D.2 Spider Deployment Instructions

This Section presents the deployment guide to install the Spider component.

D.2.1 Software Requirements

- **Windows Server** (Preferably version 2008 or newer).
- **.NET Framework** versions 3.5, 4 and 4.5.
- **Microsoft SQL Server 2008**, preferably 2012, (also works with SQL Server express but there is a limit of 4GB to the size of the database).
- Microsoft IIS web server.

D.2.2 Instructions

1. **Microsoft IIS web server**
   (a) Install IIS from the add/remove software menu of Windows server.
   (b) Open the **IIS Control panel**, and click on the **Windows Platform Installer Icon**. From there, you are able to install all **.NET** versions as well as the **MVC4** library.
   (c) Enable **Windows Communication Foundation (WCF) Library**.
   (d) Enable **HTTP activation**.
   (e) All .NET versions should have the same settings.
2. **Microsoft SQL server**
   (a) Install the Microsoft SQL server following standard installation procedure (CDROM installation setup).
   (b) Use the **BlogForever Spider database deployment script** deploy.sql which creates the initial database structure.
3. **BlogForever Spider Web component**
   (a) Extract the BlogForever spider zip file in the IIS web folder (**c:/inetpub/wwwroot/Spider** by default).
   (b) Open the IIS control panel, right click on the folder and select **Convert to Application**. (All the files beneath this folder are parts of a web application, if not it is not accessible).
   (c) Setup database connection. Edit the **web.config** file inside the Spider web folder, containing the database connection string which has to be modified according to server settings.
4. **BlogForever Web Crawler component**
   (a) Extra the BlogForever Crawler zip file in a selected folder.

(b) Note that The default file storage location of the spider executable is in a subfolder below its location.

(c) Edit **CW.CrawlerSystem.ConsoleApp.exe.config** file to setup database connection string

(d) Furthermore, inside the AppSettings there are important application variables.

i. **EntityStorage** is the folder location of the entities the crawler extracts,

ii. **SourcesStorage** is the location of the sources description,

iii. **WebRequestCache** is used to cache every HTTP request result, expiration time is 10min (this is necessary for HTTP traffic optimisation)

iv. **UserAgent** is a copy of the Opera UserAgent,

v. **mexHttpBinding** is the description of all the network services that this program provides,

vi. **baseAddress** it the address of mexHttp

vii. **SvcConfigEditor** is used to define new endpoint connections towards a service. You can use the same application to configure your client towards a service or the other way round. Example of service description is in this address: http://bf4.itc.auth.gr/Spider/SpiderService.svc

viii. maxStringContentLength and maxString should not be modified.

(e) Execute the CW.CrawlerSystem.ConsoleApp.exe to run the crawler.