Chapter 1

Introduction

1.1 Context and motivation

The notion of dataspaces arises as a consequence of the increasing growth of data sharing environments (e.g., social networks, web, intranet, etc.), enabling producers and consumers to interchange data through a network. A dataspace is defined as a virtual environment composed by documents supplied by producers, stored in autonomous persistence services, and queried by users with respect to a set of requirements over information.

The main characteristic of data in this environment is the absence of schemas [14] due to the heterogeneity within the structure and semantic of the documents. Documents can be: (i) files accessible through a network providing a direct access to their content (e.g., HTML files, multimedia, blogs, etc.); or (ii) data described by a schema and accessible through a software entity offering querying services (e.g., a DBMS or a Web service).

Clearly, a dataspace cannot be managed by a database management system (DBMS), because a DBMS is oriented to data models (structured or semi-structured) where the data intension is known. For this reason, it is necessary to build a system providing support to the management of dataspaces. This system is known as a dataspace support platform (DSSP).

A dataspace support platform (DSSP) [13] is defined as a set of logic components organized together to manage the dataspace, thus offering a transparent layer to producers and consumers (e.g., users, software entities, etc.). Consumers express requirements over data, information, and knowledge which can be satisfied by the resources of the dataspace. These requirements do not imply only data but operations on data such as integration, consolidation, exploitation or knowledge inference. Besides, obtained results may be required in a particular format such as a set of tuples or an XML document.

Given the characteristics of a dataspace, usually consumers may not have a complete global view over its content, but they may know what they require. A DSSP provides an abstract description of the dataspace's knowledge domains and their associated vocabularies. Like search engines, terms in these vocabularies can be used to retrieve the content of the dataspace.

To illustrate this situation, consider a company having a dataspace fed by its employees via web tools (e.g., blogs, forums, wikis, etc.) and business solutions (e.g. ERP, CRM, DSS, etc.). Intuitively, this dataspace is composed by data sources and applications heterogeneous in structure and intension. Additionally, consider that a finance global director (FGD) requires to obtain information related to "Production leaks for each world region". This requirement involves:

- Determining the concepts evoked by the query and their meaning. For instance, the concept "production leak" is defined as a measure determined by an aggregation data process associated to the production control, while the concept "region" is defined as the dimension for grouping data given the geographic representation of the company.
- Locating the data or data producers within the dataspace providing information to satisfy the requirement.
- Integrating data in order to provide an integrated and organized view over the results to facilitate its exploitation by users.

A DSSP provides concepts that can be used by data consumers to express a set of requirements. Also, the DSSP maintains meta-data describing the semantic relations among concepts within a vocabulary for certain documents. As presented in Figure reffig.DataspaceSources, as consumer's requirements and producer's documents are defined over the dataspace, the DSSP analyzes them to enrich the meta-data it manages (e.g., description, semantic correspondences, etc.), and uses them to refine answers associated to the predefined requirements.

The DSSP must provide meta-data to: (i) represent concepts, (ii) associate these concepts to documents previously annotated, and (ii) locate documents in distributed memory or hard

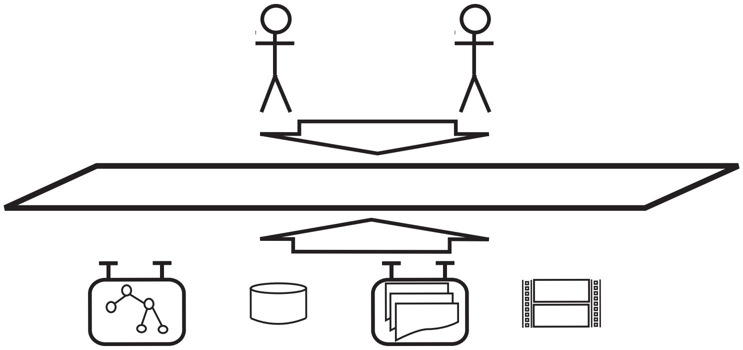

Figure 1.1: Sources feeding the meta-data managed by the dataspace

disks. As meta-data within a dataspace is incomplete, queries are processed using this partial knowledge, therefore leading to approximate answers. Similarly to search engines, approximate answers are not entirely incorrect, as search could be eventually completed if the query is refined or new documents are published in the dataspace.

Nevertheless, the content and meta-data within the dataspace must be organized: (i) physically into storage repositories, and (ii) semantically with respect to the annotations describing the resources' content and structures classifying these annotations. This way, dataspaces must be indexed according to different levels of abstraction: from concepts within an external layer (as an unified view of the dataspace's content) useful to express queries, to the the physical organization of the storage space to provide prompt answers to queries and optimize the storage space required to manage the dataspace's resources.

1.2 Problem statement and main objectives

Given a dataspace, defined as a virtual space composed by documents supplied by producers, stored in autonomous persistence services, and queried by data consumers with respect to particular requirements. It is required to assure the physical (storage media) and logical (according to their content) organization of documents in order to: (i) provide query answering in a pertinent, precise, integrated, prompt, and complete way as it is possible; and (ii) facilitate the maintenance of the dataspace's organization whenever new documents are published.

Indexing techniques used to define the physical organization of hard disks and memory are useful to solve the problem of organizing the virtual space defined over the dataspace. However, unlike classical solutions, the physical storage is not under the control of the DSSP due to the characteristics of the dataspace. This way, persistence must be ensured through a set of autonomous services providing disk space so that they can store a subset of documents from the dataspace. This implies that a document may be distributed or duplicated among different disks, so the dataspace must maintain a global vision of its direction. Routing techniques over distributed virtual memories may be useful to solve this problem.

Physical organization only allows to solve the problem related to the storage and retrieval of documents given a specific address or an identifier. However, this mechanism does not respond to the necessities required by the consumers in order to analyze the documents of the dataspace according to their content. For this reason, it is required to build a description over the dataspace's content. Indexation techniques used in search engines are adequate to solve the problem related to the content representation. However, as these indexation techniques do not consider semantics within the description of keywords, search engines are not capable to deal with the ambiguity presented over concepts within a knowledge domain. As a consequence, queries solved using this type of indexation are not relational queries and provide approximate answers with respect to their precision and recall.

Organizing terms with respect to a semantic network of concepts and relations associated to a knowledge domain may offer an organization composed by mappings between concepts. This way, the DSSP may be able to reduce the semantic ambiguity presented among terms and increase the possibility to obtain complete answers with respect to the content of the dataspace.

The problem addressed in this project resides in modeling a set of indexation structures that abstract the physical (storage space), logical (annotations over a document's content expressed e.g., as term frequency), and external (concepts within a knowledge domain) characteristics of the dataspace; and determining the way they will guarantee the pertinent and precise identification and localization of documents.

The general objective of this project consists in defining a set of indexation structures and functions to manage different levels of abstraction over the dataspace. These index structures must define a global index providing a global physical, logical, and semantical organization of the dataspace's content. Additionally, this global index must support the execution of operations such as query rewriting and data integration.

1.3 Towards an indexing service for dataspaces

An indexation service is a fundamental component of the dataspace support platform (DSSP) giving capabilities to support query rewriting and data integration within the dataspace. Given a query expressed as a disjunction or conjunction of concepts, query rewriting consists in:

- Completing the query expression with neighbor concepts generalizing, specializing or maintaining an equivalence semantic relation with the ones specified in the query.
- Rewriting the query expression by substituting the concepts with their associated terms.
- Identifying the documents associated to each term (logical address within the space) and substitute each term in the query expression with the logical address of its associated documents.
- Locate the documents according to their physical addresses within the persistence services where they are stored and substitute the logical address expressed in the query with the physical address of each document.

Data integration consists on building the answer associated to a query. Based on a set of documents retrieved by the DSSP, the indexing structure is used to organize them according to the terms used to describe them as well as the concepts associated to these terms exhibiting the possible semantic relations among them.

In this project, we propose a solution to the problem related to the indexation of a dataspace by defining a multi-level index organized in three main layers: (i) a **physical layer** specifying the storage space in which the documents persist, (ii) a **logical layer** stating the set of terms describing the documents, and (iii) an **external layer** representing a space composed by concepts and their semantic relations referred by the documents. This multi-level index is managed by an indexing service, which is a component of the DSSP and that participates in the query rewriting and information retrieval tasks presented in the query evaluation process. The main results of this work are:

Multi-level index. We defined a multi-level index offering a set of indexing structures to represent the physical distribution of the dataspace (physical layer) as well as its logical and semantical organization (logical and external layers). These structures are associated with a set of operations that allow the DSSP to exploit them in order to retrieve resources and maintain the dataspace.

Indexing service We proposed a system composed by a set components implementing operations for querying and managing the indexing structures. The indexing service exports an interface towards the DSSP mechanisms and the persistence services managing the storage spaces.

Concept tests We defined a set of use cases that allowed us to reason about the definition of a multi-dimensional index and its use within the query rewriting and data integration. These use cases refer to a business dataspace where the business vocabulary represented through an ontology is assumed. This dataspace contains documents published by its employees and stored in different local and remote persistence repositories. Queries accessing the dataspace aim to retrieve documents as well as to process them to eventually incorporate or retrieve knowledge based on their content.

1.4 Document organization

This document is composed of five chapters including this introduction. Here we give an overview of the content of these chapters.

Chapter 2 presents the state-of-art of most representative works addressing the definition of a dataspace and the main functions of a DSSP. First, the chapter presents the notion of a dataspace, its main functions, and the architecture of a DSSP. Particularly, the chapter focuses on the analysis of existing works addressing the problem related to the dataspace's indexation in different abstraction levels. Finally, the chapter discusses the necessity to explore solutions boarding the integral organization of the dataspace considering the storage devices to facilitate its access and management.

Chapter 3 describes our solution consisting of a multi-level index to organize a dataspace into three main layers: physical, logical, and external. Each layer organizes the dataspace's content with respect to a particular point of view: physical organization into hard-disks and memory, content organization, and a semantical view according to concepts. The relation among layers is explicitly defined, facilitating access to data from users to persistence services trough the query rewriting based on the different levels of indexation. This chapter describes each of these layers and presents a use case to illustrate their functionality.

Chapter 4 presents the general architecture of an indexing service to manage the multi-level index we propose. This architecture is organized in sub-services associated to each layer of the multi-level index. First, we describe the abstract architecture of the service managing each layer and its minimal operations. Then, we present an specialization of this service to satisfy the particular characteristics of each layer in order to provide management and query capabilities. In each layer, we specify a set of management strategies defined with respect to their model in order to manage the indexing structure within a collaborative context among layers. Finally, we present a set of examples describing the use of the indexing service based on the use case presented in Chapter 3.

Chapter 5 concludes this document by discussing our approach. It summarizes our contributions and enunciates several research perspectives.