Text Mining in a Pharmaceutical Company

Carola Lefrank, Roche Innovation Center Basel

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

pRED Informatics
Pharma Research & Early Development

About F. Hoffmann – La Roche Ltd.

Roche develops products offering advances in the prevention, diagnosis and treatment of disease.
As the world's leading supplier of in vitro diagnostic systems and tests, we help guide clinical decision-making in every patient-care setting, from hospitals to patients' homes.
In the Pharmaceuticals Division we are focusing on areas of significant and pressing unmet medical need. Areas like cancer, where we lead the industry both in innovations and in market share.

> *Proud as we are of our past and present achievements, however, what really excites us is the future.*
>
> **Severin Schwan**

http://www.roche.com

pRED *Informatics*
Pharma Research & Early Development

About F. Hoffmann – La Roche Ltd.

- *Key Figures 2014*

  - Group Sales: 47.5 bCHF (Pharma: 36.7 bCHF, Dia: 10.8 bCHF)
  - Over 88'000 people working together across more than 150 countries
  - Our R & D Departments employ scientists with expertise in many disciplines, such as Chemistry, Biology, Physics, Pharmacology, Clinical Medicine and Informatics

pRED Informatics
Pharma Research & Early Development

Published Scientific Content Services

Our Mission

Our service ensures ***access to external scientific content*** *(eJournals, eBooks and external databases)* ***to Roche employees worldwide****. The scope is Science, Technology and Medicine, as well as selected Business Information.*

The main activities include ***review and optimization of the Scientific Content portfolio*** *on an ongoing basis and* ***evaluation of new products*** *in close collaboration with the Roche user community (this includes the product as well as the financial perspective).*

The service team manages all ***procurement activities*** *related to the acquisition of external information, including* ***negotiations*** *with vendors, as well as* ***backcharging*** *of these services within the Roche Group.*
In addition, the ST ensures the ***integration of full text links into our database products (RocheLink)****, provides access to non-licensed material via our* ***Document Delivery service****, organize* ***trainings for end user tools****, provides guidance for* ***copyright*** *compliance and ensures maintenance of the Roche central intranet access point for external scientific content (****info.roche.com****).*

pRED Informatics
Pharma Research & Early Development

Published Scientific Content Services Numbers

2014

Resources

- Ebooks: approx. 80'000
- Ejournals: approx. 15'000
- Databases: 48
- Vendors: 176
- Contracts: approx. 300

People

- **Users** accessing external content: approx. 13'500
- **Staff**

  Elibrary and VM: 4.7 FTE

  Scientific Documentation: 5.4 FTE

  Information Science: 8 FTE

pRED *Informatics*
Pharma Research & Early Development

Roche

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

pRED Informatics
Pharma Research & Early Development

The Relevance of Text Mining

- Exponentially increasing number of publications and complexity of topics results in growing workload
  - √ for **scientists** to keep track of their area of expertise and explore new areas screening large corpora in a fact-based manner
  - √ for **research libraries** to make content accessible in depth
  - √ Text-mining allows in-depth analysis of full-text content in areas where searches on abstract level are not sufficient (e.g. Pharmacovigilance)

- Text-mining is creating a new quality in content analysis by complementing formerly established methods with new features:
  - ✓ Improved retrieval
    - ➢ Reduce search time
    - ➢ Cast a wider net (federated search)
    - ➢ Fact retrieval in place of document retrieval
  - ✓ Integration

- Literature review using automated text-mining methods is more comprehensive and systematic

pRED Informatics
Pharma Research & Early Development

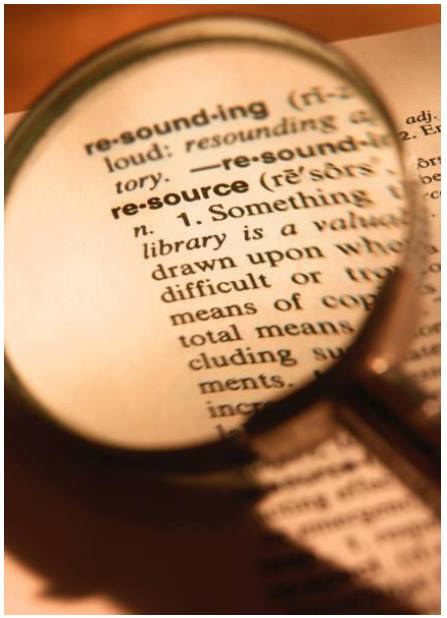
Definition

- **Textmining** describes the computer-based process of deriving relevant information from free texts by using a combination of linguistic, statistical, and machine learning techniques.

pRED Informatics
Pharma Research & Early Development

Information Sources

- **Scientific Articles (journals and books)**
  - ✓ Abstract, Fulltext
- **Patents**
  - √ WPO, EPO, USPTO and more
- **Data Feeds**
- **Competitive Intelligence Information**
  - ✓ Clinical Trial News
- **Databases**
  - √ Search Result Lists
- **Internal Information**
  - ✓ Document Management System, Intranet, Collaborative Spaces, Lab Notebooks et al.



Information Sources - Abstract vs Full Text

Abstract

- Easy access
- Clearly structured
- Easy processing

Fulltext

- Limited access/License needed
- Complete content
- Difficult processing
- Including diverse type of data

Roche Use Case

Linguamatics I2E

- Select text corpus for queries (also federated queries)
  - ✓ Complete database
  - ✓ Dataset based on database search
  - ✓ Internal data
- Run a query on I2E using pre-built queries or refine queries on-the-fly
- Receive a excel-like result list of extracted information:
  fact retrieval linking scientific entities by relationships
- Do an intellectual check of the result list and modify/ refine query if needed
- Provide a reference list with links to fulltext documents (if available) to customer

Roche

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

pRED *Informatics*
Pharma Research & Early Development

Text and Data Mining Clause for Pharma Industry

- Developed and released by ALPSP, PDR and STM Association in 2012
- Licensee may download, extract and index information from the Publisher's Content to which the Subscriber has access under this Subscription Agreement. Where required, mount, load and integrate the results on a server used for the Subscriber's text mining system and evaluate and interpret the TDM Output for access and use by Authorized Users. The Subscriber shall ensure compliance with Publisher's Usage policies, including security and technical access requirements. Text and data mining may be undertaken on either locally loaded Publisher Content or as mutually agreed.

The Association of Learned & Professional Society Publishers

STM The global voice of scholarly publishing

pRED Informatics
Pharma Research & Early Development

Challenges

- Clause needs to be implemented in multiple bilateral publisher contracts:
  - Number of contracts
  - Multiple clause versions
  - Diverse business models
  - Scepticsm against text and data mining for business reasons
- Lack of cross-publisher cooperation (e.g. unified XML format)
- Technical and financial limitations for smaller vendors

pRED Informatics
Pharma Research & Early Development

Roche

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

pRED *Informatics*
Pharma Research & Early Development

Commercial Service 1

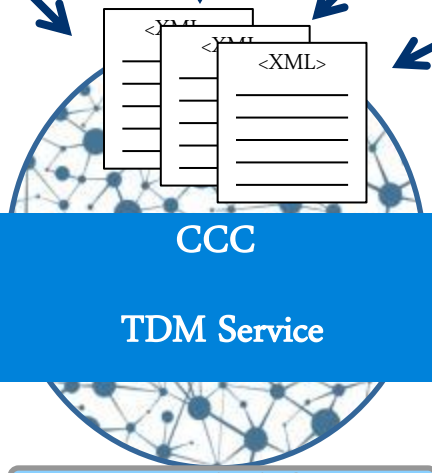
CCC - XML for Mining

- ✓ It is a growing platform of normalized full text data in XML format
- ✓ End user can create corpora including not-subscribed content
- ✓ Optional purchase functionality is available for the not-subscribed content
- ✓ Content is imported into text and data mining software
- ✓ CCC is in charge of copyright compliance
- ✓ The customer needs to have a CCC Copyright license in place
- ✓ The service is available for corporate customers

pRED Informatics
Pharma Research & Early Development

CCC - XML for Mining

1. Publishers provide content and rights



XML Article corpus

TDM Software

2. Companies create article sets using keywords or other criteria (including not-subscribed content)

3. Companies preview results including subscription status. Optional content purchase

4. Final results are imported into text mining tools

slide content provided by CCC

pRED Informatics
Pharma Research & Early Development

Commercial Service 2

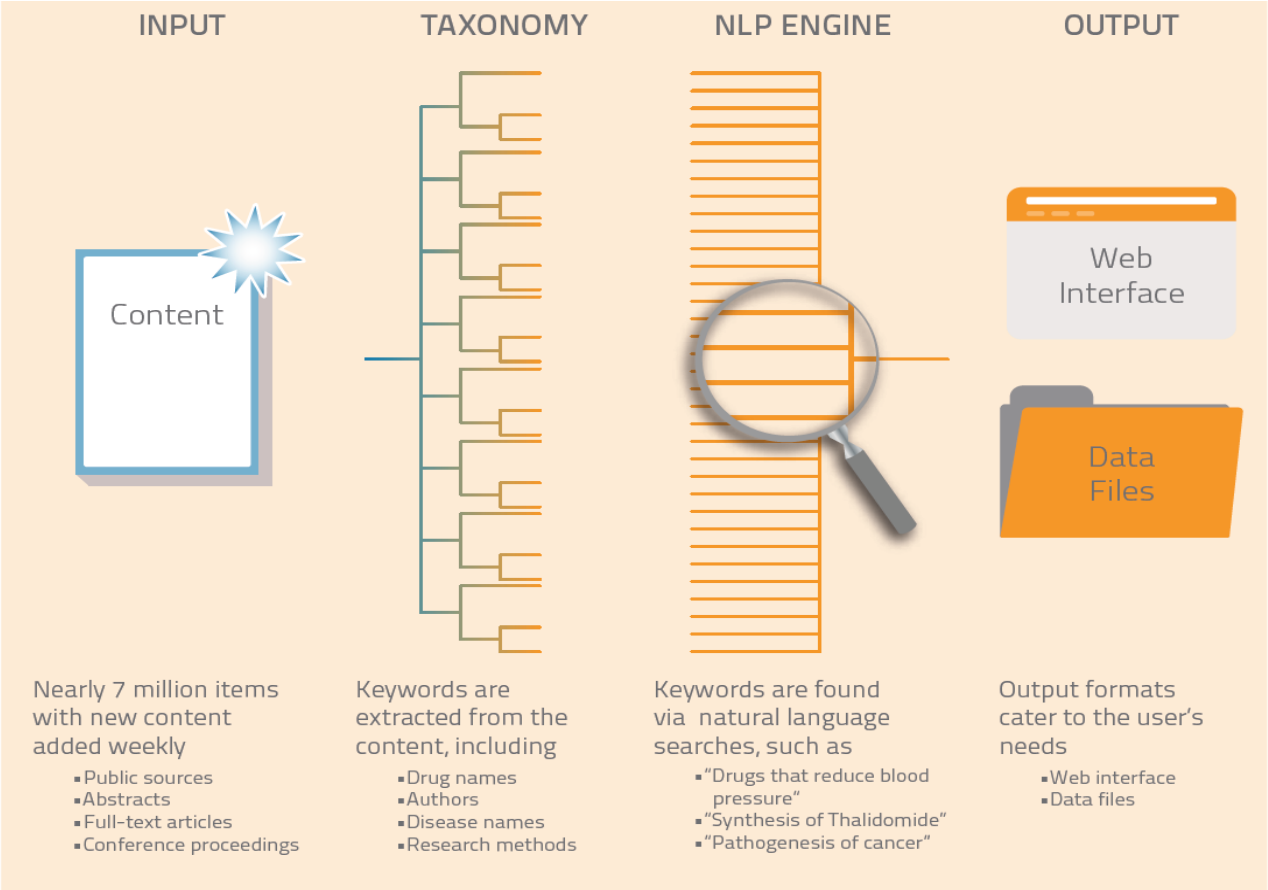
Elsevier - Text Mining for Life Sciences

- ✓ It is a growing text mining platform using abstracts, conference proceedings and full text journal articles
- ✓ Document input and output is available in multiple data formats
- ✓ Main focus is on 'interactions'
- ✓ API is under development
- ✓ Elsevier is interested in collaborating with academic institutions as well as with Pharma and Biotech

pRED Informatics
Pharma Research & Early Development

Commercial Service 2

Elsevier - Text mining Process



slide content provided by Elsevier

pRED Informatics
Pharma Research & Early Development

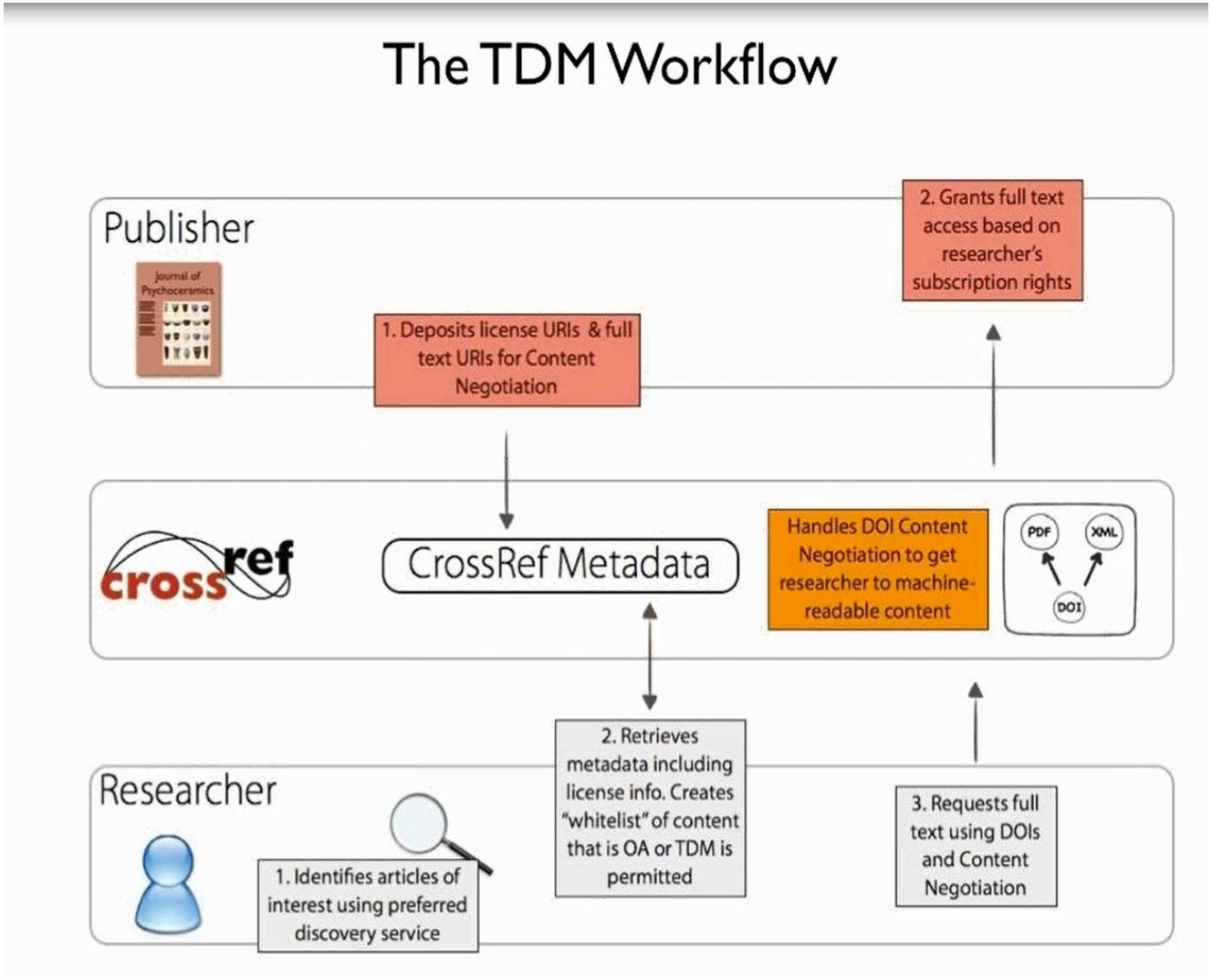
Non–Commercial Service

CrossRef

- ✓ Not-for-profit association of scholarly publishers
- ✓ CrossRef core service is DOI allocation and reference linking
- ✓ CrossRef text and data mining service:
  - DOI is the basis for a text and data mining API
  - The Cross Ref Metadata API is used to check (open access or subscribed) full text availability of content identified by CrossRef DOIs across publisher sites and regardless of the business model.
  - The API provides links to fulltext in multiple formats
  - Customer use their own tools to mine the content
- ✓ The service is currently free of charge

Non–Commercial Service

CrossRef



slide content provided by CrossRef

pRED Informatics
Pharma Research & Early Development

Roche

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

pRED *Informatics*
Pharma Research & Early Development

Summary

Our scientists need excellent laboratory, technology and **information support**

- √ Faster access to more content
- √ Better retrieval of relevant content
- √ More functionality
- √ More convenience
- √ Latest technology

All this can be improved with the help of Text Mining Technologies

pRED Informatics
Pharma Research & Early Development

Summary

We will need to manage potential constraints

- Financial
  - ✓ Budget
- Headcount
  - ✓ Number of available FTE's
  - ✓ Skills, Background
- Technical
  - ✓ System Requirements

pRED Informatics
Pharma Research & Early Development

Roche

About Us

Use of Text Mining

Negotiation Challenges

Commercial Services

Summary

Q & A

---

pRED Informatics
Pharma Research & Early Development

pRED Informatics
Pharma Research & Early Development

Roche

Doing now what patients need next

pRED *Informatics*
Pharma Research & Early Development