



Tabula DX

The Search Engine for PDF Files

Reference Guide Version 1.12
July 2008

© Copyright 2008 Aquaforest Limited

<http://www.aquaforest.com/>

CONTENTS

1	INTRODUCTION	2
2	INSTALLATION AND INITIAL CONFIGURATION	3
2.1	SYSTEM REQUIREMENTS	3
2.2	INSTALLING TABULA DX.....	3
2.3	TESTING THE INSTALLATION.....	3
2.4	TRIAL LICENSE RESTRICTIONS	4
2.5	UNINSTALLING THE PRODUCT	4
2.6	THE SAMPLE / DEMO COLLECTIONS	4
2.7	USING TABULA DX WITH IIS.....	6
2.7.1	<i>Setting Up Tabula DX with IIS 5 (Windows XP)</i>	6
2.7.2	<i>Setting Up Tabula DX with IIS 6 (Windows 2003)</i>	10
2.7.3	<i>Using Tabula DX with IIS 7 (Windows Vista, Windows 2008)</i>	13
3	SEARCH QUERY EXPRESSIONS	15
3.1	SEARCH FIELDS	15
3.2	QUERY EXPRESSIONS.....	15
4	CONFIGURING SEARCH COLLECTIONS	17
4.1	SYSTEM-WIDE SETTINGS	17
4.2	COLLECTION SETTINGS.....	18
4.3	COLLECTION ATTRIBUTES.....	18
4.4	CONFIGURING COLLECTION INDEXING	20
5	COMMAND LINE INDEXING	21
6	CUSTOMIZATION AND INTEGRATION	22
6.1	THE SEARCH URL PARAMETERS	22
6.2	CUSTOMIZING THE SEARCH INTERFACE.....	22
6.3	XML OUTPUT	22
6.4	WEB.CONFIG PARAMETERS	24
7	TABULA DX DIRECTORIES	25
8	TABULA DX AND LUCENE	26
9	PRODUCT VERSION HISTORY	27
9.1	VERSION 1.12.....	27
9.2	VERSION 1.01.....	27
10	ACKNOWLEDGEMENTS	27

1 INTRODUCTION

Designed specifically for search-enabling large collections of PDF files via a browser interface, Tabula DX offers the following benefits and features :

Complete PDF Searchability - Search on PDF bookmarks, annotations, and metadata including XMP with no limit on the number of PDF pages that will be indexed.

Ease of Use - Present users with a familiar search interface and document thumbnails.

Performance and Scalability - Tabula DX is based on the Lucene search API which has been proven to robustly support collections of millions of documents.

Customizable - Simple user interface customization via XSL.

Integration Support - Search results can be returned as pure XML from any web-based method.

Designed for IIS and ASP.Net - Tabula DX is built using C# and ASP.Net for simple integration into a Microsoft-based environment.

Simple License Model - Tabula DX is licensed per server and has no limits on the number of documents that can be indexed.

Lucene Compatible - The Tabula DX Lucene indexes are compatible with tools supporting Lucene 1.4 or later.

Simple Web-Based Administration - The administration module allows creation of PDF collections, settings and index scheduling.

2 INSTALLATION AND INITIAL CONFIGURATION

2.1 System Requirements

- Windows XP, Windows 2003, Vista, Windows 2008
- Version 2.0 of the .NET Framework
- Note : To run batch indexing jobs via the web interface a suitable user id and password will be required, This can be set using the Settings section of the Tabula DX web interface.
- Web Server : Tabula DX includes the lightweight UltiDev Cassini web server and the product is initially configured to use this. For production use IIS is recommended as a web server. See section x.x below for configuration details.

2.2 Installing Tabula DX

The source media zip file contains a setup.exe which will install Tabula DX along with the Cassini lightweight web server. Please contact support@aquaforest.com should you require any assistance.

2.3 Testing The Installation

To get started with Tabula DX, access the Tabula DX shortcut that is installed on the desktop and under the Programs menu.

© 2008 Aquaforest Limited
[Contact us](#)

Administration

- [Collections](#)
- [New Collection](#)
- [Settings](#)
- [Installation Test](#)
- [Documentation](#)

Collections

ID	Collection Name	Status	Last Updated	Index Log	Documents
1001	Sample PDF Collection	Indexed	21-Feb-2008 18:05:23	Last Log	20 Search
1002	Collection 1002	Indexed	07-Mar-2008 13:10:27	Last Log	50 Search

The main Tabula DX administration page



Installation Test Page

2.4 Trial License Restrictions

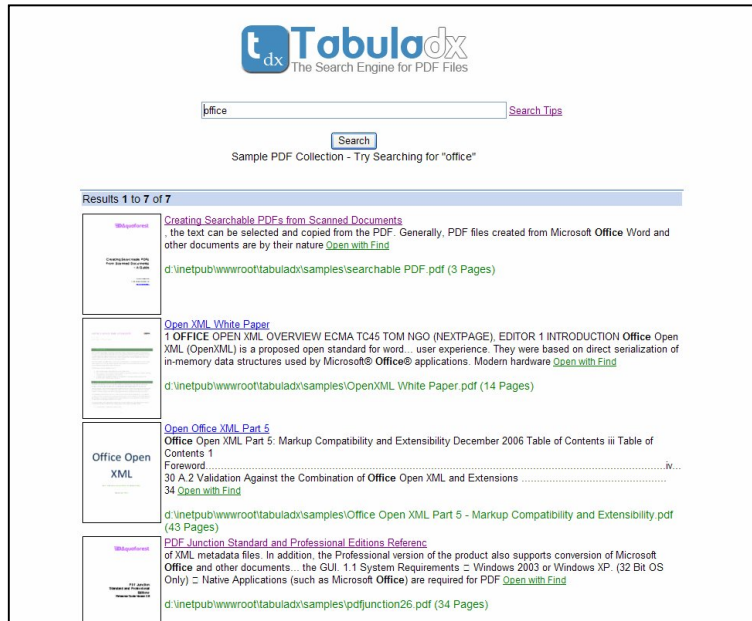
The trial license does not expire but limits the size of document collections to 1,000 documents. Please contact sales@aquaforest.com for additional assistance with trial licensing.

2.5 Uninstalling The Product

Tabula DX can be uninstalled via Windows Add / Remove Programs. UltiDev Cassini Web Server Explorer and UltiDev Cassini Web Server for ASP.Net 2.0 can be removed in the same way.

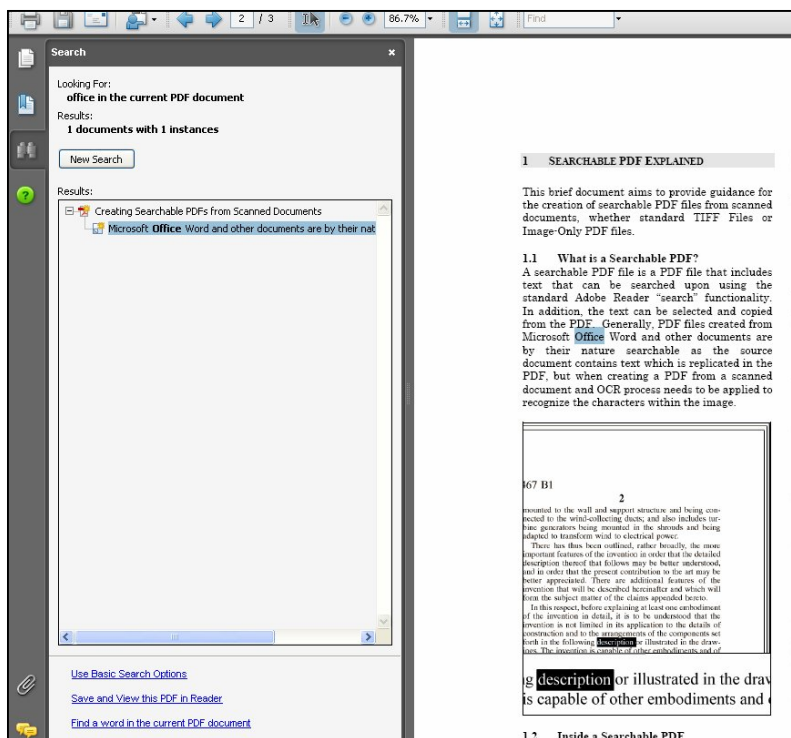
2.6 The Sample / Demo Collections

The product comes installed with a small demonstration collection of around 20 documents which can be searched as soon as the product has been installed.



Demo Collection – Sample Search Results

The PDF file can be opened in a new browser window by clicking either the document thumbnail or the title link. In addition clicking on “Open with Find” will open the PDF document with the search string passed to Adobe Reader. This will automatically open the find interface within Adobe Reader using the same query parameters. See below for an example.



Opening a PDF Document “With Find”

2.7 Using Tabula DX with IIS

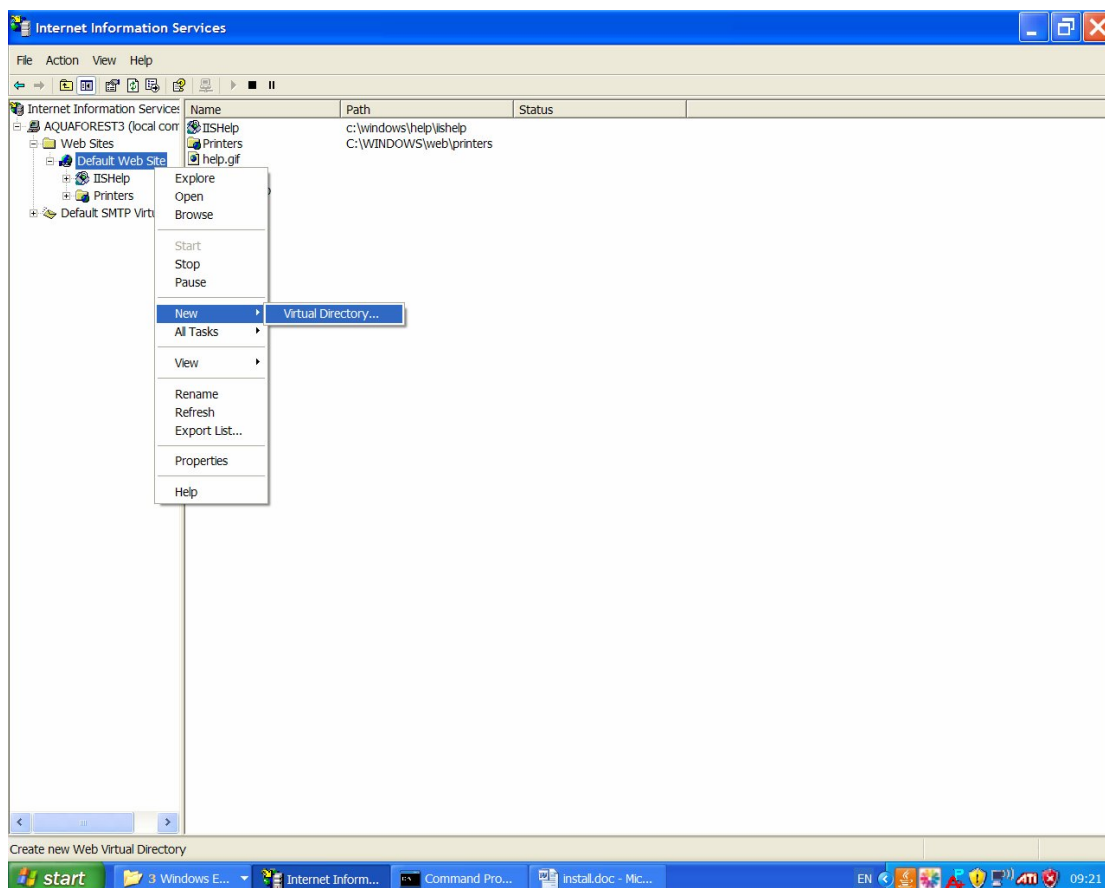
Tabula DX is supported under IIS 5 (Windows XP), IIS 6 (Windows 2003) and IIS 7 (Windows Vista, Windows 2008).


Please note the following important system requirements :


- Tabula DX needs to be running under ASP.Net 2.0
- Tabula DX does require certain file system privileges which are unlikely to be satisfied by using the default IUSR account. Therefore the Tabula DX web application should be run using Integrated Authentication or with an anonymous user configured with sufficient privilege.
- Running under IIS7 / Windows Vista will require use of the Classic ASP.Net application pool.

2.7.1 Setting Up Tabula DX with IIS 5 (Windows XP)

Create a new virtual directory called tabuladx that points to the Tabula DX install location (by default C:\Program Files\Aquaforest\Tabula DX) with a default document of main.aspx and either integrated authentication or anonymous authentication with a suitably privileged user. The screen shots below illustrate the process.




Virtual Directory Creation Wizard 


Virtual Directory Alias 

You must give the virtual directory a short name, or alias, for quick reference.

Type the alias you want to use to gain access to this Web virtual directory. Use the same naming conventions that you would for naming a directory.

Alias:

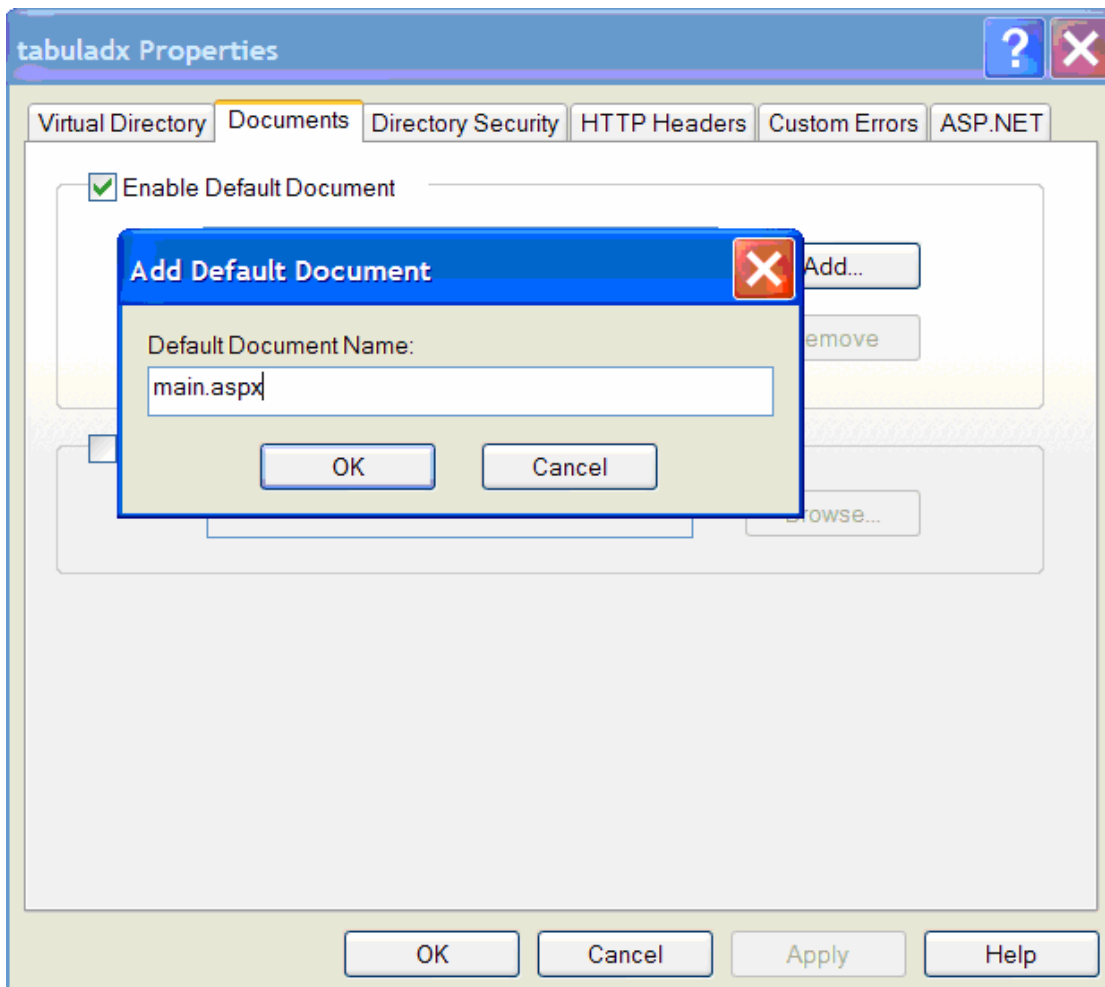
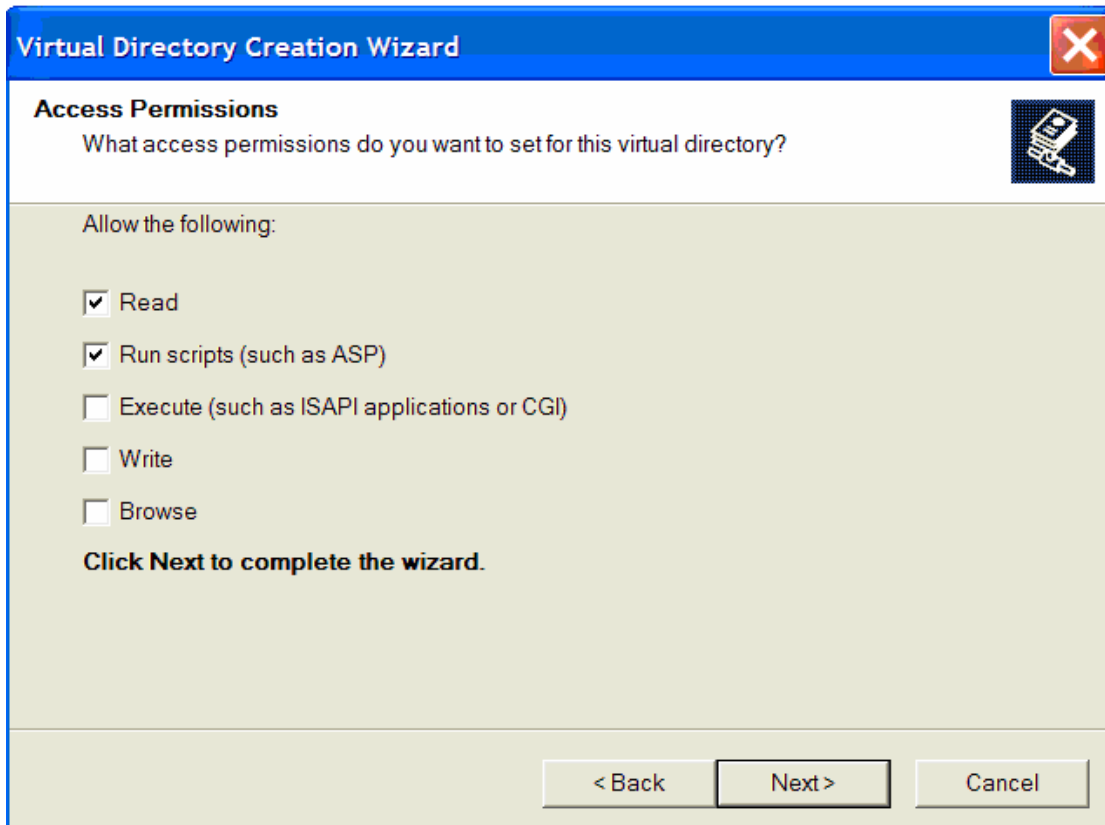
Virtual Directory Creation Wizard 

Web Site Content Directory 

Where is the content you want to publish on the Web site?

Enter the path to the directory that contains the content.

Directory:



Authentication Methods ✕

Anonymous access
 No user name/password required to access this resource.
 Account used for anonymous access:
 User name:
 Password:
 Allow IIS to control password

Authenticated access
 For the following authentication methods, user name and password are required when
 - anonymous access is disabled, or
 - access is restricted using NTFS access control lists

Digest authentication for Windows domain servers
 Basic authentication (password is sent in clear text)
 Default domain:
 Realm:
 Integrated Windows authentication

Tabula DX administration can then be accessed via <http://server/tabuladx> :

© 2008 Aquaforest Limited
[Contact us](#)



The Search Engine for PDF Files

Administration

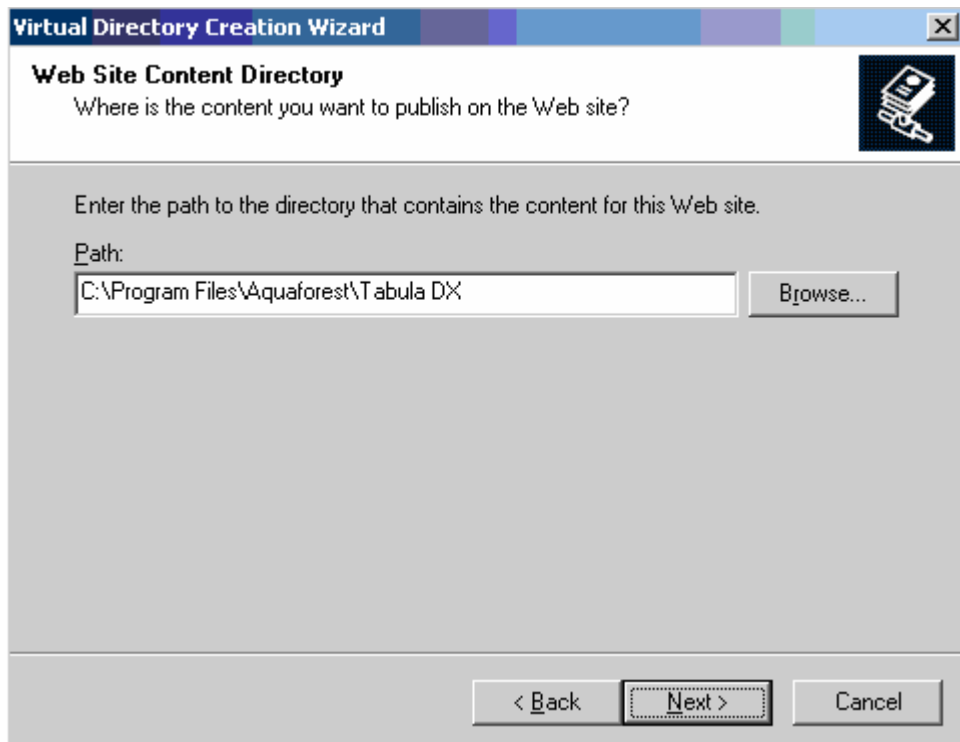
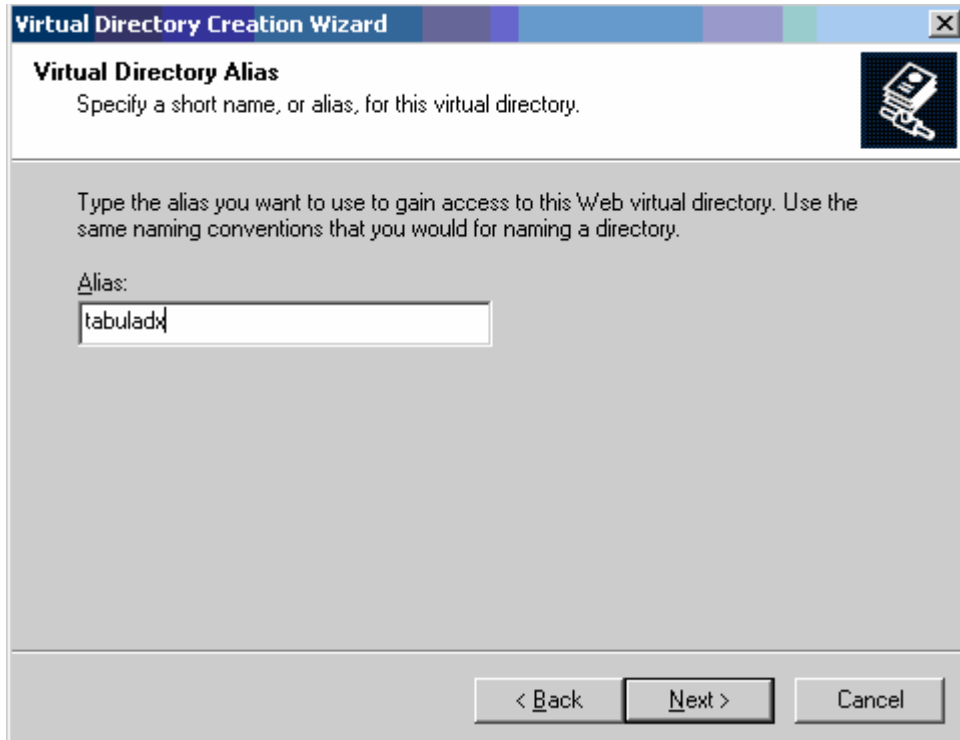
-  Collections
-  New Collection
-  Settings
-  Installation Test
-  Documentation

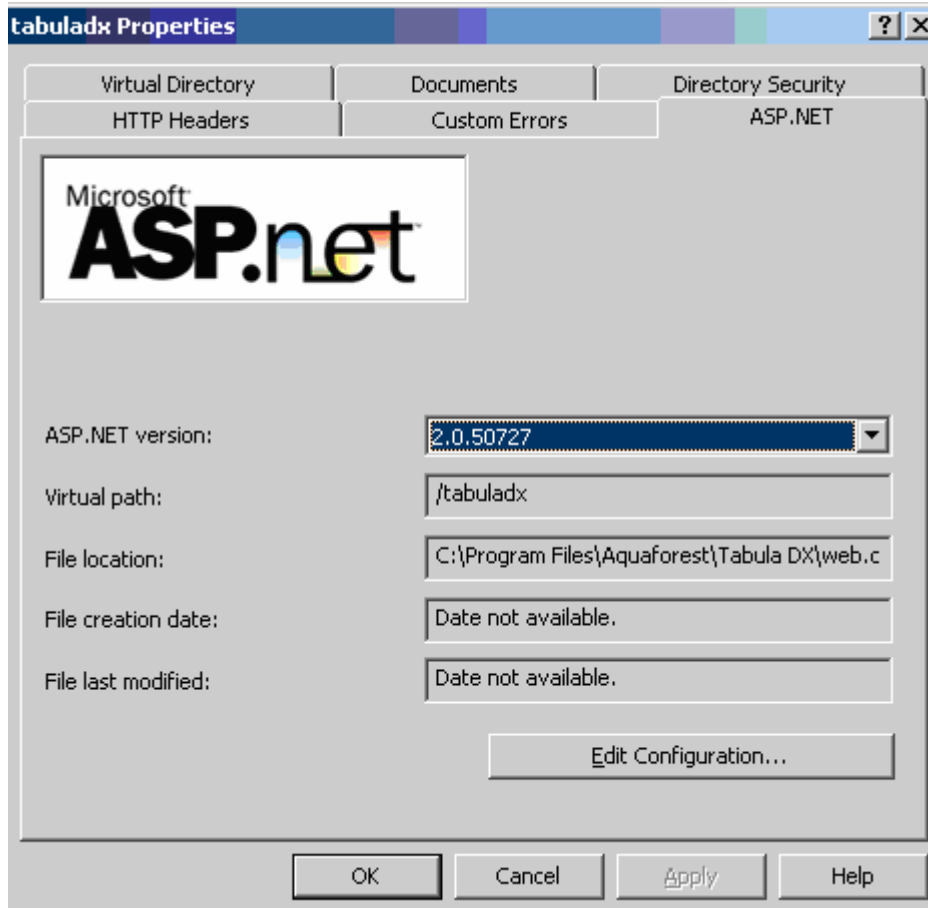
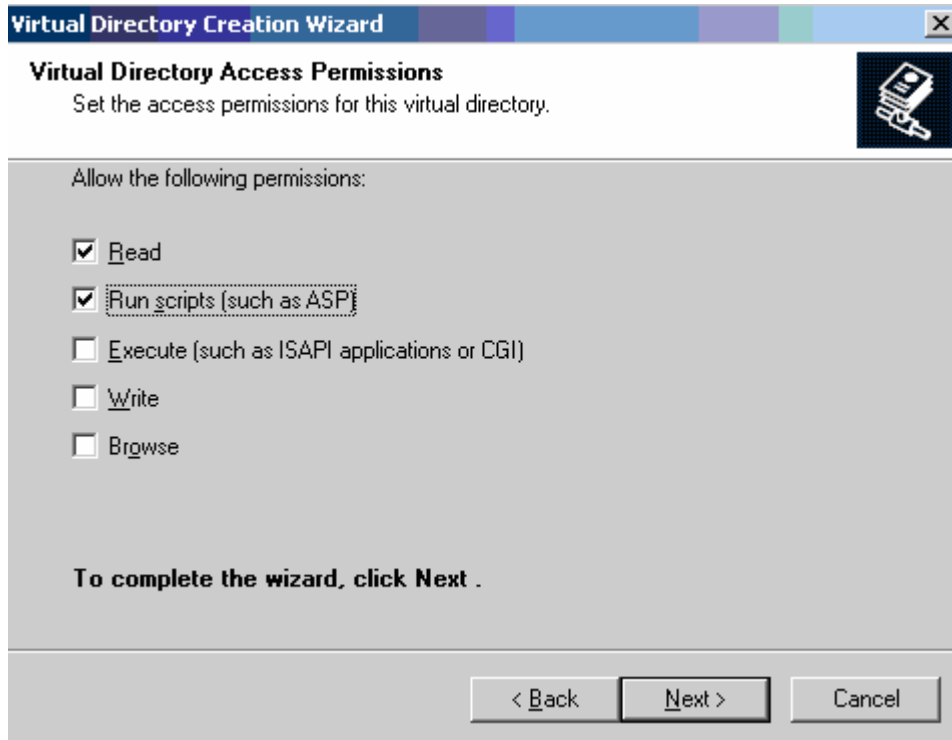
Collections

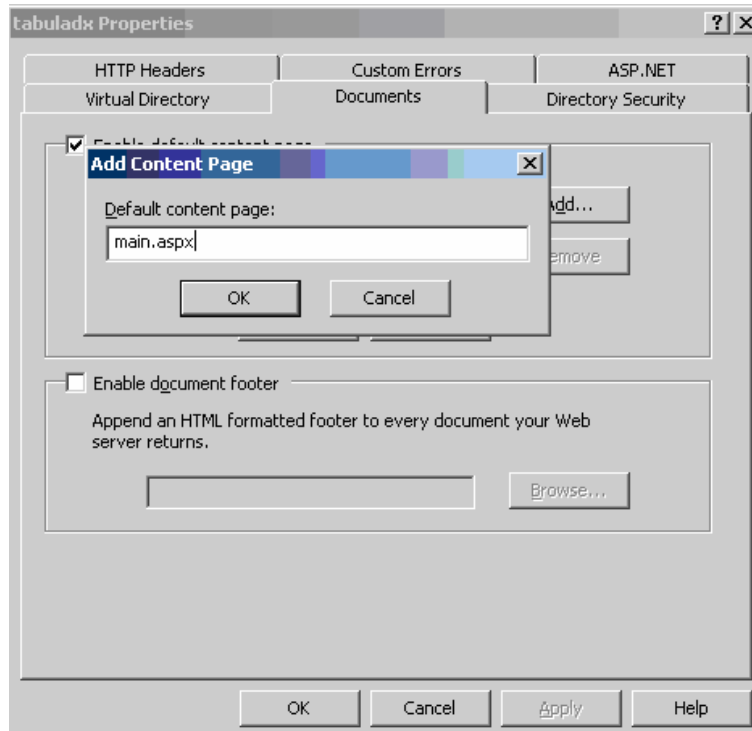
ID	Collection Name	Status	Last Updated	Index Log	Documents
1001	Sample PDF Collection	Indexed	21-Feb-2008 18:05:23	 Last Log	20  Search
1002	Collection 1002	Indexed	07-Mar-2008 13:10:27	 Last Log	50  Search

2.7.2 Setting Up Tabula DX with IIS 6 (Windows 2003)

Create a new virtual directory called tabuladx that points to the Tabula DX install location (by default C:\Program Files\Aquaforest\Tabula DX) with a default document of main.aspx and either integrated authentication or anonymous authentication with a suitably privileged user. The screen shots below illustrate the process.



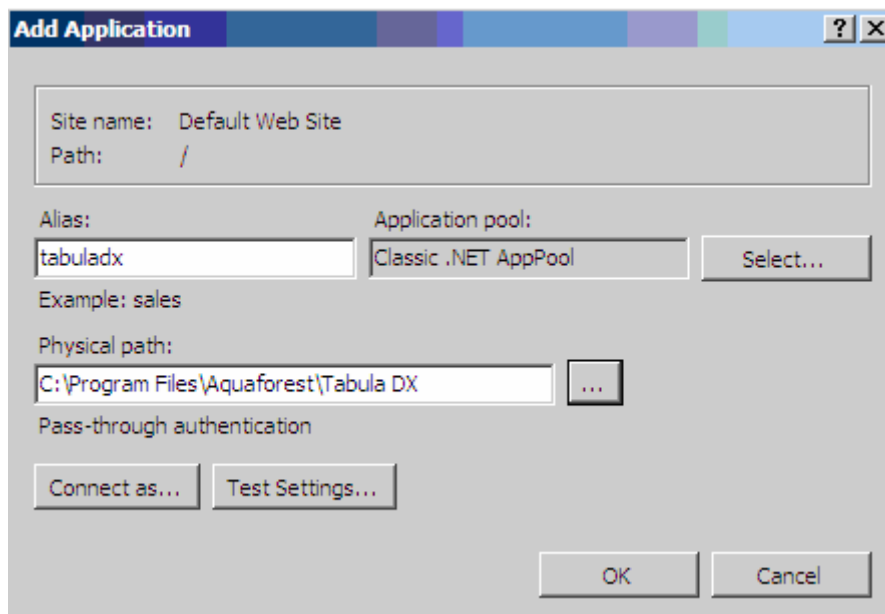
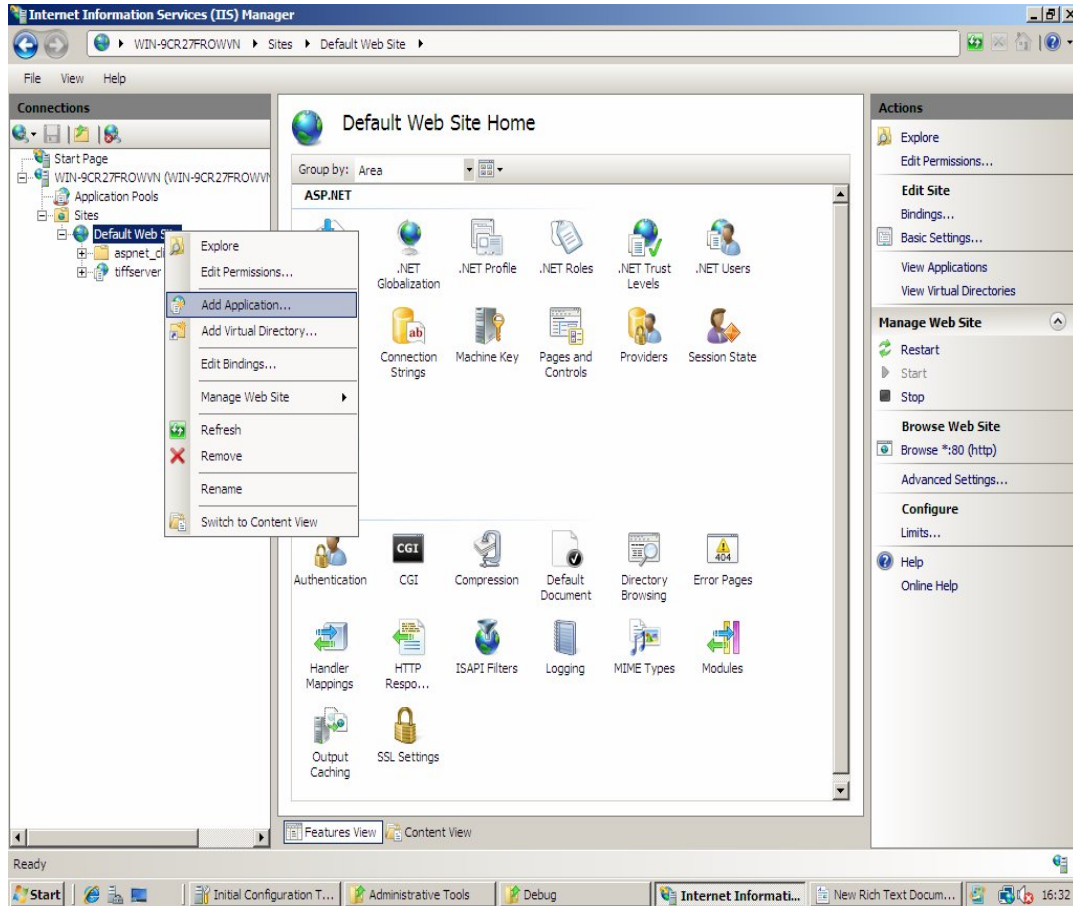




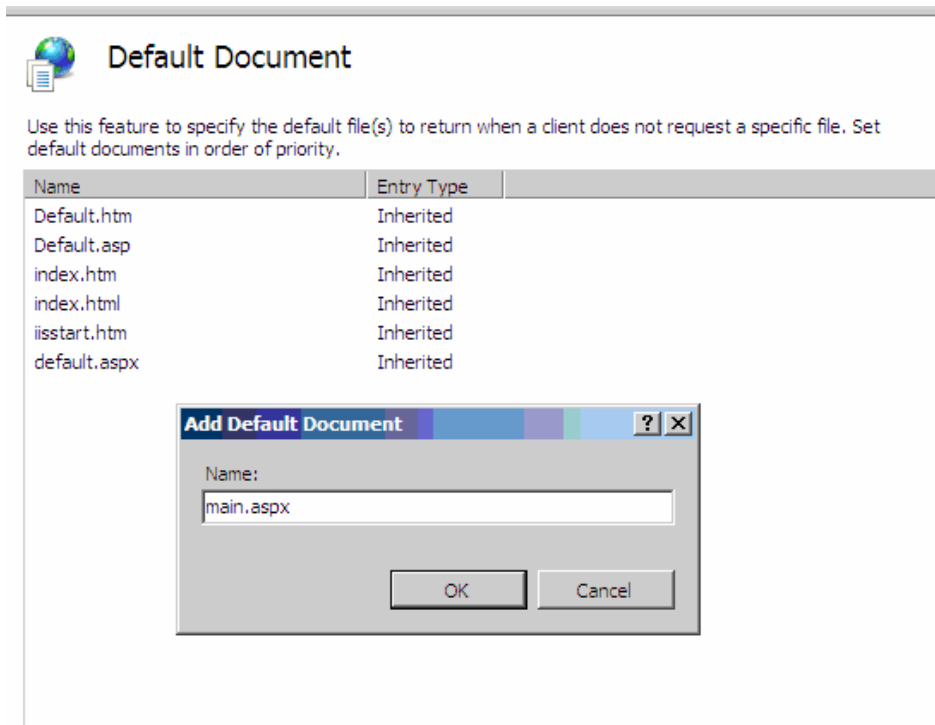
Tabula DX administration can then be accessed via <http://server/tabuladx> :

2.7.3 Using Tabula DX with IIS 7 (Windows Vista, Windows 2008)

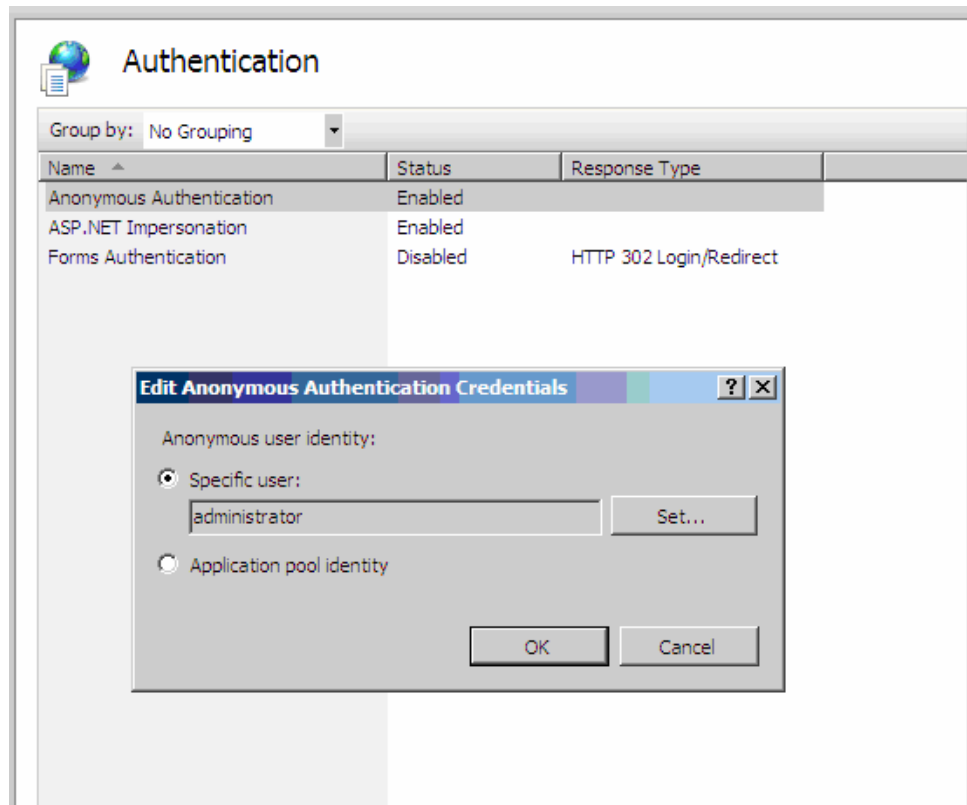
Create a new web application directory called tabuladx that points to the Tabula DX install location (by default C:\Program Files\Aquaforest\Tabula DX) and uses the Classic ASP.Net AppPool. The screen shots below illustrate the process.



Set the default document to be main.aspx :



Ensure that the application runs with a suitably privileged identity. Not necessarily as administrator but with enough privilege to write to the Tabula DX collections folder.



Tabula DX administration can then be accessed via <http://server/tabuladx> :

3 SEARCH QUERY EXPRESSIONS

3.1 Search Fields

When documents are indexed, a number of different search index fields are populated depending upon the collection configuration.

Field	Contains
Contents	The text of the document. This is the default field.
Bookmarks	
Annotations	
<i>PDF Doc Info Fields (or XMP equivalents)</i> Title Author Subject Keywords Producer Creator Creationdate Moddate	The value of these fields can be set via the Acrobat Document Properties tab.
<i>xmpfield</i>	The value of the XMP metadata field configured with the search name <i>xmpfield</i> see section X.X for details of how this is configured.
Collectionid	The collection ID of the document.
Indextime	Time stamp in YYYYMMDDHHMMSS format
Path	PDF file path
Thumbnailpath	Path of the thumbnail image (or blank if thumbails are not used in the collection)
Pages	The number of pages in the PDF document
Filesize	The size in bytes of the file
Wincreated	Time stamp in YYYYMMDDHHMMSS format
Winmodified	Time stamp in YYYYMMDDHHMMSS format

3.2 Query Expressions

Search Query	Matches documents ...
Pdf	Contains the word “pdf” in the contents of the document.
Pdf search Pdf AND search +pdf +search	Each of these expressions will find documents with both of the words “pdf” and “search” in the contents of the document.
Pdf OR search	This will find documents with either (or both) of the words “pdf” and “search” in the contents of the document.
search AND NOT pdf	Documents that contain the term search but do not contain the term PDF.
Title:china AND –title:india	The title field contains the word china but not india


(pdf or wordperfect) AND search	Documents contain the word “search” and either “pdf” or “wordperfect”.
Title:”search engines”	The title field contains the phrase search engines
Search*	Contains terms that begin with search, such as searchable, searching and search.
Search~	Contains terms that are close to the word search such as Starch
winmodified:[20070701000000 TO 20070731235959]	Contains winmodified values in the range specified

4 CONFIGURING SEARCH COLLECTIONS

4.1 System-Wide Settings

The “Settings” tab allows a number of system-wide settings to be maintained. These settings are :

© 2008 Aquaforest Limited
[Contact us](#)



Administration

- [Collections](#)
- [New Collection](#)
- [Settings](#)**
- [Installation Test](#)
- [Documentation](#)

Settings


Version	1.01.80310.01
License Key	<input type="text" value="TABULADXTRIAL"/> <small style="color: red;">Tabula DX Trial License</small>
Administrator Password	<input type="password"/>
	<small style="color: red;">*To restrict access to the Tabula DX administration site.</small>
Index Username	<input type="text" value="administrator"/>
Index User Password	<input type="password" value="•••••"/>
	<small style="color: red;">*This username and password will be used by Windows Task Manager to run the scheduled indexing tasks.</small>

Attribute	Description
License Key	Once purchased, a permanent license key may be entered here.
Administrator Password	If a password is entered, then access to the Autobahn DX admin pages will require entry of the specified password.
Index Username and Password	To run indexing jobs via the web interface a suitable user id and password will be required which is used to create the job using Windows Scheduled Tasks.

4.2 Collection Settings

Individual collections can be configured by clicking on the collection name link in the main collections page :

© 2008 Aquaforest Limited
[Contact us](#)



Administration

- [Collections](#)
- [New Collection](#)
- [Settings](#)
- [Installation Test](#)
- [Documentation](#)

Edit Collection 1001

Collection Settings

Collection Name

Description

Indexing [Configure Collection Indexing](#)

Folder Paths

File Pattern

Index Folder

Index Batch Size

Results per Page

XSL File

Always Optimize

Check for Deleted

Thumbnails

Thumbnail Folder

Index Logging

Index Logging File

Index Doc Info

Index Bookmarks

Index Annotations

Index XMP

XMP Fields

4.3 Collection Attributes

Attribute	Description
Collection Name	The descriptive name of the document collection
Folder Paths	One or more folders containing documents to be indexed.
File pattern	A pattern to be used to match the files to be indexed. Default *.pdf
Index Folder	The folder to contain the index files.
Index Batch Size	The maximum number of documents that will be indexed in a single run of the indexer.

Results per Page	Default number of results per search results page.
XSL File	Default XSL file.
Always Optimize	If checked, the indexes will be optimized at the end of each run of the indexer.
Check for Deleted	If checked, the index process will check each file in the index. If the related PDF file has been deleted, the index entry will be removed.
Thumbnails	If checked a thumbnail image of the first page of each PDF document indexed will be produced.
Thumbnail Folder	Thumbnail images will be stored in this folder. Subfolders will be automatically created to mirror the structure of the source PDF folders.
Index Logging	If checked, a log file will be created (or appended to) detailing indexing activity each time the indexer is run on this collection.
Index Logging File	The index log file name. The file will be placed in AUTOBAHN/collections/ <i>collectionid</i> /indexlog. The string %TIMESTAMP% can be included in the file name to create a unique file for each indexer run; the timestamp is of the format YYYYMMDDHHMMSS.
Index Doc Info	If checked, PDF Doc Info metadata (title, author etc) will be indexed and can be searched using query expressions such as author:shakespeare
Index Bookmarks	If checked, the text of bookmarks will be indexed and can be searched using query expressions such as bookmarks:shakespeare
Index Annotations	If checked, PDF annotations will be indexed and can be searched using query expressions such as annotations:shakespeare
Index XMP	If checked, XMP metadata will be indexed in accordance with the “XMP Fields” instructions.
XMP Fields	<p>This defines which XMP fields should be indexed, and each line defines a property of the form :</p> <p>namespace:propertyname,indexname</p> <p>For example :</p> <p>http://www.aiim.org/pdfa/ns/id/:conformance,pdfaconformance</p> <p>The above instructs Tabula DX to index the property “conformance” in the PDF/A namespace (http://www.aiim.org/pdfa/ns/id/) and index it under the name pdfaconformance. The field may be searched to find pdf/a conformant files with a query such as pdfaconformance:B</p> <p>For further information about XMP refer to the Adobe resources here : http://www.adobe.com/devnet/xmp/pdfs/xmp_specification.pdf</p>

4.4 Configuring Collection Indexing



The screenshot shows the Tabuladx web interface for editing collection 1001. The page title is "Edit Collection 1001" and the sub-section is "Collection Indexing". On the left, there is an "Administration" sidebar with links for Collections, New Collection, Settings, Installation Test, and Documentation. The main content area includes a "Return to Configure Collection Settings" link, "Manual Indexing" buttons for "Index Now" and "Clear Index Now", and "Scheduled Indexing" options. The "Scheduled Indexing" section has a dropdown menu set to "None", an empty text input field, and a "Calendar" section with three dropdown menus: "0", "00", and "AM". A "Save Schedule" button is located below these options. The top right corner of the page contains the copyright notice "© 2008 Aquaforest Limited" and a "Contact us" link. The Tabuladx logo and tagline "The Search Engine for PDF Files" are in the top left.

The configuring collection indexing page allows a collection to be indexed according to a set schedule or immediately. The indexing process analyzes the current collection index and the set of file system files, determining which files need to be indexed or reindexed.

5 COMMAND LINE INDEXING

Tabula DX collections can be indexed by using the command line interface. The `tabuladx.exe` executable can be found in the product bin folder.

```
tabuladx.exe /id=collectionid /op=operation [/debug]
```

Parameter	Description
Id	The collection ID, eg 1002
Op	The operation : index – Index the collection. The indexing process analyzes the current collection index and the set of file system files, determining which files need to be indexed or reindexed. clear – Clear the index collection The executable can is also used to set up the demo collection, this should be performed automatically by the setup process. Setupdemo – Sets up the demo collection from the template. Adjustdemo – Adjust the collection for the local location
Debug	Optional. If specified verbose output is produced.

6 CUSTOMIZATION AND INTEGRATION

Tabula DX is designed to be customized and can be easily integrated within larger solutions.

6.1 The Search URL Parameters

Tabula DX search may be accessed via the search.aspx page

Eg:

<http://localhost/tabuladx/search.aspx?collectionid=1001&query=office&xslfile=xml>

Parameter	Description	Default Value if Unspecified
Collectionid	The numeric collection id to be searched.	N/A
Query	The query string	N/A
Resultfields	The list of fields to be returned	highlight,path,title,thumbnailpath,pages
Resultstart	The result set document number of the first document to be returned.	1
Resultsperpage	The maximum number of results to be returned.	From collection configuration.
Xslfile	Set to XML to have pure XML returned (see 5.3 below) or alternatively specify an alternate XSL file.	From collection configuration.
Thumbnails	True or false. If set to false thumbnails are not to be displayed.	From collection configuration.
Indexdirectory	Index files directory	From collection configuration.

6.2 Customizing the search interface

When running a search, Tabula DX generates an XML file with details of the search results. This is passed to the browser with default of the XSL file to be used to transform the output into the search results page. By default the style/results.xsl file is used. This can be customized to suit specific needs. An alternative replaced in the collection settings page

6.3 XML Output

```
<?xml version="1.0" encoding="utf-8"?>
<searchresults>
  <search>
    <collectionid>1001</collectionid>
    <query>document guidelines</query>
    <resultfields>highlight,path,title,thumbnailpath,pages</resultfields>
    <resultstart>1</resultstart>
    <resultend>9</resultend>
    <resultsperpage>10</resultsperpage>
    <sortorder />
    <thumbnails>>true</thumbnails>
    <indexdirectory>c:\dev\scree\index</indexdirectory>
    <xslfile />
  </search>
  <hits>9</hits>
  <searchstatus>success</searchstatus>
```

```

<results>
  <result>
    <row>1</row>
    <highlight>
      <B>Guidelines</B>for Creating Archival .....
    </highlight>
    <path>c:\docs2\PDFGuideline.pdf</path>
    <title>PDF Guideline for FDA</title>
    <thumbnailpath>c:\thumbnails\PDFG.pdf.1.gif</thumbnailpath>
    <pages>6</pages>
  </result>
  <result>
    <row>2</row>
  .....
  </result>
</results>
</searchresults>

```

Attribute	Description
<search>	This contains the search attributes
<collectionid>	The collection ID
<query>	The search query
<resultfields>	The fields that will be included in <results>
<resultstart>	The result number of the first document in <results>
<resultend>	The result number of the last document in <results>
<resultsperpage>	The maximum number of results
<thumbnails>	True or False (see <thumbnailpath> if true).
<indexdirectory>	The location of the collection index directory
<hits>	The total number of results from the search
<searchstatus>	Can be : success blank (if query is blank) nomorerresults error
<results>	The result document set
<result>	An individual result
<row>	The sequence in the result set
<field>	The contents of <i>field</i> for the document
<highlight>	Document fragment, highlighted with search terms
<path>	The path of the document
<title>	Document title. If the PDF document does not have a title, the first 60 characters of the document content is used.
<thumbnailpath>	The path of the thumbnail image
<pages>	The number of pages in the PDF document.

6.4 Web.config Parameters

The web.config file contains a number of appSettings that can be adjusted. These parameters are defined below.

Setting	Description
generatedTitleLength	For PDF documents with no title, Tabula DX generates a title for search results purposes using the first <i>generatedTitleLength</i> characters in the file. Initial value : 60.
highlightMaxDocBytesToAnalyze	When constructing the “highlight” text fragment for results display, this parameter determines how many characters of text will be examined. Initial value : 100000.
highlightNumFragments	Determines how many text fragments will be used in the highlight text. Initial value : 2.
highlightDelimiter	Specifies a character string to be used to separate the highlight text fragments. Initial value : “...”
highlightLength	Defines the length of highlight text. Initial value 150.
defaultOperator	Defines whether “AND” or “OR” is the default search operator. Initial value : “AND”.
resizeThumbnails	By default thumbnails are a width of 100 pixels. They can be resized on the fly if this parameter is set to true. Initial value false.
resizeThumbnailWidth	Defines the image width if <i>resizeThumbnails</i> is set to true.

7 TABULA DX DIRECTORIES

The Tabula DX folder structure is explained below. The root folder is typically `c:\inetpub\wwwroot\tabuladx`

Folder	Description
Bin	Contains the DLLs and executables, including <code>tabuladx.exe</code>
Collections	The root folder for the search collections.
Collections/9999	The root folder for the search collections 9999. Includes the <code>config_9999.xml</code> file which holds the collection configuration.
Collections/9999/index	The default location for the collection index files.
Collections/9999/indexlog	The default location for the collection index log files.
Collections/9999/temp	The location for the collection index log files.
Collections/9999/thumbnails	The default location for the collection thumbnail files.
Config	Contains the Tabula DX config file and the template collection template file.
Docs	Contains the reference guide and license file.
Img	Contains the images used in the web interface.
Samples	Contains the sample collection documents.
Style	Default location for XSL documents. Includes <code>results.xsl</code> .
Template	The template for the sample collection.

The indexes created by Tabula DX are compatible with Lucene 1.4 or later. More information regarding Lucene can be found here : <http://lucene.apache.org/java/docs/>

Included with the product is Luke – the Lucene Index Toolbox. This can be a useful tool for analyzing the contents of indexes and running queries. Luke can be launched by running `lukeall-0.7.1.jar` in the product bin folder. Luke is Licensed under the Apache License, Version 2.0 (the "License"); you may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>

Index name: **D:\inetpub\wwwroot\tabuladx\collections\1001\index**

Number of fields: **23**

Number of documents: **20**

Number of terms: **10883**

Has deletions?: **No**

Index version: **82**

Last modified: **Mon Mar 10 14:42:46 GMT 2008**

Directory implementation: **org.apache.lucene.store.FSDirectory**

Select fields from the list below, and press button to view top terms in these fields. No selection means all fields.

Available Fields:

- <>
- <annotations>
- <author>
- <bookmarks>
- <collectionid>
- <company>
- <contents>
- <creationdate>
- <creator>
- <docid>
- <filesize>
- <indextime>

Show top terms >>

Number of top terms:

Hint: use Shift-Click to select ranges, or Ctrl-Click to select multiple fields (or unselect all).

Top ranking terms. (Right-click for more options)

No	Rank	Field	Text
1	20	<contents>	1
2	20	<collectionid>	1001
3	20	<contents>	only
4	20	<contents>	5
5	20	<contents>	from
6	20	<contents>	3
7	20	<contents>	used
8	20	<contents>	2
9	20	<contents>	4
10	20	<contents>	than
11	19	<contents>	7

Index name: **D:\inetpub\wwwroot\tabuladx\collections\1001\index**

9 PRODUCT VERSION HISTORY

9.1 Version 1.12

Reference	Change
1.12-01	Updated release includes lightweight UltiDev Cassini web server

9.2 Version 1.01

Reference	Change
1.01-01	Initial Limited Release

10 ACKNOWLEDGEMENTS

This product includes Luke - Lucene Index Toolbox (<http://www.getopt.org/luke>), Copyright 2008 Andrzej Bialecki.