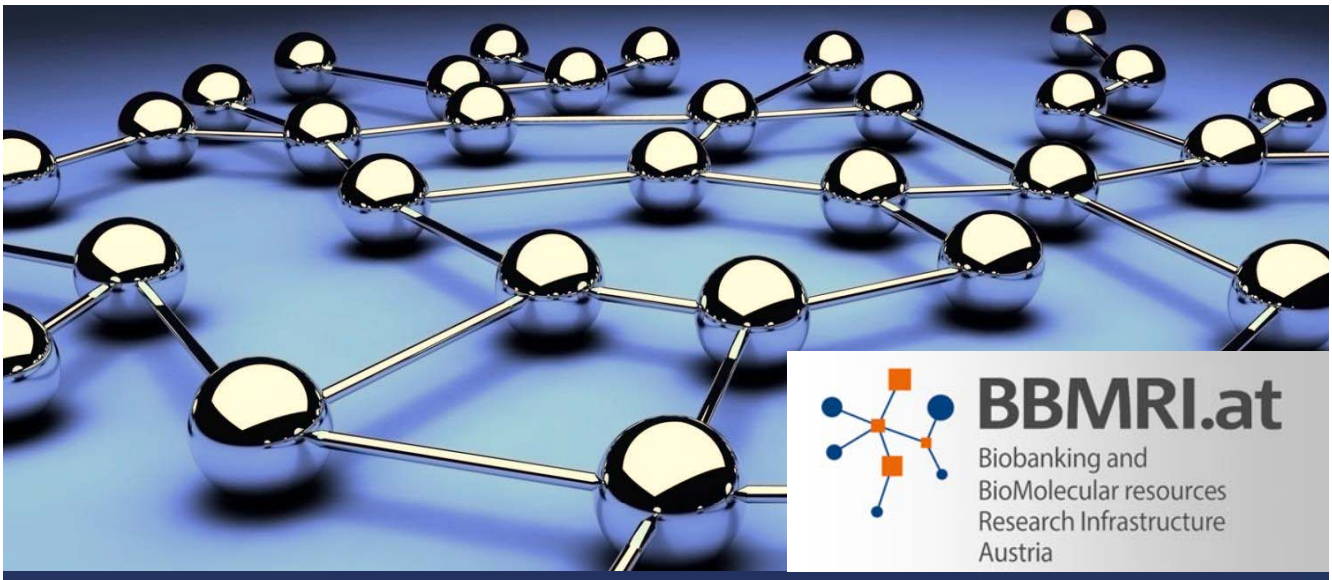


BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

# Optimizing Biobank Data Retrieval by Using IT-Based Lexical and Semantic Methods

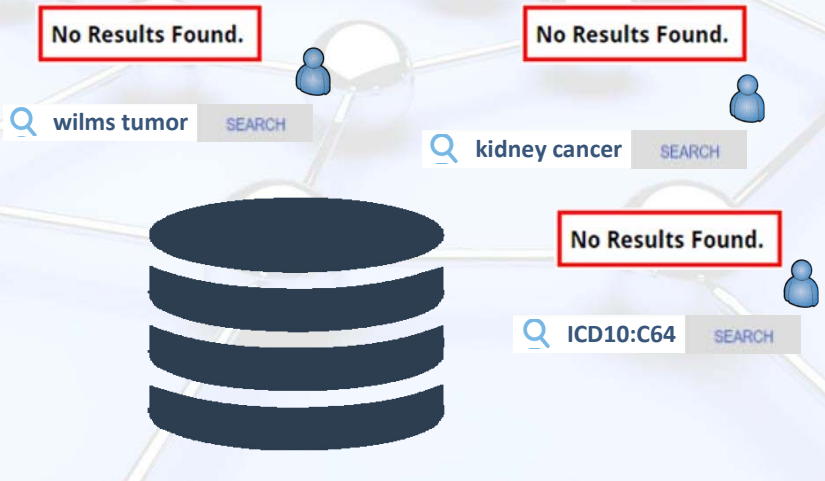
Philipp Hofer-Picout, Sabrina Neururer, Georg Göbel
philipp.hofer@i-med.ac.at

bmwfw
Funded by GZ 10.470/0016-II/3/2013

## Table of Contents

# Query Heterogeneity Problem

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria




**(i) Sample set**
Material type: FFPE, Frozen Tissue
Number of samples: 2
Diagnosis: **nephroblastoma**
...

*Biobank Information Management System

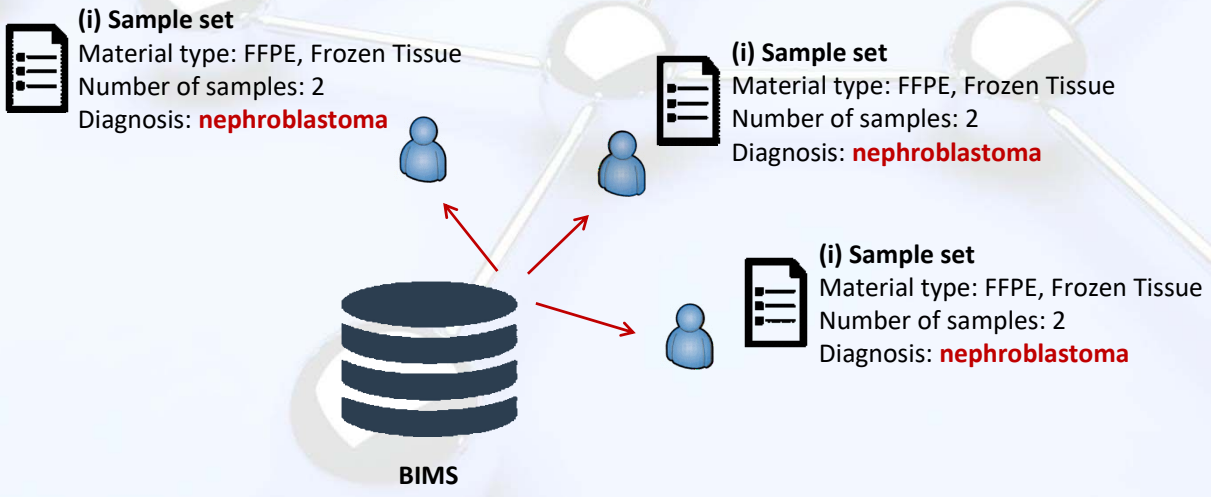




# Facilitate Data Accessibility

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

→ Optimize retrieval of desired information



# Main Steps

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

I. Identify & classify different query challenges[1,2]

↓

II. Find appropriate methods for all query scenarios

↓

III. Prototype implementation

↓

IV. Proof-of-concept

[1] Fidel, R. (1985). Moves in online searching. Online Review. https://doi.org/10.1108/eb024176
[2] Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. In Proceedings of the ASIST Annual Meeting. https://doi.org/10.1002/meet.14504701214

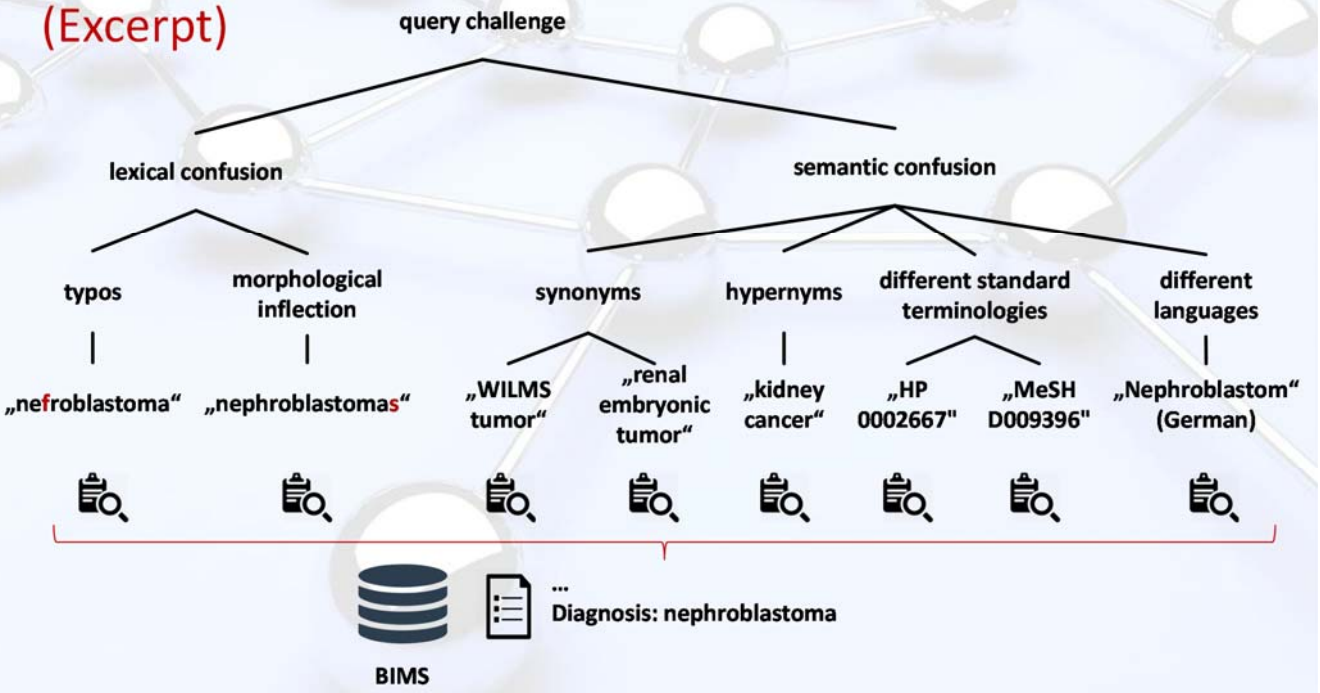




# I. Query Challenges

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

(Excerpt)



# II. Lexical & Semantic Methods

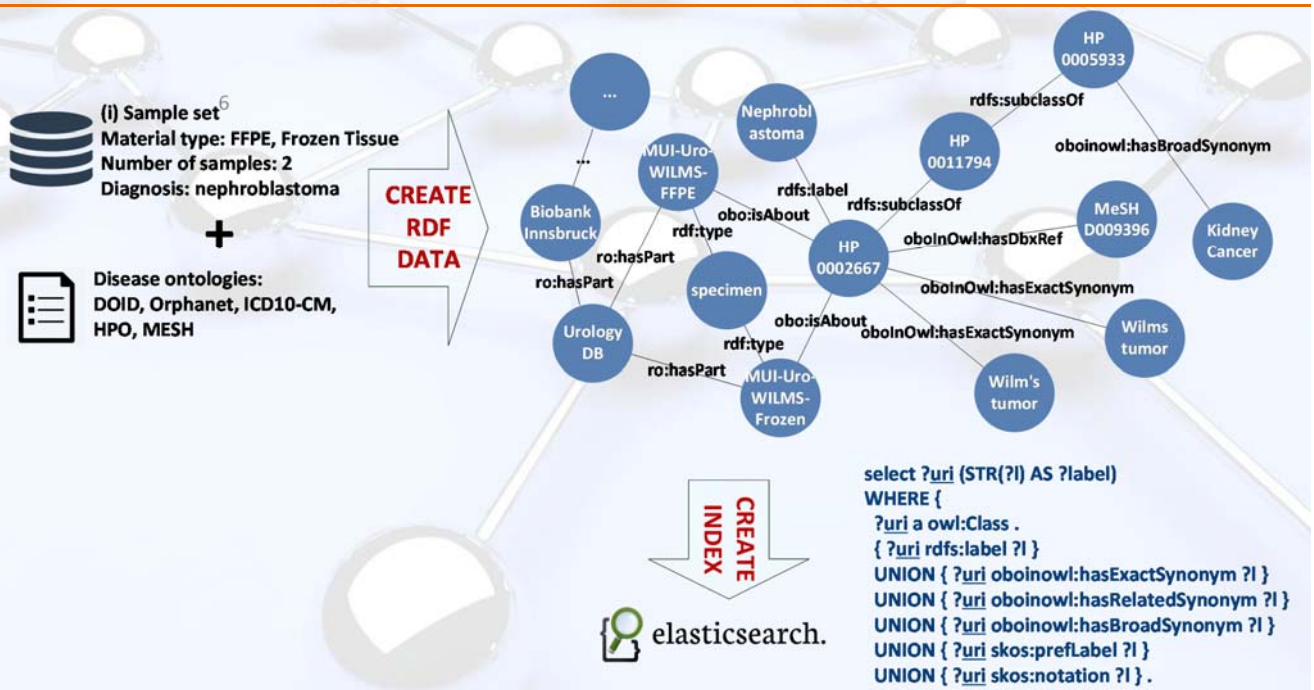
- Lexical confusion[3]
  - Vector Space Model
  - Phonetic Analysis/ Ngram/ Stemming (english)
  - Approximate String Matching (Levenstein distance)
- Semantic confusion
  - RDF Data Representation/ Graph Database[4]
  - (Bio-)Medical Standard Terminologies[5]

[3] M.S. Divya, S.K. Goyal, ElasticSearch: An advanced and quick search technique to handle voluminous data, International Journal of Advanced Computer Technology, 6 (2013), 171-175.
[4] https://www.blazegraph.com, last accessed on April 2018.
[5] Tirmizi SH, Aitken S, Moreira DA, Mungall C, Sequeda J, Shah NH, et al. Mapping between the OBO and OWL ontology languages. J Biomed Semantics. BioMed Central Ltd; 2011;2 (Suppl 1):S3.





# III. Prototype Implementation




[6] Hofer P, Fiegl H, Angerer J, Mueller- E. A Concept of a MIABIS based Register of Biosample Collections at the Medical University of Innsbruck. 2014;

# IV. Proof-Of-Concept

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

| Query class | Input String | Annotation used | Record found† | Applied method |
|---|---|---|---|---|
| misspelling | nefroblastom | HPO[a] 0002667 | Yes | Approximate string matching |
| Hypernym | Cancer of the kidneys | HPO 0002667 | Yes | hierarchical relationship |
| Synonym | Wilms tumor | HPO 0002667 | Yes | synonym relationship |
| coding standard | MeSH[b] D009396 | HPO 0002667 | Yes | cross reference HPO → MeSH |

†Expected record:

**(i) Sample set**
Material type: FFPE, Frozen Tissue
Number of samples: 2
Diagnosis: **nephroblastoma**

[a]Human Phenotype Ontology
[b]Medical Subject Headings





# Discussion & Outlook

BBMRI.at
Biobanking and BioMolecular resources Research Infrastructure Austria

- Both stemming and phonetic algorithms are more robust against typos and morphological variations
- Cross-references allow searching across different terminologies
- (Bio-)Medical standard terminologies must be valid regarding knowledge representation
- Combining different techniques would likely be the most rewarding
- A long-term evaluation about frequency of different query scenarios in productive environments will be necessary

# Literature


[1] Fidel, R. (1985). Moves in online searching. Online Review. https://doi.org/10.1108/eb024176

[2] Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. In Proceedings of the ASIST Annual Meeting. https://doi.org/10.1002/meet.14504701214

[3] M.S. Divya, S.K. Goyal, ElasticSearch: An advanced and quick search technique to handle voluminous data, International Journal of Advanced Computer Technology, 6 (2013), 171-175.

[4] https://www.blazegraph.com, last accessed on April 2018.

[5] Tirmizi SH, Aitken S, Moreira DA, Mungall C, Sequeda J, Shah NH, et al. Mapping between the OBO and OWL ontology languages. J Biomed Semantics. BioMed Central Ltd; 2011;2 (Suppl 1):S3.

[6] Hofer P, Fiegl H, Angerer J, Mueller- E. A Concept of a MIABIS based Register of Biosample Collections at the Medical University of Innsbruck. 2014;



Funded by bmwfw Bundesministerium für Wissenschaft, Forschung und Wirtschaft




# Thank you for your attention !


**Georg Goebel**
Principal Investigator
Local biobank coordinator
Biobank-IT contact
Medical Statistics, Informatics and Health Economics
georg.goebel@i-med.ac.at



**Sabrina Neururer**
Local project coordination
Medical Statistics, Informatics and Health Economics
sabrina.neururer@i-med.ac.at



**Philipp Hofer-Picout**
Researcher
Medical Statistics, Informatics and Health Economics
philipp.hofer@i-med.ac.at

**Contact**
biobank@i-med.ac.at
Tel. +43 512 9003 70912
Fax +43 512 9003 73922

**BBRMI.at**
Neue Stiftingtalstraße 2/B/6
A-8010 Graz
Austria
www.bbmri.at



Funded by bmwfw Bundesministerium für Wissenschaft, Forschung und Wirtschaft