PhUSE 2017

Paper SD07 - DRAFT

I Want it All and I Want it Now – Find What You Are Looking For… Fast!

Jörn Bilow, Entimo AG, Berlin, Germany

Abstract

Searching information in clinical data and metadata is a common, but often challenging task in pharmaceutical R&D. Traditional search mechanisms are usually based on metadata; like creation or modification date, author or type of submitted information, or context information like study ID, project or therapeutic area. In Entimo's Integrated Clinical Repository entimICE, an indexing server additionally enables search for content in documents and dataset files, including Word, Excel, PDF, CSV, tables and even SAS datasets. Based on cutting-edge search technologies like Elasticsearch, Lucene, TIKA and Akka, the tool enables researchers to look for information in the right context, based on the right content, and in a highly performant manner. The search tool uses a domain specific language (DSL) allowing researchers to use familiar terms when looking for information. Since its introduction into entimICE, various enhancements and additional capabilities have improved user experience with this tool.

1 Introduction

Since the release of the entimICE Indexing Server in 2016, new and powerful functionality is available to entimICE users. Pharmaceutical companies are utilizing it today to retrieve valuable information from their clinical data. Search and retrieval from the entimICE repository is possible for any kind of content. With the introduction of Elasticsearch as the underlying search engine, more complex searches are now possible, while availability and performance of the indexing server has improved significantly because of its load balancing and multi-tenant capabilities.

2 Indexing Server

The purpose of the Indexing Server is to provide access to any information in entimICE as fast as possible. This includes not only object metadata in the repository (e.g. names, properties, object metadata or status, comments and tags), but also attached content, such as text in PDF or Office files as well as tabular content in SAS and CSV files or in tables in the clinical repository.

2.1 Base Technology

The Indexing Server is based on rock solid and widely used open source technologies that are integrated in many enterprise search solutions. By using these technologies inside entimICE, it is possible to respect entimICE access rights (e.g. search results will never include objects not visible to the requestor in the repository tree) and to include information not accessible from outside.

The following main libraries are used in the Indexing Server:

- **Elasticsearch**: Based on Lucene, it is today's most popular enterprise search engine.
- **Apache Lucene**: Open-source information retrieval software library.
- **Apache Tika**: Content detection and analysis framework.

- **Akka**: A mature actor and concurrency library written in Scala which helps to ease the parallel and asynchronous implementation of server code.
- **Xtext**: An Eclipse-based set of tools to implement domain-specific languages (DSL) with an integrated editor that offers syntax highlighting and smart phrase selection support. It is used to provide a domain-specific search language.

The Indexing Server uses the concept of actors. Actors work in parallel and therefore utilize the resources of the underlying hardware in a more efficient way.

2.2 Index Information

Extracted information is written to the index initially and is updated automatically whenever an entimICE object or its attached content is changed.

The index is partitioned into so-called domains. In the standard configuration two domains are provided:

- The main domain for all information from the entimICE repository
- The import domain for all information from the monitored remote import directories

By configuration, more domains can be added.
Below are some examples of the information included in the index:

- Object attributes (e.g. name, comment, creation date)
- Object lifecycle information (e.g. object status)
- Object version attributes (e.g. last modification date or last change by user)
- Content of clinical data elements and structural metadata (e.g. SDTM, ADaM)
- File content (e.g. text, PDF, MS Office and numerous other popular file types)
- Content of programs, scripts, log files, listing files
- Program execution information
- Content of datasets and tables (e.g. database tables, SAS datasets, CSV datasets)
- Mappings

3 Software Demonstration

The entimICE Indexing Server demonstration will consist of several different scenarios based on typical queries against a clinical repository. It will make use of the Domain Specific Language (DSL) that translates search terms familiar to the clinical researcher into the underlying Lucene search language.

Queries and their results can be stored and reused within the entimICE environment. All search results take into account individual user access rights to the repository. Results can be stored, exported and shared with collaborating users or organizations.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Jörn Bilow
Entimo AG | Stralauer Platz 33-34 | 10243 Berlin | Germany
Email: bil [at] entimo.de
Work Phone: +49 30 520 024 100
Web: www.entimo.com