The added value of task and ontology-based markup for information retrieval

Suzanne Kabel
Robert de Hoog
Bob J. Wielinga
Anjo Anjewierden

University of Amsterdam - Social Science Informatics, Roetersstraat 15, 1018 WB Amsterdam.
Email: kabel@swi.psy.uva.nl


Abstract: In this paper we investigate how retrieving information can be improved through task related indexing of documents based on ontologies. Different index types, varying from content-based keywords to structured task-based indexing ontologies are compared in an experiment that simulates the task of creating instructional material from a database of source material. To be able to judge the added value of task and ontology related indexes, traditional information retrieval performance measures are extended with new measures reflecting the quality of the material produced with the retrieved information. The results of the experiment show that a structured task-based indexing ontology improves the quality of the product created from retrieved material only to some extent, but that it certainly improves the efficiency and effectiveness of search and retrieval and precision of use.

Keywords: ontology, information retrieval, reuse, instructional design, task related indexes.
1 Introduction

In this paper we investigate how retrieving information can be improved through structuring indexing using ontologies, and by including task related indexes. Roughly two movements or paradigms can be discerned in research on Information Retrieval (IR), a generic one focusing on improving IR through automatic indexing and statistical methods, the other on controlled vocabularies, coming from traditional library and cataloguing studies. The stream of research focusing on generic solutions led to several ways to derive a set of keywords from a document (Sparck Jones, 1999), different ways to determine the relevance of documents from term frequencies (Salton, 1991; Robertson, 1977), and to techniques such as theme generation and summarizing (Salton et al., 1994). The thesaurus-approach (e.g. Foskett, 1980), on the other hand, has recently attracted renewed attention with the rise of Internet and multimedia retrieval, the latter requiring explicit indexing as text often is absent in for example images, making automatic derivation of keywords as yet impossible. This medium mismatch can be solved using thesauri or ontologies (Meghini et al., 2001). Also in the context of the Semantic web highly semantic, ontology-based annotations play an important role (Berners-Lee, 2001). The Semantic web, by using ontologies as indexing vocabularies, aims to add semantic terms to documents that cannot directly be derived from the content of a document (Schreiber et al., 2001). Since it was recognized that not only matching terms in queries and documents, but also the user and the task he/she performs are involved in an IR system, information retrieval has become an interdisciplinary area of research that increasingly draws on information seeking research (IS) (e.g. Vakkari, 1999). Recently
the role of context, and with that user’s tasks, has received considerable attention. Despite this growing attention, context and tasks are ill-defined concepts, investigated at different levels (see Cool & Spink (2002) for an overview). At the highest level context is seen as a user’s environment, for example an organizational or work task setting, the lowest level is a linguistic level, for example investigating query disambiguation in cases where documents have the right terms but the wrong context (e.g. Lawrence, 2000). Byström et al. (2002) provide a very useful conceptual framework in which a (real-life) work task drives one or more information seeking tasks (i.e. the need for information), which may comprise one or more information retrieval tasks (i.e. consulting a source). They argue that not many researchers have empirically investigated all three task-levels and stress the need for more studies of real life tasks and the development of user-centered evaluation criteria (relevance, utility, usefulness), besides traditional retrieval performance measures of recall and precision (see also Byström, 2002).

The complexity of relevance as an evaluation criterium and the difficulties of operationalising it have been described early by Saracevic (1975), but are still pertinent. Belkin argued that evaluation begins with explicit task specification, in order to identify the criteria users apply in evaluating success (Belkin, 1995). Cool et al. (1993) studied student’s real life tasks of writing an essay and found that characteristics other than topical relevance affect a persons evaluation of a document’s usefulness. Belkin investigated humanities scholars’ ‘everyday’ tasks and goals in order to find out why and how people use texts (Belkin, 1994). Also in this study, he found that secondary characteristics of texts are routinely used for highly conceptual purposes, such as a person
seeking examples of empathy in Middle age paintings. Apart from these interesting findings, these studies investigate relevance judgments made by students, but neglect to measure what people actually do with the retrieved documents, how they apply the information they judged relevant, or how the retrieved material contributed to the result of the work task.

Whereas many studies concentrate on generic aspects of (information seeking) tasks, in pursuit of grasping the changing level of users’ understanding during the course of a search (e.g. Belkin, 1993), we investigate the benefits of explicitly providing support for a single but widely performed work task: composing lesson material. The task of instructional design has been studied extensively (see e.g. Tennyson et al., 1997 for an overview), providing sufficient foundation for a task decomposition (e.g. Merrill, 2000). The typical user group we aim at in our study is heterogeneous, representing distributed (indexers and) retrievers with varying degrees of expertise. We argue that a knowledge rich index, including knowledge of the work task, contributes to better search results, and, more importantly, to a better product created from those search results. Because a person who is interested in some topic is often not yet in a position to specify precisely what is required (Belkin, 1980), structured indexing vocabularies may help users define their problem more accurately, as these represent some (perspective on a) domain or task.

We start with a discussion of theoretic assumptions underlying our research and the use of ontologies as indexing vocabularies in section 2, followed by a description of the indexing framework that was used in the experiment in section 2.2 and an explanation of the experimental design in section 3. The results of the experiment are described in
section 4, and conclusions are drawn in section 5.

2 Theory

Taking Byström’s et al. (2002) conceptual framework of tasks as a starting point, we study different degrees of support for the work task of instructional design and subordinate information retrieval tasks, (the information seeking task remains implicit). For example, if a teacher wants to prepare lesson material to teach a biology class about, say, gorillas he/she can consult an information source containing fragments of teaching materials such as text fragments from encyclopaedias and books, images of animals, sounds etc. The relevance of the information retrieved is determined by the usefulness of the material to the work task of developing the lesson. From this perspective the performance of the information retrieval task is inextricably bound to the work task. Consequently the performance of the retrieval task will depend on taking the context of the work task into account. Let us assume that both tasks are represented by a concept space: in the information retrieval task the concepts represent the content of a domain; in the work task concepts represent the aspects of the task. For example, the teacher who creates a lesson thinks about the domain, e.g. the physical characteristics of a gorilla, but also he/she needs an introduction to that topic, some examples and some questions. The physical characteristics of the gorilla belong to the concept space of the information retrieval task; the introduction, examples and questions belong to the concept space of the work task. These spaces can have an intersection, which indicates how many of the
concepts of the work task are included in the concept space of the information retrieval task (see Figure 1). In addition these concept spaces can vary in the amount of structure. Concept spaces with less structure are flat or alphabetic lists, highly structured concept spaces tend to have an explicit hierarchy reflecting relations between concepts.

Figure 1: Intersection of the information retrieval task concept space and work task concept space

The theory we are testing in this research predicts that the larger and more structured this intersection is, the better both tasks are performed. We hypothesize that content only to a limited extent can form this intersection, because the work task most of the time will comprise more than content. Content is one part of the story; the other part is what you are going to do with this content, or, the role this content will play in the context of the work task. By enriching the concept space of the information retrieval task with concepts that are derived from the work task, the size and structure of the intersection will be increased and performance improved.

The question is what we mean by ‘performance’. We will measure performance on two criteria: effort to carry out the tasks and quality of the result of the work task. Thus more specifically the theory predicts that an increase in the size and structure of the
intersection between the two concept spaces (information retrieval and work task) will lead to a decrease in the effort needed for the tasks and an increase in the quality of the result of the work task, the composition of lesson material in our case. How can the quality of this result increase as a consequence of increasing the size and structure of the intersection? We assume that this is mediated by time. If we fix the total time available for carrying out the information retrieval task and the work task, this total time will be spent on retrieving content, selecting content and composing the instructional material (and of course some ‘noise’.) Disregarding the noise, an increase in the size and structure of the intersection will reduce the time needed for retrieving and selecting because the retrieval results are more to the point. This leaves more time for composition, which in turn will lead to a higher quality of the instructional material.

In this paper different degrees of size and structure of the intersection of the two concept spaces are compared. We compare keyword search with two forms of ontology-based search, one with a domain ontology only, and one with both domain and instructional ontologies. Domain ontologies capture the knowledge about a particular domain, for example gorillas. Keywords and domain ontologies are both means to represent the concept space of the information retrieval task of retrieving content, while the instructional ontology represents the concept space of the work task of constructing instructional material. An experiment was conducted in which subjects composed lesson material from elements in a database. Subjects search in three different conditions: a keyword index condition, a domain ontology index condition and an instructional ontology index condition. The keyword index - a flat list of terms - has the smallest and
the least structured intersection. In the ontology conditions, the structure of the intersection is increased by representing the concept spaces as ontologies, instead of keywords. The size of the intersection is further increased by enriching the concept space of the information retrieval task with concepts that are derived from the work task. Traditional retrieval measures (recall and precision) were adapted to fit the nature of the search task (satisfying instead of exhaustive retrieval) and to reflect not only the information retrieval task but also the work task. Additional measures were developed and used for quality.

2.1 Ontology-based indexing

A way to structure an indexing vocabulary is by using ontologies. Ontology, in the philosophical sense, is a theory about the nature of existence, of what types of things exist. In Artificial Intelligence, an ontology is a formal explicit specification of a shared conceptualization (Studer et al., 1998; Gruber, 1993), or, in IR and Web-researchers jargon, a document or a file that formally defines relations among terms (Berners-Lee et al., 2001). However loosely or strictly defined, the important thing is that an ontology offers a shared, structured and common understanding of some domain or task that can be communicated across people (indexer as well as retriever) and computers.

The ‘closed-vocabulary approach’ to indexing often refers to thesauri, but more recently to ontologies. A thesaurus is a controlled vocabulary, generally with a hierarchical structure, using associative relationships in addition to broader-narrower relationships. The expressiveness of relationships can vary but is normally as simple as A is related to B. Ontologies can be viewed as more complex, consisting of hierarchically
organized concepts with attributes (e.g. synonyms, abbreviations), relations (is-a and part-of), constraints and instantiations that together represent knowledge about a domain or task. The information retrieval task, in a content-based perspective, focuses on a piece of information (what is it about) and is typically described with keywords or thesauri. The work task mirrors a role-based perspective on a fragment (what role can it play in different contexts) and is typically captured in task-ontologies. A task-ontology is pragmatically oriented, applicable across domains, and allows for multiple indexing: a fragment may play more than one role in different contexts. Additional ontology types are defined in van Heijst et al. (1997).

Compared to a set of keywords, an index based on ontologies has two potential advantages: a closed vocabulary yields precise search results and a structured vocabulary supports query abstraction and refinement. There is a downside in the effort needed to develop and maintain ontologies, and in the effort needed to annotate fragments, though indexing can be done automatically to some extent (e.g. Anjewierden et al., 2001). In a different experiment that is carried out at time of writing, we measure the cost-effectiveness of extensive indexing.

2.2 The indexing framework

An indexing framework was developed to index source material (carved up into fragments), consisting of text and pictures. A fragment can be seen as a piece of content indexed by a frame with slots that need to be filled. Ontologies serve as a closed conceptual vocabulary to fill the values of the slots. To ensure flexible retrieval and reuse in an instructional setting, the indexing framework allows describing the fragments
from different perspectives represented in a domain ontology, a description ontology, and an instructional ontology. This leads to three ways of indexing:

- from the domain viewpoint with the domain ontology (in our case about gorillas);
- from the description viewpoint with a description ontology that contains general concepts for describing certain properties of the material that are independent of the domain (e.g. the type of description);
- from an instructional viewpoint with the instructional ontology that can be used to describe the type of knowledge the learner should gain and the role a fragment can play in instruction.

In this way the framework distinguishes between content-based aspects and role-based aspects of fragments. The domain and description ontologies (and of course keywords), allow describing a fragment’s content, and support the IR task, while the instructional ontology typically allows describing the role a fragment plays in an instructional setting, supporting the work task. When we classify the ontologies into different levels of generality, the domain ontology is specific; the description ontology is generally applicable; and the instructional ontology is generic for instructional tasks.

The rationale behind having a specific and a generic ontology describing content-based aspects, combined with a task-specific ontology describing role-based aspects is to enable retrieval and (re-)use in a flexible but structured way using these different indexing vocabularies. Information about a fragment is separated into what a fragment is in general terms; what a fragment is in specific terms; and what one can do with a fragment. Other role-based indexing vocabularies are envisaged, which could represent a
range of other task types.

We use a biological domain, that of gorillas. The main reason for this choice is that it is a publicly known and accessible domain where, when conducting an experiment, differences in prior knowledge of subjects are expected to be minimal. This does not hold for other domains like technical manuals or other specialized material.

A shortened version of the indexing framework is shown in Table 1, and some examples of indexed fragments are shown in Figure 2. The domain ontology allows annotating a fragment on properties of gorillas such as physical, mental, social, behavioral, and communicative properties. The description ontology allows annotating a fragment on description scope, e.g. a general or elaborate description, and on description type. Description type represents the perspective from which the fragment was originally created, like a physical or behavioral description. The instructional ontology allows indexing fragments on knowledge type and instructional role. Knowledge type stands for the kind of knowledge the learner should gain, such as conceptual or factual knowledge. Instructional role stands for the didactical purpose of the fragment or the role(s) it can play in an instructional context, like motivation, description, example, or exercise.
<table>
<thead>
<tr>
<th>Domain ontology</th>
<th>Description ontology</th>
<th>Instructional ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>Description type</td>
<td>Description scope</td>
</tr>
<tr>
<td>Animal</td>
<td>Physical description</td>
<td>General</td>
</tr>
<tr>
<td>Gorilla</td>
<td>Psychological description</td>
<td>Specific</td>
</tr>
<tr>
<td>Mountain gorilla</td>
<td>Social description</td>
<td>Brief</td>
</tr>
<tr>
<td>Animal property</td>
<td>Behavioral description</td>
<td>Elaborate</td>
</tr>
<tr>
<td>Physical property</td>
<td>Biological description</td>
<td>Overview</td>
</tr>
<tr>
<td>Size</td>
<td>Historical description</td>
<td>Detailed</td>
</tr>
<tr>
<td>...</td>
<td>Environmental description</td>
<td></td>
</tr>
<tr>
<td>Mental property</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Learning</td>
<td>Physical description</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Psychological description</td>
<td></td>
</tr>
<tr>
<td>Psychological property</td>
<td>Behavioral description</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Biological description</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Environmental description</td>
<td></td>
</tr>
<tr>
<td>Social property</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>Psychological description</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Behavioral description</td>
<td></td>
</tr>
<tr>
<td>Behavioral property</td>
<td>Biological description</td>
<td></td>
</tr>
<tr>
<td>Playing</td>
<td>Historical description</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Environmental description</td>
<td></td>
</tr>
<tr>
<td>Biological property</td>
<td>Behavioral description</td>
<td></td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Environmental description</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Genetics</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1: Indexing framework**

**Figure 2: Examples of indexed fragments**
3 Method

The hypotheses that follow from the theory, imply that the more indexing is geared towards the user task, the better the results. The experiment was designed to test the hypotheses on attributes that reflect the use of what is retrieved: efficiency, effectiveness, precision of use and quality.

Subjects have the task to create lesson material from a database filled with text and picture fragments about gorillas. They search in three different conditions, reflecting the different size and structure of the intersection between information retrieval and work task concept spaces. Subjects receive an assignment that contains a template in which didactical goals and topics to create the lesson material are set out for them. In addition, subjects have to fill in a questionnaire that determines their prior knowledge.

In the following subsections we describe the design of the experiment: the experimental conditions, the dependent variables and hypotheses, the pilot studies that are conducted before the actual experiment, the subjects, the procedure and the materials used in the experiment.

1.1 Experimental conditions

The subjects are divided in three groups, assigned to different conditions:

1. the keyword condition,
2. the domain index condition,
3. the instructional index condition.

These conditions represent different degrees in size and structure of the intersection of the concept spaces of the information retrieval task and the work task, see Table 2.
In the keyword condition subjects have to search for fragments based on keywords only. Subjects have to think up one or more terms, type them in, and browse the search results.

In the domain index condition, subjects can choose from a hierarchically ordered list of topics about gorillas (the domain ontology). This condition represents a content-based search facility: fragments are searched based on topic.

In the instructional index condition, subjects have an elaborate index at their disposal, enabling them to search not only on topic, but also on description and instructional properties of fragments. Subjects compose their query from a structured list of terms about gorillas, as well as a structured list of terms about description type, description scope, knowledge type and instructional role. This condition represents a role-based search facility: fragments are searched based on topic and on the role it can play in instruction.

Other independent variables are either kept constant or measured. The time subjects have to complete the task of composing lesson material is fixed in order to be able to compare the results between subjects. Subject’s prior knowledge, consisting of knowledge about gorillas, pedagogic experience and knowledge about computers, is assessed in a questionnaire in order to account for possible differences between

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of intersection</td>
<td>Small</td>
<td>Small</td>
<td>Large</td>
</tr>
<tr>
<td>Structure of intersection</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

**Table 2:** Size and structure of the intersection in the three conditions
conditions.

3.2 Dependent variables and hypotheses

Dependent variables are:

1. the *efficiency* of search and retrieval,
2. the *effectiveness* of search and retrieval,
3. the *precision* of use,
4. the *quality* of the lesson material.

It is hypothesized that in the instructional index condition searching and retrieving fragments is more efficient and effective, has a higher precision of use, and results in lesson material of higher quality than in the keyword and domain index conditions. A similar hypothesis is proposed for the keyword and domain index conditions: searching and retrieving fragments is more efficient and effective, has a higher precision of use and results in a higher quality of lesson material in the domain index condition than in the keyword condition. See also Table 3.

<table>
<thead>
<tr>
<th>Hypothesis 1: The instructional index condition yields a) more efficient search and retrieval; b) more effective search and retrieval; c) higher precision of use; and d) higher quality of lesson material than the keyword and domain index conditions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 2: The domain index condition yields a) more efficient search and retrieval; b) more effective search and retrieval; c) higher precision of use; and d) higher quality of lesson material than the keyword condition.</td>
</tr>
</tbody>
</table>

**Table 3: Summary of the hypotheses**

These dependent variables are discussed in more detail below.
3.2.1 **Efficiency and effectiveness of search and retrieval and precision of use**

The traditional measures of retrieval performance are recall and precision. Recall measures completeness of retrieval by dividing the number of relevant fragments retrieved by total number of relevant fragments in the database. Precision measures purity of retrieval by dividing the number of relevant fragments retrieved by total number of fragments retrieved. Recall and precision are adequate measures for exhaustive searching, i.e. ‘give me all documents about X”, while in our relatively small database with a homogeneous set of fragments searching will be performed in a satisfactory way, i.e. “give me at least one document about X that I find satisfying”. Besides, recall and precision reflect only the information retrieval task, but they do not account for the work task. As in this research the emphasis is on the added value of pragmatic or role-based markup, we introduce a third aspect related to retrieval performance: the *use* of what is retrieved.

*Efficiency* is the effort it takes to determine the relevance of fragments from some number of retrieved fragments (hits). The relevance of a fragment is determined by the user by reading it or looking at it. In other words: “How many fragments are looked at?” The assumption here is that a precise query would produce only relevant hits leading to a low number of fragments looked at. The operational definition we will use is:

\[
\text{Efficiency} = \frac{\text{total number of fragments in the database}}{\text{number of fragments looked at}}
\]

A high score on the efficiency variable means that a relatively small proportion of
fragments were looked at until satisfactory and relevant fragments were found.

*Effectiveness* is related to the result or extent to which a goal is achieved. In other words: ‘Of all fragments that are looked at, how many relevant ones are used?’ The operational definition in our context is:

\[
Effectiveness = \frac{\text{number of relevant fragments used}}{\text{number of fragments looked at}}
\]

The optimal situation for effectiveness is one in which all fragments that are looked at are used. The associated maximum value effectiveness can take is 1.

Another aspect of the use of what is retrieved is what we call the *precision of use*. In other words: ‘What percentage of everything that is used is relevant?’ Precision of use depends on precision in the conventional sense (100% precision implies 100% precision of use) and on the users’ fragment selection method. The operational definition we will use is:

\[
\text{Precision of use} = \frac{\text{number of relevant fragments used}}{\text{number of fragments used}}
\]

The optimal situation for precision of use is one in which all fragments that are used are relevant. The associated maximum value precision of use can take is 1.

### 3.2.2 Quality of lesson material

The concept ‘quality’ as such is too general to assess a piece of lesson material. We therefore defined relatively independent quality criteria, together covering quality: motivation, completeness, correctness, coherence, redundancy, and argument structure.

Motivation is a measure that accounts for the attractiveness of the introduction of the
lesson material. With completeness we check to which extent the required topics are covered, and with correctness we look at the relevance of the fragments with respect to the required topics. Coherence measures contradictions in the lesson material, while redundancy covers repetition. Argument structure refers to a structure of the lesson in the form of facts followed by elaborations. The definitions of the quality criteria are given in Table 4.

<table>
<thead>
<tr>
<th>Quality criteria</th>
<th>Definitions</th>
<th>Scoring rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td>The introduction of the lesson material is sufficiently motivating if a text fragment and a picture fragment from the norm set is used.</td>
<td>A matching combination is highly motivating.</td>
</tr>
<tr>
<td>Complete</td>
<td>A chapter in the lesson material is complete if a minimal combination of fragments from the norm set is used.</td>
<td>The use of relevant fragments is rewarded.</td>
</tr>
<tr>
<td>Correctness</td>
<td>A chapter in the lesson material is correct if no fragments that are not in the norm set are used.</td>
<td>The use of irrelevant fragments is punished.</td>
</tr>
<tr>
<td>Coherence</td>
<td>A chapter in the lesson material is coherent if fragments from the norm set are used in a non-contradictory combination.</td>
<td>The use of contradicting fragments is punished.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>A chapter in the lesson material is not redundant if fragments from the norm set are used in a non-repeating combination.</td>
<td>The use of repeating fragments is punished.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>A chapter in the lesson material has a good argument structure if fragments from the norm set are used in a way that facts precede elaboration.</td>
<td>Combinations of fact followed by elaboration are rewarded.</td>
</tr>
</tbody>
</table>

**Table 4: Quality criteria definitions**

Completeness and correctness are *soundness* measures: they measure the extent in which the lesson material contains what it should contain. The remaining criteria (coherence, redundancy and argument structure) are “clarity” or *readability* measures.
They involve combinations of fragments and require reading and possibly editing of the fragments. A subject can only score high for these criteria when two or more fragments are combined. The operational definitions we will use for these two quality dimensions are:

\[ \text{Soundness} = \text{average completeness} + \text{average correctness} \]

\[ \text{Readability} = \text{average coherence} + \text{average redundancy} + \text{average argument structure} \]

### 3.3 Pilot studies

Two pilot studies were conducted. In the first, one subject created lesson material. As a result, the indexing ontologies, the number of fragments in the database, and the assignment were adapted. In the second, 7 subjects created lesson material: 2 in the keyword condition, 2 in the domain index condition, and 3 in the instructional index condition. This led to more extensive and consistently indexed fragments, and to some technical adaptation of the search tool used in the experiment. The pilots were an indication to fix the time to carry out the task to 75 minutes.

### 3.4 Subjects

70 First year psychology students and social science informatics students volunteered to participate in the experiment. Although many of these students have some experience with preparing e.g. Powerpoint presentations, the subjects represent relatively inexperienced instructional designers. We acknowledge that having experienced teachers or instructional designers performing the task would have provided us with data representing the current realistic situation of teachers preparing lessons. However, it is also realistic to expect that not only experienced teachers but also less experienced
retrievers, learners, use semantic vocabularies for retrieval. We aim at representing the heterogeneous group of users with varying levels of expertise as can be expected in a distributed web-context. Subjects were rewarded 25 Dutch guilders (approximately 11 Euro) for participating. Additionally, to motivate the subjects, a bonus of 10 Dutch guilders was promised to subjects who performed well. In fact every subject received the bonus.

3.5 Procedure

The subjects were randomly assigned to the three experimental condition groups. In 8 experimental sessions, subjects worked in the same setting. A session took 90 minutes. In approximately 10 minutes the experimenter explained the task and the tools subjects had to use; the subjects had about 5 minutes to fill in the questionnaire and 75 minutes to create lesson material.

3.5.1 Explanation of the task

Subjects were told that the lesson is meant for first year high school students (age group 11-13 years), and that they have to prepare the lesson material for an inexperienced teacher who will hand it out for the learners to read it, a week before the actual lesson is given. The subjects are also told that the lesson material should be in English (the fragments in the database are in English), and that they should build the lesson from the fragments, instead of adding too much text of their own. Subjects were allowed to edit the fragments and type in some text, but they were to keep that to a minimum. Finally, subjects were told that the more relevant fragments they use the better, provided that they comply with the quality guidelines in the assignment.
3.5.2 Assignment

The assignment subjects received can be seen as a lesson-template they have to fill with fragments. It contains the topics subjects should treat in each chapter, with pedagogic cues such as “motivate the reader”, “give an example”, or “explain this”. Each chapter has to be accompanied with two questions, which can of course also be found in the database. The topics subjects have to treat in the lesson material are:

Introduction.

Chapter 1 Gorilla species.

Chapter 2 Physical characteristics.

Chapter 3 Group composition.

Chapter 4 Communication.

Chapter 5 Threats.

This variation in content was created to make sure that results were not influenced by limitations in the topics that subjects had to address. It is conceivable that the nature of a topic, e.g. Group composition, will generate a bias of used fragments that is due to the availability and nature of fragments. The topics and pedagogic cues are neutrally formulated to make sure the assignment was equally easy or difficult for each subject and only the different degrees of indexing would make a difference in performance.

The following quality guidelines, corresponding to the quality criteria defined in , are given to the subjects:

The introduction should motivate the reader.
The lesson material should be complete: all topics as prescribed in the assignment should be treated.

The lesson material should be coherent: the information in the lesson material should not be contradictory.

The lesson material should not be redundant: the information in the lesson material should not be repeated.

The arguments in the lesson material should be structured in such a way that facts are followed by elaboration.

3.6 Materials

A Prolog database was filled with 250 picture and text fragments that were indexed manually by two persons. The fragments were gathered from different web-sites, treat various gorilla-related topics, and differ in size (many text-fragments were about ½ a page, the largest about ¾ of a page). A search and retrieval tool, designed to fit each condition, was used for searching, browsing and copying the fragments. This application has the same look and feel as an Internet browser. The fragments had to be copied and pasted into an MSWord document. There would not have been enough time for a successful ‘retrieve-all’ strategy, as subjects had to search, retrieve, and view fragments, judge their relevance, copy and paste fragments into MS Word, and furthermore edit the resulting lesson material, all within 75 minutes, which would have left them with approximately 15 seconds per fragment.

The matching technique used in the search and retrieval tool for the keyword condition is full text search including inflections (Baayen et al., 1995), and conjunctive
(AND) queries. Pictures were indexed with a global set of keywords, as IR techniques as yet provide no other solution. Instead of using a fixed list of keywords, the keywords were added to pictures quite freely. Fragments are retrieved from the database based on a match between a term in the query and a term occurring in a text fragment, or in the global set of keywords that pictures are indexed with. Queries with more than one term match only if all terms occur in the (keyword index of a) fragment. Since the database contains relatively few fragments, terms are not weighed. To prevent a bias caused by for example alphabetic presentation order of search results, these are presented in random order.

The matching technique applied in the tools for the domain and instructional index conditions is based on a mapping between fragments that are indexed with terms from one or more ontologies and queries that are composed of terms from one or more ontologies. In the search and retrieval tool for the domain index condition, a query can contain one search term. In the search and retrieval tool for the instructional index condition, a conjunctive query can be composed of a maximum of 5 search terms. Slots that are not filled are treated as wildcards.

Figure 3 to Figure 5 show the search tools for the keyword, domain index, and instructional index conditions, respectively. The three tools have a common functionality in that they contain three browsers: in the browser at the top one can compose and submit a query; in the middle browser (“Found fragments”) one can browse through the hits on a query; and in the browser at the bottom one can view and copy the fragment that is selected in the middle browser. This design tries to eliminate differences between
conditions that could be due to variations in lay out and complexity of the different search tools.

**Figure 3:** Search tool used in the keyword condition

In the keyword condition, subjects search for fragments simply by typing in one or more keywords in the browser at the top. A list of search results is shown in the middle browser. With each found fragment, the title (for example The Mountain Gorilla) and medium of the fragment (text or picture) are shown. Subjects can browse the search results and click on a fragment they consider relevant, and they can view the selected fragment in the browser at the bottom and copy it, if needed, to the authoring environment.
In the domain index condition, subjects search for fragments by selecting a term in the drop down menu in the browser at the top, containing a collapsible hierarchy of terms about gorillas and their properties. The search results are displayed including the title, medium and topic of the fragment.
In the instructional index condition, subjects search for fragments by selecting a term in one or more of the drop down menus in the browser at the top. These menus contain collapsible hierarchies of terms. Subjects can choose from the domain index; the description scope index; the description type index; the knowledge type index; and material use (i.e. instructional role) index. With the search results, the title, medium, topic, type, scope, knowledge and use indexes are shown independently of whether they are used as search terms.

Data were obtained from several sources. Log files are kept for each subject providing information about, among other things, the number of queries, how many succeed or fail, the number of hits (fragments) on a query and the number of fragments looked at. The lessons written by the subjects in MSWord were printed and were used for assessing quality. The number of used relevant and irrelevant fragments is counted (a norm set of fragments prescribes which fragments are relevant). To assess the quality of the lesson material, three coders compare a subjects’ lesson with the norm set of fragments, guided by a coding scheme to assign values to each chapter.

4 Results

Before presenting the analysis of the results, we describe preparations that were carried out on the dataset.

4.1 Data preparation

The data preparation steps that were carried out consisted of a correction for learning curve effect, a check for inter-coder reliability and the removal of outliers.
4.1.1 Correction for learning curve effect

The time subjects had to create the lesson material was fixed. Not all subjects succeeded in fully completing the task of assembling 6 chapters. Because the nature of the search-task differs slightly in the three conditions, a learning curve could influence the quality and quantity of the lesson material. In the instructional index condition subjects have to ‘learn’ or understand the extensive conceptual index and they have to understand how to use it in an optimal way. For example, subjects have to find out that using all five search categories in one query most of the time yields no results due to the relatively small number of fragments in the database (250). A correction for learning curve effect is carried out to prevent this influencing the results. The number of relevant fragments used is corrected by dividing the number of relevant fragments used in all chapters in a certain condition by the number of subjects in that condition minus the number of subjects who did not finish that chapter. Thus, rather than use the absolute number of fragments used, we use a measure of fragments used relative to the amount of material that was produced. The formula is:

\[
\frac{\text{All relevant fragments used in all chapters X in condition Y}}{\text{Total subjects in condition Y} - \text{Subjects in condition Y who did not finish chapter X}}
\]

4.1.2 Inter-coder reliability

The quality of the lesson material was assessed by three coders. To ensure inter-coder reliability, three coders independently assessed the quality of each piece of lesson material using a coding scheme prescribing the rules to assess values for each of the quality criteria and a norm set prescribing which fragments are relevant for a chapter.

The lesson material was compared to a norm set. The norm set consists of unordered
collections of fragments that are relevant for a chapter. For example, in the norm for
chapter 2 all fragments about physical characteristics that occur in the database are listed,
in order for the coders to base their judgments on. The coding scheme then indicates
which fragments are minimally required for a chapter to be judged to be complete.

The coding scheme prescribes that coders give a rating to each chapter in all lesson
materials created by the subjects on a scale from 0 (low) to 10 (high) on completeness,
correctness, coherence, redundancy, and argument structure. Only for the introduction of
the lesson material the single criterion is the extent to which readers of the lesson material
will be motivated by it. The coding scheme contains rules that guide the assessment of
values. For example: if fragment X and Y are present in chapter 3 give a 10 for
completeness; if only X is present give a 5 for completeness; if none of the fragments X
or Y are present give a 0 for completeness. Not every criterion is applicable to every
chapter. In chapter 3 argument structure is not applicable and in chapter 5 coherence is
irrelevant, because the topics and the set of fragments in the database make it impossible
to do badly on these criteria.

As a result from using the coding scheme and the norm set, correlations between the
value assessments of the three coders are high for the most part but also partly low. The
coding scheme is refined and all three coders recoded the first chapter. With the renewed
coding scheme, an average correlation (Pearson) of .75 is found, indicating sufficient
inter-coder reliability. The remaining chapters are recoded accordingly.

4.1.3 Outliers

Of the total of 70 subjects, 9 outliers were removed from the dataset: 3 from the
keyword condition; 2 from the domain index condition; and 4 from the instructional index condition. Outliers are subjects who created lesson material that coders were unable to assign values to (for example when the subject did not follow the lesson template), and of subjects who submitted an empty query (retrieving the entire database) or only one query. One subject in the instructional index condition used only topics in 21 out of 22 queries. This subject behaved as if in the domain index condition, and has therefore been moved there. 61 Subjects remain: 19 in the keyword condition, 24 in the domain index condition, and 18 in the instructional index condition.

4.2 Analysis of results

The differences between subjects’ prior knowledge are not significant, neither is the difference in time subjects worked on a task. Thus results for the three conditions are not influenced by these variables.

Table 5 shows how the three conditions performed on the basic measures from the log files that are part of the equations for efficiency, effectiveness and precision of use.

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average # fragments</td>
<td>136.22</td>
<td>102.21</td>
<td>89.72</td>
</tr>
<tr>
<td>looked at</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average # fragments</td>
<td>21.95</td>
<td>22.83</td>
<td>21.83</td>
</tr>
<tr>
<td>used</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average # relevant</td>
<td>17.42</td>
<td>19.92</td>
<td>19.56</td>
</tr>
<tr>
<td>fragments used</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Averages for number of fragments looked at; used and relevant used per condition
We will deal separately with the related measures (efficiency etc.) and the quality of the lesson material in the subsections below.

4.2.1 Differences in efficiency of search and retrieval

Table 6 shows the number of fragments in the database and for each condition the average efficiency. The higher the average, the higher the efficiency. The highest efficiency is found in the instructional index condition: in this condition it took the least effort to find a relevant fragment.

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragments in database</td>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average efficiency</td>
<td>2.84</td>
<td>3.18</td>
<td>3.70</td>
</tr>
</tbody>
</table>

Table 6: Average efficiency per condition

Although the differences between conditions are large, an F-test shows that these differences in efficiency are not statistically significant (df=2, F=.791, p<.458).

4.2.2 Differences in effectiveness of search and retrieval

Table 7 shows the average effectiveness of search and retrieval, the higher the value the higher the effectiveness. In the instructional index condition, most fragments that are looked at are used, indicating a high effectiveness.

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average effectiveness</td>
<td>.16</td>
<td>.22</td>
<td>.27</td>
</tr>
</tbody>
</table>

Table 7: Average effectiveness per condition

An F-test shows that in general differences in effectiveness are significant (df=2,
F=4.965, p<.010). A post-hoc test (Least-significant difference) shows specifically between which conditions the differences occur. This test indicates that significant differences are found between the keyword condition and the instructional index condition (p<.003).

### 4.2.3 Differences in precision of use

Table 8 shows the average precision of use, the higher the value the higher the precision. The proportion of relevant fragments used is highest in the instructional index condition.

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average precision of use</td>
<td>0.80</td>
<td>0.88</td>
<td>0.90</td>
</tr>
</tbody>
</table>

**Table 8**: Average precision of use per condition

An F-test shows that the differences are significant (df=2, F=4.653, p<.013). A post-hoc test reveals that significant differences in precision of use are found between the keyword condition and the domain index condition (p<.026) and between the keyword condition and the instructional index condition (p<.005).

### 4.2.4 Differences in quality of lesson material

**Differences in motivation**

Table 9 shows the average scores per condition on motivation, measured for the introduction chapter of the lesson material.
<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average motivation</td>
<td>5.09</td>
<td>6.38</td>
<td>8.18</td>
</tr>
</tbody>
</table>

**Table 9:** Average scores on motivation per condition

An F-test shows that the differences between conditions are significant (df=2, F=6.585, p<.003). A post-hoc test reveals that differences between the keyword condition and the instructional index condition are significant (p<.001), and differences between the domain index condition and the instructional index condition are significant (p<.032).

**Differences in soundness and readability**

Figure 6 shows the average scores across all chapters per condition on soundness and readability. To remind the reader: soundness concerns the quantity of required fragments (completeness and correctness); readability stands for the nature of combinations of fragments (coherence, redundancy and argument structure).

**Figure 6:** Average scores on soundness and readability per condition

Across all chapters, differences in soundness and readability are not significant.
It appears that per chapter, the nature of the topic generates a bias due to the availability and nature of the fragments in the database, as was foreseen (see section 3.5.2). The variation in topics treated in the lesson material has accounted for this topic-bias. In chapters 1 and 3 significant differences between conditions are found for soundness, but not for readability. Chapter 1: keyword condition $M=6.53$, domain index condition $M=7.27$, instructional index condition $M=7.42$ ($df=2$, $F=3.130$, $p<.051$). Chapter 3: keyword condition $M=7.06$, domain index condition $M=8.19$, instructional index condition $M=8.19$ ($df=2$, $F=11.134$, $p<.000$). A post-hoc test shows that in chapter 1 the differences are found between the keyword condition and the instructional index condition ($p<.021$), in chapter 3 between the keyword condition and both ontology-based conditions (both $p<.000$).

**Differences in the separate quality criteria**

Across all chapters, differences in each of the quality criteria are not significant, although the differences in completeness and correctness are large, especially between the keyword and instructional conditions (see Appendix I for average scores on quality criteria).

In chapter 3, differences for completeness and correctness are significant. Completeness: keyword condition $M=5.24$, domain index condition $M=6.71$, instructional index condition $M=6.56$ ($df=2$, $F=10.578$, $p<.000$). A post-hoc test shows that differences are found between the keyword condition and the domain index condition ($p<.000$), and between the keyword condition and the instructional index condition ($p<.001$). Correctness: keyword condition $M=8.88$, domain index condition $M=9.67$,
instructional index condition $M=9.81$ ($df=2$, $F=3.894$, $p<.027$). A post-hoc test shows that differences are found between the keyword condition and the domain index condition ($p<.021$), and between the keyword condition and the instructional index condition ($p<.017$).

In chapter 4, differences for correctness are significant. Correctness: keyword condition $M=9.56$, domain index condition $M=9.85$, instructional index condition $M=10$, ($df=2$, $F=4.388$, $p<.018$). A post-hoc test shows that differences are found between the keyword condition and the domain index condition ($p<.047$), and between the keyword condition and the instructional index condition ($p<.006$).

### 4.2.5 Profiles for three conditions

The log files provide a profile for each condition with respect to queries. Table 10 shows the average number of queries per condition; how many of those succeed or fail on average and how many hits a succeeded query yields on average; and how many of those hits are looked at on average.

<table>
<thead>
<tr>
<th></th>
<th>Keyword condition</th>
<th>Domain index condition</th>
<th>Instructional index condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average # queries</strong></td>
<td>51.50</td>
<td>35.67</td>
<td>59.83</td>
</tr>
<tr>
<td><strong>Average # succeeds</strong></td>
<td>33.00</td>
<td>33.33</td>
<td>30.06</td>
</tr>
<tr>
<td><strong>Average # fails</strong></td>
<td>18.50</td>
<td>2.33</td>
<td>21.78</td>
</tr>
<tr>
<td><strong>Average # hits</strong></td>
<td>848.11</td>
<td>1353.96</td>
<td>745.11</td>
</tr>
<tr>
<td><strong>Average # looked at</strong></td>
<td>136.22</td>
<td>102.21</td>
<td>89.72</td>
</tr>
</tbody>
</table>

*Table 10: Average numbers of queries, succeeds, fails, hits and looked at per condition*

What is striking in the keyword condition, is the large number of fails per query
(36%), and a large number of fragments that are looked at. This profile suits the keyword condition. The large number of fails is explained by the fact that subjects have to formulate the terms in a query themselves, leading to more frequent mismatches in terminology than occurs with the availability of ontologies. The large number of fragments looked at is due to the lack of meta-information about content and role of the fragments.

The domain index-condition profile stands out against the keyword condition profile. There were relatively few queries, of which only a tiny minority fails (7%). Terms used in the index perfectly match with terms used in the query. A query cannot be composed: it contains only one search term (topic).

The instructional index-condition profile is characterized by many queries, many fails (36%), and not many hits, not many looked at. As the database contains only 250 fragments, composed queries tend to fail. The query composition task may have required some exercise from the subjects. However, the fact that queries yield not many hits, and that not many hits are looked at, shows that subjects are able to specify quite precisely what they look for.
Table 11: Overview of average scores on performance measures per condition

Table 11 summarizes the results. Significant differences are represented by + (differences between conditions are significant), * (significant difference with instructional condition) and ** (significant difference with both domain and instructional condition). The strong hypothesis 1 (the instructional index outperforms the other two) holds with respect to effectiveness, precision of use and motivation. A trend can be observed that the instructional condition outperforms the other two conditions on all variables except readability, but not all of these differences are statistically significant.

So, hypothesis 1 is rejected with respect to readability. Hypothesis 2 (the domain index outperforms the keyword index) holds only significantly for precision of use, but a trend can be observed that supports hypothesis 2 for other variables, except for readability.

5 Conclusions

Although not all effects are statistically significant, the effects of enriching the concept space of the information retrieval task with concepts that are derived from the
work task are clear: generally an instructional index allows for more efficient and more effective search and retrieval and results in higher precision of use than a keyword and a domain index. Structuring the concept space of the information retrieval task with ontologies also performs better than an unstructured keyword index only: a domain index allows for more efficient and effective search and retrieval and results in higher precision of use than a keyword index.

With respect to quality, three general conclusions can be drawn. One striking difference between conditions is found in the Introduction of the lesson material. Motivation, measured in the Introduction, is clearly highest in the instructional index condition, and motivation in the domain index condition is in turn much higher than in the keyword condition. The Introduction differs from the other chapters in that the assignment for the Introduction consists of a pedagogical cue only (‘motivate the reader’) while no specific topic is required. The single quality criterion on which the Introduction is judged is whether it motivates the reader. The strong effect in the instructional index condition on the quality of the Introduction is evidence that instructional markup is very effective when the information retrieval task and the work task coincide.

The other chapters are of a different character than the Introduction. For these chapters the pedagogical cue in the assignment is less obvious; a topic is required, shifting the focus of the author from pedagogical cues to topics, and the results are judged on multiple quality criteria (completeness, correctness etc.). The results with respect to quality in the remaining chapters are therefore not as marked as in the introduction.

The second conclusion with respect to quality concerns soundness. For both
ontology-based conditions soundness is approximately equal, but they are much higher than in the keyword condition. This proves that using ontologies as closed conceptual indexing vocabularies in relatively small and homogeneous databases improves the use of the retrieved material in terms of completeness and correctness. The small difference between the ontology-based conditions indicates that retrieving fragments with the right topic is sufficient to obtain good soundness.

Finally, for readability, the ontology-based conditions performed less well than the keyword condition. This opposite effect could be explained by the fact that readability is more concerned with how users treat retrieved fragments and less with searching and retrieving fragments. The measures for readability; coherence, redundancy and argument structure, are measures for the use of the retrieved fragments. They are related to the work task of *composing and styling* the overall lesson material rather than to the task of searching and retrieving fragments that reflect their *instructional role*, independent of their location in the lesson material. The instructional markup related to the work task did not account for the readability measures: fragments are not annotated with coherence, redundancy and argument structure-properties. Thus not all concepts in the concept space of the work task are part of the intersection. In order to achieve good readability one could consider adding indexes that can help the author in composing the lesson.

An alternative explanation for readability being higher in the keyword condition than in the ontology-based conditions could be that in the ontology-based conditions more fragments are used than in the keyword condition. Despite the fact that only combinations of fragments are rewarded in the judgment of readability, the use of more fragments
increases the a priori probability to create more incorrect combinations.

This experiment shows that retrieval performance improves in many respects when the vocabulary is structured with ontologies and aspects of the work task are included. In future research we will investigate the extent to which an instructional ontology can offer further support to the work task, e.g., by using instructional scenario’s.

6 References


Acknowledgements

We would like to take this opportunity to thank the anonymous reviewer for contributing to improving this article. This research was partly supported by NWO (Netherlands Organisation for Scientific Research).
# 7 Appendix I Data for the quality criteria

<table>
<thead>
<tr>
<th>Keyword condition</th>
<th>Chapter 1</th>
<th>Chapter 2</th>
<th>Chapter 3</th>
<th>Chapter 4</th>
<th>Chapter 5</th>
<th>All chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>4.32</td>
<td>6.39</td>
<td>5.24</td>
<td>4.00</td>
<td>4.00</td>
<td>4.79</td>
</tr>
<tr>
<td>Correctness</td>
<td>8.74</td>
<td>8.89</td>
<td>8.88</td>
<td>9.56</td>
<td>9.00</td>
<td>9.01</td>
</tr>
<tr>
<td>Coherence</td>
<td>6.11</td>
<td>8.50</td>
<td>6.20</td>
<td>8.67</td>
<td></td>
<td>7.37</td>
</tr>
<tr>
<td>Redundancy</td>
<td>8.44</td>
<td>7.13</td>
<td>5.00</td>
<td>9.33</td>
<td>8.67</td>
<td>7.71</td>
</tr>
<tr>
<td>Argument structure</td>
<td>5.11</td>
<td>4.56</td>
<td></td>
<td>5.92</td>
<td>4.22</td>
<td>4.95</td>
</tr>
<tr>
<td>Soundness</td>
<td>6.53</td>
<td>7.64</td>
<td>7.06</td>
<td>6.78</td>
<td>6.50</td>
<td>6.90</td>
</tr>
<tr>
<td>Readability</td>
<td>6.56</td>
<td>6.73</td>
<td>5.60</td>
<td>7.97</td>
<td>6.44</td>
<td>6.76</td>
</tr>
<tr>
<td>Average</td>
<td>6.54</td>
<td>7.09</td>
<td>6.33</td>
<td>7.5</td>
<td>6.47</td>
<td>6.82</td>
</tr>
</tbody>
</table>

Table 12 Average scores on quality criteria in the keyword condition

<table>
<thead>
<tr>
<th>Domain index condition</th>
<th>Chapter 1</th>
<th>Chapter 2</th>
<th>Chapter 3</th>
<th>Chapter 4</th>
<th>Chapter 5</th>
<th>All chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>5.20</td>
<td>5.38</td>
<td>6.71</td>
<td>4.95</td>
<td>4.75</td>
<td>5.40</td>
</tr>
<tr>
<td>Coherence</td>
<td>6.18</td>
<td>8.44</td>
<td>5.45</td>
<td>8.35</td>
<td></td>
<td>7.11</td>
</tr>
<tr>
<td>Redundancy</td>
<td>7.09</td>
<td>8.50</td>
<td>4.91</td>
<td>8.76</td>
<td>8.92</td>
<td>7.64</td>
</tr>
<tr>
<td>Argument structure</td>
<td>4.09</td>
<td>3.44</td>
<td></td>
<td>5.29</td>
<td>4.69</td>
<td>4.38</td>
</tr>
<tr>
<td>Soundness</td>
<td>7.27</td>
<td>7.17</td>
<td>8.19</td>
<td>7.40</td>
<td>6.84</td>
<td>7.37</td>
</tr>
</tbody>
</table>
Table 13 Average scores on quality criteria in the domain index condition

<table>
<thead>
<tr>
<th>Domain index condition</th>
<th>Chapter 1</th>
<th>Chapter 2</th>
<th>Chapter 3</th>
<th>Chapter 4</th>
<th>Chapter 5</th>
<th>All chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>5.79</td>
<td>6.80</td>
<td>5.18</td>
<td>7.47</td>
<td>6.81</td>
<td>6.47</td>
</tr>
<tr>
<td>Average</td>
<td>6.38</td>
<td>6.94</td>
<td>6.69</td>
<td>7.44</td>
<td>6.83</td>
<td>6.86</td>
</tr>
</tbody>
</table>

Table 14 Average scores on quality criteria in the instructional index condition

The higher the value, the better the score. Maximum score is 10.