

# MySpace: Personalized Work Space in AstroGrid

C.L. Qin<sup>1</sup>, A.C. Davenhall<sup>2</sup>, K.T. Noddle<sup>3</sup> and N.A. Walton<sup>4</sup>

## Introduction

### Abstract

MySpace is a component of AstroGrid, a system for accessing astronomical archives which is being developed in the UK. MySpace provides AstroGrid users with scratch space for storing temporary or permanent datasets, typically the results of queries submitted to large databases, and other transient files. The novel feature of the MySpace system is that the scratch space is geographically dispersed, typically with stores at the various sites hosting AstroGrid services. Users can access and navigate MySpace seamlessly and easily, the network details of the individual stores being hidden. MySpace is a fully integrated component of the AstroGrid system, written in Java and communicating via Web services. It is under active development and its current state and future plans are described. Functionality similar to that of MySpace seems likely to be a reasonably common requirement in distributed systems, and the experience gained with MySpace should be applicable elsewhere.

### Background

Astronomy is an observational science which progresses by accumulating observations of celestial phenomena. Data archives of past observations, both from systematic sky surveys and disparate observations of individual objects are important. Such data have been collected for hundreds, indeed thousands, of years. However, with modern telescopes and detectors the volume of data being acquired is vastly greater than hitherto with those volumes increasing all the time. Numerous data archives are available, though they are extremely heterogeneous both in content and data format. Moreover, they are usually accessed using bespoke software which is specific to the archive. Such software is often idiosyncratic and provides only limited functionality. Also, the archives are rarely interoperable. AstroGrid<sup>5</sup> is a UK e-Science project to address some of these deficiencies. The problem of accessing astronomical archives is not specific to the UK, of course, and AstroGrid is co-operating with the IVOA<sup>6</sup> and a number of similar initiatives overseas in the development of the so-called global 'Virtual Observatory'.

### Why MySpace ?

Accessing astronomical archives basically consists of performing database searches to yield files of results in e.g. VOTable<sup>7</sup> format. Astronomers may wish to download such files to their own computers for further analysis, or to perform further queries on the results, either in isolation or in combination with searches of other distributed archives. However, it is anticipated that increasingly the volume of data will be such that moving them around the network will become increasingly costly and time consuming. This is the drive to Grid technologies which seek to move the processing to the data – a key element of AstroGrid. In any event, work space is required for storing

---

<sup>1</sup> Dept. of Physics and Astronomy, University of Leicester, University Road, Leicester, LE1 7RH.

<sup>2</sup> Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ.

<sup>3</sup> Dept. of Physics and Astronomy, University of Leicester, University Road, Leicester, LE1 7RH.

<sup>4</sup> Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge, CB3 0HA.

<sup>5</sup> See URL: <http://www.astrogrid.org/>

<sup>6</sup> International Virtual Observatory Alliance. See: URL: <http://www.ivoa.net/>

<sup>7</sup> See URL: <http://cdsweb.u-strasbg.fr/doc/VOTable/>

such files, and software is needed for managing and accessing these files. MySpace is AstroGrid's system for this purpose. Astronomers use one MySpace interface to access geographically dispersed data sets and work on them.

## Design

### Concepts

We briefly describe some of the concepts used in the MySpace system.

<b>Data Source</b>	A file or Database table that is stored within MySpace Server. In this paper, we refer to it as VOTable.
<b>Metadata</b>	Data that is stored within MySpace Manager, such as data source size, data source creation date, owner of the data source, location of the data source etc.
<b>MySpace Registry</b>	Sub-Component of MySpace Manager where Metadata are held.
<b>Community</b>	A single physical location in the federation of data archives that form AstroGrid. Each Community will have only one MySpace System installed.
<b>Containers</b>	Hierarchical view of many structured Data Sources.
<b>Expiry period</b>	Each MySpace System has a configurable integer representing the period that a data source can exist.
<b>Publishing</b>	Some of the data sources may be allowed to persist indefinitely. The MySpace System will move these data sources to a permanent storage area and make them publicly available to all AstroGrid users. Only Administrators can delete published data sources.

### Design overview

MySpace's design goal is to use open-source software products to build distributed, collaborative and centrally-controlled data storage/processing system, that provides a simple and user-friendly interface. MySpace will eventually be deployed at institutions around the globe.

It is envisaged that Astrogrid will include several MySpace systems, providing work space for different purposes. For example, one use is to provide large caches close to DatasetAccess (data archives) for the storage of intermediate results. Another use is to provide astronomers with their own longer-term work space, where they can keep results and 'work-in-progress'. However, each MySpace system has the same basic structure, irrespective of how it is being used: it comprises one *MySpace manager* and one or more *MySpace servers*. All of these components can be geographically dispersed, both from each other and from other components of the AstroGrid system.

MySpace is primarily intended to provide (temporary) work space and thus it is necessary to retire old files. Each MySpace system has a configurable *expiry period* that is applied to all data sources created by a user of a MySpace system. Towards the end of a data source's expiry period, the user will be warned that the data source is about to expire (e.g. by displaying it in red in MySpace Explore), and the user then can choose to 'extend lease' of this item, 'publish' it, or delete it.

### MySpace System Components

#### Components

We now describe the main components of the MySpace system (see Figure 1):

**MySpace Manager:** MySpace Manager is invoked by users when they use MySpaceExplor to access or manipulate data sources in MySpace systems. It incorporates a *registry* comprising the Metadata describing the data sources in the MySpace system. Typically there will be one MySpace Manager per Community.

**MySpace Server:** MySpace Servers are repositories and are where the Data Sources are kept. They are invoked by the MySpace Manager to copy, delete etc. a named Data Source. Users do not know about, or interact directly with, MySpace Servers. From a usage perspective, there are two different kinds of MySpace Server:

Cache MySpace Server is the MySpace Server that stores transient Data Sources (e.g. The results from intermediate queries).

Community MySpace Server is the MySpace Server that stores Data Sources for users within its Community.

A single MySpace Server can act both as a Cache Server and a Community Server. Data Sources have a different expiry period depending upon the user to which they belong. In general, cached data sources have shorter expiry period than data sources stored on Community Servers.

**MySpace Explore:** This is a user interface displayed within a web browser that allows astronomers to interactively browse/edit their data sources within MySpace Systems. A user sees data sources in a MySpace Explore in a hierarchical tree view format similar to a Unix directory structure (though the analogues of directories are called 'Containers'). This notional hierarchy of Containers is distributed across the (geographically dispersed) MySpace Servers. MySpace Explore should display a simple and intuitive representation of the hierarchy. The notional hierarchy of containers does not correspond to the actual directory structure on MySpace Servers (which is flat in Iteration Two). Rather, the structure is stored and maintained by MySpace manager.

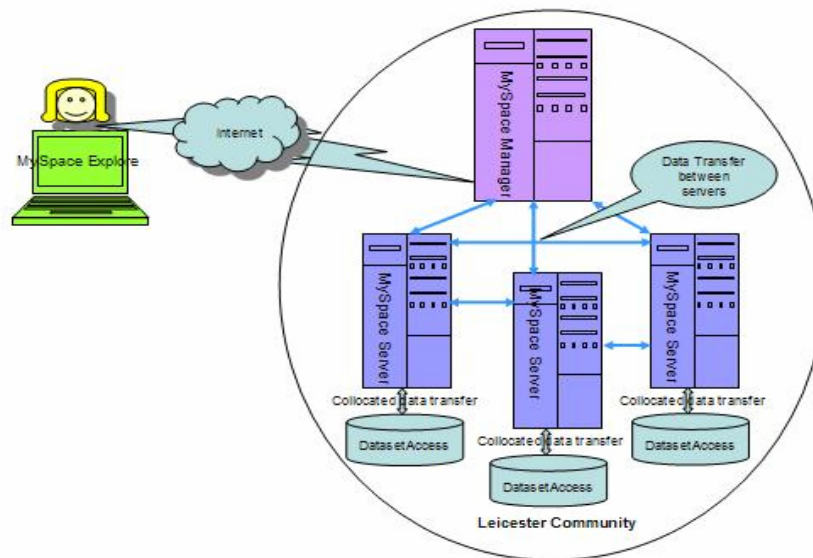


Figure 1. An overview of a fully installed MySpace System (DatasetAccess is another sub-system within AstroGrid system and is shown here for clarity.)

### **Technology Choices**

MySpace is a component of AstroGrid and is being implemented using exactly the same technologies as the rest of the system. AstroGrid is being written in Java and its components communicate via Web services (this is true of the internal components within a MySpace System as well as MySpace communicating with external components). Insofar as practical, only non-commercial packages are used in

AstroGrid (for example, those licensed using GPL), so that AstroGrid can be deployed without incurring license fees. Web services are implemented using Axis<sup>8</sup> from the Apache Software Foundation<sup>9</sup>.

Astronomers do not install any bespoke AstroGrid software on their desktop computer in order to access AstroGrid. Rather, they use a standard Web browser to access an AstroGrid portal. This portal is implemented using Cocoon<sup>10</sup>, also from Apache. Strictly speaking, MySpace Explore is part of the portal and is also written in Cocoon using MVC pattern.

## Current Implementation and Supported Functionalities

AstroGrid is being developed in a sequence of three-month iterations. The software release following the completion of Iteration Two on 31 July 2003 included the first version of MySpace System. This version had the following functionality:

- ⑧ create container,
- ⑧ import file,
- ⑧ copy file,
- ⑧ move file,
- ⑧ export file,
- ⑧ delete file or container,
- ⑧ look up details of a single file,
- ⑧ look up details of all files matching an optionally wild-carded file name,
- ⑧ a rudimentary MySpace Explore.

The basic framework is present and we plan to continue to develop the system.

## Enhancements and Open Issues

During the remaining iterations of AstroGrid additional functionality will be added to MySpace. Some of the enhancements are conceptually straightforward and merely involve implementing functions which are not yet written, for example allowing copying between MySpace servers and improving the facilities for file import and export. Similarly MySpace Registry is currently implemented as a flat file and we plan to replace it with a set of database tables.

Conversely, some of the desired features pose more substantial problems and require further thought. These open issues include the following items.

- ⑧ Currently the MySpace Server stores files. However, in AstroGrid results are usually generated by querying databases and results of such queries can be represented as new database tables as well as files. Thus, we hope to incorporate DBMS tables of results into the MySpace Server.
- ⑧ Currently the access control to data sources in MySpace is rudimentary. In the wider AstroGrid system the access which a user enjoys to facilities and resources is controlled by a complex set of permissions based on his membership of one or more groups of collaborators. We plan to tie this wider system into the access control of individual data items within MySpace.

---

<sup>8</sup> see URL: <http://ws.apache.org/axis/>

<sup>9</sup> see URL: <http://www.apache.org/>

<sup>10</sup> see URL: <http://cocoon.apache.org/>

- ⑥ Transferring large VOTables using Internet. There are several transfer protocols (Http/Ftp/Sftp), each with its own limitations, either in terms of handling the transfer of large (sometimes gigabyte-sized) files across internet or security beyond firewalls. We are currently working on a solution to this.
- ⑥ In order that free access to (sometimes huge) distributed data sources does not cause network resources to become overloaded, we would need to investigate the possibility of smart algorithms for controlled replication of (parts of) data sources.

Although these are challenging, we are planning to address above new features/issues in future iterations.