Lotus Developer Domain
The Technical Resource for Lotus software
LDD Today

Level: Intermediate
Works with: Discovery Server
Updated: 04-Jun-2002

by
Wendi
Pohs

So you've heard all the hype about the Discovery Server, how it's a state-of-the-art tool that can keep track of all your documents and your personnel, regularly checking for updates and additions. You think that sounds really neat. And you have a real problem to solve in your organization: Your company has just acquired another company, across the pond, and you want to identify other people and groups who work like you do. The software is installed, and now it's time to start feeding it data. Where do you start?

This article offers advice on how to select and configure various types of data repositories for the Discovery Server. We'll take you from the high-level analysis to recommendations about how to fill out the Discovery Server's Data Repository and Field Map forms. This article assumes familiarity with Discovery Server and how it works. For more information about Discovery Server, see the *LDD Today* article **A preview of Lotus Discovery Server 2.0**. For information about K-maps, see the *LDD Today* article **Creating meaningful K-map taxonomies**.

Selecting repositories

Before you select data repositories for your K-map, you should consider the kinds of questions your users are likely to ask. For example, if an organization builds a K-map for its support employees, it might combine product documentation databases with information in a call tracking system so that the employees can see both procedural and troubleshooting information at the same time. You might decide to combine technical project tracking databases with marketing and competitive information, so that technical employees can have direct access to information about the market or about the competition. But no matter what problem you decide to solve, it's good practice to answer the following questions before you start.

What type of content is appropriate to use to create a K-map?
Sources with good metadata and lots of text work best. Talk to the people who know the content in their area; they'll know which databases contain the best information. You may not even want to spider all the databases in your organization. If you're using the Discovery Server to find people with similar affinities, for example, you might only want to spider content that is rich in technical project detail. You may not need to spider repositories that only contain project schedules.

Which data types work best?
While it may be tempting to use many data repositories at first, we've found that a bit of careful analysis up front often saves time later. Internet Web sites, for example, can contain useful information, but they can also contain pages of links to other sites or lots of advertising. It's good practice to determine the signal-to-noise ratio for the sites you decide to spider. If there's more rich content than advertising or links on a Web site, then that Web site is a good candidate to spider. You also want to be sure that the Web site contains good metadata.

The Discovery Server supports several data repository types. Here they are in the order we've found them to be most useful:

© Copyright IBM

1. Generic Notes databases with rich metadata (such as databases with Title, Author, and Category fields) and large, rich text Body fields
2. Lotus Domino.doc databases
3. Lotus QuickPlaces
4. Windows-compatible file system files with rich metadata (If you use a Microsoft Word file, for example, make sure that the properties fields contain up-to-date Author and Title information.)
5. Web sites with good metadata and lots of text (If in doubt about a site's metadata, open a representative page on the site, right-click the page, and choose View Source, then look for Author or Title tags.)

How should the content be organized before you spider?
Often you will not want to spider all the information in any one data repository. Some of it may be out of date, or it may contain administrative information, or it may contain documents about topics that aren't pertinent to your users. The Discovery Server lets you take advantage of existing methods to create the subsets of data you need. In a Notes database, for example, you can decide to spider only certain views or folders. But more on that later.

Are existing fields, categories, and keywords useful?
The Discovery Server lets you create Field Maps to take advantage of existing fields, categories, and keywords in your data repositories. A document's author may be called an Author in one database and Owner or Creator in another database. When you use Field Maps, you tell the Discovery Server which fields denote authorship in the data repositories you select.

Examine your data repositories and check the names of the author, title, and category fields before you start. Sometimes database managers define overview or abstract fields. You can map these fields to the Summary field in the Field Map form. Database or Web site designers may already have designated Category or Topic fields in the repositories you spider. Often these fields contain the names of predefined categories. You can map these fields to be Keywords in a Field Map form.

In a Notes database you may have documents that contain more than one Body field. You use the Field Map forms to indicate that the spider should consider all these Body fields together.

You should take advantage of all the existing information you can when you fill out a Field Map form. And don't worry, you can modify the Field Map form later if you discover a field you may have missed.

Getting started

Start by selecting three or four repositories that are rich in text and metadata fields. Notes discussion databases are good repositories to start with because they often combine good metadata (Subject, Author, and Category fields) and large rich text Body fields. They also can contain views or selection formulas that the Discovery Server can reuse to organize content.

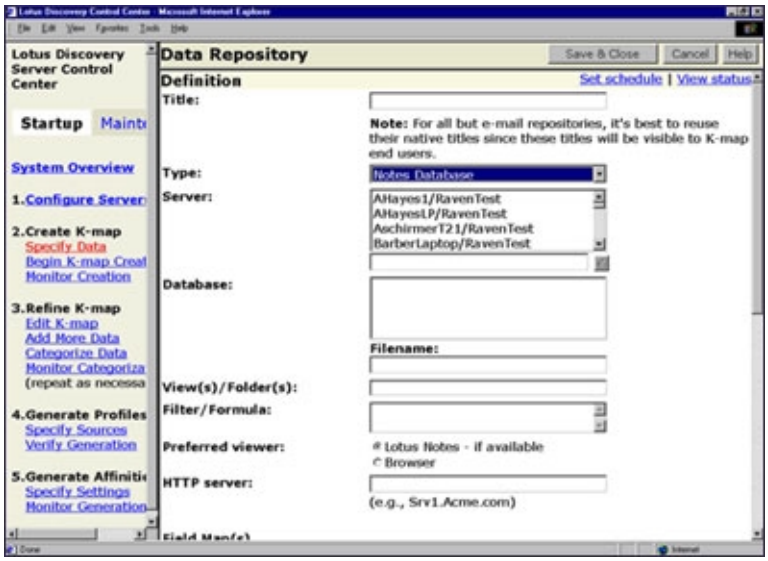
Use the Notes client to look at possible candidate databases before you spider. Make sure your server can access the databases and individual documents you need. Spiders collect access control lists (ACLs) for both repository and document-level access (if document access restrictions exist). The servers running the Notes spiders need Reader access to the repositories to spider and collect ACL information. Here are a couple of other caveats:

- If a document has a reader list that does not allow spider access, the document will not be spidered.
- If the repository's server is in a foreign Domino domain, you'll have to enter that server's name manually in the Server Name field on the Data Repository form, assuming cross-certification and connection records are already in place to allow connections to the foreign domain.
- Think about how your users access these documents today. Some organizations use the Notes client exclusively while others access Notes databases through a browser. You select a preferred viewer for each data repository you define.

Filling out the Data Repository form
You use Data Repository forms to specify which repositories you want the Discovery Server to spider. After you complete the form, you can also consult it to monitor a repository's status.

To open the Data Repository form, go to the Startup menu and click Specify Data. Then click Add Data Repository to see a blank form. Here's the Data Repository form for a Domino database:

© Copyright IBM



The Data Repository forms are customized for each data repository type, as follows:

For Domino databases:

- The View(s)/Folder(s) option allows you to limit the amount of data from the repository that gets spidered. You select the view or folders that contain the information you need.
- The Filter/Formula option allows you to enter a formula or filter to apply to either the entire repository or to the view(s)/folder(s) specified above. The format of the filter/formula must be in Notes formula syntax when the repository type is Notes. This field is editable after the repository has been queued for spidering (although edits once it's queued will not affect the current run). Check the "Process all documents during next run" box to apply the edits to the entire repository.

For Domino.doc:

- The Server option allows you to choose a server in the Server list box or type a server name in the text box and click the checkmark button. Doing this populates a new Library list box with only the Domino.Doc library databases on the chosen server (based on the list in the ddadmin.nsf database on that server).
- The Library option allows you to choose a library in the Library list box or type a library name in the text box and click the checkmark button. Doing this populates the File Cabinet list box with only the cabinet files associated with the selected library.
- The File Cabinet option displays all File Cabinets associated with the specified Library. Selecting one displays the filename in the Filename text box.
- The Versions option includes the "Latest versions of documents only/Include all versions of documents" radio button options. "Include latest versions of documents only" is the default.

For QuickPlace:

- The Server option allows you to choose a server in the Server list box or type a server name in the text box and click the checkmark button. Doing this populates the QuickPlace list box with only the main.nsf files from the QuickPlace directories on the selected server.
- The QuickPlace option displays all main.nsf files from the QuickPlace directories on the specified server, except for those found in the default QuickPlace, Help, and Tutorial directories. Selecting one displays the filename in the Filename text box.

For Web repositories:

- The Spider linked pages option allows you to determine how much of a Web site you want to spider. You may want to spider two levels, for example, if you know the pages on the site contain many links to other sites.
- The Follow links to other servers option is another way to create a subset of a Web site's data. You probably don't want to follow all the links on an Internet Web site.

For Windows compatible file systems:

- The Spider sub-directories option allows you to limit the subdirectories you spider. It's often a good idea to limit them at first so that you get an idea about the value of the data in the directories and subdirectories. Administrators are often surprised by the number of files in any one directory.

© Copyright IBM

- Although the Discovery Server can access files on other file systems, like Novell or Samba, it can't get ACL information from these systems. In these cases, we provide an override capability to allow you to spider content without ACL information. If the spider can't spider an unknown file system, you can select an ACL override on the Data Repository form.

You may decide you need to change some of these options once you see the data the spider collected. If you need to do this, most of these fields are editable on the Data Repository forms.

Checking it out

After you see that the spiders have finished processing, you can search for documents using the K-map user interface to be sure your data is ready to display to your end users. Administrators are often surprised by the title and author information that appears after an initial spider run. At one site, for example, we were happily startled to see a group of documents authored by "Another satisfied Notes user." While this was encouraging news, it wasn't accurate author information, so we went back to the database owner to ask for more information. It turned out that he had defined another field that automatically captured the author's name, so we decided to spider that field instead. We created a new Field Map to reflect our new understanding about this data repository and modified the Data Repository record to "Process all documents during next run."

Defining Field Maps

You can access the Field Map form while you're completing the Data Repository form. As we mentioned previously, this is the form you use to determine which fields you want to include in your K-map. You may decide not to include all the fields in the documents in your repositories, or you may decide to map several fields together. The $Global Field Map is our best guess at the range of field names that could be contained in most documents. You can also create additional Field Maps that can be customized to fit specific fields in the data repositories you spider.

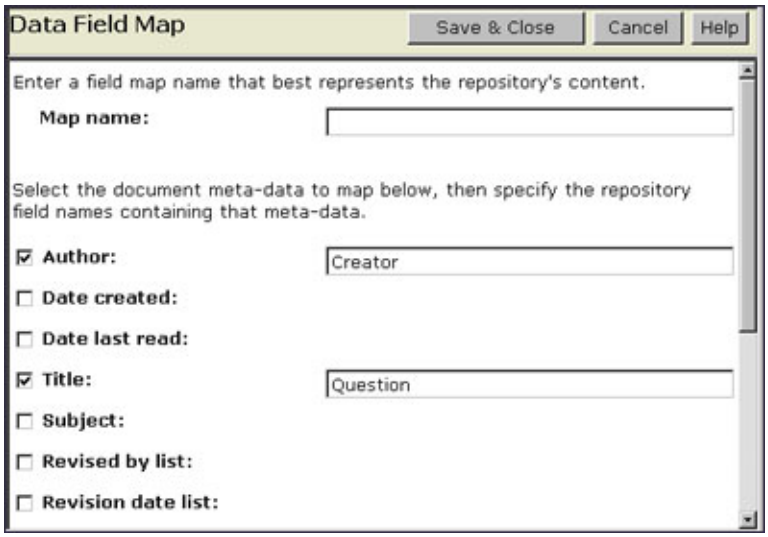
The $Global field map contains the following default values:

Metadata content types	Possible originating repository field names
Title	Title, PageTitles, h_name, PR_SUBJECT
Body	Body, PageBody, Mission, Main_Remarks, Response_Remarks, PR_BODY
Author	From, Author, DocAuthor, h_originator, h_alternatename, PR_SENDER_NAME
Date created	Created, CreateDate, DateComposed, TimeCreated, PR_CREATION,TIME
Revision date list	$Revisions
Revised by list	$UpdatedBy
Date last read	LastRead
Subject	NewLetter, Subject, DescriptionOfDocument, PR_SUBJECT
Keywords	Keywords, Categories
Summary	Summary
Directed to	DirectedTo, Reviewers, ReviewerList, People

We include a checkbox to the left of every field on the form. Checking or unchecking this box will display or hide the text boxes to the right, so you will only see the fields that are mapped for that repository. Hiding a text box clears the values, so re-enabling the mapping is easy. By default, in a new, customized Field Map form, all checkboxes will be unchecked. You can choose the fields that make the most sense for your data.

Here's an example of a Field Map for a repository that contains author information in a field named Creator and title information in a field named Question. Users will see values from these fields when they access document information through the K-map user interface.

© Copyright IBM



A stitch in time

As you can probably surmise from reading this article, a little up-front analysis can save you lots of time when you're selecting and specifying data repositories to spider. If you analyze the formats and existing organization of your data beforehand, you can use the Data Repository forms and Field Maps to get your best data into the Discovery Server fast. With a little planning, your end users won't have to wade through lots of irrelevant information before they find the content they really need.

ABOUT THE AUTHOR

Wendi Pohs is a principal taxonomy specialist on the Discovery Server team and the author of a book about knowledge management methodologies, *Practical Knowledge Management: The Lotus Knowledge Discovery System*, published by IBM Press. Wendi joined Lotus Development Corporation in 1996 and has worked on various projects as a spec writer, online help designer, and user assistance manager. Prior to joining Lotus, Wendi worked at the American Mathematical Society and at Digital Equipment Corporation. Wendi received her BA and MILS degrees from the University of Michigan.

© Copyright IBM