

# **Annotating and Indexing Scientific Articles**



# with Rare Diseases

Hosein Azarbonyad<sup>1</sup>, Zubair Afzal<sup>1</sup>, Max Dumoulin<sup>1</sup>, Rik Iping<sup>2</sup>, George Tsatsaronis<sup>1</sup>

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

Methods	Results		
<ul> <li>Collection of Scientific Articles</li> <li>Scopus database</li> <li>Publications published in the past 10 years</li> <li>36 million records</li> </ul>	<ul> <li>Evaluation Strategy</li> <li>Scopus articles published in the Netherlands were indexed using TERMite</li> <li>For four big medical centers in the NL, indexed articles</li> </ul>		
Scopus Search Sources Lists SciVal > Quick Link Test > (?) Create account Sign in Start exploring Discover the most reliable, relevant, up-to-date research. All in one place.	were manually sent to them to determine the winning engine by domain experts		
Documents     Authors     Affiliations     Search tips ()	Medical Center         #wins SciBite         #wins String Matrching		
Search within Search documents *	Erasmum MC (N=276) 132 15		
Article title, Abstract, Keywords	Leiden MC (N=73) 27 14		

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

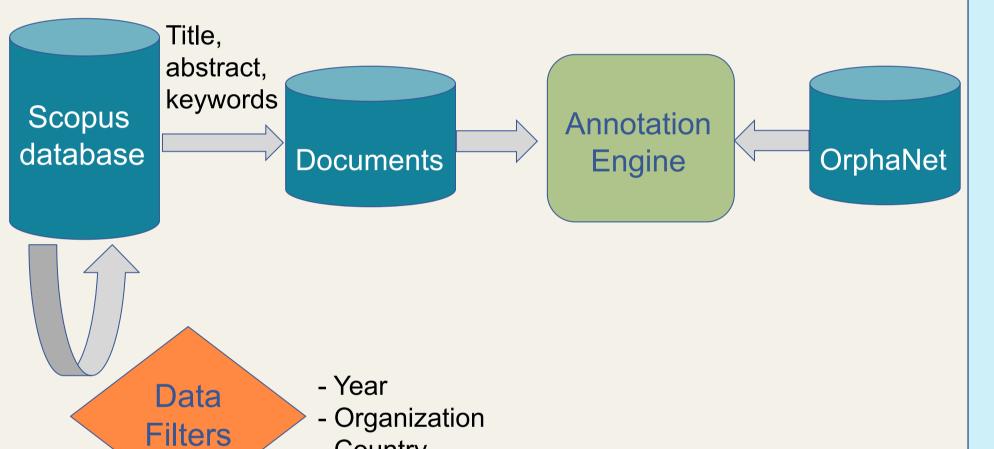
+ Add search field 拄 Add date range Advanced document search >

## **Rare Diseases Taxonomy**

orphanet

- 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**



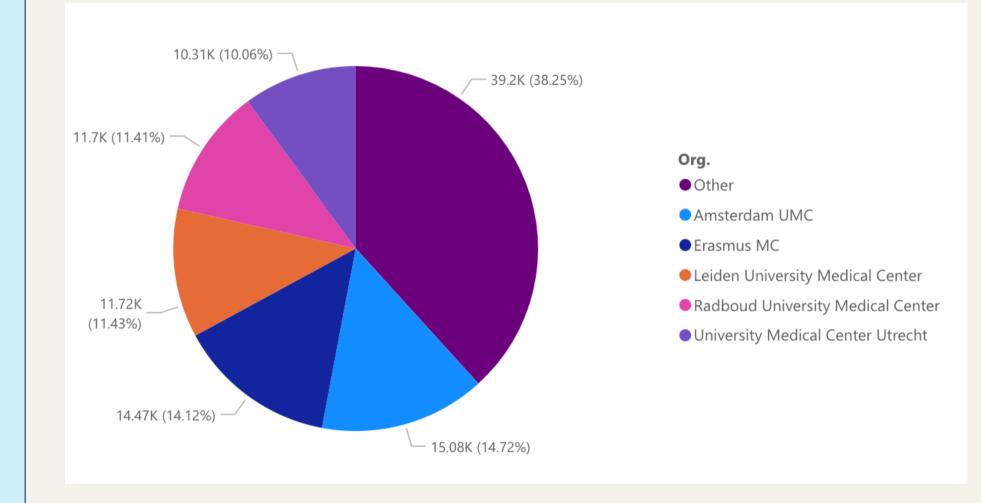
Utrecht MC (N=88)	72	2
Amsterdam MC (N=89)	14	26

- 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



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

## - Country

## **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

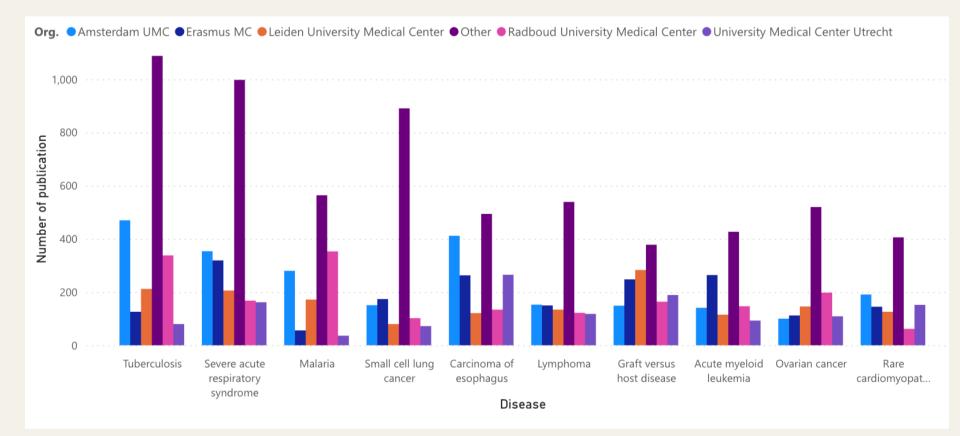
Summary

Treatment of dermatologic connective tissue disease and autoimmune blistering disorders in pregnancy Autoimmune skin disease occurs in pregnancy, and treatment is often required to control both maternal disease and fetal outcomes. Here we present the available safety data in pregnancy and lactation for medications used to treat autoimmune skin diseases, including cutaneous lupus erythematosus, dermatomyositis, morphea and systemic sclerosis, pemphigus vulgaris, pemphigus foliaceus, and pemphigoid gestationis. A PubMed search of the English-language literature using keywords, "pregnancy" "rheumatic disease," and "connective tissue disease" was performed. Relevant articles found in the search and references were included. Reasonable evidence supports the careful and cautious use of topical steroids, topical calcineurin inhibitors, systemic corticosteroids, hydroxychloroquine, and azathioprine in pregnancy. Case reports or clinical experience suggest intravenous immunoglobulin, dapsone, phototherapy, rituximab, and plasmapheresis may be safe. Several treatment options exist for autoimmune skin disease in pregnancy and lactation, and should be considered when treating these patients.

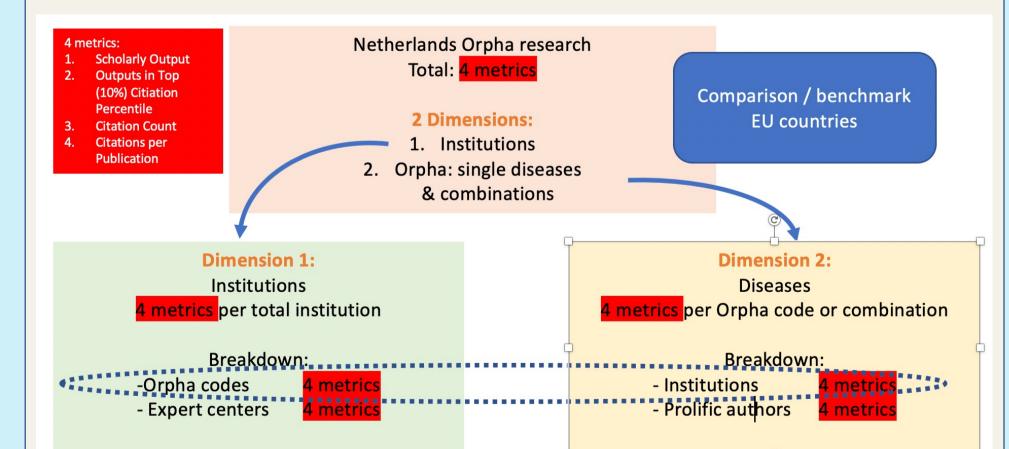
#### TERMite Hits

Poor quality (ambiguous) hits shown in red/grey and should be treated with much caution, often users filter these out of their results as they are of low ac				
Туре	Name	ID	Hit Synonyms	
ORPHAN	Pemphigus vulgaris	ORPHA704	pemphigus foliaceus; pemphigus vulgaris	
ORPHAN	Subacute cutaneous lupus eryt hematosus	ORPHA163525	cutaneous lupus erythematosus	
ORPHAN	Rare cutaneous lupus erythem atosus	ORPHA535	cutaneous lupus erythematosus	
ORPHAN	Systemic sclerosis	ORPHA90291	systemic sclerosis	
ORPHAN	Pemphigus foliaceus	ORPHA79481	pemphigus foliaceus	
ORPHAN	Discoid lupus erythematosus	ORPHA90281	cutaneous lupus erythematosus	
ORPHAN	Dermis disorder	ORPHA79377	skin disease; skin diseases	

## Most researched diseases in the NL in the past five years



## Sketch of a solution displaying Rare Disease research



## 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
- 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

## 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