Annotating and Indexing Scientific Articles with Rare Diseases

Hosein Azarbonyad[1], Zubair Afzal[1], Max Dumoulin[1], Rik Iping[2], George Tsatsaronis[1]
1 Elsevier, The Netherlands
2 Erasmus MC, The Netherlands

Introduction

- In Europe 30 million people are suffering from a rare (or orphan) disease, a disease that occurs in less than 1 per 2,000 people
- Rare disease patients are entitled to the best possible health care, constituting the efficient organization of the respective clinical care and scientific literature imperative

Which are the excellence centers that could best treat patients for certain rare diseases?

Which are the key research initiatives for the various different rare diseases?

- Answering such questions requires **deep bibliometrical and scientometrical analysis** that can be based in the efficient **annotation and indexing** of the respective scientific literature
- We use a novel methodology based on SciBite's TERMite text annotation engine to annotate and index any scientific text with taxonomical concepts that describe rare diseases from the OrphaNet taxonomy

Objectives and Challenges

Main Objective

Map research outputs (articles in Scopus) to concepts in OrphaNet (Orphan Diseases taxonomy) to track research on rare diseases

Medical centers can use the output to:

- Show-case their research output to be recognized as an expert centre
- Develop, execute and evaluate research strategies into rare diseases with reliable evidence
- Analyze and evaluate research in rare diseases externally and internationally
- Get funding in the research areas (rare diseases) they are expert at
- Recruit, retain and promote talented researchers and faculty members

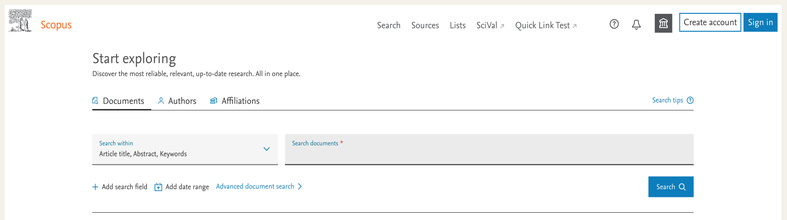
Challenges

- Some rare diseases are only rare in a specific part of the population
- Some of the rare diseases are very similar conceptually and their differences are very difficult to recognize especially in the context of a medical or clinical (scientific) articles
- The OrphaNet taxonomy, as any taxonomy, might be incomplete in certain areas, and its structure might not be homogeneous in granularity across all the parts of the taxonomy
- Polysemy and synonymy of the string surface appearance of rare diseases in text may still hinder the applicability of any annotation engine

Methods

Collection of Scientific Articles

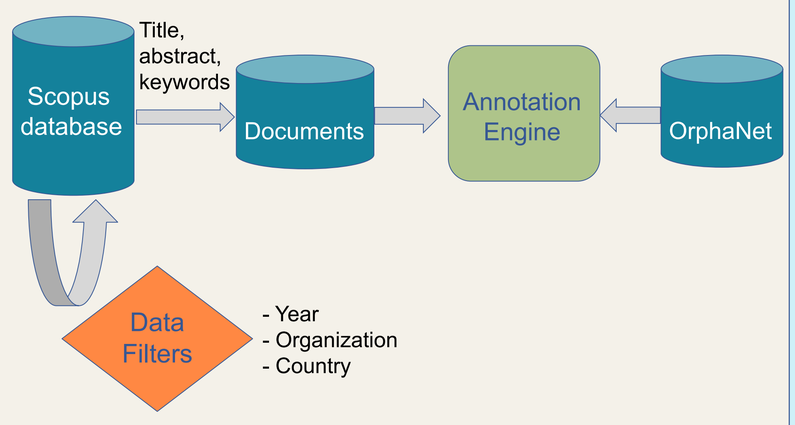
- Scopus database
- Publications published in the past 10 years
  - 36 million records



Rare Diseases Taxonomy

- A structured vocabulary for rare diseases capturing relationships between diseases, genes and other relevant features
- Created by French National Institute for Health and Medical Research in 1997
- 9,287 concepts organized in a hierarchy
- Updated monthly and follows the standard guidelines on deprecation of terms

Annotation Approach

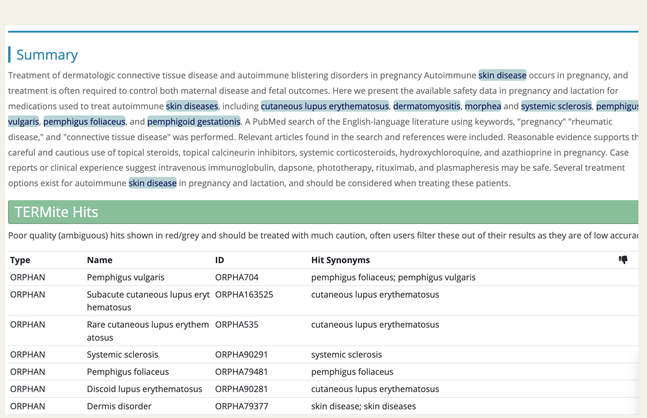


Annotation Engine

String matching
- Matching rare diseases names and synonyms with documents
- Simple regex matching of cleaned text and concepts

SciBite's TERMite annotation tool

- TERMite is an NER tool that rapidly scans and semantically annotates raw text (up to 1 million words per second) with entities from over 50 biopharma and biomedical topics



- Covers 98% of OrphaNet concepts
  - String matching is used to cover the remaining 2%
- Synonym search in combination of fuzzy matching
- Disambiguates detected entities for a document by means of:
  - Additional relevant ontologies such as gene-disease relations
  - Entity disambiguation techniques such as the relationship between detected entities
- All articles assigned to a code are propagated and assigned to the ansestor codes in the taxonomy tree

Results

Evaluation Strategy

- Scopus articles published in the Netherlands were indexed using TERMite
- For four big medical centers in the NL, indexed articles were manually sent to them to determine the winning engine by domain experts

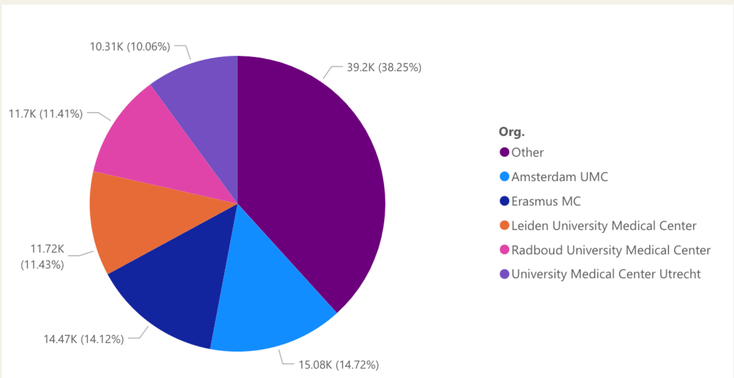
Medical Center	#wins SciBite	#wins String Matrching
Erasmum MC (N=276)	132	15
Leiden MC (N=73)	27	14
Utrecht MC (N=88)	72	2
Amsterdam MC (N=89)	14	26

- TERMite has a higher annotation performance compared to string matching
- For many articles, the codes assigned by TERMite and string matching are the same
  - Many Orpha concepts tend to be unique and specific which are detectable by string matching in articles

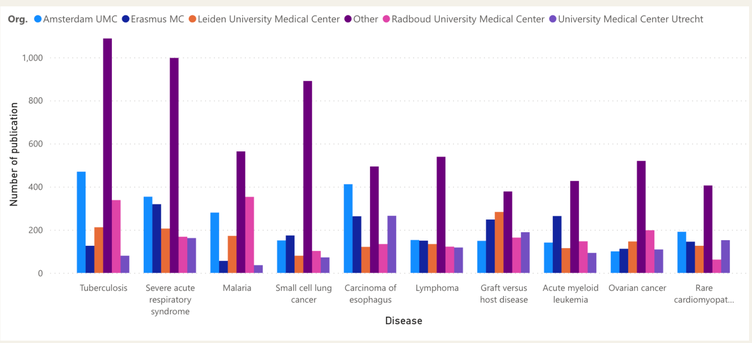
Statistics of Scopus publications on rare diseases

Metric	Number of pub.
Matched records for the Netherlands (Non-unique)	104,136
Matched records for the Netherlands (unique)	66,940
Matched records for EU (Non-unique)	1,048,423
Matched records for EU (unique)	672,162
Matched records for the world (Non-unique)	3,663,867
Matched records for the world (unique)	2,459,516

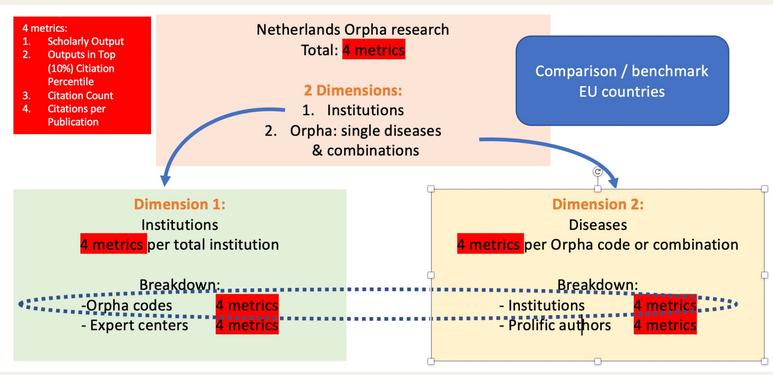
Most active research centers in rare diseases research in the NL in the past five years



Most researched diseases in the NL in the past five years



Sketch of a solution displaying Rare Disease research



Conclusions

- Compared to the existing engines for indexing rare diseases, TERMite has the highest coverage (98%) for the OrphaNet taxonomy
- TERMite can address some of these challenges associated with indexing articles with rare disease, in combination with advanced NLP and Text Mining techniques
- The combination of TERMite and Scopus results in a rich dataset of scientific articles indexed with rare disease
  - This can be the basis for bibliometrics analyses using the wealth of metadata and reference linking that Scopus provides

h.azarbonyad@elsevier.com