GeS 2007 German e-Science

Available online at http://www.ges2007.de

The German Astrophysical Virtual Observatory

Knowledge networking for astronomy in Germany and abroad

Gerard Lemson[1,2], Wolfgang Voges[1], Joachim Wambsganss[2], on behalf of the GAVO team
[1] Max-Planck-Institut fur extraterrestrische Physik, Garching
[2] Astronomisches Rechen-Institut, Zentrum für Astronomie, Heidelberg
Email: {gerard.lemson, wolfgang.voges}@mpe.mpg.de,
jkw@ari.uni-heidelberg.de
phone: (+49 089) 30000 3316, *fax:* (+49 089) 30000 3569

Abstract

We describe the work of the German Astrophysical Virtual Observatory (GAVO). GAVO is the German node in the world-wide international virtual observatory (VObs) effort aimed at making geographically distributed astrophysical data and applications more readily available to the community and to find ways of combining these in an interoperable network so as to create new ways of doing science. In this presentation we give a short overview of the ideas behind the virtual observatory and describe the efforts of the International Virtual Observatory Alliance (IVOA). We describe then in detail the activities of GAVO, both in its first pilot phase and in the recently initiated second phase. Here we will pay particular attention to the effort spearheaded by GAVO to include results of large scale computer simulations in the VObs efforts, both national and international.

1 Knowledge networking for astronomy: the Virtual Observatory

In recent years, advances in instrumentation have made it possible to perform astronomical all-sky surveys that produce huge catalogues containing hundreds of millions of objects, extracted from image catalogues containing many terabytes of data. These surveys have mapped the sky in many wavelength regimes, from radio through infrared and optical to X-ray. In the coming years, surveys are planned that will map the whole sky in a couple of days, producing catalogues of petabyte sizes, containing billions of objects. Similarly, advances in computer technology as well as numerical algorithms have allowed numerical simulations producing many tens of terabytes of data.

The data reduction and analysis of such large data sets require new techniques and cannot be performed by individuals or small groups anymore. Instead, the data will generally be made available to the astronomical community, something that is actually more and more forced upon the data pro-

ducers by their funding agencies. Regarding the size of the data, the old style publication of data of allowing file downloads is no longer feasible. Instead new ways of online publication have to be developed. The first such efforts were concentrated around individual archives and data centers, allowing users to obtain observations of parts of the sky by querying a remote catalogue online, instead of requiring a new telescope observation.

It was quickly realised that if a common approach was followed, the individual nodes could be linked together and similar queries might produce multi-wavelength views of the sky, opening a door to the general astronomical community for a kind of research that so far was only possible in large collaborations.

This common approach to data publication has been coined the *Virtual Observatory (VObs)* and the ideas have been extended from data publication to include more generic online (computational) services and even robotic telescopes.

2 Standardisation: International Virtual Observatory Alliance (IVOA)

It is generally accepted that in order to allow independently created and distributed services and data archives to interoperate so as to facilitate new scientific research, it is not sufficient to just link the computational infrastructure as is the target of the Grid approach. On top of this one needs to standardise on messaging protocols and data descriptions so that users can access the different nodes in the same way. To this end the International Virtual Observatory Alliance (IVOA) was formed. Its mission is to *facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory*[1]. As of January 2007, the IVOA consists of 16 funded VObs projects from Armenia, Australia, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the United Kingdom, and the United States.

The work of the IVOA has been organised in a number of working groups, aimed at creating standards for data description and modeling, data query and access protocols, resource registries and grid and web services. The first

[1] http://www.ivoa.net/pub/info/

results were standards for tabular data formats (VOTable), some simple query protocols, both for source catalogues and image archives (Simple Cone Search, Simple Image Access) and a first attempt at a simple semantical description for dataset attributes (Uniform Content Descriptors). A resource registry model was created and implemented which allows users to search for data sets and services of interest. Currently efforts are underway for standardising access to spectral catalogues and for developing a query language for astronomical databases and common protocols for astronomical services, both on the Grid and as simple web services. A relatively new point of interest in the VObs concerns theoretical data products. This will be discussed in more detail below.

3 The Virtual Observatory in Germany: GAVO

Within Germany, the VObs efforts are represented by the German Astrophysical Virtual Observatory (GAVO) project, which is sponsored by the BMBF and is currently in its second phase of funding. GAVO was designed for building a platform to support modern astronomical research in Germany, and was recommended as one of the highest priority projects in the DFG Denkschrift Astronomie 2003 [1]. Several German astronomical institutes have partnered in order to contribute their own data archives and relevant expertise, to develop ideas and tools to store, manipulate, process, and exploit this collection of data archives, and to act as primary contact points for all people interested in using the VObs, ranging from professional scientists, teachers and students at high-schools or universities, to amateur astronomers.

In GAVO's pilot phase (GAVO-I, 2003-2006), work concentrated on four main areas: archive technology and publication, data mining and knowledge discovery in federated astronomical archives, theory in the VObs, and Grid-computing. The overarching goal was to support the process of scientific discovery in the era of huge distributed scientific databases. Due to the small team-size of GAVO-I we could only investigate what the requirements for this were, and follow and participate in the international efforts and implement relatively small prototypes.

The second funding phase of GAVO (GAVO-II) has started in the second half of 2006 and includes the following institutes: Astronomisches Rechen-Institut, Zentrum für Astronomie Heidelberg (ARI/ZAH), Universität Tübingen (UT), Technische Universität München-Informatik (TUM), the Max-Planck-Institut für extraterrestrische Physik (MPE) and the Max-Planck-Institut für Astrophysik (MPA) in Garching and the Astrophysikalisches Institut Potsdam (AIP). In GAVO-II we are advancing from prototyping and now produce more long-term services for the astronomical community in

Germany and abroad. Below we describe a few activitities and devote a special section to the work on the incorporation of theoretical data in the VObs that GAVO is spearheading.

3.1 Standards work

GAVO is active in the IVOA as the official representative of the German astronomical community, in particular in the efforts on developing data modeling standards and data access protocols, the development of an astronomical data query language and especially in the efforts dealing with standards for theory.

A practical application of these efforts is the implementation of IVOA standards on targeted datasets provided by the German community. Currently implementations exist of the Simple Image Access Protocol (SIAP, [2]) for the ROSAT[2] All-Sky Survey (RASS) images and pointed observations[3], of the Simple Cone Search (SCS, [3]) on the RASS source catalogues and photon event list[4] and of the Simple Spectral Access Protocol (SSAP, [4]) on the spectral follow up of the X-ray sources in the Chandra Deep Field South [5][5].

An interesting aspect of the SSAP protocol is its use of an involved data model for describing spectral data sets. The design of this model is still in progress in the IVOA. GAVO will investigate its applicability to spectra resulting from X-ray observations. These are rather different from the more standard optical spectra for example, requiring extra data products in their interpretation, which will require amendments to the protocol. We intend to investigate this further with the X-ray experts at the MPE, one of the members of GAVO.

In a similar project GAVO will publish theoretical stellar spectra produced by the Tübingen group via the SSAP protocol. Though the data products are similarly structured as observed spectra, simulated spectra have very different provenance (i.e. history how the data was obtained) which must be described by the model and handled by the SSAP protocol. This includes the fact that in general no position on the sky and no temporal information is

[2] http://www.mpe.mpg.de/xray/wave/rosat/index.php?lang=en

[3] http://www.g-vo.org/rosat/SIAP_start

[4] http://www.g-vo.org/rosat/pages/RASSConeSearches.jsp

[5] http://www.g-vo.org/ssa/

available, and also the meaning of uncertainties/errors is very different from the observed counterparts.

3.2 ARI/ZAH Virtual Observatory Expertise Center

One of the core goals of GAVO-II is to create a VObs data and expertise center at ARI-ZAH. One of its tasks will be to host smaller datasets (images, spectra and catalogues) and/or data set descriptions (metadata) from German institutes who do not have the resources and/or expertise to maintain an online presence themselves. We will develop tools to assist users in uploading their resources and most importantly to describe these according to the relevant IVOA standards. The site will also serve as the portal through which these data sets can be accessed with user friendly web pages implementing the standards. It is also planned to implement value added services such as source catalogue cross-matching to these datasets, something that will be facilitated by the coordinated storage of the metadata and the standard access protocols to the underlying data.

A first result in this effort is based on a tool which we call the *FITS Ingester*. This tool facilitates the convenient mapping of archives of FITS[6] files to IVOA standards. It is based on a two stage process. In the first stage the catalogue's metadata are completely ingested into a special purpose relational database. This database stores the FITS keywords in a relational format which allows, in the second stage, to map the key words to a relational representation of the IVOA standard data models. This mapping can now be represented in many cases as a single SQL statement. The advantage of this over other methods is that SQL is much more easily used in discussions with the domain experts who will always have to be involved in the process.

3.3 Other efforts, collaborations and outreach

At AIP in Potsdam the results from the RAVE survey are housed and made available through VObs means. GAVO is collaborating with the AstroGrid-D[7], the project within the D-Grid initiative developed by the astronomical community. In this context we plan to further develop tools initiated in GAVO-I for cross-correlating two or more different catalogues following a peer-to-peer approach developed at TUM in collaboration with the Johns Hopkins University (JHU) in Baltimore, USA. Another project in this effort is the continuation of a cluster finder tool that combines X-ray observations

[6] http://fits.gsfc.nasa.gov/standard21b.html

[7] http://www.gac-grid.de/

with optical galaxy catalogues[8]. Important in GAVO-II will also be outreach to the German astronomical community. This will be done through workshops and consultation. The VObs expertise center to be established at ARI will play an important role in this effort.

4 Theory in the Virtual Observatory

One of the original core interests of the GAVO project is the publication of results of computer simulations in the virtual observatory. What it means to "publish simulations" in a VObs context is not obvious, but one of the main requirements of the Theoretical VObs is to be a bridge between observational and simulation archives by enabling access to either in a uniform manner. This is often referred to as the Theory-Observational Interface.

4.1 Standardisation: IVOA theory interest group

The comparison between publication of simulated archives and observational archives (for which a lot of work has been done already), is not straightforward. One reason is that most current VObs standards for accessing observational archives such as the Simple Cone Search (SCS) and the Simple Image Access Protocol (SIAP) are based on positional searches on the sky. The reason that positional queries play a key role in the "observational VObs" is that most correlations between observations in different wavelength regimes start with matching by position. An important task people hope to achieve is the identification of sources in different catalogues as being different observations of the same object.

Positional queries of this type are hardly ever relevant for querying simulations. The theory-observational interface must therefore be based on correlations of a different nature. They will most likely be based on common identification of the class of an object and its physical parameters. An example of this might be mass and temperature of a galaxy cluster simulated using a hydrodynamical code, or the relative orientation of a pair of galaxies involved in a merger event. Another possibility for both simulation-simulation and simulation-observation correlations is offered by statistical quantities. Examples of these are luminosity functions in a cosmological simulation, structural properties of simulated galaxies or clusters of galaxies or the properties of observed versus simulated spectra. Such observations, written down

[8] http://www.g-vo.org/portal/tile/products/services/clusterfinder/index.jsp

in a whitepaper [6] urged the IVOA to form a special interest group for theory, which is chaired by GAVO. Its aim is to[9]

- Provide a forum for discussing theory specific issues in a VObs context.
- Contribute to other IVOA working groups to ensure that theory specific requirements are included.
- Incorporate standard approaches defined in these groups when designing and implementing services on theoretical archives.
- Define standard services relevant for theoretical archives.
- Promote development of services for comparing theoretical results to observations and vice versa.
- Define relevant milestones and assign specific tasks to interested parties.

Currently efforts are underway to define a data access protocol for simulations together with a model for the metadata describing these simulations and an effort at extending the *semantic vocabulary*[10] with terms and concepts particular to simulation data products and their underlying theoretical models.

4.2 Virtual telescopes

GAVO has also been active in the field of theory beyond the IVOA process. It is important to test ideas for standardisation of online data and service publication in practice by implementing prototypes of the proposals. In particular, GAVO has created a number of applications implementing the concept of the "virtual telescope". This is a tool mimicking observations of real objects by virtually "observing" simulation results. These virtual telescopes produce results that can be directly compared to real observations, both in content and in format. The idea behind this concept is that it is easier to produce such products from simulations, containing the projection and instrumental effects that are introduced in a realistic observation, than to take such effects out of the real observation, for the goal to produce products that could be directly compared to the theoretical data. This is an important tool in the theory-observational interface.

GAVO has created three prototype implementations that approximate the virtual telescope concept. The Planck group at the MPA[11] has offered a simulator for the cosmic microwave background, which allows users to se-

[9] http://www.ivoa.net/twiki/bin/view/IVOA/IvoaTheory

[10] http://www.ivoa.net/twiki/bin/view/IVOA/IvoaSemantics

[11] http://planck.mpa-garching.mpg.de/

lect cosmological and observational parameters and produces results that are similar to those the Planck satellite will produce[12]. This web application is designed according to a simple architecture that is easily applied to other legacy applications. Figure 1 illustrates two other such tools built by GAVO producing simulated images of object catalogues and X-ray clusters, extracted from a large cosmological simulation and a set of hydrodynamical simulations, respectively.

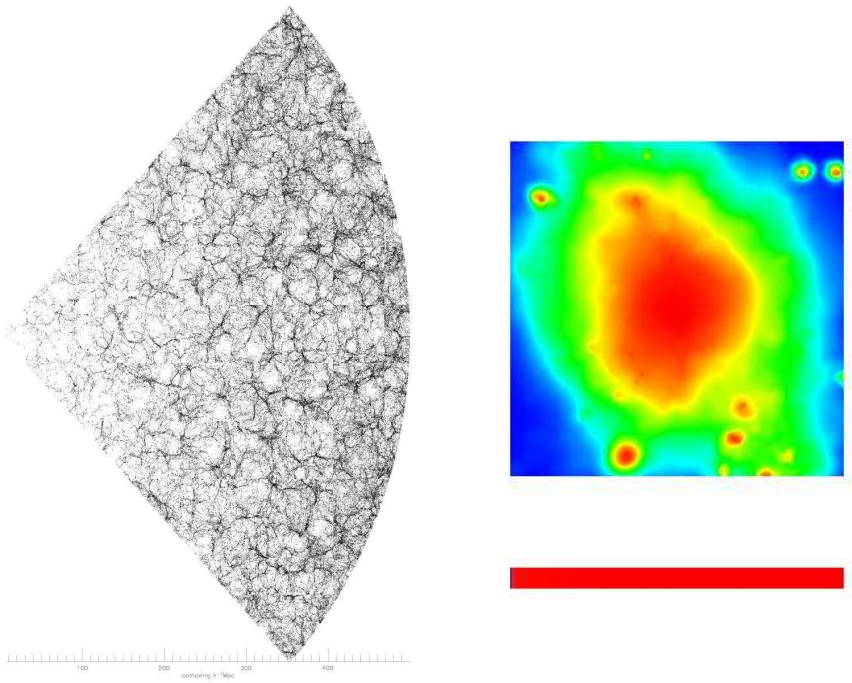

Figure 1: Two examples of results of virtual telescopes on the GAVO website. Left a synthetic galaxy catalogue, extracted from a cosmological simulation (http://www.g-vo.org/mpasims/MoMaf2), right a projected X-ray surface brightness of a simulated galaxy cluster (http://www.g-vo.org/hydrosims/).

These applications are pure prototypes, in that so far they are hardly producing interesting scientific results. Partly this is for computational reasons: the GAVO hardware is not able to house and publish datasets that are large enough to serve the requirements of the users. Also, to produce results of interest requires higher resolution "virtual observations", with consequently higher CPU demands. For this reason GAVO will collaborate with the As-

[12] http://www.g-vo.org/planck

troGrid-D, and thus get access to the infrastructure of the German D-Grid project. This will allow us to store complete datasets, sampling a much larger range of simulation parameters. It will also give us access to more and more powerful CPUs, so that we can produce more realistic results, and it will allow us to publish more complex and realistic telescope models so that real science can be done with them soon.

4.3 Theory archives: the Millennium database

A major and very successful effort of GAVO dealt with the archiving and publication of simulation results in a relational database. Contrary to observational archives, this technology is very new for simulation archives. One reason to pursue this was to evaluate whether scientific analysis can benefit from relational database technology. Another reason was to attempt the publication of simulation products in ways that are similar to those of observational archives, especially also since the IVOA query language is very much influenced by SQL. As the subject of publication we used the post-processing products of the largest cosmological simulation run to date, the so called *Millennium* simulation [7], run at the Rechenzentrum Garching, and provided to us by the MPA. GAVO has taken up the task of data modeling, database management and the creation of a web site giving access to the database[13].

The data products that can be queried in the database are halo catalogues and density fields derived from the raw particle distributions. Furthermore we have model galaxy catalogues calculated using semi-analytical techniques [8]. We provide catalogues for the individual simulation outputs, as well as catalogues created using simulated observations. The modeling for this database was especially interesting as it provides new features that are generally not encountered in observational databases. In particular we required relational models for the evolution of objects in so called merger trees and also we needed new models for multi-dimensional spatial indexes. The storage model for trees in relational databases is new and can be generalized to other tree structures [9].

The user support for database access is mirrored on that from the popular SkyServer web site[14] of the Sloan Digital Sky Survey. As a particular feature it provides registered users with their own database which they can use to

[13] http://www.g-vo.org/Millennium

[14] http://cas.sdss.org/dr5/en/

store query results on the server. This MyDB is a prototype for the VOSpace concept being developed in the IVOA grid working group [10].

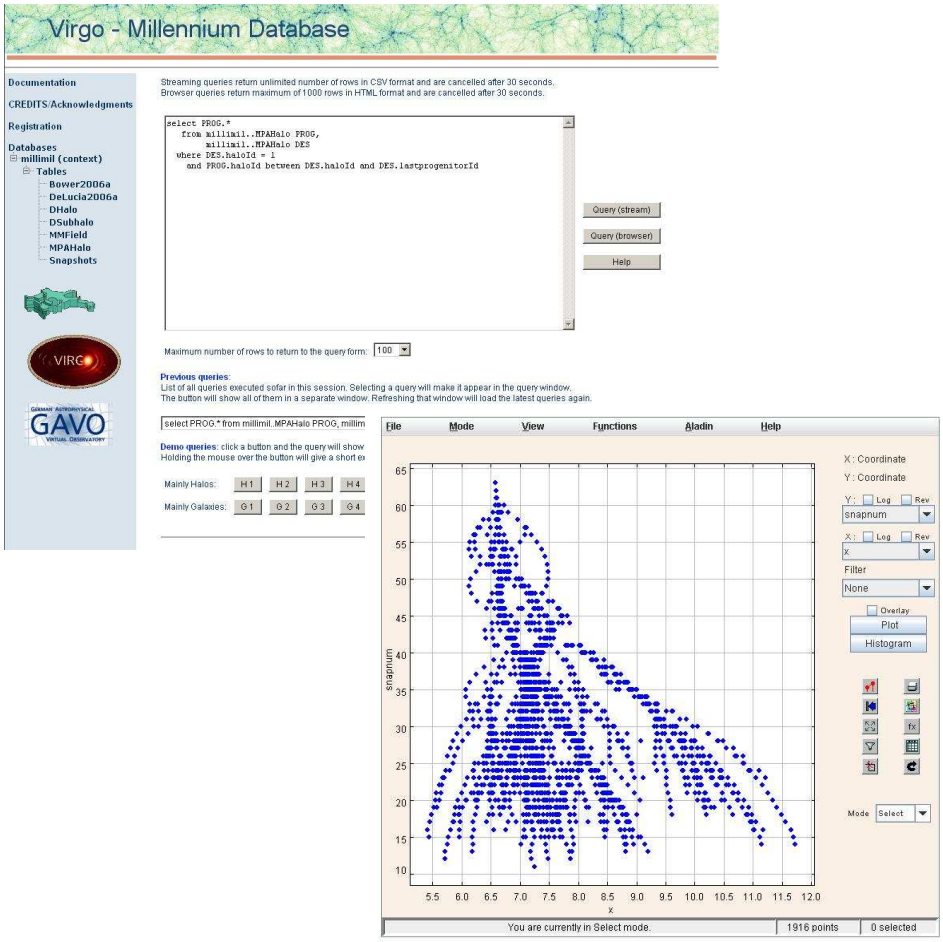

Figure 2 Screen shot of the public web site giving access to a small version of the full Millennium database. This small version has exactly the same design as the large version and allows users to test their queries before applying it to the full database. Users can view results of the queries using a VOPlot applet created by VO-India.

The website has been available in a public and private version since August 2006 and was published in the online Los Alamos preprint server[11]. Figure 2provides a screen shot of the web site and a typical result visualized using the VOPlot[15] applet, directly available from the web site. Currently, for this service we have some 120 registered science users at GAVO and have had over a million individual, successful SQL query requests. Such large num-

[15] http://vo.iucaa.ernet.in/~voi/voplot.htm

bers of requests are only achievable when web based tools such as this are made explicitly available in a form that allows use from within scripts. This was a major requirement of our web service design and has clearly paid off. Especially in small teams, where advanced graphical software skills are rare, it is an easy and cheap alternative to allow users to download data into their tools of choice, as opposed to developing such tools oneself.

Acknowledgements

GAVO is supported by a grant from the German Federal Ministry of Education and Research (BMBF) under contract 05 AC6VHA. We thank our collaborators of MPA in the various projects: Martin Reynecke from the Planck group, Klaus Dolag for the hydro simulations and visualization tools, Jeremy Blaizot for his Momaf code. We thank Volker Springel, Simon White, Gabriella de Lucia and Jeremy Blaizot, the Virgo consortium for their participation the creation of the Millennium database and Alex Szalay (JHU) for discussions about its design.

References

[1] *Denkschrift*: *Status und Perspektiven der Astronomie in Deutschland 2003-2016*, Deutsche Forschungsgemeinschaft (Editor), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim 2003. ISBN: 3-527-27220-8
[2] http://www.ivoa.net/Documents/latest/SIA.html
[3] http://www.ivoa.net/Documents/latest/ConeSearch.html
[4] http://www.ivoa.net/twiki/bin/view/IVOA/SsaInterface
[5] Szokoly, G. et al, 2004, The Astrophysical Journal Supplement Series, Volume 155,
pp. 271-349.
[6] Lemson, G. and Colberg, J., 2004, *Theory in the VO*, IVOA whitepaper, http://www.ivoa.net/pub/papers/TheoryInTheVO.pdf
[7] Springel V. et al. 2005, Nature 435, 629
[8] DeLucia, G. & Blaizot J. 2006, Monthly Notices of the Royal Astronomical Society, Volume 375, pp. 2-14.
[9] Lemson G. and Springel, V. Astronomical Data Analysis Software and Systems XV ASP Conference Series, Vol. 351, Edited by C. Gabriel, C. Arviset, D. Ponz, and E. Solano. San Francisco: Astronomical Society of the Pacific, 2006, p.212
[10] http://www.ivoa.net/twiki/bin/view/IVOA/VOSpaceHome
[11] Lemson, G. and the Virgo Consortium, 2006,
http://xxx.lanl.gov/abs/astro-ph/0608019