Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Resource Registries for the Virtual Observatory

Raymond Plante

National Center for Supercomputing Applications (NCSA), University of Illinois Urbana-Champaign, Urbana, IL 61801

Gretchen Greene, Robert Hanisch

Space Telescope Science Institute (STScI)

Thomas McGlynn

Goddard Space Flight Center, NASA

William O'Mullane

Johns Hopkins University (JHU)

Roy Williams

California Institute of Technology (Caltech)

Ramon Williamson

National Center for Supercomputing Applications

Abstract. Data discovery will be a core utility of the Virtual Observatory (VO). Registries that contain high-level descriptions of resources such as archives and services are essential for making data discovery efficient in a distributed environment. We review a framework architecture for VO registries currently under development within a International Virtual Observatory Alliance (IVOA) working group. We present an overview of a prototype implementation of the framework developed as part of the National Virtual Observatory (NVO) project. We illustrate how institutions can publish descriptions of their resources within their own registries. Other registries specialize in harvesting these descriptions to centralized locations where users may search them. We show how our prototype registry supports the NVO's first publicly released service, a Data Inventory Service.

1. A Common, Global Approach to Resource Discovery

Registries are an important linchpin for the Virtual Observatory (VO): they provide the mechanism for discovering the resources available to applications. By resources, we primarily mean data and services, but we can view other things

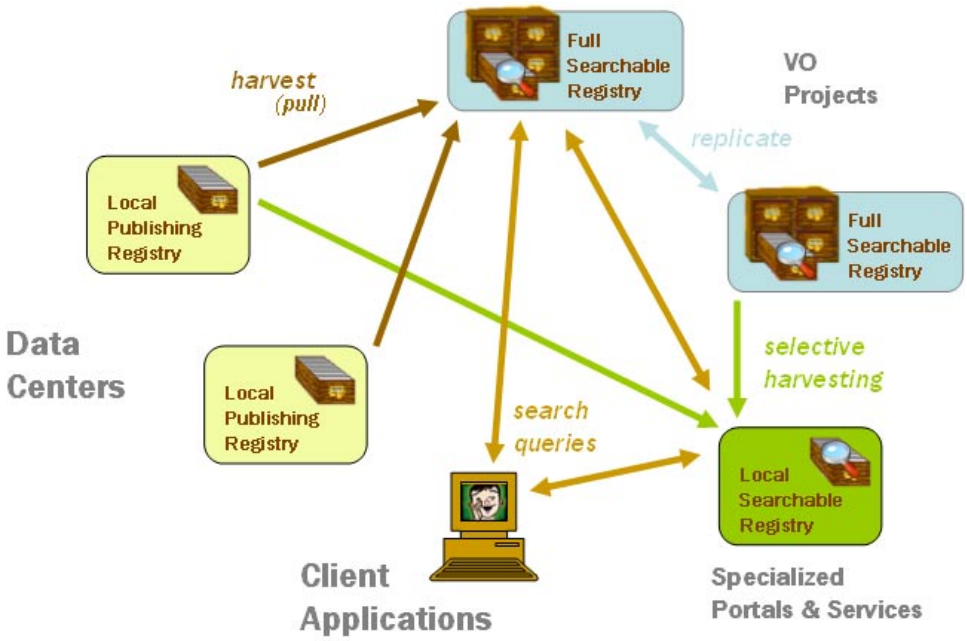

Figure 1. A distributed model for registries.

as "resources", such as organisations, projects, and software. A *registry* is simply a list of resource descriptions, expressed in terms of structured metadata to enable automated processing and searching.

The International Virtual Observatory Alliance[1] has established a Registry Working Group that is actively developing a common framework for resource registries. This framework must address a number of requirements. The most important function of the framework is to allow users to select resources likely to pertain to a scientific question based on various characteristics: e.g. the type of resource (catalogs, image archive, educational resources), coverage in space, time, and frequency, and where the data comes from and who curates it. The framework must recognize that resources come and go, and so it must adapt accordingly. A distributed system not only avoids depending on a single point of failure, it allows for multiple views of what is available in the VO. The framework must also help preserve the data providers' control over their data by letting them control what gets registered, what's included in the description, and when it gets updated. In particular, registries should integrate well with a provider's existing resource management tools (e.g. GLU). Finally, the framework must allow extension to describe new types of resources in the future.

The distributed model for registries developed by the Registry Working Group that addresses these requirements is illustrated in Figure 1. It features three types of registries. The *full searchable registry* is intended for use end-user applications and contain all resource descriptions known to the VO. These

[1] `http://www.ivoa.net`

registries are filled by collecting descriptions from many *local publishing registries* through a process called "harvesting." Publishing registries are distinguished from searchable ones in that, as the name implies, they do not support searching; they simply expose resource descriptions to the searchable ones (see Williamson & Plante 2004). There can be multiple full searchable registries; thus, there needs to be a "replicating" process to keep them synchronized. The third type of registry is the *local searchable registry*. These are intended for searching by end-user applications but do not contain everything that is known to the VO. Rather, they might specialize in a particular type of resource or scientific topic, e.g. resources related to supernova research. Thus, these might carry out "selective harvesting" to populate themselves.

2. The NVO Prototype

The US-based National Virtual Observatory (NVO) project has put together a prototype implementation of this framework. The purpose of the prototype is to support a real end-user application, the Data Inventory Service (DIS; McGlynn et al. 2004). This service gives user a listing of what data is available for some location of the sky; the first step in this inventory service is to search a registry for services serving catalog and image data.

2.1. Resource Metadata

The IVOA Registry Working Group is developing a standard set of metadata for describing resources[2]. The standard comes in two parts. The first is a prose document that defines the basic concepts (Hanisch et al. 2004). The second is a set of XML Schemas used to encode the concepts into XML.

The XML version of the metadata takes advantage of XML Schema's object-oriented modeling capabilities. The core concepts are defined in the "VOResource" schema. At the center of the model is a generic `Resource` which contains metadata concepts that apply to all resources. More specific "resource classes" extend `Resource` by inheriting the core metadata and adding additional, specialized concepts. Examples of specific resource classes include `Organisation`, `DataCollection`, `Service`, and `Registry`. These extensions are defined in separate schemas so that applications can pick and choose which extensions to use. With these schemas, data providers can describe various kinds of research organisations (e.g. data centers, observatories, and missions), their data collections and archives, and a variety of services. In particular, not only can they describe emerging VO standard services like Cone Search and Simple Image Access,[3] they can also describe their existing browser-based and CGI services.

2.2. Publishing Registries

We established two prototype publishing registries: one at Caltech and one at NCSA. Each featured a web page form that data providers could use to register

[2] `http://www.ivoa.net/twiki/bin/view/IVOA/IVOARegWp03`

[3] `http://www.us-vo.org/standards.html`

their resources. The motivation behind the development of these registries was two-fold. First, we needed to build a listing of all known Simple Image Access implementations, which we had not had before. Second, we wanted to use these prototypes to develop techniques for making the registration process easier for data providers. A more detailed discussion of the NCSA registry is described in this volume by Williamson and Plante (2004). Both implementations store the descriptions entered by users as XML documents. These descriptions are exposed to a searchable registry via the Protocol for Metadata Harvesting using Open Archives Initiative[4] (OAI; see discussion in Williamson and Plante 2004).

It's worth noting that the specific manner in which a data provider publishes data descriptions will likely depend on the number of resources being described. If the provider only has a few resources that are fairly static, then the easiest thing to do will likely be to just go to a site like those set up at NCSA and Caltech and fill out the forms; such "public" registry sites would manage the descriptions of the resources on behalf of the provider. If the provider has a moderate number of resources to register (say, a few tens) that may be changing somewhat with time, she may wish to take greater control over the descriptions; in this case, they may install a generic registry (e.g. VORegistry-in-a-Box, Williamson and Plante 2004) at their own site. If the provider has a very large number of resources, which are perhaps highly dynamic, he may wish to implement the OAI interface himself, perhaps plugging directly into his existing database or other tools used for managing resources.

2.3. The Searchable Registry

A prototype searchable registry was set up (jointly) at JHU and STScI (described in further detail in this volume by Greene et al. 2004). It collects resource descriptions from the publishing registries at Caltech and NCSA and loads them into a relational database. Searching is provided to client applications through a web interface. In this prototype, the interface is fairly specialized to the needs of the Data Inventory Service; however, we expect this to be standardized through the course of development by the Registry Working Group.

At the moment, there is no automated mechanism for regularly polling the publishing registry for new records. Instead, the web service interface has a special operation for initiating the harvesting process.

References

Greene, G., O'Mullane, W., Hanisch, R., & Gaffney, N. 2004, this volume, 285.

Hanisch, R., Linde, T., Plante, R., Richards, A., Auden, E., Noddle, K., Greene, G., & O'Mullane, W. 2004, this volume, 273.

McGlynn, T. Lee, J., Hanisch, R., O'Mullane, W., Greene, G. 2004, this volume, 319.

Williamson, R. & Plante, R. 2004, this volume, 334.

[4] `http://www.openarchives.org`