Knowledge-Based Search for Oncological Literature

Vít Nováček, Tudor Groza, Siegfried Handschuh
DERI, National University of Ireland, Galway
IDA Business Park, Galway, Ireland
e-mail: FirstName.LastName@deri.org

***Abstract*—Using the current state of the art in life science publication search (e.g., PubMed), one can efficiently search for resources containing particular key-words or their combinations. It is impossible to search for abstract concepts and expressive relations between them (e.g., *type of*, *different from* or *part of*), though. Nevertheless, such a more expressive—semantic—search could largely reduce the efforts related to finding appropriate answers in biomedical articles. In this paper we identify challenges related to building a semantic publication search engine. Then we describe the architecture and usage principles of a tool tackling them. Eventually, we report on the tool's deployment on oncological literature data and preliminary tests with domain experts.**

I. Introduction

Although online publishing allows for very efficient global knowledge dissemination, the search capabilities currently offered by the state of the art tools (like PubMed, MEDLINE or Elsevier ScienceDirect) are often insufficient. Essentially, we produce the new data faster than we are able to interpret them. As an illustration, let us imagine we want to get more information on various types of the breast cancer and study the publications that are relevant to such knowledge. It is not possible to search for the *type* relationship per se in the current search engines. One may try to use for instance rather advanced full-text query `("type" OR "is a") AND "breast cancer"` (search for documents containing either of the first two terms, and the last term at the same time). However, even advanced full-text search results are often practically useless: (1) They may be incomplete, since the fact a concept is a type of another one is not necessarily expressed only using the *type* or *is a* terms. Moreover, breast cancer may be referred to by its synonyms as well. (2) The results may contain a lot of irrelevant hits, since the occurrence of the search terms does not mean there are also any breast cancer sub-types present. (3) Detailed manual analysis of the result set necessary to get the answers is very often impossible. For instance, the given example query retrieves more than $20,000$ publications on PubMed, which is clearly too much even if one uses a result filtering.

The community of life science researchers and practitioners demands more "semantic" solutions that would enable efficient answering of expressive queries on the biomedical data [10], [16]. However, extraction of the necessary knowledge from free text is hardly feasible in large scale if done purely manually [3]. Methods for automated knowledge extraction 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 [3]. In the following, we show this is the major and largely unsolved problem even in the case of cutting-edge solutions targeting more expressive search in life science publications.

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.

A. 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 (therefore we used them as a base-line in the user-centric experiment reported in Section IV). 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.

FindUR [15], Melisa [1] and GoPubMed [8] are ontology-based front-ends to a traditional publication full-text search. They allow either for effective restriction and intelligent visualisation of the query results (GoPubMed), or for focusing the queries onto particular topics based on an ontology (FindUR and Melisa). FindUR and Melisa use a Description Logics [2] ontology built from scratch and a custom ontology based on MeSH (cf. http://www.nlm.nih.gov/mesh/), respectively. 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. None of the tools, nevertheless, offers querying for or browsing of arbitrary publication knowledge – terms and relations not present in the systems' rather static ontologies simply cannot be reflected in the search.

Textpresso [16] enables searching for relations between concepts in particular chunks of text (namely for gene-to-gene interactions). However, the underlying ontologies and their instance sets have to be provided manually. Moreover, the system's scale regarding the number of publications' full-texts and concepts covered is quite low.

B. Contributions and Structure

From the overview of the state of the art in the field, it is obvious that the biggest challenge is a reliable automation

of more expressive content acquisition. None of the related systems addresses this problem appropriately, which makes them either poorly scalable, or difficult to port to a new domain. We have set to tackle this challenge with a prototype knowledge-based publication search engine – CORAAL (*COntent extended by emeRgent and Asserted Annotations of Linked publication data*). It can easily employ legacy domain resources or even work without any human intervention. The tool is based on a recently developed framework for more efficient exploitation of automatically extracted knowledge [17]. We combined the framework with a repository for semantically inter-linked publications [12] in order to provide for comprehensive combination of full-text and knowledge-based search.

The rest of the paper is organised as follows. Section II describes the data used in the current CORAAL deployment, as well as the tool's architecture and underlying technical principles. The user perspectives of the tool are covered in Section III. Section IV reports on assessment of CORAAL with oncology domain experts. We discuss the delivered work and outline future directions in Section V.

II. Data and Methods

A. Inputs and Outputs

1) Input: As of March 2009, 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 provided within the Elsevier Grand Challenge competition (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 the journals in the following list: 1) *FEBS Letters*; 2) *Biochemical Pharmacology*; 3) *Cancer Genetics and Cytogenetics*; 4) *Cell*; 5) *Trends in Cell Biology*; 6) *Experimental Cell Research*; 7) *Controlled Clinical Trials*; 8) *Molecular Aspects of Medicine*; 9) *Advanced Drug Delivery Reviews*; 10) *Gene*; 11) *Trends in Genetics*; 12) *Genomics*; 13) *Leukemia Research*; 14) *Journal of Microbiological Methods*; 15) *Trends in Microbiology*; 16) *Journal of Molecular Biology*; 17) *Oral Oncology*; 18) *European Journal of Pharmacology*. 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).

2) Output: CORAAL exposes two data-sets as an output of the publication processing:

- First, 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 [14] describing the input articles. This included $247,392$ publication titles and $374,553$ authors (both from full-texts and references processed).
- Second, 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.

Thanks to the improved knowledge representation back-end, generation of the output data-sets from the input articles took two days (as opposed to four days in the semifinal CORAAL prototype). Query evaluation on the produced content takes usually fractions and at most units of seconds[1].

B. Architecture of our Solution

In order to support comprehensive search functionalities, we propose to complement a standard (full-text) publication search approach with advanced services catering for semantic search. By semantic search we mean querying for and browsing of expressive statements capturing relations between concepts in the respective source articles.

Our particular solution—the CORAAL prototype—is built on the top of the KONNEX framework (a semantically interlinked publication repository [12]) and the EUREEKA library (enables integration and exploitation of automatically extracted knowledge [17]). CORAAL runs in a client-server mode. In order to work with the tool, one only needs a web browser. Everything else is handled by the server, quite similarly to the classical search engines (e.g., Google) from the user's point of view. The technical architecture of CORAAL is depicted in Figure 1.

[1]These results were achieved on a single server machine (which is not exclusively dedicated to CORAAL). There are still reserves regarding scalability even with the current implementation, however, for processing data two and more orders of magnitude larger, a distributed solution would be much better.

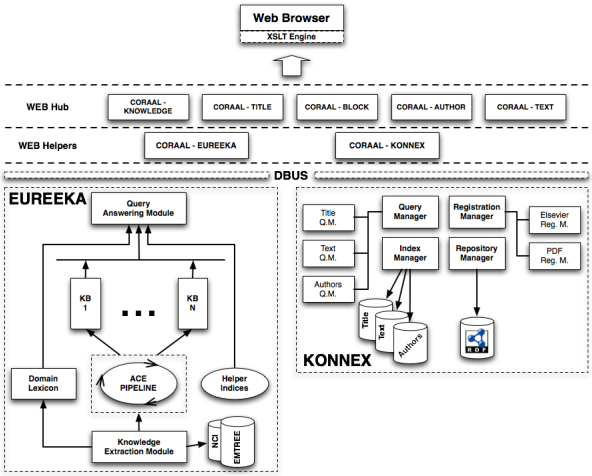

Fig. 1. Architecture of our solution

EUREEKA provides for knowledge extraction from text and other knowledge resources (e.g., ontologies or machine readable thesauri like NCI and EMTREE) via the *knowledge extraction module*. The extraction process possibly updates the *domain lexicon* and produces new knowledge being processed in the *ACE* pipeline (*ACE* stand for Addition, Closure, Extension; see Section II-E for examples). The pipeline caters for incremental addition, expansion and refinement of the emergent extracted knowledge within particular *knowledge bases*, which may be multiple if we want to represent particular contexts of the domain of interest separately. The knowledge bases are exposed to consumers via a semantic *query answering module*, optimising the retrieval and sorting the results using *helper indices* on the stored data. KONNEX tackles the integration of the extracted publication text and meta-data (in the form of 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* modules, possibly composed of sub-modules handling particular data or query types.

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 by the web browser itself via the MIT's Exhibit faceted browsing front-end. This solution results in CORAAL being a pure Semantic Web [4] 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 as well.

C. Knowledge Extraction

The publications, their meta-data and full-text are stored and indexed within the KONNEX framework for linked publication data processing [12]. After parsing the input XML article representations, the XML meta-data and structural annotations are quite straightforwardly integrated in the KONNEX RDF repository. Full-text information regarding the articles' content, titles, authors and references is managed using multiple Lucene IR indices (cf. http://lucene.apache.org/java/docs/).

For extraction of knowledge in the form of subject-predicate-object triples, we use a simple, yet already quite productive NLP-based heuristics similar to the technique described in [13]:

1) we identify sentences in the raw text and split them to particular word tokens
2) we tag the words in sentences by a probabilistic part-of-speech tagger
3) we chunk-parse the tagged sentences using a generic probabilistic shallow parser; the resulting plain (i.e., non-nested) chunks are of three types: NP (noun phrase – noun sequences, possibly with modifiers and/or grouped by coordinate conjunctions), VP (verb phrase – verbs only), PP (prepositional phrase – prepositions only)
4) for every $NP\ (PP\ NP)^*\ VP\ PP?\ NP\ (PP\ NP)^*$ chunk sequence[2] present in a sentence, we assume that the part preceding the verb phrase expresses a subject, the succeeding part an object and the verb phrase itself a predicate (i.e., property or relation holding between the subject and object)
5) we use several additional heuristics in order to generate the actual triple terms:
   - the head verb of VP is used as the predicate term (a preferred name based on WordNet [9] verb synsets used if possible)
   - if the predicate part of the chunk sequence is in the $VP\ PP$ form, the VP head verb with the respective preposition are used as the predicate term; if VP without a consequent PP is followed by $NP\ (PP\ NP)^+$, the head noun of the first NP and PP preposition are attached to the predicate terms in order to form a more specific predicate; an additional triple expressing the *is a* relation between the specific and verb-only generic predicate is generated
   - the remaining $NP\ (PP\ NP)^+$ sequences are merged together to provide a basis for the subject and object term construction
   - if there are modifiers or other nouns attached to a head noun, additional triple is generated in order to capture the *is a* relationship between the modified noun and the noun itself (based on heuristics discussed for instance in [6])
   - if there is an enumeration of terms in a noun phrase, additional triples are generated in order to capture mutual negative *is a* (*not a*) relationships (i.e., disjointness; based on a heuristics explained for instance in [19]);

[2] *, $^+$ and ? mean zero or more, one or more and zero or one repetitions of the preceding expression, respectively.

enumerations in either subject or object noun phrase result in multiple basic triples, too

D. Knowledge Representation Principles

A compact representation of concepts and knowledge bases we construct from the emergent extracted statements is given by Definition 1. The compact representation consequently allows for a straightforward specification of soft integration and similarity notions, which enable publication knowledge merging and its approximate querying in CORAAL.

Definition 1: **Concept** is a square matrix $\mathbf{A}$ with elements $a_{i,j} \in [-1, 1], i, j \in I$, where I is an index set. Let M be a set of all concepts, L a set of lexical expressions that may refer to concepts in M and L^* a set of fuzzy sets [21] defined on the L universe. We define **lexical interpretation** λ as a bijection $\lambda : M \rightarrow L^*$. **Index assignment** binding the concepts and indices together is then a bijection $ind : M \rightarrow I$ such that $ind(\mathbf{A}) = ind(\mathbf{B})$ iff $\lambda(\mathbf{A}) = \lambda(\mathbf{B})$. Regarding concept equivalence, we call concepts **strongly equal**, $\mathbf{A} = \mathbf{B}$, iff $a_{i,j} = b_{i,j}$ for $\forall i, j \in I$, and **weakly equal**, $\mathbf{A} \simeq \mathbf{B}$, iff $ind(\mathbf{A}) = ind(\mathbf{B})$. **Empirical knowledge base** is a tuple $(K, I_K, L_K, ind_K, \lambda_K)$, where $K \subseteq M, I_K \subseteq I, L_K \subseteq L$ and ind_K, λ_K are the respective specific index assignment and lexical interpretation mappings.

Note that we do not distinguish between "classes" and "individuals" in the traditional sense (i.e., sets and elements in a domain universe, respectively). A concept can be empirically considered to be an "individual" as long as it has no sub-types, but it can suddenly become a "class" when a sub-type concept is newly introduced to it. Therefore everything is just a concept and finer-grained ontological distinctions are left to particular applications of the basic principles.

The sets M, L^*, I can be understood as the conceptual, symbolic and real world domain, respectively, in the semiotic triangle [18] perspective (considering I as a set of unique identifiers of entities existing in universe). The λ, ind mappings (and their inverses) can then be understood along the symbolisation and reference relations in the triangle. The intuition behind the fuzzy character of λ is the fact that concepts are usually referred to by more than one lexical expression. Moreover, these expressions have uneven degrees of relevance w.r.t. the particular concept (e.g., the expression *"a reasoning erected primate"* is perhaps a bit more relevant to the *"human"* concept than the expression *"a bipedal animal without feathers"*, while the *"human"* expression is one of the most relevant).

The introduced matrix notation for concepts is convenient due its conciseness, however, we can use also a more human-readable and explanatory *statement notation*, closely following the standard RDF(S) terminology [5]. A concept $\mathbf{A}$ can be expanded as a conjunction

$$<s : p_1 : o_1>^{a_{p_1,o_1}} AND <s : p_2 : o_2>^{a_{p_2,o_2}} AND \ldots AND$$
$$<s : p_n : o_n>^{a_{p_n,o_n}}$$

of particular *subject : predicate : object* statements[3]. $s = ind(\mathbf{A})$ and n is the number the of non-zero elements in $\mathbf{A}$. p_i, o_i are the row and column indices, respectively, of the particular non-zero matrix element. The element values a_{p_i,o_i} represent the degrees of certainty about the fact that the actual relation $ind^{-1}(p_i)$ holds (or does not hold for $a_{p_i,o_i} < 0$) between $ind^{-1}(s)$ and $ind^{-1}(o_i)$.

Example 1: Consider for instance the concept `T-cell leukemia`, being certainly a `type of` the concept `disease` and certainly *not* a `type of` (i.e., different from) the concept `infection` according to a human expert. The respective concept matrix $\mathbf{A}$ may look like this (omitting the zero elements):

SAMPLE-URI#1	SAMPLE-URI#3	SAMPLE-URI#4
SAMPLE-URI#2	1.0	-1.0

The corresponding statement notation would be:

$$<\texttt{SAMPLE-URI\#1 : SAMPLE-URI\#2 : SAMPLE-URI\#3}>^{1.0} AND$$
$$<\texttt{SAMPLE-URI\#1 : SAMPLE-URI\#2 : SAMPLE-URI\#4}>^{-1.0}$$

The SAMPLE-URI#2, SAMPLE-URI#3 and SAMPLE-URI#4 indices correspond to matrices $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$, respectively, regarding the ind mapping. The lexical interpretation λ is defined as follows then: $\lambda(\mathbf{A}) = (S_1, \mu_1), \lambda(\mathbf{B}) = (S_2, \mu_2)$, $\lambda(\mathbf{C}) = (S_3, \mu_3), \lambda(\mathbf{D}) = (S_4, \mu_4)$, where $(S_1, \mu_1), (S_2, \mu_2)$, (S_3, μ_3) and (S_4, μ_4) are fuzzy sets such that $\mu_1(\texttt{T-cell leukemia}) = 1$, $\mu_2(\texttt{type of}) = 1$, $\mu_3(\texttt{disease}) = 1, \mu_4(\texttt{infection}) = 1$.

The interpretation of the element values in the concept matrices is deliberately left unrestricted, since the degrees are meant to capture heuristic confidence bound to emergent statements. The confidence may be based on statistics, however, it may also be based on arbitrary measures assigned either by people, or by various knowledge extraction algorithms, therefore no particular mathematical formalism can be used for the degree interpretation in general. Users can, however, impose a specific semantics on degrees later on when consuming the data stored in an empirical knowledge base. The most straightforward approach is interpreting the absolute values bound to the positive and negative facts as membership degrees, following the fuzzy sets formalism [21]. Intuitionistic fuzzy sets [7] can be used for more straightforward interpretation of positive and negative degrees as membership and non-membership measures, respectively.

Concepts stored in a knowledge base can arbitrarily change in the emergent settings, which is supported by the following operation (note that the possible updates of the ind_K, λ_K mappings are technical issues dependent on the particular implementation and thus omitted in the definition).

Definition 2: **Concept change** $\Delta_{u,v} : M^2 \rightarrow M, u, v \in [0, 1]$ is a parametrised binary operation such that $\Delta_{u,v}(\mathbf{A}, \mathbf{B}) = \mathbf{C}$, where $c_{i,j} = \nu_{u,v}(a_{i,j}, b_{i,j})$. $\nu_{u,v} : [-1, 1]^2 \rightarrow [-1, 1]$ is a so called **element change** function defined as follows: (1), for $x \neq 0$, $\nu_{u,v}(x, 0) = x$; (2), for $y \neq 0$, $\nu_{u,v}(0, y) = y$; (3), for $x, y \neq 0$, $\nu_{u,v}(x, y) = F(ux, vy)$, F being an ordered weighted averaging (OWA) operator[4].

[3]Note that without loss of generality, URIs may serve as concept indices in the statements. Consequently, ind^{-1} de facto plays a role of the URI dereference. To facilitate readability, we provide simply lexical terms instead of indices or URIs in the examples throughout the paper, though.

[4]Defined in [20] as $F(a_1, \ldots, a_n) = \sum_{j=1}^{n} w_j b_j$, where b_j is the j-th largest of the a_i and w_j are a collection of weights (also called a weight vector) such that $w_j \in [0, 1]$ and $\sum_{j=1}^{n} w_j = 1$. Note that we use the additional u, v weights in order to explicitly capture the relative relevance of the $\Delta_{u,v}$ first and second argument independently from their relative sizes.

The change operation can be understood as a simple, yet useful formal model of cognitive learning and attitude change as studied in psychology [11], i.e., acceptance, rejection or modification of the attitude on a topic—a relation between two concepts in our case—according to the current content of the knowledge base and persuasive emergent facts. The purpose of the u, v parameters is to reflect the "persuasion potential" (weight) of particular knowledge sources being incorporated w.r.t. the already known content. For instance, they may be set as $u = 1, v = 1$ for presumably correct and equally trusted knowledge from manually created ontologies, $u = 0.9, v = 1$ for input knowledge from a resource more trusted than the actual content, or $u = 1, v = 0.75$ for less trusted learned data.

Using an appropriate selection of the F operator (see [20] for details on possible choices), one can model various types of concept change. Possible practically relevant choices are, e.g., maximum (strict preference of positive or more certain facts) or weighted arithmetic mean (using u, v as the respective weights).

Example 2: Imagine that we learn from a natural language text that `T-cell leukemia` is different from `acute granulocytic leukemia` with confidence 0.8, and that `T-cell leukemia` is a type of `infection` with confidence 0.2. Assuming the 0.8 relevance for the learning algorithm when compared to the 1.0 relevance of human expert, the `T-cell leukemia` concept from Example 1 has to be updated using $\Delta_{1.0,0.8}$ regarding the new findings. The changed concept is described as follows then (degrees computed using the dynamic weighted mean OWA operator):

$$\texttt{<T-cell leukemia : type : disease>}^{1.0}\ AND\ \texttt{<T-cell leukemia : type : acute granulocytic leukemia>}^{-0.8}\ AND\ \texttt{<T-cell leukemia : type : infection>}^{-0.46}$$

Besides direct concept incorporation by the change operation, one has to be able to aggregate multiple concepts evenly as well. This is particularly useful for instance when merging concepts from multiple equally trusted sources before their actual incorporation into the known content, or when aggregating intermediate inference results.

Definition 3: **Concept aggregation** is a function $\bigcirc : 2^M \to 2^M$, $\bigcirc(X) = \{\odot(S_l) | S_l \in S\}$, where S is a set of the equivalence classes on X w.r.t. $\simeq$. $\odot : 2^M \to M$ is a function aggregating weakly equal matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ into a matrix $\mathbf{B}$ with elements $b_{i,j} = F(x_1, \ldots, x_k)$, where $x_1, \ldots, x_k$ are the respective non-zero $a_{i,j}$ values among $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and F is an OWA operator.

Example 3: An aggregation of the `T-cell leukemia` concept updated in Example 2 with the following concept of the same relevance (e.g., learned, too, but from different data)

$$\texttt{<T-cell leukemia : type : acute granulocytic leukemia>}^{-0.5}\ AND\ \texttt{<T-cell leukemia : type : infection>}^{-0.8}$$

would result in this update (degrees computed using arithmetic mean OWA operator):

$$\texttt{<T-cell leukemia : type : disease>}^{1.0}\ AND\ \texttt{<T-cell leukemia : type : acute granulocytic leukemia>}^{-0.65}\ AND\ \texttt{<T-cell leukemia : type : infection>}^{-0.63}$$

Crucial for the basic inference services in our approach is the notion of similarity. We formalise it using metrics on M.

Definition 4: Semantic metrics class Ω is a set of parametrised functions $\delta_H : M^2 \to \mathbb{R}_0^+$ for all $\emptyset \subset H \subseteq I^2$ such that: (1), δ_H is a metric on M; (2), in order to compute $\delta_H(\mathbf{A}, \mathbf{B})$, only elements $a_{i,j}, b_{i,j}$ with $(i,j) \in H$ are taken into account. We can define a partial ordering $\preceq$ on Ω, such that $\delta_{H_1} \preceq \delta_{H_2}$ iff $H_1 \subseteq H_2$. Dually to distance, we define graded **concept similarity**[5] as $\sigma_H : M^2 \to (0, 1], \sigma_H(\mathbf{A}, \mathbf{B}) = \frac{1}{1+\delta_H(\mathbf{A},\mathbf{B})}$. A partial ordering $\sqsubseteq$ on the set of all similarities can be defined as $\sigma_{H_1} \sqsubseteq \sigma_{H_2}$ iff $\delta_{H_1} \preceq \delta_{H_2}$.

An example metrics is $\delta_H(\mathbf{A}, \mathbf{B}) = \sum_{(i,j)\in H} |a_{i,j} - b_{i,j}|$, or a normalised alternative $\delta_H(\mathbf{A}, \mathbf{B}) = \frac{1}{|H|} \sum_{(i,j)\in H} |a_{i,j} - b_{i,j}|$. The dependence on the H set allows for graded modelling of specific distances, influenced only by certain relations instead of all relations possible. The specificness of the particular distances (or the dual similarities) is directly related to the $\preceq$ ordering, i.e., if $\delta_{H_1} \preceq \delta_{H_2}$, then δ_{H_1} is more specific than δ_{H_2}. Specific similarities are particularly useful when we want to retrieve content from a knowledge base – e.g., all concepts being type of a disease and treated by radiological methods. We can form a respective query concept and check the knowledge base for matrices with specific similarity regarding the two query properties higher than a given threshold. Comparison regarding all possible properties would possibly retrieve much smaller set of appropriate answer concepts for large knowledge bases with many properties present, which is not the intuitively expected behaviour.

We can distinguish certain prominent types of similarity functions according to the H parameter. First, let $\sigma(\mathbf{A}, \mathbf{B}) = \sigma_{I^2}(\mathbf{A}, \mathbf{B})$ be a similarity *between* $\mathbf{A}$ and $\mathbf{B}$ (a general comparison). Second, let $\overleftarrow{\sigma}(\mathbf{A}, \mathbf{B}) = \sigma_{\{(i,j)|b_{i,j}\neq 0\}}(\mathbf{A}, \mathbf{B})$ and $\overrightarrow{\sigma}(\mathbf{A}, \mathbf{B}) = \sigma_{\{(i,j)|a_{i,j}\neq 0\}}(\mathbf{A}, \mathbf{B})$ be a similarity of $\mathbf{B}$ *to* $\mathbf{A}$ and $\mathbf{A}$ *to* $\mathbf{B}$, respectively (a specific comparison of either $\mathbf{B}$ to $\mathbf{A}$, or $\mathbf{A}$ to $\mathbf{B}$, based on the respective non-zero elements). Third, let $\overline{\sigma}(\mathbf{A}, \mathbf{B}) = \sigma_{\{(i,j)|a_{i,j}\neq 0 \wedge b_{i,j}\neq 0\}}(\mathbf{A}, \mathbf{B})$ be an intersection similarity between $\mathbf{A}$ and $\mathbf{B}$ (a comparison based only on the elements $\mathbf{A}$ and $\mathbf{B}$ have in common). Quite clearly, $\overline{\sigma} \sqsubseteq \overrightarrow{\sigma} \sqsubseteq \sigma, \overline{\sigma} \sqsubseteq \overleftarrow{\sigma} \sqsubseteq \sigma$.

Example 4: The similarity $\overrightarrow{\sigma}$ of the following concept (with the subject indicating a variable)

$$\texttt{<?X : type : disease>}^{1.0}\ AND\ \texttt{<?X : type : acute granulocytic leukemia>}^{-1.0}$$

to the concept

$$\texttt{<T-cell leukemia : type : disease>}^{1.0}\ AND\ \texttt{<T-cell leukemia : type : acute granulocytic leukemia>}^{-0.65}\ AND\ \texttt{<T-cell leukemia : type : infection>}^{-0.63}$$

is about 0.851 when using the $\delta_H(\mathbf{A}, \mathbf{B}) = \frac{1}{|H|} \sum_{(i,j)\in H} |a_{i,j} - b_{i,j}|$ distance as a basis for the similarity computation. The respective σ similarity *between* the concepts is about 0.753 then. Note that both similarities are relatively high, suggesting that `T-cell leukemia` might be an instance of `?X` to a certain degree.

The gradual concept similarities are employed by lightweight inference services of two basic types: 1) *retrieval* of similar concepts (quite straightforward); 2) fixed-point *rule-based materialisation* of implicit relations, complex *querying*

[5]The duality w.r.t. the distance is ensured by the conformance to two intuitive conditions – inverse proportionality and equality to 1 when the distance is 0.

(similarity as a basis for soft variable unification and for approximate fixed-point computation). The inference algorithms have anytime behaviour and it is possible to programmatically adjust their completeness/efficiency trade-off. Proper elaboration of the inference is out of scope here, however, we cover it in a technical report [17], which addresses also implementation details of the knowledge base storage.

E. CORAAL Workflow Example

In the following we exemplify how the extracted knowledge is processed in CORAAL. Initially we extract triples, encoding three types of ontological relations between concepts: taxonomical—*is a/type*—relationships, difference of concepts (i.e., negative *is a/type* relationships) and generic relations (e.g., *part of* or *plays role in*). We extend the extracted triples to quads by attaching scores based on term frequencies in the input corpus. An example of a sentence and part of respective extracted knowledge follows[6]:

> *The rate of T-cell leukemia, acute granulocytic leukemia and other hematologic disorders in the studied sample was about three times higher than average.* ⇝ `(T-cell leukemia, is a, leukemia, 1.0), (T-cell leukemia, is a, acute granulocytic leukemia, -0.6), (T-cell leukemia, is a, hematologic disorder, -0.6),...`

There is one obvious mistake in the extracted knowledge – *T-cell leukemia* actually is a *hematologic disorder*. However, CORAAL can remedy that. The seed knowledge base imported from the NCI and EMTREE domain thesauri contains the following knowledge (crisp, therefore with 1.0 degrees only):

```
(leukemia, is a, Hematopoietic and Lymphoid Cell
Neoplasm, 1.0), (Hematopoietic and Lymphoid Cell
Neoplasm, is a, Hematologic and Lymphocytic
Disorder, 1.0), (Hematologic and Lymphocytic
Disorder, same as, hematologic disorder, 1.0)
```

Thus, thanks to the EUREEKA inference engine that is currently employing a modification of RDFS general-purpose entailment rules [5], we know that `(leukemia, is a, hematologic disorder, 1.0)` according to domain experts. Therefore, the *T-cell leukemia*, a type of *leukemia*, should also be a *hematologic disorder*. The automatically extracted noisy knowledge has much lower relevance than the presumably precise domain resources, and CORAAL can make use of it when aggregating new content into its emergent knowledge base. Assuming the opinion of experts who created the thesauri is five times more relevant than the extracted knowledge in case of conflicts, the eventually incorporated statement computed by concept aggregation is `(T-cell leukemia, is a, hematologic disorder, 0.73)`. The erroneous emergent fact stating that *T-cell leukemia* actually is not a *hematologic disorder* has been automatically repaired to large extent.

Besides the emergent knowledge refinement, CORAAL is also able to extend the extracted concepts by additional relations using light-weight analogical reasoning. For instance, it is able to find that *acute granulocytic leukemia* is related to *myeloproliferative disorders* and *myelomonocytic leukemia*. The abstract, yet already quite useful associations are directly based on the concept similarities.

[6]The degree values are illustrative. The negative degree corresponds to negation of the respective relation. The difference in absolute values of the positive and negative degrees (i.e., 1.0 vs. -0.6) corresponds to different confidence measures provided by the algorithms responsible for the extraction of particular statements. The confidence in correctness of disjointness is usually lower in practice since the respective algorithms are less precise than those for extraction of sub-type relationships.

III. Using CORAAL

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.

Examples of particular queries possible in CORAAL for various types of search are as follows:

- publication **knowledge** search:
  - *concepts* – use just the concept name, i.e., respective term(s). Examples: `lymphoblastic leukemia`, `chemosensitizer`
  - *statements* – use the `S : P : O` syntax, where `S, P, O` are subject, predicate, object expressions, respectively. The expressions may be either in the form of a concept name, or in the form of a variable (anything starting with `?`, possibly even `?` alone The only limitation is that there may be at most one variable in a query statement. Checks for concepts satisfying a feature, or for relations between concepts. Example: `? : is a : breast cancer, p53 : ? : early carcinogenic events, rapid antigen testing : part of : ?`
  - *conjunctive statements* – use the $\texttt{St}_1$ `AND` $\texttt{St}_2$ `AND ...AND` $\texttt{St}_n$ syntax, where $\texttt{St}_1$, ..., $\texttt{St}_n$ are statements. At most one variable identifier (either in subject or in object positions) is allowed to appear in a conjunctive query. Checks for concepts satisfying multiple features. Examples: `rapid antigen testing : part of : ? AND ? : is a : clinical study, ? : blocks : binding site AND ? : mediator of activation : reactive oxygen metabolite`
  - *negative statements* – use the `NOT St` syntax, where `St` is a statement. The `NOT` key word may be used even inside the statement. Checks for concepts explicitly satisfying a negative feature. Examples: `NOT ? : is a: penicillin, acute granulocytic leukemia : NOT is a : chronic neutrophilic leukemia`
  - *complex queries* combining the above. Examples: `? : NOT is a :mouse AND ? : is a : animal, ? : as : complementary method AND ? : NOT type : polymerase chain reaction`
- publication text, title and author **full-text** search:
  - terms in the traditional full-text search syntax, i.e., term names plus wild-cards like * or ? and boolean key-words like `AND`, `OR` or `NOT`. Examples: `"breast cancer"`, `carci*`, `"breast cancer" AND p53`

CORAAL itself 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).

After pointing a browser to the URL, the main search interface will appear as shown in Figure 2. The tabs correspond to the particular types of search – the *Knowledge* tab serves for publication knowledge search using the query language specified above, while the *Text, Title, Authors* tabs realise full-text search for the respective publication (meta)data. Figure 2 shows how a query for *knowledge* is constructed simply by typing it into the search box.

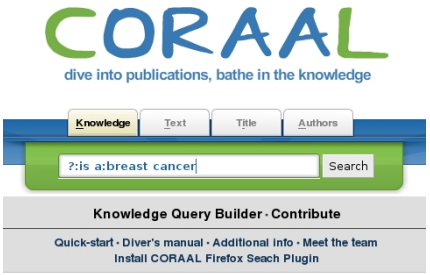

Fig. 2. Asking a query – direct

CORAAL can also assist the user when asking a query. After clicking on the *Knowledge Query Builder* link, one can use a form with auto-completion capabilities providing for guided query build-up on the actual content of the underlying knowledge base (Figure 3).

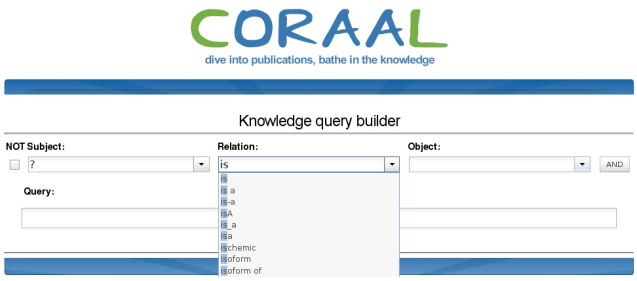

Fig. 3. Asking a query – assisted

The answers are displayed as particular statements provided with several types of meta-information: (1) *source* provenance – articles relevant to the statement; (2) *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); (3) *certainty* – how certain the system is that the statement holds and is relevant to the query (values in $[0, 1]$); (4) *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. Using a particular filtering, one can quickly focus only on statements of a particular interest. Such a specific focus can be seen in Figure 4. Note that the `HAS PART` relation has rather general semantics in the knowledge extracted by CORAAL, i.e., its meaning is not strictly mereological in the physical sense, it can refer also to, e.g., conceptual parts or possession of entities. Similarly for the `PART OF` relation.

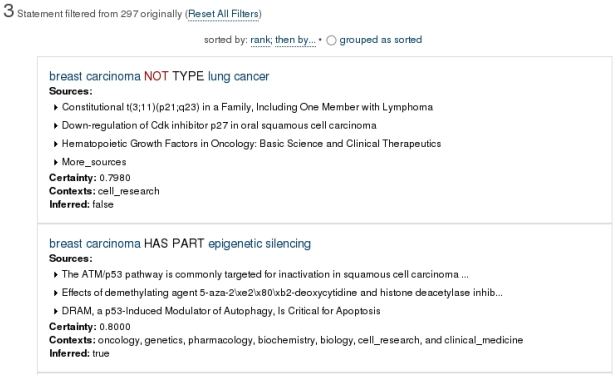

Fig. 4. Answer display – focused

Article provenance summaries of particular statements can be displayed in-line as shown in Figure 5.

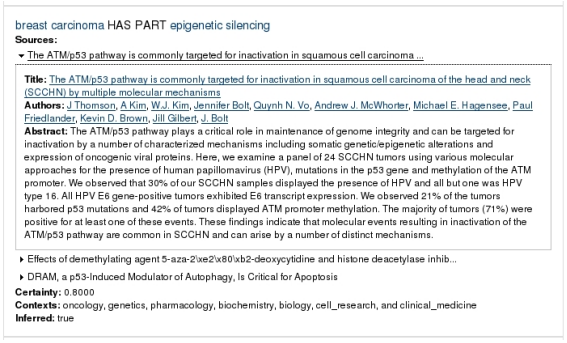

Fig. 5. Answer display – in-line provenance info summary

The results of the traditional full-text search can be filtered by the associated knowledge, too, in CORAAL. For instance, Figure 6 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 for instance for finding candidate experts on certain topics.

Fig. 6. Filtering author search

IV. Preliminary Tests with Domain Experts

During the CORAAL prototype development, we continually collaborated with several biomedical experts, who formed a committee of sample users and evaluators. Before the final stages of the Elsevier Grand Challenge, we prepared five tasks to be worked out with CORAAL and a base-line application (ScienceDirect or PubMed). Our hypothesis was that users should perform better with CORAAL than with the base-line, since the tasks were focused rather on structured knowledge than than on a plain text-based search[7].

[7]For instance, the users were asked to find all authors who support the fact that the `acute granulocytic leukemia` and `T-cell leukemia` concepts are disjoint, or to find which process is used as a `complementary method`, while being different from the `polymerase chain reaction`, and identify publications that support their findings.

The average level of evaluation tasks' direct similarity to the day-to-day agenda of users was approximately 4 on the $1-6$ scale (from least to most relevant), meaning that the tasks had tangible relation to the practice. The success rate of task accomplishment was 60.7% and 10.7% when using CORAAL and the base-line application, respectively. This clearly confirms our hypothesis.

Besides evaluating the users' performance in sample knowledge-based search tasks, we were interviewing them regarding the overall usability of the CORAAL interface. The most critical issue was related to the query language – half of the sample users were not able to construct appropriate queries directly sometimes. However, CORAAL offers also the form-based query builder that assists the user as illustrated in Section III. Using this feature, users performed up to six-times faster and 40% more efficiently than with purely manually constructed queries.

The expert users also had slight problems with too general, obvious or irrelevant results when presented only with a plain non-interactive unsorted list of answer statements provided by CORAAL. These concerns are addressed by the following particular features in the user interface: (i) *relevance-based sorting* of concepts and statements [17] – the proportion of relevant statements present among the results increases towards the top of the answer list; (ii) intuitive *faceted browsing* functionality – support for fast and easy reduction of the displayed results to a sub-set with certain features (i.e., statements having only certain objects or authors writing about certain topics). The solutions were considered as mostly sufficient regarding the sample users' concerns (an average $4.\overline{6}$ score on the $1-6$ scale going from least to most sufficient).

V. Conclusions and Future Work

We have presented the architecture and usage principles of CORAAL – a unique combination of a semantic publication repository [12] and a framework for automated exploitation of the knowledge contained in publication texts [17]. We deployed CORAAL on a sample oncological literature data and showed that the experts were able to perform much better with the tool than with state of the art publication search engines. We also reported on how we have reflected the evaluators' feedback in the recent agile development aiming at improved user experience.

The key strength of CORAAL is its ability to perform with straightforwardly imported legacy vocabularies (e.g., biomedical thesauri), or even without them if necessary. The knowledge from textual resources is extracted and processed purely automatically and can be easily queried in an expressive, yet intuitive way. The faceted browsing of the results allows for fast focus on statements of particular interest, which ensures usability despite of some remaining noise in the automatically acquired knowledge. CORAAL can be cost-efficiently deployed wherever expressive querying of knowledge scattered across vast amounts of unstructured textual data is required (e.g., patient records, to name one example). None of the similar tools we know of can be deployed and used in such an easy and intuitive way across arbitrary domains.

In future, we are going to extend the scalability of the tool to millions of publications and beyond, utilising federated knowledge storage and querying. We also want to provide user-friendly and reliable means for exploitation of the wisdom of the crowds. This will in effect enable for filtering out most of the imprecision possibly present in the emergent knowledge handled by CORAAL.

References

[1] J. M. Abasolo and M. Gómez. Melisa: An ontology-based agent for information retrieval in medicine. In *SemWeb2000*, 2000.

[2] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider. *The Decription Logic Handbook: Theory, implementation, and applications*. Cambridge University Press, Cambridge, USA, 2003.

[3] S. Bechhofer et al. Tackling the ontology acquisition bottleneck: An experiment in ontology re-engineering, 2003. At http://tinyurl.com/96w7ms, Apr'08.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Sci. Am.*, 5, 2001.

[5] D. Brickley and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004. Available at (Feb 2006): http://www.w3.org/TR/rdf-schema/.

[6] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning taxonomic relations from heterogenous sources of evidence. In P. Buitelaar, P. Cimiano, and B. Magnini, editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, pages 59–73. IOS Press, 2005.

[7] G. Deschrijver, C. Cornelis, and E. E. Kerre. On the representation of intuitionistic fuzzy t-norms and t-conorms. In *Trans. on Fuzzy Systems*. IEEE, 2004.

[8] H. Dietze et al. Gopubmed: Exploring pubmed with ontological background knowledge. In *Ontologies and Text Mining for Life Sciences*. IBFI, 2008.

[9] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[10] C. Goble. State of the nation in data integration. In *Proceedings of the WWW2007/HCLSDI Workshop*. ACM Press, 2007.

[11] A. G. Greenwald. Cognitive learning, cognitive response to persuasion, and attitude change. In *Psychological Foundations of Attitudes*, pages 147–169. Academic Press Inc., New York, 1968.

[12] T. Groza, S. Handschuh, K. Moeller, and S. Decker. KonneXSALT: First steps towards a semantic claim federation infrastructure. In *The Semantic Web: Research and Applications (Proceedings of ESWC 2008)*, pages 80–94. Springer-Verlag, 2008.

[13] A. Maedche and S. Staab. Discovering conceptual relations from text. In *Proceedings of ECAI 2000*. IOS Press, 2000.

[14] F. Manola and E. Miller. *RDF Primer*, 2004. Available at (November 2008): http://www.w3.org/TR/rdf-primer/.

[15] D. L. McGuinness. Ontology-enhanced search for primary care medical literature. In *Proceedings of the Medical Concept Representation and Natural Language Processing Conference*, pages 16–19, 1999.

[16] H. M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11), 2004.

[17] V. Nováček. Towards an efficient knowledge-based publication data exploitation: An oncological literature search scenario. Technical Report DERI-TR-2009-03-23, DERI, NUIG, 2009. Available at http://tinyurl.com/csh3rf.

[18] C. K. Ogden and I. A. Richards. *The Meaning of Meaning*. Mariner Books, 1989.

[19] J. Voelker, D. Vrandecic, Y. Sure, and A. Hotho. Learning disjointness. In *Proceedings of ESWC'07*. Springer, 2007.

[20] R. R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190, 1988.

[21] L. A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.