# A First Tutorial on Dataspaces

Michael Franklin
U. of California, Berkeley

Alon Halevy
Google, Inc.

David Maier
Portland State University

## 1. INTRODUCTION

Dataspace systems offer services on data without requiring upfront semantic integration. In sharp contrast with existing information-integration systems, dataspaces systems offer best-effort answers even before semantic mappings are provided to the system. Dataspaces offer a *pay-as-you-go* approach to data management. Users (or administrators) of the system decide where and when it is worthwhile to invest more effort in identifying semantic relationships. As such, dataspaces offer services on the data *in place*, without losing the context surrounding the data.

The concept of dataspaces was proposed previously in [7, 10]. Dataspaces provide a target system architecture around which we could unify some of the relevant ongoing work in the community. The system architecture also enables identifying additional research challenges for achieving the above goals.

This tutorial will survey the motivations for dataspaces, relevant work in the community, and the progress that has been achieved in the recent years.

## 2. OUTLINE

The tutorial covers the following topics.

### 2.1 Motivation for dataspaces

We begin by describing how datspaces differ from database systems and current information-integration systems. We introduce the following motivating scenarios that are used throughout the tutorial to illustrate the concepts of dataspaces: personal information management, managing scientific data, and data integration on the Web. We also introduce the main logical components of a dataspace system and introduce the services we strive to support with it.

### 2.2 Existing research directions

There are many areas that have been researched in the data management and related communities that are relevant to dataspaces. In this section, we briefly overview some of these and show how they contribute to a dataspace system. We also highlight some of the most important developments in these areas in recent years. These areas include:

- schema matching and mapping
- reference reconciliation
- database profiling
- provenance and lineage
- information extraction
- keyword search on databases

### 2.3 Familiarization and customization

The very front end of dataspace management deals with locating and understanding the available sources, determining what is true about them, and modifying them to make them more suitable to the task at hand.

In this early stage, it is useful to have tools that require no explicit input in terms of schemas or mappings, yet can work at scale with unfamiliar datasets. One such tool is QUARRY, which supports efficient browse and query operations on data where there is no explicit schema (but where there are implicit patterns) [11, 12]. Once initial hypotheses are formed about sources, it is useful to validate those suppositions as a basis for documenting source semantics and determining useful transforms on the data. The InfoSonde framework [12] defines a module structure consisting of three components: A *probe* verifies or deduces a particular property of a a source. The *switch* provides alternative customizations based on the outcome of the probe. One or more *check* routines ensure that any selected customization remains valid in the face of data update.

### 2.4 Schema-less services

We will discuss techniques that can be applied to a data source or combination of sources without an a priori schema for those sources, with particular attention to scalable approaches.

### 2.5 Uncertainty in data integration

Modeling and reasoning about uncertainty in data, queries and semantic mappings are critical for management of dataspaces. Managing uncertainty enables the system to work in a principled fashion when it does not know everything about the data, which is the common case in dataspaces. We cover basic uncertainty formalisms and describe some recent work on modeling probabilistic schema mappings [6] and probabilistic mediated schemas [18].



PVLDB '08, August 23-28, 2008, Auckland, New Zealand

## 2.6 Pay-as-you-go integration

One of the key principles underlying dataspaces is that users and administrators improve the semantic cohesion of the dataspace with time, focusing on the most beneficial efforts. We will present techniques for incremental structure extraction and query from unstructured data [4]. We also discuss techniques based on decision theory for determining where a dataspace system should seek help from users (i.e., how to make pay-as-you-go most effective) [13], and techniques to boostrap a pay-as-you-go dataspace [18].

In another approach, Vaz Salles et al. [17] present iTrails as one style of incremental customization of a dataspace that supports gradual improvement in the degree of source integration. An iTrail is essentially a statement of semantic equivalence of two access paths through the data, and their addition allows queries to obtain more complete answers from a dataspace.

## 2.7 Indexing dataspaces

Indexing collections of heterogeneous data that have not been semantically reconciled raises some interesting challenges. There have been several efforts in the community that considered the problem of schema-less indexing, and we cover some of them here.

# 3. PROMINENT DATASPACES

Finally, we cover several prominent examples of dataspace projects.

**Personal Information Management**: Personal information management was one of the initial motivations for dataspaces. We cover several projects, such as iMemex [2] and Semex [5].

**Personal Health Information**: While hospitals, health plans and clinics are attempting to integrate patient information, it is generally from their internal perspective. A patient interacting with multiple providers may still see a fragmented information space. We illustrate the issues using medication standards in the context of a consolidated medication record.

**eScience**: A scientific research group or community implicitly defines a dataspace of interest, but often one whose conceptual structure is evolving in parallel with scientific understanding of the domain. We will discuss approaches for managing broad classes of scientific information even when the organization of that information is in flux (e.g., [1, 8, 9, 15, 16]).

**The Web**: The web presents one of the most interesting (and rather unique) dataspaces. We describe some recent experience dealing with different versions of this dataspace. In particular, we look at current efforts to crawl the deep web [14], and to analyze large collections of HTML tables on the Web [3]. We also describe recent mashup creation tools that address some of the goals of dataspaces.

# 4. REFERENCES


[1] A. Ananthanarayan, R. Balachandran, R. L. Grossman, Y. Gu, X. Hong, J. Levera, and M. Mazzucco. Data webs for earth science data. *Parallel Computing*, 29(10), October 2003.

[2] L. Blunschi, J.-P. Dittrich, O. Girard, S. K. Karakashian, and M. A. V. Salles. The imemex personal dataspace management system (demo). In *CIDR*, 2007.

[3] M. J. Cafarella, A. Halevy, Z. D. Wang, E. Wu, and Y. Zhang. Webtables: Exploring the power of tables on the web. In *Proc. of VLDB*, 2008.

[4] E. Chu, A. Baid, T. Chen, A. Doan, and J. Naughton. A relational approach to to incrementally extracting and querying structure in unstructured data. In *Proc. of VLDB*, 2007.

[5] X. Dong and A. Halevy. A platform for personal information management and integration. In *CIDR*, 2005.

[6] X. L. Dong, A. Y. Halevy, and C. Yu. Data integration with uncertainty. In *Proc. of VLDB*, 2007.

[7] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. *SIGMOD Record*, 34(4):27–33, 2005.

[8] R. Grossman. Data integration via universal keys. Position paper, Workshop on Information Integration, Philadelphia, PA, October 2006.

[9] R. Grossman and M. Mazzucco. Dataspace: A dataweb for the exploratory analysis and mining of data. *IEEE Computing in Science and Engineering*, 4(4), July/August, 2002.

[10] A. Halevy, M. Franklin, and D. Maier. Principles of dataspace systems. In *Proc. of PODS*, 2006.

[11] B. Howe, D. Maier, and L. Bright. Smoothing the ROI curve for scientific data management applications. In *Proceedings of CIDR*, 2007.

[12] B. Howe, D. Maier, N. Rayner, and J. Rucker. Quarrying dataspaces: Schemaless profiling of unfamiliar information sources. In *Workshop on Information Integration Methods, Architectures, and Systems, Cancun, Mexico*, April 2008.

[13] S. Jeffery, M. Franklin, and A. Halevy. Pay-as-you-go user feedback in dataspace systems. In *Proc. of SIGMOD*, 2008.

[14] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep-web crawl. In *Proc. of VLDB*, 2008.

[15] V. Markowitz. Biological data management in a dataspace framework. Position paper, Workshop on Information Integration, Philadelphia, PA, October 2006.

[16] V. M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, N. Ivanova, and N. Kyrpides. The integrated microbial genomes (img) system: A case study in biological data management. In *Proc. of VLDB*, 2005.

[17] M. A. V. Salles, J.-P. Dittrich, S. K. Karakashian, O. R. Girard, and L. Blunschi. iTrails: Pay-as-you-go information integration in dataspaces. In *Proc. of VLDB*, 2007.

[18] A. D. Sarma, X. L. Dong, and A. Y. Halevy. Bootstrapping pay-as-you-go data integration systems. In *Proc. of SIGMOD*, 2008.