IVOA Search Interface Study

D. J. Burke

May 5, 2008

Contents

1 Introduction **4**

2 Faceted Navigation **4**

3 Prototype **7**

4 Main project **12**
4.1 Choice of browser 12
4.2 VizieR search application 13
4.2.1 Using the VizieR browser 14
4.2.2 Lessons learnt 14
4.3 DataScope search application 18
4.3.1 Lessons learnt 18

5 Conclusions **19**
5.1 Data access 19
5.2 Data modeling 20
5.3 Faceted browsing 20
5.4 Future Work 21

A A demonstration of the VizieR registry browser **26**
A.1 Starting with a text search 26
A.2 Starting with a class 27

B Namespaces **39**

C The IVOA Registry **40**

D The prototype browser **42**
D.1 Registry data . . . 42
D.2 Facets . . . 46

E Resource Description Framework **50**

F RDF modeling of IVOA Registry data **53**
F.1 VizieR Registry Browser . . . 53
F.2 Comments and notes on the conversion . . . 55
F.3 Mapping between IVOA and RDF schema . . . 57

G Longwell configuration for IVOA Registry data **61**
G.1 Defining the data types that can be searched . . . 61
G.2 Facet configuration . . . 62
G.3 Data Display . . . 63

H RDF modeling of NVO DataScope Application data **68**
H.1 Accessing the DataScope data . . . 68
H.2 Modeling the metadata . . . 68
H.3 Modeling the data . . . 71
H.4 Comments and notes on the conversion . . . 78

Listings

1 Resource from STScI Service Locator . . . 42
2 Resource from VizieR service . . . 43
3 JSON version of `ivo://CDS/VizieR/J/A+AS/135/511/table1` . . . 47
4 N3 representation of example RDF statement . . . 51
5 Declaring classes and instances in RDF . . . 52
6 RDF/N3 version of `ivo://CDS/VizieR/J/A+AS/135/511/table1` . . . 54
7 Example facade for a Tabular Sky Service Resource . . . 61
8 Facet definition for Tabular Sky Service records . . . 62
9 Fresnel lens definition for Tabular Sky Service Resources . . . 63
10 RDF/N3 representation of HST Previews metadata . . . 70
11 RDF/N3 representation of HST Previews table data . . . 73
12 RDF/N3 representation of UCD values . . . 75

List of Figures

1 Examples of facets . . . 5
2 Facet contents depends on context . . . 6

3 Date formats when used in a facet . . . 7
4 Current list of filters applied to a search . . . 8
5 Examples of faceted browsing in Astronomical databases . . . 9
6 VizieR search results . . . 9
7 STScI registry search results . . . 10
8 The prototype application . . . 11
9 Starting point of `Longwell` searches for VizieR records . . . 15
10 Searching the VizieR data base using `Longwell` . . . 16
11 Demo: Start page of the registry browser . . . 28
12 Demo: Starting with a text search . . . 29
13 Demo: A single match . . . 30
14 Demo: The Chandra facility . . . 31
15 Demo: Combining text and facet filters . . . 32
16 Demo: Following the Bibliographic Code link . . . 33
17 Demo: The VizieR web page for a table . . . 34
18 Demo: Expanding a facet search . . . 35
19 Demo: The intersection of multiple facet restrictions . . . 36
20 Demo: A new search using the set of Resources . . . 37
21 Demo: A new search using the set of VizieR keywords . . . 38
22 Simple RDF graph . . . 50
23 `Longwell` display of a `vor:TabulasSkyService` record . . . 66
24 `Longwell` display of a `vor:BibliographicReference` record . . . 67
25 Example of resources found by a DataScope search . . . 69
26 Starting point of `Longwell` searches for DataScope resources . . . 78
27 `Longwell` display of a `vor:DatascopeResource` record . . . 79
28 `Longwell` display of a `vor:VOTable` resource . . . 80
29 `Longwell` display of a `vor:VOTableColumn` resource . . . 80

List of Tables

1 Namespaces and prefixes used in this document . . . 39
2 Relationship of facet name to the VizieR data . . . 47
3 Mapping from OAI-PMH to RDF for VizieR records . . . 59
4 Number of items created from the VizieR data set . . . 60
5 Values of the instrument field of the DataScope resources . . . 72
6 Data types found in the VOTables . . . 77
7 Unit conversion used for VOTable . . . 77

Abstract

The use of faceted navigation was investigated for searching two Astronomical data sets accessed using International Virtual Observatory Alliance protocols: the VizieR catalog and data within 0.1 degrees of the Antennae galaxies. Demonstration applications were created for both samples, and areas for future study were identified.

1 Introduction

The Astronomical community has recently seen a huge increase in the number of on-line resources available to them. The advent of large surveys — such as the Sloan Digital Sky Survey and the Great Observatories Origins Deep Survey ([27], [9]) — is one of the primary causes of this growth. There has also been significant work by projects such as the NASA Astrophysics Data System, NASA/IPAC Extragalactic Database, High Energy Astrophysics Science Archive Research Center, and the VizieR Catalog service ([1], [19], [10], [31]) to use the standards created by the International Virtual Observatory Alliance ([13]) and other bodies to provide programmatic access to their contents. It is now possible for Astronomers to access a large fraction of the resources they need — be it data, documentation, or code — without leaving their computer; however, present interfaces are complex and baroque and provides a significant barrier to their use by many researchers. As the standards have matured, and the amount of data has increased, it is an opportune time to work on improving the search mechanisms, with the aim of providing a seamless integration of the "Virtual Universe" into the work flow of Astronomers.

2 Faceted Navigation

In many current Astronomical search engines, the User Interface contains a single set of options for constraining the search; once these choices have been made the results are provided with limited, or no, means for the user to directly refine or expand their search (e.g. Figures 5a and 6). The use of faceted navigation for such searches would provide this flexibility as well as improve the user's understanding of the available data.

Faceted navigation of data was pioneered by the Flamenco Search Interface project ([6]), which was designed to help users navigate in a flexible manner through large quantities of data without getting lost. By exposing category metadata in the form of hierarchical facets, users are able to refine or expand their search. This interface technique is used to some degree by many Electronic Commerce web sites ([35]) and Figure 1 shows four facets that might be shown in a search of Astronomical data tables. In this particular example the faceted navigation is context dependent, in that the facet values list the number of matching items and are updated to reflect the currently selected sample; this gives the user immediate visual feed-back on how their choices will change the search results.

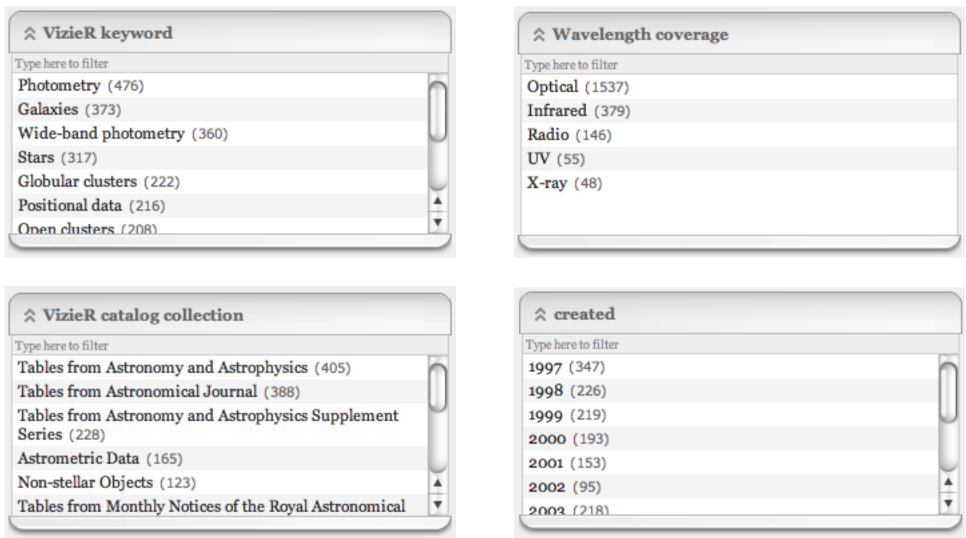


Figure 1: Four example facets — "Keyword", "Wavelength", "Catalog", and "Created" — with their corresponding values, such as "Photometry", "Optical", "Tables from Astronomy and Astrophysics", and "1997". Each facet value includes the number of matching items, which means that a user can see how the data is distributed within a facet. In most of the facets the values are ordered by decreasing frequency, but other options are possible. This example uses the `Longwell` browser, which is described in section 4.1, to filter and display the results; other User Interface designs are possible.

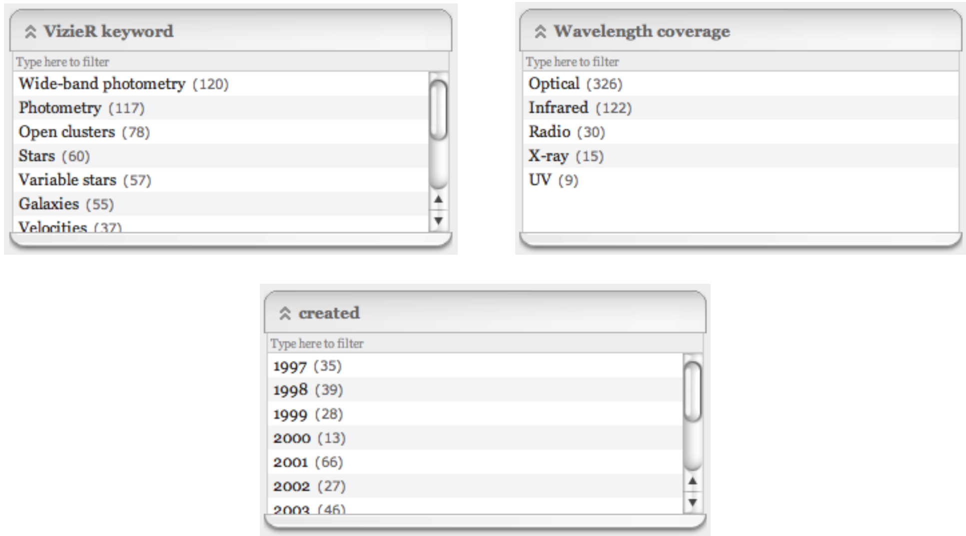


Figure 2: The facet display after the user has selected the "Tables from Astronomy and Astrophysics" entry from the "Catalog" facet shown in Figure 1. The "Catalog" facet is no-longer displayed, since all matching entries have the same value for this facet.

In context-dependent faceted browsing, the choice of facets, and their contents, depends on the previous selections made by the user. This is shown in Figure 2, where the user has selected the "Tables from Astronomy and Astrophysics" entry from the Catalog facet in Figure 1. The content of the facets has changed to reflect the new set of matching items, and the numeric count adjusted so as to indicate to the user how the items are distributed within each facet. The addition of context-dependent information to the display can greatly increase the computation cost of a search, where the increase depends on a number of factors, including the size of the search space, the reproducibility of searches (if many searches are similar then the results can be cached), and the design of the engine used to provide the faceted browsing capabilities.

The contents of a facet can also vary depending on the data type and range of the facet. In Figures 1 and 2 the "Created" facet covered a large time range and so only displayed year values. When the range becomes smaller the facet provides a finer-grain display of the results, as shown in Figure 3, where the results are now split by month within the year 1998.

When searching an unfamiliar database, it is important to be able to remove as well as add search constraints. This facility can be provided in a number of ways; one example is shown in Figure 4, where a list of filters is given, along with the ability to delete them if required. This Figure also illustrates two other features that may be useful when searching: faceted browsing can be combined with other search constraints, such as a text search for

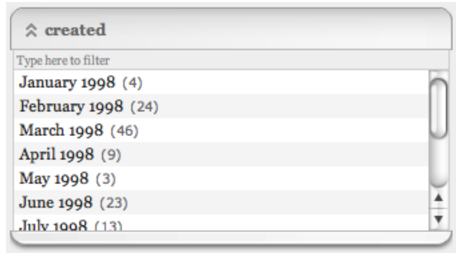


Figure 3: An example of how the facet display can vary depending on the data type and range of the facet. In this case by restricting the search to a single year (1998), the "created" facet automatically groups the data by month, rather than year as in previous examples, such as Figure 2.

a word like "magnitude", and that a filter on a facet can be applied so as to match one or more selected items. The first feature is particularly important for Astronomical searches, where a number of extra search types may be required, such as numerical ranges like "B-V > 0.2" or a positional search within a given radius around a point on the sky. The latter capability is particularly useful for facets whose elements are disjoint, where a record can only contain one of the items.

The idea of using faceted browsing for Astronomical data is not new, and it is already in use to some degree. However, it is not generally used in the context-dependent manner described previously; instead various facets are shown as fixed lists, as in the VizieR Advanced Search[1] page, or only a very restricted set of facets are used, such as the STScI Registry Search[2] shown in Figure 5.

3 Prototype

As a proof of concept for the use of faceted browsing of Astronomical data, both to see whether the interface itself was viable and that the existing data was suitable for such a search, a prototype application was developed using version 1 of the `Exhibit` web-application framework ([5]). This framework was chosen because it was designed to make it very easy for a user to create a faceted browser for a small set of data. All that is needed is to convert the data to be browsed into JavaScript Object Notation (JSON) format[3] and write the interface pages in HTML.

The data used in the prototype service was taken from the Virtual Observatory Data and Service Locator service run by STScI and Johns Hopkins University for the US NVO project. The query used was "`bcg magnitude`", using `AND` to combine the keywords, which

[1] http://webviz.u-strasbg.fr/viz-bin/VizieR
[2] http://nvo.stsci.edu/VORegistry/
[3] http://www.json.org/

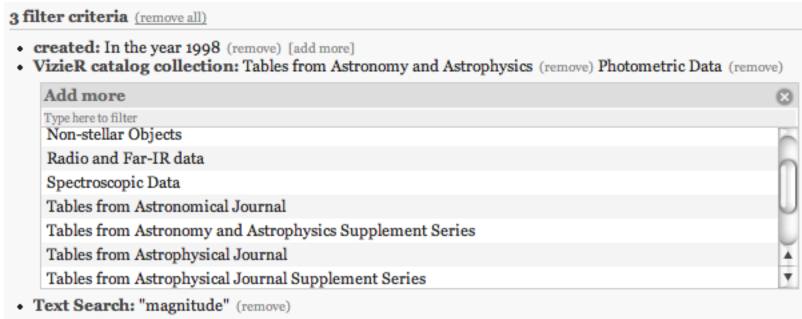
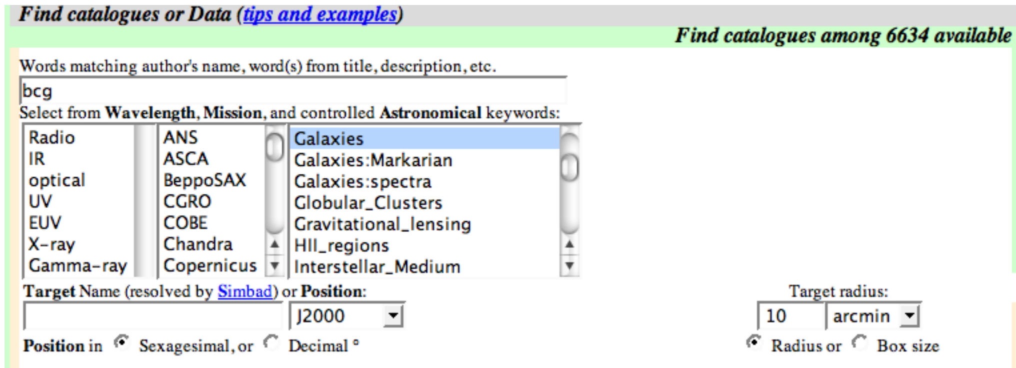


Figure 4: The list of filters — from facets and other search types — that the user has selected. The "Catalog" filter has has its "add more" link selected, which allows a user to add other items from that facet to the search. This has been used to expand the filter on this category to search for all items that match either "Tables from Astronomy and Astrophysics" or "Photometric Data". Also shown in the list is a non-faceted search constraint, in this case the text search on the word "magnitude".

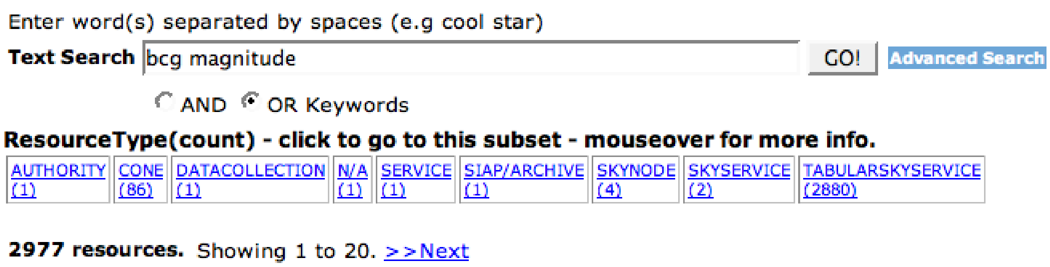
produced 25 matching resources: the first two matching records from this service are shown in Figure 7.

Appendix D describes the steps taken to create the prototype, including examples of the data used, in both XML and JSON formats. For this particular choice of search terms — namely "bcg" and "magnitude" — all the matching records are from the VizieR Catalogue service. The XML data was obtained from the VizieR site using IVOA-approved protocols, and therefore forms a useful subset of the Astronomical data and schema provided by the community as part of the "Virtual Observatory". This data was converted to JSON format, using simple heuristics, for display by `Exhibit`. To simplify the conversion the information on the tabular data of each record — such as the column names and their data types — was not used.

Examples of the main parts of the prototype interface are shown in Figure 8; in order to save space not all the information is shown. The left-hand column contains the currently selected sample, along with controls to change the order and display of this sample (these are provided by the `Exhibit` framework). A simple display, consisting of the main elements from each record, was chosen, along with links to the actual data on the VizieR web site (the `Description` and `Data` items on the Catalog line), together with a simple bit of Javascript to allow the abstract display to be hidden or revealed by selecting the "+" icon. The right-hand column contains the facets, along with the current selection. In Figure 8b the "Galaxies" keyword has been chosen, which cuts down the sample from 25 to 5 in the left column. The facets in the right-hand column have also changed to indicate the composition of the new selection (the "Keyword" facet retains all its original values which allows the user to select multiple keywords in their search).



(a) VizieR Advanced Search



(b) STScI Registry Search

Figure 5: The two figures show examples of faceted browsing in existing Astronomical searches. Figure 5a shows a simple form of faceted browsing, where the facet display is not context sensitive; no indication of the matches for each item is given and once a search is made the results can not be refined. The search results shown in Figure 5b include a facet for the "ResourceType" field which does include the number of matches for each item, but here this is the only facet provided. Both figures only show part of the interface provided by these services, chosen to highlight the use of facets; the excluded areas provide extra search and display capabilities.

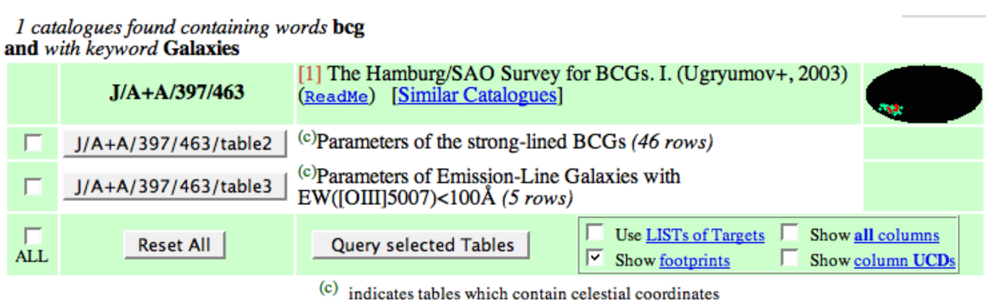


Figure 6: The results from VizieR for the search shown in Figure 5a.

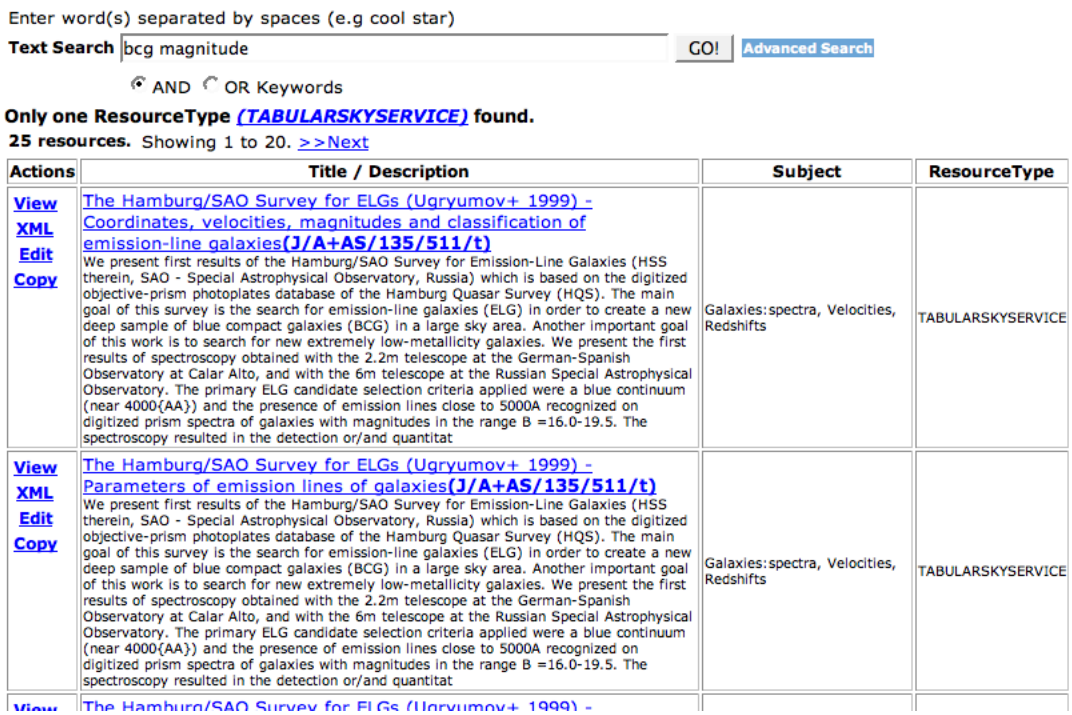
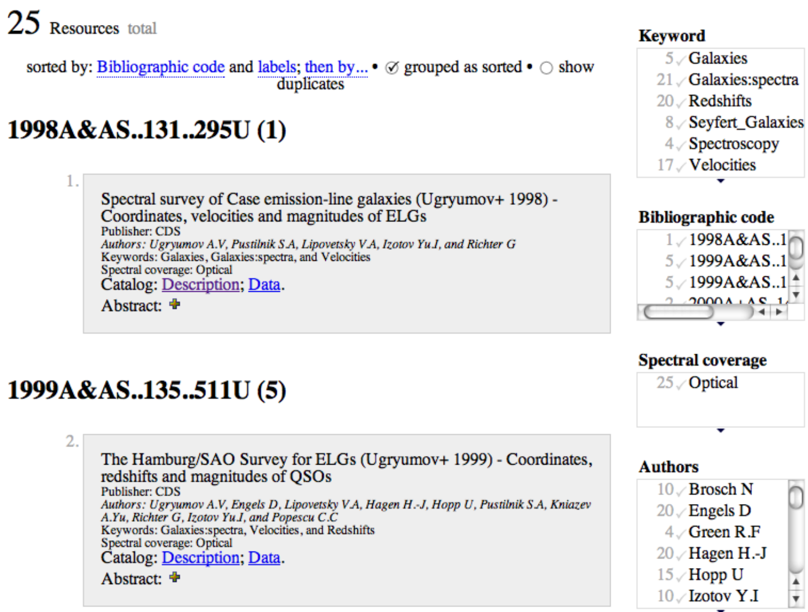
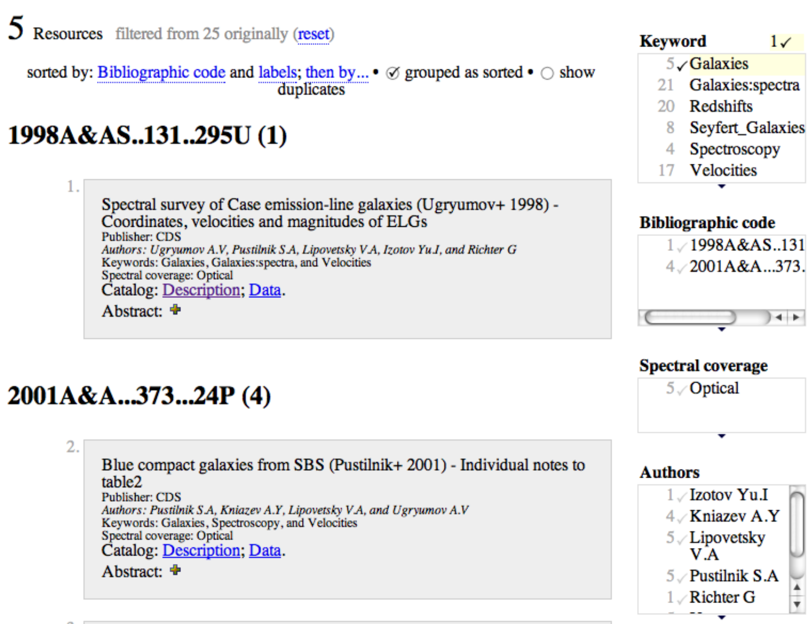


Figure 7: The first two results from the STScI registry search for the words "bcg" and "magnitude". Unlike the example shown in Figure 5b, no facet for `ResourceType` is included in the interface since all the matches have the same value for this field, namely `TABULARSKYSERVICE`.



(a) Initial display from the prototype



(b) Results from selecting the "Galaxies" facet value in the prototype

Figure 8: The main part of the `Exhibit`-powered prototype application. Figure 8a shows the starting point of the search and Figure 8b shows how the faceted interface changes when a choice is made.

The prototype served its two primary goals: is it possible to access data from the Astronomical archives in a form amenable to use and display by a faceted-browsing application, and does the addition of facets to a database search provide extra power to the Astronomer. The following issues were encountered whilst developing the prototype:

1. Most of the development time was spent in creating a useful model of the data for the faceted browser.

2. Given the small size of the chosen database — only 25 items — and the limitations of the `Exhibit` framework used to implement the prototype, it was not possible to get an in-depth understanding of how much improvement faceted browsing can provide for Astronomical searches.

3. The registry records may not contain all the data stored in the original database. As an example, the VizieR web interface provides a list of related links for each record but this information is not included in the registry data.

4. Some of the data stored in the registry is not suitable for display without some degree of clean up or validation.

4 Main project

The initial proposal was to apply faceted browsing to a search of a document-rich Astronomical database, and, towards the end of the project it was decided to see how well the technique would work on a data set which contained more numeric data. Both applications were based on data primarily accessed from IVOA Registries (Appendix C). The aim was to develop a web-based system since this removes much of the cost of developing a separate user interface, and is an approach that Astronomer's are used to. The disadvantage to this is that it may limit the expressiveness of the search application.

4.1 Choice of browser

The `Exhibit` web framework used in the prototype provided all the mechanics for faceted browsing, requiring only the data to display and the creation of a HTML web page. This meant that time could be spent on modeling the data rather than developing a User Interface for the application, since this was handled by the user's web browser. Since the framework runs as a client-side application, using the browser's Javascript engine, performance quickly degrades as the number of items increases, and it was found to essentially be useless by the time 1000 records was reached. The same group that wrote `Exhibit` also provide a Java web application called `Longwell` that also requires limited customization to produce a working faceted browser, but that can scale to much larger data sets than `Exhibit`, since the processing is now done on the server side ([17]).

The underlying data model used by `Longwell` is based on the Resource Description Framework language ([23]) developed by the World-Wide Web Consortium (W3C); a brief description of the salient points of this language is provided in Appendix E. This language is designed to represent information about resources in the World Wide Web, and is one of the corner stones of recent work on the Semantic Web ([28]). This therefore provides an opportunity to see how well the schema and data models provided by the IVOA map to those used by the general Semantic Web community, and to see what can be gained by considering Astronomical data as a set of linked resources.

An exhaustive search of alternate systems for faceted browsing was not made since time was limited. The following alternative engines were briefly looked at:

Aduna AutoFocus Server It appeared difficult to adapt the system to index data that is not in one of the supported formats, which would be required here. The home page is `http://www.aduna-software.com/products/autofocus_server/overview.view`.

Flamenco There appeared to be no development of the system, and little community involvement, after its initial release. The home page is `http://flamenco.berkeley.edu/`.

Apache Solr While this system, which is built on top of the Lucene search engine, appears powerful, it appeared to require more work than `Longwell` to create a simple faceted browser. The home page is `http://lucene.apache.org/solr/`.

mSpace It was not clear what the license was for this system. There also appeared to be little to no development going on for it. The home page is `http://mspace.fm/`.

/facet There appeared to be little to no development on the system. The home page is `http://slashfacet.semanticweb.org/`.

4.2 VizieR search application

The first application was designed to see how well faceted search could extend and enhance an existing, popular, Astronomical search interface. The VizieR service, at the Centre de Données astronomiques de Strasbourg (CDS), collates a large number of data sets from Astronomical journals and projects and provides a search interface to this ensemble, and was chosen since:

- It is a commonly used facility by the Astronomical community.
- It contains a mix of text data — such as the abstracts from journal papers and the description of tables — coupled with more structured data, such as keywords, data types, and data values, typical of Astronomical data sets.
- The database is provided as an IVOA Registry, which means that it is a good test of data access, handling, and modeling.

- `Longwell` has been used by the DSpace[4] project to provide faceted browsing for data sets with a similar structure to those used by VizieR.

The two main areas of work were the modeling of the data in RDF and the configuration of `Longwell` for best use of this data; the steps taken are described in Appendices F and G respectively. Appendix F.2 discusses issues found during the conversion of the VizieR data set to a RDF model.

The final sample contained 14075 records, along with the associated ancillary records shown in Table 4, and was 29 Mb in size. The application speed was perfectly usable when run on an OS-X MacBook pro, although it was only tested with a local connection.

4.2.1 Using the VizieR browser

Appendix A contains a simple demonstration of using the VizieR browser, highlighting the main points of the interface. This section describes the details of the browser.

The starting point of the browser is shown in Figure 9. At this point a user can enter either a text string or select one of the classes of data to start searching. For this application all the RDF classes that were developed during the data modeling were included to see whether they were useful. Other than using confusing names in the interface — such as "IVOA Resources" — this set of classes did allow the user to navigate the data base; this is not necessarily the case, as will be shown in section 4.3. Figure 10a shows the interface after the user has selected the "Resources" class (which refers to the Tabular Sky Service resources in the VizieR Registry) and shows the facets that have been defined for the class (Appendix G.2). A listing of constraints is provided at the top of the page; this provides feed-back to the users on what they have previously done and allows them to either delete existing filters or to expand the sample by including other values (e.g. Figure 4). The facet list and text-search box on the right are used to further restrict the search.

Figure 10b shows the interface after several selections have been made; the sample size is now small enough to have individual entries included in the main display, and much of the contents of the individual facets have been removed since they are no longer relevant. The full display for an individual record is shown in Figure 23; Appendix G.3 describes how the contents and appearance of these pages are defined. For this application the column data included in the Registry data — i.e. the contents of the `vs:table` element in Listing 2 — were not included in the RDF model to simplify the conversion (Appendix F.1).

4.2.2 Lessons learnt

The major items learnt from the development and use of this application are:

[4] http://www.dspace.org/

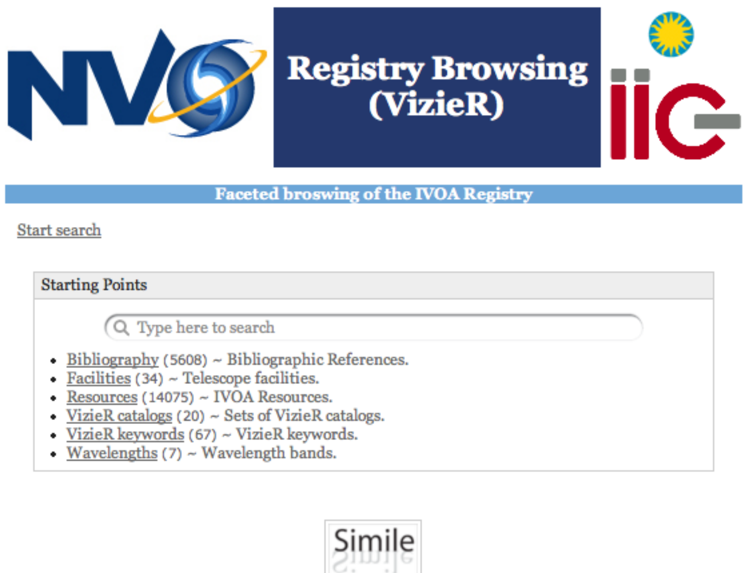
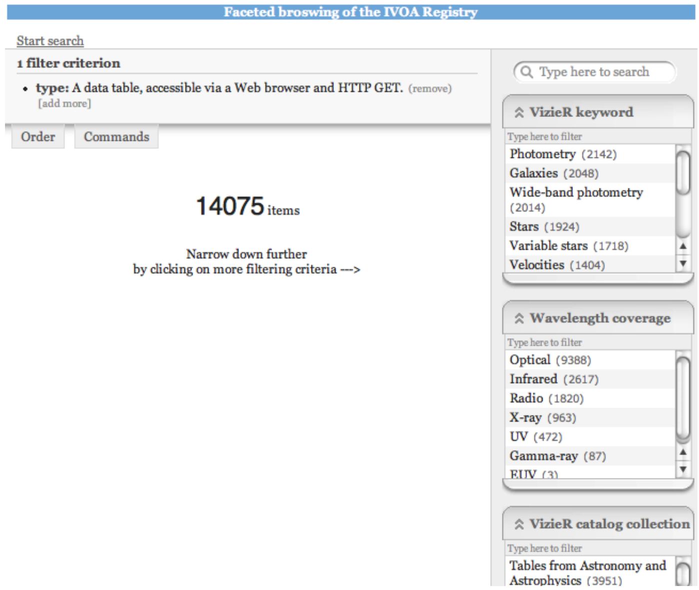
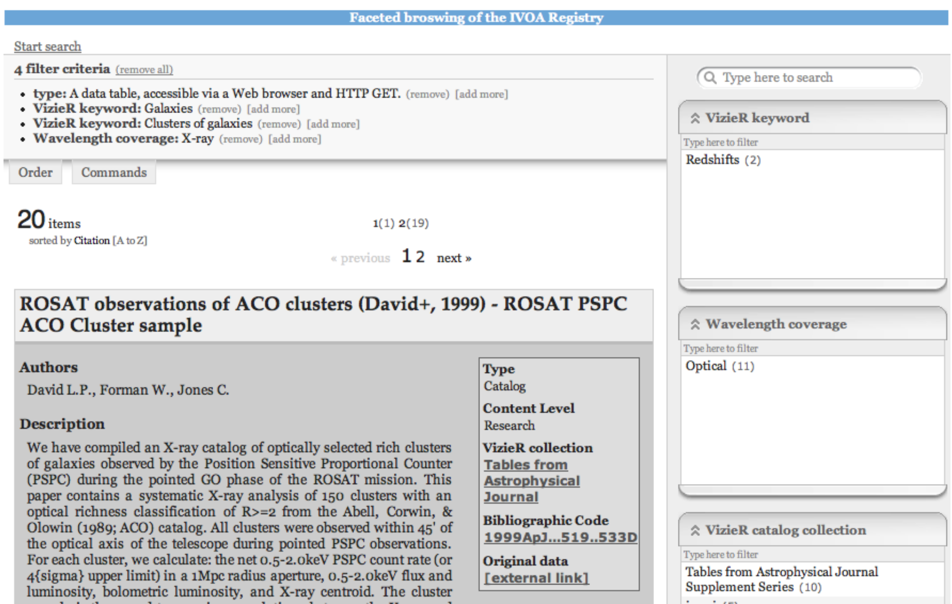


Figure 9: The starting point of the `Longwell` application when configured to browse the VizieR registry data. The "Starting Points" section is the entry point to the search engine, and is created by `Longwell` according to what facades have been configured (Appendix G.1). A simple banner was added to the display, based on the NVO template for web services, although the documentation, help, and other informational links were excluded.



(a) VizieR browser selecting the "Resources" class



(b) Multiple facet restrictions

Figure 10: Examples of a search using `Longwell` with the the VizieR data base. Figure 10a shows a truncated version of the results of selecting the "Resources" entry from Figure 9. When further restrictions are added, such as selecting the "X-ray" value of the "Wavelength coverage" facet, the display changes to something similar to Figure 10b.

- The RDF modeling of IVOA Registry data is relatively simple in general, since the underlying models are very simple (namely, that objects can be described as properties attached to resources, even for abstract concepts).

- The IVOA Registry model provides several elements (which were converted into RDF predicates) that were obvious candidates for facets such as keyword and wavelength (which correspond to the `subject` and `waveband` elements in Table 3). Several other elements, such as `contributor` and `contentLevel`, contained a single value; for more varied data sets (e.g. from multiple registeries), these fields would also make useful facets.

- The most useful addition to the current model would be to separate out the author list into individual values. A data model such as that provided by the FOAF project ([7]) could be used for this purpose.

- Extra value can be provided using the RDF model by making subjects into resources in their own right, and adding extra information to them (e.g. a description of what wavelength range is meant by "X-ray") is relatively easy and can be beneficial to the user.

- The IVOA meta data standard does not allow all the information from the publisher to be encoded. This was noticed in two areas with the VizieR data set:

    1. Text could contain extra mark up, which makes it hard to correctly display to the user without some indication of what form of mark up is being used (e.g. Figure 23). This could be resolved by explicitly forbidding such mark up, by adding extra information to the XML schema to indicate the presence and type of mark up, or by allowing some subset of XHTML to be included in these elements.
    2. The registry records do not contain as much information as the original service does; for the VizieR records there are links between tables and lists of extra keywords that are not included in the registry version. It is not obvious whether the XML schema can be enhanced to include this information.

- The facets provide a simple and obvious way to combine search constraints using "and", which makes it easy for a user to drill down into the search space. It is also easy to delete filters, and this need not be done in the order they were added, which allows the user to expand their search results. However, the user interface for combining filters using "or" is not immediately obvious or natural, and is limited to combining values that are in the same facet.

4.3 DataScope search application

The development and use of the VizieR search interface showed that `Longwell` could be used to search text-rich data sets, but did not provide any information on how well numerical resources, such as tables and images, would be handled. To rectify this, a numerically-rich data set was identified and a `Longwell` configuration was developed for this new sample.

The NVO DataScope application[5], which queries IVOA resources for matching data given a region of sky, was used to create the data sample. The existing interface — parts of which are shown in Figure 25 — already uses a simple form of faceted browsing to arrange the data results by wavelength and instrument type (e.g. imaging or spectroscopy); the question was whether the inclusion of more facets would improve the search experience for an Astronomer. Since there was little time left for this phase of the project, the DataScope application was used to search a specific location on the Sky — corresponding to the interacting pair of galaxies known as the Antennae galaxies — and the resulting data set was converted into RDF and displayed in `Longwell`. Appendix H describes the process of data acquisition and modeling, as well as example output; since the screen shots are similar to those developed for the VizieR data sample (e.g. Figures 9 and 10) they are not included here.

4.3.1 Lessons learnt

As it was not possible to resolve all the issues related to the handling and display of numerical data, the results are not as polished as those for the VizieR search system. The project did highlight the following issues:

- The registry data model had to be enhanced and expanded from that used in the VizieR application since the data sample was more heterogeneous. This meant that the contents of XML elements from the registry, such as `contentLevel`, could now be modeled using RDF classes rather than just as string literals.
- The quality of VOTables provided by the Astronomical services varied greatly, and while they were all well-formed XML files, there were a number of validation failures when comparing them to the VOTable XML schema, including namespace issues, incorrect attribute names, and incorrect element content.
- The addition of more classes and facets to the interface (compare Figure 26 to 9) did not lead to an improved search experience, as the extra options often made it harder to navigate through the data set. This was in part due to poor naming of classes, and the separation of the table meta data from the column information.

[5]http://heasarc.gsfc.nasa.gov/vo/

- The interface did not provide any display of the numeric data from the tables, so users could not actually access the column data from the application. It was not clear whether the problems encountered in displaying this data were down to poor choices in the data modeling or were either limitations or bugs in `Longwell`.
- There is no support in `Longwell` for numeric facets other than treating the data as single values as is done with text data. One option would have been to create a facet that automatically grouped data into ranges, but there was no time to develop this idea. It was also not clear how best to deal with range filters, such as $k > 20$, within `Longwell`.
- Since Astronomical data sets may be provided by multiple publishers, care needs to be taken to amalgamate these results, to avoid confusing the user. In the sample analyzed here, this was complicated by the fact that for at least one such match the retrieved tables were not the same, in that the number of columns and rows did not agree.

5 Conclusions

The project consisted of three basic phases: can the Virtual Observatory data be accessed, how well can this data be modeled using RDF, and does faceted browsing enhance the search experience for Astronomers?

5.1 Data access

The development of IVOA standards, such as the Registry Data Model and VOTable format, has meant that it is possible for Astronomical Data Centers to provide information using interfaces and formats that can be easily accessed and manipulated via clients such as those developed in this project. The main issue is in how well supported these standards are.

The April 2008 move to version 1.0 of the registry schema should reduce, if not remove completely, issues encountered in the processing of the registry data (Appendix F.2), as the community has a better understanding of both the process and data models. Similarly, as data providers get more experience and feed-back, and as the standards stabilize, many of the issues seen with the VOTable support are expected to be cleared up; in fact some of the problems discussed in Appendix H.3 have since been fixed. However, Virtual Observatory clients should still expect to receive incorrectly formated data, and develop techniques to extract as much usable information as possible.

The IVOA community has been divided over the amount of information that should be recorded in the registry records: should it just be a system which identifies resources or should it also include detailed meta data about them? The initial registry implementations

— those that followed version 0.10 of the meta data schems — tended to follow the latter approach whereas the intention is for the new v1.0 compliant registries to be light weight, and the detailed meta data is accessible directly from the services themselves using the proposed VO Support Interfaces system. For the search interfaces presented here, this change would complicate the process of data access and ingest, but should not change the user interface. A related problem is in how data providers can provide extra information that can not be expressed using the IVOA schema, and how will services that access this information make use of it?

5.2 Data modeling

Since the underlying concepts behind the IVOA Registry Data Model and the RDF language are similar — namely that items can be identified and described using simple statements — then the conversion between the IVOA XML schema and RDF was relatively straight forward. It was also possible to enhance the search data by describing values as instances of a RDF class rather than just use a string literal. For example, the `em.IR.J` [6]Unified Content Descriptor value was annotated with a description — in this case "Infrared radiation between 1.0 and 1.5 microns" — and it can be used to find all other items that refer to it, so enhancing the links between items.

The RDF vocabularies described in Appendices F and H were built for the specific purpose of modeling the data for use in `Longwell`. They are therefore not complete, since they do not model all components of either the registry or VOTable schema, just those elements needed for searching and displaying the data sets. They are also not particularly suited for use in other applications since the data structures used (namespaces, class design, predicates) are specific to the project; general purpose vocabularies were used when possible but this only covered a small number of cases. Although not available at the time of the project, there is now a proposed RDF vocabulary for version 1.0 of the IVOA Registry schema ([16]).

5.3 Faceted browsing

The two applications developed in this project show that the available data can be viewed using a faceted browser, and that the use of facets does enhance the search interface. However, it is clear that a search interface which relies solely on facets is insufficient for many use cases; the `Longwell` browser itself already includes a text search filter, and the project identified numerical searches — whether as a facet, where the values are arranged into bins[7], or the ability to apply a range filter — as important for Astronomical data sets. One issue for existing search engines will be how to integrate faceted searching into their code bases.

[6]UCD

[7]There has been recent work developing a facet like this in version 2 of the `Exhibit` framework.

The design of `Longwell` requires that the choice of what predicates are used as facets is made when the application is configured (e.g. Appendix G.2). For the two samples used in this project, the facets were chosen manually, based primarily on what predicates were available, and what the range of each predicate was (there is little point in providing a facet if there are too few or too many distinct values). The resulting facets — such as "keyword" and "UCD value" — were often found to be useful, but a more systematic method for choosing the facets could be useful. There is existing work on automatic and semi-automatic classifiers that may prove useful for such work (e.g. [2], [21]).

The `Longwell` browser provides a user interface in which it is easy to combine restrictions using a logical "and" to drill down into a data set. However, it has limited support for using logical "or" to combine filters, and the user interface for this is in no way obvious. It also has no support for negation, so that search terms can not be excluded from the search results. These are not necessarily intrinsic limitations to the technique of faceted browsing, although the design of a simple interface for these concepts may be challenging (e.g. [21]).

5.4 Future Work

During the project a number of items were discussed that could be profitably looked at for ways to improve Astronomical searches; they are not all related to faceted browsing.

- Since the contents of the Virtual Observatory, as defined by the Astronomical archives and data services, are not constant, then search results will change with time for the same query. It would be very useful to a user to see what has changed — whether values have been added, deleted, or updated — since their last search on a topic.

- Can a search be saved? The data set created by a search can be saved, even if it may be a cumbersome process (e.g. Appendiz H.1), but the constraints used to make that search — such as the facet values used — are not. This information could be used to save time, since they would make it easy to repeat a search at a later time, find similar data by changing certain parts of the search (e.g. location or wavelength coverage), or perhaps even to recommend other data sets.

- Given the sample sizes from Astronomical searches, it would often be useful to rank the search results so that the "best" matches are highlighted. There has been some work on such ranking of Astronomical catalogues ([4]) and individual data providers, such as the Chandra archive, provide such rankings for their own catalogs, but there has been little, if any, development towards such a system for large searches.

- The data tables contain limited semantic information about their columns — in the form of Uniform Content Descriptor (UCD) and Unique type (UType) attributes — that could be used to enhance search applications. UCDs were modeled as a RDF class in the DataScope application (Section 4.3) which allowed users to search for all tables that contain a "photometric magnitude" or "Proper motion", but this does not make full use of the information encoded by these terms.

- Since the end of this project there has been activity in the semantics group of the IVOA, with several Astronomical vocabularies in RDF — including ones based on keyword values — being proposed for general use. Such work would help in the data modeling of Registry resources, and may provide guidance on facet choice ([22]). This work also offers the potential for improving the search interface by providing contextual knowledge and disambiguation, so that the term "BCG" would also match "Blue Compact Galaxy" and "Brightest Cluster Galaxy".

- There are other ways to visualize the facet choices than those used by the `Longwell` engine, such as cluster maps (Aduna AutoFocus) and the graphs of TouchGraph[8].

[8]http://www.touchgraph.com/technology.html

References

[1] NASA Astrophysics Data System,
http://ads.harvard.edu/

[2] Castanet, (Semi) Automatic Creation of Facet Hierarchies,
http://flamenco.berkeley.edu/castanet_demo.html

[3] Dublin Core Metadata Initiative,
http://dublincore.org/

[4] M. Dolensky, B. Rino, "Ranking Query Result Sets", IVOA Note

[5] Exhibit web application framework,
http://simile.mit.edu/exhibit/

[6] The Flamenco Search Interface Project,
http://flamenco.berkeley.edu/

[7] Friend of A Friend vocabulary (FOAF),
http://www.foaf-project.org/

[8] The Fresnel display vocabulary for RDF,
http://www.w3.org/2005/04/fresnel-info/

[9] Great Observatories Origins Deep Survey (GOODS),
http://www.stsci.edu/science/goods/

[10] High Energy Astrophysics Science Archive Research Center (HEASARC)

[11] info:bibcode/ - NASA ADS Bibliographic code URI,
http://info-uri.info/registry/docs/misc/faq.html#ex_bibcode

[12] NASA/IPAC Infrared Science Archive (IRSA),
http://irsa.ipac.caltech.edu/

[13] International Virtual Observatory Alliance (IVOA)

[14] R. Plante, T. Linde, R. Williams, and K. Noddle, "IVOA Identifiers", IVOA Recommendation

[15] K. Benson et al., "IVOA Registry Interfaces", IVOA Working Draft

[16] N. Gray, "An RDF version of the VO Registry", IVOA Note

[17] Longwell faceted browsing web application,
http://simile.mit.edu/wiki/Longwell

[18] Notation 3 specification,
http://www.w3.org/DesignIssues/Notation3.html

[19] NASA/IPAC Extragalactic Database (NED),
http://nedwww.ipac.caltech.edu/

[20] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH),
http://www.openarchives.org/pmh/

[21] E. Oren, R. Delbru, and S. Decker, "Extending faceted navigation for RDF data", Proceedings of the International Semantic Web Conference (ISWC), pp. 559-572. Nov. 2006.

[22] O Suominen, K Viljanen, and Eero Hyvnen, "User-centric Faceted Search for Semantic Portals", Proceedings of the European Semantic Web Conference ESWC 2007, Innsbruck, Austria, Springer, June 4-5, 2007.

[23] Resource Description Framework (RDF),
http://www.w3.org/TR/rdf-primer/

[24] RDF Schema Vocabulary (RDFS),
http://www.w3.org/TR/rdf-schema/

[25] Robustness Principle from RFC 793,
http://en.wikipedia.org/wiki/Robustness_Principle

[26] R. Hanisch et al., "Resource Metadata for the Virtual Observatory", IVOA Recommendation

[27] Sloan Digital Sky Survey (SDSS),
http://www.sdss.org/

[28] The Semantic Web,
http://www.w3.org/2001/sw/

[29] S. Derriere et al., "An IVOA Standard for Unified Content Descriptors", IVOA Recommendation

[30] Uniform Resource Name Syntax (URN),
http://tools.ietf.org/html/rfc2141

[31] The VizieR Catalog Service,
http://vizier.u-strasbg.fr/

[32] R. Plante et al., "VOResource: an XML Encoding Schema for Resource Metadata", IVOA Recommendation

[33] F. Ochsenbein et al., "VOTable Format Definition", IVOA Recommendation

[34] IVOA Data Services schema, version 0.5,
`http://www.ivoa.net/xml/VODataService/v0.5`

[35] Web Design Practices,
`http://www.webdesignpractices.com/navigation/facets.html`

A A demonstration of the VizieR registry browser

This appendix presents a simple run-through of the VizieR registry browsing application described in Section 4.2. Figure 11 shows the starting point of the application, which includes the search interface in the center of the window together with the banner — based on the NVO application template — and the browser controls (all the images shown here were taken using the `Firefox` web browser running on an OS-X machine). Since the application was running locally the URL used in these examples (`http://127.0.0.1:8080/`) is not what would be used in a production environment.

The user can start a search by either entering a word into the search box or by selecting one of the six links ("Bibliographic References" to "Wavelength bands"), as described below.

A.1 Starting with a text search

Figure 12 shows the result of entering the text "chandra" into the search box on the start page. The results page consists of three main areas (other than the header): the main search results on the left-side of the screen, the facets and text search box on the right of the screen, and the filter list above the search results.

In this example, the search has produced 270 matches, and the main display area shows the start of the first match, together with options for paging through the whole sub-sample. The facets and text box on the right allow the sample to be pruned by applying more filters, whilst the filter list lets the user know where they are (i.e. what filters have been applied) and gives them the ability to remove some of these restrictions. The details for a single record can be found by clicking on its title (Figure 13); note that this display contains no facets, since there is only a single match, or any indication of how the record was found (i.e. the set of filters is not included).

The type facet in Figure 12 shows that one of the matches is an observatory facility whereas the remaining items are data tables. Figure 14 shows the result of selecting the "observatory facility" value in this facet. The facility class was one of the RDF classes introduced during the data modeling stage (Appendix F) and so does not contain any information other than the name of the facility; in a production environment the database could be extended to include RDF statements such as the home page and a brief description of the facility. The "Referers" section at the bottom of the record has been expanded to show all the items which refer to this item; in other words all the RDF subjects that have a predicate whose object is the Chandra observatory facility. This makes it possible for users to easily find related items, although the User Interface is somewhat clumsy and not self explanatory.

Figure 15 shows what happens if the "data table" entry is chosen in Figure 13, followed by "Atomic data" value in the VizieR keyword facet. The filter list now shows all three of these stages, and can be used to remove or change the constraints. The facet display has been updated: the type facet has been removed and the remaining facets reflect the contents of the current sample. The facets can deal with incomplete information — i.e. each matching record does not have to contain the facet predicate — as shown by the VizieR keyword facet (only 1 out of 22 matches contains a VizieR keyword).

The links within each record can take you either to other entries in the database, such as the "Bibliographic Code" entry (Figure 16), or to external sites, such as the "Original data" link (Figure 17). As with Figure 14, the bibliographic data shown in Figure 16 was added during the data modeling stage; in this case the entry was extended to include extra information, such as the publishing information and NASA/ADS link, as this was easily extracted from the bibliographic code entry (Appendix F.2).

Figure 18 shows the filter list being used to expand the search sample. In this case the `[add more]` entry on the VizieR keyword line has been selected which reveals a list of all the keyword values. Selecting one of these will add the filter using a union rather than intersection, so choosing "Early-type stars" will select all data tables which contain the text "chandra" and have VizieR keywords of either "Atomic data" or "Early-type stars". Figure 19 shows the result of this selection.

A.2 Starting with a class

Searches can also be started by selecting a class rather than with a text search. Figures 20 and 21 show the results of selecting the "Resources" and "VizieR keywords" links, respectively, from the start page (Figure 11). For the resources selection, no results are shown since there are too many matches; further filters have to be applied to get the sample down to a reasonable size before the individual entries will be included. The keyword selection shows a class added, as with the facility (Figure 14) and bibliographic (Figure 16) entries, during the data modeling stage, hence the lack of information for each record. After the initial selection, the same steps can be taken as with the text search (Section A.1).

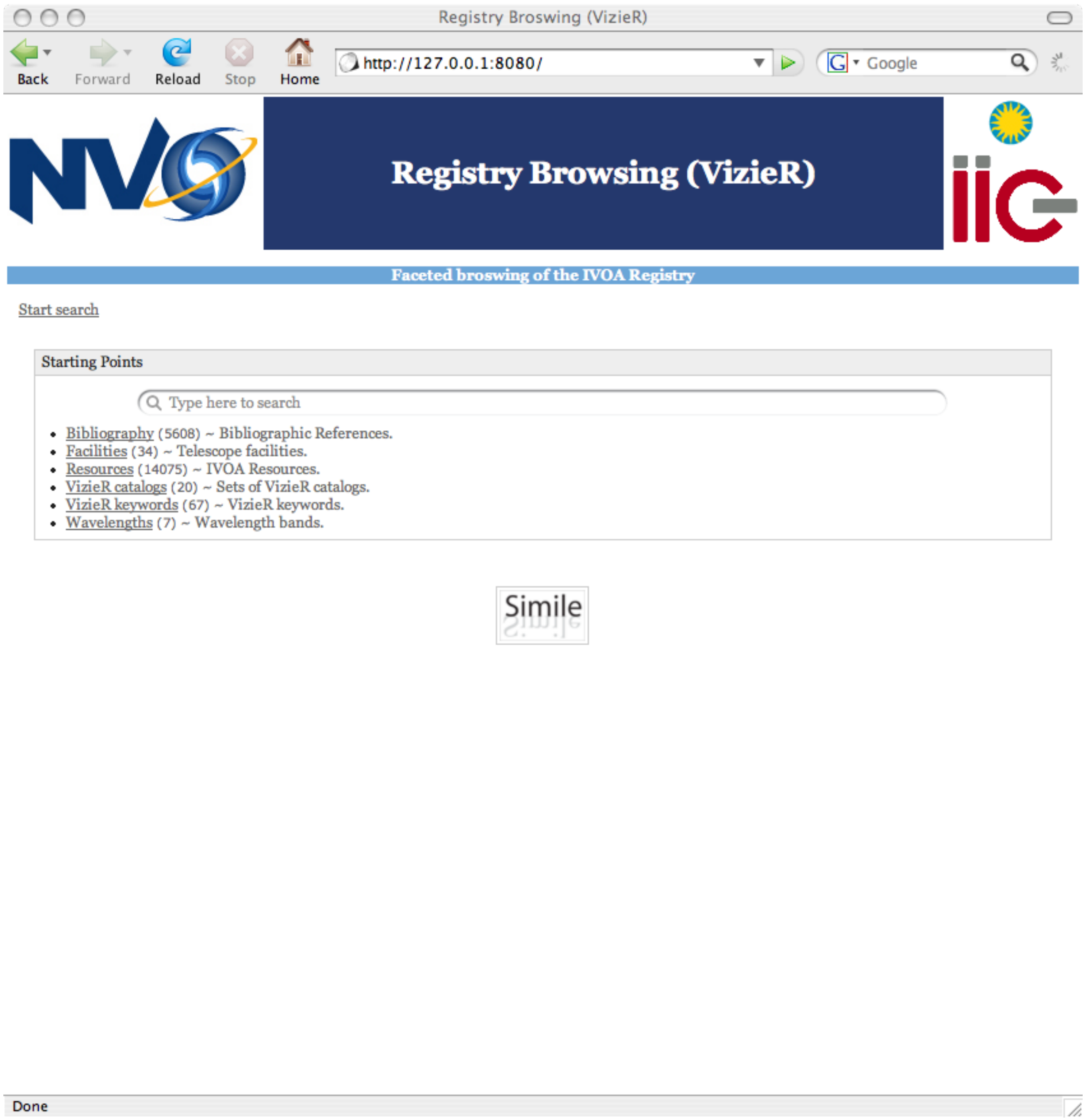


Figure 11: The start page of the VizieR registry browser.

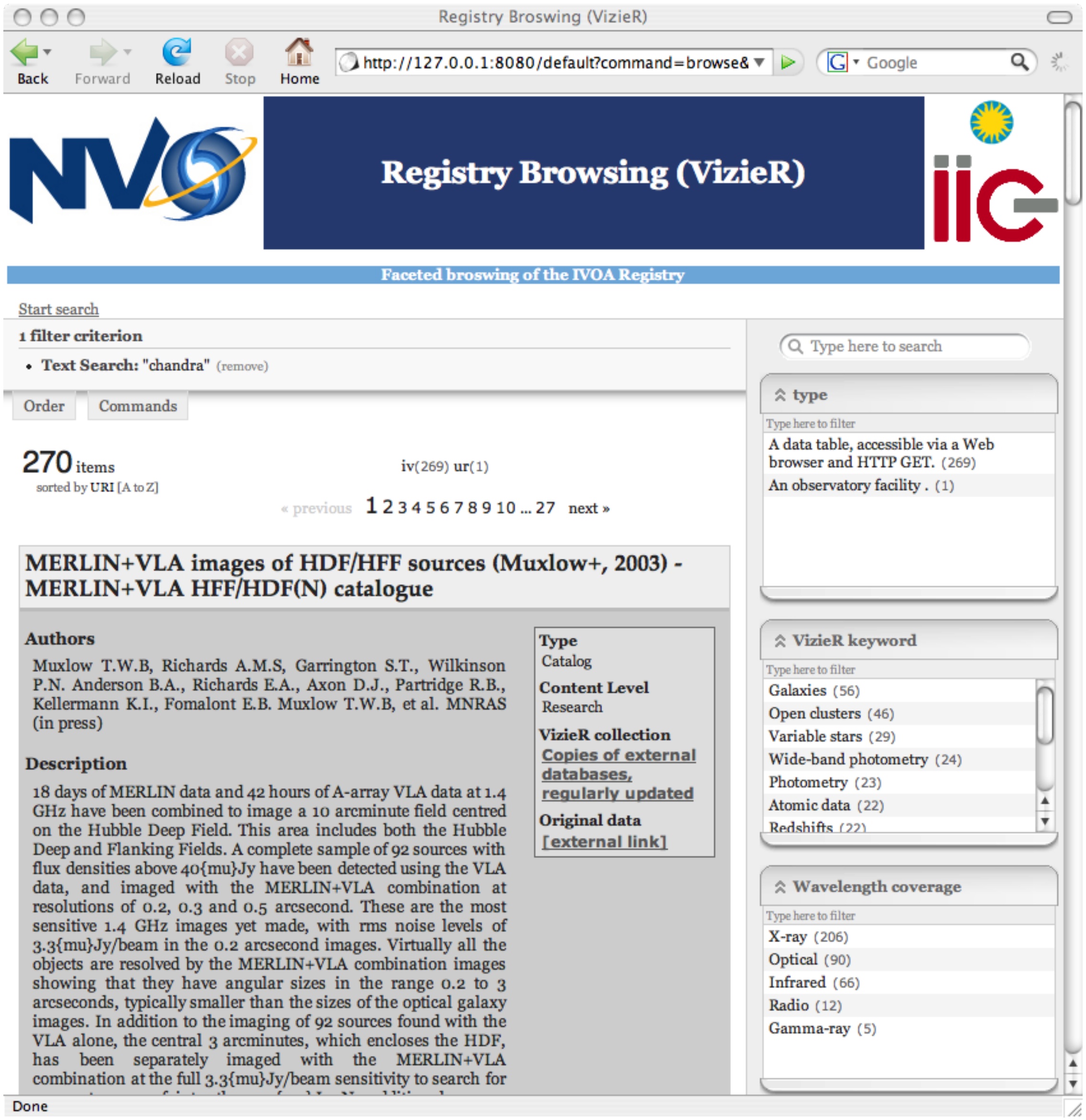


Figure 12: The results from a text search for the word "chandra". The search results are shown on the left side of the page whilst the facets are shown on the right.

Figure 13: The contents of the first match from the list shown in Figure 12.

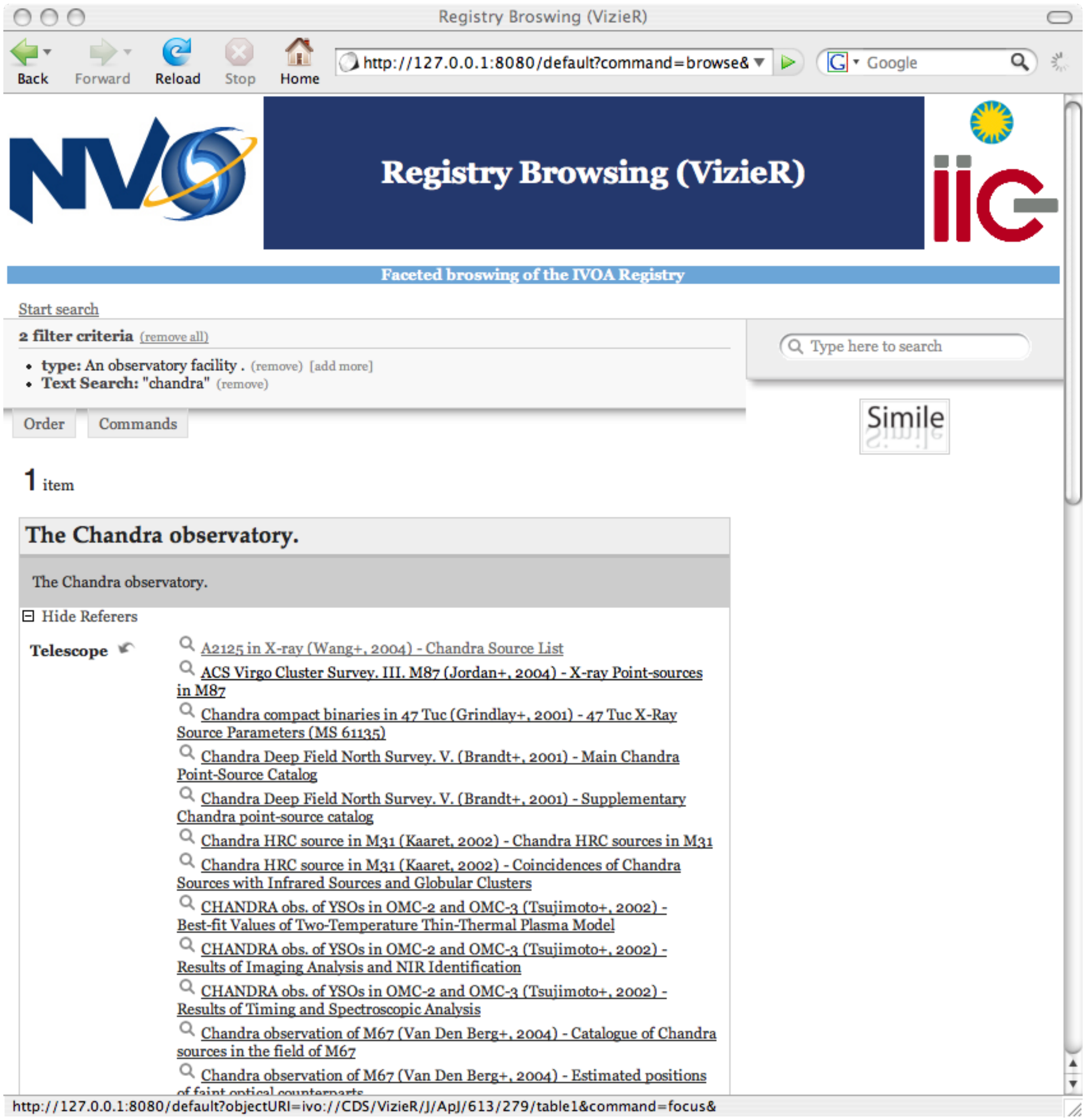


Figure 14: Selecting the "observatory facility" value in the "type" facet shown in Figure 12 results in a single match, and hence no facets. The "Show referers" link has been selected to show all items that refer to this facility.

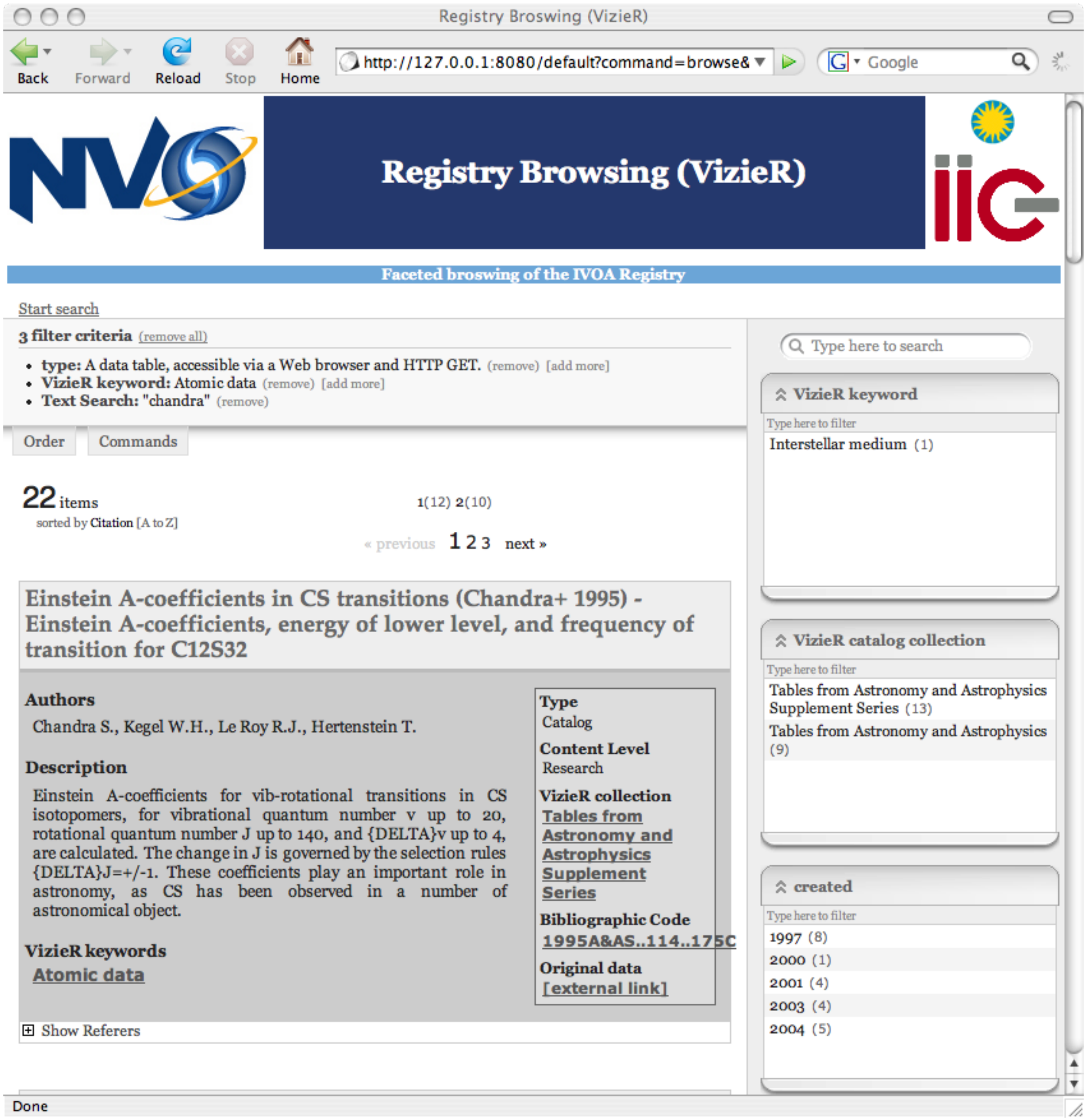


Figure 15: The search results from combining the text search for “chandra” (Figure 12) with two facet restrictions: a type of “data table” and the “Atomic data” keyword. A list of the current set of filters is shown above the matching resources; this list can be used to remove or add new filters.

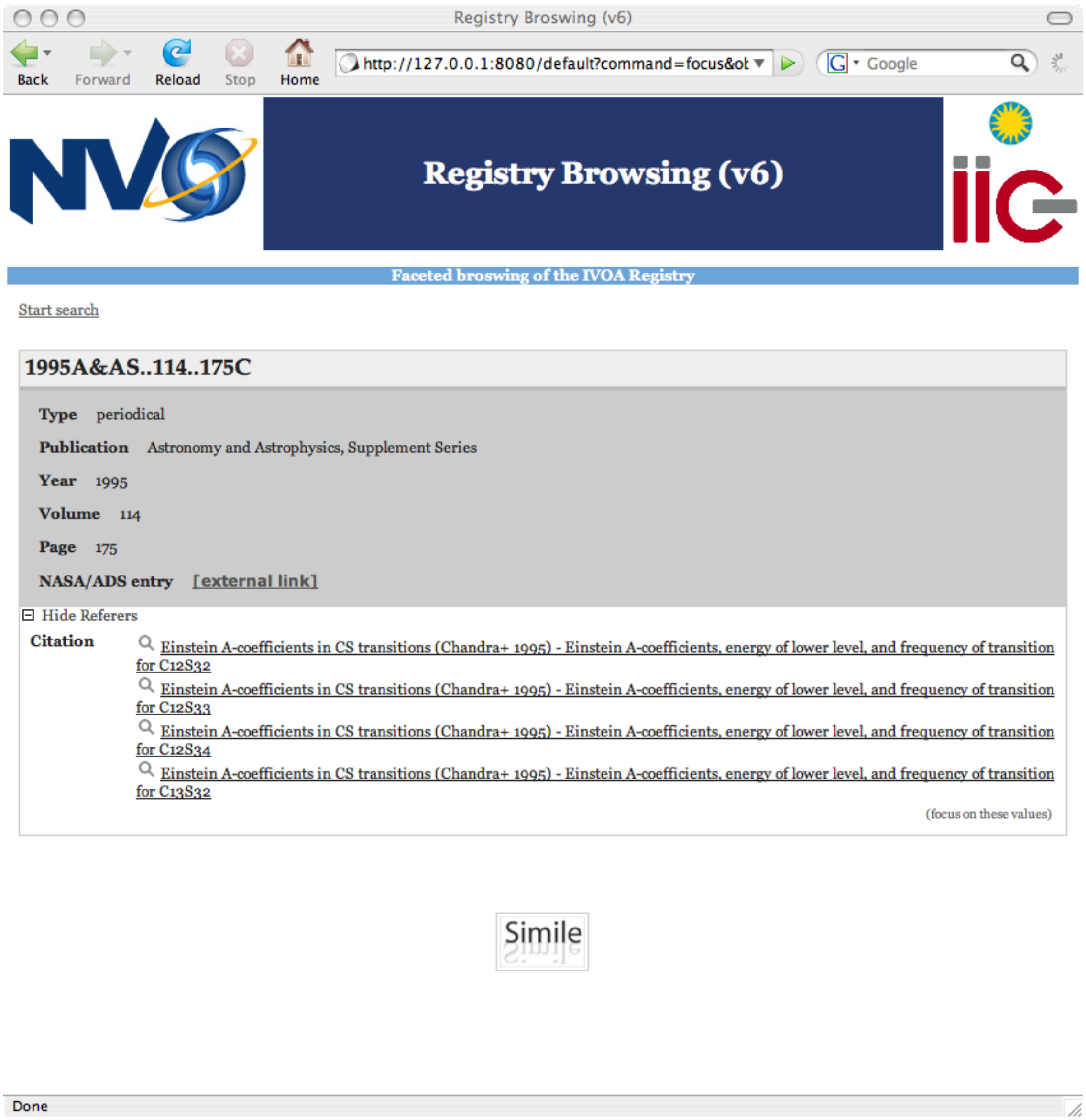


Figure 16: The display after selecting the "Bibliographic Code" link from the first record in Figure 15. The "Referers" link has been selected to show all the tables from this paper.

Catalogue: <A HREF="http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A%2BAS/114/175/table2a">J/A+AS/114/175/ta...

Back Forward Reload Stop Home http://cdsarc.u-strasbg.fr/viz-bin/Cat?J/A+AS/114 Google

Detailed Description of J/A+AS/114/175

CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · Tutorial · Developer's corner

Individual files can be fetched (via http) from here
or via ftp from: ftp://cdsarc.u-strasbg.fr/pub/cats/J/A+AS/114/175

The catalogue can be queried from: http://vizier.u-strasbg.fr/viz-bin/VizieR?-source=J/A+AS/114/175

This standardized document follows the rules of the Standard Description for Astronomical Catalogues. *For those having to work on this catalogue using FORTRAN programs, this* `ReadMe` *file can be transformed into a f77-compliant program which reads the files making up this catalogue, in two versions: either a* f77 program loading whole files into arrays, *or a* f77-program reading each data file line by line.

```
J/A+AS/114/175      Einstein A-coefficients in CS transitions    (Chandra+ 1995)
================================================================================
Einstein A-coefficients for vibrational transitions in CS
     Chandra S., Kegel W.H., Le Roy R.J., Hertenstein T.
    <Astron. Astrophys. Suppl. Ser. 114, 175 (1995)>
    =1995A&AS..114..175C      (SIMBAD/NED BibCode)
================================================================================
ADC_Keywords: Atomic physics

Keywords: ISM: molecules - molecular data - star: abundances


Description:
    Einstein A-coefficients for vib-rotational transitions in CS
    isotopomers, for vibrational quantum number v up to 20, rotational
    quantum number J up to 140, and {DELTA}v up to 4, are calculated. The
    change in J is governed by the selection rules {DELTA}J=±1. These
    coefficients play an important role in astronomy, as CS has been
    observed in a number of astronomical object.

File Summary:
--------------------------------------------------------------------------------
       FileName      Lrecl  Records   Explanations
--------------------------------------------------------------------------------

× ReadMe              80        .    This file
× table2a             49    23660    Einstein A-coefficients, energy of lower level,
                                    and frequency of transition for C12S32
× table2b             49    23660    Einstein A-coefficients, energy of lower level,
                                    and frequency of transition for C12S33
× table2c             49    23660    Einstein A-coefficients, energy of lower level,
                                    and frequency of transition for C12S34
× table2d             49    23660    Einstein A-coefficients, energy of lower level,
                                    and frequency of transition for C13S32
--------------------------------------------------------------------------------
```

Done

Figure 17: The VizieR page for the resource is found by following the “Original data” link from the first match shown in Figure 15.

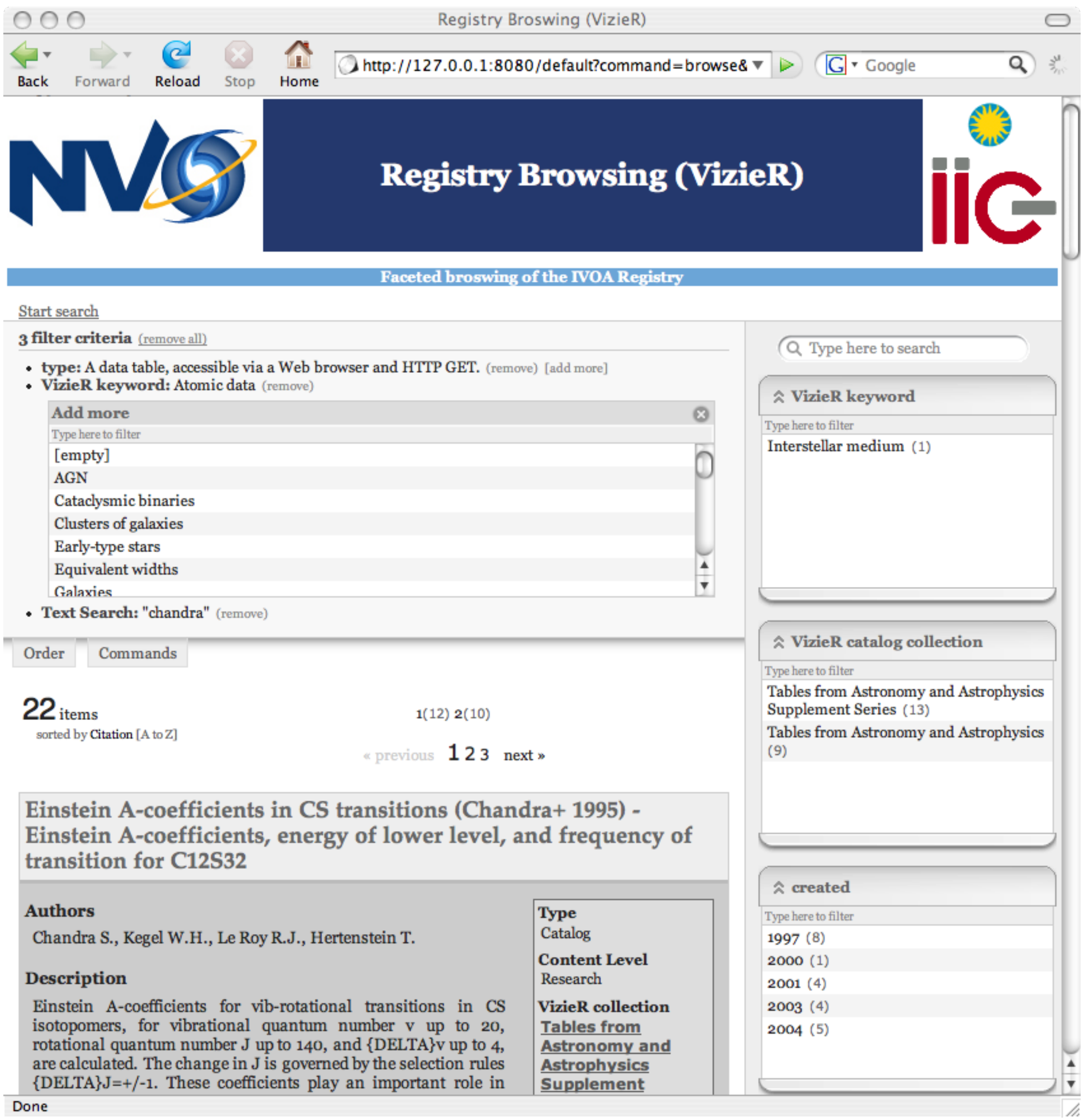


Figure 18: The filter list can be used to expand the search results by combining results using the union rather than the intersection of the search sets. In this case the VizieR keyword filter from Figure 15 has been expanded using the `[add more]` link.

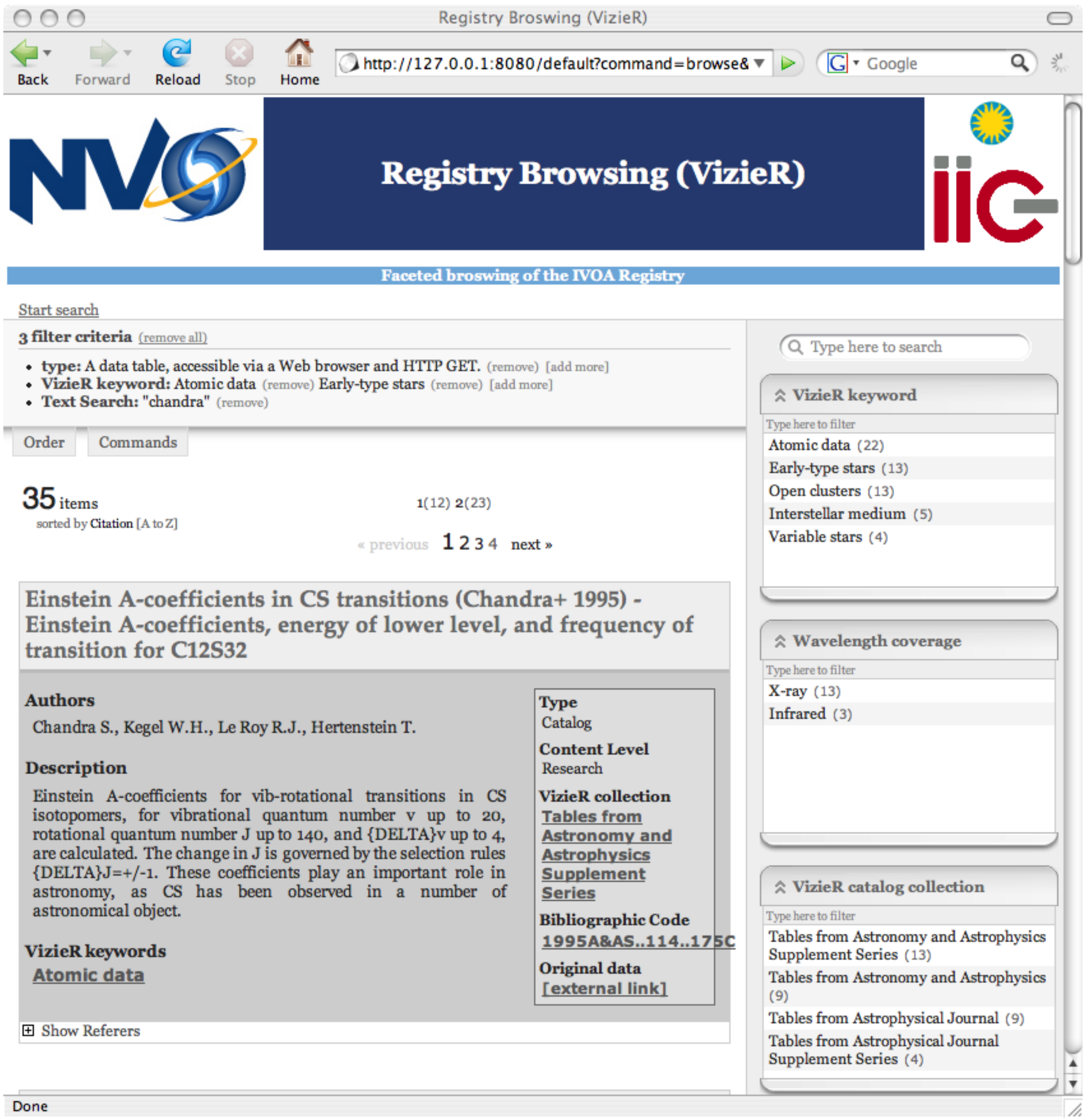


Figure 19: The result of combining the "Atomic data" and "Early-type stars" VizieR keyword filters from Figure 18.

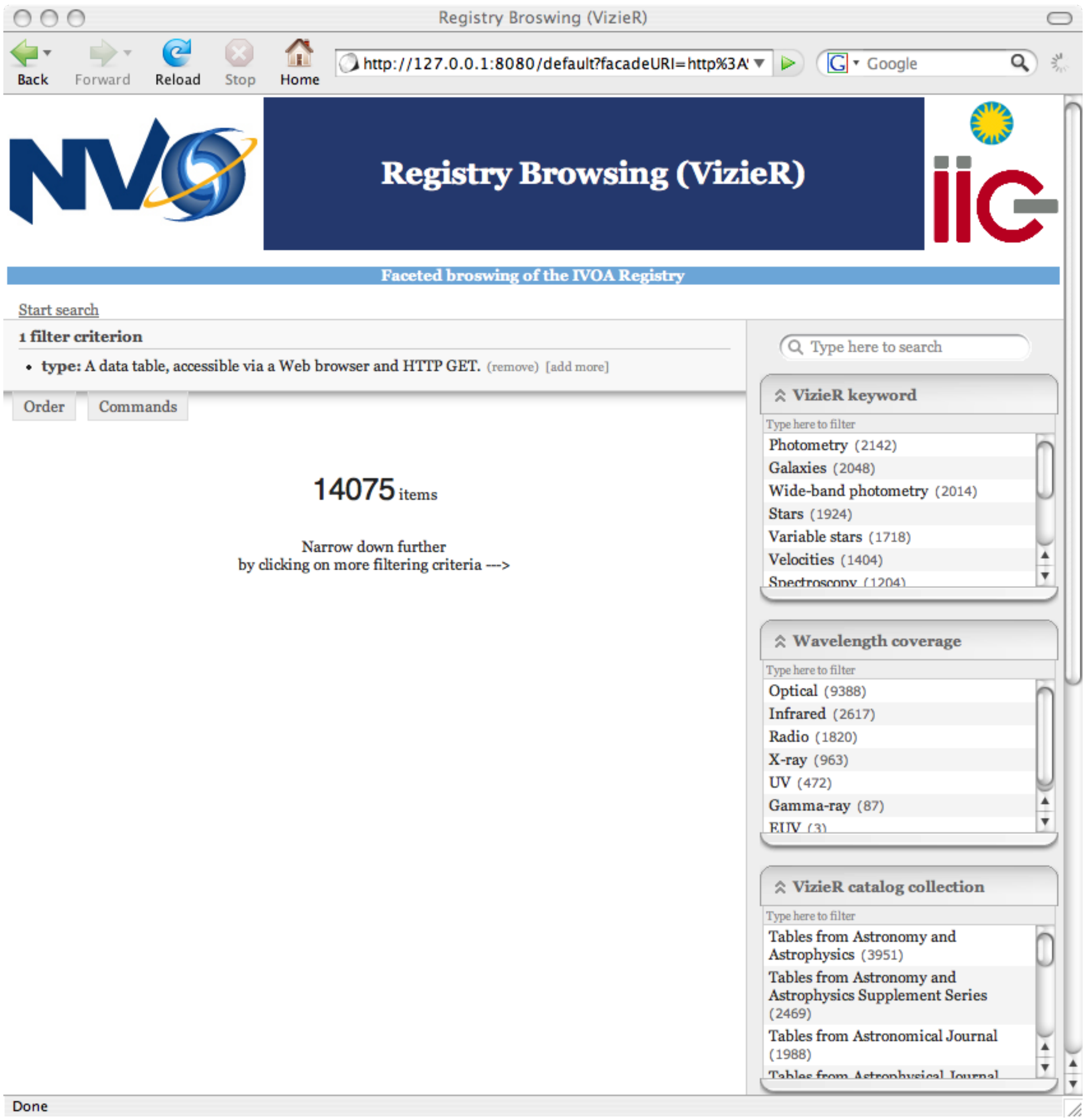


Figure 20: A new search has been started, this time selecting the "Resources" link in Figure 11. No individual items are displayed yet since there are too many records that match.

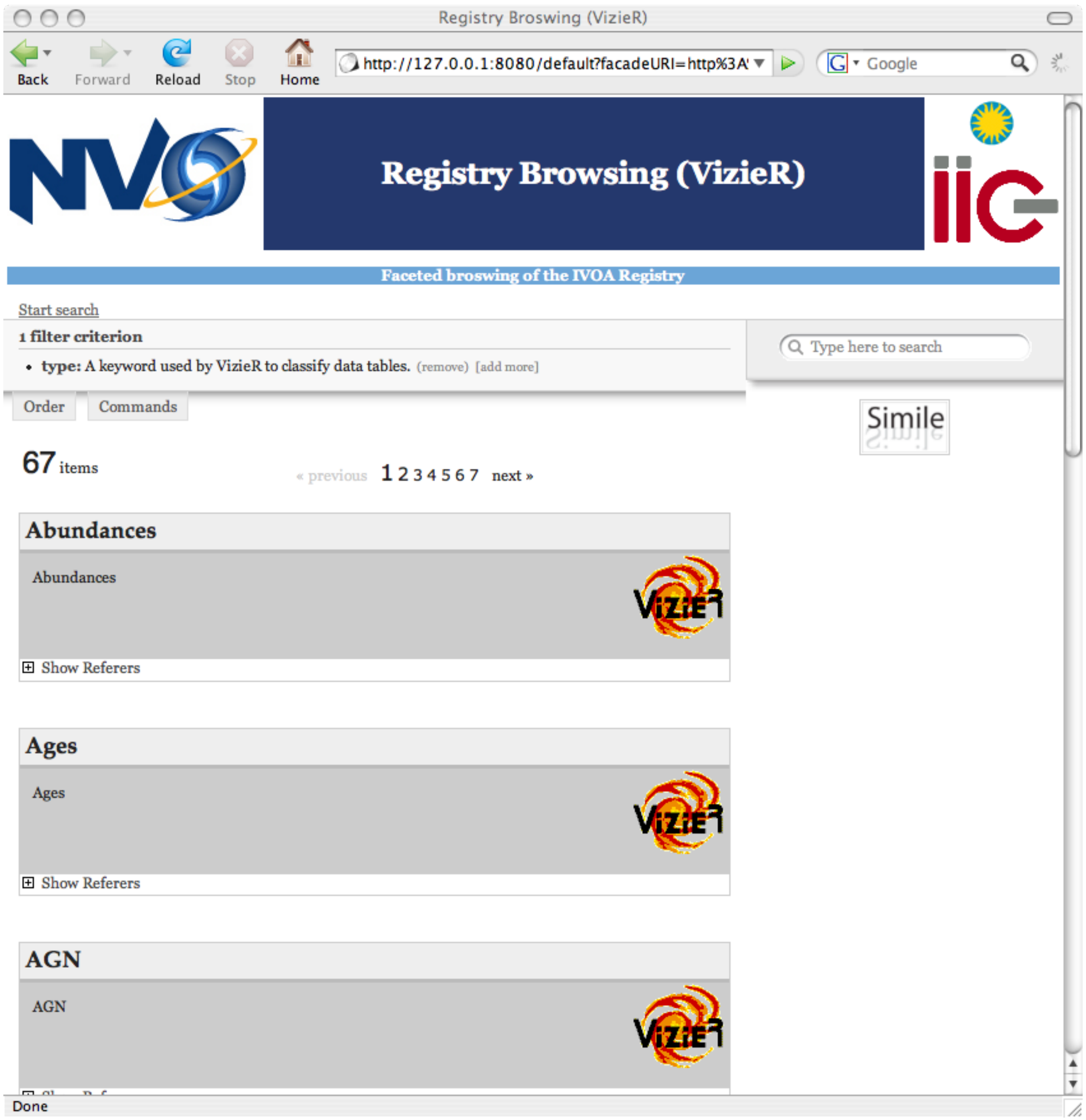


Figure 21: A new search has been started, this time selecting the "VizieR keywords" link in Figure 11.

B Namespaces

Table 1 lists the namespaces and their prefixes used in this document.

Prefix	Namespace	Type
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Standard
rdfs	http://www.w3.org/2000/01/rdf-schema#	Standard
xsi	http://www.w3.org/2001/XMLSchema-instance	Standard
dc	http://purl.org/dc/elements/1.1/	Standard
dct	http://purl.org/dc/terms/	Standard
oai	http://www.openarchives.org/OAI/2.0/	Standard
foaf	http://xmlns.com/foaf/0.1/	Standard
vr	http://www.ivoa.net/xml/VOResource/v0.10	IVOA
vs	http://www.ivoa.net/xml/VODataService/v0.5	IVOA
fresnel	http://www.w3.org/2004/09/fresnel#	Standard
facets	http://simile.mit.edu/2006/01/ontologies/fresnel-facets#	Longwell
flair	http://simile.mit.edu/2005/04/flair#	Longwell
facade	http://simile.mit.edu/2005/04/flair/facades#	Longwell
vor	urn:sao-iic-facets:voresources#	Development
dsp	urn:sao-iic-facets:display#	Development
key	urn:sao-iic-facets:keyword#	Development
dtyp	urn:sao-iic-facets:datatype#	Development

Table 1: The table lists the namespaces and their prefixes used in the document. The Type column indicates the stability of the namespace and its attendant vocabulary: “Standard” means that this is commonly used; “IVOA” means that the namespace is an IVOA standard; “Longwell” means that the namespace is used by the `Longwell` application; and “Development” means that the namespace was created during the project.

C The IVOA Registry

The IVOA has recognized that the metadata about Astronomical data and services is vitally important to any effort to allow wide-spread access and use of these resources. It has produced a document describing a model of such metadata, and the relationship with current and future web standards ([26]). This model presents a hierarchical system which is used to manage metadata, built around the concept of a "resource", which can essentially be anything that has an identity and can be described. This concept is used as a basis for describing organizations, data sets, data services, as well as more abstract ideas such as the area of sky covered by a survey or the instrumental set up used for an observation.

In order to describe resources, an XML schema has been developed ([32]); Listing 2 shows an example of a resource encoded using this schema. The original model re-uses many of the concepts from the Dublin Core MetaData Initiative ([3]), although the elements are within an IVOA namespace, such as `vr:content` and `vr:description`. The other required component is a standardized means of identifying a resource; the IVOA have introduced the IVOA identifier framework ([14]) which allows users to register globally-unique identifiers for their resources. There are two formats used to serialize these identifiers: one is defined using the XML Schema language as an XML complex type, and the other is written as a Uniform Resource Identifier using the `ivo:` scheme.

Data providers can create "registries", which contain a set of resources that allow external users to search and access the IVOA resource data. The mandated interface for these registries ([15]) provides access via both SOAP and REST interfaces; the former is described using a Web Services Description Language document and the latter uses the Open Archives Initiative Protocol for Metadata Harvesting ([20]). These registries are themselves registered in a "Registry of Registries" so that users can easily find them.

Two issues of relevance to this project are:

- The exact definition of metadata — namely what information should be stored in a registry — has been a contentious issue in the IVOA. The current approach is that the registry entries should be light weight, and that the services they refer to should be queried for further information using the nascent VO Support Interfaces (VOSI) proposal from the IVOA Grid and Web Services working group. For the search applications developed in this proposal it does not matter — other than efficiency — where the information is stored, just that the information is available in a machine-readable format.

- The VOResource schema used in both the VizieR and DataScope search applications (v0.10) has been deprecated, although it is currently in wide-spread use. The plan is to upgrade the majority of registries to the new version of the schema (v1.0) by the end of APril 2008. This update means that the XML to RDF mappings developed for the project, such as those given in Table 3, will need to be re-written. A proposed

RDF model of registry data using the v1.0 schema was released in September 2007 ([16]), and so could be used for the mapping.

D The prototype browser

D.1 Registry data

The STScI/JHU Virtual Observatory Data and Service Locator web site[9] was used to find all resources that contained both the words “bcg” and “magnitude”. The first two matching records are shown in Figure 7. The XML versions of the 25 matches were downloaded, using the links provided in the search results, and Listing 1 shows the contents of one of the records. Inspection of these records showed that they did not contain sufficient information for browsing; for instance the author list is missing and the description is truncated.

Listing 1: The XML version of one of the resources returned by the STScI Virtual Observatory Data and Service Locator service when searching for the words “`bcg magnitude`”. The document has been changed to simplify the display: the default namespace was changed from `http://www.us-vo.org` to `http://www.ivoa.net/xml/VOResource/v0.10`; the XML Schema and Schema instance namespace declarations were removed; the addition of the `vizcat` entity reference; and line breaks and white space have been added.


```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE ArrayOrResource [
<!ENTITY vizcat "http://vizier.u-strasbg.fr/cgi-bin/Cat" >
]>
<usvo:ArrayOfResource
    xmlns:usvo="http://www.us-vo.org"
    xmlns="http://www.ivoa.net/xml/VOResource/v0.10">
  <usvo:Resource updated="1999-03-17">
    <title>The Hamburg/SAO Survey for ELGs (Ugryumov+ 1999) - Coordinates,
    velocities, magnitudes and classification of emission-line
    galaxies</title>
    <shortName>J/A+AS/135/511/t</shortName>
    <identifier>ivo://CDS/VizieR/J/A+AS/135/511/table1</identifier>
    <curation>
      <publisher ivo-id="">CDS</publisher>
      <creator />
      <contributor>NOT PROVIDED</contributor>
      <version>NOT PROVIDED</version>
    </curation>
    <content>
      <subject>Galaxies:spectra</subject>
      <subject>Velocities</subject>
      <subject>Redshifts</subject>
      <referenceURL>&vizcat;?J/A+AS/135/511/table1</referenceURL>
      <contentLevel>Research</contentLevel>
      <description>
We present first results of the Hamburg/SAO Survey for Emission-Line
Galaxies (HSS therein, SAO - Special Astrophysical Observatory, Russia)
```


[9] http://nvo.stsci.edu/VORegistry/

```
which is based on the digitized objective-prism photoplates database of the
 Hamburg Quasar Survey (HQS). The main goal of this survey is the search
for emission-line galaxies (ELG) in order to create a new deep sample of
blue compact galaxies (BCG) in a large sky area. Another important goal of
this work is to search for new extremely low-metallicity galaxies. We
present the first results of spectroscopy obtained with the 2.2m telescope
at the German-Spanish Observatory at Calar Alto, and with the 6m telescope
at the Russian Special Astrophysical Observatory. The primary ELG candidate
 selection criteria applied were a blue continuum (near 4000{AA}) and the
presence of emission lines close to 5000A recognized on digitized prism
spectra of galaxies with magnitudes in the range B =16.0-19.5. The
spectroscopy resulted in the detection or/and quantitat
      </description>
    </content>
  </usvo:Resource>
</usvo:ArrayOfResource>
```

The records do however contain the unique IVOA identifier for each resource — it is `ivo://CDS/VizieR/J/A+AS/135/511/table1` for the resource shown in Listing 1 — which meant that it was possible to obtain the original data from the VizieR servive, this time using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[10] interface, which is supported by all the IVOA registries. The OAI-PMH `GetRecord` verb was used to extract the data for the individual records, using the IVOA identifier, and restricting the output to those elements in the `ivo_vor` namespace. The data returned for resource `ivo://CDS/VizieR/J/A+AS/135/511/table1` is shown in Listing 2.

Listing 2: The XML document returned by the VizieR data service for the resource listed in Listing 1. The document has been changed to simplify the display: the default namespace was changed from `http://www.openarchives.org/OAI/2.0/` to `http://www.ivoa.net/xml/VOResource/v0.10`; the namespace declarations — and corresponding `xsi:schemaLocation` values — were moved to the root element; entity references for `vizurl`, `cgibin`, and `tbl` were added; and line breaks and white space have been added.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE OAI-PMH [
<!ENTITY vizurl "http://vizier.u-strasbg.fr/" >
<!ENTITY cgibin "&vizurl;cgi-bin/" >
<!ENTITY tbl "J/A+AS/135/511/table1" >
]>
<oai:OAI-PMH
    xmlns:oai="http://www.openarchives.org/OAI/2.0/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:vr="http://www.ivoa.net/xml/VOResource/v0.10"
    xmlns:vs="http://www.ivoa.net/xml/VODataService/v0.5"
    xmlns:cs="http://www.ivoa.net/xml/ConeSearch/v0.3"
```

[10] http://www.openarchives.org/OAI/openarchivesprotocol.html

```
    xmlns="http://www.ivoa.net/xml/VOResource/v0.10"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
       http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
     http://www.ivoa.net/xml/VOResource/v0.10
       http://www.ivoa.net/xml/VOResource/VOResource-v0.10.xsd
     http://www.ivoa.net/xml/VODataService/v0.5
       http://www.ivoa.net/xml/VODataService/VODataService-v0.5.xsd
     http://www.ivoa.net/xml/ConeSearch/v0.3
       http://www.ivoa.net/xml/ConeSearch/ConeSearch-v0.3.xsd">
<oai:responseDate>2007-02-23T21:22:17Z</oai:responseDate>
<oai:request oai:metadataPrefix="ivo_vor"
    oai:identifier="ivo://CDS/VizieR/J/A+AS/135/511/table1"
    oai:verb="GetRecord">&cgibin;registry/vizier/oai_v0.10.pl</oai:request>
<oai:GetRecord>
  <oai:record><oai:header>
    <oai:identifier>ivo://CDS/VizieR/J/A+AS/135/511/table1</oai:identifier>
    <oai:datestamp>1999-03-17T07:11:09Z</oai:datestamp>
    <oai:setSpec>J.A_AS</oai:setSpec></oai:header>
  <oai:metadata>

<Resource xsi:type="vs:TabularSkyService" status="active">
<title>The Hamburg/SAO Survey for ELGs (Ugryumov+ 1999) - Coordinates,
velocities, magnitudes and classification of emission-line galaxies</title>
<shortName>J/A+AS/135/511/t</shortName>
<identifier>ivo://CDS/VizieR/J/A+AS/135/511/table1</identifier>
<curation>
  <publisher ivo-id="ivo://CDS">CDS</publisher>
  <creator>
  <name>Ugryumov A.V., Engels D., Lipovetsky V.A., Hagen H.-J., Hopp U.,
  Pustilnik S.A., Kniazev A.Yu., Richter G., Izotov Yu.I., Popescu
  C.C.</name>
  </creator>
  <contributor>Patricia Bauer [CDS]</contributor>
  <date role="creation">1999-03-17T07:11:09Z</date>
  <version>03-Dec-1998</version>
  <contact>
    <name>CDS support team</name>
    <address>CDS, Observatoire de Strasbourg, 11 rue de l'Universite,
    F-67000 Strasbourg, France</address>
    <email>question@simbad.u-strasbg.fr</email>
  </contact>
</curation>
<content>
  <subject>Galaxies:spectra</subject>
  <subject>Velocities</subject>
  <subject>Redshifts</subject>
  <source format="bibcode">1999A&amp;AS..135..511U</source>
  <referenceURL>&cgibin;Cat?&tbl;</referenceURL>
  <type>Catalog</type>
  <contentLevel>Research</contentLevel>
```

```
  <description>
We present first results of the Hamburg/SAO Survey for Emission-Line
Galaxies (HSS therein, SAO - Special Astrophysical Observatory, Russia)
which is based on the digitized objective-prism photoplates database of the
 Hamburg Quasar Survey (HQS). The main goal of this survey is the search
for emission-line galaxies (ELG) in order to create a new deep sample of
blue compact galaxies (BCG) in a large sky area. Another important goal of
this work is to search for new extremely low-metallicity galaxies. We
present the first results of spectroscopy obtained with the 2.2m telescope
at the German-Spanish Observatory at Calar Alto, and with the 6m telescope
at the Russian Special Astrophysical Observatory. The primary ELG candidate
 selection criteria applied were a blue continuum (near 4000{AA}) and the
presence of emission lines close to 5000A recognized on digitized prism
spectra of galaxies with magnitudes in the range B =16.0-19.5. The
spectroscopy resulted in the detection or/and quantitative spectral
classification of 74 emission-line objects. Of them 55 are newly
discovered, and 19 were already known as galaxies before. 11 of the latter
have redshifts and are known ELGs. For most of the known galaxies emission
line ratios were measured for the first time and an improved classification
 is presented. 47 objects are classified as BCGs, one as Sy2 galaxy, six as
 probable LINERs, and four as new QSOs. The remaining galaxies do not show
significant H{beta} and [OIII]4959,5007{AA} emission lines, and are likely
either low-ionization starburst or dwarf amorphous nuclear starburst
galaxies.
  </description>
</content>

<vs:coverage>
  <vs:spectral><vs:waveband>Optical</vs:waveband></vs:spectral>
</vs:coverage>

<interface xsi:type="vr:WebBrowser">
  <accessURL use="full">&cgibin;VizieR-2?-source=&tbl;</accessURL>
</interface>
<interface qtype="GET" xsi:type="vs:ParamHTTP">
  <accessURL
  use="base">&vizurl;viz-bin/votable/-dtd/-A?-source=&tbl;</accessURL>
  <vs:resultType>text/xml+votable</vs:resultType>
</interface>

<vs:table>
  <vs:name>J/A+AS/135/511/table1</vs:name>
  <vs:description>Coordinates, velocities, magnitudes and classification of
   emission-line galaxies</vs:description>
  <vs:column><name>recno</name> <unit> </unit> <ucd>RECORD</ucd>
    <description>Record number within the original table (starting from
    1)</description>
  </vs:column>
  <vs:column><name>Name</name> <unit> </unit> <ucd>ID_MAIN</ucd>
    <description>Name of objects</description>
```

```
  </vs:column>
  <vs:column><name>RAB1950</name> <unit>"h:m:s"</unit>
  <ucd>POS_EQ_RA_MAIN</ucd>
    <description>Right ascension (B1950)</description>
  </vs:column>
  <vs:column><name>DEB1950</name> <unit>"d:m:s"</unit>
  <ucd>POS_EQ_DEC_MAIN</ucd>
    <description>Declination (B1950)</description>
  </vs:column>
  <vs:column><name>HRV</name> <unit>km/s</unit> <ucd>VELOC_HC</ucd>
    <description>Heliocentric radial velocity</description>
  </vs:column>
  <vs:column><name>n_Bmag</name> <unit> </unit> <ucd>NOTE</ucd>
    <description>[*] *: Magnitudes are taken from the APM, Maddox et al.,
    Cat. &lt;J/MNRAS/246/433&gt;</description>
  </vs:column>
  <vs:column><name>Bmag</name> <unit>mag</unit> <ucd>PHOT_MAG_B</ucd>
    <description>Magnitude in B filter</description>
  </vs:column>
  <vs:column><name>BMAG</name> <unit>mag</unit>
  <ucd>PHOT_ABS-MAG_BAND</ucd>
    <description>Absolute B-magnitude</description>
  </vs:column>
  <vs:column><name>Type</name> <unit> </unit> <ucd>CLASS_MISC</ucd>
    <description>Classification (1)</description>
  </vs:column>
  <vs:column><name>Names</name> <unit> </unit> <ucd>ID_ALTERNATIVE</ucd>
    <description>Other names from NED</description>
  </vs:column>
</vs:table>
</Resource>
</oai:metadata></oai:record></oai:GetRecord></oai:OAI-PMH>
```


D.2 Facets

The data from VizieR was analyzed and a small set of facets chosen; Table 2 shows how these facets relate to the data provided by the service. The conversion was mainly automatic, in that the VizieR-provided data could be direclty used; the only change needed was to separate out the author list from a single, comma-separated, list of names into individual entries, one per author. The conversion was simplified by the fact that the number of resources was small and formed a homogeneous set — since each record represents a `TabularSkyService` resource which is defined by version 0.5 of the VO Data Services schema ([34]) to be "a service that interacts with one or more specified tables having some coverage of the sky, time, and/or frequency" — so that many of the complexities required to model the full set of IVOA schema could be avoided. For the prototype no attempt was made to include information about the table contents – such as column names, units,

Facet name	Location
Keyword	`Resource/content/subject`
Bibliographic code	`Resource/content/source[@format='bibcode']`
Spectral coverage	`Resource/vs:coverage/vs:spectral/vs:waveband`
Authors	`Resource/curation/creator/name`†
Data Type	`Resource/content/type`
Data Level	`Resource/content/contentLevel`
Data publisher	`Resource/curation/publisher`

† The `name` element contains a single, comma-separated, string of author names which had to be converted into individual values.

Table 2: The relationship between the facets chosen for the prototype browser and the original data. The Location column gives an XPath expression defining the position of the data relative to the `/oai:OAI-PMH/oai:GetRecord/oai:metadata` element, as shown in Listing 2. The default namespace for the expressions is the the VOResource namespace (`http://www.ivoa.net/xml/VOResource/v0.10`), and `vs` refers to the VODataService namespace (`http://www.ivoa.net/xml/VODataService/v0.5`).

descriptions, or Unified Content Descriptors (UCD)[11] values – in the search interface.

Version 1 of the `Exhibit` web-application framework was used to browse the data. This framework provides a simple means for browsing small amounts of structured data since all that is required is the conversion of the data to JavaScript Object Notation (JSON) format[12] and the creation of an HTML page. The faceted browsing is provided by JavaScript code and is run entirely client-side. This meant that it was very easy to set up, and the dataset could be searched using modern web browsers such as Firefox, Safari, and Opera. The start of the JSON-format data is shown in Listing 3, and mainly consists of keyword, value pairs representing the names and contents of the XML elements of the Resource element.

Listing 3: The JSON data file used to represent the VizieR matches. The `types` and `properties` entries control the appearance of the User Interface provided by `Exhibit`, in particular the facet labels. The `items` array contains the actual data; the listing contains an abbreviated version of the entry for the record `ivo://CDS/VizieR/J/A+AS/135/511/table1`.

```
{
 types: { "Resource": { pluralLabel: "Resources" } },
```

[11]http://www.ivoa.net/Documents/latest/UCD.html
[12]http://www.json.org/

```
 properties: {
  subject: { label:   "Keyword" },
  contentSource: { label:   "Bibliographic code" },
  creators: { label: "Authors" },
  spectralCoverage: { label: "Spectral coverage" },
  contentType: { label: "Data Type" },
  contentLevel: { label: "Data Level" },
  publisher: { label: "Data publisher" }
 },

 items: [
   {
   type: "Resource",
   ivoid: "ivo://CDS/VizieR/J/A+AS/135/511/table1",
   tableID: "J/A+AS/135/511/table1",
   catalogID: "J/A+AS/135/511",
   label: "The Hamburg/SAO Survey for ELGs (Ugryumov+ 1999) - Coordinates,
   velocities, magnitudes and classification of emission-line galaxies",
   shortName: "J/A+AS/135/511/t",
   publisher: "CDS",
   creators: [
    "Ugryumov A.V", "Engels D", "Lipovetsky V.A", "Hagen H.-J", "Hopp U",
    "Pustilnik S.A", "Kniazev A.Yu", "Richter G", "Izotov Yu.I", "Popescu
    C.C" ],
   subject: [ "Galaxies:spectra", "Velocities", "Redshifts" ],
   contentType: "Catalog",
   contentLevel: "Research",
   contentSource: "1999A&AS..135..511U",
   contentSourceFormat: "bibcode",
   spectralCoverage: [ "Optical" ],
   description: "We present first results of the Hamburg/SAO Survey for
   ..."
   },
   ...
] }
```

Figure 8a shows a truncated version of the initial page of the prototype, where the records are displayed on the left of the screen and the facets are shown on the right. Each facet contains a list of possible values, preceeded by the number of resources that contain that value, so there are 5 resources with the keyword “Galaxies”. The main display of the resources, within the colored boxes on the left, contains the information on each resource, includes links to the VizieR-provided services (“Description” and “Data”), and hides the abstract from the user by default (selecting the + symbol displays the contents). The records have been automatically grouped by the bibliographic code — the `contentSource` field in the JSON data — although `Exhibit` provides controls for changing the grouping and ordering of the selected data. To reduce development time, the value of the bibliographic code was not converted from the input format — such as `1999A&AS..135..511U` — to a more user-friendly form.

Several of the relations used for facets in the prototype, such as “Wavelength” and “Keywords”, are provided as filters on the main VizieR search page (Figure 5a). The principal advantage to the faceted approach is that the facets provide extra information to the user on how the data is distributed amongst the facet values, as can be seen by comparing Figures 8a and 5. This advantage disappears when the facet does not provide any useful discrminatory power for the current dataset; an example being the “Spectral coverage” facet in Figure 8a, where all the resources contain the same value of “Optical”.

Figure 8b shows how the display reacts when a user selects an item in a facet; in this case the “Galaxy” value in the “Keyword” facet. The main display is now restricted to the five resources that match this selection, and the values in the other facets are updated to show the range of values in the sub-sample. The “Keyword” facet retains the values of the full sample which allows values in this facet to be combined using “or” rather than “and”. With this interface it is possible to further qualify the search, by adding additional constraints on the facet values, or to broaden the search by removing the constraints. This can be compared to the traditional interface, as used by VizieR and shown in Figure 6, where there is limited choice for further exploration once a search has been performed. The STScI registry search does make basic use of the faceted search idea in that it will produce a facet for the `ResourceType` field when multiple values are present in the search results, as shown in the example in Figure 5b. This is not present in the original search (Figure 7) since the results all have the same `ResourceType` value.

E Resource Description Framework

The Resource Description Framework, commonly referred to as RDF, is designed to represent information in the World Wide Web, and its specification has reached recommendation status by the World Wide Web Consortium. There are six documents provided by the consortium that describe the semantics, syntax, and use of RDF; the RDF Primer document is one of these and provides links to the other five ([23]). A thorough understanding of the background and capabilities of RDF is not required for this report. What follows is a summary of the salient points of RDF as used by `Longwell`.

RDF represents knowledge about a subject as a graph of nodes and arcs which represents resources, their properties, and values. These graphs are stored as RDF statements which consist of three elements (and so are often referred to as triples), namely

Subject The item being discussed.

Predicate The property or characteristic of the subject.

Object The value of the property.

So, the statement "NGC 4038 is a member of the Antennae Galaxies" can be modeled in RDF as having a subject of "NGC 4038", predicate of "member of", and object "Antennae Galaxies", as shown in Figure 22. Although this example has used text to represent the elements of the RDF triple, the subject and predicate are actually stored as Uniform Resource Identifiers (URIs), and the object may be a URI or a plain text value, with an optional datatype. The use of URIs allows rich linking of datasets, since multiple statements can be made about the same subject, and these statements can come from multiple data sources. In many cases the URIs will use the http scheme, but this is not required, and for the data modeling in this project the universal resource name scheme was used, using a private namespace identifier of `sao-iic-facets` (e.g. Listing 6).

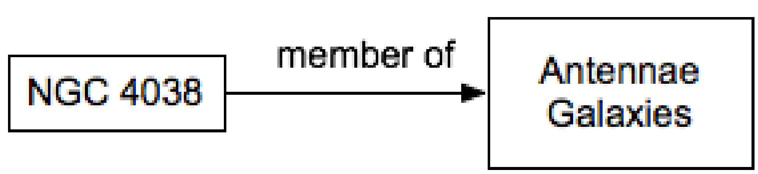


Figure 22: A simple RDF graph representing the statement "NGC 4038 is a member of the Antennae Galaxies". The arc between the nodes "NGC 4038" and "Antennae Galaxies" represents the predicate "member of" and is drawn from the subject to the object of the statement.

The RDF data model operates under the "Open World" assumption, which allows it to handle incomplete, or uncertain, knowledge at the expense of validation, since it states that if a statement can not be inferred from the current data set then it can not be assumed to be false. This assumption was not found to make a difference in the faceted browsing applications that were developed for the project, since the applications did not need to infer knowledge from the existing data. It could however be an issue for more-ambitious systems which try to combine data from multiple sources, or allow inferences based on the user's search request.

There are several data serialization formats for RDF that are supported by `Longwell`. For the project, and the examples used in this document, the Notation 3 (N3) form is used, since it is much more compact and readable than the XML format. The statement shown in Figure 22 could be written in N3 as shown in Listing 4, which also includes statements which give a human-readable label for each resource using the `label` predicate from the RDF Schema vocabulary ([24]). The document begins with two namespace definitions, indicated by the lines starting with the `@` character, which map the symbols `ex` and `rdfs` to the given URIs (as with XML documents the symbols are syntactic elements and it is only the URI that they are bound to which has any meaning). In this document the URI bound to the `ex` symbol has been chosen for illustrative purposes only, as are any of the elements within it such as `ex:memberOf`, whereas the `rdfs` namespace is a World Wide Web Consortium standard. In this format, single statements are written as `Subject Predicate Object .` and multiple statements with the same subject can be combined using the `;` symbol to avoid repeating the subject. Further details on the format can be found at the Notation 3 specification page ([18]).

Listing 4: The RDF statement shown in Figure 22 can be expressed in Notation 3 format as shown in this listing. Additional statements have been added, using the `rdfs:label` predicate, to add human-readable labels to each resource (the optional `@en` suffix is used to indicate that a string is written in the English language).

```
@prefix ex:            <urn:sao-iic-facets:example#> .
@prefix rdfs:          <http://www.w3.org/2000/01/rdf-schema#> .

ex:antennae_galaxies   rdfs:label "Antennae Galaxies"@en .

ex:ngc4038             ex:memberOf ex:antennae_galaxies ;
                       rdfs:label "NGC 4038" .
```

Although not shown in Listing 4, subjects can be considered to belong to one or more kinds or categories in a manner similar to the concept of Object-Orientated programming. In the RDF model the `rdf:type` predicate (where the `rdf` namespace refers to `http://www.w3.org/1999/02/22-rdf-syntax-ns#`), is used to declare that the subject is an instance of a given category, or class. For example, Listing 5 says that the URI `ex:ngc4038` is an instance of the `ex:Galaxy` class. In the RDF data model, resources such as `ex:ngc4038` can be instances of multiple classes, as shown in Listing 7. Although not a requirement, there is a strong convention — followed in this document – that class names begin with a capital letter, such as `ex:Galaxy`, whereas property and instance names begin with a lowercase letter, such as `ex:ngc4038`.

Listing 5: A resource representing the class of "Galaxy" is defined and given the URI `ex:Galaxy`. This class is then used to indicate that the `ex:ngc038` resource from Listing 4 represents a galaxy.

```
@prefix ex:             <urn:sao-iic-facets:example#> .
@prefix rdf:            <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:           <http://www.w3.org/2000/01/rdf-schema#> .

ex:Galaxy       rdf:type rdfs:Class ;
                rdfs:label "Galaxy"@en .

ex:ngc4038      rdf:type ex:Galaxy .
```

The configuration of `Longwell` is mostly done via RDF statements, as discussed in Appendix G, and the data that is browsed is required to be in RDF format (Appendix F).

F RDF modeling of IVOA Registry data

The contents of the VizieR registry were downloaded in XML format using the OAI-PMH service point[13] on February 23 2007. The `Identify`, `ListMetadataFormats`, `ListSets`, and `ListRecords` verbs were used, with the latter request restricted to elements in the IVOA VO Resource namespace (a constraint of `metadataPrefix='ivo_vor'`; the alternative would have been to use the standard `oai_dc` namespace which would have provided minimal information about each resource using elements from the Dublin Core set of namespaces). The request returned 14115 separate resources (i.e. `vr:Resource` elements), The format used for the `ListRecords` request is essentially the same as that used for `GetRecord`, which was discussed in Appendix D.1, and so the two formats will be considered to be the same in the discussion below (the differences are only in the metadata sections describing the OAI-PMH request and not in the IVOA Resource section).

F.1 VizieR Registry Browser

For the VizieR registry browsing application, each input record has the same type, namely the `TabularSkyService` resource from the IVOA Data Services schema ([34]). This simplified the initial conversion, since the IVOA registry schema are, in general, well suited to a basic conversion approach, with each resource being made an instance of the RDF class `vor:TabularSkyService`, the XML element names are turned into predicates — using elements from standard schema where applicable or inventing new URIs when not — and the XML element values are converted to RDF string literals. As the data was explored, and the browser developed, new RDF classes were added to enrich the browsing, and the relevant RDF objects changed from string literals to URIs. For this application the VODataService tabular information — the contents of the `vs:table` element, which includes the names and types of the columns — were excluded from the conversion.

Four types of URI were used to represent the resources, both as subjects and predicates. The first form comes from the IVOA resources themselves, as they all contain an identifier element whose value looks something like `ivo://CDS/VizieR/J/A+AS/135/511/table1` (this particular identifier represents the resource shown in Listing 2). These URIs were used to represent the RDF version of the resource; i.e. they were used as the subject of all the statements about a particular resource. The second form were the elements of commonly-used schema, such as the `dc:description` and `dct:created` predicates from the Dublin Core Metadata Initiative vocabularies ([3]). The third form were `http` URIs which were used to represent existing resources that could be accessed via a URL; examples included the URLs of VizieR catalog pages or the NASA/ADS page for a paper. The last type were the predicates and objects that were created during the data modeling. The Uniforn Resource Name ([30]) scheme was used for these, using an un-registered name-space identifier, to ensure that the URIs did not conflict with other extant resources. Examples include

[13] http://vizier.u-strasbg.fr/cgi-bin/registry/vizier/oai_v0.10.pl

`vor:TabularSkyService`, for representing a Tabular Sky Service Resource from the IVOA registry, and `key:agn` to indicate the X-ray keyword used by VizieR (the namespaces are listed in Table 1). In a production system these URIs would preferably use either IVOA-created or internet-standard vocabularies.

Listing 6 shows how the VizieR record for the entity `ivo://CDS/VizieR/J/A+AS/135/511/table1` (Listing 2) was modeled. Elements of standard vocabularies ([3]) were used to model many of the text elements, i.e. those with string literal subjects, such as `dc:title` and `dc:creator`. The remaining relationships were modeled using predicates specific to this project, such as `vor:contentLevel`, `vor:bibName`, and `vor:vizierkeywordLink`. The range of these predicates included both string literals and URIs.

Listing 6: RDF/N3 version of the IVOA Resource shown in Listing 2. Since several of the relations refer to URIs rather than string literals — such as `vor:vizierkeywordLink` — examples of these classes are included. The contents of the `dc:description` predicate have been elided since there has been no change over the value shown in Listing 2. String literals which may span multiple lines are enclosed by pairs of `"""` characters, rather a single quotation symbol, and the & character has been replaced by its hexadecimal value (%26) in URIs.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .
@prefix dct:    <http://purl.org/dc/terms/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix bib:    <urn:sao-iic-facets:bibliography#> .
@prefix key:    <urn:sao-iic-facets:keyword#> .
@prefix set:    <urn:sao-iic-facets:set#> .
@prefix wave:   <urn:sao-iic-facets:waveband#> .

<ivo://CDS/VizieR/J/A+AS/135/511/table1>
    rdf:type vor:TabularSkyService;
    vor:referenceURL
      <http://vizier.u-strasbg.fr/cgi-bin/Cat?J/A+AS/135/511/table1>;
    vor:bibLink bib:bc1999A_AS__135__511U;
    vor:vizierkeywordLink
      key:galaxies_spectra, key:redshifts, key:velocities;
    vor:viziersetLink set:j_a_as ;
    vor:spectralwavebandLink wave:optical;
    dc:contributor "Patricia Bauer [CDS]";
    dc:creator """Ugryumov A.V., Engels D., Lipovetsky V.A.,
Hagen H.-J., Hopp U., Pustilnik S.A., Kniazev A.Yu., Richter G.,
Izotov Yu.I., Popescu C.C.""";
    dc:publisher "CDS";
    dc:title """The Hamburg/SAO Survey for ELGs (Ugryumov+ 1999) -
Coordinates, velocities, magnitudes and classification of
```

```
emission-line galaxies""";
    dct:created "1999-03-17T07:11:09Z";
    vor:contentLevel "Research";
    vor:type "Catalog";
    dc:description
      """We present first results of the Hamburg/SAO Survey ...""" .

bib:bc1999A_AS__135__511U
    rdf:type vor:BibliographicReference ;
    vor:adsCFALink
<http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1999A%26AS..135..511U> ;
    vor:bibcodeLink <info:bibcode/1999A%26AS..135..511U> ;
    vor:bibName "Astronomy and Astrophysics, Supplement Series" ;
    vor:bibYear "1999" ;
    vor:bibVolume "135" ;
    vor:bibPage "511" ;
    vor:bibClass "periodical" ;
    rdfs:label "1999A&AS..135..511U" .

key:galaxies_spectra rdf:type vor:VizieRKeyword ;
    dc:subject "galaxies:spectra"@en ;
    dc:description "The galaxies:spectra VizieR keyword."@en ;
    rdfs:label "Galaxy spectra"@en .

set:j_a_as rdf:type vor:VizieRSet ;
    dc:subject "J.A_AS" ;
    rdfs:label
      "Tables from Astronomy and Astrophysics Supplement Series"@en .

wave:optical rdf:type vor:SpectralWaveband ;
    dc:subject "optical"@en ;
    dc:description "Optical wavelengths."@en ;
    rdfs:label "Optical"@en .
```

F.2 Comments and notes on the conversion

The modeling only had to be sensitive to the needs of a single application, since the RDF version of the registry data was only intended to be used by `Longwell`. This meant that is was often easier, and quicker, to create new predicates rather than use existing vocabularies. This is particularly true for the "bibliographic" relations — those beginning `vor:bib` in Listing 6 — since there are a number of proposed vocabularies in use to model bibliographic data but none appeared to be in prominent use at the time of the project. As an aid to development, predicates whose range was a URI, rather than string literal, were generally named with the suffix `Link` or `URL`, such as `vor:vizierkeywordLink` and `vor:referenceURL`.

The single biggest improvement to the conversion would have been to split up the author list — the `dc:creator` predicate — so that each author was represented by a RDF node, preferably using the FOAF vocabulary ([7], e.g. as a `foaf:Person`). This was not done in this project since the format used to represent individual names is not standardized enough to make processing the `vr:Resource/vr:curation/vr:creator/vr:name` element of the IVOA registry items a simple task. Similarly, there are open questions on how best to represent the author list; the simplest approach, which would likely have been adequate for `Longwell`, would have been to ignore the ordering, but this would not be suitable for many applications. Since the authors were identified by string literals, searches could be made on an Author's name in `Longwell`, but it was not possible to restrict a search to a particular Author, or to find out all tables that they had published if their name was not unique.

It was simple to convert some elements into separate resources, rather than string literals. As an example, the `vr:Resource/vr:content/vr:subject` elements from the original resource were converted into instances of the `vor:VizierKeyword` class. Labels and descriptive text were added to the instances using the `dc:subject`, `dc:description`, and `rdfs:label` predicates which are used by `Longwell` when displaying the items. The remaining elements were left as string literals either because it made no sense to convert to a URI (e.g. the title of a resource) or because in this situation the user would not have benefited from the conversion. Examples of the latter are the `dc:publisher`, `vor:contentLevel` and `vor:type` predicates shown in Listing 6. For the dataset assembled here, these predicates all had the same values — "CDS", "Research", and "Catalog" respectively — and so provided no discriminatory power if used as a facet. For larger, or more varied, samples these relations could benefit from the use of URIs as their objects: the DataScope browser (Appendix H), since it sources data from multiple repositories, does use URIs rather than literals to represent some of these relations.

Data for the ancillary classes, such as `vor:BibliographicResource` and `vor:Facility`, were extracted from the OAI-PMH data — normally by extracting the unique values — and then converted to RDF/N3 form. This process was mostly automatic, but additional manual steps were taken to add meaningful labels to some of the items. The bibliographic information used to create instances of the `vor:BibliographicReference` class was generated in a separate step. The bibliographic code for each resource — i.e. the contents of the `vr:Resource/vr:content/vr:source[@vr:format="bibcode"]` element — was found and fed to a small Perl script. This used the `Astro::Bibcode` module[14] to extract the relevant information, such as journal name, year, page, and create the bibligraphic objects. This conversion was not possible in all cases due to errors in the input data, one example being the record whose bibliographic code was `2003MNRAS.(in.press`. To avoid confusion, a private URI was used to represent the items rather than the URI extracted from the `vr:source` element (i.e. `bib:bc199A_AS__135__511U` rather than `info:bibcode/1999A&AS..135..511U`), even though the latter is defined by the Astrophysics Data System as a standard identifier for the resource ([11]).

[14] http://search.cpan.org/dist/Astro-Bibcode/

There is no guarantee in RDF that the name of a node — such as wave:optical and set:j_a_as — has any meaning in and of itself; it is the set of predicates and values that are attached to the node that provide the semantics for the label. When creating nodes it was therefore only necessary that there was no collision; in practice this was achieved using a mechanical conversion based on the dc:subject or rdfs:label value, where the text was converted to lower case and then normalized by converting illegal characters to _ ([30]) (this was partly necessary to work around an issue with the support for identifiers in RDF/N3 files in some of the Java libraries used by Longwell).

A close look at the vr:Resource/vr:content/vr:description element from Listing 2 shows that the text actually includes some simple markup, primarily "{AA}" to indicate the Å symbol, and examples from other resources include "3x10^-16^ergs/s/cm^2^" and "(2004, Cat. <J/A+A/426/11>)". There is no way for the publisher of the records to indicate either what form of mark up is being user or even if the entries contain additional mark up (since the content type of such elements is likely to be restricted to strings). For an individual Registry publisher, such as CDS in this example dataset, it may be possible to find out, or guess, the syntactic rules for mark up and so either filter it out or convert it into a more useful form. This would not be feasible if the application were to include data from multiple publishers.

It was found that the VizieR records, accessible via the web site, contained extra information beyond that provided in the registry records. This information provides, in general, extra value to the records and is likely to have been generated by the VizieR team; examples include extra keywords and links to related data sets, as can be seen by looking at the VizieR page[15] for the catalog shown in Listing 2. The reason for some of this omission is that the IVOA resource schema may not support the inclusion of such information. It does mean, however, that it may not possible to match the information content available to users of the home services of a particular registry provider.

The conversion from the VizieR data, extracted using the OAI-PMH protocol, to vor:TabularSkyService items in RDF/XML format was performed using a simple XSLT document. The ancillary classes, and their instances, such as vor:Facility, were created using a variety of methods, since such conversion generally required some manual intervention. A small number (40) of records from the input data were excluded because they referred to meta data about the VizieR collection rather than any Astronomical catalogs; they all had vr:subject elements with a value of _META_.

F.3 Mapping between IVOA and RDF schema

Table 3 gives the mapping used when converting the VizieR registry records, retrieved using the OAI-PMH protocol, to RDF form for use in Longwell. Table 4 shows the number of items of each class that were created from this data set, and Figure 9 shows how Longwell

[15]http://vizier.u-strasbg.fr/cgi-bin/Cat?J/A+AS/135/511/table1

displays this data. The total size of the RDF database was 29 Mb and the size of the OAI-harvested input XML file was 82 Mb.

XML record location	RDF predicate	RDF class
oai:setSpec†	vor:viziersetLink	vor:VizieRSet
vr:Resource[@xsi:type="vs:TabularSkyService"]‡		vor:TabularSkyService
vr:title	dc:title	
vr:curation/vr:creator/vr:name	dc:creator	
vr:curation/vr:publisher	dc:publisher	
vr:curation/vr:contributor	dc:contributor	
vr:curation/vr:date[@vr:role="creation"]	dct:created	
vr:content/vr:subject	vor:vizierkeywordLink	vor:VizieRKeyword
vr:content/vr:source[@vr:format="bibcode"]	vor:bibcodeLink	vor:BibliographicReference
vr:content/vr:referenceURL	vor:referenceURL	
vr:content/vr:type	vor:type	
vr:content/vr:contentLevel	vor:contentLevel	
vr:content/vr:description	dc:description	
vs:facility	vor:facilityLink	vor:Facility
vs:coverage/vs:spectral/vs:waveband	vor:spectralwavebandLink	vor:SpectralWaveband
vr:shortName		
vr:curation/vr:publisher/@vr:ivo-id		
vr:curation/vr:version		
vr:curation/vr:contact		
vr:interface		
vs:table		

† The base is /oai:OAI-PMH/oai:ListRecords/oai:record/

‡ The base is /oai:OAI-PMH/oai:ListRecords/oai:record/oai:metadata/

Table 3: The mapping between elements from the XML data retrieved from the VizieR registry (the "XML record location" column, given as an XPath expression), the RDF predicate used to model the relation (if any), and the RDF class used to model the data (if any). Unless otherwise noted the base (i.e. XPath context node) for the XML locations is /oai:OAI-PMH/oai:ListRecords/oai:record/oai:metadata/vr:Resource/. Elements that were not converted are listed with no values in both of the RDF columns. The value of the vr:identifier record is used as the URI that identifies the vor:TabularSkyService instance. Not all records contained all the elements given in this Table.

RDF class	Number of items
`vor:TabularSkyService`	14075
`vor:BibliographicReference`	5608
`vor:VizieRKeyword`	67
`vor:Facility`	34
`vor:VizieRSet`	20
`vor:SpectralWaveband`	7

Table 4: The columns give the RDF class name and the number of instances of the class created from the VizieR data set.

G Longwell configuration for IVOA Registry data

As Longwell is a Java web application, some basic configuration is done using Java property files, but the main logic used to define what items can be browsed, and how the data is displayed to the user, is done with a combination of RDF statements and Cascading Style Sheet rules. There are three such areas which are described below. The configuration used is based on the basic Longwell profile; the newer "longwell-csi" profile, which includes design elements from the Exhibit framework, was not available in the version of Longwell used for this project (2.4.2).

G.1 Defining the data types that can be searched

Subjects which are instances of the flair:Facade class are used to define the set of data types that can be searched by Longwell. Listing 7 shows how the vor:TabularSkyService class is defined to be one of the top-level classes for searching, as well as providing several text labels that are used in the interface. The resulting start page is shown in Figure 9.

Listing 7: Facade definition for Tabular Sky Service Resources. The statements use a mix of standard namespaces (rdf, rdfs, and dc), Longwell-specific namespaces (flair and facade) and project-specific ones (vor).

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix flair:  <http://simile.mit.edu/2005/04/flair#> .
@prefix facade: <http://simile.mit.edu/2005/04/flair/facades#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .

vor:TabularSkyService
        rdf:type        rdfs:Class ;
        rdfs:label      "A resource representing a data table."@en .

facade:vorTSSResource
        rdf:type                flair:Facade ;
        rdf:type                flair:QueryBasedFacade ;
        dc:description  "A facade for browsing IVOA Resources."@en ;
        flair:shortLabel        "Resources"@en ;
        flair:contentDescription        "IVOA Resources."@en ;
        flair:restriction [
                flair:predicate         rdf:type ;
                flair:object            vor:TabularSkyService ; ] .
```

G.2 Facet configuration

When a user selects a set of results to display, through either the text search or data type choice shown in Figure 9, the system has to chose what properties of the selected items should be displayed as facets. This is defined using statements such as the ones shown in Listing 8, which uses the `Longwell` `fresnel-facets` vocabulary to limit the primary set of facets for instances of the `vor:TabularSkyService` class to the following, using the `facets:facets` predicate: `vor:vizierkeywordLink`, `vor:spectralwavebandLink`, `vor:viziersetLink`, and `dct:created`. This set of facets is shown in Figure 1. The `facets:hides` predicate is used to mark those predicates that not suited for use as a facet; this normally happens when the values of the relationship are unique to each record. For this listing this includes items such as the title of the record and the URL of the VizieR page. Any remaining predicates of `vor:TabularSkyService` resources — i.e. those not mentioned in either the `facets:facets` or `facets:hides` statements — will be included at the end of the facet display, with their contents hidden by default (these facets can be expanded by the user if required).

Listing 8: Facet definition for Tabular Sky Service records.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix facets: <http://simile.mit.edu/2006/01/ontologies/fresnel-facets#>
.
@prefix dc:     <http://purl.org/dc/elements/1.1/> .
@prefix dct:    <http://purl.org/dc/terms/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix dsp:    <urn:sao-iic-facets:display#> .

dsp:tssFacets rdf:type facets:FacetSet ;
        rdfs:label "Facets for TabularSkyService resources."@en ;
        facets:types ( vor:TabularSkyService ) ;
        facets:facets (
                vor:vizierkeywordLink
                vor:spectralwavebandLink
                vor:viziersetLink
                dct:created ) ;
        facets:hides (
                dc:description
                dc:title
                dc:contributor
                vor:referenceURL
                vor:adsCfALink ) .
```

G.3 Data Display

The most complicated part of configuring Longwell is the system used to define the display of resources, namely what information is provided to a user once a set of records has been identified. Longwell uses the Fresnel display vocabulary ([8]) to declare what properties of a resource are to be displayed — via the fresnel:Lens class — and their formatting, using the fresnel:Format class and Cascading Style Sheet rules. A partial example of the statements used to control the appearance of vor:TabularSkyService resources is given in Listing 9.

Listing 9: Fresnel lens definition for Tabular Sky Service Resources. Longwell uses items which are instances of the fresnel:Lens class to determine what properties of a resource are displayed to the user; in this case the example defines the display elements for items that belong to the vor:TabularSkyService class. This lens controls which properties are displayed, which should be hidden, and allows for limited customization of the output via the fresnel:use and fresnel:sublens properties. The styling of the display is defined using CSS rules, the location of which are given by the fresnel:stylesheetLink predicate of the associated fresnel:Group.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix facets: <http://simile.mit.edu/2006/01/ontologies/fresnel-facets#>
.
@prefix fresnel:        <http://www.w3.org/2004/09/fresnel#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .
@prefix dct:    <http://purl.org/dc/terms/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix dsp:    <urn:sao-iic-facets:display#> .

dsp:tssLens rdf:type fresnel:Lens ;
        rdfs:label "Lens for TabularSkyService resources."@en ;
        fresnel:purpose fresnel:defaultLens ;
        fresnel:classLensDomain vor:TabularSkyService ;
        fresnel:showProperties (
                vor:type
                vor:contentLevel
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property vor:viziersetLink ;
                fresnel:use dsp:vsetLabelFormat ]
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property vor:bibLink ;
                fresnel:sublens dsp:bibLabelLens ]
                vor:referenceURL
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property dc:creator ;
                fresnel:use dsp:creatorFormat ]
```

```
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property dc:description ;
                fresnel:use dsp:descriptionFormat ]
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property vor:vizierkeywordLink ;
                fresnel:use dsp:vkeyLabelFormat ]
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property vor:spectralwavebandLink ;
                fresnel:use dsp:vwaveLabelFormat ]
                [ rdf:type fresnel:PropertyDescription ;
                fresnel:property vor:facilityLink ;
                fresnel:use dsp:facLabelFormat ] );
        fresnel:hideProperties ( dc:title rdf:type ) ;
        fresnel:group dsp:grp .

dsp:grp rdf:type fresnel:Group ;
        fresnel:resourceStyle "vo-resource" ;
        fresnel:propertyStyle "vo-property" ;
        fresnel:stylesheetLink "styles/vor.css" .

dsp:vsetLabelFormat rdf:type fresnel:Format ;
        fresnel:propertyFormatDomain rdfs:label ;
        fresnel:propertyStyle "vor-info" ;
        fresnel:label "VizieR collection";
        fresnel:group dsp:grp .

dsp:creatorFormat rdf:type fresnel:Format ;
        fresnel:propertyFormatDomain dc:creator ;
        fresnel::propertyStyle "vor-single" ;
        fresnel:label "Authors";
        fresnel:group vord:gr .

dsp:bibLabelLens rdf:type fresnel:Lens ;
        fresnel:purpose fresnel:labelLens ;
        fresnel:classLensDomain vor:BibliographicReference ;
        fresnel:showProperties ( rdfs:label ) ;
        fresnel:hideProperties  ( dc:subject ) ;
        fresnel:group dsp:gr .
```

The `fresnel:showProperties` predicate lists the relations of the given class, as given by the `fresnel:classLensDomain` relation, that should be displayed, and allows simple formatting rules to be applied, such as what the text of the label should be and what CSS class names will be added to the HTML elements. Figure 23 shows the output of `Longwell` for the VizieR record given in Listing 2. CSS rules have been used to group the first five elements — Type to Original data — into a side bar, with the main area used to display the authors, description, and associated information such as dspectralwavelength coverage. Links are given for those predicates such as `vor:bibLink`, whose object has been given as a resource rather than a text literal. This allows users to explore the data, since they can select these links to see what associated information is available; in this case selecting the "Bibliographic Code" entry takes the user to the item shown in Figure 24.

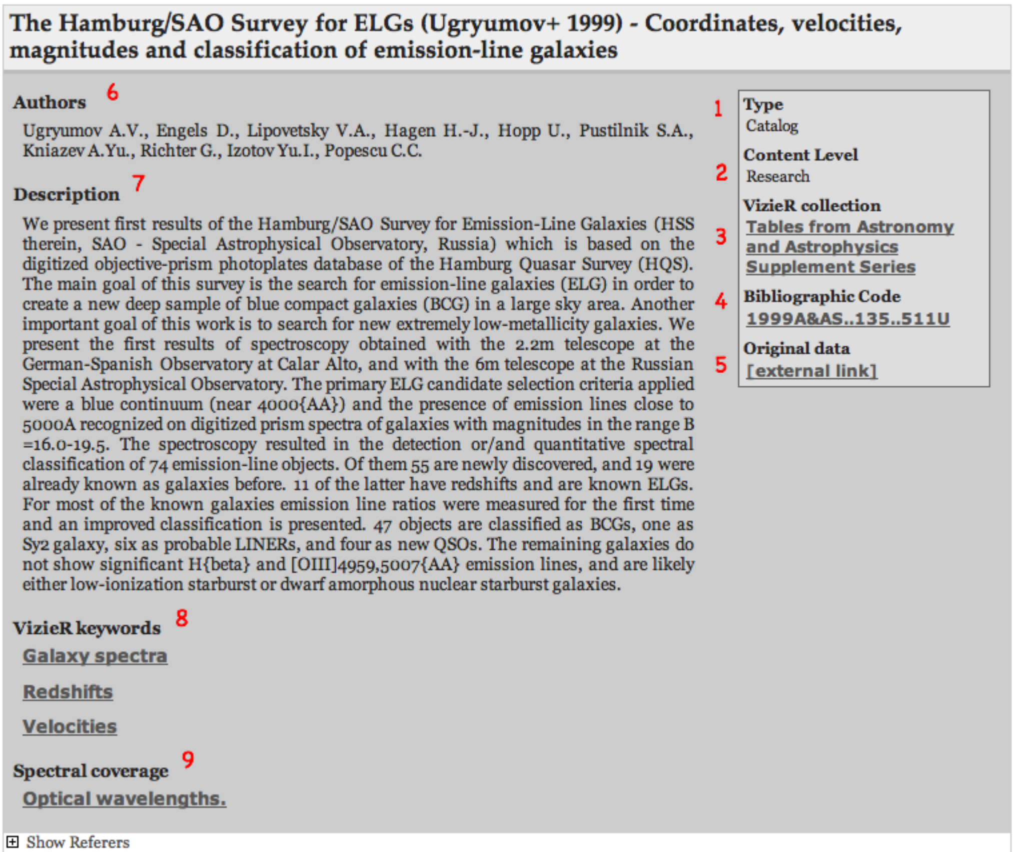


Figure 23: The `Longwell` display of the VizieR record given in Listing 2. The numbered figures are used to indicate how the display elements are related to the `fresnel:showProperties` predicate from the lens definition given in Listing 9. The numbers 1 to 9 refer to the properties `vor:type`, `vor:contentLevel`, `vor:viziersetLink`, `vor:bibLink`, `vor:referenceURL`, `dc:creator`, `dc:description`, `vor:vizierkeywordLink`, and `vor:spectralwavebandLink` respectively (this particular record does not contain a `vor:facilityLink` predicate). As there are no items that refer to this record, the "Show Referers" section has been left closed.

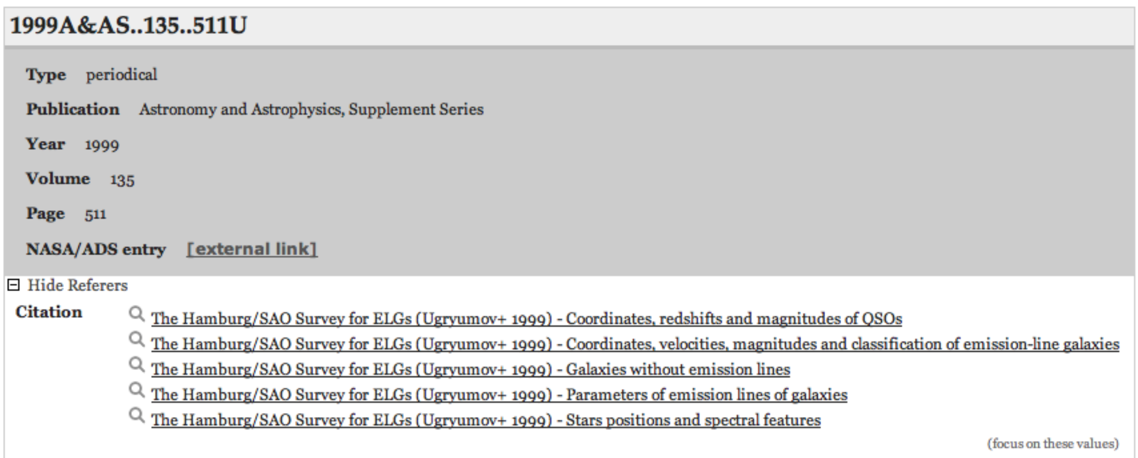


Figure 24: Following the "Bibliographic Code" link from the display shown in Figure 23 takes the user to this record, which is for a record of the `vor:BibliographicReference` class. In this case the lens, and associated CSS rules, is much less complicated, since the information stored is much simpler (just a sub-set of the bibliographic information of the paper from which the table was taken). The "Referers" section has been opened up to display all the tables that refer to this paper (i.e. these are all the tables taken from this paper); selecting one of these links would take the user back to a display similar to Figure 23.

H RDF modeling of NVO DataScope Application data

This appendix describes the modeling and display of data taken from the NVO DataScope search page for a single astronomical query. Since this part of the project was not completed, the results are much more experimental than the VizieR browser presented in Appendix F.

H.1 Accessing the DataScope data

The NVO DataScope application[16] was used to search for data within 0.1 degrees of the target name `Antennae` (which resolved to a postion of `12`^{h} `01`^{m} `52.48`^{s} `-18`$^{\circ}$ `52' 02.9"`). Figure 25 shows some of the resources returned by the search.

For each resource, such as `HST Previews`, an XML representation of the tabular data (Figure 25b) — which uses the IVOA VOTable format ([33]) — and a text version of the metadata (Figure 25c) were retrieved. Although many of the resources contain image data, such data is only provided by calling an external service (i.e. by referencing a HTTP URL), and so the image data was not included in the process (information about the images, such as size and location, and the URL to access the data, are stored in the VOTables).

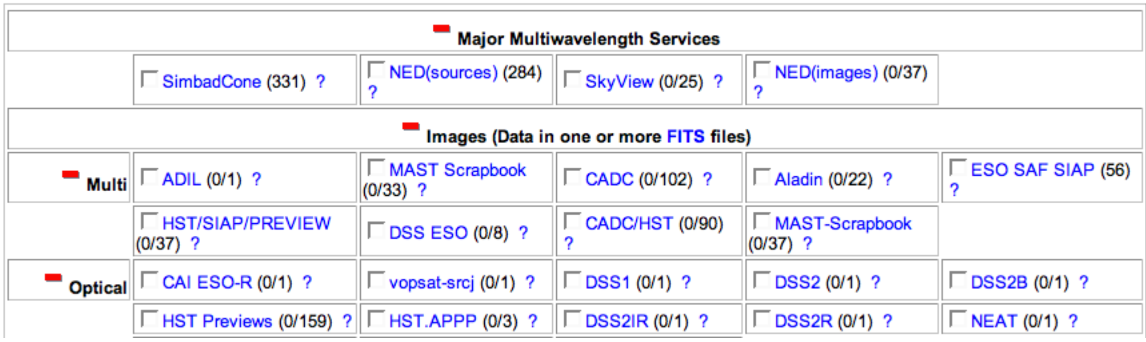
The retrieval step had to be done manually since DataScope is not intended to be used in a non-interactive manner. This resulted in 144 data files and 143 metadata files, as one resource did not contain any metadata; this table was therefore removed from the sample. A further data set was found to contain too many rows (8130) for the processing system and was also removed from the sample; it does not affect the scientific integrity of the sample since this record gave the pointing direction of the HETE-2 satellite as a function of time while it fell within 0.1 degrees of the Antennae galaxies. The final sample consists of 142 resources.

To simplify the modeling, no attempt was made to record information about those services that either did not return data (e.g. because they did not cover the requested part of the sky or they contained no data for this location), or those services that failed (e.g. the connection timed out for some reason or there was an error). This is a limitation of the current approach, since it is important to distinguish between "a source was not detected at a particular wavelength" and "this survey does not cover the area of sky in which the source is located".

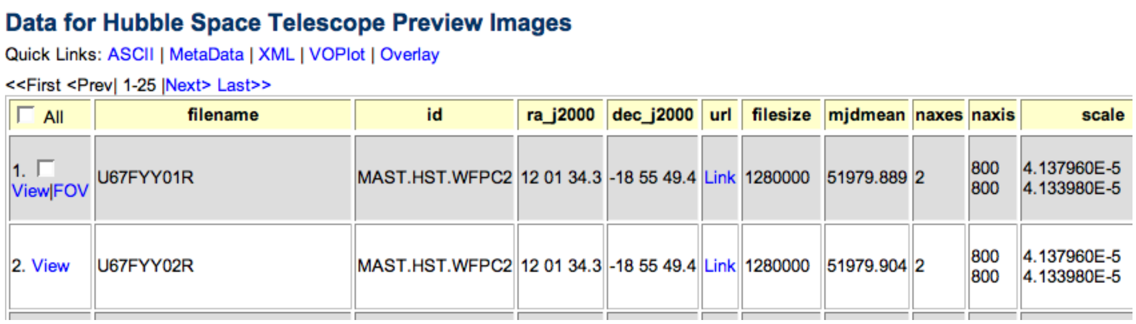
H.2 Modeling the metadata

The metadata document for a resource contains much of the IVOA resource information for the service — such as Title, Description, and Coverage information — in a simple text format. This was converted to RDF using a Perl script to extract the data, and the RDF vocabulary from the VizieR Registry browser (Appendix F) was re-used where

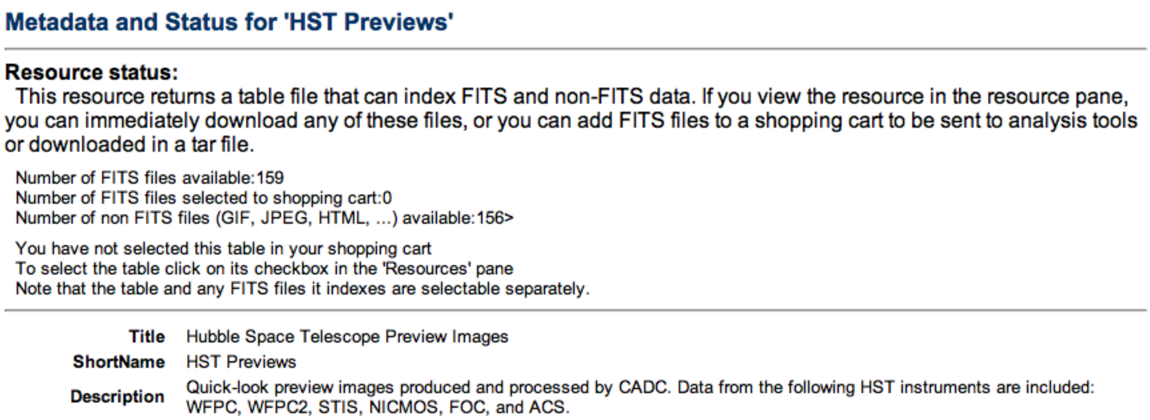
[16] http://heasarc.gsfc.nasa.gov/vo/



(a) DataScope Resource display



(b) DataScope Data display for the `HST Previews` resource



(c) DataScope Metadata display for the `HST Previews` resource

Figure 25: The results of a DataScope search for the name `Antennae`. Figure 25a shows the start of the resource list, which indicates what databases contain data for the search term. The display is arranged so that similar resources — such as optical imaging data — are listed together, in a basic form of faceted browsing. Figure 25b shows part of the data display for the `HST Previews` resource, and Figure 25c shows part of the metadata for this resource.

possible. Some predicates were changed so that they referred to URIs rather than string literals, such as `dc:publisher` and `vor:typeLink`; and some relations required new predicates, e.g. `vor:instrument`. Further consolidation between the two vocabularies — for example combining the VizieR "keyword" entries with the DataScope "subject" ones — may make sense but there was no time to attempt this. One file contained invalid mark-up (unescaped HTML code) which had to be manually removed before it could be processed. A converted file is shown in Listing 10.

Listing 10: RDF/N3 version of the metadata for the HST Previews resource shown in Figure 25c.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix wave:   <urn:sao-iic-facets:waveband#> .
@prefix fac:    <urn:sao-iic-facets:facility#> .

@prefix dsr:    <urn:sao-iic-facets:datascope#> .
@prefix vlvl:   <urn:sao-iic-facets:validationlevel#> .
@prefix clvl:   <urn:sao-iic-facets:contentlevel#> .
@prefix srv:    <urn:sao-iic-facets:service#> .
@prefix type:   <urn:sao-iic-facets:type#> .
@prefix pub:    <urn:sao-iic-facets:publisher#> .
@prefix sub:    <urn:sao-iic-facets:subject#> .

dsr:hst_previews rdf:type vor:DatascopeResource ;
        vor:spectralwavebandLink wave:optical ;
        vor:facilityLink fac:stsci ;

        vor:typeLink type:archive ;
        vor:contentlevelLink clvl:research ;
        dc:publisher pub:mast ;
        vor:subjectLink sub:uv , sub:optical , sub:infrared_astronomy ;

        vor:tableLink <file://HST_Previews.150.xml> ;
        vor:validationlevelLink vlvl:2 ;
        vor:servicetypeLink srv:siap_archive ;

        vor:referenceURL <http://archive.stsci.edu/> ;
        vor:idLink <ivo://mast.stsci/siap/hst.previews> ;
        vor:serviceURL <http://archive.stsci.edu/siap/search.php?id=HST> ;

        vor:instrument "WFPC, WFPC2, STIS, NICMOS, FOC, ACS" ;
        dc:title "Hubble Space Telescope Preview Images" ;
        dc:contributor "Archive Branch, STScI" ;
        vor:shortName "HST Previews" ;
```

```
        dc:creator "CADC" ;
        vor:version "1.0" ;
        dc:description """Quick-look preview images produced and processed
by CADC. Data from the following HST instruments are included: WFPC, WFPC2,
STIS, NICMOS, FOC, and ACS.""" .
```

The `vor:tableLink` predicate does not match the pattern used by other statements since it refers to a `file` URI; this was used to explicitly link the metadata for a table to the representation of the table itself. Such a link could have been inferred, since the table data contains a reference to the metadata statements, but the inclusion of an explicit statement was beneficial during development. From the document shown in Listing 10, it can be seen that the `vor:instrument` predicate refers to a comma-separated list of instrument names. This was originally intended to be replaced by a set of URIs, but it was found that the range of descriptions for this field were too varied to allow such a conversion during the project (Table 5 lists the values found in this dataset).

H.3 Modeling the data

As the VOTable document structure ([33]) is moderately complex, and it was not clear how well `Longwell` would handle tabular data (both in terms of how it would display the data and the computational resources needed to store and process the information), the conversion to RDF was done in a piecemeal manner, with elements added as needed. Although the VOTable format is XML, the conversion did not use XSLT stylesheets, as with the VizieR registry entries, but a Perl script that used version 0.9 of the `Astro::VO::VOTable` module[17] to parse the VOTable data. This approach was taken since the process required some data conversion and manipulation which was easier to perform in Perl than XSLT.

The first complication is that of the namespaces and versions used for the VOTable documents. At the time of the project version 1.1 of the schema was the IVOA recommendation, with version 1.0 still in widespread use. Only 5 of the 142 documents used a namespace declaration: 1 used version 1.1 of the schema, 2 used version 1.0, and 2 used a URL that was not correct[18]. There were 134 documents that used a DOCTYPE declaration to indicate that they used version 1.0 of the format, and there were three that had neither a namespace or DOCTYPE declaration. The version numbers also varied, in that the correct values of `1.1` and `1.0` were only used 1 and 23 times respectively, whereas the incorrect value of `v1.0` was used 28 times, `1` was used once, and 69 documents used the wrong attribute for the version (they used `ID="v1.0"`)[18]. For the conversion software, these issues had to either be worked around (e.g. by looking for elements in various namespaces) or ignored where possible.

The data tables were modeled as an instance of the `vor:VOTable` class — which represents metadata about the table, such as the contents of the `DESCRIPTION` and `RESOURCE/PARAM`

[17] http://heasarc.gsfc.nasa.gov/classx/votable/

[18] This error was reported and has since been corrected.

Text	Count
Far and Near Ultraviolet microchannel plates, grism spectrograph	2
1 IR camera, 2 Spectrometers, 1 spectrophotopolarimeter	1
2MASS 3-channel camera	1
48” Oschin Schmidt Telescope	2
4k x 4k CCD camera	1
ACIS, HRC	1
ACS	1
ACS, WFPC2, WFPC, STIS, GHRS, FOS, FOC, HSP, FGS	1
Digitized film	1
ESO Schmidt Telescope, MAMA microdensitometer	1
FOC	1
FOS	1
FUV, NUV	1
GALEX	1
GHRS	1
HST/WFPC2	1
NICMOS	1
PMM Scans	1
STIS	1
Schmidt telescope	7
Spectrograph	2
VLA	2
Various	3
WFPC	1
WFPC, WFPC2, STIS, NICMOS, FOC, ACS	1
WFPC2	1
WFPC2, STIS, FOC, UIT	1
X-Ray cameras	1
null	3

Table 5: The set of values for the instrument field of the metadata for the Datascope resources, together with the number of times the text occurred in the 142 metadata entries.

XML elements — together with `vor:VOTableColumn` instances which represent the individual columns in the tables. The columns were labeled using the attributes of the relevant `FIELD` element; of the 2333 columns, 1536 contained a `name` attribute which was used, and the remaining 797 used the `ID` attribute (172 columns contained both attributes). Listing 11 shows part of the RDF for the HST Previews table: most of the links from the table to the columns have been removed, as have most of the columns themselves, to save space. The two major additions over previous RDF models developed in this project[19] are the use of blank nodes for the range of the `vor:paramLink` predicates, and the use of RDF lists — indicated by `( element element ... )` — to store the column data.

Listing 11: RDF/N3 version of the table data for the HST Previews resource shown in Figure 25b. Some data has been elided (indicated by ...) to save space.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix dsr:    <urn:sao-iic-facets:datascope#> .
@prefix vot:    <urn:sao-iic-facets:votable#> .
@prefix dtyp:   <urn:sao-iic-facets:datatype#> .
@prefix ucd:    <urn:sao-iic-facets:ucd#> .

<file://HST_Previews.150.xml> rdf:type vor:VOTable ;
        vot:metadataLink dsr:hst_previews ;
        rdfs:label "File: HST_Previews.150.xml" ;
        vot:votableDescription """Multimission Archive at STScI (MAST)
archive.stsci.edu  send questions to: archive@stsci.edu""" ;
        dc:description "MAST SIAP_KEYWORDS Search: 315 row(s) returned!" ;
        vot:paramLink [
                rdf:type vor:Param ;
                vot:ID "id" ;
                vot:name "id" ;
                vot:value "MAST.HST.%" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:name "pos" ;
                vot:value "180.468667,-18.867472" ;
                vot:datatype dtyp:float ;
                vot:ucdLink ucd:pos_eq ;
                vot:unit "degrees" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:name "size" ;
                vot:value "0.1500000007450581" ;
                vot:datatype dtyp:float ;
```

[19] These features are used in the configuration of `Longwell` (e.g. Listing 9).

```
                vot:ucdLink ucd:stat_fit_param ;
                vot:unit "degrees" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:ID "requestid" ;
                vot:name "requestid" ;
                vot:value "DS1178221239705" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:ID "ra" ;
                vot:name "ra" ;
                vot:value "180.38940836174 .. 180.54792563826" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:ID "dec" ;
                vot:name "dec" ;
                vot:value "-18.942472000373 .. -18.792471999627" ; ] ;
        vot:paramLink [
                a vor:Param ;
                vot:ID "verb" ;
                vot:name "verb" ;
                vot:value "1" ; ] ;
        vot:columnLink <file://HST_Previews.150.xml#col1> ;
...
        vot:columnLink <file://HST_Previews.150.xml#col28> .

<file://HST_Previews.150.xml#col2> rdf:type vor:VOTableColumn ;
        dc:title "id column of HST_Previews.150.xml" ;
        rdfs:label "id" ;
        vot:tableLink <file://HST_Previews.150.xml> ;
        dc:description """The instrument ID follows the convention:
Archive.Satellite.Instrument""" ;
        vot:name "id" ;
        vot:datatype dtyp:char ;
        vot:ucdLink ucd:meta_id ;
        vot:ucdLink ucd:instr ;
        vot:arraysize "*" ;
        vot:columndata ( "MAST.HST.WFPC2" "MAST.HST.WFPC2" ... ) .

<file://HST_Previews.150.xml#col3> a vor:VOTableColumn ;
        dc:title "ra_j2000 column of HST_Previews.150.xml" ;
        rdfs:label "ra_j2000" ;
        vot:tableLink <file://HST_Previews.150.xml> ;
        dc:description """Should be center of image. May not match the
Target RA in the FITS header of some HST #observations.""" ;
        vot:name "ra_j2000" ;
        vot:datatype dtyp:double ;
        vot:ucdLink ucd:pos_eq_ra ;
        vot:ucdLink ucd:meta_main ;
        vot:unit "degrees" ;
```

```
        vot:ref "myJ2000" ;
        vot:columndata ( "180.393" "180.393"  ... ) .

...
```

Blank nodes were introduced to see how well this technique worked, both for modeling the data and in how well `Longwell` could display such data. The parameter section of the VOTable was chosen for this experiment since there seemed to be little use in having explicit names for these objects. It turned out that `Longwell` could display these items as well as named nodes, since the class (`vor:Param`) could be used to identify the contents, but that it was not obvious how to use them as facets, and the project finished before this issue could be resolved.

The column data was included in the conversion, but this ended up not being used in `Longwell` due to problems encountered in displaying RDF lists (the first element would be displayed but not the rest of the data). It was unclear whether this was a modeling issue or a bug in `Longwell`, and time constraints meant that this issue was not fully investigated. One known issue with the model used is that the order of the columns within a table is not guaranteed to be preserved; this could be changed by adding further statements to these lists.

Values in VOTables may contain a `ucd` attribute which provides a semantic meaning to the value, by use of the IVOA Unified Content Descriptors controlled vocabulary ([29]). The data tables were found to use values from either the original format (UCD1) or the new format (UCD1+); values from the `VOX`/`NVOX` "vocabulary" introduced by the Simple Imaging Access Protocol (SIAP) were also found in tables using both UCD1 and UCD1+ keywords. Rather than handle all cases, the UCD1 keywords were converted to UCD1+ values using the suggested mapping[20] and the `VOX` values added to the list of UCD1+ symbols. With descriptions added from the approved list of UCD values[21], the UCD values were modeled using statements like those shown in Listing 12. Additional statements were planned to be added to indicate the grouping of the UCD values (e.g. to say that `ucd:pos_eq_ra` is a member of the positional data group of keywords) and the syntactical rules for the token. The handling of UCD keywords as shown in Listing 11 is not ideal since it loses the ordering information which is an important part of the semantics of these tokens. As none of the table columns contained unique-type (utype) attributes, which are described in the IVOA Characterisation Data Model[22] document, no RDF model was developed to handle these values.

[20] http://vizier.u-strasbg.fr/UCD/lists/ucd1-ucd1p.txt
[21] http://www.ivoa.net/Documents/latest/UCDlist.html
[22] http://www.ivoa.net/Documents/latest/CharacterisationDM.html

Listing 12: RDF/N3 version of the Unified Content Descriptor values (UCD1+). Included are definitions of the class used to model UCDs (`vor:UCD`) and the predicate that links to the terms (`vot:ucdLink`). The description for the UCD class has been shortened to save space.

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:     <http://purl.org/dc/elements/1.1/> .
@prefix dct:    <http://purl.org/dc/terms/> .

@prefix vor:    <urn:sao-iic-facets:voresources#> .
@prefix vot:    <urn:sao-iic-facets:votable#> .
@prefix ucd:    <urn:sao-iic-facets:ucd#> .

ucd:pos_eq_ra rdf:type vor:UCD ;
        dc:description "Right ascension in equatorial coordinates" ;
        rdfs:label "pos.eq.ra" .

ucd:meta_main rdf:type vor:UCD ;
        dc:description "Main value of something" ;
        rdfs:label "meta.main" .

vor:UCD rdf:type rdfs:Class ;
        dc:title "Unified Content Descriptors" ;
        dct:provenance """Description taken from
http://cdsweb.u-strasbg.fr/doc/UCD/inVizieR.htx""";
        dc:description """The UCD (Unified Content Descriptors)  is a
categorization of any parameter present in a column of a table or catalogue
in a predefined scheme. ...""" .

vot:ucdLink rdf:type rdf:Property ;
        rdfs:range    vor:UCD ;
        rdfs:label    "Type" .
```

Data types for parameters and columns were modeled as members of the `vor:DataType` class, and Table 6 lists the instances that were needed. The "Int16" and "Int32" values are not valid according to the VOTable schema but were found in one file so included in the conversion. A similar conversion was performed for units, and the two main issues found were: firstly, virtually all files used the correct attribute name of `unit` but two files used `units` instead, and secondly, a range of values were found for the same unit, as shown in Table 7. Instead of The units could have been modeled by specifying them as XML Schema Datatype constraints on the values — both parameter and column — themselves, but it was not clear how well this was supported by `Longwell` and also whether it would be useful to have this information available as a predicate for use in a facet.

The last problem that was found with the VOTables was that the `LINK` element — used to indicate a URL of some other form of reference — was not always used correctly; the

Type	URI
Char	`dtyp:char`
Double	`dtyp:double`
Float	`dtyp:float`
Int	`dtyp:int`
Int16	`dtyp:int16`
Int32	`dtyp:int32`
long	`dtyp:long`
Short	`dtyp:short`
Unsigned byte	`dtyp:unsignedbyte`

Table 6: Data type values from the VOTables and their mapping as URIs.

Original	Converted
deg	degrees
degree	degrees
"d:m:s"	degrees (sexagesimal notation)
dd mm ss	degrees (sexagesimal notation)
hrs	hours
"h:m:s"	hours (sexagesimal notation)
hh mm ss	hours (sexagesimal notation)
yr	years
s	seconds
arcmin	arcminutes
arcsec	arcseconds
m	meters
pix	pixels
[0.1arcmin]	0.1 arcminute
0.1arcmin	0.1 arcminute
km/sec	km/s
deg/pix	degrees/pixel
arcsec/yr	arcseconds/year
mas.yr-1	mas/year
mas/yr	mas/year
erg/s/cmˆ2	erg/cmˆ2/s

Table 7: Some VOTables used different values for the same units, so they were converted to use a single value. The Original column gives the unit values found in the VOTable files and the Converted column gives the value used in the RDF version of the tables.

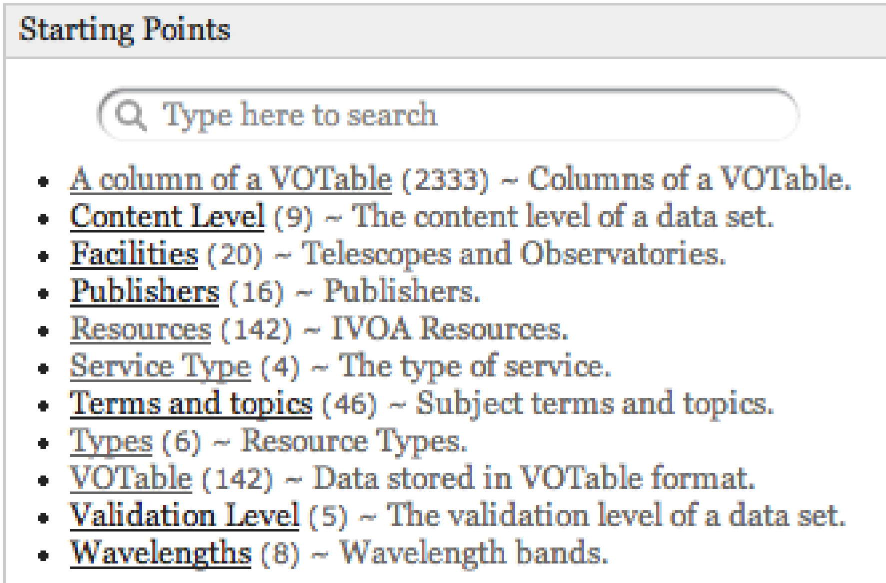


Figure 26: The starting point of the `Longwell` application when configured to browse the NVO DataScope data. The number of data classes has increased over that used for modeling the VizieR data (Figure 9), and these options do not include all classes used to model the data (e.g. it does not show the 109 entries used to represent UCD values found in the tables).

URL should be given using the `href` attribute but was stored as a `HREF` child element of the `LINK` element in one file.

H.4 Comments and notes on the conversion

Figure 26 shows the starting point of the DataScope browser. The main classes of interest are the "Resources", "VOTable", and "columns" which give, respectively, meta data about a service, general information about the tabular data, and information about an individual column in a table. Figures 27, 28, and 29 show examples of these three classes.

Since the DataScope application was not completed, and the number of predicates and classes used is larger than that for the VizieR application (Appendix F.3), a mapping between the input data and RDF output is not included here. The total size of the RDF database was 4 Mb; the total size of the 142 VOTables, in XML format, was 3 Mb and most of the data was included in the RDF version.

The RDF modeling extended that used for the VizieR application in that more objects were created as URIs rather than string literals; this was in part because the data was

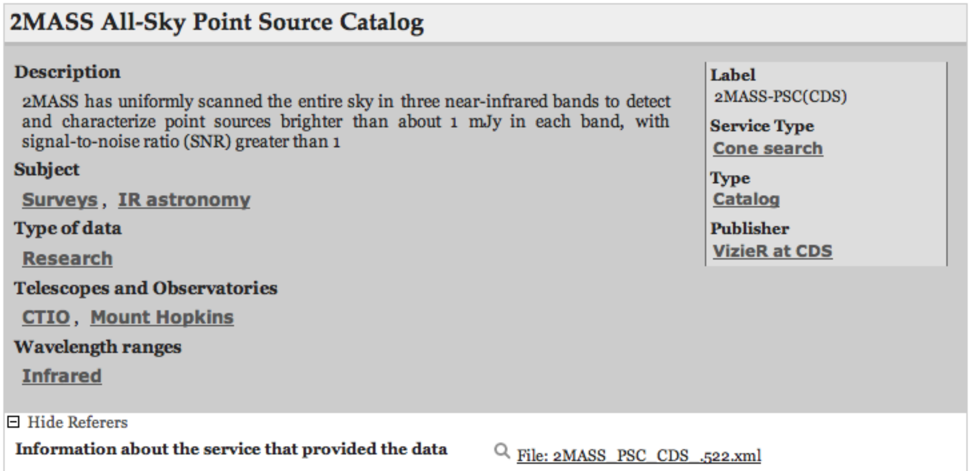


Figure 27: The `Longwell` display for the metadata for the data from the CDS version of the 2MASS point-source catalog. A similar layout and look was used to the VizieR registry application (e.g. Figure 23). The metadata section describes the resource — here a "cone search" — that provided the data; the actual tabular data for this record is shown in Figure 28.

more heterogeneous and so it made more sense to create RDF classes for objects such as the content level field of an IVOA resource (for the VizieR data set this particular field always had the same value of "Research", whereas with the DataScope sample there were nine different values). This effort provided extra information to the user since it provides links between resources, in that a user can go from an item, such as the one shown in Figure 27, to the definition of a term like "Infrared wavelength range", and from there back to all resources that link to this term, by using the "Show Referers" link provided at the bottom of each match. In this application there was minimal effort spent on adding extra information to the ancillary objects, such as the CTIO facility or Infrared wavelength range shown in Figure 27, since this was not the primary aim of the project.

As has previously been mentioned, some items included direct links to relevant information rather than having them be inferred by `Longwell` — for example instances of the `vor:DatascopeResource` and `vor:VOTable` classes contain explicit links to each other via the `vor:tableLink` and `vot:metadataLink` predicates — which simplifies the user interface, since the information can be included in the main display rather than hidden away and only accessible via time-consuming calls back to the database (e.g. the "Show Referers" link). Not all such links are provided to the user when they probably should; the most notable absence is that at present the metadata about a table (e.g. Figure 27) does not provide a direct link to the tabular data, and to get at this link the user has to expand the "Show Referers" section.

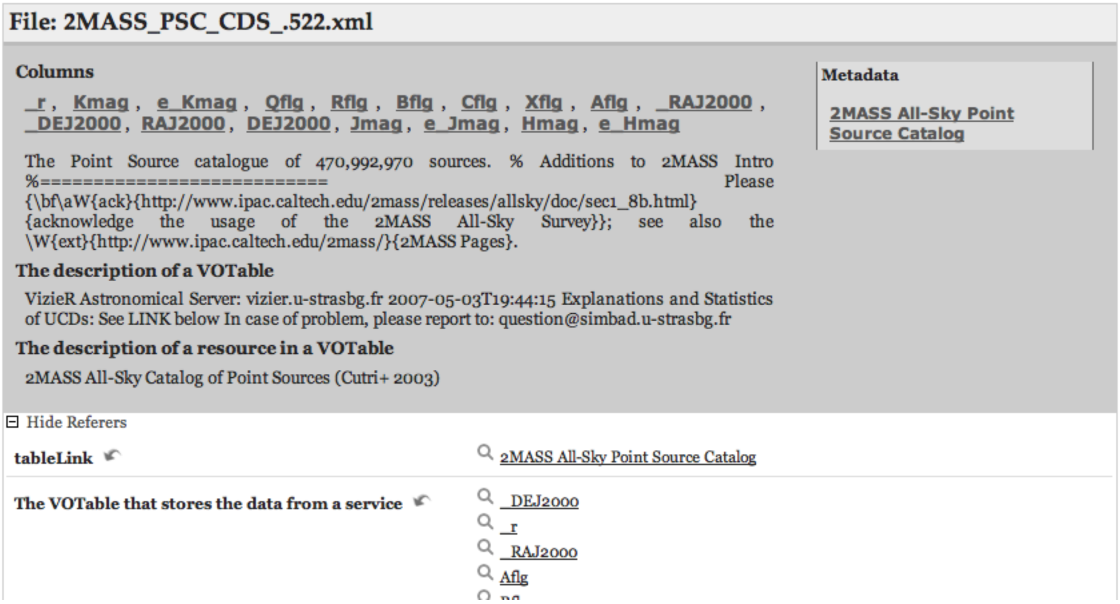


Figure 28: The display of the tabular data for the 2MASS point-source catalog entry whose metadata is shown in Figure 27. Some of the descriptive text for this table contains layout (white space) and LaTeX-like mark up that has not been removed or converted to HTML. For this particular resource, the Fresnel lens displays all the links to other data resources (i.e. table columns and metadata), so there are no new links in the "Referers" section (which is only partially shown in this Figure).

Figure 29: The `Kmag` column from the 2MASS table shown in Figure 28. The actual data stored in this column, although included in the RDF model, is not shown due to problems with the display of list-like data in `Longwell`.

Some of the effort needed to convert the data, in particular the metadata about a table, was due to the method used to obtain this data; namely the extraction of ASCII-format data from the DataScope interface. This technique was used as it was the simplest method available at the time, but overall it would have been better to query the IVOA Registries for the data directly, using a scheme similar to that used for the VizieR application. This would have also increased the information available, since the ASCII format excluded some fields, and would have allowed a more direct re-use and extension of the registry vocabulary described in Appendix F.

As discussed in Appendix H.3, a reasonable amount of time was spent dealing with converting the VOTable format; issues included namespace differences, incorrect names or values for attributes, invalid mark up, and incomplete or misleading values (such as UCD values of `"?"` or content levels set to `"test"`). Some of the issues were reported and have been fixed since the project finished, but it is likely that applications which handle VOTable documents will need to be robust when parsing their contents (c.f. the Robustness Principle [25], commonly referred to as Postel's law).

Some text elements were found to include extra mark up, which can lead to an unreadable or confusing display, as shown in Figures 28 and 29. This issue was also noted for the VizieR browser (Appendix F.2).

The use of blank nodes for representing some parts of the VOTable structure, such as parameter values, was a limited success. The main problem that occurred was in how to display the information modeled as a blank node to the user; it was not clear how well `Longwell` supported this feature of RDF since some display issues were definitely down to poor data modeling, such as missing labels, that were not able to be resolved during the project. There were similar unresolved problems in getting `Longwell` to display the column data, so although this information is present in the RDF version, it is not available to the user.

Although the data browser was not completed, its use did highlight the following issues:

- Without an understanding of the RDF data model, it was easy to get lost when searching, since it was not always clear what the different types of resource meant. Several partial solutions were identified but not implemented: reduce the number of classes by merging some resources, primarily the metadata about a table and the table data itself; change the display (such as the style or layout) of different classes to increase the visual distinction; and reduce the number of classes that can be directly searched (e.g. Figure 26).

- Large Astronomical data sets may be made available via multiple sites to reduce access times. This can result in the same data being returned multiple times to the user, causing potential confusion. As an example, the data set referenced in Figure 27, the "2MASS All-Sky Point Source Catalog", is provided by the VizieR service and is a copy of the data from the NASA/IPAC Infrared Science Archive

([12]), but the IRSA version of this catalog does not contain the "Telescopes and Observatories" information shown in this Figure. When the tables returned by the services are compared, the set of columns does not match, nor do the number of returned items (the latter appears to be due to a difference in the maximum radius used in the cone search).

- The inclusion of extra information in the search interface, whether as extra classes that could be searched or new facets to help restrict the search, did not appear as useful as it did for the VizieR browser. The simple approach of "make everything a facet" does not produce a powerful interface, and additional search elements — such as range filters like "K $>$ 20" or positional filters — are required.