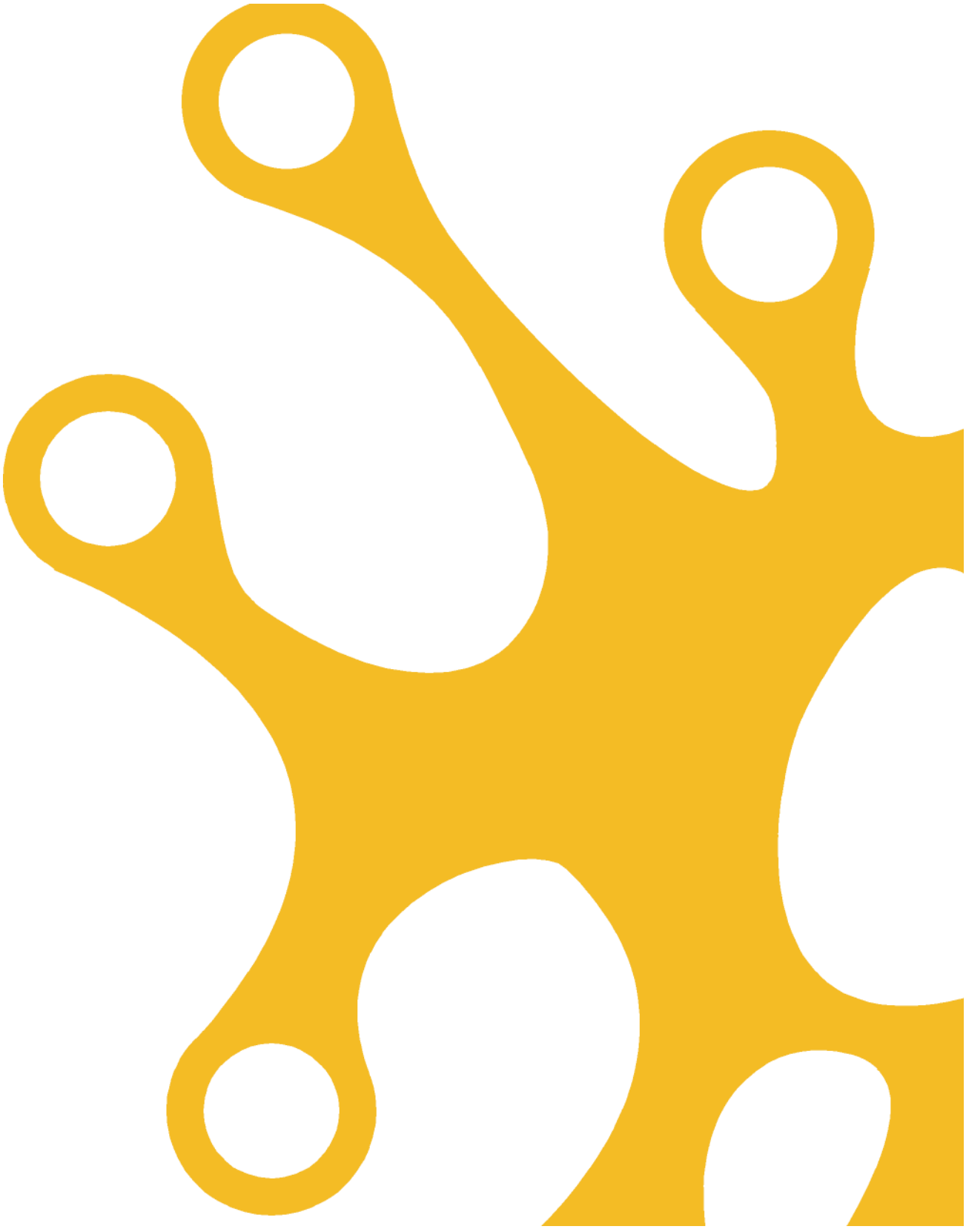
Gemini UDS

Advanced Internet Searching

Version 1.3 – User's Guide



Gemini Unified Datamining System User's Guide for Windows and Mac OS X.

Snowtide Informatics Systems, Inc.
P.O. Box 346
South Hadley, MA 01075

http://www.snowtide.com

Produced and printed in the United States.

Table of Contents

Introduction

Searching for information is hard; don't let anyone try to tell you anything different. Fundamentally, searching for information is hard because of at least two facts:

> (a) *The amount of available information in the world is growing so quickly (exponentially, in fact) that no known technology can keep up with that growth in its entirety,*
>
> and
>
> (b) *The number of available information resources (like search engines, subscription services, databases, etc) has also grown over the past 5-10 years at such a clip that humans can't keep up.*

If you regularly search for anything more complex than sports statistics (not that there's anything wrong with that!), you know these things to be true. You can spend hours wading through page after page after page looking for the data that you need. You use so many different information resources on a regular basis that it's hard to keep track of their individual strengths, weaknesses, and idiosyncrasies.

The Gemini Unified Datamining System (Gemini UDS) was developed to give you the means necessary to cope with and overcome these facts of our information society. This guide will bring you through the steps of setting up and using Gemini UDS in order to massively boost your productivity while searching for information.

Finally, a quick note about the screenshots and images in this manual: we have put together a mix of screenshots from both the Windows and Mac OS X versions of Gemini UDS so that you will have some point of reference regardless of the platform that you use.

With that, we will leave you to explore this guide and Gemini UDS . . .

The Snowtide Team

Chapter 1 Getting Started

This chapter will outline what Gemini UDS needs to run properly, how to get it installed and running, and will go through the steps you'll need to take to run your first search using Gemini UDS. Later chapters will examine this process in great detail.

System Requirements

In order to use Gemini UDS effectively, we recommend that you have one of the following system configurations:

Mac OS X with:

- A G3 or G4 based Macintosh running at 233Mhz or faster.
- Mac OS X version 10.1.3 or later[1] (with the Java Update 1 patch[2])
- 256MB of RAM or more
- At least 40MB of free hard drive space.

Windows 98SE, ME, 2000 (SP2 or later), NT 4 (SP 5 or later), or XP with:

- Version 1.3 or later of the Java Runtime Environment
- At least a Pentium II system running at 233Mhz or faster.
- 256MB of RAM or more
- At least 40MB of free hard drive space.

Windows users may download the Java Runtime Environment (JRE) from Sun Microsystem's website through this page: http://java.sun.com/j2se/1.4/download.html. The default Gemini UDS distribution comes with a JRE built-in, so doing this is typically unnecessary. Gemini UDS will not run without a suitable JRE installed and properly configured (Mac OS X comes with the JRE already installed).

Gemini UDS may be used on any Java-capable with v1.3 or later of the JRE, but Snowtide does not officially support Gemini UDS on platforms other than Mac OS X and Windows.

It is strongly recommended that you use Gemini UDS on a system that is connected to the Internet or your network by a high-speed connection: DSL, cable modem, LAN, and faster ISDN connections are all ideal. Of course, using Gemini UDS on even faster connections can be beneficial. Gemini UDS can certainly be used over a modem connection, but it will be limited by that communications bottleneck.

There is no special installation needed to start using Gemini UDS. Just download Gemini UDS from Snowtide's website, or copy it from a distribution CD to your local hard drive. After opening the

[1] Gemini UDS will not run properly on versions of Mac OS X before 10.1.3, and is completely unsupported on those versions. Updating to version 10.1.3 is easy; either run Software Update through your System Preferences, or download the combination 10.1.3 update from http://docs.info.apple.com/article.html?artnum=120103 and install it manually.

[2] The Java Update 1 patch for Mac OS X can be obtained via the Software Update panel in System Preferences, or it can be downloaded from http://docs.info.apple.com/article.html?artnum=122000 and installed manually.

Gemini UDS directory, you should see a set of files and directories. You can start Gemini UDS by double-clicking on the Gemini UDS executable (or by running the startGemini.sh script on Unix-based systems).

Initial Setup

The first time you start Gemini UDS, it will do an initial check of your system. Once it has finished that, it will display the End User License Agreement, which you must agree to before using Gemini UDS.

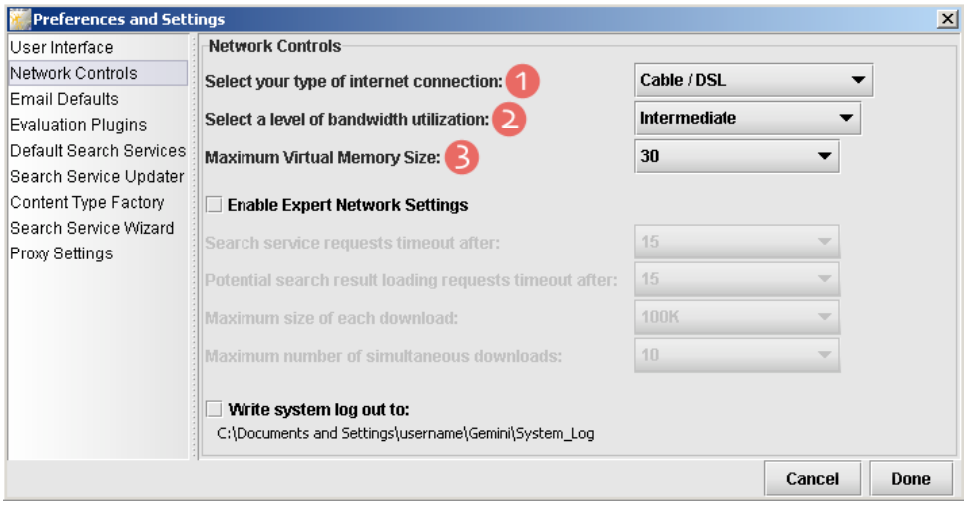
There are a couple of things that you should probably do before using Gemini UDS for the first time. First, you should give Gemini UDS an idea of what kind of Internet connection you are using; this will ensure that it tailors its operation to your type of Internet connection.

To do this, open the Preferences window (Windows users should look under the Edit menu; Mac OS X users should look under the Gemini UDS menu to the left of the Apple menu). When the Preferences window opens, select "Network Settings" from the list on the left. This will show you Gemini UDS' network settings screen.

Popup Help

When you first start up Gemini UDS, you'll see that there are a number of blue icons scattered around the interface that look like question marks (like the one above). Those icons are buttons that, when clicked, will pop up a window containing tips and help about the part of the interface the icon is in or near. Using these popups is a great way to learn how to use Gemini UDS enough to get started without having to read through this entire user's guide.

To turn the popup help on or off, select "Toggle Help Popups" in the Help menu.



These settings are very important to Gemini UDS' proper and efficient functioning, so you should be sure to set them appropriately.

1. *Select your type of internet connection.* If you do not see your type of internet connection, choose the one that is closest to it. If you are not sure what to select, ask your system administrator or contact your ISP for more information.

2. *Select a bandwidth utilization level.* This setting tells Gemini UDS how much of your internet connection to use. For example, if you select the "Intermediate" utilization level, Gemini UDS will limit its activity so that it uses around 2/3 of your connection's available bandwidth. "Maximum" corresponds to Gemini UDS using as much bandwidth as possible, while "Minimum" corresponds to Gemini UDS using around 1/3 of your connection's available bandwidth. ("Intermediate" is usually fine to get started with.)
3. *Select a virtual memory size.* Gemini UDS uses its own virtual memory scheme that allows it to process very large amounts of information even if your system does not normally have the resources to do so. The default selection of 30MB is generally fine for most purposes; if you plan on using Gemini UDS with a very large collection of sizable documents, increasing this setting is recommended.

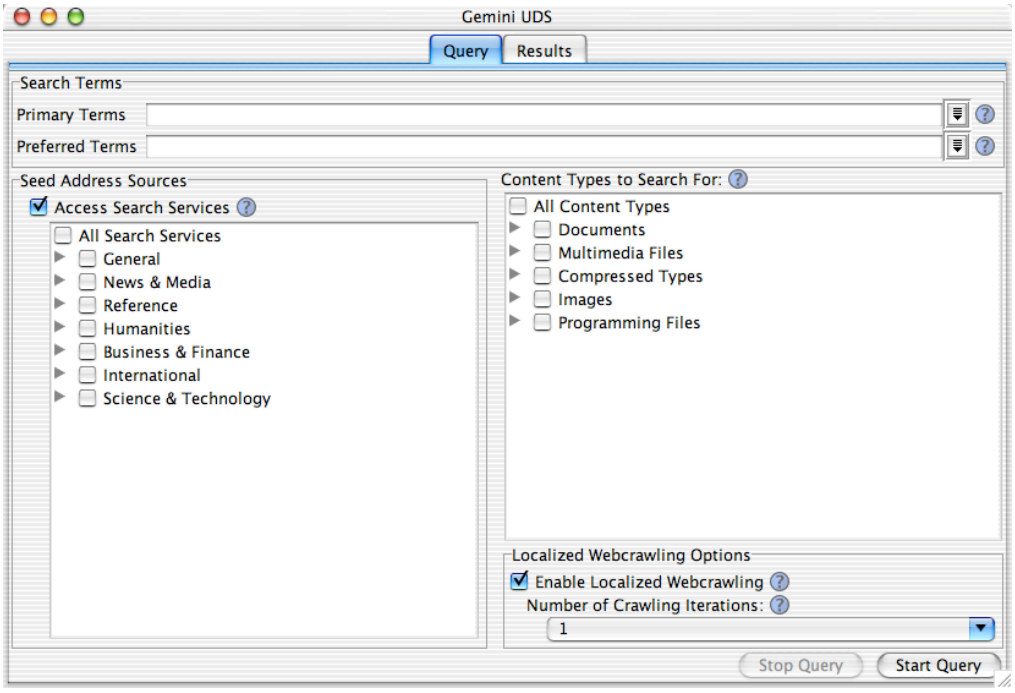
Gemini UDS also allows you to control more specific, "expert" network settings, which are not covered here. See Chapter 3 for more information about expert network settings.

The other setting that needs to be taken care of before using Gemini UDS is its proxy configuration. If you do not use a proxy to connect to the internet, you can skip this step. (Ask your system administrator if you are unsure about this.) See page 32 for information on how to set up Gemini UDS to use your school's or company's proxy server.

That is all the setup needed to start using Gemini UDS. Many more configuration options are available to you in Gemini UDS; they are covered extensively in Chapter 3. Now let's do a search to introduce you to the basic elements of Gemini UDS' interface.

Your First Search

When you first start Gemini UDS, you should see the standard query interface:

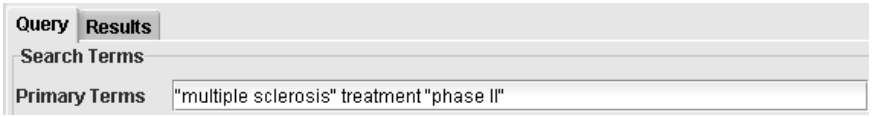


There are four parts to the standard Gemini UDS query interface: search terms, search services, content types, and localized webcrawling. All of these areas have many advanced features that you'll want to take advantage of in order to get the most out of Gemini UDS; please see Chapter 2 for details about those features.

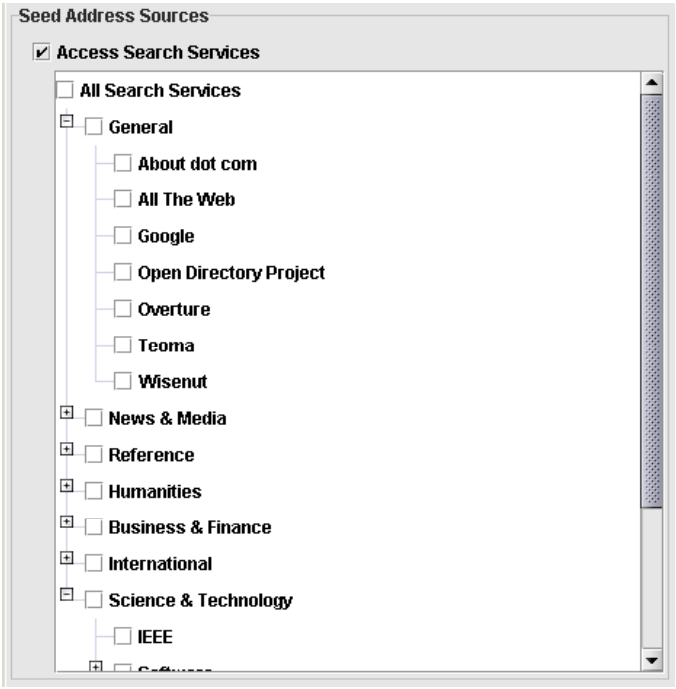
The area that you'll be most familiar with is the search terms area, which, for now, you can treat just like the keyword boxes in most search engines. For now, let's just use the Primary Terms field, which is most like search engines' keyword boxes. In the Primary Terms field, simply enter a couple of keywords describing what you're looking for:

What is a *search service*?

A search service is any kind of information source that you can access over the World Wide Web (or over your private network using Web protocols). This includes search engines (like Google, Yahoo, and Excite), specialty sites (like the New York Times or Pubmed), subscription services, or a database that you have available to you that has a web "front-end" (see *Integrating Information Resources* on page 33 for information on how to integrate your preferred search services into Gemini UDS).



Once you've entered your keywords, you need to select which search services to access. All of the search services Gemini UDS can find on your system are collected in the hierarchical “tree” display at the left of the query interface:



Here you can select either a complete category of search services (News & Media), a subcategory of them (Science & Technology/Medicine), or you can select individual search services that you think will provide useful seed addresses.

Once you have entered some keywords into the Primary Terms field and selected one or more search services, you can start your query by clicking the 'Start Query' button in the lower right corner of the Gemini UDS window. The other sections of the query interface, content types and the Localized Webcrawling controls, are not strictly necessary to perform a query; they simply allow you to control more advanced features that you'll learn about later.

Once you start your query, the Gemini UDS window will switch to the Results pane, where you can monitor Gemini UDS' activities and view the results that it has gathered for you. Near the bottom of that pane, you'll see a small status bar that indicates what Gemini UDS is doing, and how far along it is in its current task (for right now, ignore the buttons to the right):

Now that your query is underway, let's step back for a moment and take in the "big picture" of what Gemini UDS has done so far:

1. It formatted what you entered in the Primary Terms field and sent those formatted terms to each search service you selected.

2. The seed addresses returned by those search services were then loaded. At this point, Gemini UDS weeded out and discarded those results that could not be accessed.

3. The content of the addresses that did load properly were and are being analyzed to determine if they are relevant to your query. It is during this analysis that the addresses are assigned a rating reflecting how relevant Gemini UDS believes them to be to your query.

4. The addresses that were determined to be relevant to your query are streamed into the results display area for you to view and work with.

What are *seed addresses*?

Seed addresses are the raw results returned by the search services that you have selected. As you know, those results are often filled with "bad links", or they point to pages that have moved or otherwise changed enough so that they're no longer relevant to your query. In Gemini UDS, those seed addresses are just starting points for the Localized Webcrawling process. As part of that process, Gemini UDS confirms that those seed addresses are accessible, and that they're relevant to your query, ensuring that you don't have to wade through outdated, dead, or irrelevant results.

As Gemini UDS finds results that meet your query's parameters, they will be appended to the list of results in the table above the status bar. Once you see some results listed, you can click on the blue underlined link of each result, which will open that search result in your browser. You may also perform sub-searches on those results and filter them; please see page 22 for more about results filtering and subsearches.

After Gemini UDS processes the seed addresses it extracts from the search services you selected, its job isn't done yet. Then, its Localized Webcrawling process automatically scans through the most relevant results, and gathers the links that are held by the content of those results. Those links are then used as another set of seed addresses that are loaded, analyzed, and then made available to you as more search results. This process essentially eliminates the work that you would otherwise have to do in following links on relevant pages. So, if you didn't change any of the Localized Webcrawling settings from their defaults when you started your query, Gemini UDS will crawl down four levels of links automatically, and display any newly-discovered pages along with the results extracted from the search services you selected. (Please see Appendix A for more information about Localized Webcrawling.)

Unix and Browsers

Most of the time, you will want the results found by Gemini UDS to be opened in your browser for further examination and viewing. This interfacing between Gemini UDS and the browser is largely automatic on the Windows and Mac OS X platforms.

However, most Unix operating systems are not so automatic in their browser management, so you will find that some configuration may be necessary in order to get Gemini UDS to interface with your browser properly.

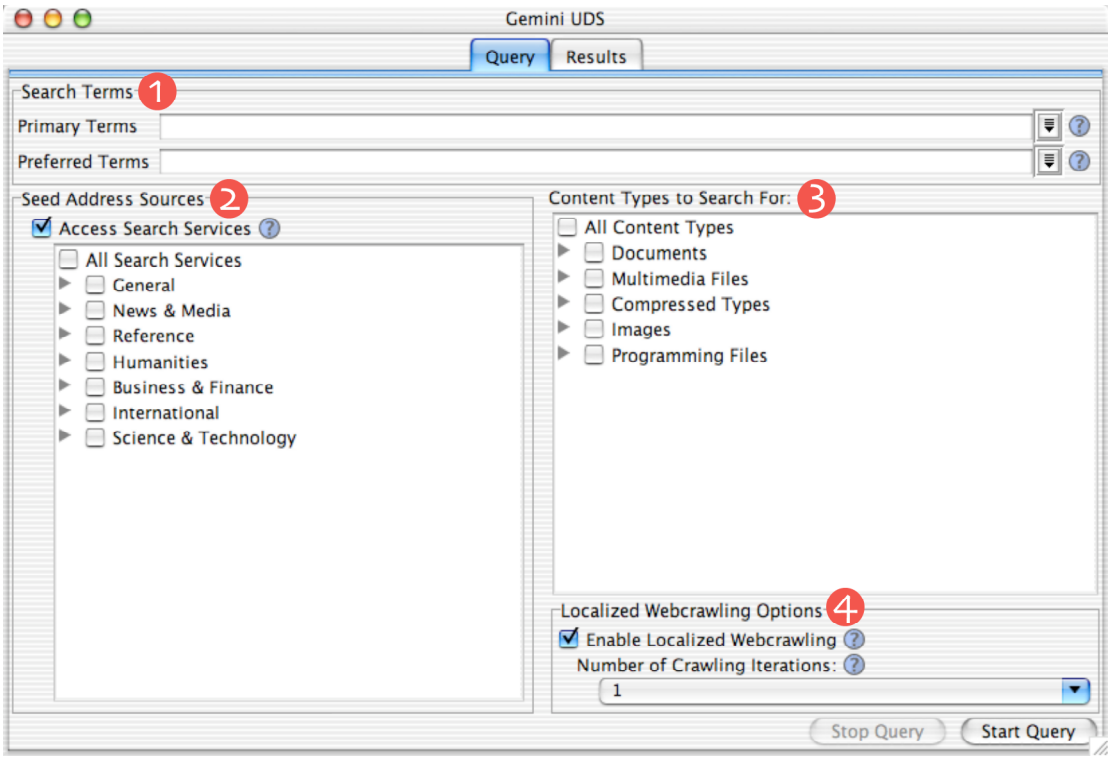
See page 28 for instructions on how to configure Gemini UDS to use your preferred browser in Unix.

We hope this introduction has given you a small taste of what Gemini UDS does, and hints at the wonderful things that Gemini UDS can do for you as you fully utilize its features and power. The rest of this guide is dedicated to a point-by-point explanation of each of those features; reading through all of it will give you the know-how to utilize Gemini UDS to its fullest potential.

This chapter will focus on the mechanics of forming search queries within Gemini UDS, meaning that we will cover how to go about entering your query and using the tools provided to simplify that process. This chapter does not go into how to form *good* queries – that is a task that can vary widely from topic to topic depending on what you are searching for.

Forming your query properly is the most important part of Gemini UDS, as your query will determine everything that Gemini UDS does for you. Knowing how to use Gemini UDS' query interfaces properly will greatly improve your productivity when using Gemini UDS.

The first time you start Gemini UDS, you will see the standard query interface, which should look something like this on your screen (major areas of the interface are numbered):



There are four sections to the standard query interface, with each section affecting a different aspect of Gemini UDS' search process:

1. **Search Terms:** This area is where you input your keywords and other expressions that describe what you're looking for.
2. **Seed Addresses:** Here you control which search services should be accessed in order to get an initial set of results.
3. **Content Types:** Here you can specify which types of special content should be searched for.
4. **Localized Webcrawling Options:** Here you can control certain aspects of the Localized Webcrawling process.

Each of these sections contains a broader set of functionality in the advanced query interface; we'll look at the differences between the standard and advanced interfaces a little later. Let's go through each section and detail each of its purposes and capabilities.

Search Terms

There are two areas for search terms, Primary and Preferred. What you enter in these fields will inform Gemini UDS as to what you are looking for, and your entries will be used to determine relevancy of each potential search result (i.e. whether it is shown to you, what its relevancy ranking should be, etc).

(Note: The search terms fields may be labeled differently if you have selected a different Evaluation plugin to use; see page 34 for more information about Evaluation plugins. *All of the statements about relevancy in this section assume you are using Gemini UDS' default plugin; other plugins may determine relevancy in completely different ways.*)

In short, these fields accept regular keywords (words like 'dog', 'cancer', or 'trains'), as well as phrases surrounded by double-quotes (such as "dog food" and "new york") and boolean expressions (like 'dog NOT cat' and 'grass OR hay'). Both fields accept the same syntax; see Appendix B for a full explanation of the syntax and advanced features it provides, such as support for regular expressions and weighted search terms.

There are some very important yet subtle differences between terms entered in the Primary field versus those entered in the Preferred field:

1. Your input to the Primary Terms field will be formatted and sent to your selected search services; what you input into the Preferred Terms field is not sent to the services.
2. A potential search result must satisfy all of the terms in the Primary field for that potential result to be considered relevant.
3. A potential search result must satisfy all of the terms in the Preferred field for that potential result to be considered to contain Preferred Terms.

Some explanation of these differences is warranted. The first point is relatively simple: only Primary Terms are used to query the search services you select, while Preferred Terms are used exclusively for analysis within Gemini UDS. The importance of this distinction will become apparent as we discuss the other differences between the Primary and Preferred Terms fields.

When Gemini UDS analyzes a potential search result's content, it looks for each of your Primary search terms in that content. If the content contains all of those terms (i.e. it satisfies all of the entries in the Primary field), then that potential result will be confirmed as relevant and will therefore be shown to you in the results display area. Otherwise, it is discarded and not shown as a result.

The third point is important in three ways. First, search results whose content contains Preferred Terms will be rated *much* higher than results that do not, but their inclusion in the final set of relevant results are not affected by whether or not they contain those terms. In short, results are not *required* to contain Preferred Terms, but those that do will be rated much higher than others.

Another way in which the second point above is important is that the Localized Webcrawling system in Gemini UDS will follow links in content that contains Preferred Terms before it follows links from content that doesn't contain those terms. Therefore, if two relevant results are "available" to be crawled by the Localized Webcrawling system, and only one of them has content that contains your Preferred Terms, then that result will be crawled first. (See Appendix A for more information about Localized Webcrawling and how it works.)

Third, whether or not a search result contains Preferred Terms can be very important when filtering results. Details on how to tailor your query to take advantage of this can be found on page 22.

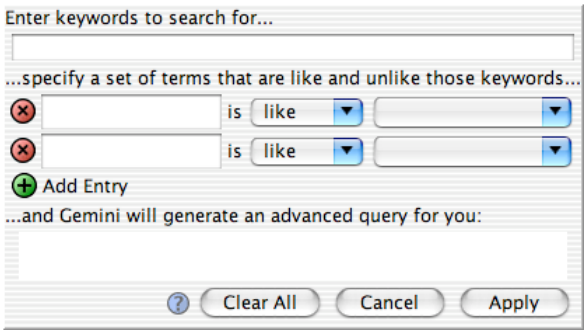
Finally, it is important to remember that only if a search result's content *satisfies* the terms in the Preferred Terms field will that result be "marked" as containing Preferred Terms. This means that, for example, when searching for information about different types of bicycles, you should not simply enter a string of keywords into the Preferred Terms field like "mountain road recumbent unicycle". Doing so will result in only results containing each of those words being marked as containing Preferred Terms (which probably wouldn't be your intent). In such a case, using a boolean expression to indicate that you are looking for results containing *any* of those terms would be more appropriate (such as "mountain OR road OR recumbent OR unicycle"), causing results containing any of those terms to be marked as containing Preferred Terms.

But what if you don't know how to use boolean expressions effectively (or you don't know what boolean expressions are at all)? The next section talks about Gemini UDS' Query Assistant, a tool that can help you to use boolean expressions, regardless of your skill level.

The Query Assistant

As you know, Gemini UDS' search terms fields support boolean operands (AND, OR, and NOT). Research has shown that search queries that utilize boolean operands effectively are much more successful; unfortunately, few people know how to use boolean operands properly, even though they could benefit from them tremendously.

For this reason, Gemini UDS includes the Query Assistant, an interface that helps you to build effective, well-structured boolean queries. (It can also be a timesaver for some advanced users that need to build very complex boolean queries.) Both of the search terms fields (Primary and Preferred) have Query Assistants associated with them; they can be opened by clicking on the button that is next to each search term field (an example of which is shown to the right). When a Query Assistant first opens, it will look something like this:



The Query Assistant helps you to build boolean queries by allowing you to use simple words ("like" and "unlike") to represent the relationship between search terms. To start using the Query Assistant, just enter a set of search terms in the top text field and hit return (those terms can be anything the search terms fields can accept; see Appendix B for a full explanation of the syntax).

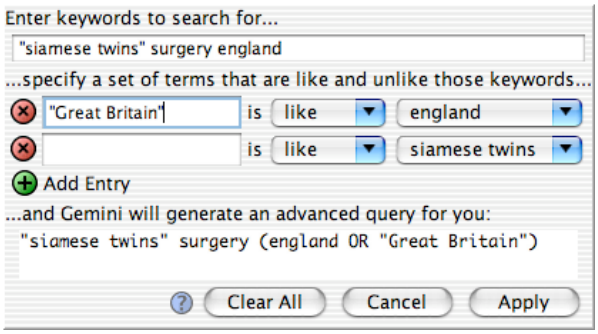
Building a boolean query from that start consists of three steps that lead to you essentially filling in blanks to create a complete, true sentence:

1. Enter a word, phrase, or other search term in one of the smaller text fields on the left of the Assistant's window (these fields also accept the full range of syntax described in Appendix B).

2. Then select the search term that is to be related to what you typed in Step One from the rightmost pull-down menu that is on the same row as the text field you typed in in Step One.

3. Finally, select the appropriate word from the middle pull-down menu that makes the sentence on that row true.

This example showing the result after going through this process should make things clear:



The initial query "'siamese twins' surgery england" was entered in the top text field. Since "Great Britain" and "england" are essentially interchangeable (ignoring for the moment Wales, Scotland, etc.), we would like to have them treated as such. So "Great Britain" was entered in the first text field on the left, "england" was selected in the rightmost pull-down menu (since it is the term we wish to relate to "Great Britain), and "like" was selected from the middle menu since the two terms' meanings are nearly identical. (Alternatively, we could have chosen "unlike" from the middle menu to indicate that "Great Britian" and "England" were not very similar; this would result in a very different boolean query.)

After doing these things, Gemini UDS generates a boolean query in the bottommost text area; clicking the "Apply" button will copy that boolean query to the main query interface. If this query were used, it would find many more relevant results than the original entry ("'siamese twins' surgery england") because many potential search results probably talked about "Great Britain" instead of "england".

As you are building boolean queries using the Query Assistant, you can use the red "X" icons next to each row (called an "Entry") to delete that row, and you can use the "Add Entry" button (the "+" icon) to add a row. This enables you to build boolean queries with many sets of related terms. Clicking the

"Clear All" button will remove all of the Entries; clicking the "Cancel" button will dismiss the Query Assistant without affecting anything in the main query interface.

Search Services

The second area in the Gemini UDS standard query interface as shown on page 11 is where you can select which search services to submit your query to to start the search process. (See the sidebar on page 8 for a definition of what a search service is.) Gemini UDS ships with a large set of search services that cater to a range of popular interests, as well as a set of general purpose search services that can be used to start searches on nearly any topic.

The search services are displayed in this area in a hierarchical way, with individual search services organized within categories (or subcategories in some cases). You can therefore select individual search services, or entire categories, which will cause Gemini UDS to access all of the search services contained in that category.

When you start your search, Gemini UDS formats your Primary search terms appropriately for each search service, and then queries the search services with the properly-formatted set of search terms (see the *Important Note* to the right for more about this formatting). The search services will return a set of results, just as if you were accessing them manually through your web browser; those raw results are called *seed addresses*, which will be loaded and whose content will be analyzed by Gemini UDS to determine which of them are relevant to your query.

(There are other ways to provide seed addresses to Gemini UDS, as discussed later in this chapter in the section about the advanced query interface. See Appendix A for more about seed addresses and how they fit into the Localized Webcrawling process that drives Gemini UDS).

As a final note, remember that you can use Gemini UDS to access and process results from essentially any search service that uses standard web protocols. See *Integrating Information Resources* on page 33 for details.

Important Note

At this point, you know that Gemini UDS sends your search terms to each search service that you've selected. It should be noted that not everything that you type into the search terms fields is guaranteed to be sent to or "understood" by the search services you selected.

The reason for this is that there are no standards for what kinds of queries search services must accept and process properly. A simple case of this is quoting: most search services recognize search terms surrounded by double quotes to be phrases (terms that must be found in search results adjacent to each other), but some do not.

Therefore, when you enter search terms containing phrases, Gemini UDS will not send that phrase to any search service(s) you've selected that cannot process phrases – the quotes will be stripped, making the parts of the phrase(s) simple keywords in the search service(s) eyes. For example, '"dog food"' will be sent as just 'dog food' to search services that don't support phrases.

See Appendix B for a full discussion of Gemini UDS' query syntax.

Content Types

The third area in the Gemini UDS standard query interface as shown on page 11 allows you to select which content types Gemini UDS should search for. The content types are arranged much like the

search services are, with similar or related content types grouped into categories; you can select either individual content types or entire categories.

Selecting a content type will not require that content type to be present in the content of the search results that Gemini UDS returns to you. Rather, it causes Gemini UDS to check each search results' content for the content type(s) you selected. (Gemini UDS considers any links or embedded references to your selected content types to be sufficient to "mark" a result as containing those content types.)

Any content types that are found are then noted in the search results listing. You can then use the sorting and filtering tools provided by Gemini UDS to quickly get a clear picture of which results contain the content you're looking for (see the sections starting on page 22 for details on sorting and filtering results).

The content type selection area is the only portion of the query interface that is the same in the standard and advanced interfaces.

Localized Webcrawling Options

The fourth area in the Gemini UDS standard query interface as shown on page 11 allows you to control two aspects of the Localized Webcrawling process: whether it is to be used for the next search, and how many iterations it should be allowed to process.

If you disable Localized Webcrawling, then the Gemini UDS search process becomes much simpler: it will submit your Primary search terms to the search services you've selected, retrieve the raw results, analyze those raw results' content, and display the analyzed and rated results in the results display area. Doing this might be helpful if you are doing "exploratory" searching (if you aren't sure of the proper terms to use to find what you need, etc), or if you are accessing custom search services that would not benefit from the application of the Localized Webcrawling process (a rare condition, but possible).

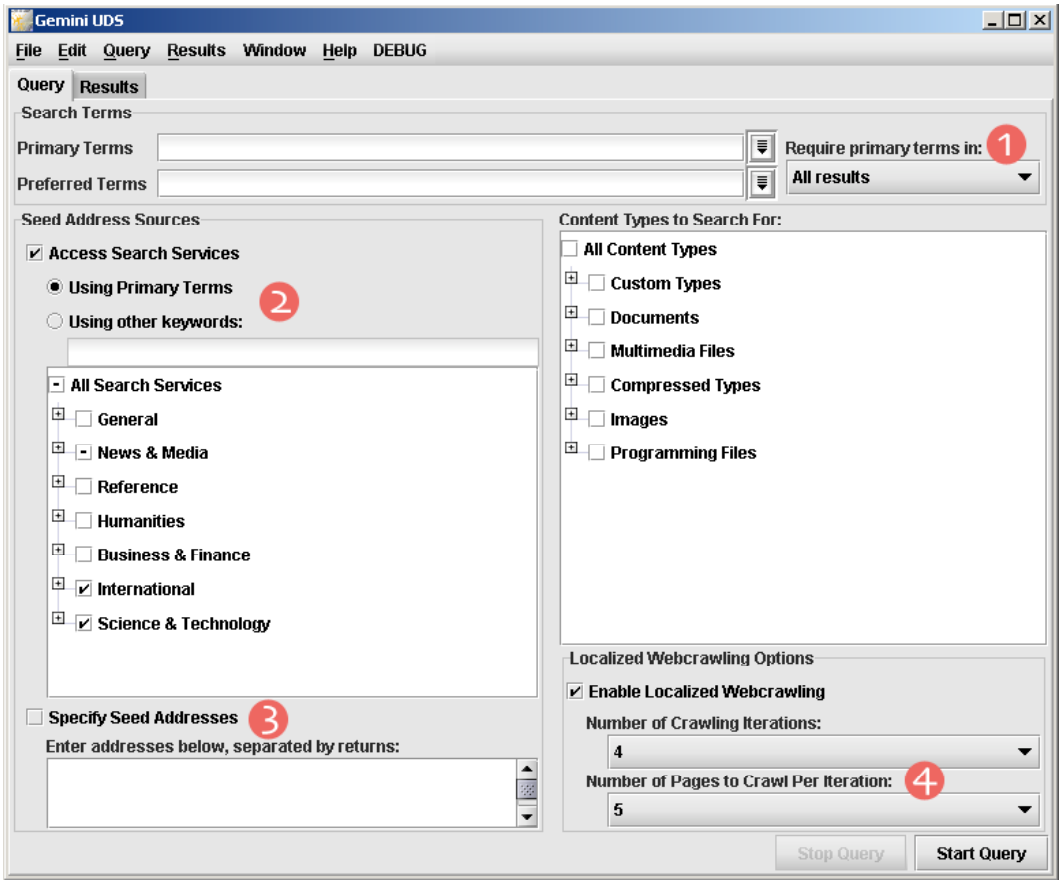
You can also control how many iterations the Localized Webcrawling process will be allowed to perform. An iteration is essentially one cycle through the process of loading a set of potential search results, analyzing them to determine their relevancy, and following links in the most relevant results to find additional results not returned by the search services you selected.

Until you feel comfortable with how the Localized Webcrawling process works, there is little need to change the default settings for this area (Localized Webcrawling enabled, iterations set to 4); those settings are a good starting point for using Gemini UDS effectively.

It is recommended that if you wish to use these options effectively and Gemini UDS to its fullest potential, that you review Appendix A. That section details the Localized Webcrawling process that drives the Gemini UDS search process; knowing the main points of how it works will enable you to craft your queries to best take advantage of its capabilities.

While the standard query interface in Gemini UDS is very useful, it hides some very powerful parts of Gemini UDS in order to simplify things for beginning users or those who do not need to control those more robust features. For everyone else, the advanced query interface is probably the best choice; especially if you understand in broad strokes how the Localized Webcrawling process works, it gives you a level of control that is inaccessible in the standard interface.

You can switch to the advanced query interface by selecting its entry in the Query menu. It should look something like this:



As you can see, there are four regions that differ between the standard query interface and its advanced cousin. Only the marked regions are different in the advanced query interface; other elements of the interface are identical to the standard type as described in previous sections, so we shall only discuss the differences here.

The first region of difference is in the Search Terms area. This pull-down menu allows you to control how Gemini UDS treats Primary search terms. By default, every result that is provided to you in the results display area is required to satisfy what you enter in the Primary Terms field. The pull-down menu in the first region indicated above allows you to change that so that Primary Terms are only required in the content of results found directly through search services, and no others.

The consequence of this is that while results coming directly from search services will be required to satisfy those Primary Terms, any results that are found via the Localized Webcrawling process after analyzing those initial results will not be required to satisfy your input in the Primary Terms field. This will cause the number of total results found to be much greater, but it also gives the Localized Webcrawling process the opportunity to follow links on pages that do not contain Primary Terms (and therefore would never have been considered relevant).

The second and third regions of difference are in the Seed Addresses area. The third region is simpler, so we'll discuss it first. In this region, you may manually provide seed addresses to Gemini UDS. To do this, simply check of the "Specify Seed Addresses" checkbox, and enter one or more URLs in the text area below that checkbox. There should be one URL per line (each one separated by a single return).

These seed addresses are treated just like seed addresses extracted from search services, with the exception that links in their content are *always* crawled in the Localized Webcrawling process. This is excellent if you would like to search a single site or a set of pages on a site (long FAQ pages that contain hundreds of links are ideal): Gemini UDS will follow all of the links held by the content of the seed addresses you provided, and subject those new found pages to the same loading and evaluation process as every other potential search result.

The second region in the advanced query interface affects how Gemini UDS interacts with the search services you select. By default, Gemini UDS uses your input into the Primary Terms field as the basis of its submissions to the search services you select. However, there are instances where this may not be appropriate because Primary Terms are also used as the main way of determining the relevancy of potential search results. In those instances, you can opt to provide another set of search terms that will be used exclusively for querying the search services you've selected. Those "other keywords" as they are called in the interface will not be used at all in the evaluation and analysis of potential search results; they are only meant to provide an extra level of control in how Gemini UDS accesses your search services. See the sidebar to the right for an example of how this feature would be very useful.

An example of using "Other Keywords"

A simple example of when the "Other Keywords" feature would be useful is when searching for specific chemical compound names. Many search services do not index or catalog such names because of their obscurity (such as *3,3,4,4, -Tetrachlorotetrahydrothiophene-1,1-Dioxide*), so putting such a term in the Primary Terms field would not be helpful – the search services you select that do not catalog that term would return no results or irrelevant results, giving Gemini UDS little to work with. However, if you know that that chemical compound is used in wastewater treatment (don't take our word for that), you can put "wastewater treatment chemicals" in the "other keywords" field, send those terms to the search services, and leave *3,3,4,4,-Tetrachlorotetrahydrothiophene-1,1-Dioxide* in the Primary Terms field. That will bring back a ton of results from the selected search services, but only those that contain that chemical compound's name will get through Gemini UDS' analysis and be shown to you in the results display area.

The fourth and final difference between the standard and advanced query interfaces is the addition of a second Localized Webcrawling option, how many pages to crawl per iteration in the Localized Webcrawling process.

As we discussed earlier, an iteration in that process consists of loading a set of potential search results, analyzing them, and then crawling the most relevant of them to find links to more results not provided by the selected search services. A potential issue in this process is that it might find so many links to

new results that your computer system and network would be overloaded; many webpages can contain 20, 40, even 100 links, which can add up quickly and place a strain on your available resources.

To deal with this, the Localized Webcrawling process crawls only a certain number of pages in each iteration, 5 by default. That means that in each iteration, the 5 most relevant search results will be crawled. Lowering this number will pace the Localized Webcrawling process somewhat, forcing it to do less in each iteration; raising the setting does the opposite. What you set this to really depends on what kind of content you expect Gemini UDS to have to deal with. Web-based content in general is best handled with the default setting. Searching specialized sets of data that contain few links to other pages would benefit from having a higher setting here, while other data that is highly cross-referenced should have this setting lowered.

Once search results start streaming into Gemini's results display area, you can start working with them. Much of the time, this will involve what you might do with search results delivered by any other tool: viewing the results that are provided by opening them in your browser. However, Gemini UDS has a number of tools you can use to manage those results more intelligently that are unavailable through any other tool. In this chapter, we will cover each of those tools and how to use them.

Controlling the Results Display

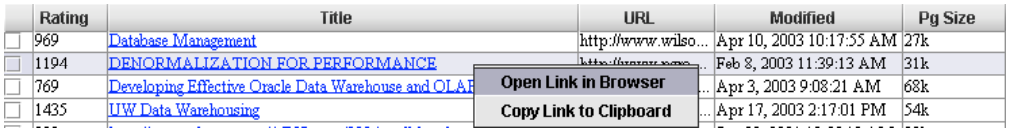
Let's look at the most common things you'll be doing with results first: opening them in your browser, moving around within them, and changing how they are presented to you.

So let's assume you've run a search, and you are presented with a set of results; a couple of them might look like this:

	Rating	Title	URL	Modified	Pg Size
☐	969	Database Management	http://www.wilso...	Apr 10, 2003 10:17:55 AM	27k
☐	1194	DENORMALIZATION FOR PERFORMANCE	http://www.pgro....	Feb 8, 2003 11:39:13 AM	31k
☐	769	Developing Effective Oracle Data Warehouse and OLAP Applications	http://www.dba-o...	Apr 3, 2003 9:08:21 AM	68k
☐	1435	UW Data Warehousing	http://www.washi...	Apr 17, 2003 2:17:01 PM	54k

Opening a single search result's URL in your browser is simple: just click on the hyperlink (the blue underlined text), and Gemini UDS will open its corresponding URL in your browser (on Windows and Mac OS X, Gemini UDS will pass the URL to your operating system, which will prompt your preferred browser to open that URL. Gemini UDS must pass the URL directly to a browser on Unix-based systems; see page 28 for instructions on how to configure Gemini UDS to do this properly in Unix.)

Alternatively, you can bring up a contextual menu for that result. To do this, right-click on a row (you can also use Ctrl-click on Mac OS X), and the contextual menu for that result will be shown:

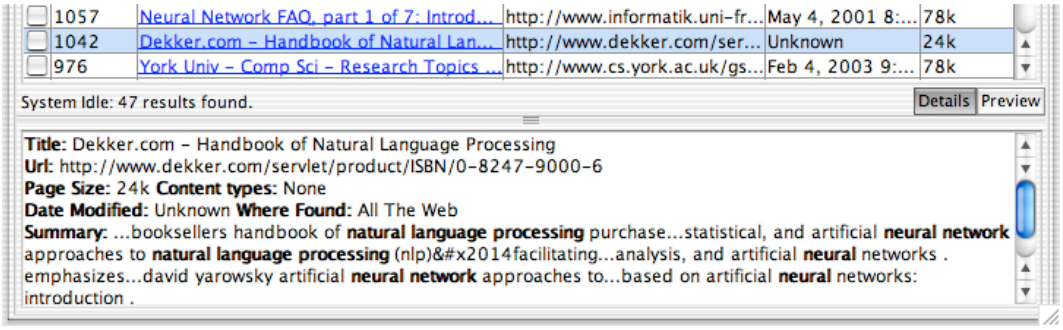


That menu allows you to either open that result in your browser, or to copy that result's URL to your clipboard (which is equivalent to selecting text with your mouse and issuing a "copy" command).

In many cases you will want to open more than one result's URL at a time; this can be very helpful, as it eliminates the need to switch back and forth between Gemini UDS and your browser when viewing results. To do this, simply select the results you wish to open by checking the checkbox next to those results. Then go to the Results menu, and select "Open Selected in Browser"; this action will cause Gemini UDS to instruct your browser to open all of the results you selected in new windows. (Note: some versions of Internet Explorer do not support this command, and will simply display the last result you selected in an already-open window.)

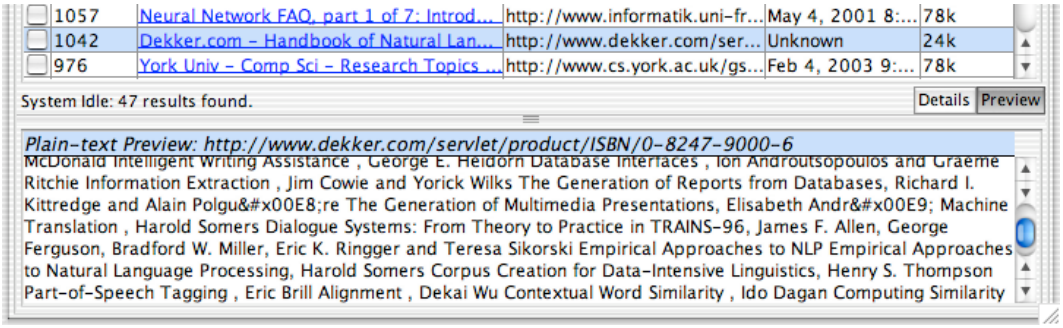
As you may have noticed, Gemini UDS displays a great deal of information about each search result,

from its relevance rating to its content's size to when the content was last modified. If you prefer, even more information can be shown about each search result by selecting a search result's row in the table and clicking the "Details" button at the lower right of the results area (you can also double-click on a search result row to open the details pane corresponding to that row). Doing so will result in this type of display:



The details pane provides even more information about a search result. In addition to the typical bits about the URL and page title, the details pane tells you where the result was found (either through a search service, or via Localized Webcrawling), which of the content types you selected in your query are referred to by the search result's content,

Next to the button that opens the details pane, there's a button labeled "Preview" that will open up a plain-text preview pane. Opening it results in a display like this:



The text in the preview pane is the complete content of the search result page you have selected in the results table, with all of the formatting and images removed (to make it easier and faster to display and scan as you are clicking through many sets of results). Having access to such a preview right in the results display can help you to select which results you wish to view in full in your browser. Previews are not available from results that you have opened from saved query files.

You can choose the font used to render the results table and the details and preview panes in the Gemini UDS preferences, as described in the section *User Interface* on page 27.

Relevance Ratings

You'll quickly notice while using Gemini UDS that it does not *normalize* (keep within a certain range, like 1-100) the ratings it assigns to each search result to indicate relevancy. The default Gemini UDS Evaluator plugin does not normalize these ratings so as to provide you with a better indication of relevancy than a simple number between 1 and 100 (which can be very constraining and useless when your result sets reach into the hundreds). Other Evaluator plugins may normalize ratings at their discretion.

Sorting Results

Most search engines and other search tools don't allow you sort your results in different ways; usually, they restrict you to seeing your results in the order of their ranking. Gemini UDS enables you to sort your results in a number of different ways that make it easy for you to get the information you need in the order you need it in.

Gemini UDS enables you to sort your results in ten different ways. Changing how the results are sorted is simply a matter of clicking in the results table's header on the criterion by which you wish to sort. So, to sort your results by their rating, just click on the header at the top of the rating column.

You may choose to sort the results in an ascending or descending "direction" for each of the following criteria:

Rating: Selecting this method will cause Gemini UDS to sort your results by their relevance rating, which is determined by the evaluator you used for the query that generated the results.

Title: Selecting this method will cause Gemini UDS to alphabetically sort your results by the results' pages' titles.

Date Modified: Selecting this method will cause Gemini UDS to sort your results according to when their content was last changed.

Page Size: Gemini UDS tracks how large each search result's content is in kilobytes (KB). Selecting this method will cause Gemini UDS to sort your results by their content's size.

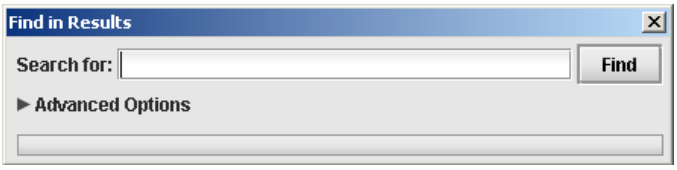
URL: Selecting this method will cause Gemini UDS to sort your results alphabetically by their URLs.

Note that you can sort on any of these criteria in descending or ascending order. The first time you sort on a search results column, the results are sorted in ascending order, starting with the least result (lowest rating, smallest results page content, etc) and ending with the greatest result (highest rating, largest content, etc). You can reverse this ordering by clicking on a column header that has already been sorted. An arrow will appear to the right of the header title indicating how the results have been sorted.

Searching Within Results

By design or by accident, you will eventually get too many results back from a search in Gemini UDS. In times like this, you will find Gemini UDS' subsearch feature to be your best friend.

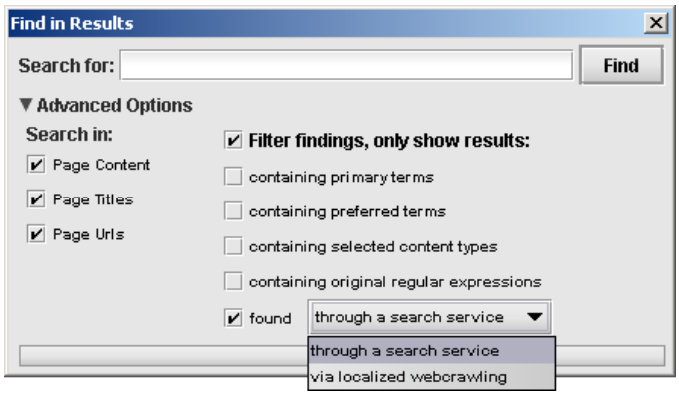
Once Gemini UDS has begun to stream results into the results area, you can perform subsearches within those results. To do so, select the "Find in Results" item from the Results menu. That will bring up the subsearch dialog:



Here you can enter a set of keywords to search for only within the results you have showing in the results display. The subsearch will (by default) require that the keywords you enter be found in either the title, URL, or actual content of the search results that are showing. The results that match that subsearch will then be shown in a new results area.

NOTE: you cannot search through the results' content if those results have been loaded from a saved query file; in such circumstances, only the title and URL of the results will be searched.

If you want to control how the subsearch is conducted and filter the results of the subsearch, just click on the Advanced Options button, which will reveal the rest of the subsearch dialog:



Here you can specify where the entered keywords should be found (again, in the title, URL, or content of the current search results, or any combination thereof), and you can also require that subsearch results match certain other criteria. These criteria are called filters in the dialog, and they match many of the properties that you can view in the results display areas:

- **Containing primary terms:** Selecting this filter will cause the subsearch to show only those results whose content contains the Primary Terms you specified in your query.
- **Containing preferred terms:** Selecting this filter will cause the subsearch to show only those results whose content contains the Preferred Terms you specified in your query.
- **Containing selected content types:** Selecting this filter will cause the subsearch to show only those results whose content refers to one or more of the content types you selected in your query.
- **Containing original regular expressions:** Selecting this filter will cause the subsearch to show only those results whose content matches one or more of the regular expressions you entered into the primary or preferred terms fields in the query.
- **Were found (via Localized Webcrawling / through a search service):** Selecting this filter will cause Gemini UDS to show only those results that were found using the method selected in the dropdown menu

Let us take the preferred terms filter as an example, as it is one of the most useful. As you saw in the section "Forming Your Query", Gemini UDS provides you with two areas in which to define search terms, Primary and Preferred (refer back to that section for a full explanation of the difference between Primary and Preferred Terms). In many cases, you will run a search that finds many results that contain the Primary Terms you entered, but only a few that contain the Preferred Terms (exactly the point of separating those two fields). However, it's not immediately obvious which results contain those crucial Preferred Terms. Selecting the preferred terms filter in the subsearch dialog will cause only those

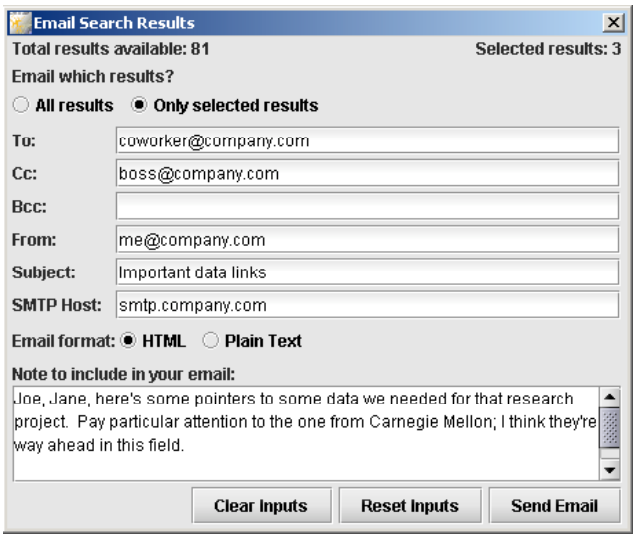
results that contain your Preferred Terms to be shown in the new results area, solving the problem of finding the results that contain those preferred terms instantly.

You may also combine filters to further narrow the set of results that are found by the subsearch. In addition, you can perform subsearches on sets of results that were found using a subsearch. This provides a great deal of flexibility when working through a set of results, as you are able to narrow down the result set in multiple ways very easily by, for example, applying a filter to your initial results, then searching the titles of the results of that filtering, then applying another filter to that subset, etc., etc. Gemini UDS does not limit how many subsearches you can perform on a set of results.

Emailing Search Results

Emailing search results to colleagues and coworkers has always been a very common activity, but it can be a very time-consuming one as well: tediously copying-and-pasting between a web browser and your email client is not the most pleasant task, and the results rarely look neat or professional. Thankfully, Gemini UDS enables you to email your search results in a simple, no-hassle way that will result in professional-looking results being viewed by your recipients.

Once your search is completed, you can email your results by selecting "Email Results" from the Results Menu. Doing so will bring up this dialog:



Most of these fields correspond to those you would fill out when composing an email message in your normal email client. (For an explanation of what these fields mean and how you can set default values for them to avoid typing things like your own email address in repeatedly, see the *Email Defaults* section on page 30).

Note: Your email will be sorted and filtered exactly as it appears in the results display area you are viewing when you select the email results command.

There are a couple of elements in this dialog that you may not be familiar with. At the top of the dialog is a count of how many results are currently available (i.e. how many were found by your search) and

how many you've selected (results may be selected by checking off the boxes next to each search result). Below that, you can select whether to email all of the available results, or only those that you've selected. So, if you've selected three results out of a set of 50, only those three results will be emailed to your recipients.

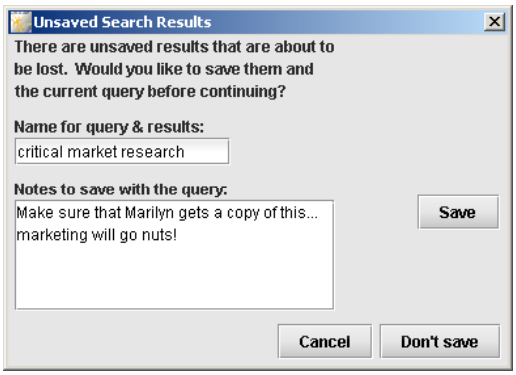
You can also include a short note with the email that will appear above the results by typing in the text field at the bottom of the email dialog. This note can be used for a greeting and perhaps a small explanation of what the results pertain to.

You can send your email by clicking the "Send Email" button. Or, if you've made a mistake, you can click the "Clear Inputs" button to erase all of the entries and make them blank, or your can click the "Reset Inputs", which will reset them to the defaults you've saved in Gemini UDS' Preferences (see page 30 for more information about setting email defaults).

Saving and Opening Results

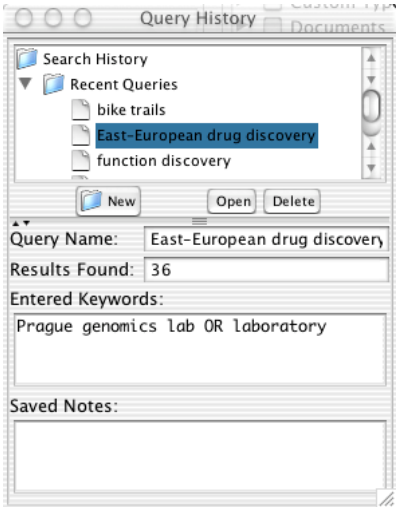
Nothing is more frustrating than finally finding what you've been searching for for an hour than to have *something* happen that causes you to lose that critical information: lightning strike, program malfunction, your cat stepping on the keyboard, etc. For those (and many more) reasons, Gemini UDS includes the ability to save and retrieve your search results (along with the query that led to them).

Saving results is very simple. When your search is complete, you can select the "Save Query and Results" item under the File menu. This will bring up a dialog from which you can save the results currently in the results display area and the query that resulted in those results:



Here you can give a name to the query/results file, and add some short notes for your later use. Clicking "Save" will then save the query and results to your hard drive. Each new set of query and results you save will be placed in the "Recent Queries" directory that is accessible via the Query History window.

Opening your saved queries and results is done through the Query History window. To open the Query History, select "Show Query History" from the Query menu. A new window will be opened that will look something like the image below.



The Query History window has two parts: the file manager above, and the info viewer below. You can use the file manager to move, delete, and rename files and folders just as you would in the Mac OS X Finder, the Windows Explorer program, or in any of the many Unix file managers.

When you click on a file (the Query History window will only show you files that contain a query and results), the info viewer portion of the window will update to show you some summarized data about that query and results file, including the name you gave to it when it was saved, the number of results it holds, which terms were used in the query, and the notes that you saved along with the query.

You can then open a query and its results either by double-clicking on its file in the file manager portion of the Query manager, or you can select the file and click on the "Open" button that sits between the file manager and the info viewer. Opening a query and its results causes two things to happen:

- The query entry interface in Gemini UDS' main window is set to have the parameters that were used to obtain the saved results.
- The saved results are loaded into the results display area, where they can be viewed, opened, sorted, and filtered just as if you had run the search moments before.

Once you've opened a query and results file, you can switch between the Query and Status/Results tabs as necessary to examine the saved query and search results.

Finally, you also have the option of exporting your results as a regular HTML file that you can view using your browser, save in another location for future reference, or whatever. To export your results in this way, just choose the "Export Results to HTML…" item in the File menu. Then use the file dialog that appears to navigate to where you'd like to save the HTML file, enter a name for the file, and click the "Export" button. Your results will then be saved as HTML where you indicated.

It is very clear that each person's search and information retrieval needs are likely to be very unique. For this reason alone, Gemini UDS was designed to be very customizable. However, it goes above the call of duty by enabling you to:

- Integrate external search engines and information sources into your Gemini searches
- Search for *custom* content types
- Utilize customized information analysis plugins suited to the kind of data you're working with
- Create new information analysis plugins specialized for your specific industry, problem, or field of expertise.

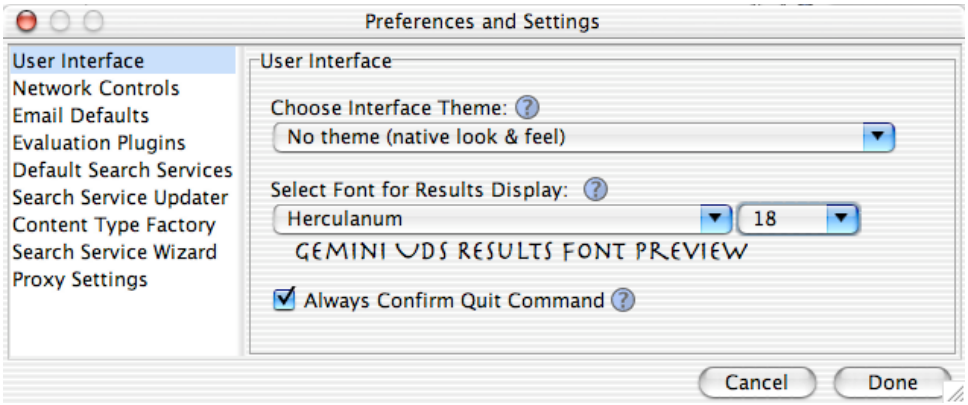
This chapter of the user's guide will take you through each of these aspects of Gemini UDS' customization features, as well as delve into the more ordinary (but no less useful) configuration options that Gemini UDS provides. Some of these options are complex and powerful, and are therefore broken out into their own sections within this chapter.

Advanced Configuration Options

Essentially all of Gemini UDS' configuration options are accessed via the Preferences window. This window can be opened by selecting it from the Edit menu (Mac OS X users can open the Preferences window by selecting the Preferences item in the Gemini UDS application menu to the left of the Apple menu). When you first open it, the left side of the Preferences window should contain a listing much like the one shown to the right (the "Web Browser Config" item will only appear on Unix systems).

This listing is of the set of available "control panels" in the Preferences window; clicking on each one leads to a different set of options that you can use to affect how Gemini UDS behaves. Please note that the Evaluation Plugins, Search Service Wizard, and Content Type Factory panels are special cases that are covered later in this chapter.

User Interface



This panel (called "User Interface" in the listing on the left side of the Preferences window) is where you can change a number of options associated with how Gemini UDS looks. We'll run through each of these in turn:

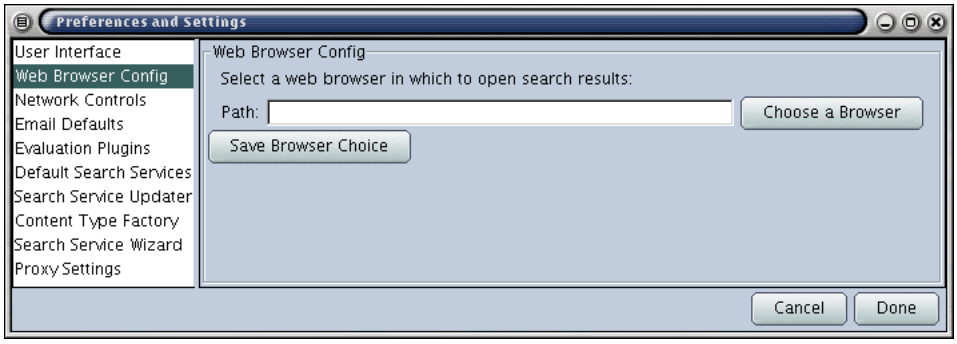
- **Choose Interface Theme:** here you can select which interface look and feel you wish to use. Gemini UDS ships with a set of themes; once you select one, restart Gemini UDS, and it will appear with your selected theme activated.
- **Select Font for Results Display:** here you can select which font type and size are used to render the results display. Once you choose a font from the pull-down menu, a preview of that font's appearance will appear below the menu.
- **Always Confirm Quit Command:** If this checkbox is selected, then Gemini UDS will always ask for confirmation before quitting; otherwise, it will quit immediately when it is ordered to do so.

Web Browser Config (Unix only)

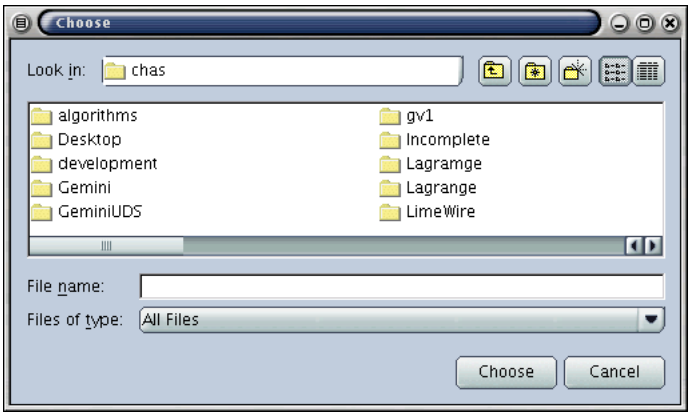
Opening search results in your browser is essentially an automatic process on Mac OS X and Windows: Gemini UDS simply asks the operating system to open a URL, and the OS passes that information on to your preferred browser (usually Internet Explorer on Windows; Mac OS X users can select their preferred browser in the Internet panel in System Preferences).

Unix does not have such an ability, so Gemini UDS must communicate directly with a browser on that platform. Gemini UDS by default attempts to open results by passing the result's associated URL to the Netscape Navigator browser (usually located at */usr/bin/netscape* or */usr/local/netscape*). However, if you do not have Netscape installed, installed at those locations, or if you wish to use a different browser, you must configure Gemini UDS to "point to" your preferred browser before it can open any results in that browser.

Therefore, when Gemini UDS is running on Unix (or any non-Windows or non-Mac OS X platform), an additional preferences panel is available, called "Web Browser Config" in the listing on the left side of the Preferences window:

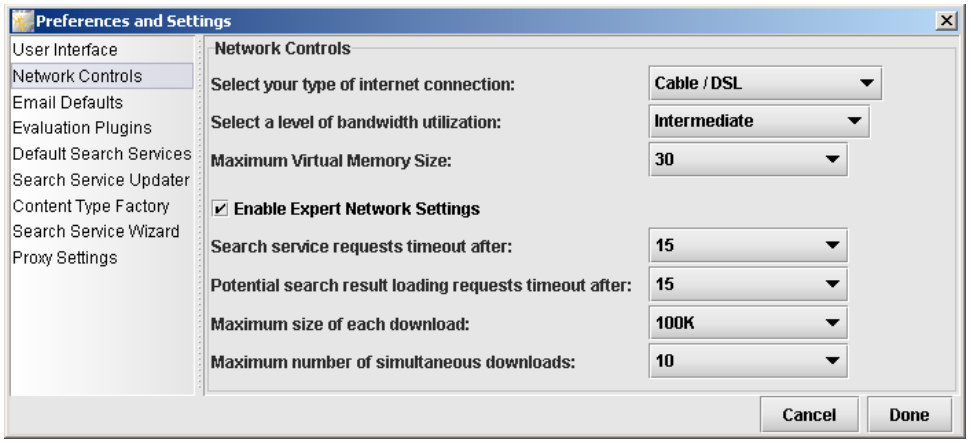


Here you can provide the path where your preferred browser resides by typing the path in manually or by selecting the browser from a graphical dialog. You can bring up the graphical dialog by clicking the "Choose a Browser" button, which will bring up a dialog that looks like this:



Navigate to where your browser is installed, and select it; when the dialog closes, the browser's path should be showing in the text field. Finally, click on the "Save Browser Choice" button to save your selection; simply clicking the "Done" button in the Preferences window will not save this setting. After that, when you click on a link in the results display, Gemini UDS will open the browser you specified and "tell" it to open that link's URL.

Network Controls



This panel (called "Network Controls" in the listing on the left side of the Preferences window) is where you can control all aspects of how Gemini accesses and utilizes your network connection. This first set of network controls are used if you do not use the expert network settings described a little later:

- **Select your type of internet connection:** here you should select what kind of internet or network connection you have.
- **Select a level of bandwidth utilization:** here you should select how much of your available bandwidth Gemini UDS should feel free to use. If you find that your connection becomes very slow while using Gemini UDS "in the background", you can lower Gemini UDS' usage of the connection's resources with this control.
- **Maximum Virtual Memory Size:** here you should select the maximum amount of space (in megabytes, MB) that Gemini UDS may use to temporarily store data about potential search results. If you plan on searching through very large or very many documents, increasing this value is recommended.

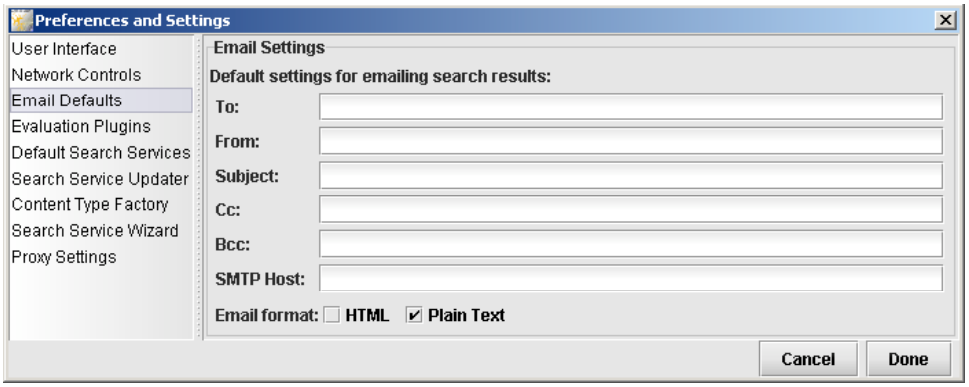
Unless you manually set the expert network settings as described below, Gemini UDS will use the above settings to make educated guesses about how it should access and utilize your internet or network connection. The expert settings are:

- **Enable Expert Network Settings:** If this checkbox is selected, then Gemini UDS will adhere to the selections made in the expert network settings instead of those made in the simpler network settings as described above.
- **Search service requests timeout after:** This setting controls the maximum number of seconds Gemini UDS will wait for a response from all of the selected search services before pushing forward in the search process. If a search service does not respond within the selected time limit, then Gemini UDS will not be able to extract potential search results from that service.
- **Potential search result loading requests timeout after:** This setting controls the maximum number of seconds Gemini UDS will wait for a response from each potential search result. If a potential search result's content cannot be downloaded in the specified time, then it is assumed to be inaccessible (i.e. a "dead link") and is not included in any further processing.
- **Maximum size of each download:** This setting controls the maximum amount of content Gemini UDS will download per potential search result. For example, if a potential search result's content is actually 287K, and this setting is set to 150K, then Gemini UDS will download the first 150K of that content and determine the content's relevancy based on that segment. This is done to ensure that Gemini does not get "stuck" downloading overly large files and draining system resources in the process.
- **Maximum number of simultaneous downloads:** This setting controls how many potential search results will be downloaded simultaneously. If you have a faster connection, a higher setting will make Gemini UDS much faster; if this setting it set too high, your network connection (and your system as a whole) may suffer.

Remember to click the "Done" button at the bottom-right corner of the Preferences window when you're finished changing settings; otherwise, your changes will not be saved.

Email Defaults

Gemini UDS allows you to send search result listings to your friends, colleagues, and coworkers from within Gemini UDS without laboriously cutting-and-pasting (or otherwise transferring) the results to your email client. (See page 24 for more information.) However, there is much that can (and should) be set up prior to sending email messages; this panel (called "Email Defaults" in the listing on the left side of the Preferences window) allows you to make those settings, which will be the default ones used when emailing search results:



In other words, the entries you make here will be automatically inserted into the 'Email Results' dialog, saving you from retyping that information every time you email results to someone.

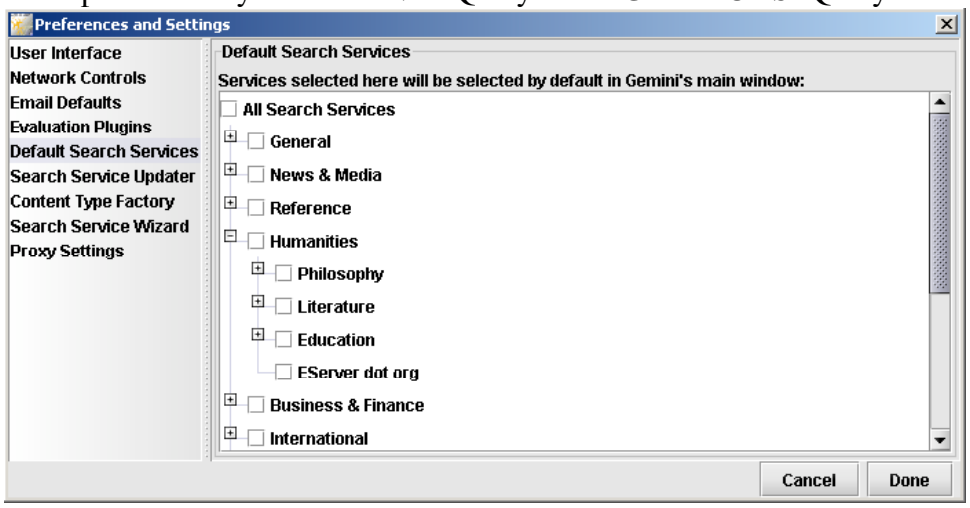
The possible settings are as follows:

- **To:** Here you can input a list of default email addresses you would like to send results to often. This list of email addresses should be separated by commas (i.e. "user@host.com, John Harvard <jharvard@cambridge.org>").
- **From:** This should be your name and email address, like so: Gail Merrill gmerrill@company.com
- **Subject:** You can enter a default subject for your emails here.
- **CC:** This field is essentially the same as the To: field. Recipients listed here are considered "secondary" to those listed in the To: field.
- **BCC:** This field also acts like the To: field, but the recipients listed here will not be known to anyone else listed in the To: or CC: fields, and vice-versa.
- **SMTP Host:** enter the hostname of your outgoing email server here. This can essentially be copied directly from the corresponding field in your email client's configuration. Ask your system administrator or ISP about this if you do not know what your SMTP Host is.
- **Email format:** Gemini UDS can format emails in two ways, either plain text (which will work in all email clients) or HTML (also called rich text). HTML is generally a safe choice unless you know your recipients are using older or text-based email clients.

As always, click the "Done" button at the bottom-right corner of the Preferences window when you're finished changing settings; otherwise, your changes will not be saved.

Default Search Services

This panel (called "Default Search Services" in the listing on the left side of the Preferences window) is where you can control which search services are selected by default in the query entry interface when Gemini UDS starts up and when you select 'New Query' from Gemini UDS' Query menu.

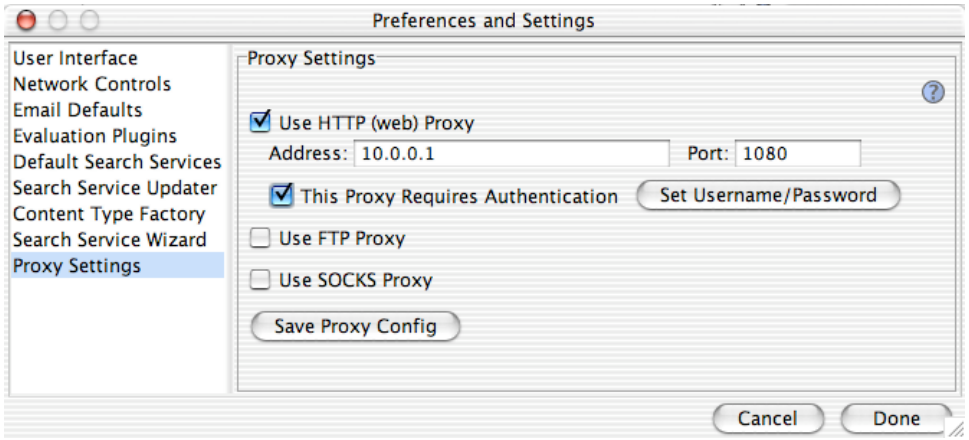


The hierarchy of search services you see in this preferences panel is identical to what you see in the query interface, and it works identically as well. To select a set of default search services, simply select

them in this panel, and click the "Done" button at the bottom-right corner of the Preferences window when you're finished changing settings; otherwise, your changes will not be saved.

Proxy Settings

This panel (called "Proxy Settings" in the listing on the left side of the Preferences window) is where you can control how Gemini UDS accesses and utilizes proxy servers that your school or company might have installed. (If you do not know what proxy servers are, chances are you do not need to set anything in the Proxy Settings panel; if you're not sure, ask your system administrator.)



Gemini UDS support three different types of proxies:

- HTTP (web) proxies are the most common, and are used exclusively for accessing data using the HTTP protocol (such as web pages).
- FTP proxies are used exclusively for accessing data using the FTP protocol, which is most common when accessing larger non-text files.
- SOCKS proxies can be used to access data using a range of different protocols (including HTTP and FTP).

It should be noted that Gemini UDS currently supports SOCKS v4 proxies, but not SOCKS v5 proxies.

Configuring any of these types of proxies is easy, and the same process is used regardless of the type of proxy being configured:

1. Check off the checkbox next to the proxy server type that you need to configure.
2. Enter the address (host) and port the proxy uses.
3. If necessary, enter a username and password for accessing the proxy by selecting the "This proxy requires authentication" checkbox and then clicking on the "Set Username/Password" button. This will bring up a dialog where you may enter your authentication information. Note that not all proxies require authentication.
4. Click the "Save Proxy Configuration" button; this is necessary to save the configuration. In this case, simply clicking the "Done" button in the Preferences window is not sufficient for saving this information.

NOTE: Configuring a SOCKS proxy in this panel will cause Gemini UDS to use that proxy for *all* network communications, even if you've also configured HTTP and/or FTP proxies as well. This is because SOCKS proxies can handle both of those protocols as well as many more.

Once you've saved the proxy configuration you've entered, Gemini UDS will use the enabled proxy(ies) for all network activity, including searching, the Search Service Updater, and the Application Updater.

Integrating Information Resources

The 'Unified' part of *Gemini Unified Datamining System* refers to the fact that Gemini UDS unifies and integrates a number of different information sources so that you can access all of them from a single interface – the one provided by Gemini UDS. These information sources can be public search engines to subscription services to private databases and resources that only you and your colleagues have access to.

This integration has traditionally been impossible: each information source (called a *search service* in this guide and in Gemini UDS) uses its own "protocol", a way to communicate between itself and its clients (typically a web browser). Since there are no standards for what information can or should be passed between search services and their clients, it is very difficult to use a single tool to access them all in an easy, automated fashion that can boost productivity. Despite the difficulties involved, Gemini UDS can access virtually any search service that you can access with a web browser in an automated, easy, and more productive way.

You can create a new search service by using the Search Service Wizard, which is in the listing on the left side of the Preferences window. The Search Service Wizard includes extensive directions, which we won't repeat here. Simply follow the on-screen instructions when using the Search Service Wizard, and Gemini UDS will be able to create a search service description for your custom information source; it will be accessible to you in the "Custom Services" category in the query interface.

The Search Service Wizard has two limitations. First, it does not currently handle query forms that contain frames; if the information source you wish to integrate into Gemini UDS uses frames, then it is advisable that you determine exactly which web page the query form is on, and provide that page's URL to the Wizard instead of the "parent" URL that contains frames.

Secondly, the Wizard supports one of the two possible authentication schemes regularly associated with search services. The first, which is fully supported, is called HTTP Authentication. You can tell whether or not a search service uses HTTP Authentication by what happens when you attempt to access the service: if your browser pops up a window asking for a username and password, then the service is using HTTP Authentication. The second authentication scheme, which is not currently supported, is where you enter your username and password into a regular webpage – no special window is displayed by your browser.

Defining and Using Custom Content Types

One of the most powerful tools that Gemini UDS offers is the ability to find essentially any type of content in a set of search results. This means that not only can Gemini UDS deliver excellent search

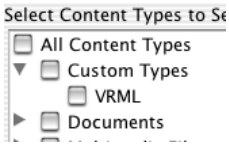
results, but it can easily inform you of what kind(s) of content are embedded in or linked from each search results page. This can be a huge timesaver when you need to find a particular kind of file, but it is in a format that cannot be easily analyzed by Gemini (such as multimedia files, or proprietary formats like Microsoft Word documents).

Gemini UDS comes with dozens of content type "descriptions", which allow you to search for the most popular content types; those types appear on the right side of the query interface, which you can select individually or by category.

Gemini UDS also allows you to create your own content type "descriptions", which will appear alongside all of the content types that come with Gemini UDS. Creating a custom content type is easy:

1. Open Gemini's preferences window (Edit Menu ▸ Preferences on Unix and Windows, Gemini UDS ▸ Preferences on Mac OS X).
2. Select "Content Type Factory" from the listing on the left.
3. Enter the name of the new content type in the first field.
4. Enter the file extensions that are usually associated with that content type's files. For example, JPEG images, which are used widely on websites, usually have a '.jpg' or '.jpeg' file extension. **Note: These file extensions should be separated by a single space and not include any periods, like so: "jpg jpeg".**
5. Optionally, you may enter a description of this new content type, which will appear over its entry in the query interface when tooltips are on.
6. Click "Save Content Type".

Repeat this process for each content type you wish to create. The result of this process is twofold. First, a content type description is saved on your hard drive. Secondly, when you close the Gemini UDS Preferences window (by clicking the "Done" button) and each time you start Gemini UDS from that point forward, your new custom content type will appear under the "Custom Types" category in the Gemini UDS query interface:



After that point, your custom content type will always be available; when you perform a search with that content type selected, Gemini UDS will search all of the search results' content for instances of that content type and indicate its presence in the results display.

Evaluation Plugins

Most search engines use a proprietary scheme for evaluating potential search results; this scheme is responsible for which results you see (and don't see) when you use a particular search engine, and the order in which those search results are presented to you. Gemini UDS is similar in that it comes with an evaluation engine (called a plugin) that examines each potential search result's content and other characteristics to determine if it is relevant to your query.

The difference with Gemini UDS is that you can change which plugin is used to conduct this evaluation, and you can use and develop custom plugins to solve specific information analysis problems. Being able to select the most appropriate evaluation plugin for your particular search or research problem can lead to remarkably improved results compared to using an inappropriate evaluation scheme.

Selecting A Plugin

Choosing which plugin to use is simple:

1. Open Gemini's preferences window.
2. Select "Evaluation Plugins" from the listing on the left.
3. Choose the plugin that you wish to use from the list of evaluation plugins at the top of the window (most well-written plugins will include a description of what the plugin does in the text area at the bottom of the window).
4. Click the "Done" button in the lower-right corner of the Preferences window.

This process will load the plugin you selected, which will be used for all of your searches until you change the selected plugin. Please note that each plugin may analyze potential search results very differently than other plugins, including the default plugin that Gemini UDS ships with. Please refer to the documentation that came with your plugin to determine if it is appropriate for the searching and information analysis you wish to perform.

Installing Plugins

All evaluation plugins (except for the default one shipped with Gemini UDS) reside in the "Plugins" directory in your Gemini UDS application directory. When you wish to use a new plugin, simply drag its *.class* file into that Plugins directory, and restart Gemini UDS – the new plugin should then appear in the Evaluation Plugins section in the Gemini UDS preferences.

Developing New Plugins

Most of the time, Gemini UDS' default plugin is sufficient for the searching that you need to do. However, there may come a time where you need to perform some more complicated or specialized analysis on potential search results.

For example, if you are doing specialized polymer research, and you have integrated your research database into Gemini UDS, you would be able to analyze the data in that database far more effectively with a custom plugin designed for that task instead of a plugin that is more general-purpose in its functionality.

For just this sort of occasion, Snowtide has published the full specification for building custom Evaluation plugins. This specification is a Java-based API; in very broad strokes, developing a plugin involves writing your specialized analysis code, interfacing with the Gemini UDS Evaluator API, and placing the compiled *.class* file in the Plugins directory.

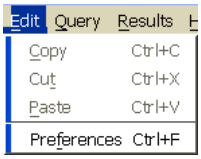
Snowtide Informatics maintains a developer resource center online that has links to documentation of the Evaluator API and tutorials on how to get started writing custom Evaluation plugins. This resource center is available at http://www.snowtide.com/developer/.

This chapter will detail all of the menu commands available within Gemini UDS, provide their corresponding keyboard shortcuts for each computing platform, and give a brief description of what they do. Nearly all of these commands are part of a set of functionality within Gemini UDS that is completely explained elsewhere in this guide.

File Menu

	Keyboard Shortcuts			
Menu Item	**Mac OS X**	**Windows**	**Unix**	**Purpose**
About Gemini UDS (Unix and Windows only)	None	None	None	Shows the Gemini UDS "about box". Mac OS X users can se the about box by selecting "About Gemini UDS" from the Gemini UDS menu to the left of the Apple menu.
Close	⌘-W	Ctrl-W	Ctrl-W	Closes the current active Window. Not available when the main Gemini UDS window is the current window.
Export Results as HTML	⌘-R	Ctrl-R	Ctrl-R	Allows you to export the current results as a regular HTML file.
Save Query and Results	⌘-S	Ctrl-S	Ctrl-S	Allows you to save your current query and search results in a format that can be later loaded back into Gemini UDS. You may only select this item if no search is being performed.
Quit (Windows & Unix only)	⌘-Q	Ctrl-Q	Ctrl-Q	Quits Gemini UDS. If you have unsaved results, Gemini UDS will prompt you to save them before quitting. Mac OS X users can quit by selecting "Quit" from the Gemini UDS menu to the left of the Apple menu.

Edit Menu



	Keyboard Shortcuts			
Menu Item	**Mac OS X**	**Windows**	**Unix**	**Purpose**
Copy	⌘-C	Ctrl-S	Ctrl-S	Copies the current selection to the clipboard.
Cut	⌘-X	Ctrl-X	Ctrl-X	Copies the current selection to the clipboard and deletes the current selection.
Paste	⌘-V	Ctrl-V	Ctrl-V	Inserts the contents of the clipboard into the current text field.
Preferences	None	None	None	Opens Gemini UDS' Preferences window. This item is in the Edit menu only on Windows and Unix; Mac OS X users can open the Preferences window by selecting the Preferences item from the Gemini UDS application menu (next to the Apple menu).

Query Menu

	Keyboard Shortcuts			
Menu Item	**Mac OS X**	**Windows**	**Unix**	**Purpose**
New Query	⌘-N	Ctrl-N	Ctrl-N	Switches the view to the query entry interface, and resets all fields to their defaults.
Start Query	⌘-Enter	Ctrl-Enter	Ctrl-Enter	Starts a new search with the current entries in the query entry interface.
Stop Query	⌘-.	Ctrl-.	Ctrl-.	Stops any currently-running search.
Use Standard Interface	None	None	None	Switches to the standard query entry interface. If this item is unavailable, you are already using the standard interface.
Use Advanced Interface	None	None	None	Switches to the advanced query entry interface. If this item is unavailable, you are already using the advanced interface.
Show Query	⌘-Y	Ctrl-Y	Ctrl-Y	Opens the Query History window,

History				through which you can view, open, and manage queries and results that you've saved previously.

Results Menu

	Keyboard Shortcuts			
Menu Item	**Mac OS X**	**Windows**	**Unix**	**Purpose**
Email Results	⌘-E	Ctrl-E	Ctrl-E	Opens the Email Results dialog, which allows you to email your selected results to friends, coworkers, and colleagues.
Open Selected in Browser	⌘-B	Ctrl-B	Ctrl-B	Opens all of the selected results in your browser (note that not all browsers support this command).
Use Selected as Seeds	⌘-D	Ctrl-D	Ctrl-D	Starts a new query that has the selected results' URLs inserted as seed addresses (requires use of the advanced query interface).
Clear Result Selections	None	None	None	Clears all of the selections made in the results area.
Hide Search Result Pages	⌘-U	Ctrl-U	Ctrl-U	Toggles whether or not the results display will show search result pages that it has found via the Localized Webcrawling process. Generally, these pages are very numerous, and are best used as a part of the Localized Webcrawling Process.
Reset Table Column Layout	None	None	None	Returns the size and ordering of the result table columns back to their defaults.
Find in Results	⌘-F	Ctrl-F	Ctrl-F	Opens a subsearch dialog for searching within a set of found results. See page 22 for details.

Help Menu

	Keyboard Shortcuts			
Menu Item	**Mac OS X**	**Windows**	**Unix**	**Purpose**
Show Tooltips	⌘-W	Ctrl-W	Ctrl-W	Turns on tooltips.
Toggle Help	⌘-T	Ctrl-T	Ctrl-T	Toggles the help popups that

Popups				appear next to many components in the Gemini UDS interface. These popups appear as a yellow question mark, and can be viewed by clicking on that question mark.
Check for Gemini UDS Update	None	None	None	Opens the Gemini UDS Application Updater, which allows you to automatically check for and install updates to Gemini UDS.

Appendix A Localized Webcrawling Overview

This Appendix offers an overview of how Localized Webcrawling works, the process that drives Gemini UDS. For a perspective on how Localized Webcrawling compares to "traditional" webcrawling, other existing information processing systems, and what Localized Webcrawling can do for you in terms of productivity, see <http://snowtide.com/home/products/Gemini/lw.jsp>.

Localized Webcrawling is a process by which a set of hyperlinked resources can be processed exhaustively **with direction and guidance provided by the user**. In the case of Gemini UDS, Localized Webcrawling automates the processing of hundreds or thousands of web pages each time a user provides a query. This automation involves three main phases:

- **Loading** - The content of a set of seed addresses/URLs is loaded and stored.
- **Evaluation** - The content of the loaded seed addresses/URLs is analyzed and evaluated to determine each address' relevancy to the user's query. This analysis can range from simple matching of keywords to determining whether or not a particular address contains some non-text content types to complex semantic or natural language processing.
- **Crawling** - Links to other URLs are extracted from the content of the most relevant addresses analyzed in Step 2; these new links serve as the seed addresses that are used to start another cycle of Localized Webcrawling (i.e. go back to Step 1).

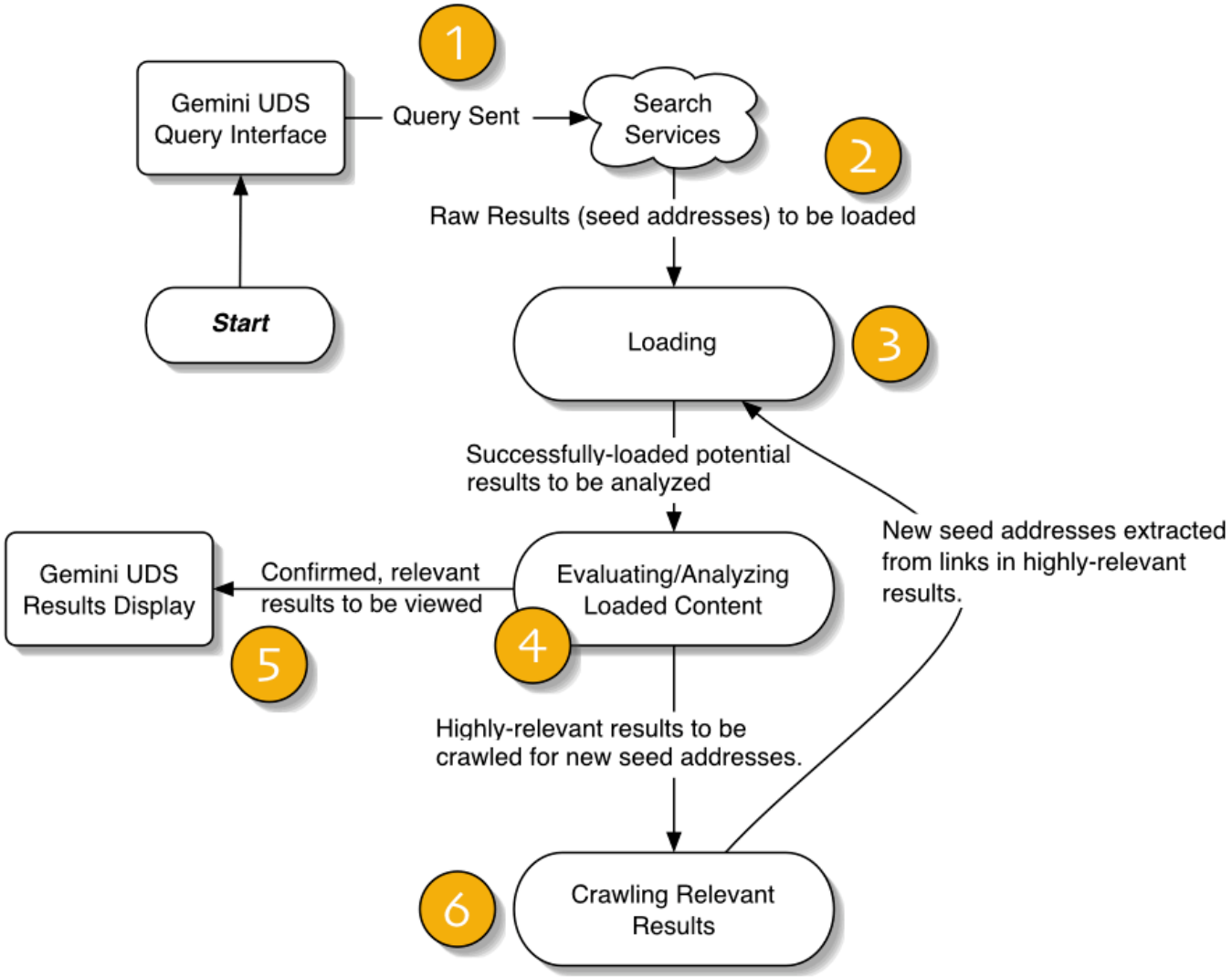
This process mirrors the process that you must engage in when searching for information using "regular" search engines and services. The only difference is that instead of manually clicking through link after link looking for relevant pages, Localized Webcrawling does everything without your constant oversight and instruction once you've provided it with a query.

To better explain this process, the diagram on page 42 provides an illustration of it. The rest of this Appendix refers to that diagram.

When you start a search with Gemini UDS, the first thing it does is format your query, which is sent to each of the search services you selected (step 1 in the diagram). As the search services reply to Gemini UDS' query, it extracts all of the raw results from those responses (represented as step 2 in the diagram); the addresses of those raw results become the seed addresses for the Localized Webcrawling process. It is at this point that any seed addresses you provided manually are added to the mix. (See page 9 for a definition of what seed addresses are.)

Once all of the seed addresses are gathered from search services and your manual input, each of them are loaded in step 3. This process accomplishes two things: it retrieves the content of each seed address (which may become a confirmed relevant search result after analysis), and it weeds out any seed addresses that cannot be accessed. This eliminates all of the "dead links" that you might otherwise have to wade through when using those selected search services.

When all seed addresses have been given a chance to load, those that have successfully been loaded are then evaluated and analyzed in step 4. Here, your selected Evaluation plugin goes to work determining which seed addresses are relevant to your query. As the plugin finds relevant results, they are immediately sent to the Gemini UDS results display (step 5).



Once all of the loaded seed addresses are analyzed, any that were found to be irrelevant are discarded. Then a subset of the relevant results (those that were determined to be the most relevant) are then crawled in step 6. This crawling process gathers all of the links in those highly-relevant results' content. Those links then become a new set of seed addresses, which are sent back to step 3.

As you can see, this process is repetitive, iterative to be exact. That is why the advanced query interface allows you to specify how many iterations the Localized Webcrawling process should run for. Specifically, each iteration consists of one cycle of loading, evaluating, and crawling.

Appendix B Search Term Field Syntax

This Appendix details the acceptable syntax that may be used in the Primary and Preferred Search Terms fields that are used in Gemini UDS. By understanding and using the syntax elements described here properly, you will give Gemini UDS the raw material it needs to do its job most effectively.

The search terms fields accept each of the following elements; each element **must** be separated from other elements by a space. Do not use commas, semicolons, '+' or '-' signs, or any other operators in front of the elements you input into the fields – doing so will include those symbols in your query, something that you likely do not want to do.

Acceptable search term elements:

- Keywords
- Phrases/Quotes
- Boolean Expressions
- Weighted Keywords
- Regular Expressions

Let's go through each one and detail how to properly use them and cover some examples. As we progress through each search term element, we'll add to the same example, a query meant to find information about multiple sclerosis treatments (a kind of neuromuscular disorder).

Note: much of the information in this Appendix is based on the default Gemini UDS Evaluation plugin. If you use a different plugin with Gemini UDS, it may treat these query elements differently; consult the documentation for your plugin to see if it differs from what is stated here.

Keywords

Keywords are simple words like 'dog', 'cat', 'endometriosis', 'store', etc. They are the foundation for the syntax, and your most direct way of letting Gemini UDS know what you wish to find. There are certain "special" characters and a couple of words that you may not use as part of a keyword; they are reserved for use in other parts of the syntax, and are discussed in full later in this Appendix. The reserved characters are:

- Double quotes "like this" – used to define phrases and literals
- Parentheses (like these) – used to group search terms for various purposes
- Asterisks (*) – used to define weights
- Forward slashes (/) – used to define regular expressions
- The words "and", "or", and "not" – used to form boolean expressions

While you cannot use these words and characters directly, you can use them in your search terms if you surround them with double quotes; see the Phrases/Quotes section for details.

Note: the default Gemini UDS Evaluation plugin treats search terms in a case-insensitive manner. In other words, 'dog' and 'DOG' are identical as far as that plugin is concerned; other plugins may or may not be case sensitive.

So, our example query might start like this, using only simple keywords for now:

```
multiple sclerosis treatment
```

This is a perfectly valid query that will yield good results. As we progress through discussing each element of the Gemini UDS syntax, we'll enhance this query to be the best it can be.

Phrases / Quotes

Gemini UDS supports phrases using double quotes. A phrase is a set of keywords joined using double quotes; using a phrase in your query tells Gemini UDS that the keywords that make up the phrase must be found in each search result's content in the exact order as they appear in the phrase. For example, entering the phrase:

"dog food"

in the Primary Terms field will require that 'dog' and 'food' must be found in each search result's content together. Notice that the phrase is "surrounded" by double quotes; indicating a phrase using any other characters (like 'single quotes') will not work.

Double quotes can also be used to search for words or symbols that are otherwise reserved for special purposes within the search terms fields' syntax. For example, asterisks are used to weight keywords (a feature that is discussed later in this Appendix), and are therefore not normally allowed as a regular search term. However, you can use double quotes to indicate that you wish to have the asterisk used as part of a search term:

"wild*star"

Without the quotes, that search term would cause an error in Gemini UDS because of the asterisk. The other characters and words normally reserved by the search term syntax may be used as normal parts of search terms in this way.

We can enhance our example query by making a couple of the existing keywords into a phrase. We want to find information about multiple sclerosis, but there are other kinds of sclerosis (tuberous, primary lateral, etc.), and we don't want to find pages that talk about those disorders that just happen to also contain our 'multiple' keyword. So, we should make 'multiple sclerosis' into a phrase that will be required in the content of all of our results:

```
"multiple sclerosis" treatment
```

This will essentially guarantee that we do not get results that are actually about other kinds of sclerosis disorders.

It should be noted that not all search services support the creation of phrases; in these cases, Gemini UDS accesses the search service(s) in question using a version of your search terms without the double quotes. However, all analysis done by Gemini UDS conforms to the purpose and spirit of the quoted search terms.

Boolean Expressions

Within the context of Gemini UDS, boolean expressions are a way of describing what should, should not, and what may be found in the course of a search (boolean expressions and boolean algebra have much wider connotations in other contexts that we won't discuss here). This opens up a much broader range of capabilities available to you in forming queries because you are no longer restricted to defining only what *must* be found. For assistance in forming boolean expressions, please see page 13 of this user's guide.

To illustrate the contrast concretely, let's consider an example. Suppose you are looking for information on dog and cat food. Well, you can't simply type in "dog food" and get good information on cat food, nor can you type in "cat food" and get information about dog food. What you need is a way of describing both; the boolean OR operand does that:

dog OR cat food

This query will return information about both kinds of food because Gemini UDS will require only one of the terms separated by the OR operand, so both sets of information are considered relevant. Gemini UDS currently support three boolean operands, AND, OR, and NOT. Here are some generic examples for each one and how they affect what is and is not relevant:

keyword1 AND *keyword2* – both keywords must be present in the content for it to be relevant
keyword1 OR *keyword2* – either one or both of the keywords must be present in the content for it to be relevant
NOT *keyword2* – *keyword2* may not be present in the content for it to be relevant

Important: Unless another boolean operand is present, the Primary and Preferred Terms fields assume internally that there is an AND operand between each of your search terms. Therefore, these two queries are equivalent:

dog food Doberman "basset hound" dog AND food AND Doberman AND "basset hound"

This has the effect of requiring all of the search terms you provide, unless you specify otherwise.

You may also use parentheses to group search terms so that boolean operands apply to a set of terms instead of just one. For example:

(Doberman "basset hound") OR (dog food)

This query will return pages that contain either 'Doberman' and '"basset hound"', or 'dog' and 'food', or both sets of terms. You can nest parenthetical boolean expressions arbitrarily deep, and you may use boolean expressions in conjunction with phrases, regular expressions, and weighted keywords.

It should be noted that not all search services support boolean operands, and some of those that do do not support a full range of them (many do not recognize NOT, for example). Because of this, it is not guaranteed that all of your boolean operands will be sent to your selected search services; in cases where Gemini UDS knows that a search service does not support a certain operand, it makes its best effort to send a query to that search service that is most like the spirit of your "real" query. Regardless of support for boolean expressions by search services, Gemini UDS will always analyze potential search results using any boolean expressions you provide in the most rigorous way possible.

So, back to our main example. Multiple Sclerosis is often referred to using the acronym MS or M.S. It would be great if our query could reflect that, so that pages that use only the acronym would be considered relevant. This is easy with boolean operands:

```
("multiple sclerosis" OR MS OR M.S.) treatment
```

Weighted Keywords

Gemini UDS provides a very unique syntactic element, that of keyword weights. By default, every search term you provide to Gemini UDS has a weight of 1 associated with it. By changing this weight, you can cause Gemini UDS to give higher precedence to the differently-weighted search terms, which has a corresponding affect on how results that contain those weighted search terms are rated.

Weighting a search term is done by appending an asterisk (*) to the term, directly followed by an integer between 1-99. Any spaces between the term, the asterisk, or the integer will result in an error, or undesired results. Here's a simple example:

dog*5 cat

In the process of running this query, Gemini UDS will give a much higher rating to pages that contain greater instances of the 'dog' term than pages that contain an equal number of instances of the 'cat' term. You can think of weighting as a sort of multiplicative factor: if a page was given 5 "points" for the number of 'dog' instances it contained, and 10 "points" for the number of 'cat' instances it contained, its rating would be 35 by virtue of the '*5' weighting being applied to the 5 points received for 'dog'. (This point system is highly simplified, but it suffices for the example.)

These weightings are relative to each other within the same query, and can be applied to essentially any search term, including phrases and boolean expressions. So, this is a perfectly valid (and useful) sort of query:

(Denver football)*2 OR "John Elway"*5

What this is saying to Gemini UDS because of the relative nature of the ratings is: "find information about either 'Denver football' or John Elway, but I prefer to see pages about John Elway first". You should also notice that the '*2' weighting is applied to all of the terms enclosed by parentheses; the query below is therefore equivalent to the one above:

(Denver*2 football*2) OR "John Elway"*5

If you had a mathematical background, you might say that weights are distributive when used with a set of search terms grouped with parentheses.

It should be noted that no search services support weights like this, so Gemini UDS does not send the weights you assign to search terms to the search services you select.

Back to our main example. Suppose we are primarily looking for treatment options for multiple sclerosis, and therefore want pages containing such information rated higher than those that contain less about treatment. This can be done easily with a weighting that will favor results that contain greater instances of the 'treatment' term:

```
("multiple sclerosis" OR MS OR M.S.) treatment*4
```

Regular Expressions

Gemini UDS accepts one additional element in the Primary and Preferred Terms fields: Perl5-compatible regular expressions. We won't go into how to form regular expressions here; vast tomes have been written about using regular expressions (otherwise known as *regexes* or *RE*'s) and applying their power and versatility in finding patterns within text. What we will cover is how to add regular expressions to your Gemini UDS query.

(**Note for advanced users**: if you use regular expressions in Gemini UDS much, you'll probably want to know that it uses the Apache ORO package to process them <http://jakarta.apache.org/oro/index.html>. As with every regular expression package, the ORO package has a couple of peculiarities that you should be aware of if you're working with sophisticated regexes.)

You can provide a regex wherever you might provide another kind of search term. The only difference is that regular expressions **must** be surrounded by forward slashes in order to let Gemini UDS know what they are. For example:

```
/d.g food/
```

will return results where the results' content contains the pattern indicated by the regex between the forward slashes (in this case a 'd' followed by any other character followed by 'g food').

Virtually no search services support regular expressions, so Gemini UDS does not attempt to send them to your selected search services. Therefore, you must have some other kind of search terms in the Primary Terms field (or in the "Other Keywords" field in the advanced query interface) if you wish to put regexes in the Primary Terms field (and therefore require that your results contain matches to the regex' pattern(s)).

At this time, the number of matches to a regex pattern within a certain potential search result's content does not affect that result's relevance rating. Also note that the search-and-replace functions available in many regular expression packages are essentially meaningless within Gemini UDS – it makes no sense to try to evaluate an expression that by its nature alters the content of search results.

Finally, let's add a regular expression to our main example query. Instead of the psychological or concrete physical aspects of multiple sclerosis, suppose we wish to see information about its neurological facets. The problem with 'neurological' is it might not be in the content of pages that contain terms like 'neurology', 'neurologist', or 'neuro-opthamology', information we would probably want to see. Regular expressions are perfect for this:

```
("multiple sclerosis" OR MS OR M.S.) treatment*4 /neuro.* /
```

The regular expression in the above query will require that all results contain *some* term that starts with 'neuro' and has a space after it. Using this, we'll get pages that are about the neurological aspects of multiple sclerosis, but we won't have to specify what sort of neurological aspects – they will all be included.

Finally, it is recommended that you use the "other keywords" feature in the advanced query interface (see page 18) if you are going to use regexes in the Primary Terms field. This is because (as was mentioned earlier) regular expressions are not sent to search services due to a lack of support for them. So if we used the above example query in the Primary Terms field, it is possible that we would not get back *any* results because the search services we selected didn't know about our 'neuro' requirement. Putting "multiple sclerosis neurology treatments" in the "other keywords" field would remedy this problem.