# Topic driven access to scientific handbooks

Caracciolo, C.

APPENDIX B

# Glossary

**C-precision** Evaluation measure for retrieved segments, introduced in Chapter **??**. Let $S$ be a retrieved segment. The C-precision of $S$, $C_p(S)$, is defined as the proportion of relevant paragraphs included in $S$:

$$C_p(S) = \frac{|Relevant\ in\ S|}{|S|}, \tag{B.1}$$

where $|S|$ is the total number of paragraphs in $S$.

**DTD** Document Type Definition [XML, 1998]. The purpose of a DTD is to define the legal building blocks of an XML document. It defines the document structure with a list of legal elements.

**EoE** Early onset Error. Evaluation measure for retrieved segments, introduced in Chapter **??**. It measures the proportion of irrelevant material included in the segment preceding the first relevant paragraph ($r_1 - s_1 > 0$).

$$EoE = \min\left\{1, \min_{R:s_1 \leq r_1} \frac{r_1 - s_1}{|S|}\right\}, \tag{B.2}$$

**LoE** Late onset Error. Evaluation measure for retrieved segments, introduced in Chapter **??**. It measures the *proportion of missed relevant paragraphs* at the beginning of the segment. Analogously, if $s_1 - r_1 > 0$ the LoE measures the proportion of relevant paragraphs missed at the beginning of the segment (Figure **??** (c)). More

formally:

$$LoE = \min\left\{1, \min_{R:r_1 \leq s_1} \frac{s_1 - r_1}{|R|}\right\}, \tag{B.3}$$

where $|R|$ stands for the number of relevant paragraphs at the beginning of segment $S$.

**TF.IDF** Term Frequency by Inverse Document Frequency. Weighting schema that balances the weight coming from the number of occurrences of a term in a document, with its frequency in the entire collection [Salton and Buckley, 1988]. One way to balance these frequencies is the following:

$$w_{i,D} = tf_{i,D} * \log\left(\frac{N}{df_i}\right),$$

where $tf_{i,D}$ is the frequency of term $i$ in document $D$, $N$ is the total number of documents in the collection, and $df_i$ is the number of documents containing the term $i$.

**Vector Space Model** In a vector space model, the similarity between a document $D$ and a query $Q$ is computed by means of the cosine similarity:

$$SIM(D,Q) = \frac{\sum_t (w_{t,D} * w_{t,Q})}{\sqrt{\sum w_{tD}^2 * w_{tQ}^2}}$$

where $w$ is a weight assigned to each term (i.e., $_{t,D}$ is the weight assigned to term $t$ in document $D$). The $TF.IDF$ is one of the weighting schema that can be used.

**OKAPI** Weighting schema that does not only considers the frequency of the query terms, but also the average length of documents in the collection and the length of the document under evaluation. It combines IDF weightings with corpus-specific sensitivities to the lengths of the document's retrieved [Robertson and Walker, 1994]. In this thesis we used the BM25 variant of Okapi, which can be expressed as follows:

$$w_i = f(tf_i) * tf_{q,i} * \log\frac{N - df_j}{df_j}$$

where:

$$f(tf_i) = \frac{(k_1+1)tf_i}{K + tf_i},$$

and $K = k_1((1-b) + b * \frac{dl}{avgdl})$ where $dl$ and $avgdl$ are the document length and average document length respectively. $k_1$ and $b$ are global parameters that may be tuned on the basis of evaluation data.

**Precision** Evaluation measure for IR. It gives the proportion of retrieved documents that are relevant:

$$Precision = \frac{|Relevant \cap Retrieved|}{|Retrieved|}$$

**Recall** Evaluation measure for IR. It gives the proportion of relevant documents that are retrieved:

$$Recall = \frac{|Relevant \cap Retrieved|}{|Relevant|}$$

**RQL** RDF Query Language [Karvounarakis et al., 2002, RQL, 2003]. RQL is a query language for RDF and RDFS that allows one to query the RDF and RDFS taken as graphs, and specify edges and nodes for retrieval. RQL showed to have some limitations, for example it does not distinguish variables and URI and does not remove duplicates in the results.

**SeRQL** Sesame RDF Query Language [SeRQL, 2005]. This query language for RDF and RDFS is developed by Aduna as part of Sesame [SESAME, 2005]. It is very similar to RQL, but it addresses some of limitations of RQL.

**XML** The Extensible Markup Language (XML) [XML, 1998] is a *metalanguage*, i.e., a language for describing other languages, which lets one design customized markup languages for limitless different types of documents. Various communities have developed their own XML, including chemistry [CML, 1997] and mathematics [W3C, 2001]. XML is content oriented, so that layout and content issues are separated. The rules of combination of the elements in an XML document can be either implicit in the document, or described in an external document (a Document Type Definition (DTD), or a Schema). The XML data structure is a tree.

**RDF** The Resource Description Framework (RDF) language [RDF, 1999] is a metadata-oriented language whose fundamental concepts are: resources, properties and statements. A statement is an object-attribute-value triple, which makes it especially suitable for encoding metadata. The data model of RDF is the triple, or the graph. This data model allows us to represent objects and their properties in a straightforward manner and it is actually simpler than the XML data model (that only allows strict trees of nested elements).

**RDFS** RDF Schema. RDF allows us to encode complex metadata graphs, but it does not specify the semantics associated with these graphs. In other words, the graph results from a collection of statements without "commitment" to a specific ontology. RDFS [RDFS, 2004] tries to fill that gap by extending the RDF data model in order

to allow hierarchical organization of properties. RDFS adds to RDF the definition of *subproperties*, and the grouping of concepts into *classes*.[1]

**OWL** Web Ontology Language (OWL) [OWL, 2004]. The limitations of RDF and RDFS include the following: properties only have local scope, classes cannot be disjoint, nor can they be combined in a boolean way. Also, transitivity of properties is not expressible, nor is it possible to impose cardinality restrictions to property values. In practice, in order to have a total inclusion of RDFS in OWL, we should allow primitives for "the class of all classes" and for "the class of all properties," which would make the underlying reasoning problems undecidable. For these reasons, OWL was defined as three different sublanguages, with varying expressive power and, consequently, varying computational complexity. OWL Full is fully upward-compatible with RDF, both syntactically and semantically, with the drawback of being undecidable. OWL DL (where DL stands for Description Logics) restricts OWL Full in a way that is fully translatable into a description logic: it loses compatibility with RDF and RDFS, but it allows for efficient reasoning support. Finally, OWL Lite further restricts OWL DL to a sublanguage that is easy to grasp and implement, at the cost of lower expressivity. At the moment, it is considered the most promising language and the best suited to address the special needs of semantically oriented approaches to information management and retrieval.

[1] Despite the name, then, RDFschemas differ somewhat from XML schemas (such as DTD or XML Schema [XMLS, 2001], in that they do not define a permissible syntax, but operate at the semantic level and are therefore appropriate for writing ontologies.