PHARMA SEMANTIC SEARCH

CONNECTING REGULATORY INFORMATION TO INTERNAL R&D DATA VIA A KNOWLEDGE GRAPH

LSWT | 5.7.2022
Matthias Jurisch
BROX IT-SOLUTIONS
BACKGROUND AND GOAL
TECHNICAL CHALLENGES
SOLUTION AND IMPLEMENTATION
VALUE FOR THE CUSTOMER
brox IT-Solutions GmbH

IT-Consulting

• Founded: 26.11.1998
• CEO: Hans-Christian Brockmann
• Headquarter: An der Breiten Wiese 9 | 30625 Hannover
• Verticals: Automotive & Manufacturing, focused on Process- and IT-Consulting
• Research Partners: University Leipzig, DFKI
• Departments:
  - IT Sourcing Management
  - IT Architecture & Infrastructure
  - IT Lifecycle Management
  - Information Management
Our Customers
Background

Drug Development and Regulatory Process in the Pharma Domain

- Challenge: Regulatory documents and internal R&D data are not linked
  - E.g., internally a substance is referred to as “candidate 2917493” – not external name
  - Not linking the data can cause significant issues
Goal

Requirements

Integrate data from regulatory and R&D Domain to

- ensure data quality of submission documents
- getting information on which substances are registered in which countries
- directing research effort to areas that result in products

Provide frontend for exploring the data

- Users: no data-science/analytics background
- Use known UI metaphors
Technical Challenges

Integrating data from the R&D and regulatory domain

- Input: Text-mined documents
- Data cleansing required
- Matching to internal (RDF) master data on substances and legal entities
- Result needs to be integrated to other sources (-> knowledge graph)

Making the data available to non-technical users via a front-end

- Use an interaction pattern that was known to users - search engine
- Allow a faceted search over RDF data
Architecture

Implementation

Regulatory Submission Documents → Text-Mining → XLS-Files → Data Cleansing and Linking

Graph Database
- Organizational Graph Data
- Document Graph Data
- R&D Graph Data

Search Frontend
- Elastic Search
- Searchkit

Sparql Endpoint
Data Integration

Implementation

• ETL-Software was used to extract data from text mining results and create a graph.
• Matching text mining results via linking patterns created with SPARQL queries.
• Data was then ingested into an RDF database into a document graph, linked to the other graphs.
Search Frontend

Implementation

- Frontend building block: Elastic
- Ingestion:
  - Rdflib
  - Elastic library for python
- Frontend: searchkit
  - fast implementation
  - easy and accessible templates.
Value for the Customer

Risk and cost reduction, new streams of revenue

• Reduce Risk:
  • Discover Inconsistencies between regulatory and R&D data

• Cutting Costs:
  • Connecting products, substances, and legal entities that are allowed to sell them in the graph. Without the graph: hours or even days of manual work searching through documents.
  • Implementation done within weeks (a graph was already present for internal R&D data; tools like pentaho, graph databases, searchkit facilitate quick prototyping)

• New Streams of Revenue:
  • Regulatory data can be filtered by country, internal substance identifiers, related company, ... -> overview of the current market access of the company
Conclusion

What did we learn today?

• Data issues in the pharma regulatory process
• What do users in that domain need?
  • Data Integration
  • Simple UI
• How to do that?
  • ETL tools
  • Graph database
  • Data cleansing
  • Standard search solutions
• Why is this usefull?
  • Cost reduction: no manual information integration
  • Risk reduction: reduces regulatory risks
THANKS
FOR YOUR ATTENTION

Dr. Matthias Jurisch
Manager Information Management Unit

brox IT-Solutions GmbH
m: +49 173 3969952
e: mjurisch@brox.de