System Description: Reaching Deeper into the Life Science Bibliome with CORAAL

Vít Nováček, Tudor Groza, Siegfried Handschuh

DERI, National University of Ireland, Galway
IDA Business Park, Galway, Ireland
e-mail: FirstName.LastName@deri.org

Abstract. Despite being a flourishing field, regarding search, the contemporary online scientific publishing properly exploits mostly raw publication data (bags of words) and shallow meta-data (authors, key words, citations, etc.). The much needed economical mass exploitation of the knowledge implicitly contained in publication texts is still largely an uncharted territory. Within our long-term ambition to quell the lions there, we have made the first step with the CORAAL prototype presented in this report. The tool essentially *extracts* asserted publication meta-data together with the knowledge implicitly present in the respective text, *integrates* the emergent content and *exposes* it via a multiple-perspective search&browse interface.

1 Introduction

Digital content processing has no doubt introduced a whole lot of new possibilities of dealing with scientific publications. It makes knowledge much more open and exploitable than in the old "paper times". However, one still needs to go manually through a lot of possibly irrelevant content very often before actually finding the right answers. If we are to make the next step, it is necessary to process knowledge (i.e., concepts and their mutual relations), and not just data or shallow meta-data (i.e., chunks of free text, titles or authors).

Substantial automation of such meaning-intensive information processing is hardly possible with the current industry-strength technologies (e.g., full-text search), since they lack proper support for extraction, representation and processing of knowledge implicitly present in texts. As an illustration, imagine for instance finding a support of the claim that *acute granulocytic leukemia* is different from *T-cell leukemia*. With the current solutions, it is easy to find articles that contain both or either of the terms, however, the number of results may be quite high (e.g., 593 on PubMed). It is tedious or even impossible to go through all of them in order to find out which of them actually mention the two leukemias being different.

Methods for automated knowledge extraction than can dig more than mere key words from text exist, however, their results are deemed to be to too noisy and sparse to be exploited by the current state of the art without significant manual post-processing [1]. We have recently researched a novel framework for effortless exploitation of automatically extracted knowledge that makes use of similarity-based knowledge representation and respective light-weight inference services [2]. We combined the framework with our repository for semantically inter-linked publications [3], delivering a prototype knowledge-based publication search engine – CORAAL (*COntent extended by emeRgent and Asserted Annotations of Linked publication data*). The tool essentially *extracts* asserted publication meta-data together with the knowledge implicitly present in the respective text, *integrates* the emergent content with existing domain knowledge and *exposes* it via a multiple-perspective search&browse interface. This way we allow for fine-grained publication search combined with convenient and effortless large scale exploitation of the knowledge that is associated with and hidden in the publication texts.

1.1 Motivating Use Case

The decision to apply CORAAL to the life sciences domain was motivated by our cooperation with clinicians and researchers at Masaryk Oncology Institute in Brno, Czech Republic. When figuring out how to apply our text mining and light-weight knowledge representation expertise to their needs, we articulated a particular "translational biomedical knowledge integration" use case. Essentially, the oncologists were interested in how the information in patient records links to relevant statements in the biomedical literature, and vice versa. The critical requirements are, firstly, lack or minimisation of manual efforts necessary for the knowledge extraction and integration, and, secondly, efficient, exhaustive and intuitive querying of large amounts of such integrated knowledge. CORAAL presents a first step towards a practical realisation of the use case, tackling the extraction, integration and exploitation of the publication knowledge.

1.2 Related Work Overview

The state-of-the-art applications like ScienceDirect or PubMed Central require almost no effort in order to expose arbitrary life science publications for search. However, the benefit they provide is rather limited when compared to cutting-edge approaches aimed at utilising also the publication knowledge within the query construction and/or result visualisation. Such innovative solutions may require much more a priori effort in order to work properly, though.

GoPubMed [4] is an ontology-based front-end to a traditional publication full-text search. It allows for effective restriction and intelligent visualisation of the query results. GoPubMed dynamically extracts parts of the Gene Ontology (cf. `http://www.geneontology.org/`) relevant to the query, which are then used for restriction and a sophisticated visualisation of the classical PubMed search results. Nevertheless, GoPubMed does not offer querying for or browsing of arbitrary publication knowledge – terms and relations not present in the system's rather static basal ontology simply cannot be reflected in the search.

Textpresso [5] enables searching for relations between concepts in particular chunks of text (namely for gene-to-gene interactions). The tools described in [6, 7] support complex ontology-based querying for relationships identified in free texts (related to dengue literature and lipidospehere, respectively). However, the underlying ontologies have to be provided manually for all the above systems. Moreover, their scale regarding the number of publications' full-texts and concepts covered is quite low.

Regarding more recent and/or scalable tools for expressive search in biomedical literature, EBIMed [8], novo|seek (`http://www.novoseek.com`), iHOP [9] and u-Compare [10] are among the most relevant. EBIMed performs classical full-text search and annotates the results with possible associations between co-occurring protein, gene, drug or species names. novo|seek provides for synonymy resolution and disambiguation of the search terms, in effect extending the coverage and precision of the classical full-text search. iHOP allows for browsing publications along links induced from occurrence of same gene or protein names across different resources. Finally, u-Compare allows for drag-and-drop style building of custom NLP workflows in order to analyse life science publications. All the tools focus mostly on the extraction and/or annotation of entities (and possibly also their relationships), however, they do not exploit any underlying knowledge representation allowing for, e.g., reasoning (which is essential for sophisticated querying).

From the overview of the state of the art in the field, it is obvious that the two biggest challenges are: 1. reliable automation of more expressive content acquisition; 2. representation of the acquired content allowing for exhaustive and expressive querying. None of the related systems addresses both problems appropriately. On the other hand, CORAAL, the system we present in this paper, can easily employ legacy domain resources or even work without any human intervention. Apart of full-text indexing, the extracted statements are integrated into a light-weight knowledge base with similarity-based inference capabilities that allow for refinement, augmentation and querying of the stored facts. In effect, CORAAL delivers a comprehensive combination of full-text and knowledge-based search in life science publications that is to large extent complementary to the related state of the art.

1.3 Structure of the Paper

The rest of the paper is organised as follows. Section 2 presents the essentials of the processing pipeline in CORAAL and outlines the relevant technological aspects of the tool. The deployment and user perspectives of CORAAL are covered in Section 3. We summarise the delivered work and outline future directions in Section 4.

2 CORAAL Essentials

In order to provide comprehensive search capabilities in CORAAL, we decided to complement a standard (full-text) publication search approach with advanced services catering for knowledge-based search. By knowledge-based search we mean querying for and browsing of expressive statements capturing relations between concepts in the respective source articles. CORAAL is built on the top of two substantial research outputs of our group at DERI – the KONNEX [3] and EUREEKA [2] frameworks. The former is used for storing and querying of publication full-text and meta-data. The latter serves for exploitation of the knowledge (i.e., concepts and their mutual relations) implicitly contained in the publication texts by means of knowledge-based search.

2.1 Data Processing Outline

EUREEKA provides for knowledge extraction from text and other knowledge resources (e.g., ontologies or machine readable thesauri). It employs two data structures. Firstly, a lexicon takes care of mapping extracted lexical labels to unique entity identifiers and vice versa (i.e., dealing with synonymy resolution and disambiguation). The extraction process continually updates the second data structure in EUREEKA – a light-weight knowledge base supporting arbitrary extensions by custom rules, as well as similarity-based reasoning and querying of the stored data. The knowledge bases are exposed to consumers via a query answering interface. The interface allows for evaluating conjunctive argument-relation-argument queries with negation and variables. It returns extracted and/or inferred statements relevant to the query, sorted according to their intrinsic relevance degree (derived from all particular relationships in the knowledge base using a generalised IR ranking algorithm). Theoretical and technical details of EUREEKA and its deployment in CORAAL are provided in [2].

KONNEX tackles the integration of the extracted publication text and metadata, represented as RDF graphs in a triple store. Operations related to data registration (inclusion and integration with the stored content), repository maintenance, full-text query processing and indices are handled by respective manager KONNEX modules, possibly composed of sub-modules handling particular data or query types. Details are given in [3].

2.2 Technological Aspects

There are several conceptually separate modules in CORAAL, moreover, EUREEKA is written in the Python programming language, while KONNEX in Java. Therefore we utilise an inter-process communication layer implemented using the D-BUS framework (cf. `http://en.wikipedia.org/wiki/D-Bus`). On the top of the core-level EUREEKA and KONNEX APIs, a set of helper web services rests. These manage the user requests and forward the data returned by the core APIs to the web hub, which is a set of Java servlets handling particular types of search. The servlets produce machine-readable RDF representing answers to user queries. The RDF has XSL style sheets attached in order to render the results in a human-readable form via the Exhibit faceted browsing web front-end (cf. `http://www.simile-widgets.org/exhibit/`). Such a solution results in CORAAL being a pure Semantic Web application, as the data-flow between the core infrastructure and the other modules is strictly based on RDF graphs.

While being presented in a human-readable form in the browser, the produced data can be directly analyzed by an application or fetched by a crawler.

3 Deploying the System

3.1 Processed Data

We have processed 11,761 Elsevier journal articles from the provided XML repositories that were related to cancer research and treatment. The access to the articles was kindly provided within the Elsevier Grand Challenge competition we participated in (cf. `http://www.elseviergrandchallenge.com`). The domain was selected so due to the expertise of our sample users and testers from Masaryk Oncology Institute in Brno, Czech Republic. We processed articles evenly distributed across journals dedicated to the following topics: oncology, genetics, pharmacology, biochemistry, general biology, cell research, and clinical medicine. From the article repository, we extracted the knowledge and publication metadata for further processing by CORAAL. Besides the publications themselves, we employed legacy machine-readable vocabularies for the refinement and extension of the extracted knowledge (currently, we use the NCI and EMTREE thesauri – see `http://www.cancer.gov/cancertopics/terminologyresources` and `http://www.embase.com/emtree/`, respectively).

CORAAL exposes two data-sets as an output of the publication processing: 1. We used a **triple store** containing publication meta-data (citations, their contexts, structural annotations, titles, authors and affiliations) associated with respective full-text indices. The resulting store contained $7,608,532$ of RDF subject-predicate-object statements describing the input articles. This included $247,392$ publication titles and $374,553$ authors (both from full-texts and references processed). 2. We employed a custom EUREEKA **knowledge base** with facts of various certainty extracted and inferred from the article texts and the seed life science thesauri. Directly from the articles, $215,645$ concepts were extracted (and analogically extended). Together with the data from the initial thesauri, the domain lexicon contained $622,611$ terms, referring to $347,613$ unique concepts. The size of the emergent knowledge base was $4,715,992$ weighed statements (ca. 99 and 334 extracted and inferred statements per publication in average, respectively). This number is significantly smaller than in the case of the semifinal prototype. However, this is due to a full integration of the knowledge from formerly separate contexts, the data themselves are still the same. The contextual meta-knowledge related to the statements (like provenance information) amounts to more than $10,000,000$ additional statements should it be expressed in RDF triples.

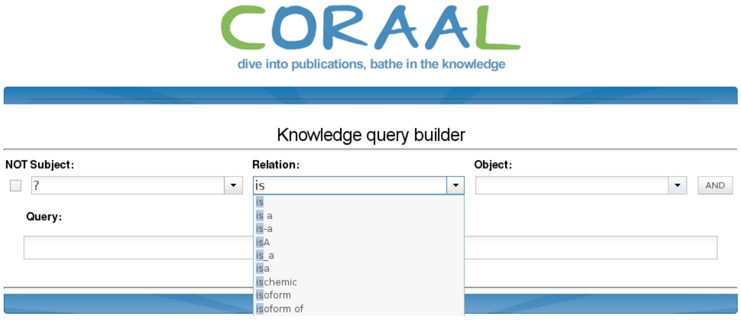
3.2 Accessing and Using CORAAL

CORAAL can be accessed at `http://coraal.deri.ie:8080/coraal/`. The following browsers have been tested with CORAAL and are known to work on most desktop configurations and operating systems: (1) Firefox (versions 2.x, 3.x and newer); (2) Internet Explorer (versions 7.x and newer; in most cases only on Windows Vista, though); (3) Opera (versions 9.6 and newer); (4) Safari (versions 3.1 and newer); (5) Google Chrome (all versions).

For the user interface of our system, we employed the MIT's state-of-the-art Exhibit framework (cf. `http://www.simile-widgets.org/exhibit/`). It supports faceted browsing (cf. `http://en.wikipedia.org/wiki/Faceted_browser`) of the knowledge-based search results, letting users to conveniently focus on the relevant answers. Similarly, we allow for faceted browsing of the classical full-text search results that are tightly integrated with the knowledge extracted or inferred from the respective articles. Both types of search (knowledge-based and full-text) are illustrated in the rest of the section.

Knowledge-Based Search We expose the content of the CORAAL knowledge base via a query-answering module. It returns answer statements sorted according to their relevance scores and similarity to the query (see [2] for the details of the process). Answers are provided by an intersection of publication provenance sets corresponding to the respective statements' subject and object terms. The module currently supports queries in the following form: $t \mid s : (NOT\ )?p : o(\ AND\ s : (NOT\ )?p : o)^*$, where NOT and AND stands for negation and conjunction, respectively. s, o, p may be either variable—anything starting with the `?` character or even the `?` character alone—or a lexical expression. t may be lexical expressions only. The ? and * wildcards mean zero or one and zero or more occurrences of the preceding symbols, respectively, $|$ stands for or. Only one variable name is currently allowed to appear within a single query statement and across a statement conjunction.

Statement queries can be asked either directly in the CORAAL *Knowledge* tab, or one can use an interface for assisted query construction with auto-completion capability, as seen in Figure 1. The auto-completion is context-



Knowledge query builder

NOT Subject: □ ?

Relation: is

Object:

Query:

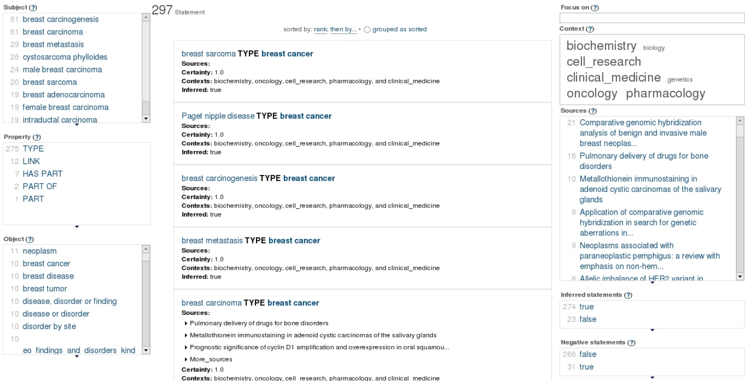
Fig. 1. Asking a query – assisted

sensitive – if you settle, e.g., on a subject, only relations (properties) actually associated with the particular subjects in the knowledge base are offered. Negative or conjunctive statements are created using the *NOT*, *AND* check-box or button, respectively. This allows for very convenient query construction even for users who are not sure how exactly should they ask the system.

The knowledge search results (Figure 2) are displayed as particular statements provided with several types of meta-information:

- *source* provenance – articles relevant to the statement
- *context* provenance – sub-domain of life sciences the statement relates to (determined according to the main topic of the journal that contained the articles the statement was extracted from)
- *certainty* – how certain the system is that the statement holds and is relevant to the query (values between 0 and 1)
- *inferred* – whether the statement was inferred or not (i.e., directly extracted)

One can filter the answer statements based on their particular elements (subjects, properties and objects), associated meta-information and their negativity. The filtering boxes have a contextual help associated with them – after pressing the (?) symbol, an inline hint on the usage of the respective box is opened. Using a particular filtering, one can conveniently focus only on statements of interest. Article provenance summaries of particular statements can be displayed in-line as shown in Figure 3.



Subject (?)

61	breast carcinogenesis
61	breast carcinoma
29	breast metastasis
26	cystosarcoma phylloides
24	male breast carcinoma
20	breast sarcoma
19	breast adenocarcinoma
19	female breast carcinoma
19	intraductal carcinoma

Property (?)

275	TYPE
12	LINK
7	HAS PART
2	PART OF
1	PART

Object (?)

11	neoplasm
10	breast cancer
10	breast disease
10	breast tumor
10	disease, disorder or finding
10	disease or disorder
10	disorder by site
10	eo findings and disorders kind

297 Statement

sorted by: rank; then by... • ○ grouped as sorted

breast sarcoma TYPE **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

Paget nipple disease TYPE **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast carcinogenesis TYPE **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast metastasis TYPE **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast carcinoma TYPE **breast cancer**
Sources:
▸ Pulmonary delivery of drugs for bone disorders
▸ Metallothionein immunostaining in adenoid cystic carcinomas of the salivary glands
▸ Prognostic significance of cyclin D1 amplification and overexpression in oral squamou...
▸ More_sources
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine

Focus on (?)

Context (?)

biochemistry biology
cell_research
clinical_medicine genetics
oncology pharmacology

Sources (?)

21	Comparative genomic hybridization analysis of benign and invasive male breast neoplas...
16	Pulmonary delivery of drugs for bone disorders
10	Metallothionein immunostaining in adenoid cystic carcinomas of the salivary glands
9	Application of comparative genomic hybridization in search for genetic aberrations in...
9	Neoplasms associated with paraneoplastic pemphigus: a review with emphasis on non-hem...
8	Allelic imbalance of HER2 variant in

Inferred statements (?)

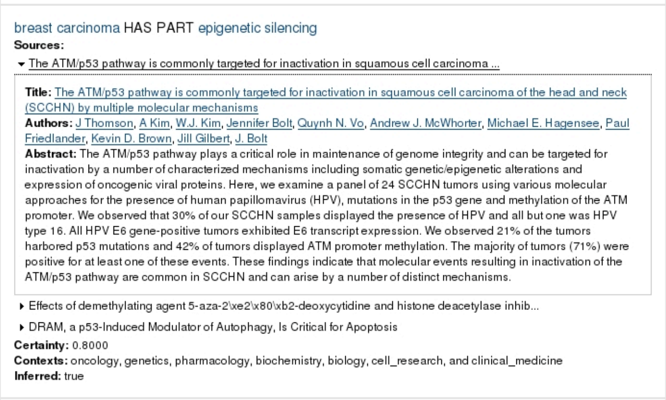
274	true
23	false

Negative statements (?)

266	false
31	true

Fig. 2. Answer display – full

After clicking on the concept names, statements related to it are displayed (similarly to the previously shown answer display). Additionally, we provide a box with further information on the concept (Figure 4). Currently we automatically fetch an abstract, image and biomedical classification details from the respective Wikipedia entries (looking up the synonyms of the concept in



breast carcinoma HAS PART epigenetic silencing

Sources:

▾ The ATM/p53 pathway is commonly targeted for inactivation in squamous cell carcinoma ...

Title: The ATM/p53 pathway is commonly targeted for inactivation in squamous cell carcinoma of the head and neck (SCCHN) by multiple molecular mechanisms

Authors: J Thomson, A Kim, W.J. Kim, Jennifer Bolt, Quynh N. Vo, Andrew J. McWhorter, Michael E. Hagensee, Paul Friedlander, Kevin D. Brown, Jill Gilbert, J. Bolt

Abstract: The ATM/p53 pathway plays a critical role in maintenance of genome integrity and can be targeted for inactivation by a number of characterized mechanisms including somatic genetic/epigenetic alterations and expression of oncogenic viral proteins. Here, we examine a panel of 24 SCCHN tumors using various molecular approaches for the presence of human papillomavirus (HPV), mutations in the p53 gene and methylation of the ATM promoter. We observed that 30% of our SCCHN samples displayed the presence of HPV and all but one was HPV type 16. All HPV E6 gene-positive tumors exhibited E6 transcript expression. We observed 21% of the tumors harbored p53 mutations and 42% of tumors displayed ATM promoter methylation. The majority of tumors (71%) were positive for at least one of these events. These findings indicate that molecular events resulting in inactivation of the ATM/p53 pathway are common in SCCHN and can arise by a number of distinct mechanisms.

▸ Effects of demethylating agent 5-aza-2\xe2\x80\xb2-deoxycytidine and histone deacetylase inhib...

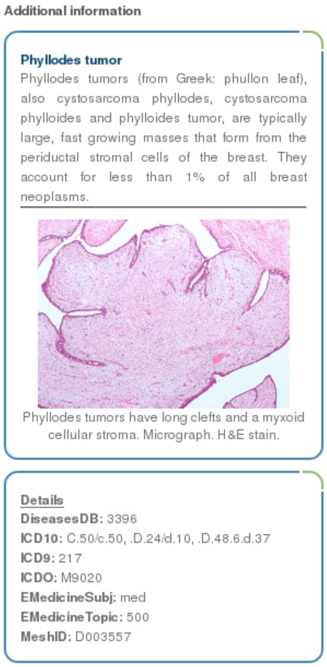
▸ DRAM, a p53-Induced Modulator of Autophagy, Is Critical for Apoptosis

Certainty: 0.8000

Contexts: oncology, genetics, pharmacology, biochemistry, biology, cell_research, and clinical_medicine

Inferred: true

Fig. 3. Answer display – in-line provenance info summary



Additional information

Phyllodes tumor

Phyllodes tumors (from Greek: phullon leaf), also cystosarcoma phyllodes, cystosarcoma phylloides and phylloides tumor, are typically large, fast growing masses that form from the periductal stromal cells of the breast. They account for less than 1% of all breast neoplasms.

Phyllodes tumors have long clefts and a myxoid cellular stroma. Micrograph. H&E stain.

Details

DiseasesDB: 3396

ICD10: C.50/c.50, .D.24/d.10, .D.48.6.d.37

ICD9: 217

ICDO: M9020

EMedicineSubj: med

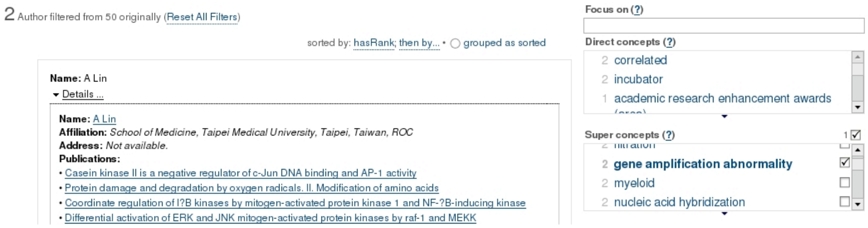
EMedicineTopic: 500

MeshID: D003557

Fig. 4. Single concept view – additional information box

question). However, other online sources, such as the KEGG service (cf. `http://www.genome.jp/kegg/kegg2.html`) can be easily incorporated.

Full-Text Search Besides the knowledge-based search, CORAAL supports also classical full-text search, which results in article paragraphs containing the queried terms. Links to the detailed view of the respective publication are provided, as well as citation contexts for the queried terms (i.e., publications referenced from the adjacent text). The full-text search results can be filtered, for instance by authors or citation contexts. One can also filter publications according to the concepts they contain, but also according to the associated indirect (inferred) general topics or specific instances. Besides publications, one can search for titles and authors. The respective results can be filtered, too. Figure 5 shows list of authors corresponding to the "Lin" name filtered only to those who have written an article concerned with the "gene amplification abnormality" topic. This feature of CORAAL can be used when looking for candidate experts on certain topics of interest.



2 Author filtered from 50 originally (Reset All Filters)

sorted by: hasRank; then by... • ○ grouped as sorted

Name: A Lin

▾ Details ...

Name: A Lin
Affiliation: *School of Medicine, Taipei Medical University, Taipei, Taiwan, ROC*
Address: *Not available.*
Publications:

- Casein kinase II is a negative regulator of c-Jun DNA binding and AP-1 activity
- Protein damage and degradation by oxygen radicals. II. Modification of amino acids
- Coordinate regulation of I?B kinases by mitogen-activated protein kinase 1 and NF-?B-inducing kinase
- Differential activation of ERK and JNK mitogen-activated protein kinases by raf-1 and MEKK

Focus on (?)

Direct concepts (?)

2 correlated
2 incubator
1 academic research enhancement awards (area)

Super concepts (?) 1 ☑

2 nitration ☐
2 **gene amplification abnormality** ☑
2 myeloid ☐
2 nucleic acid hybridization ☐

Fig. 5. Filtering the author search results

4 Conclusions and Future Work

With CORAAL, we started to implement the requirements of the use case outlined in Section 1.1. We have processed non-trivial amount of publication data purely automatically, extracted and processed the knowledge hidden in the articles, and exposed it conveniently to the users. The next and rather straightforward step is to integrate the publication knowledge with knowledge in patient records. For that, an appropriate parser for the patient records is essentially enough, as the EUREEKA engine employed in CORAAL is agnostic w.r.t. the origin of the input data. In the long-term perspective, one major task lies ahead of us, though. To make CORAAL really useful in practice, we still need to reduce the noise level in the exposed knowledge. Perhaps the easiest way is to utilise the wisdom of the crowds by supporting secure, error-prone and unobtrusive user involvement in the knowledge base updates (namely by (in)validation of existing statements, introduction of new statements and submission of new rules refining the domain semantics).

Acknowledgments This work has been supported by the 'Líon', 'Líon II' projects funded by Science Foundation Ireland under Grants No. SFI/02/CE1/I131, SFI/08/CE/I1380, respectively. We acknowledge much appreciated help from Ioana Hulpus, who developed the initial user interface for CORAAL. Very special thanks goes to the people who have actively participated in the continuous prototype evaluation and testing, namely to (in an alphabetical order): Doug Foxvog, Peter Gréll, MD, Miloš Holánek, MD, Matthias Samwald, Holger Stenzhorn and Jiří Vyskočil, MD.

References

1. Bechhofer, S., et al.: Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering (2003) At `http://tinyurl.com/96w7ms`, Apr'08.
2. Nováček, V., Decker, S.: Towards lightweight and robust large scale emergent knowledge processing. In: Proceedings of ISWC'09, Springer (2009) In press.
3. Groza, T., Handschuh, S., Moeller, K., Decker, S.: KonneXSALT: First steps towards a semantic claim federation infrastructure. In: The Semantic Web: Research and Applications (Proceedings of ESWC 2008), Springer-Verlag (2008) 80–94
4. Dietze, H., et al.: Gopubmed: Exploring pubmed with ontological background knowledge. In: Ontologies and Text Mining for Life Sciences, IBFI (2008)
5. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biology **2**(11) (2004)
6. Rajapakse, M., Rajaraman, K., Ang, W.T., Veeramani, A., Schreiber, M.J., Baker, C.J.O.: Ontology-centric integration and navigation of the dengue literature. Journal of Biomedical Informatics **41**(5) (2008) 806–815
7. Baker, C.J.O., Rajaraman, K., Ang, W.T., Veeramani, A., Low, H.S., Wenk, M.: Towards ontology-driven navigation of the lipid *bibliosphere*. BMC Bioinformatics **9**(S-1) (2008)
8. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., Stoehr, P.: Ebimed - text crunching to gather facts for proteins from medline. Bioinformatics **23**(2) (2007) 237–244
9. Hoffmann, R., Valencia, A.: A gene network for navigating the literature. Nature Genetics (2004)
10. Kano, Y., Jr., W.A.B., McCrohon, L., Ananiadou, S., Cohen, K.B., Hunter, L., ichi Tsujii, J.: U-compare: share and compare text mining tools with uima. Bioinformatics **25**(15) (2009) 1997–1998