CORAAL – Dive into Publications, Bathe in the Knowledge

Vít Nováček * and Tudor Groza and Siegfried Handschuh and Stefan Decker
Digital Enterprise Research Institute, National University of Ireland Galway, IDA Business Park, Dangan, Galway, Ireland

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 CORAAL (*COntent extended by emeRgent and Asserted Annotations of Linked publication data*), an award-winning prototype presented in this article. 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. This way we allow for convenient diving into publications and bathing in the knowledge related to the particular texts.

Key words: knowledge acquisition, linked publication data, emergence, knowledge integration, life sciences, semantic search

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 author names).

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.

To remedy the shortcomings of the current solutions, the future publishing paradigms should support decomposed machine-readable content that goes beyond mere text locked in rather monolithic publications. The envisioned content should allow for expressive inter-linking of scientific artefacts, thus unfolding new dimensions of browsing and data integration. It should be amenable to robust autonomous extraction and meaningful inference of implicit domain knowledge present in the text in order to expose it for search, too.

* Corresponding author. Tel: +353 (0)91 495738
Email addresses: vit.novacek@deri.org (Vít Nováček), tudor.groza@deri.org (Tudor Groza), siegfried.handschuh@deri.org (Siegfried Handschuh), stefan.decker@deri.org (Stefan Decker).

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 (3). 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 (11). We combined the framework with our repository for semantically inter-linked publications (6), 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 associated with and hidden in the publication texts.

In this system overview article, we describe how we implemented and deployed CORAAL for the Elsevier Grand Challenge (Section 2 and 3, respectively). Section 3 also discusses preliminary evaluation of the system with sample users and comments on particular ways of practical CORAAL application. Summary of related systems is given in Section 4, Section 5 concludes the article then.

2. Implementation

2.1. *Architecture*

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 semantic search. By semantic 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 (6) and EUREEKA (11) 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 semantic search.

CORAAL runs in a client-server mode. In order to work with the tool, you only need your 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. EUREEKA provides for knowledge extraction from

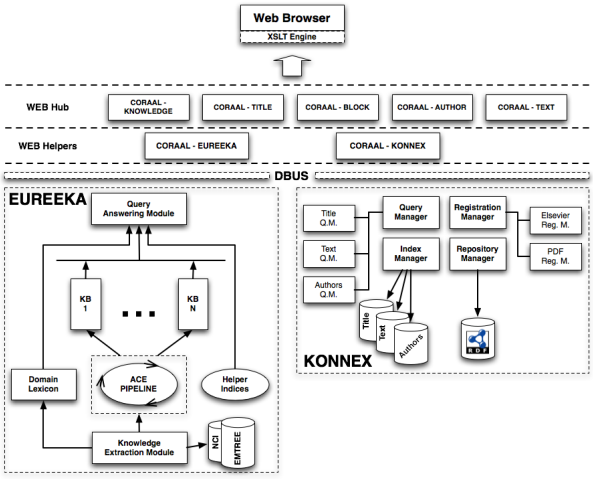

Fig. 1. CORAAL architecture

text and other knowledge resources (e.g., ontologies or machine readable thesauri) via the *knowledge extraction module*. The extraction process possibly updates the *domain lexicon* and produces new knowledge being processed in the *Addition-Closure-Extension* (ACE) pipeline (see Section 2.2 for details). After being processed by the pipeline, new facts are added into 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*. Optimisation of the retrieval and sorting of the results makes use of *helper indices*, representing for instance relevance scores of particular stored statements. KONNEX tackles the integration of the extracted publication text and meta-data, represented as RDF graphs (8), 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 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.

2.2. *Technological Groundwork*

The publications, their meta-data and full-text are stored and indexed within our KONNEX framework for linked publication data processing (6). After parsing the input XML representations of Elsevier articles, 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 are managed using multiple Lucene IR indices (cf. `http://lucene.apache.org/java/docs/`).

Exploitation of the publication knowledge is tackled by our novel EUREEKA framework for emergent (e.g., automatically extracted) knowledge processing (11). The framework de facto builds on a simple triple model (8). However, we extended the subject-predicate-object triples by positive or negative heuristic certainty measures and organised them in so called conceptual matrices, concisely representing every positive and negative relation of an entity to other entities. Metrics can be easily defined on the conceptual matrices. The metrics then serve as a natural basis for gradual concept similarities that define basic light-weight empirical semantics in EUREEKA (11). On the top of the similarity-based semantics, we implemented simple, yet quite practical inference services of two basic types: (i) *retrieval* of knowledge similar to an input concept, and/or its *extension* by means of similar stored content; (ii) fixed-point rule-based *materialisation* of implicit relations, and/or complex *querying* (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. Technical details of the solution are out of scope regarding this article, but one can find them in (11).

We applied our EUREEKA prototype to: (i) automated extraction of machine-readable knowledge bases from particular life science article texts; (ii) integration, refinement and extension of the extracted knowledge within one large emergent knowledge base; (iii) exposure of the processed knowledge via a query-answering and faceted browsing interface, tracking the article provenance of particular statements.

For the initial knowledge extraction, we use a NLP-based heuristics stemming from (7; 12) in order to process chunk-parsed texts into subject-predicate-object-score quads. The scores were derived from aggregated absolute and document frequencies of subject/object and predicate terms. The extracted quads encode three major types of ontological relations between concepts: (I) taxonomical—*type*—relationships; (II) concept difference (i.e., negative *type* relationships); (III) "facet" relations derived from verb frames in the input texts (e.g., *has part*, *involves* or *occurs in*). We impose taxonomy on the latter, considering the head verb of the respective phrase as a more generic relation (e.g., *involves expression of* was assumed to be a type of *involves*). Also, several artificial relation types were introduced to restrict the semantics of some most frequent relations. Namely, (positive) *type* is considered transitive and anti-symmetric, and *same as* is set transitive and symmetric. Similarly, *part of* is assumed transitive and being inverse of *has part*. Note that the *has part* relation has rather general semantics within the extracted knowledge, i.e., its meaning is not strictly physically mereological, it can refer also to, e.g., conceptual parts or possession of entities.

The emergent quads are processed as follows (details of the particular steps and the underlying principles are described in (11)):

(I) *addition* – The extracted quads are incrementally added into an emergent knowledge base K, using a fuzzy aggregation of the respective conceptual matrices. As a seed defining the basic domain semantics (i.e., synonymy and core taxonomy of K), we used the EMTREE and NCI thesauri.
(II) *closure* – After the addition of new facts into K, we compute its materialisation according to RDFS entailment rules (4) ported to the format specified in (11).
(III) *extension* – the extracted concepts are analogically extended using similar stored knowledge.

We expose the content of the eventual knowledge base via a query-answering module. It returns answer statements sorted according to their relevance scores and similarity to the query (11). 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. A realistic example of queries and CORAAL answers is given in Section 3.2.

3. Deployment

3.1. *Data*

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: (i) *FEBS Letters*; (ii) *Biochemical Pharmacology*; (iii) *Cancer Genetics and Cytogenetics*; (iv) *Cell*; (v) *Trends in Cell Biology*; (vi) *Experimental Cell Research*; (vii) *Controlled Clinical Trials*; (viii) *Molecular Aspects of Medicine*; (ix) *Advanced Drug Delivery Reviews*; (x) *Gene*; (xi) *Trends in Genetics*; (xii) *Genomics*; (xiii) *Leukemia Research*; (xiv) *Journal of Microbiological Methods*; (xv) *Trends in Microbiology*; (xvi) *Journal of Molecular Biology*; (xvii) *Oral Oncology*; (xviii) *European Journal of Pharmacology*. From the article repository, we extracted the knowledge and publication meta-data 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).

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 (8) describing the input articles. This included $247,392$ publication titles and $374,553$ authors (both from full-texts and references processed). Apart of the triple store, we employed a custom EUREEKA knowledge base (11) 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 later on). 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). The contextual meta-knowledge related to the statements, namely provenance information, amounted to more than $10,000,000$ additional statements (should it be expressed in RDF triples). Query evaluation on the produced content takes usually fractions and at most units of seconds.

3.2. *Asking Queries, Browsing Answers*

In CORAAL, you can ask classical full-text queries, or knowledge-based queries like: `? : type : breast cancer, rapid antigen testing : part`

`of : ? AND ? : type : clinical study, acute granulocytic leukemia : NOT type : T-cell leukemia, p53 : ? : early carcinogenic events, ...` Detailed description of the querying is given in the CORAAL quick-start document at `http://smile.deri.ie/projects/egc/quickstart`.

Answers in CORAAL are presented as a list of either query-conforming statements (for the knowledge-based search), or resources (publication titles, paragraphs or author names for the full-text search). The statement results can be filtered based on their particular elements (e.g., subjects, properties and objects), associated meta-information and the fact whether they are negative or not. The resource results can be filtered according to the concepts associated with them (both extracted and inferred) and additional meta-data (e.g., authors or citations present in the context of the resulting paragraphs). Using the filtering (i.e, faceted browsing), one can quickly focus the set of results only to particular items of interest.

Example result for the `? : type : breast cancer` query (give me all types of breast cancer) is displayed in Figure 2. The result is already focused to subjects *having* benign histological features, or being different from (i.e., not being type of) endometriosis.

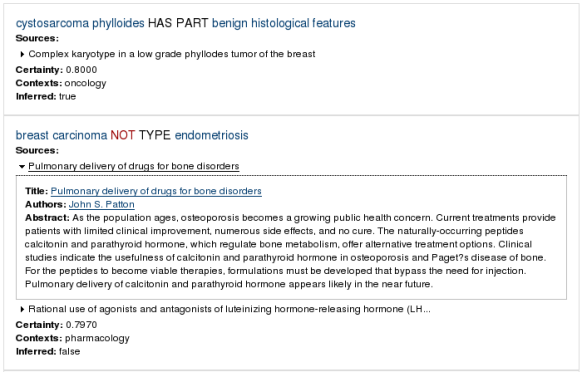

Fig. 2. Focused answer example

The particular types of meta-information associated with statements are: (I) *source* provenance – articles relevant to the statement; (II) *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); (III) *certainty* – a real number meaning how certain the system is that the statement holds and is relevant to the query (values between 0 and 1; derived from the absolute value of the respective statement degree and from the actual similarity of the statement to the query); (IV) *inferred* – a boolean value determining whether the statement was inferred or not (the latter indicating it was directly extracted). More can be checked out at `http://coraal.deri.ie:8080/coraal`.

3.3. *Evaluation Overview*

3.3.1. *Knowledge Quality*

To evaluate the quality of the knowledge served by CORAAL, we picked 100 random concepts and generated 100 random statement queries based on the actually extracted content of the oncological literature knowledge base. We let the domain experts vote on the relevance of respective concept and statement queries to their day-to-day work. We used the ten most relevant ones (five concept-only and five statement queries) to evaluate the answers provided by CORAAL.

We used the traditional notions of precision, recall and F-measure (2) for the answer quality evaluation. Details on how we computed the necessary gold standard with an assistance of the sample users are given in (11). For a base-line comparison, we processed the extracted knowledge using mere incorporation of the respective crisp facts into a state-of-the-art RDF store with inference and querying support. Summing up the evaluation described in (11), CORAAL clearly outperformed the base-line. We computed two sets of measures concerning quality of the answer statements and relevance of the respective provenance articles. The improvement of CORAAL over the base-line was at least two-fold and at most eight-fold for the respective F-measures.

The absolute CORAAL results may still be considered rather poor when compared to the gold standard generated by the users (i.e., F-measures for concept queries around 0.2). However, one must realise that the construction of the gold standard only for the 10 sample queries took almost two working days of an expert committee. The CORAAL knowledge base was produced purely automatically in about the same time for much larger amounts of data involving hundreds of thousands of concepts. The queries take seconds to evaluate and one can find many relevant (even if not all) answers very quickly due to the relevance-based sorting of the results (the first 10 results contained more than 67% of relevant answers in average, while in between the 200. and 400. result, only about 5% were consid-

ered correct). The evaluation committee unequivocally considered the ability of CORAAL to perform purely automatically as an acceptable trade-off for the detected noise in the results.

3.3.2. *Continuous Tests with Users*

Before the final stage of the current CORAAL prototype development, we arranged four biomedical experts as a committee of sample users. We prepared five tasks to be worked out with both CORAAL and a base-line application (ScienceDirect or PubMed). Our hypothesis was that the 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 [1] .

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.

Apart of the positive results of the preliminary CORAAL version, we have got quite some negative comments indicating potentially serious usability problems. We did a detailed analysis of the user feedback collected during the preliminary CORAAL evaluation (before the challenge semifinal) and another workshop with the domain experts (after the semifinal). Two most critical issues were identified: (i) the scattered and loosely integrated presentation of the knowledge search results; (ii) lack of guided knowledge query construction that would take the actual knowledge base content into account. A remedy for these issues was implicitly or explicitly demanded by all the sample users participating in the CORAAL evaluation or (re)development. The former has been addressed by the new back-end and consecutive integral display of the results. The latter has been resolved within the knowledge query builder form with context-sensitive auto-completion Both solutions were unequivocally considered by the sample users as fully implementing their requests. Moreover, due to the improvements and increased intuitiveness of the interface, the users were able to perform up to six-times faster and 40% more efficiently than with the old CORAAL version. They were also able to use the tool after a 2-minute presentation of the query language, relying only on the online contextual help from then on.

Less critical, but still important were requests for extensions of the query language (mainly regarding support for variables in the predicate position allowing for direct exploration of arbitrary relations between particular concepts). The syntax was extended accordingly in the current version.

The expert users also had slight problems with too general, obvious or irrelevant results presented. These concerns were addressed by the following particular improvements: (i) improved *relevance-based sorting* of concepts and statements – more relevant statements present in the top results; (ii) the intuitive *faceted browsing and filtering* functionality of the new interface – 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 improvements 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).

3.4. *Application Areas*

Regarding a particular CORAAL application, we have identified the following possibilities following discussions with our sample users:
(I) *knowledge-based retrieval* of publications – E.g., finding articles describing an arbitrary relation between a cancer type and a gene.
(II) *automated tagging* of articles – Supported by the association of the most relevant topics (i.e., super-type concepts) to publications. Note that the tags come both from the seed thesauri and from the most general concepts in the article corpus. Basically any life science vocabulary can be incorporated for the tagging together with or instead of the currently used NCI and EMTREE thesauri.
(III) rudimentary automated *expert finding* – Filtering of authors based on the topics they write about.
(IV) semi-automated *population of the standard biomedical vocabularies* by the publication knowledge – Directly supported by the integration of the extracted statements into the seed domain thesauri.

[1] 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.

Further manual curation would be needed for the exported content, though, to tackle possible noise. (V) application of CORAAL as a *general-purpose publication knowledge back-end* with arbitrary services implemented on the top of it – E.g., textual entailment service checking for whether the statements present in article A are consistent with the article B and to which extent, or the services extending the other prototypes implemented in the challenge.

The range of possible applications is wider for CORAAL than for any similar system we know of. The CORAAL applicability is extended by its high portability due to the automation of the knowledge extraction, integration and refinement processes. This supports the prototype's indubitable potential for a further development into a truly production system.

4. Related Work

Approaches tackling problems related to those addressed by the core technologies powering CORAAL are analysed in (11; 6). Here we offer an overview of systems targeting similar problems to those tackled by our framework. Figure 3 organises relevant applications in a plot with two axes – *effort* and *benefit* (the placement is only orientational, though, as it does not reflect any formal measure related to the particular systems). The *effort* axis indicates how much more or less manual effort must the creators and/or maintainers of a tool spend before it can perform sufficiently, or before it can be ported to a new domain. The *benefit* axis reflects how much benefit users get when searching for the knowledge hidden in publications with a tool.

Fig. 3. Informative comparison of selected systems

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). 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 (9), Melisa (1) and GoPubMed (5) 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 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. On the other hand, CORAAL works on any domain and extracts arbitrary knowledge from publications automatically, although the offered benefits may not be that high due to possibly higher level of noisiness.

Textpresso (10) is quite similar to CORAAL concerning searching for relations between concepts in particular chunks of text. However, the underlying ontologies and their instance sets have to be provided manually, whereas CORAAL can operate with or even without any legacy ontology. Moreover, the system's scale regarding the number of publications' full-texts and concepts covered is much lower than for CORAAL.

From the overview of the related cutting-edge systems, it is obvious that the biggest challenge is a reliable automation of more expressive content acquisition. Contrary to CORAAL, none of the related systems addresses this problem appropriately, which makes them either poorly scalable, or difficult to port to a new domain. This is why we were not even able to use the related systems for a base-line comparison in our domain-specific application scenario – we simply could not adapt them so that they would be able to perform reasonably, both due to technical difficulties and lack of necessary resources.

5. Conclusions and Future Work

We have delivered a solid and self-contained piece of innovative work in the form of the CORAAL prototype and technologies that power it. We have processed non-trivial amount of data purely automatically and the indicative tests with real users have proven that we are on the right path regarding our vision. Already now we are able to effortlessly extract and process the knowledge hidden in large legacy repositories and offer it conveniently to the users who seek for it.

However, two important things we have foreseen still remain to be accomplished in the long-term perspective:
(I) *Utilising the wisdom of the crowds* – support for intuitive and unobtrusive dynamic 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.
(II) *Making the step from CORAAL to a CORAAL reef* – proposal and implementation of a distributed peer-to-peer model covering multiple CORAAL installations autonomously communicating with each other (e.g., asking for answers when no answer is available locally or exchanging appropriate rules to improve the local semantics).

After incorporating the capabilities of the prospective CORAAL reefs into the ecosystem of the current online publishing, we can instantly realise the exciting future, exploiting the huge body of knowledge scattered over millions of scientific articles out there much more intelligently than possible today.

Acknowledgments This work has been supported by the 'Líon', 'Líon II' projects funded by SFI under Grants No. SFI/02/CE1/I131, SFI/08/CE/I1380, respectively. We acknowledge the help from Ioana Hulpus, who developed the initial user interface for CORAAL. Big thanks goes to our evaluators: Doug Foxvog, Peter Gréll, MD, Miloš Holánek, MD, Matthias Samwald, Holger Stenzhorn and Jiří Vyskočil, MD. We also appreciated the challenge judges' feedback that helped to streamline the final prototype a lot. Last but not least, we acknowledge the prompt and professional support provided by Noelle Gracy, Anita de Waard and numerous other people at Elsevier, B.V. regarding the challenge organisation.

References

[1] J. M. Abasolo, M. Gómez, M.: Melisa: An ontology-based agent for information retrieval in medicine, in: Proceedings of the First International Workshop on the Semantic Web (SemWeb2000), 2000.

[2] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

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

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

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

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

[7] A. Maedche, S. Staab, Discovering conceptual relations from text, in: Proceedings of ECAI 2000, IOS Press, 2000.

[8] F. Manola, E. Miller, RDF Primer, available at (November 2008): `http://www.w3.org/TR/rdf-primer/` (2004).

[9] D. L. McGuinness, Ontology-enhanced search for primary care medical literature, in: Proceedings of the Medical Concept Representation and Natural Language Processing Conference, 1999.

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

[11] V. Nováček, Towards lightweight and robust large scale emergent knowledge processing, Tech. Rep. DERI-TR-2009-06-18, DERI, NUIG, available at `http://tinyurl.com/lsjl6b` (2009).

[12] J. Voelker, D. Vrandecic, Y. Sure, A. Hotho, Learning disjointness, in: Proceedings of ESWC'07, Springer, 2007.