Britannica Concept Search for Lucene.Net
Technical Manual

Version 2.0.1.0

February, 2012

Britannica Concept Search for Lucene.Net - Technical Manual

Version 2.0.1.0, February, 2012

System specifications and the information contained in this document are subject to change without notice.

Encyclopaedia Britannica

Natural Language Technology Division

331 N.LaSalle Street,

Chicago, Illinois

USA 60654

Tel: (312) 205-6440

Fax: (312) 294-2162

E-mail: nltsupport@eb.com

Website: http://corporate.britannica.com/nlt/

Table of Contents

Chapter 1

Introduction

1.1. Welcome

Thank you for choosing the Britannica Concept Search for the Apache Lucene.Net of Apache Lucene™.

Once installed, users have access to the powerful capabilities of Britannica Concept Search to perform thorough, yet efficient morphological searches (see the next two sections for additional details).

This document describes the technical steps for installing and integrating Britannica Concept Search with the Apache Lucene.Net.

1.2. Morphological Search

Morphological search basically means that the analysis of a text, whether on a query or upon indexing, is done based on all inflected forms of the stems in the text. For elaborated information on morphological search as well as on the concept of stem, see the Britannica Concept Search User Manual for a specific language.

Britannica Concept Search is a robust solution provided by Encyclopaedia Britannica's Natural Language Technologies (NLT) Division for analyzing languages characterized by a complex and rich morphology such as Hebrew, Arabic and Persian.

Britannica Concept Search for Lucene.Net enables Lucene.Net-based search solutions to support morphological searches (both inflectional and derivational), synonym search, and cross language search.

1.3. Entity Extraction

Entity extraction is the process of identifying semantically significant tokens in input texts and extracting them into categorized structures. Britannica’s Intelligent Content Analysis (ICA), integrated within Britannica Concept Search, identifies and marks entities including names of people, place names, organizations, military terms, temporal expressions, and other categories of words and phrases in input texts.

Britannica's Natural Language Technologies (NLT) Division also specializes in providing cross-language capabilities to deal with foreign language texts (e.g. by English speakers who do not speak these languages). ICA offers cross-language entity extraction, the extraction of entities appearing in foreign language texts, and presenting and publishing them in English. ICA can also offer faceted navigation.

1.4. Notes on Britannica Concept Search Editions

Installation is the same for all three editions of Britannica Concept Search. The following is a summary of the features available in each edition of the Britannica Concept Search software:

Concept Search				
Standard Edition	░	▒	▓	Morphological Search & Context Analysis Exact Match Search
Enterprise Edition		▒	▓	Expanded Morphological Search Synonym Search English Search Cross Language Search Proper Name Transcription Search
Ultimate Edition			▓	Entity Search Faceted Navigation

1.5. Additional Help

For further explanations and clarifications, please contact us by e-mail at nltsupport@eb.com.

Chapter 2

Installation

This chapter describes the Britannica Concept Search installation procedure. The Attivio installation folder (e.g. `C:\ConceptSearch_LuceneNet`) will be referred to as `%B_LUCENE_HOME%` in the rest of this document.

This chapter contains the following sections:

- Prerequisites & Technical Information
- Installing Britannica Concept Search for Lucene.Net
- Licensing

2.1. Prerequisites & Technical Information

In order to use Britannica Concept Search for Lucene.Net, the following prerequisites are required:

- A Microsoft.NET installation of version 3.5.
- A working Lucene.Net installation (*dll*) of version 2.9.2.
- A maximal memory footprint of 100 MB in RAM needed by Britannica Concept Search. The total memory requirement should include this memory size in addition to Lucene.Net's own requirements. The memory footprint is not affected by loading Britannica Concept Search by multiple search threads.
- Disk space of 200MB for the complete installation of Britannica Concept Search.

Britannica Concept Search for Lucene.Net is written in .Net, and its deliverable consists of *dll* files. In addition, the deliverable contains several *dll* and database files (using proprietary data formats) that include language-specific lexicons with their grammatical rules.

2.2. Installing Britannica Concept Search for Lucene.Net

▸ To install the Concept Search for the Lucene.Net base software package

There is only one *.exe* file to be downloaded for the base Concept Search. The other *.exe* files are for the language packs.

1. Download the appropriate installation *exe* files listed below from the provided software kit.
2. Install the base Concept Search for Lucene.Net package by running the program `concept_search_lucenenet_base_<platform>.<version>.exe`
   During the installation, click **Next** and accept all defaults.
3. Install the language packs of your choice by using the installers named
   `concept_search_lucenenet_<language code>_<platform>.<version>`.exe

   For example:
   `concept_search_lucenenet_ar_win.2.0.0.0` is an Arabic language pack for Windows.

Provide a valid license key (see Licensing below for additional details).

Language packs must be installed to the same directory where you installed the base Concept Search.

2.3. Licensing

A valid license is necessary in order to use the Britannica Concept Search system. The licensing steps are performed by manually invoking the licensing command-line utility for the applicable product after it is installed. The license utility for each language package can be found inside the respective folder under the installation directory, in a subfolder called *license*. For example, the license utility for the Hebrew package can be found at under *Hebrew/license* and is called *HeCSKey.exe*.

In order to license a product:

1. Run the appropriate licensing utility from the command line. It will output a unique Britannica machine code for the computer on which the software is being installed.
2. To obtain a license, please copy the machine code and e-mail it to your Encyclopaedia Britannica contact or to the e-mail support line: *nltsupport@eb.com*. You will be provided with the appropriate license key.
3. When you receive the license key in a return e-mail from technical support, run the licensing utility again, this time passing it the license key as a parameter. You will receive the message "*License has been successfully installed.*"

Chapter 3

Integration

This chapter describes how to integrate the Britannica Concept Search with Lucene.Net. The sections in this chapter are:

- Enabling Concept Search on Your Lucene.Net Project
- Testing Concept Search on a Sample Lucene.Net Project
- Enabling Indexing
- Performing Queries

3.1. Enabling Concept Search on Your Lucene.Net Project

The following procedure is required to enable concept searching in Lucene.Net. The following *dll*s, residing in the *lib* or *lib64* folders (depending on the relevant architecture), should be put in the GAC (the GAC folder can be reached by typing *assembly* from the *Run* box):

- BLuceneNet_WbFacade.dll
- BLuceneNet_WbWrapper.dll
- BLuceneNet.dll

If signed, the Lucene.Net *dll* should also be put in the GAC.

Upon completion of the previous steps, you can start your .Net project as normal. All handling of data of the above configured language(s) will now be redirected to Britannica Concept Search's Tokenizer and Analyzer, which extend the Lucene.Net's Tokenizer and Analyzer.

3.2. Testing Concept Search on a Sample Lucene.Net Project

The *samples* folder includes a sample application called *BLuceneNet_Tester.exe*, demonstrating the basic capabilities of Concept Search for Lucene.Net. It can be used in order to perform a sanity check on the installation.

The *samples* folder contains a folder for each language package: Hebrew, Arabic and Farsi. Inside each such folder there exists the following:

- A folder named *filesToIndex<language>* containing several simple *txt* files, to be indexed by the sample application.
- A *txt* file named *query<language>* with several queries to be performed on the created index.
- *bat* files named *RunTester_Index_<language >_<architecture>.bat* for indexing using the sample application.
- *bat* files named *RunTester_Query_<language >_<architecture>.bat* for executing queries using the sample application.

The appropriate *bat* file for indexing should be run first. It will create an index for the aforementioned files in an *index* folder, residing under the *filesToIndex<language>* folder. Then, the appropriate *bat* file is to be run. It will execute the queries on the created index and output an *xml* file named *HeBLuceneNet_<time stamp>.xml*, consisting of the results of executing the queries.

3.3. Enabling Indexing

Having performed the procedure(s) above, indexing is now available for using with Britannica Concept Search. To enable indexing, please follow the Lucene.Net documentation.

> If your project already has indexed documents, please remove and re-index them in order to make them available to Concept Search.

3.4. Performing Queries

Once your data has been indexed using Britannica Concept Search, you can now perform queries, once again please follow the Lucene.Net documentation.

Appendix A
Glossary

The following terms may be encountered in your use of this product:

Term	Definition
CLR	Stands for Common Language Runtime, the virtual machine component of Microsoft .Net framework. This component is responsible for compiling the intermediate language code ("bytecode") into native code. It also provides other services such as memory management.
Corpora	A set of texts that comprise a database on which a morphological search engine indexes and performs queries. Short for text corpus.
GAC	Stands for Global Assembly Cache, a .Net assemblies (i.e. *dll* or *exe* files) cache for Microsoft's CLR platform. Assemblies residing in the GAC must have strong name, which allows for side-by-side execution of different code versions. This prevents pitfalls like the situation of "*dll* hell".
Morphology	The structure and form of words in language, including inflection, derivation, and the formation of compounds.
Strong Key	See Strong Name.
Strong Name	A mechanism used in the Microsoft .Net framework to uniquely identify a component (e.g. *dll* file), as a measure against the situation of "*dll* hell", in which the existence of more than one component with the same naming but with different versions leads to many conflicts.
Thread	A way for a program to split itself into two or more simultaneously running tasks. Short for thread of execution.
Tokenizer	A tool that breaks text up into tokens.