Chapter 5

Conclusions and perspectives

The main objective of this thesis project was to propose an indexation system adapted to a dataspace capable to organize resources with respect to their physical location and content. Unlike classic data management system (e.g., relational DBMS), resources within a dataspace are virtually stored in a shared virtual memory offering a global view over a set of hard-disks. This way, the main challenge of a dataspace resides in managing this virtual memory so that resources are always available and managed without regarding about the persistence space capabilities. This indexation level (physical layer) allows to organize in a more efficient way the storage space managed by the dataspace.

Oppose to classic database management systems, resources in the dataspace may lack of data models defining the structure of the data they contain or describing their content. Similarly to search engines, resource's content in a dataspace must be indexed in order to be queried and retrieved. This indexation level (logical layer) provides a global view over the content of resources within the dataspace without regarding of their format (e.g., documents, blogs, Web sites, etc.) and provides capabilities to express queries. However, the semantic of terms composing the index is not described, thus presenting ambiguity among terms during query processing.

To solve this deficiency, a dataspace may be associated to a semantical layer (external layer) describing the meaning(s) of terms with respect to a specific knowledge domain. The external layer is crucial to query processing as it allows the DSSP to guarantee that precise and pertinent resources will be retrieved. The external layer represents a semantical organization of resources similar to a semantic cache, thus reducing the ambiguity presented during query evaluation and eventually optimizing the resource's retrieval task.

This thesis project is a first exploratory attempt to provide a solution to the indexation problem presented in dataspaces. We have identified the necessities related to this problem, particularly the requirement to merge diverse existing techniques from databases and information retrieval. Also, we have analyzed the architecture proposed for a dataspace management system and identified the role of the indexing service. The rest of this chapter focuses con describing the main results obtained during the development of this project and our main contributions.

5.1 Results and contributions

The following list describes the main results of this thesis project:

- **Dataspace state-of-art**. We defined a survey of current works within dataspaces under three main aspects. First, we presented the notion of dataspaces as a new abstraction for data management [] and the first attempts to provide an architecture to build a dataspace management system. Then, our state-of-art described existing works in databases aiming to provide a solution to the problem related to the indexation of resources using diverse levels of abstraction []. Finally, we presented an overview of the *pay-as-you-go* paradigm as an effective alternative to address the problems related to the absence of schemas within a dataspace (*No Schema Approach*) in data integration and query processing.

- **Multi-level index**. We have defined a indexing structure characterizing each abstraction level of a dataspace. From the physical layer described as the set of persistence services composing the storage space; the logical layer represented as a set of terms describing the content of documents in the dataspace; to the external layer defined with a set of ontologies providing a semantical and homogeneous view of the document's content and associated to diverse knowledge domains.

- **Indexation service**. We proposed a mechanism composed by three managers (one for each layer) that are articulated together to: (i) answer queries over the index in a prompt and efficient way, and (i) give maintenance to the indexing structures in order to guarantee the validity of results.

The main contribution of this thesis project is a multi-level indexing strategy that integrates: (i) the physical indexation of resources with adapted data structures and (ii) the semantical indexation based on the resource's content and the organization of concepts with respect to one or multiple knowledge domains. Furthermore, we proposed the notion of neighborhood and cache as a strategies to optimize query evaluation over each indexing layer.

5.2 Perspectives

The perspectives of this project are organized in two main categories: theoretical and experimental. The rest of this section describe these categories.

- **Experimental.** When indexation is addressed, the main challenge is to improve the organization of resources within the persistence support in order to improve the response time related to the document retrieval. Once an indexation strategy has been proposed, it must be evaluated in order to validate its performance when accessing resources. This work has not covered the experimental validation of the multi-level index. To achieve this, we consider it is necessary to explore the definition of an experiment based on social networks like Facebook, or web based encyclopedias like Wikipedia.

- **Theorethical.** It is necessary to define a formal specification (mathematical model) of the indexation layers in order to estimate if these layers allow us to characterize and organize resources of the dataspace in a correct and complete way. Through this specification, it will be possible to define performance measures over the layers to define a cost-based model.