An Information Retrieval Ontology for Information Retrieval Nanopublications*

Aldo Lipani, Florina Piroi, Linda Andersson, and Allan Hanbury

Institute of Software Technology and Interactive Systems (ISIS)
Vienna University of Technology, Austria
{surname}@ifs.tuwien.ac.at

Abstract. Retrieval experiments produce plenty of data, like various experiment settings and experimental results, that are usually not all included in the published articles. Even if they are mentioned, they are not easily machine-readable. We propose the use of IR nanopublications to describe in a formal language such information. Furthermore, to support the unambiguous description of IR domain aspects, we present a preliminary IR ontology. The use of the IR nanopublications will facilitate the assessment and comparison of IR systems and enhance the degree of reproducibility and reliability of IR research progress.

1 Motivation

An important part of information retrieval research consists of running retrieval experiments, beginning with choosing test collections, selecting indexing algorithms, tuning parameters, evaluating outcomes and concluding with publishing summaries of the results in conference or journal articles. A research article, however, "is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete [data] set of instructions which generated the figures" [3]. Making available the necessary components to reproduce IR research results is beneficial for the IR community, most of all to the authors of the published research [6].

The map of availability solutions for IR experiments has currently an island-like geography. Tools like EvaluatIR[1] or Direct[2] concentrate on IR system comparison by examining their outputs in retrieval experiments. Details about the IR systems are not available through these tools and experiments cannot be cited *per se*. The same can be said about the music IR domain where experiments and comparisons of algorithm results are available since 2005[3] or about the myExperiment community dedicated to sharing scientific workflows and packets of research objects, like data and/or algorithms[4].

* This research was partly funded by the Austrian Science Fund (FWF) project number P25905-N23 (ADmIRE).

[1] http://evaluatir.org
[2] http://direct.dei.unipd.it/
[3] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[4] http://www.myexperiment.org/

In life-sciences, where high throughput experiments are not uncommon, the need to publish supplemental information to research articles has led to the development of nanopublications. Nanopublications offer the possibility to publish statements about data and experiments, together with references that establish the authorship and provenance of the statements in a machine-readable format.

The content of a nanopublication is expressed using ontologies which ensure a common understanding on the published statements/assertions. In information retrieval research, ontologies are mostly used to improve the retrieval accuracy in some given domain [10,4]. We present, here, the outline of an IR ontology that can be used in creating nanopublications on statements about IR.

We advocate, thus, the publication of supplemental material for IR publications, in form of IR nanopublications, with the ultimate goal that such publications will make the assessment of research progress on any given IR topic quick and reliable, and significantly improve the reproducibility of IR research results.

We underline that this is preliminary work to present the concept, with changes to our proposed IR ontology and IR nanopublications being expected.

2 Ontology Description

The IR domain is affected by a lack of formality caused, not least, by how research results are published. For example, important information is omitted for the sake of brevity, or because it is considered implicit in the publication's context; or new names for well-known concepts are introduced, making them ambiguous [13, Chapter 1]. It is, therefore, difficult to reconcile results published over longer periods of time. We believe that the design of an IR domain specific ontology is a natural solution to this issue.

With the ontology we describe here[5] we aim at a formal representation of the concepts in the IR domain, establishing a common discourse ground for the publication of (meta-)data that forms the basis of research articles. The IR domain ontology we propose consists of a vocabulary of concepts specific to the IR domain and the relationships between them. With it we want to model the evaluation activities taking place in this domain. This is in line with what a domain ontology should contribute to the respective domain [7].

Our methodology to establish the ontology is a mix of top down and bottom up approaches. First, manually parse a number of publications—more than 50—from the NTCIR, CLEF, and TREC series of publications, as well as by now classic teaching books (e.g. [11]), in order to identify taxonomy categories. Second, on a collection of documents (e.g. the almost 5,000 CLEF publications we have access to) compute the noun phrase termhood (e.g. C-value [5]). And third, manually go through the top terms in the termhood list to create ontology individuals. We present here the outcome of the first step of our methodology, as steps two and three are work in progress. The proposed IR ontology, devel-

[5] http://ifs.tuwien.ac.at/~admire/ir_ontology/ir

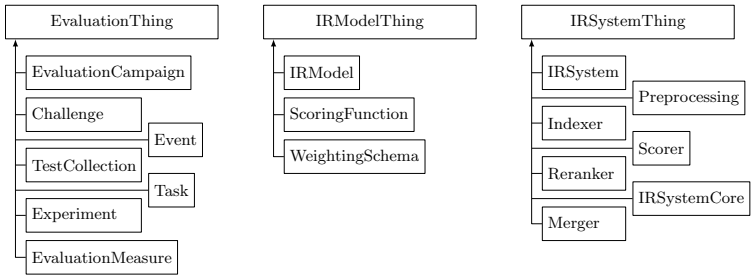

Fig. 1. Fragment of the IR taxonomy

oped using the Protégé framework[6] and the OWL Web Ontology Language, is composed of three sections that represent the following fundamental aspects of the IR research: evaluation, IR models and IR systems. The three main concept categories, are as follows (see also Figure 1):

`EvaluationThing`**:** models concepts like evaluation measures, experiments and evaluation campaigns with their events and challenges;

`IRModelThing`**:** models the theoretical concepts of the IR models with their theoretical definitions of scoring functions, weighting schemata, etc.;

`IRSystemThing`**:** models concepts used in describing concrete IR systems with their constituent parts and components which are usually instances of the theoretical concepts modeled by `IRModelThing` conceps.

The `IRSystemThing` section is closely coupled with the `IRModelThing` and the `EvaluationThing` sections by concept relationships which, first, make explicit the theoretical foundations of the IR systems modelled by the `IRModelThing` section, and, second, explicit a system's assessment and presence in evaluation campaign events modelled by the `EvaluationThing` section.

Modelling IR Evaluation Activities. Evaluating IR systems is an important and well-developed aspect of information retrieval research [9,12] which aims at objective measurements of IR algorithms and technique improvements.

`EvaluationThing`'s subclasses define three related concepts: i) `Experiment` models the execution and assessment of a given IR system on an evaluation setup (collection of documents, a set of queries and a set of relevance judgements, measures); ii) `EvaluationCampaign` models a series of evaluation events, such as the TREC, NTCIR, CLEF, FIRE, and ROMIP; and iii) `EvaluationMeasure` models the measures available for the performance evaluation of an IR system (Precision, Recall, MAP, Precision@n, etc.).

In addition to these three concepts—modelled as `EvaluationThing` subclasses—other concepts are present in this category: the `TestCollection` class, whose elements are `TestCollection` components (`Collection`, `Groundtruth`, and `Topics`), experiment components (`Run`, `TestCollection`, and `Score`), to name a few.

[6] http://protege.stanford.edu, supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

An `EvaluationCampaign` consists of one or more `Event`s (TREC-1, TREC-2, etc.), with each `Event` one or more `Challenge`s are associated (AdHoc Track, Robust Track, etc.). A `Challenge` is an area of research focus aiming to solve a particular retrieval problem. A `Challenge` is part of an `Event` and to each `Challenge` one or more `TestCollection`s can be associated. Evaluation measures are used to assess the performance of a given information retrieval system on a test collection. The `EvaluationMeasure` category models the function and properties of the different measures (parametric, set-based, ranking vs. non-ranking, etc.).

Modelling the IR Model. Models of information retrieval form the foundation of IR research, being the result of theoretical and empirical analysis of specific IR problems. It is this kind of analysis that contributes to the definitions of weighting schemata and scoring functions, as well as to their interpretations.

Weighting schemata, like TF-IDF, LM1, RSJ, etc., are in essence a way of representing selected elements in a collection of documents within an index. Scoring functions provide the means to make comparisons between a given topic and the (previously indexed) collection documents.

In our proposed ontology we list many IR models and weighting schemata, connected with scoring functions.

Modelling the IR System. In the proposed ontology, the `IRSystemThing` category models the structure and the particular software components of an IR system. At the same time, the ontology allows us to express the interplay between an `Experiment`, a `TestCollection`, and the realization of an `IRModelThing` via an `IRSystemThing`. This realization is defined by the relationships between the `IRModel` subclasses (`WeightingSchema`, `ScoringFunction`) and the `IRSystem` subclasses (`Indexer`, `Scorer`, etc.). This design allows us to make explicit IR systems based on more than one IR model.

3 Nanopublications in IR

One of the driving ideas behind nanopublications is the need to disseminate information about experiments and experimental data, and, more importantly, do it in a way that can be attributed and cited. In essence, nanopublications are the smallest unit of publication, containing an assertion which is uniquely identified and attributed to an author [1]. A nanopublication should contain two main parts: an *assertion*, which is the scientific statement of the nanopublication, expressed as a triple <subject, predicate, object>, and the *provenance*, which documents how the statement in the assertion was reached. The provenance usually include supporting meta-data (algorithms, data sets, etc.) and attribution meta-data (authors, institutions, etc.).

Besides the main nanopublication parts mentioned above, there are currently no established standards for the format and additional content of the nanopub-

lications. The Concept Web Alliance[7] advocates a Named Graphs/RDF format which allows to later aggregate nanopublications about some research topic [8].

Below is an example nanopublication describing `IR-Experiment-1` produced by `IR-System-1` running on `TestCollection-1`, with a MAP score of 0.24.

```
@prefix :  <http://www.example.org/nanopub/this-ir-example > .
@prefix np:  <http://www.nanopub.org/nschema# > .
@prefix ir:  <http://ifs.tuwien.ac.at/~admire/ir_ontology/1.0/ir# > .
@prefix pav:  <http://purl.org/pav/ > .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema# > .

: {
 : a np:Nanopublication ;
   np:hasAssertion :IR-Experiment-1 ;
   np:hasProvenance :Provenance ;
   np:hasPublicationInfo :PublicationInfo .  }

:IR-Experiment-1 {
 :Exp-1 a ir:Experiment ;
        ir:hasExperimentComponent :Run-file ;
        ir:hasExperimentComponent :TestCollection-1 ;
        ir:hasScore :MAP-Exp-1 .

 :Run-file a ir:Run ;
           ir:belongsToIRSystem :IR-System-1 .

 :TestCollection-1 a ir:TestCollection .
 :IR-System-1 a ir:IRSystem .

 :MAP-Exp-1 a ir:Score ;
            ir:measuredByEvaluationMeasure ir:MeanAveragePrecision ;
            ir:hasValue 0.24 .  }

:Provenance {
 : pav:derivedFrom <http://dx.doi.org/example/doiID > .  }

:PublicationInfo {
 : pav:authoredBy <http://orcid.org/author-orcid-id > .
 : pav:createdOn "2013-10-02T10:47:11+01:00"^^xsd:dateTime .  }
```

In our view, a collection of IR nanopublications can be used in a natural language question and answering system (Q&A). In this application, the IR ontology will be used as an intermediary layer contributing to the natural language understanding module [2]. Such a system will be able to answer requests like: *'Give me all retrieval experiments which used Solr on the* CLEF–IP *collection and have a MAP score higher than 0.2'.*

The purpose of this system is not just to return a list of papers containing these words. When existing, we want to have also the nanopublications containing additional data about the experiments, the IR indexing and weighting components, the tuning parameters, etc., together with authorship and publication information. The 'Solr' and 'CLEF–IP' named entities can be identified with the help of the ontology, and assigned to their parent classes. Using a reasoner, then, we can instantiate vague concepts. In this example, we can reason that Solr uses Lucene as a search engine and infer that experiments where Lucene was used, but Solr is not mentioned, may be of interest. In the same example we would be able to distinguish between the four versions of the CLEF–IP col-

[7] http://www.nbic.nl/about-nbic/affiliated-organizations/cwa/

lection (2009–2012), each closely related to specific tasks, not all using MAP as an evaluation measure.

4 Future Work

The IR ontology we presented is in its infancy. Our next steps are to extend and consolidate it, validate it through examples, revisiting design phases as needed.

At this phase issues like ontology completeness and maintenance, (central) locations of IR nanopublications, etc. are not dealt with. We expect that discussion rounds with the IR researcher community, either at conferences or dedicated workshops, will contribute towards a solution commonly agreed on. IR nanopublications will then provide means to make experimental data citable and verifiable, as part of the final steps of the operational chain in IR experimentation.

By encouraging researchers in the IR domain to (nano)publish details about their experimental data we encourage them to contribute to their work being reproducible, giving more weight and credibility to their own research statements.

References

1. Guidelines for nanopublication. http://nanopub.org/guidelines/working_draft/. last retrieved: May 2014.
2. James Allen. *Natural Language Understanding (2Nd Ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1995.
3. J. Buckheit and D. L. Donoho. *Wavelets and Statistics*. Springer-Verlag, Berlin, New York, 1995. chapter 'Wavelab and Reproducible Research'.
4. M. Fernàndez, I. Cantador, V. Lòpez, D. Vallet, P. Castells, and E. Motta. Semantically enhanced Information Retrieval: An ontology-based approach. *J Web Semant*, 9(4):434–452, 2011.
5. K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *Int J Digit Libr*, 3(2):115–130, 2000.
6. J. Freire, P. Bonnet, and D. Shasha. Computational Reproducibility: State-of-the-art, Challenges, and Database Research Opportunities. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 593–596, New York, NY, USA, 2012. ACM.
7. A. Gómez-Pérez, M. Fernandez-Lopez, and O. Corcho. *Ontological engineering*, volume 139. Springer-Verlag, 2004.
8. P. Groth, A. Gibson, and J. Velterop. The Anatomy of a Nanopublication. *Inf Serv Use*, 30(1-2):51–56, January 2010.
9. D. Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011.
10. Z. Li, V. Raskin, and K. Ramani. Developing Engineering Ontology for Information Retrieval. *J Comput Inform Sci Eng*, 2008.
11. C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
12. S. Robertson. On the history of evaluation in IR. *J Inform Sci*, 34(4):439–456, 2008.
13. T. Roelleke. Information Retrieval Models: Foundations and Relationships. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(3):1–163, 2013.