INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

A Survey of dataspaces and indexing

Sonia Sharma[1], Kannu Wadhwa[2]

Jmit Radaur, [1]Assistant Professor (CSE)
soniasharma@jmit.ac.in
Jmit Radaur, [2] Kurukshetra University
kannuwadhwa14@gmail.com

Abstract: *Dataspaces are an abstraction in data management that aims to overcome some of the problems encountered in data integration system. The aim is to reduce the effort required to set up a data integration system by relying on existing matching and mapping generation techniques, and to improve the system in "pay-as-you-go" fashion as it is used. Labor-intensive aspects of data integration are postponed until they are absolutely needed. Traditionally, data integration and data exchange systems have aimed to offer many of the purported services of dataspace systems. Dataspaces can be viewed as a next step in the evolution of data integration architectures, but are distinct from current data integration systems in the following way. Data integration systems require semantic integration before any services can be provided. Hence, although there is not a single schema to which all the data conforms and the data resides in a multitude of host systems, the data integration system knows the precise relationships between the terms used in each schema. As a result, significant up-front effort is required in order to set up a data integration system. This paper reviews dataspaces, some of the prominent dataspaces and indexing dataspaces.*

Keywords: *dataspaces, indexing, schema, attribute frequency.*

1. INTRODUCTION

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. In section 2 we discuss existing search directions, then some of the prominent dataspaces and summarise indexing.

2. EXISTING SEARCH DIRECTIONS

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

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

A DataSpace Storage Platform as per the architectural goal would create efficiently queryable association between data objects in different participants (say documents in different formats) and improve access to data sources that have limited access patterns and to enable answering queries without accessing the actual data source.

3. PROMINENT DATASPACES

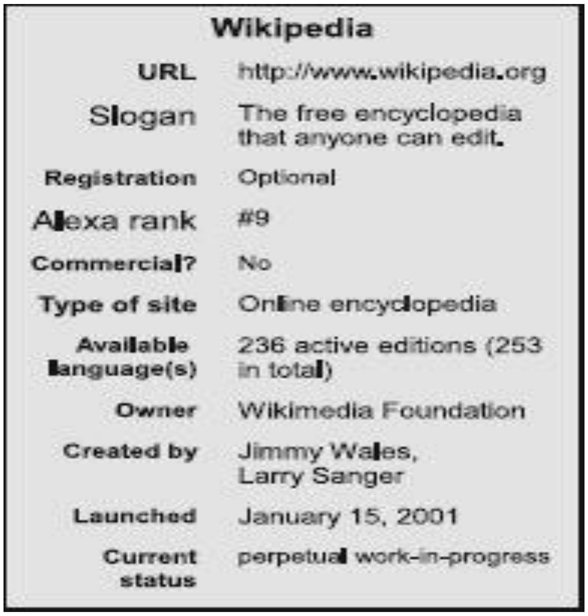
DATASPACE UK : DataSpace UK Ltd specialize in all aspects of document management and archive storage whether your organization is large or small privately owned or central government run storing paperwork, data, media tapes or electronic images; services to assist organizations and effectively manage documents and archive storage workflow. The head office is in the North West of England but no matter where you are based be it Warrington, Chester, Manchester, Liverpool, London, Scotland – pretty much anywhere in the UK, DataSpace can manage your document and archive storage smoothly and efficiently. Also a secure online system called FileLive has been developed where it is possible to track and trace physical documentation and securely link to the scanned image of the documents online. [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.

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.

4. INDEXING DATASPACES

Data records extracted from Web are often heterogeneous and loosely structured 9]. The concept of dataspaces is proposed in [1, 6] which provides a co-existing system of heterogeneous data. Instead of traditionally integrating the heterogeneous data into an uniform interface in advance, dataspaces offer a best-effort integration in a pay-as-you-go style [10,11]. In particular, the semantic schema of dataspaces may not be fully explored or identified. Therefore, dataspace systems rely on schema less services to access the data before identifying semantic relationships [2]. Examples of these interesting dataspaces are now prevalent, especially on the Web [3]. In Wikipedia,1 each article usually has a tuple with some attribute value pairs to describe the basic structured information of the entry. For instance, as shown in Figure 1, a tuple describing the Nikon Corporation may contain attributes like{founded: Tokyo Japan (1917); industry: imaging; products: cameras . . . }. The attributes of tuples in different entries are various (i.e., heterogeneous), while each tuple may only contain a limited number of attributes (i.e., sparse). Thus, all these tuples form a huge dataspace in Wikipedia.

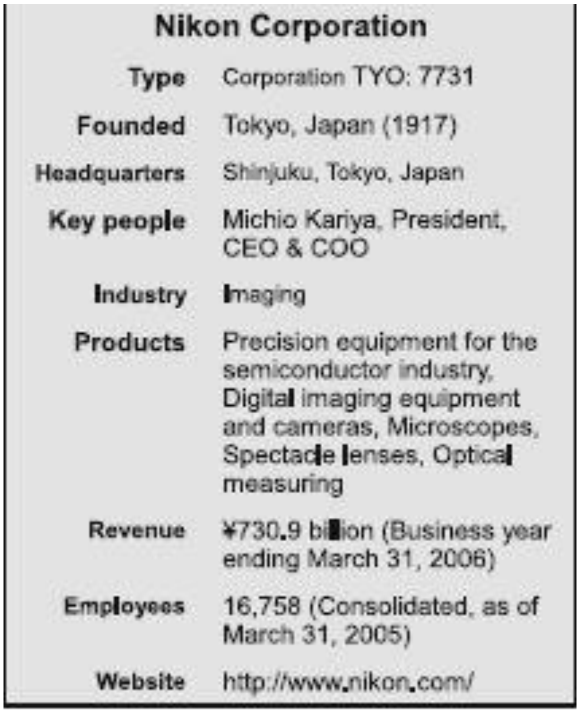


Wikipedia

URL	http://www.wikipedia.org
Slogan	The free encyclopedia that anyone can edit.
Registration	Optional
Alexa rank	#9
Commercial?	No
Type of site	Online encyclopedia
Available language(s)	236 active editions (253 in total)
Owner	Wikimedia Foundation
Created by	Jimmy Wales, Larry Sanger
Launched	January 15, 2001
Current status	perpetual work-in-progress

Figure 1: Tuple from Wikipedia

Given another example, Google Base2 is a very large, self-describing, semi structured, heterogeneous database. As shown in Figure 3, each tuple consists of several attribute value pairs and can be regarded as a tuple in the dataspaces. According to observations, there are total 5,858 attributes in 307,667 tuples (random sample items), while most of these tuples only have less than 30 attributes individually. Therefore, the dataspace is extremely sparse.



Nikon Corporation

Type	Corporation TYO: 7731
Founded	Tokyo, Japan (1917)
Headquarters	Shinjuku, Tokyo, Japan
Key people	Michio Kariya, President, CEO & COO
Industry	Imaging
Products	Precision equipment for the semiconductor industry, Digital imaging equipment and cameras, Microscopes, Spectacle lenses, Optical measuring
Revenue	¥730.9 billion (Business year ending March 31, 2006)
Employees	16,758 (Consolidated, as of March 31, 2005)
Website	http://www.nikon.com/

Figure 2: Tuple describing Nikon Corporation

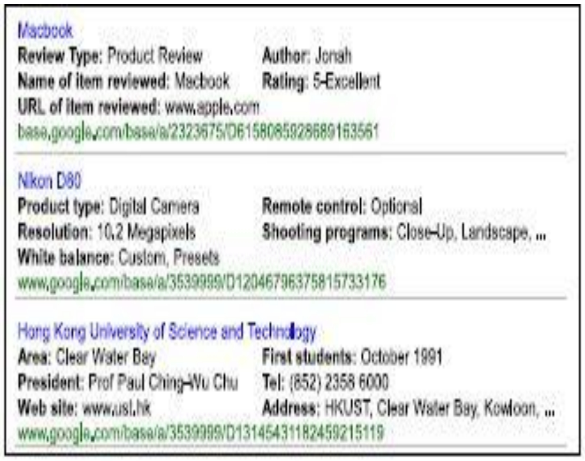
The traditional tables in a relational database mainly focus on specific domains, with a limited number of attributes or columns, while dataspaces are in universal agents like the Web, with loose or even no limitations on the attributes [8]. Therefore, comparing with the traditional tables, we intuitively conjecture the characteristics of dataspaces.

– **Heterogeneous** Rather than a specific domain, the items come from universal areas. The contents, e.g., attributes and values in the tuples, are various.

– **Sparse** Although the entire dataspace is in a high dimensions space of attributes,

each single tuple may only have a very small set of attributes.

– **Large scale** The data might be contributed from the Web around the world. Up till 2008, there have been 2,330,427 articles in English Wikipedia. The queries over dataspaces can be abstracted to two typical operators, i.e., and or queries, and the query returns the candidate answers that satisfy all the query predicates. For example, a query may specify predicates like industry = imaging and products = cameras. The returned tuples should contain these two values in the corresponding attributes, respectively. Moreover, the query over dataspaces may not only searches the specified attributes, e.g.



Macbook
Review Type: Product Review Author: Jonah
Name of item reviewed: Macbook Rating: 5-Excellent
URL of item reviewed: www.apple.com
base.google.com/base/a/2323675/D6158085928689163561

Nikon D80
Product type: Digital Camera Remote control: Optional
Resolution: 10.2 Megapixels Shooting programs: Close-Up, Landscape, ...
White balance: Custom, Presets
www.google.com/base/a/3539999/D12046796375815733176

Hong Kong University of Science and Technology
Area: Clear Water Bay First students: October 1991
President: Prof Paul Ching-Wu Chu Tel: (852) 2358 6000
Web site: www.ust.hk Address: HKUST, Clear Water Bay, Kowloon, ...
www.google.com/base/a/3539999/D13145431182459215119

Website attribute in a query Website = www. wikipedia.org, but also the heterogeneous attributes correlated to the specified attributes, 3 such as URL. That is, the query returns tuples with either Website = www.wikipedia.or or URL=www.wikipedia.org. This or query returns the candidate answers that may only satisfy some predicates.

Finally, the returned candidate tuples are then ranked according to their (similarity) scores to the query. Dataspace queries can be optimized e.g., by using materialization [10] or semantic data dependencies [11]. As a complementary aspect of query optimization, in this paper, we also focus on the indexing of dataspaces.

The attribute frequency in the dataspaces follows the Zipf law like distribution. This interesting observation motivates us to explore the successful inverted index in information retrieval to manipulate the dataspaces. Each pair of attribute label and value maps to a token, then the inverted index can be extended to manipulate dataspaces. However, this basic technique, which also serves as a baseline approach in this study, falls short in the efficiency consideration. Given a query, all the referred tuple lists have to be merged for ranking, which is quite expensive in terms of I/O and CPU cost. In fact, the main bottleneck of inverted list like index is the time cost to merge the tuple lists of large sizes. Rather than the whole bunch of result tuples, a typical query may only be interested in the top-k answers in the real application [4]. Thus, the aggregating and ranking for those low score tuples are wasting.

The key concept used here is indexing of dataspace using a tool and the words that are searched the most is put in cache thus the next time it is accessed it is accessed from the cache thus reducing the access time and improving the performance. Also cache replacement policy is used which makes room for incoming data thus keeping only those words in cache that are most frequently accessed.

5. CONCLUSION

The notion of dataspaces is an abstraction arisen from the requirements presented over recent years after the development of multiple environments such as web, business, and government and in general any environment in which great amounts of data are produced and consumed.

Dataspaces shift the emphasis to a data co-existence approach providing base functionality over all data sources, regardless of how integrated they are. For example, a DataSpace Support Platform (DSSP) can provide keyword search over all of its data sources, similar to that provided by existing desktop search[1-2] systems. When more sophisticated operations are required, such as relational-style queries, data mining, or monitoring over certain sources, then additional effort can be applied to more closely integrate those sources in an incremental fashion. Similarly, in terms of traditional database guarantees, initially a dataspace system can only provide weaker guarantees of consistency and durability. As stronger guarantees are desired, more effort can be put into making agreements among the various owners of data sources, and opening up certain interfaces (e.g., for commit protocols).

REFERENCES

[1] Michael Franklin, Alon Hacvey, David Maier "From databases to dataspaces, A new abstraction for information management" ACM SIGMOID Record December 2005

[2] Michael Franklin, Alon Halvey, David Maier "A First Tutorial on Dataspaces"

[3] Xin Dong, Alon Halvey, "Indexing dataspaces,"

[4] Shaoxu Song, Lei Chen, "Indexing Dataspaces with partitions", Springer, Volume 16, Issue 2, pp 141-170

[5] Wikipedia, Dataspaces http://en.wikipedia .org/wiki/Dataspaces

[6] Alon Halevy, Michael Franklin, David Maer "Principle of database system"

[7] Wikipedia, Pyloric Stenosis, http://en.wikipedia.org/wiki/Pyloric_stenosis, retrieved June 15, 2008.

[8] Li, Q., Chen, J., Wu, Y.: Algorithm for extracting loosely structured data records through digging
strict patterns. World Wide Web 12(3), 263–284 (2009)

[9] Jeffery, S.R., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go user feedback for dataspace systems. In: SIGMOD Conference, pp. 847–860 (2008)

[10] Salles, M.A.V., Dittrich, J.-P., Karakashian, S.K., Girard, O.R., Blunschi, L.: Itrails: pay-as-

yougo information integration in dataspaces. In: VLDB, pp. 663–674 Sarma,

[11] A.D., Dong, X., Halevy, A.Y.: Bootstrapping pay-as-you-go data integration systems. In: SIGMOD Conference, pp. 861–874 (2008)