw or other forms into content suitable for the GVO.

NRC · CNRC

The economics of the GVO: What does it take to buy in?

There has never been a free ride in science before and there is not going to be a free ride now.

If you don't contribute then your service will be fully duplicated elsewhere.

NRC · CNRC

GVO

- Its not about money
- Its about people with the scientific and technical vision to see the path
- Its about science
- Its about software
- Its about people who develop scientific software

Budgets should be 2/3 people and 1/3 hardware.

Staffing the GVO

Two models that need to be abandoned:

1) Get a couple of grad students to do the software

2) One person with a machine can do some neat things

Why?

- Too many facets to the problem.
- Complex storage systems (mix of DVD,CD,DLT, magnetic disks)
- Databases for process control, catalogue queries, request handling
- Scripts: IRAF, Perl, shell, SQL
- Code: c,c++,Java, fortran
- System administration
- Database administration

Too many skills needed
Too many things to do
Need to design software

3a. Random thoughts Distributed versus centralized facilities

- Low-level data can be in a distributed form.
- Can effective querying be done on distributed catalogues?
- Can distributed processing be done practically?
- Is centralised expertise (CS and Science) preferable to isolated outposts?
- Is a centralised multi-wavelength centre the way to go?

3b. Random thoughts Sharing and Protecting data

In planning the GVO we need to discuss frankly the issue of proprietary data rights.

The U.S. has had a 3 day meeting on the NVO where proprietary rights were NEVER MENTIONED.

ESO is restricting some data rights to member countries.

Is the sharing of data a motivation for GVO?

Is the conviction that sharing data will be more productive and more efficient a motivations for GVO?

4. CADC: Our first steps down the path toward GVO

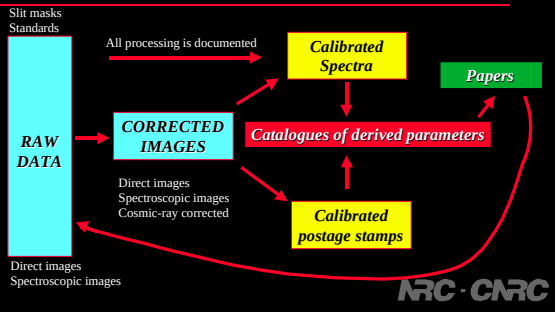
I. The First Steps

- *Building the CFHT Data Archive*
- *HST Data Archive*

  On-the-fly recalibration
- *JCMT Data Archive*

  SCUBA recalibration
- *Gemini Data Archive*

The CNOC Archive as a prototype VO



Contributions by CADC

- Design for a virtual observatory
- System for querying massive databases
- Distributed processing system

NRC · CNRC

Content Contributions by CADC

- WFPC2 datasets and catalogues

  Recalibration, stacking, photometric/astrometric calibration, catalogues
- CFH12k datasets and catalogues

  Calibration, stacking, photometric/astrometric calibration, catalogues
- CFHT MegaCam Surveys

  600 nights(?) of CFHT time for a package of surveys

  Calibration, stacking, photometric/astrometric calibration, catalogues

NRC · CNRC

CADC plan

- Take all of the content that we have produced and link it with a few other selected catalogues (e.g. 2MASS, SDSS, IRAS,ROSAT)
- Construct a large catalogue database with good querying tools
- Link the catalogue entries with the (local) data the came from
- Gradually extend the capabilities

We must regularly produce valuable new products while our ultimate goal is the full capability of the GVO

NRC · CNRC

The path

The first step on the path is to produce content that is worth connecting to the GVO.

The second step is to try to connect it locally

NRC · CNRC

Suggestions for discussion

- The IVO needs to be a merging of existing facilities rather than creation of a new one from scratch
- Distributed systems have many difficulties. Centralisation has many benefits

The End

NRC · CNRC