

PhUSE 2017

Indexing and
Searching a Clinical
Data Repository

Jörn Bilow

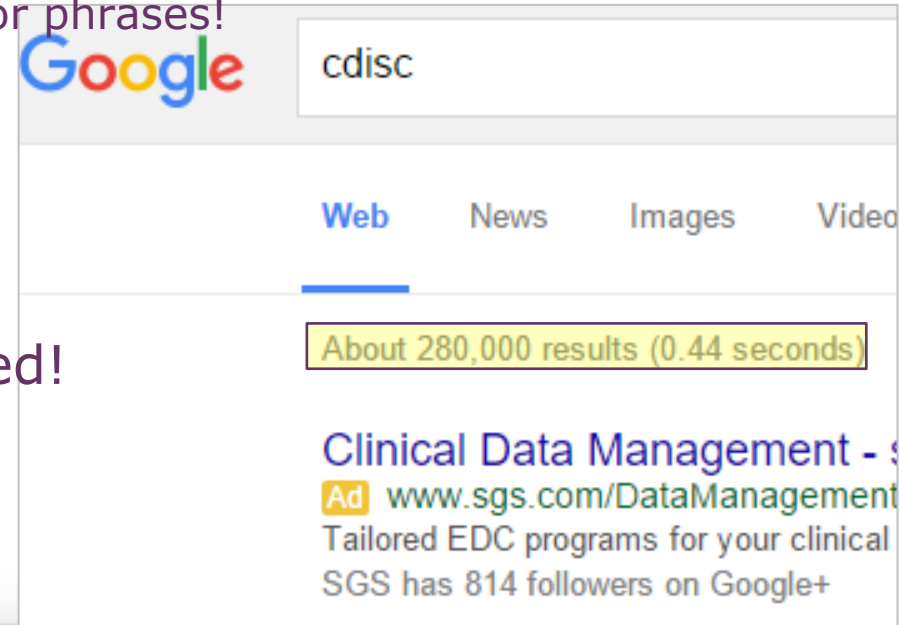
- [illegible]

- **“I want it all and I want it now”**

- Finding what you are looking for ... fast

- **Scope and Limits of Indexing Technology**

- Fast search results even for large amounts of data
- Text usually stored as words without stop-words
 - no exact search for sentences or phrases!
- Noise when getting results
 - often too many matches!
- Changes are not seen immediately!
- Appropriate resources needed!



■ Rock-solid open source libraries:

- Elasticsearch
- Apache Lucene
- Apache TIKA
- Akka
- Xtext for search language



■ Characteristics:

- Highly parallel indexing
- Multitenant-capable, "sharding" of indexes, load-sharing
- Various extractors (SAS, CSV, PDF, metadata) can work in parallel, different index domains can be written in parallel
- Client/batch process changes do trigger non-disturbing operations (no client impact) for background re-indexing on the server
- Data stored as "fuzzy" match, metadata as "exact" match

- **Indexing targets (configurable):**
 - All textual files (text and log files, also SAS programs)
 - PDF files, Excel sheets, XML files
 - SAS datasets, ClinRep datasets, CSV files as tables
 - Repository object metadata (e.g. variables, study metadata)
 - Process metadata (e.g. mappings and exec traces)
 - Links
- **Infrastructure requirements:**
 - Each change triggers a re-indexing process for the affected node(s) – on the server (no client delay)
 - Large datasets might take time to be indexed depending on the available resources
 - SAS server resources must be assigned to index SAS datasets
 - Direct access to entimICE Filesystem for Indexing Server process recommended
 - Clustering: Multitenant setup fully supported

- **Configuration options:**
 - Custom extractors (possible also with any SAS program)
 - Include/exclude sets
 - Multiple remote import directories
- **Remote import dialog supported:**
 - Check before import (on multiple directories)
- **Summary files:**
 - "icix" extension files contain custom-specific extracts and references to sources
- **All entimICE metadata included:**
 - Allows to combine data and metadata in queries for precise searches
- **Access rights management**
 - Search results according to user's role and group
 - No violation of governance, e.g. blinding rules

- **Demo of various searches:**
 - Text, data, metadata
 - Combine content and metadata search
 - Range searches with numbers and dates
 - Search for tables with certain columns
 - Search exec traces

- **Post-Processing search UI (“Find and filter”)**
 - Offer details about search and filter hits based on this (deselect hits)
- **Integration with other search engines**
 - Enterprise search can also find entimICE resources or entimICE has access to enterprise search
- **Support of SAS-License-Free Reading**
 - Commercial product S-JDBC may be used by customers for SAS-independent reading of SAS datasets

Questions and Answers

