Data behind a Publication

—

An astronomer´s view

Berlin 4 Open Access
March 30, 2006

Wolfgang Voges
Max Planck Institute for Extraterrestrial Physics
with contributions from Carlson, Genzel, Hasinger, Lemson, Springel, Szalay

Abstract:
The amount of data being produced yearly in astronomy will soon reach the Peta-byte limit. Observational data covers the whole frequency or wavelength regime (radio, infrared, optical, ultraviolet, X-ray, and gamma-ray) and is collected from both ground-based telescopes and instruments on satellites. At the same time, simulations on supercomputers produce equally daunting quantities of theoretical data. Innovative methods and tools are needed to ingest, digest and concentrate this data before publishing the results and triggering a new cycle of knowledge discovery.
This talk addresses a new conception of the relationship between data and publications. It is increasingly important to make not only the final, condensed results available, but also the raw data, detailed methods, and complete output. This supplementary information, which may consist of lists, catalogues, image collections, or computer programs, has to be in electronic form to be used effectively.
It is now common policy in the field of astronomy that data are made public soon after they have been recorded, processed, or calculated. This accords with the spirit of publicly funded research, enhances the quality of the science, discourages sloppy or unethical conduct, and encourages rapid progress.

Outline

- Astronomical data
- The Virtual Observatory and the GRID
- Publishing the Data
- Closing remarks

Astronomical data

Processed Astronomical data

- Images (photometric data, position, brightness, colour)
- Spectra (low/high dispersion, diff. elements, Doppler motion…)
- Time series (movement, pulsations, supernova..)

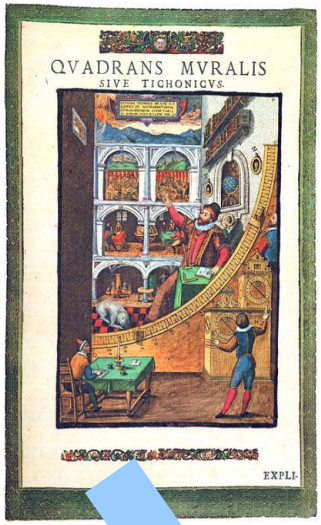
Derived Astronomical data

- **Object Cataloques** (position, flux, type, ID…)

Computer simulations

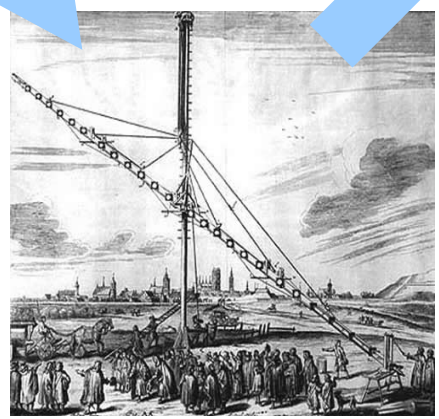
The growth of astronomical catalogs

Tycho Brahe at work



1,000 stars
several pages

invention of the telescope (1609)



Hevelius' 150-foot long refractor

1,000,000 stars
several volumes

invention of photograph (1884)

Hale 200-inch reflector

1,000,000,000 stars
80 GB
a library

invention of the digital camera (1990)

10 Tbytes
Library of Congress
- every night!

Pan-STARRS

background info: not presented during the talk

Star catalogs:

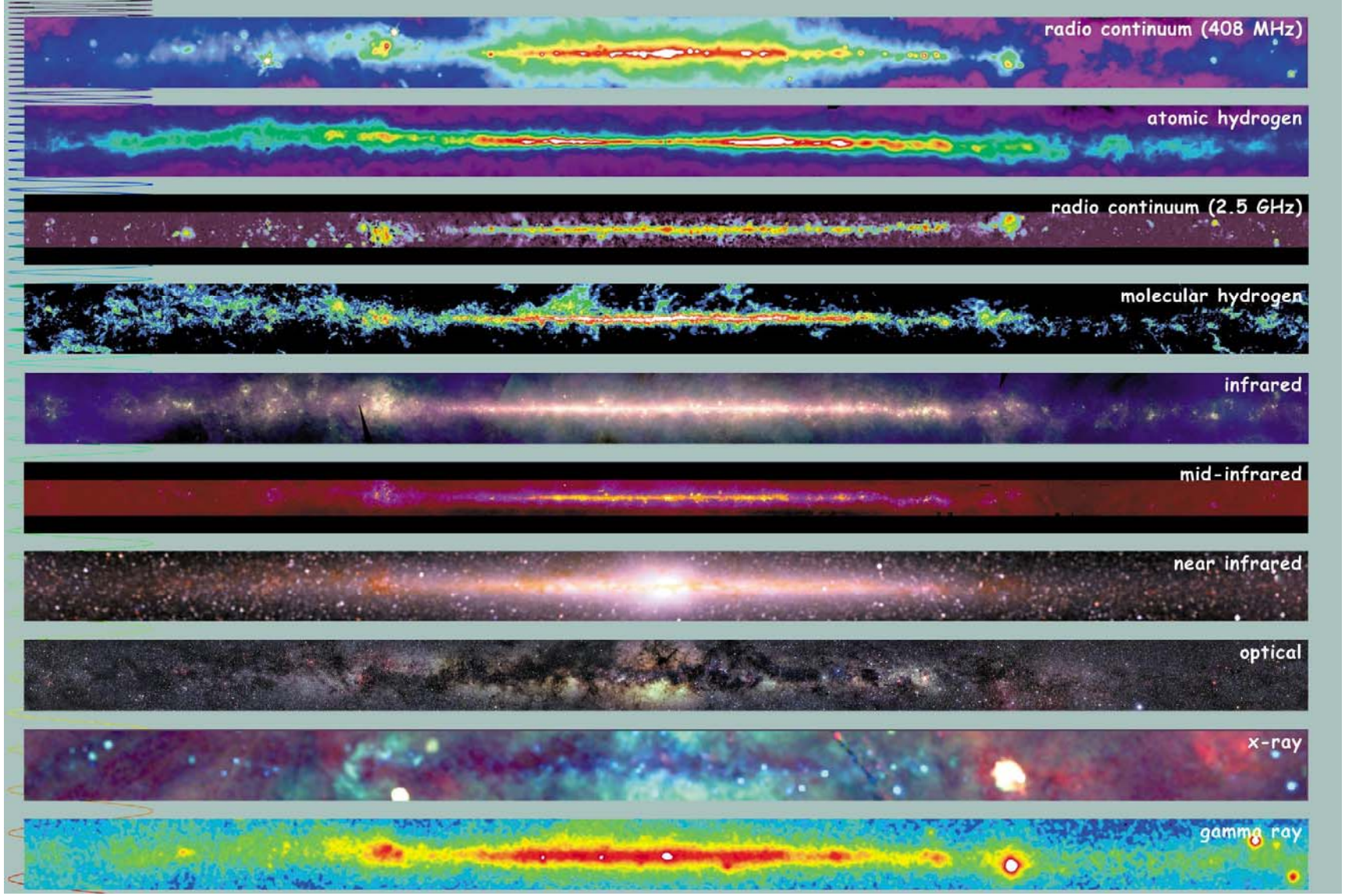
ca. 340 BC Gan De (China)
300 BC Timocharis of Alexandria
130 BC Hipparchus (of Rhodes)
120 AD **Ptolemy** (83-161 AD) compiled his monumental "Almagest", a catalog of 48 constellations and **1022 stars**.
1540 Allessandro Piccolomini (1508-78), *De le Stelle Fisse*
1603 Johann Bayer (1572-1625), *Uranometria*; note this appeared still before the introduction of the telescope in astronomy in 1609.
1661 Johann Hevelius (1611-1687), *Sternverzeichnis*
1679 Edmond Halley (1656-1742) compiled the first southern star catalog
1725 John Flamsteed, *Stellarum Inerrantium Catalogus Britannicus*.
1762 James Bradley (1693-1762), Star Catalog.
1821-**1835** Friedrich Wilhelm Bessel (1784-1846) determined accurate positions for **32,000** stars
1852-**1859** *Bonner Durchmusterung* (BD), **320,000 stars.** Included visual photometric estimates.
1892 *Cordoba Durchmusterung of the Southern Sky*. 120,000 stars. Included visual photometric estimates.
1904-1908 *Göttinger Aktinometrie. A pioneering work in the area o photographic stellar photometry.*
1891-1950 *Catalogue astrographique.* Photographic. Over **4.6 million stars**.
1918-1924 *Henry Draper Catalog.* The monumental early work in the area of spectral classification. 225,300 stars.

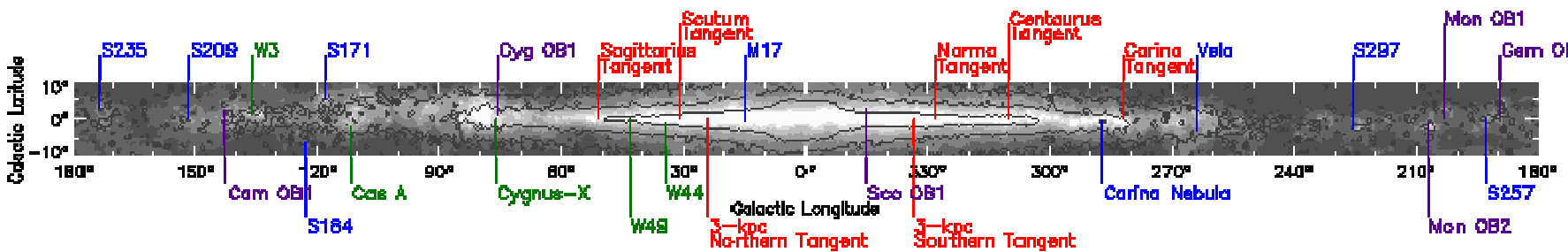
The largest object catalogue:
USNO-B1.0 **(U.S. Naval Observatory)** Positions, proper motions, magnitudes in various optical passbands, and star/galaxy estimators for **1,042,618,261** objects. The data were obtained from scans of 7,435 Schmidt plates taken for the various sky surveys during the last 50 years.

The Strasbourg astronomical Data Centre (CDS) collects and distributes astronomical data catalogues, related to observations of stars and galaxies, and other galactic and extragalactic objects. Catalogues about the solar system bodies and atomic data are also included.

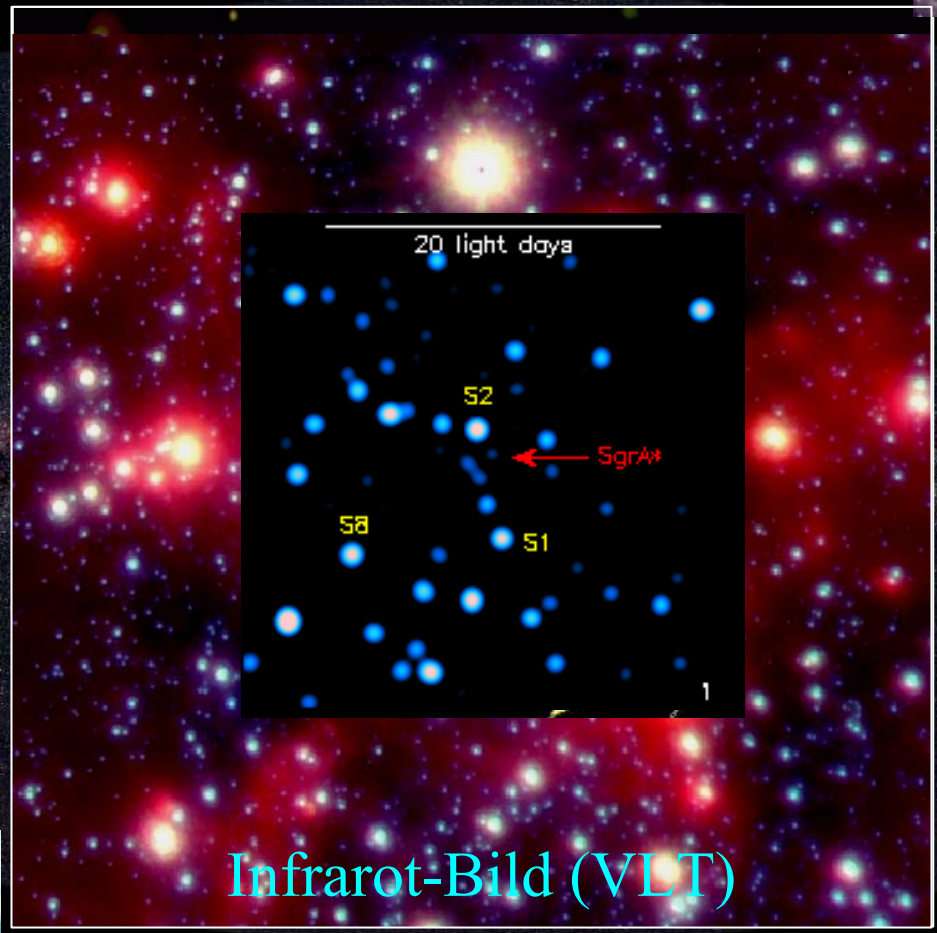
6475 Catalogues are available from CDS, of which **5643** are available on-line. They contain approx. 3,000,000,000 objects of 0.5 Tbyte. (CDS database on 15 mirror sites around the world!!!)

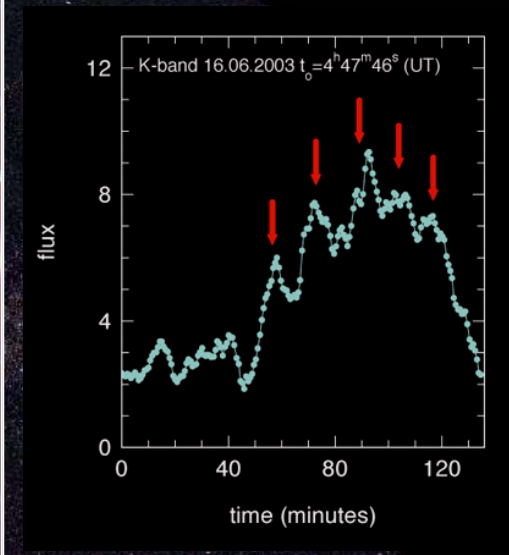
Since January 1993, Tables from articles published in *Astronomy & Astrophysics* are prepared by and made available at CDS.


radio continuum (408 MHz)
atomic hydrogen
radio continuum (2.5 GHz)
molecular hydrogen
infrared
mid-infrared
near infrared
optical
x-ray
gamma ray


S235
S209
W3
S171
Cyg OB1
Sagittarius Tangent
Scutum Tangent
M17
Norma Tangent
Centaurus Tangent
Carina Tangent
Vela
S297
Mon OB1
Gem OB1
Galactic Latitude
10°
0°
−10°
180°
150°
120°
90°
60°
30°
0°
330°
300°
270°
240°
210°
180°
Cam OB1
S184
Cas A
Cygnus−X
W49
W44
3−kpc Northern Tangent
Sco OB1
3−kpc Southern Tangent
Carina Nebula
Mon OB2
S257
Galactic Longitude

Our Galaxy

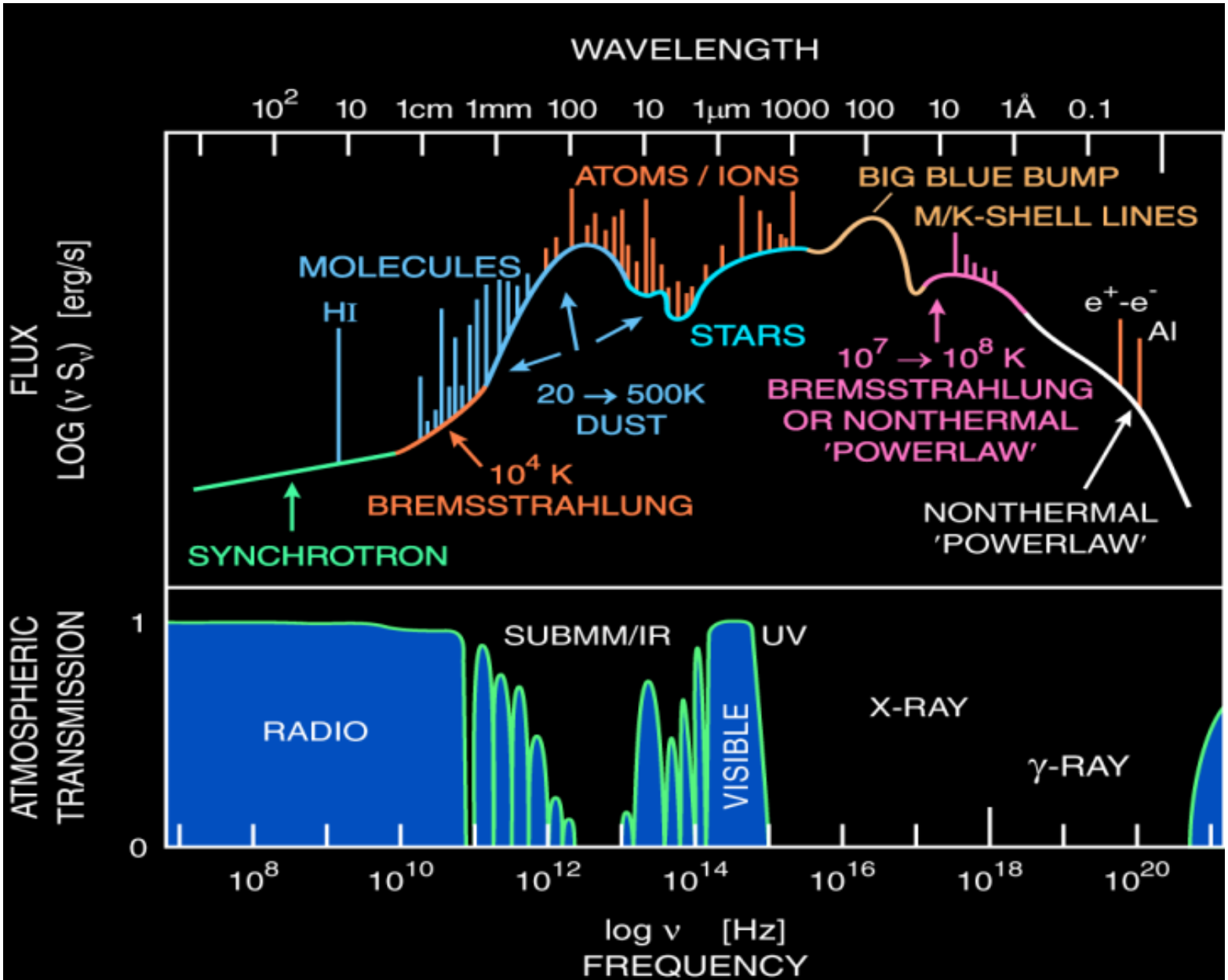


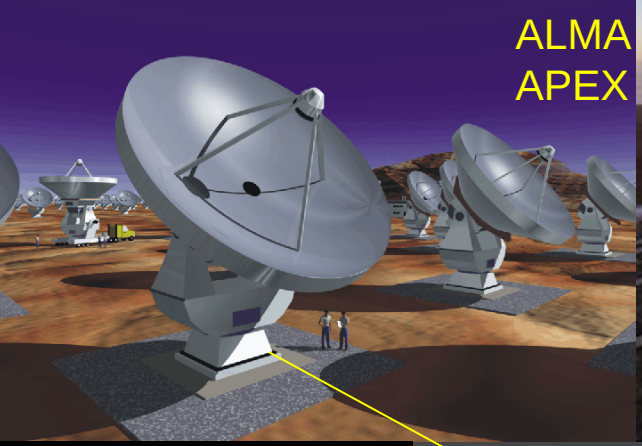


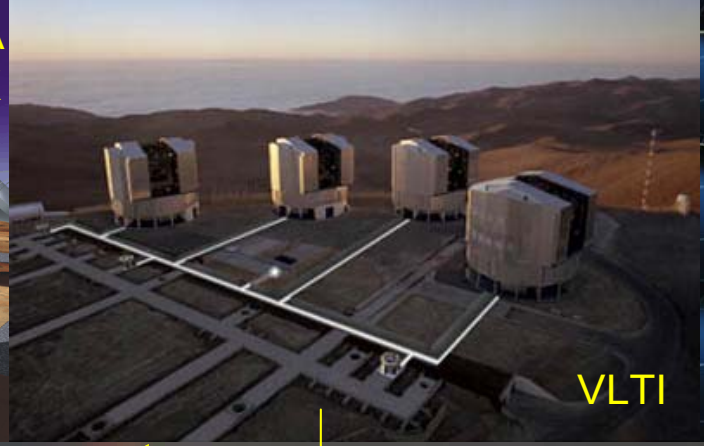
Stellares Schwarzes Loch

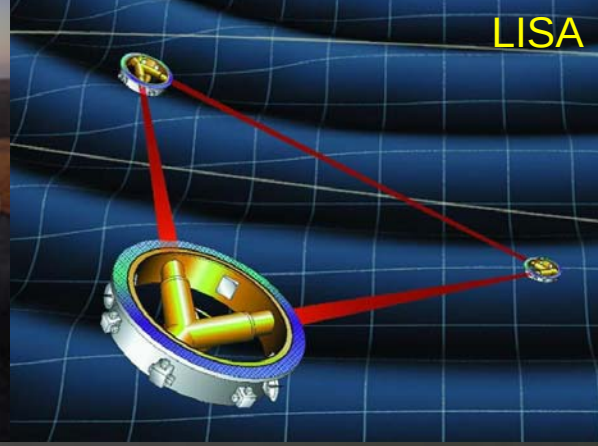
Rapidly rotating black hole with 3-4 million solar masses

Multi-λ Information

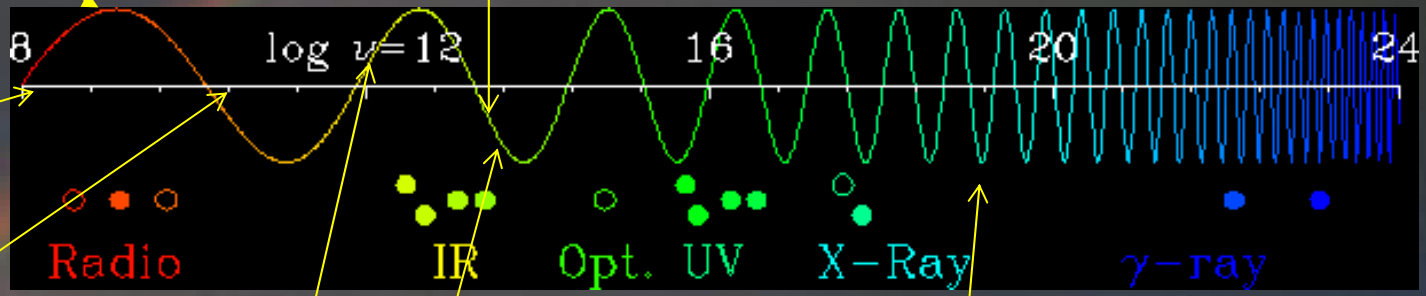


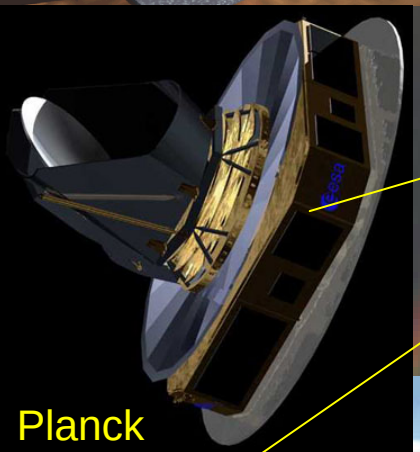




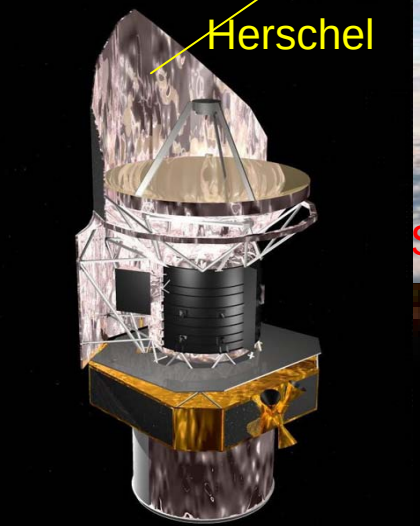


1 Oktave

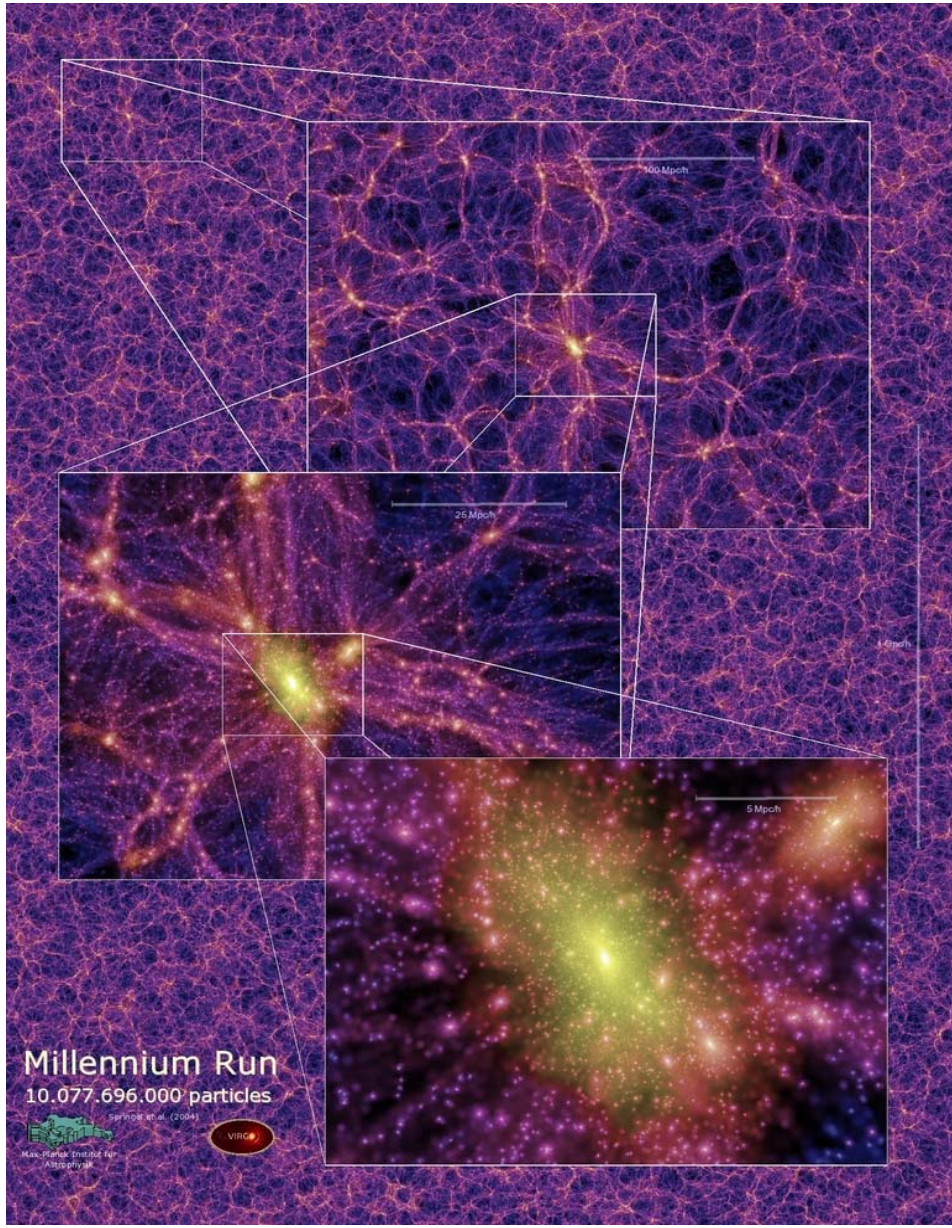




56 Oktaven (7 „Grand Pianos")



The Millennium Run used more than 10 billion particles to trace the **evolution of the matter distribution** in a cubic region of the Universe over **2 billion light-years** on a side. It kept the supercomputer at the Max Planck Society's Computer centre in Garching busy for **more than a month**. By applying sophisticated modeling techniques to the **25 Tbytes** of **stored output**, Virgo scientists have been able to create evolutionary histories both for the 20 million or so galaxies which populate this enormous volume and for the super-massive black holes which occasionally power quasars at their hearts. **By comparing such simulated data to large observational surveys, one can clarify the physical processes** underlying the build-up of real galaxies and black holes.
The illustration shows a projected density field for a 15Mpc/h thick slice of the redshift = 0 output, a massive cluster of galaxies. The overlaid panels zoom in by factors of 4 in each case, enlarging the regions indicated by the white rectangles.



A 80-second movie of the VIRGO project:
The movie shows the dark matter distribution in the universe at the present time, based on the Millennium Simulation. By zooming in on a massive cluster of galaxies, the movie highlights the morphology of the structure on different scales, and the large dynamic range of the simulation. The zoom extends from scales of several Gpc (Giga parsec ~ 3.3 billion light-years) down to resolved substructures as small as ~ 10 kpc.

Run millenium-simulation 1024x768 millennium2

(the movie is available under http://www.mpa-garching.de/galform/virgo/millennium/index.shtml)
(http://www.mpa-garching.mpg.de/galform/data_vis/millennium_sim_1024x768.avi)
Credit: Springel et al. 2005, Nature, 435, 629

A 2-minute movie of the VIRGO project:
A 3-dimensional visualization of the Millennium Simulation shows a journey through the simulated universe. On the way, we visit a rich cluster of galaxies and fly around it. During the two minutes of the movie, we travel a distance for which light would need more than 2.4 billion years.

Run millenium flythru-fast millennium2

(The movie is available under http://www.mpa-garching.de/galform/virgo/millenium/index.shtml)
(**http://www.mpa-garching.mpg.de/galform/data_vis/millennium_flythru_fast.avi**)
Credit: Springel et al. 2005, Nature, 435, 629

Outline

- Astronomical data
- The Virtual Observatory and the GRID
- Publishing the Data
- Closing remarks

Definition of a Virtual Observatory

A collection of **integrated (distributed) astronomical data archives** and **software tools** (**open source**) that utilize **computer networks** to create an environment in which research can be conducted.

Several countries have national virtual programs that will **combine** existing databases from **ground-based** and **orbiting observatories** and make them **easily accessible** (**open access**) to researchers ++.

Virtual Observatory (2)

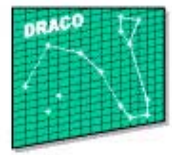
As a result, **data** from all the world‘s major observatories will be **available to all** users and to the public. This is significant not only because of the **immense volume** of astronomical data but also because the data on stars and galaxies has been compiled from observations in a **variety of wavelengths** – radio, infrared, optical, X-ray, gamma ray and more.

Each wavelength can provide **different information** about a celestial event or object, but also **requires a special expertise to interpret**.

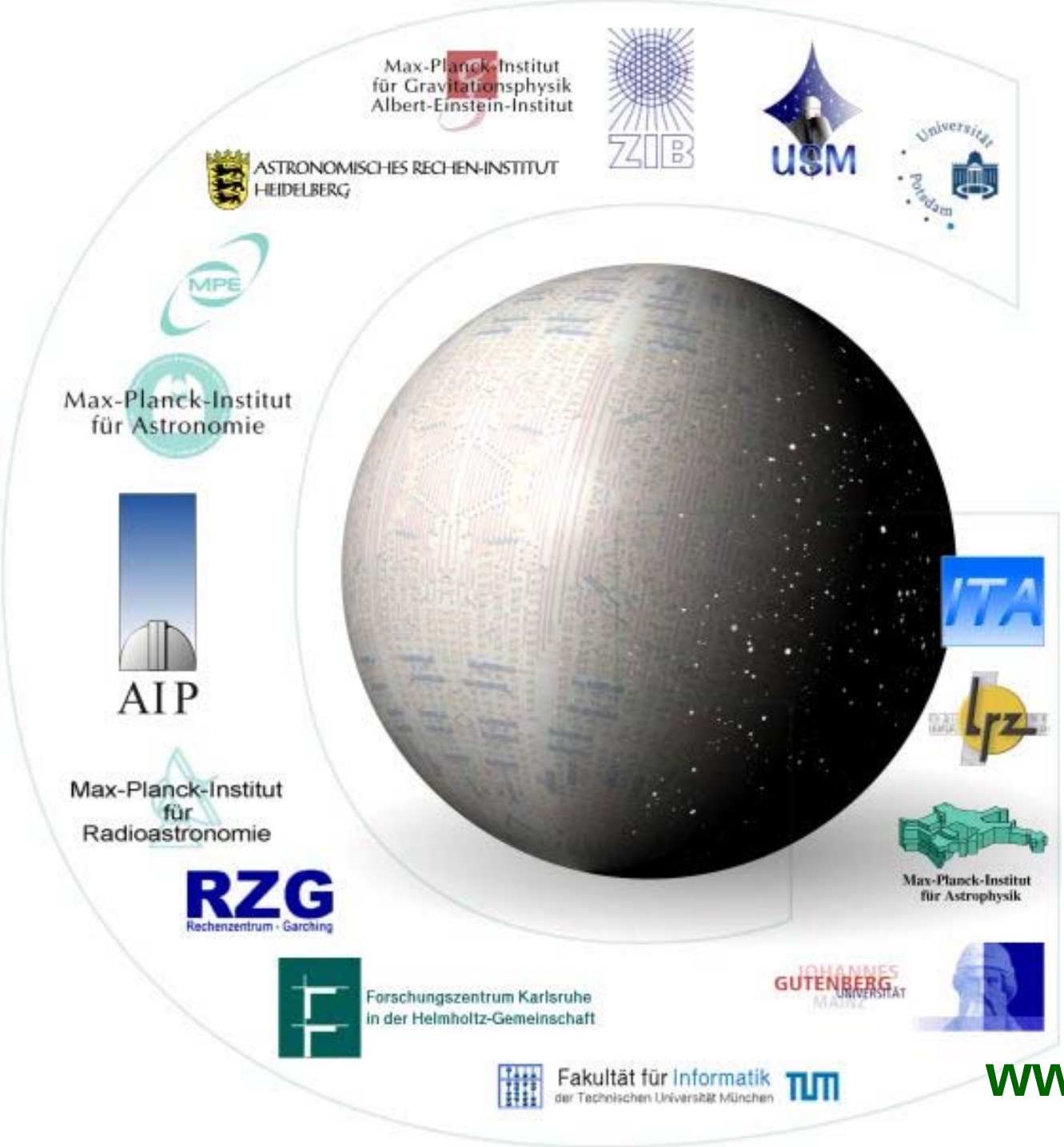
In a virtual observatory environment, all of this **data** is **integrated** so that it **can be synthesized** and used in a given study.

International Virtual Observatory Alliance

www.ivoa.org



www.g-vo.org



German Astronomical Community GRID

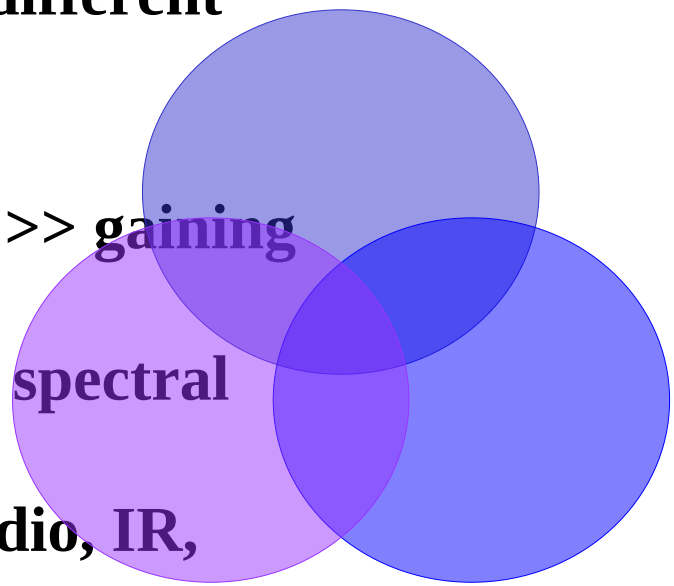
AstroGrid-D

www.gac-grid.org

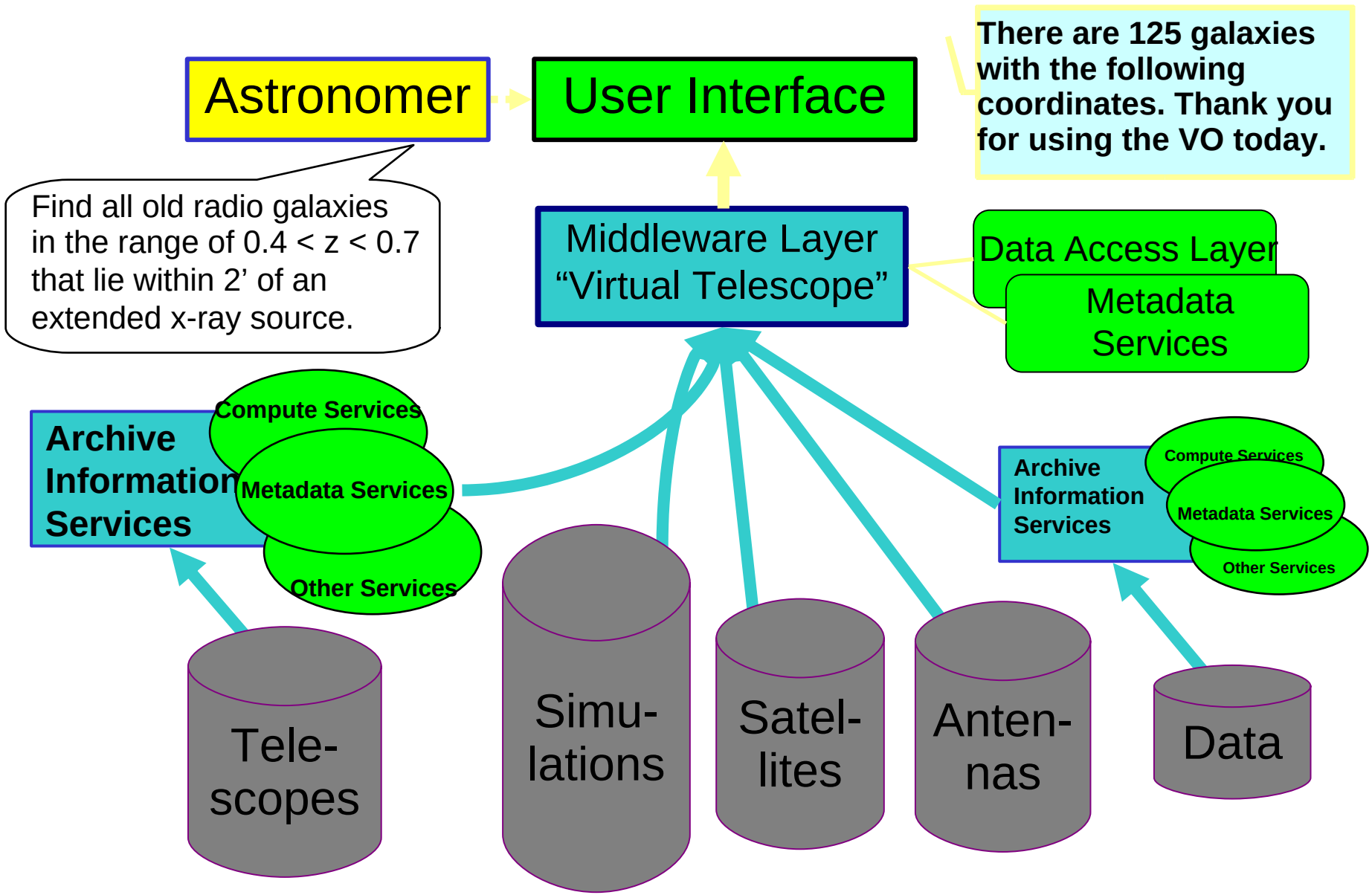
Making Discoveries

- Where and how are discoveries made?

  - **At the edges and boundaries of different disciplines**
  - **Collecting more data**
  - **Going deeper (longer exposures >> gaining sensitivity)**
  - **Using data of higher spatial and spectral resolution**
  - **Combining more colours (i.e. radio, IR, UV,optical, x-ray, gamma-ray)**
  - **Adding data from the time domain**
  - **Comparing observational data and simulations**



The Virtual Observatory

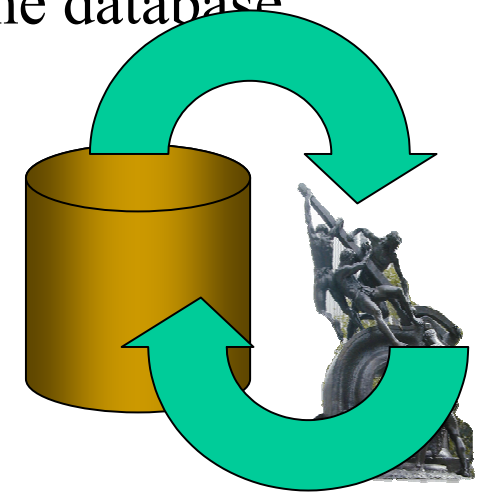


Smart Data

- If there is too much data to move around,

  take the analysis to the data!

- **Do** all data **manipulations at the database**
  - Build custom procedures and functions in the database
- Automatic parallelism guaranteed
- Easy to build-in custom functionality
  - Databases and procedures are being unified
  - Example: temporal and spatial indexing
  - Pixel processing
- Easy to reorganize the data
  - **Multiple views, each optimal for certain analyses**
  - Building hierarchical summaries is trivial
- Scalable to petabyte datasets



active databases!

Astronomy in the VO and GRID World

- Virtual Observatory and GRID activities are **complementary**.
- Common goal: **linking huge data archives** and providing **fast access** to them
- Requires **common data structures, formats, access protocols, registration of data, and tools** – in short, **standards**.
- GRID computing enables the **use of dispersed computing power** and **data storage**.
- The same concepts can be used to link experiments, telescopes, and simulations.
- **VO: OA to data and services GRID: OA to resources**

Main VO Challenges

- **How to avoid trying to be everything for everybody?**
- Database connectivity is essential
  - **Bring the analysis to the data**
- Core web services, higher level applications on top
- Use the **90-10 rule**:
  - **Define** the **standards** and **interfaces**
  - Build the framework
- Build the **10% of services** that are **used by 90%**
  - Let the **users build the rest** from the components
- Rapidly changing "outside world"
- **Make it simple**!!!

Special features in Astronomy and Astrophysics

- **Nature of the data**
  - huge quantities, very heterogeneous, dispersed world-wide
  - in principle largely open but in practice hard to find
  - both observations and simulations
- **Tradition and culture**
  - Invention of the World-Wide-Web by CERN scientists, early adoption and spread by astronomers and astrophysicists
  - Initiative for the Virtual Observatory and establishment of the IVOA
  - Preprint philosophy and extensive use of search engines
- **Current developments**: Astronomical journals in the USA, in cooperation with ADEC, ADS and IVOA, are developing standards for references to on-line data sets in publications. See

http://vo.ads.harvard.edu/dv/

Outline

- Astronomical data
- The Virtual Observatory and the GRID
- Publishing the Data
- Closing remarks

Publications in research

- The production and use of publications is central to scientific research.
- The **content** of a publication includes ...
  - identifying information (title, date, authors, addresses)
  - text (abstract, outline, various sections)
  - figures (diagrams, maps, photographs) *
  - tables (in astronomy, often catalogues of objects) *
  - references to the data used and to other publications
- To **access** this content efficiently requires archives, data bases, and tools, which together permit ...
  - search on title, date, authors, keywords (metadata)
  - full text search with boolean expressions
  - on-line availability of full text and supporting data

* if data too huge often published only partly with a link to full lists, catalogues, image collection etc.

Technical requirements

- Traditionally, publications are submitted to an established journal and printed on paper.
- Open access journals and, to some extent, preprint servers, serve a similar purpose but have the advantages of free access and full text searches.
- To replace traditional journals, it is essential that they be
  - centrally registered
  - indefinitely and reliably available
  - stored and maintained in long term archives

This is a technical challenge!

Scope of documentation needs

Planning and carrying out a project requires documentation – in electronic form – of ...

- communication
  - correspondence (emails, faxes, letters)
  - protocols and notes from meetings, telephone calls, and video conferences
- planning
  - schedules of milestones and deadlines
  - workflows, action items, bug tracking
- results
  - raw data, log books, lab books, photographs
  - processed data, program versions, processing parameters
  - reports, deliverables, publications, presentations

>>>> **eLAB** ??

Virtual Observatory and Refereeing

“VO leads to bad science”

- It is often claimed that, if it becomes too easy for non-specialists to get access to data products, they may draw conclusions that are unfounded . They are not able to take fully into account how the data came to be.
- Only experts can use the data properly, and they already know how to get the data they need.

But

- **Bad science is done by bad scientists, not by bad data**
- This is a problem with or without a VO
- The solution is a referee system, where experts filter publication submissions
- Difficult, but very important

Help the referees

- By making **all data** underlying a publication available,
- in **raw** form and
- in **processed** form,
- together with the **analysis tools** used to go from raw to processed data.

How ?

- Standardized publication of data
- Hyperlinks from within publication
- Tagged with metadata
- Stored persistently
- Also useful for the readers of the final publication of course ...

Tasks

- **Define standards**
  - IVOA for example
- Provide **storage**
- Provide **CPU**
- Provide **services** and **tools**
- **Enforce standards** :
  no publication if data can not be accessed
- somewhat harsh, but a good starting point, science is supposed to **rely on reproducibility**

Publications as data

• The Virtual Observatory and GRID projects of the world-wide astronomy community are developing **generic standards, data models, and tools** for registering, archiving, processing, and visualising large, heterogeneous data sets.

• With close cooperation, these can be synergetically **transferred to other disciplines**.

• This is **especially true for publications**, which can be seen as a special sort of data.

• The concept of a publication should include **auxiliary (supplementary) information and supporting data** that are not directly included in the publication.

Outline

- Astronomical data
- The Virtual Observatory and the GRID
- Publishing the Data
- Closing remarks

Closing remarks

OA issues have been treated within the astronomy community for a long time (but traditionally different for ground-based and space-born instrumentations).

OA policy is a cultural issue which needs to be resolved.
Not every astronomical observatory has an archiving policy which is in favour of OA.

The Virtual Observatory and GRID activities are making data, services and resources freely available to all researchers ++.
(open issue is the accounting of extraordinary resources)

The publication of the data evaluation process and the results together with supplementary data still needs to be standardized.

It is now almost common policy in the field of astronomy that data are made public soon after they have been recorded, processed, or calculated.
This accords with the spirit of publicly funded research, enhances the quality of the science, discourages sloppy or unethical conduct, and encourages rapid progress.

A very personal last remark:

I am not just excited by meeting so many people from other disciplines here, but also by discovering that we share similar problems across the spectrum from the humanities to my field of astronomy.

Last but not least I see here the possibility and the need for collaborating more closely on common problems and for looking together for solutions which would serve both communities.

The Max Planck Digital Library could be the place to get this done.

The end