FatWire SOFTWARE

Site Capture

Version 1.1

Installation Guide

Document Publication Date: *Jan. 31, 2012*

FATWIRE CORPORATION PROVIDES THIS SOFTWARE AND DOCUMENTATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. In no event shall FatWire be liable for any direct, indirect, incidental, special, exemplary, or consequential damages of any kind including loss of profits, loss of business, loss of use of data, interruption of business, however caused and on any theory of liability, whether in contract, strict liability or tort (including negligence or otherwise) arising in any way out of the use of this software or the documentation even if FatWire has been advised of the possibility of such damages arising from this publication. FatWire may revise this publication from time to time without notice. Some states or jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

Copyright © 2012 FatWire Corporation. All rights reserved.

The release described in this document may be protected by one or more U.S. patents, foreign patents or pending applications.

FatWire, FatWire Content Server, FatWire Engage, FatWire Satellite Server, CS-Desktop, CS-DocLink, Content Server Explorer, Content Server Direct, Content Server Direct Advantage, FatWire InSite, FatWire Analytics, FatWire TeamUp, FatWire Content Integration Platform, FatWire Community Server and FatWire Gadget Server are trademarks or registered trademarks of FatWire, Inc. in the United States and other countries.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. AIX, AIX 5L, WebSphere, IBM, DB2, Tivoli and other IBM products referenced herein are trademarks or registered trademarks of IBM Corporation. Microsoft, Windows, Windows Server, Active Directory, Internet Explorer, SQL Server and other Microsoft products referenced herein are trademarks or registered trademarks of Microsoft Corporation. Red Hat, Red Hat Enterprise Linux, and JBoss are registered trademarks of Red Hat, Inc. in the U.S. and other countries. Linux is a registered trademark of Linus Torvalds. SUSE and openSUSE are registered trademarks of Novell, Inc., in the United States and other countries. XenServer and Xen are trademarks or registered trademarks of Citrix in the United States and/or other countries. VMware is a registered trademark of VMware, Inc. in the United States and/or various jurisdictions. Firefox is a registered trademark of the Mozilla Foundation. UNIX is a registered trademark of The Open Group in the United States and other countries. Any other trademarks and product names used herein may be the trademarks of their respective owners.

This product includes software developed by the Indiana University Extreme! Lab. For further information please visit http://www.extreme.indiana.edu/.

Copyright (c) 2002 Extreme! Lab, Indiana University. All rights reserved.

This product includes software developed by the OpenSymphony Group (http://www.opensymphony.com/).

The OpenSymphony Group license is derived and fully compatible with the Apache Software License; see http://www.apache.org/LICENSE.txt.

Copyright (c) 2001-2004 The OpenSymphony Group. All rights reserved.

You may not download or otherwise export or reexport this Program, its Documentation, or any underlying information or technology except in full compliance with all United States and other applicable laws and regulations, including without limitations the United States Export Administration Act, the Trading with the Enemy Act, the International Emergency Economic Powers Act and any regulations thereunder. Any transfer of technical data outside the United States by any means, including the Internet, is an export control requirement under U.S. law. In particular, but without limitation, none of the Program, its Documentation, or underlying information of technology may be downloaded or otherwise exported or reexported (i) into (or to a national or resident, wherever located, of) any other country to which the U.S. prohibits exports of goods or technical data; or (ii) to anyone on the U.S. Treasury Department's Specially Designated Nationals List or the Table of Denial Orders issued by the Department of Commerce. By downloading or using the Program or its Documentation, you are agreeing to the foregoing and you are representing and warranting that you are not located in, under the control of, or a national or resident of any such country or on any such list or table. In addition, if the Program or Documentation is identified as Domestic Only or Not-for-Export (for example, on the box, media, in the installation process, during the download process, or in the Documentation), then except for export to Canada for use in Canada by Canadian citizens, the Program, Documentation, and any underlying information or technology may not be exported outside the United States or to any foreign entity or “foreign person” as defined by U.S. Government regulations, including without limitation, anyone who is not a citizen, national, or lawful permanent resident of the United States. By using this Program and Documentation, you are agreeing to the foregoing and you are representing and warranting that you are not a “foreign person” or under the control of a “foreign person.”

FatWire Site Capture Installation Guide
Document Publication Date: Jan. 31, 2012
Product Version: Version 1.1

FatWire Headquarters
FatWire Corporation
330 Old Country Road
Suite 303
Mineola, NY 11501

Table of

Contents

About This Guide

This guide contains procedures for installing and configuring FatWire Site Capture, a web application that downloads dynamically published websites to support archiving and rapid deployment operations.

Who Should Use This Guide

This guide is intended for installation engineers with experience installing and configuring enterprise-level software, such as application servers and databases. Also required is a general administrator's experience with FatWire Content Server and the FatWire Web Experience Management Framework.

Related Documents

See the following documents in the FatWire documentation set:

- *FatWire Site Capture Administrator's Guide*
- *FatWire Web Experience Management Framework Administrator's Guide*

Conventions

The following text conventions are used in this guide:

- **Boldface** type indicates graphical user interface elements that you select.
- *Italic* type indicates book titles, emphasis, or variables for which you supply particular values.
- `Monospace` type indicates file names, URLs, sample code, or text that appears on the screen.
- **`Monospace bold`** type indicates a command.

Third-Party Libraries

FatWire Content Server 7.6 patch 2 and its applications include third-party libraries. For additional information, see *FatWire Content Server 7.6 Patch 2: Third-Party Licenses*.

Chapter 1

Welcome

This chapter presents an overview of the FatWire Site Capture application and the installation process.

This chapter contains the following topics:

Introducing FatWire Site Capture

FatWire Site Capture is a web application that integrates with FatWire Content Server through the FatWire Web Experience Management (WEM) Framework to capture dynamically published websites for evaluation, compliance purposes, high availability requirements, and other types of scenarios.

Crawls can be initiated manually from the Site Capture interface, or they can be triggered by the completion of a Content Server RealTime publishing session. In each scenario, the crawler captures the site in one of the following modes, depending on the user's selections:

- Static mode: The site is stored as files that are ready to be served. Only the latest capture is kept.
- Archive mode: The site is stored in a zip file. A pointer in the Site Capture database enables archive management from the Site Capture interface.

About This Guide

This guide contains procedures for installing and configuring Site Capture to support:

- Static and archive capture initiated manually, from the Site Capture interface.
- Static and archive capture triggered by the completion of a Content Server RealTime publishing session. Setting up publishing-triggered site capture is an option.
- Administrative users and developers. The Site Capture application is designed for general administrators of the Content Server system on which the Site Capture application will be running. Developers will write advanced crawler configuration code – for example, code that triggers Site Capture to execute a post-crawl command such as copying statically captured sites to a web server's doc base.

Before You Begin

Users of this guide must have experience installing and configuring enterprise-level software, such as application servers and databases. Also required is a general administrator's experience with Content Server and the WEM Framework.

- To complete the procedures in this guide, you must be a Content Server general administrator who belongs to the `RESTAdmin` security group.
- Download the Site Capture release notes and *Supported Platform Document* for information about supported operating systems, application servers, databases, and browsers.

  Note

  All product documentation is available on the e-docs site at `http://support.fatwire.com`, a password-protected site. Accounts can be requested at the same URL.

- Read this guide to acquire an understanding of the Site Capture installation process. The basic steps are configuring the Site Capture application server, running the

installer to create the Site Capture `war` file, deploying Site Capture, completing post-installation and verification steps, and if necessary, setting up publishing-triggered site capture.

- On all systems, set the `JAVA_HOME` variable to point to a valid installation of JDK 1.6.
- Prepare the Site Capture installation components:
  - Ensure you have a dedicated and fully functional Content Server 7.6 Current Patch installation running the WEM Framework, which uses Central Authentication Service (CAS). Site Capture must communicate with this Content Server system in order to run.
  - Ensure you have a dedicated host machine on which to install the Site Capture application server and the application itself. During the installation process, you will configure Site Capture to communicate with the Content Server/WEM Framework system, which runs on its own host machine (as described above).
  - Decide whether your Site Capture application will be running as a single application or in clustered mode. For a diagram of a single-server installation, see Figure 1, on page 10. For a diagram of a clustered installation, see Figure 2, on page 10.
    - Install or reuse the following components:
      - Install a supported Site Capture application server on the dedicated host machine. For clustered installations, install the application server for each Site Capture instance.
      - For a clustered installation, install a load balancer on a host machine of your choice. **The Site Capture installation directory must be a shared directory, accessible to all cluster members**.
      - To store archived sites, you can either reuse Content Server's database or install a dedicated, supported Site Capture database on a host machine of your choice.
- Decide whether you will be running publishing-triggered site capture. If so, you will need a Content Server source and target system:
  - Both Content Server systems must be version 7.6 current patch.

    The target system must be running the WEM Framework, which provides the REST API and WEM SSO API. The APIs enable the target system to communicate with the Site Capture application at the end of the publishing session in order to start the required crawlers. The Content Server target system will then receive crawler invocation status from the Site Capture application and sends that information to the source Content Server system. Both Content Server systems record the status information in their own log files (`futuretense.txt`, by default).
  - You will integrate the source and target Content Server systems with your Site Capture installation. For some of the possible configurations, see Figure 3, on page 11 and Figure 4, on page 11.
- Decide whether to install sample crawlers (**recommended**). For more information about the sample crawlers, see the first note on page 20.
- You have the option to install Site Capture silently or graphically. The silent installer provides help and sample values for every piece of information that needs to be set.

Figure 1: Single-Server Installation

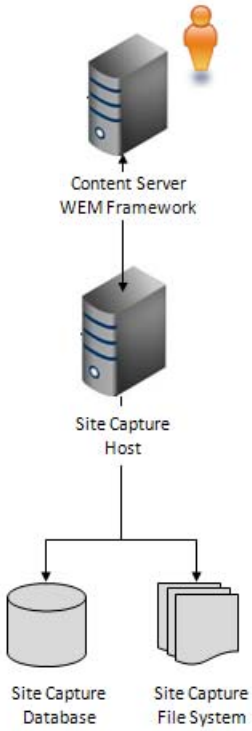


Figure 2: Clustered Installation

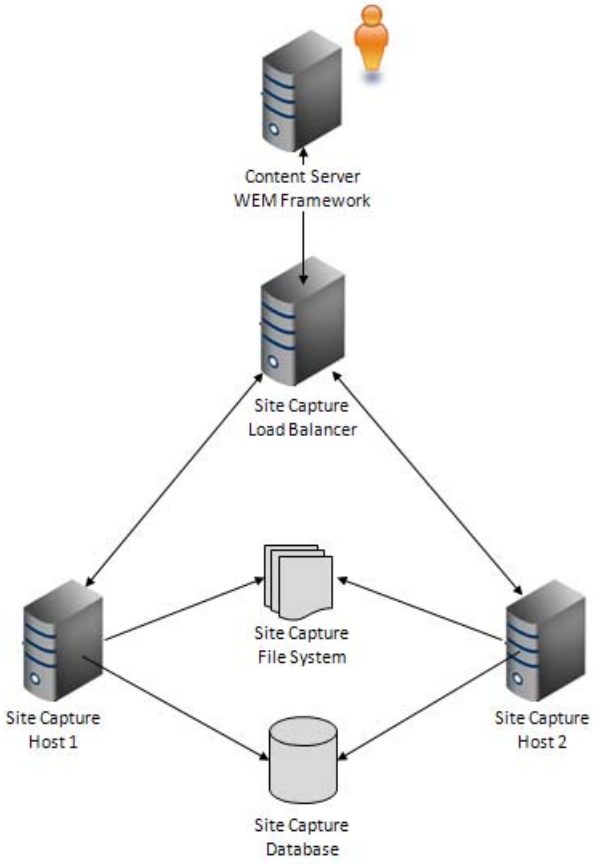
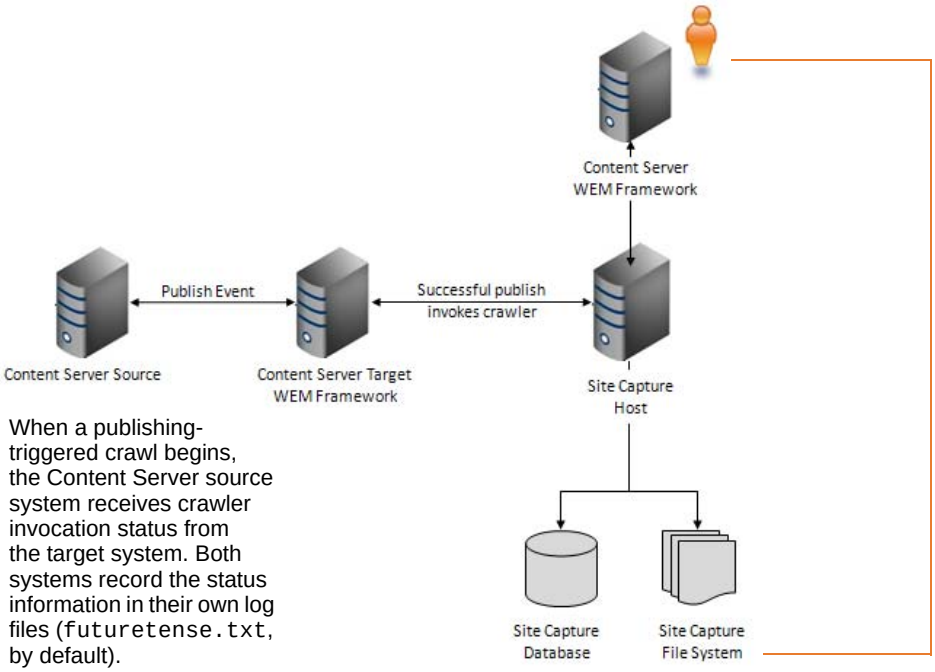


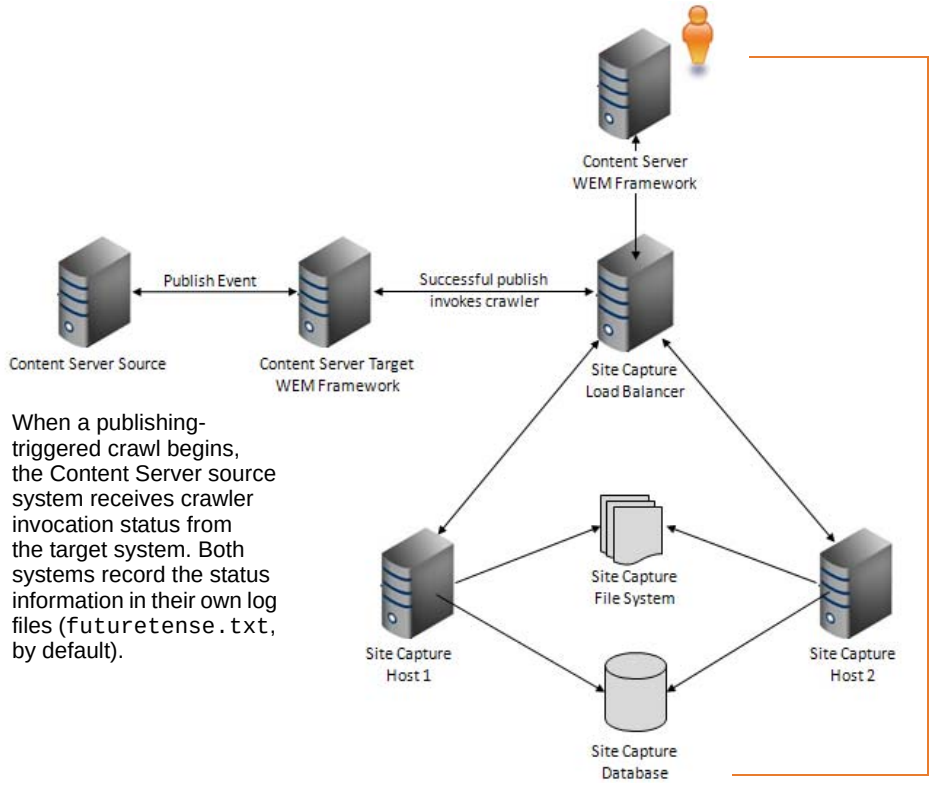
Figure 3: Single-Server Installation Enabled for Publishing-Triggered Capture



When a publishing-triggered crawl begins, the Content Server source system receives crawler invocation status from the target system. Both systems record the status information in their own log files (`futuretense.txt`, by default).

Single-server installation from Figure 1

Figure 4: Clustered Installation Enabled for Publishing-Triggered Capture



When a publishing-triggered crawl begins, the Content Server source system receives crawler invocation status from the target system. Both systems record the status information in their own log files (`futuretense.txt`, by default).

Clustered installation from Figure 2

Installation Worksheet

When configuring the Site Capture application server(s), you will specify parameters that you will also reuse when setting up the installer to create the Site Capture `war` file. Record the parameters in the following table:

Installation Type	Installation Parameter
All types	Content Server's host name (or IP address). This is the Content Server on which Site Capture will run as a WEM application:
	Content Server's port number:
	Content Server context root:
	Content Server general administrator's user name:
	Above user's password:
	CAS host name (or IP address):
	CAS port number:
	CAS context root:
Single-Server	Application server's host name (or IP address):
	Application server's port number:
	Application server's datasource name:
Clustered	Load balancer's host name (or IP address):
	Load balancer's port number:

Next Step

Continue to Chapter 2, "Configuring the Application Server."

Chapter 2

Configuring the Application Server

Site Capture can run as a single application or in a clustered installation.

This chapter contains the following topics:

Configuration Options

In this chapter, you will configure an application server to support Site Capture running as a single instance or in cluster mode. Complete one of the following procedures:

- Configuring Tomcat Application Server
- Configuring JBoss Application Server
- Configuring WebLogic Application Server
- Configuring WebSphere Application Server

Configuring Tomcat Application Server

If you are creating a clustered installation, complete the following steps on all cluster members.

To configure Tomcat

1. Open `server.xml` in `<apache-tomcat-6.0.29 directory>\conf` and do the following:
   a. Replace port `8080` with the Site Capture port (on which the application server is listening):

   Note

   For clustered installations, specify a unique port number for each cluster member.

   ```
   <Connector port="8080" protocol="HTTP/1.1"
      connectionTimeout="20000" redirectPort="8443" />
   ```

   b. Configure the data source for Site Capture.

   Note

   For clustered installations, configure the same data source on each cluster member.

   Refer to Table 1, on page 15 for information about configuring the driver class name and URL for your selected driver.

   Below is a sample DB2 configuration:

   ```
   <Context path="" docBase="ROOT" debug="5" reloadable="true"
      crossContext="true">
   ```

```
        <Resource name="scDataSource"
           auth="Container"
           type="javax.sql.DataSource"
           url="jdbc:db2://<hostname>:<port>/CLUSTDB"
           driverClassName="com.ibm.db2.jcc.DB2Driver"
           username="xxxxx"
           password="xxxxxxx"
           maxActive="20"
           maxIdle="10"
           maxWait="-1" />
     </Context>
```

Table 1: Database Driver Parameters

Database Driver	Parameter	Value
SQL server	DriverClass	`net.sourceforge.jtds.jdbcx.JtdsDataSource`
	Required `.jar` files	`jtds-1.2.jar`
	URL	`jdbc:jtds:sqlserver://<server>:<dbport>/<dbname>` **Ex:** `jdbc:jtds:sqlserver://10.120.14.22:1433/SC70`
DB2	DriverClass	`com.ibm.db2.jcc.DB2Driver`
	Required `.jar` files	`db2jcc.jar, db2cc_license_cu.jar`
	URL	`jdbc:db2://<hostname>:<dbport>/<dbname>` **Ex:** `jdbc:db2://10.120.16.30:50000/SC70`
Oracle	DriverClass	`oracle.jdbc.driver.OracleDriver`
	Required `.jar` files	`Ojdbc6.jar`
	URL	`jdbc:oracle:thin:@//<hostname>:1521/<dbname>` **Ex:** `jdbc:oracle:thin:@//godzilla.fatwire.com:1521/SC70`

2. Copy the required database driver jar file to `<apache-tomcat-6.0.29 directory>\lib`.
3. Once you have completed configuring the application server(s), continue to Chapter 3, "Installing FatWire Site Capture."

Configuring JBoss Application Server

If you are creating a clustered installation, complete the following steps on all cluster members.

To configure JBoss

1. Create a JBoss instance by duplicating the `default` folder in `<jboss_install_dir>\server` and naming the new folder `<jboss_instance_name>`.
2. Configure the data source:

> **Note**
>
> For clustered installations, configure the same data source on each cluster member.

   a. Create a new datasource file named `<database_type>-ds.xml` and place it in the `<jboss_install_dir>/server/<jboss_instance_name>/deploy` directory.

      1) Paste the `xml` code shown below:

```
<?xml version="1.0" encoding="UTF-8"?>
<datasources>
   <local-tx-datasource>
      <jndi-name>jdbc/scDataSource</jndi-name>
      <connection-url><URL></connection-url>
      <driver-class><Driver Class></driver-class>
      <user-name><User></user-name>
      <password><Password></password>
   </local-tx-datasource>
</datasources>
```

      2) Edit the code:

         - Specify the datasource name.
         - Replace the **`<URL>`** and **`<Driver Class>`** with the URL and driver class shown in Table 1, on page 15.
         - Replace the **`<User>`** and **`<Password>`** with the credentials of the user that connects to the database.

   b. Place the required database driver into the following folder:

```
<jboss_install_dir>/server/<instancename>/lib
```

3. Set the Site Capture port.

> **Note**
>
> For clustered installations, specify a unique port number for each cluster member.

You have several ways to set the Site Capture port. For example, open `server.xml` and replace port `8080` with the Site Capture port:

```
<Connector port="8080" address="${jboss.bind.address}"
   maxThreads="250" maxHttpHeaderSize="8192"
   emptySessionPath="true" protocol="HTTP/1.1"
   enableLookups="false" redirectPort="8443" acceptCount="100"
   connectionTimeout="20000" disableUploadTimeout="true" />
```

4. Once you have completed configuring the application server(s), continue to Chapter 3, "Installing FatWire Site Capture."

Configuring WebLogic Application Server

If you are creating a clustered installation, complete the following steps on all cluster members.

To configure WebLogic

1. Create a domain in WebLogic and configure the administrator user name and password for the domain.
2. Start the WebLogic Admin Server from the command prompt:
   - Windows: `startweblogic.cmd`
   - `Unix:  ./startweblogic.sh`
3. Log in to the console.
   a. Create a managed server `<managed_server_name>` in the domain you just created.

   Note

   For clustered installations, specify a unique port number for each cluster member.

   b. Create a data source and map it to the managed server `<managed_server_name>`.

   Note

   For clustered installations, configure the same data source on each cluster member.

4. Once you have completed configuring the application server(s), continue to Chapter 3, "Installing FatWire Site Capture."

Configuring WebSphere Application Server

If you are creating a clustered installation, complete the following steps on all cluster members.

To configure WebSphere

1. Create a WebSphere application server instance.

   Note

   For clustered installations, specify a unique port number for each cluster member.

2. Create a data source for the newly created server instance.

   Note

   For clustered installations, configure the same data source on each cluster member.

   a. Copy the required database driver into `<websphere_install_dir>/websphere/AppServer/universalDriver/lib`. (refer to Table 1, on page 15 for the required driver).

   b. Log in to the WebSphere console and do the following:

      1) Add the user name and password of the database by creating an alias (go to **Security > Global security > Java Authentication and Authorization Service > J2C authentication data**).

      2) Create a JDBC provider and data source (go to **Resources > JDBC**).

3. Once you have completed configuring the application server(s), continue to Chapter 3, "Installing FatWire Site Capture."

Chapter 3

Installing FatWire Site Capture

Site Capture can be installed graphically or silently. You will first create the `war` file using the installer, then manually deploy the `war` file and test the installation.

This chapter contains the following topics:

Installation Steps

To install Site Capture, you will complete the following basic steps:

1. Run the silent or GUI installer to create the Site Capture `war` file (`ROOT.war`).
2. Deploy the Site Capture application.

Detailed steps are provided below.

Note

- If you are creating a clustered installation:
  - Run the installer once. You will deploy the `ROOT.war` file (and `ROOT` folder) on all cluster members, as shown in this chapter.
  - The Site Capture installation directory must be a shared directory, accessible to all other cluster members.
- We recommend installing the sample crawlers. Procedures for quickly testing and using the Site Capture installation are based on the **Sample** crawler, described below.
  - **Sample** is a basic crawler that can be easily configured to capture any dynamic site. The configuration step amounts to setting the crawler's start URI in the crawler's configuration file.
  - **FirstSiteII** is a crawler with advanced configuration code for capturing Content Server's FirstSiteII sample website. The code demonstrates the implementation of various methods and interfaces that are used to control the crawler's site capture process, in this example, downloading a dynamic site (FirstSiteII) as a static site.

  Crawlers can be easily deleted should you no longer need them.

Running the Silent Installer

Note

If you wish to run the GUI installer instead of the silent installer, skip to "Running the GUI Installer," on page 21.

1. Unzip the `sitecapture.zip` file into the desired location on the server.
2. Configure the `omii.ini` file (located at the root level in the extracted folder) according to instructions provided in the file, and save the file.

Reminder

We recommend installing the sample crawlers. For information about the crawlers, see the first note on this page.

3. Open the `Install.ini` file (located at the root level in the extracted folder), add the line `loadfile=omii.ini` to the first section, and save the file.
4. Open a command prompt and run one of the following commands:
   - Windows: **`scInstall.bat -silent`**
   - Linux: `sh` **`scInstall.sh -silent`**

   You will see the following message in the command window after installation is successful – `"Installation Finished Successfully"`
5. Wait until the installer process exits normally and shows the command prompt again. The `ROOT` folder and the `ROOT.war` file will be created in the `webapps` folder of the installation directory specified in the `omii.ini` file.
6. Deploy the Site Capture `ROOT.war` file. For instructions, see one of the following sections:
   - Deploying on Tomcat Application Server
   - Deploying on JBoss Application Server
   - Deploying on WebLogic Application Server
   - Deploying on WebSphere Application Server

Running the GUI Installer

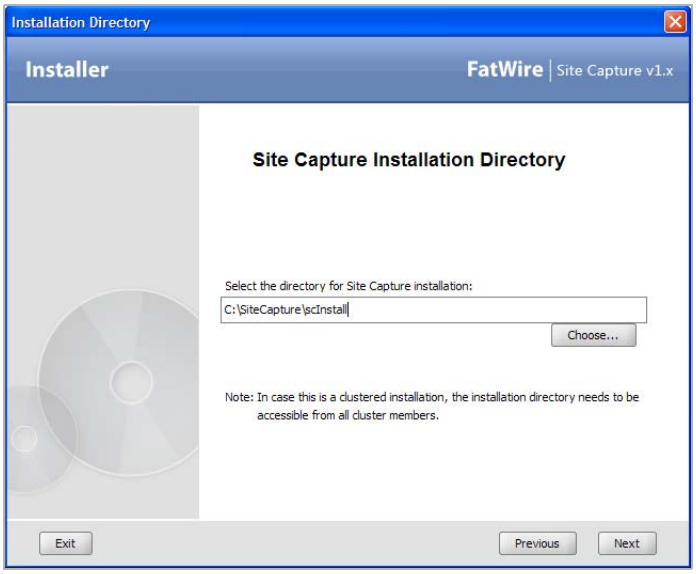
Note

If you wish to run the silent installer instead of the GUI installer, go to "Running the Silent Installer," on page 20.

1. Download `sitecapture.zip` to a directory on your Site Capture server and extract it into a temporary directory. The extracted folder contains the `scInstall.bat` and `scInstall.sh` files, which are needed to run the Site Capture installer.
2. Execute the following installer script:
   - On Windows: `scInstall.bat`
   - On Unix: `scInstall.sh`

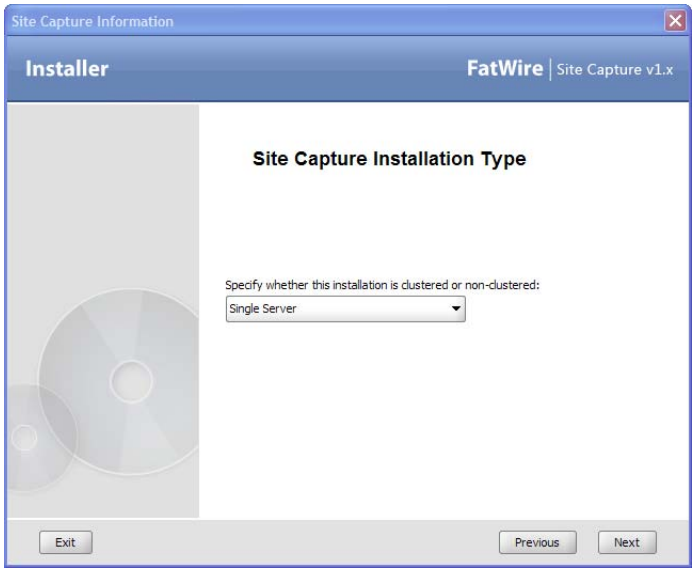
The installer “Welcome” screen is displayed:

3. Specify the path to the directory where Site Capture will be installed and click **Next**.

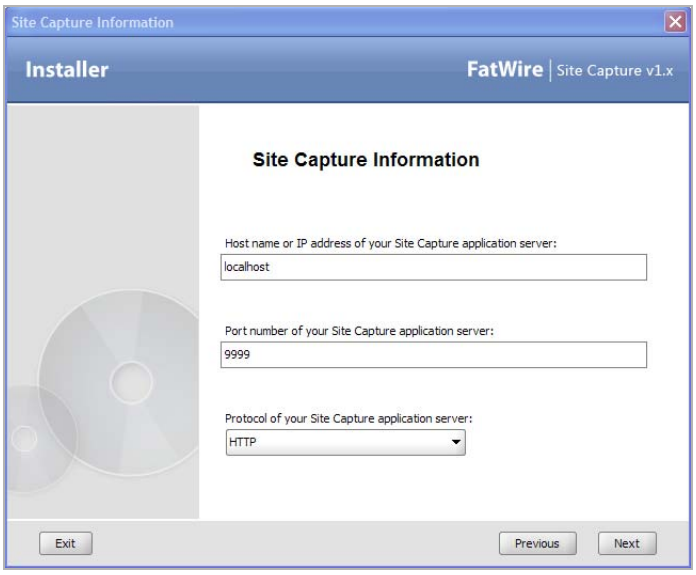


4. Select the product to be installed (Site Capture) and click **Next**.

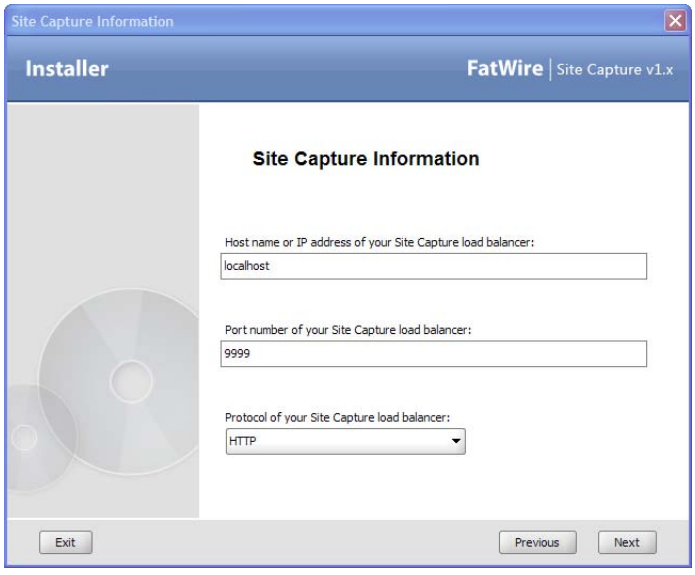
5. Select either **Single Server** or **Cluster**, and click **Next**.



6. Enter information about the Site Capture host machine:
   - Single server installation:
     - Enter the host name (or IP address) of the machine running the Site Capture application server.
     - Enter the port number of the Site Capture application server (the same number that you specified in the application server configuration steps).
     - Specify the protocol to be HTTP or HTTPS.



   - Clustered installation:
     - Enter the host name (or IP address) of the machine running the load balancer.
     - Enter the port number of the load balancer.
     - Specify the protocol to be HTTP or HTTPS.

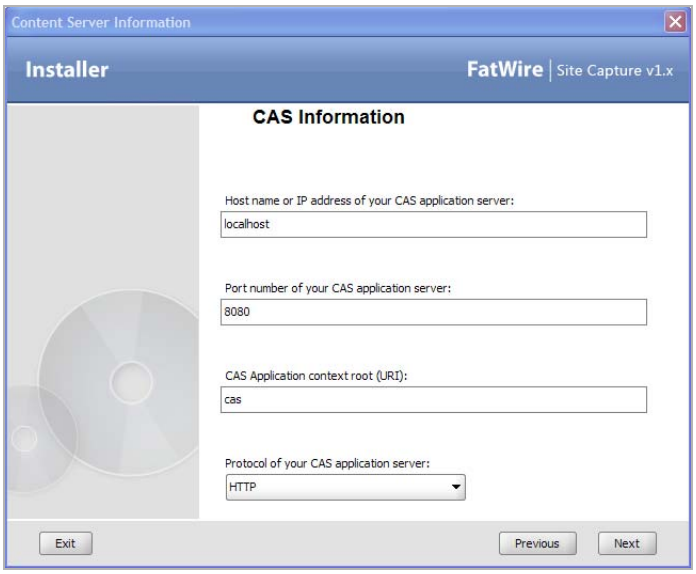


7. Enter information about the Content Server system on which Site Capture will run as a WEM application:
   - Content Server host name or IP address
   - Port on which Content Server is listening
   - Protocol of the Content Server application server (HTTP or HTTPS)
   - Context root of the Content Server application

8. Enter the Content Server administrator's credentials:
   - Current user name of the Content Server general administrator (who belongs to the `RESTAdmin` security group)
   - Current password

9. Enter information about the CAS application:

- CAS host name (or IP address)
- Port number
- Protocol of the CAS application server (HTTP or HTTPS)
- CAS context root



10. Select the application server on which Site Capture will be installed.

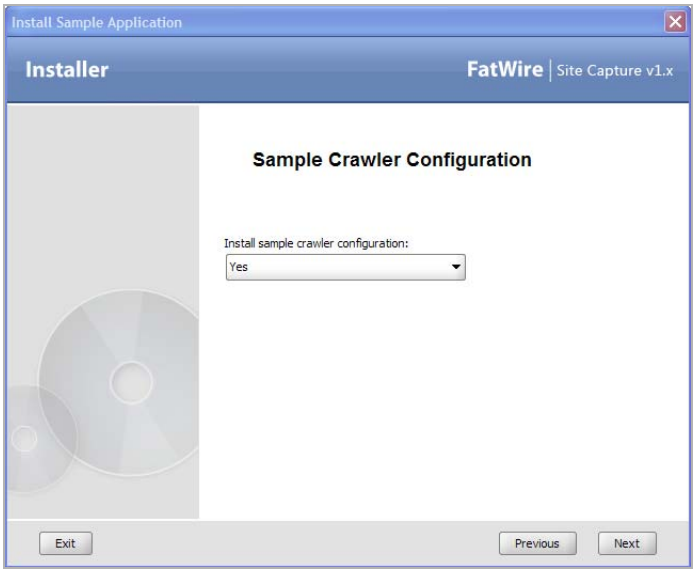
> **Note**
>
> The installer will create a `ROOT.war` file, which you will deploy on the application server.

11. Enter database information:

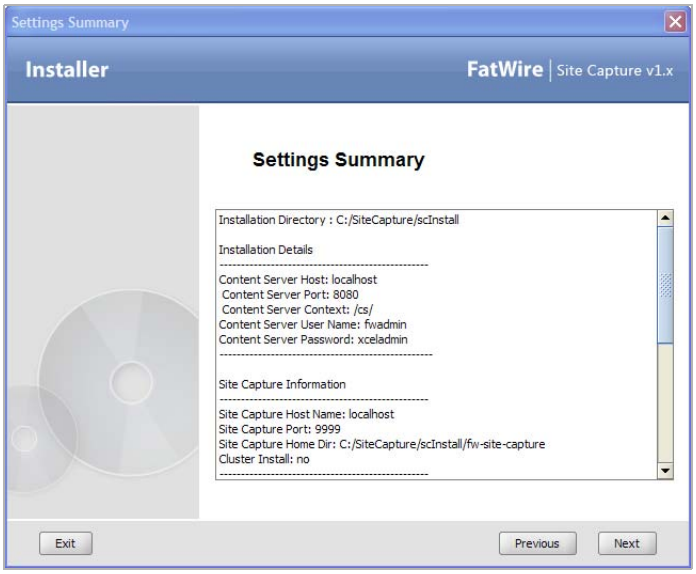
- Select the database to which Site Capture will connect (select either Site Capture's database or Content Server's database).
- Specify the name of Site Capture's data source exactly as it is configured in the application server (in Chapter 3, "Installing FatWire Site Capture").

12. Select whether to install the sample crawlers:

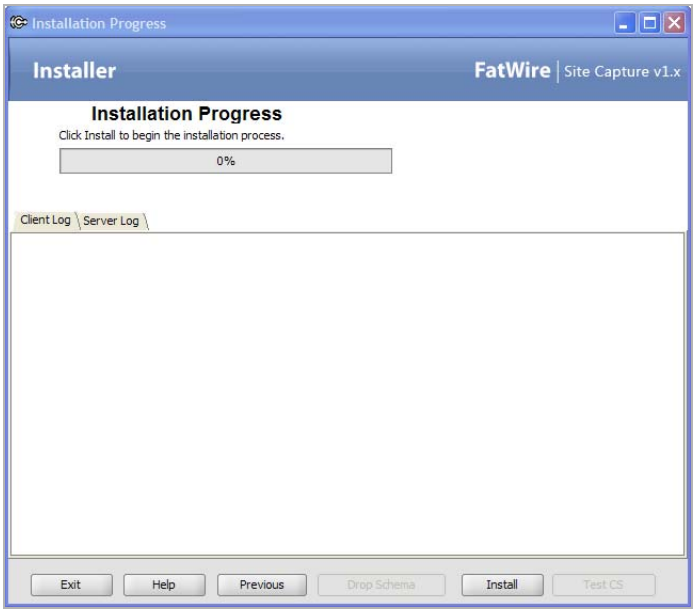
> **Reminder**
>
> We recommend installing the sample crawlers. For information about the crawlers, see the first note on page 20.



13. Review the settings summary to verify the data you have entered. To edit your settings, click **Previous** to return to the relevant screens.



14. Once you have confirmed your settings, click **Install**.



When the installation process successfully ends, the installer displays the following message: `"Site Capture war file was successfully created"`

The Site Capture file system was also created. The `war` file, called `ROOT.war`, is located in the `<SC_INSTALL_DIR>/fw-site-capture/webapps` folder. For information about the Site Capture file system, see Chapter , "Reference: FatWire Site Capture File System."

15. Deploy the Site Capture `ROOT.war` file. For instructions, see one of the following sections:
    - Deploying on Tomcat Application Server
    - Deploying on JBoss Application Server
    - Deploying on WebLogic Application Server
    - Deploying on WebSphere Application Server

Deploying Site Capture

Note

- For clustered installations, complete the deployment steps on each cluster member.
- The Site Capture application must have a context root of `/`.

Deploying on Tomcat Application Server

1. Copy the `ROOT` folder in `<SC_INSTALL_DIR>\fw-site-capture\webapps\` and overwrite the `ROOT` folder in the `<apache-tomcat-6.0.29 directory>\webapps\` directory.
2. To start Site Capture:
   a. Ensure that Content Server is running.
   b. Start the application server (run the `startup.bat` or `startup.sh` file in the `<apache-tomcat-6.0.29 directory>\bin` folder).
3. Continue to "Post-Installation Steps," on page 32.

Deploying on JBoss Application Server

1. Delete the `ROOT.war` folder in `<jboss_install_dir>\server\<jboss_instance_name>\deploy\jboss- web.deployer\` and create an empty `ROOT.war` folder in `<jboss_install_dir>\server\<jboss_instance_name>\deploy`.
2. Go to the Site Capture installation folder, `<SC_INSTALL_DIR>\fw-site-capture\webapps\ROOT` copy the sub-folders (`META-INF`, `resources`, `WEB-INF`) into `ROOT.war` created in the previous step.
3. To start Site Capture:
   a. Ensure that Content Server is running.
   b. Start the application server (run the `run.bat` or `run.sh` file in `<jboss_install_dir>\bin`):

      `run.bat -c <jboss_instance_name> -b <site_capture_host_IP>`
4. Continue to "Post-Installation Steps," on page 32.

Deploying on WebLogic Application Server

1. Copy the files named `antlr-2.7.7.jar` and `commons-lang-2.5.jar` in Site capture's `WEB-INF\lib` folder:

   `<SC_INSTALL_DIR>\fw-site_capture\webapps\ROOT\WEB-INF\lib`

   and paste them into the WebLogic installation path:

   `<Weblogic Home>\wlserver_10.3\common\lib`.
2. Set the pre-class path in the `setdomainEnv.cmd` or `setdomainEnv.sh` file (located in `<Weblogic Home>\user_projects\domains\<domain name>\bin`) for the Site Capture domain.

Below is an example of where the changes need to be made:

- **On Windows** (in `setdomainEnv.cmd`):

```
set WL_HOME=E:\<WL HOME\Weblogic\wlserver_10.3

set PRE_CLASSPATH=%WL_HOME%\common\lib
   \antlr-2.7.7.jar;%WL_HOME
%\common\lib\commons-lang-2.5.jar(in windows machine)

for %%i in ("%WL_HOME%") do set WL_HOME=%%~fsi
```

- **On Linux** (in `setdomainEnv.sh`):

```
WL_HOME="/root/Oracle/Middleware/wlserver_10.3"

PRE_CLASSPATH=$WL_HOME/common/lib/antlr-2.7.7.jar:
$WL_HOME/common/lib/commons-lang-2.5.jar

export WL_HOME
```

3. Deploy the Site Capture application:
   a. Go to the deployment section in the console and select the path to the `ROOT` folder located in `<SC_INSTALL_DIR>\fw-site-capture\webapps`.
   b. Select the managed server as the deployment target.
4. To start Site Capture:
   a. Ensure that Content Server is running.
   b. Start the managed server `<managed_ server_name>`. For example:

```
./startmanagedweblogic.sh <managed_ server_name>
http://<admin_server_hostname>:<admin_server_port>/
```

5. Continue to "Post-Installation Steps," on page 32.

Deploying on WebSphere Application Server

1. In the WebSphere Admin interface, do the following:
   a. Go to the path **Application** > **Application Types** > **WebSphere enterprise applications** and select the `ROOT.war` file located in `<SC_INSTALL_DIR>\fw-site-capture\webapps\`.
   b. Complete the deployment process.
2. To start Site Capture:
   a. Ensure that Content Server is running.
   b. Start the application server.
3. Continue to "Post-Installation Steps," on page 32.

Post-Installation Steps

When Site Capture is installed and the application server is started, Site Capture starts and automatically registers itself as a WEM Framework application enabled on the Admin Site. In addition, Site Capture adds the necessary tables and records to its database, unless the database is Oracle.

To complete and verify your Site Capture installation, complete the steps in the following sections:

- If Site Capture Is Using an Oracle Database
- Verifying the Site Capture Application in the WEM Framework
- Authorizing Users to Work with Site Capture

If Site Capture Is Using an Oracle Database

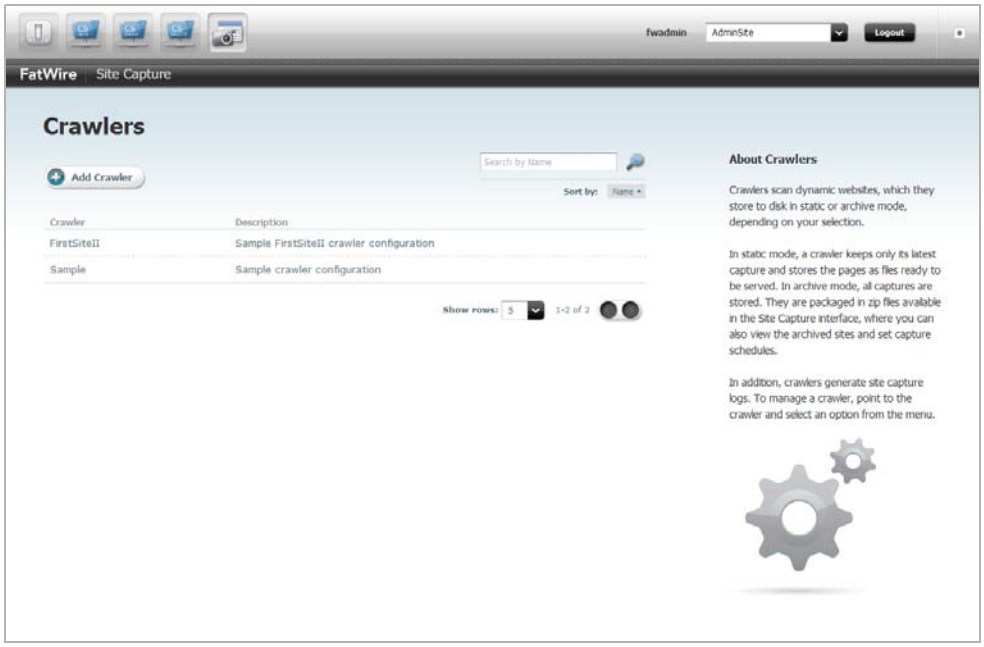
When Site Capture uses the Oracle database, the sql script that creates the schema must be run manually. The script, named `crawler_oracle_db.sql`, is available in the `<SC_INSTALL_DIR>/Sql-Scripts` folder.

Verifying the Site Capture Application in the WEM Framework

Ensure that the Site Capture application is available in the WEM Admin interface:

1. Log in to the WEM Framework as a general administrator.
2. Navigate to the AdminSite and select the WEM Admin application.
3. In the menu bar, click **Apps** and verify that **Site Capture** is listed on the “Apps” page.
4. Navigate to the Site Capture application (represented by this icon: ).

   The home page opens. If you chose to install the sample crawlers, they are listed on the home page as **Sample** and **FirstSiteII**.



Authorizing Users to Work with Site Capture

The Site Capture application is designed for Content Server general administrators and developers with general administrator's rights. All such users have access to the AdminSite, where the Site Capture application is registered and running. (Access is granted via the `GeneralAdmin` role on the AdminSite and the user's membership in the `RestAdmin` security group. For more information about user authorization and WEM Framework administration, see the *WEM Framework Administrator's Guide*.)

Site Capture users also require administrative access to its host machine, especially to retrieve statically captured sites and crawler logs.

Next Steps

- If you wish to set up publishing-triggered site capture, see Chapter 4, "Enabling Publishing-Triggered Site Capture."
- To get started with Site Capture, see the *Site Capture Administrator's Guide* for information about navigating the interface and writing crawler configuration code.

Chapter 4

Enabling Publishing-Triggered Site Capture

You have the option to enable publishing-triggered site capture once the Site Capture application is installed.

This chapter contains the following topics:

Integrating Site Capture with Content Server's Publishing Process

Site capture can be triggered by the completion of a RealTime publishing session. If you wish to enable publishing-triggered site capture, complete the steps in this section after you have installed the Site Capture application (Chapter 3). The steps show you how to integrate Content Server's publishing system to communicate with your installed Site Capture application. For some of the possible system configurations, see Figures 3 and 4 on page 11.

To integrate Site Capture with Content Server's publishing process

1. On the Content Server source system:
   a. Deploy the `fw-crawler-publish-listener-1.0.jar` file to the `<cs_deploy>/WEB-INF/lib` folder.
   b. Unzip the `fw-crawler-publish-listener-1.0-elements.zip` file and import `FW_PublishingEventRegistry.html` using CatalogMover.

      This step creates a `RemoteElementInvokingPublishingEventListener` record in the `FW_PublishingEventRegistry` table in the Content Server database, which will allow publish events to call the `InvokeCrawler` element on the Content Server target system.
   c. Restart the Content Server source system.
2. On the Content Server target system:

   Note

   The target Content Server system must be running the WEM Framework.

   a. Using CatalogMover, import `ElementCatalog.html` and `SiteCatalog.html` from the unzipped `fw-crawler-publish-listener-1.0-elements.zip` file extracted in step 1b above (on the Content Server source system).

      This step imports the `InvokeCrawler.jsp`, which is used to start the crawler(s) in the Site Capture application.

      Note

      The crawler(s) must be defined in the publishing destination definition for Site Capture and in the Site Capture application. For more information, see "Next Steps," on page 37.

   b. Copy the `crawler.properties` file (in the `/<cs_deploy>/WEB-INF/classes` folder) and configure the properties listed below:
      - `sc.url:` Do one of the following:

        For a single-server installation, specify the URL of the Site Capture application:

        `sc.url=http://<sitecapturehost:sitecaptureport>/__admin`

For a clustered installation, specify the URL of the load balancer:

`sc.url=http://<loadbalancerhost:loadbalancer>/__admin`

- `cas.url=http://<cashost:casport>/cas`

  Specify the CAS application that is pointed to by the Site Capture application:

- `cs.username=<RestAdmin User>`

  Specify the user name of the Content Server general administrator exactly as it was specified during the Site Capture installation process:

- `cs.password=<Password>`

  Specify the above user's password exactly as it was specified during the Site Capture installation process:

c. Deploy the `fw-crawler-publish-listener-1.1.jar` file to the `<cs_deploy>/WEB-INF/lib` folder on the target Content Server system.

3. You have completed the integration process. Continue to "Next Steps" for a summary on setting up a publishing-triggered site capture operation.

Next Steps

At this point, you have completed the integration process. However, for publishing-triggered site capture to work, the following conditions must also be satisfied:

- A RealTime publishing destination definition must be configured on the source system to name the crawler(s) that will be invoked to capture the newly published site. The definition must also specify the crawlers' capture mode.
- The crawler(s) named in the step above must exist in the Site Capture application. In addition, the `CrawlerConfigurator.groovy` file for each crawler must specify at least a valid starting URI and link extraction logic for the crawler.

Note

Once the above configuration steps are completed, you will publish the site from the source Content Server system to the target Content Server system. When publishing ends, site capture begins and proceeds as follows:

1. The source Content Server system calls the `InvokeCrawler` element on the target system.
2. The target Content Server system communicates with the Site Capture application and invokes the crawler(s).
3. The Content Server target system communicates crawler invocation status to the Content Server source system. Both the source and target systems record the status information in their own log files (`futuretense.txt`, by default).

   At the same time, the invoked crawlers capture site resources either statically or in archive mode, depending on your settings in the publishing destination definition.

When enabling publishing-triggered site capture, you can configure as many publishing destination definitions and invoke as many crawlers as necessary. When you are ready to proceed with the configuration steps above, refer to the *Site Capture Administrator's Guide* for instructions. In the same guide, you will find information about navigating the Site Capture interface, setting up a site capture operation, and coding a crawler's configuration file to control the site capture process.

Appendix A

Reference: FatWire Site Capture File System

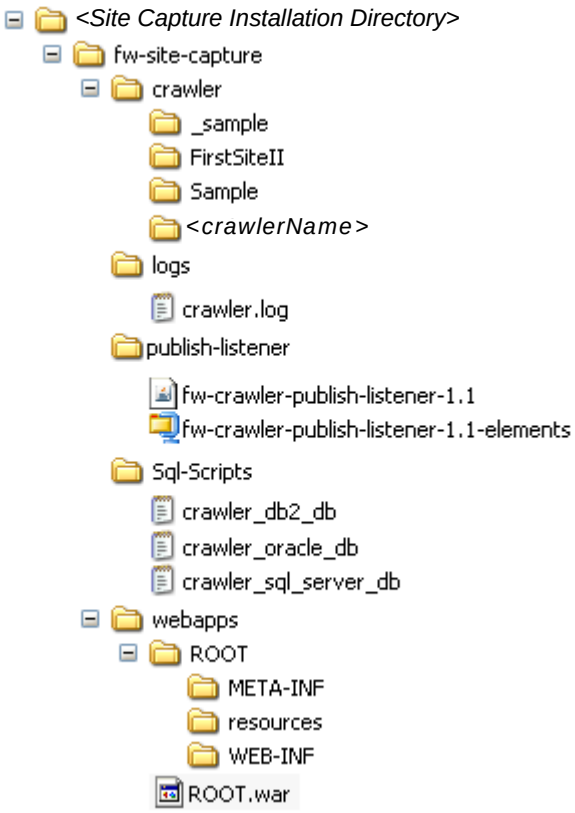
The Site Capture file system is created during the Site Capture installation process and stores installation-related files, property files, sample crawlers, and sample code used by the FirstSiteII crawler to control its site capture process. The file system also provides the framework in which Site Capture organizes users' custom crawlers and their captures.

This chapter contains the following topics:

General Directory Structure

Figure 1 shows Site Capture's most frequently accessed folders to help administrators find commonly used Site Capture information. All folders, except for `<crawlerName>`, are created during the Site Capture installation process (for information about `<crawlerName>` folders, see page 41 and "Custom Folders," on page 42).

Figure 1: Site Capture File System



Folder	Description
`/fw-site-capture`	The parent folder.
`/fw-site-capture/crawler`	Contains all Site Capture crawlers, each stored in its own crawler-specific folder.
`/fw/site-capture/crawler/_sample`	Contains the source code for the "FirstSiteII" sample crawler. For information about the code, see the *Site Capture Administrator's Guide*. **Note:** Folder names beginning with the underscore character ("_") are not treated as crawlers. They are not displayed in the Site Capture interface.

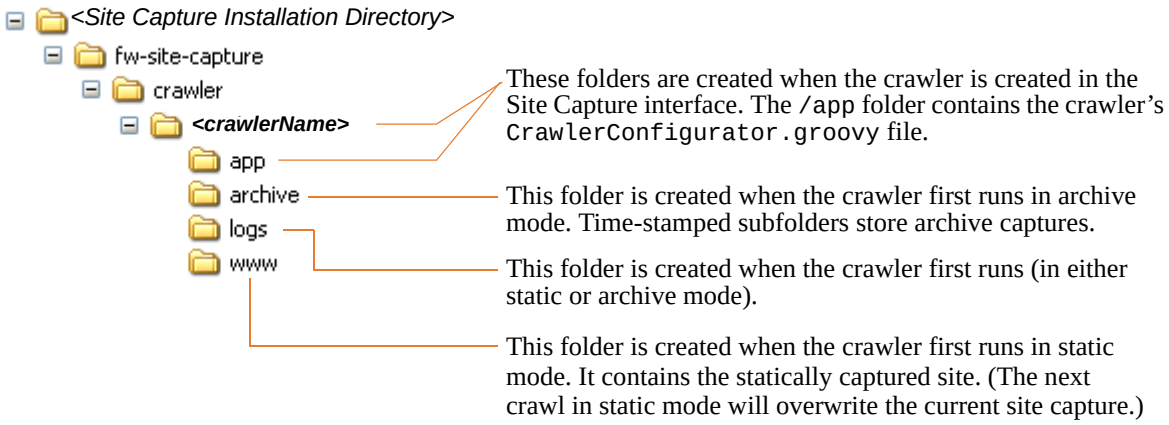
Folder	Description
`/fw-site-capture/crawler/<crawlerName>`	Represents a crawler. This custom folder is created by Site Capture for every crawler that a user defines in the Site Capture interface. This folder is used to organize the crawler's configuration file, archive captures, static captures, and logs. For a summary of the `<crawlerName>` folder structure, see "Custom Folders," on page 42. For detailed information about the folder structure, see the *Site Capture Administrator's Guide*. **Note:** The `FirstSiteII` and `Sample` folders are examples of a `<crawlerName>` folder, as explained in the next two rows of this table.
`/fw-site-capture/crawler/FirstSiteII`	Represents a sample crawler named "FirstSiteII." This folder is created only if the "FirstSiteII" crawler was installed during the Site Capture installation process. The `FirstSiteII` folder contains an `/app` folder, which stores the `CrawlerConfiguration.groovy` file specific to the "FirstSiteII" crawler. The file contains advanced configuration code for capturing Content Server's FirstSiteII dynamic sample website as a static website, and in this manner, demonstrating the use of various methods and interfaces in the `BaseConfigurator` class. When the "FirstSiteII" crawler is invoked in static or archive mode, subfolders are created within the `/FirstSiteII` folder. For more information about the folder structure, see the *Site Capture Administrator's Guide*.
`/fw-site-capture/crawler/Sample`	Represents a crawler named "Sample." This folder is created only if the "Sample" crawler was installed during the Site Capture installation process. The `Sample` folder contains an `/app` folder, which stores the `CrawlerConfiguration.groovy` file specific to the "Sample" crawler. The file contains basic configuration code for capturing any dynamic site. The code demonstrates the use of required methods (such as `getStartUri`) in the `BaseConfigurator` class. When the "Sample" crawler is invoked in static or archive mode, subfolders are created within the `/Sample` folder. For more information about the folder structure, see the *Site Capture Administrator's Guide*.
`/fw-site-capture/logs`	Contains the `crawler.log` file, a system log for Site Capture.
`/fw-site-capture/publish-listener`	Contains the following files needed for installing Site Capture for publishing-triggered crawls: • `fw-crawler-publish-listener-1.1-elements.zip` • `fw-crawler-publish-listener-1.1.jar`

Folder	Description
`/fw-site-capture/Sql-Scripts`	Contains the following scripts, which create database tables that are needed by Site Capture to store its data: • `crawler_db2_db.sql` • `crawler_oracle_db.sql` • `crawler_sql_server_db.sql`
`/fw-site-capture/webapps`	Contains the `ROOT/WEB-INF/` folder.
`/fw-site-capture/webapps/ROOT/WEB-INF`	Contains the `log4j.xml` file, used to customize the path to the `crawler.log` file.
`/fw-site-capture/webapps/ROOT/WEB-INF/classes`	Contains the following files: • `sitecapture.properties` file, where you can specify information for the Content Server application that is running the WEM Framework and Site Capture. The information includes Content Server's host machine name (or IP address) and port number. • `root-context.xml` file, where you can configure the Site Capture database.

Custom Folders

A custom folder is created for every crawler that a user creates in the Site Capture interface. The custom folder, `<crawlerName>`, is used to organize the crawler's configuration file, captures, and logs, as summarized in Figure 2.

Figure 2: Site Capture's Custom Folders: `<crawlerName>`



For more information about custom folders, see the *Site Capture Administrator's Guide.*