

DBKDA 2011 Panel Discussion:

Will Dataspaces Make Data Integration Obsolete?

Moderator:

Fritz Laux, Reutlingen Univ., Germany

Panelists:

Kazuko Takahashi, Kwansei Gakuin Univ., Japan

Lena Strömbäck, Linköping Univ., Sweden

Nipun Agarwal, Oracle Corp., USA

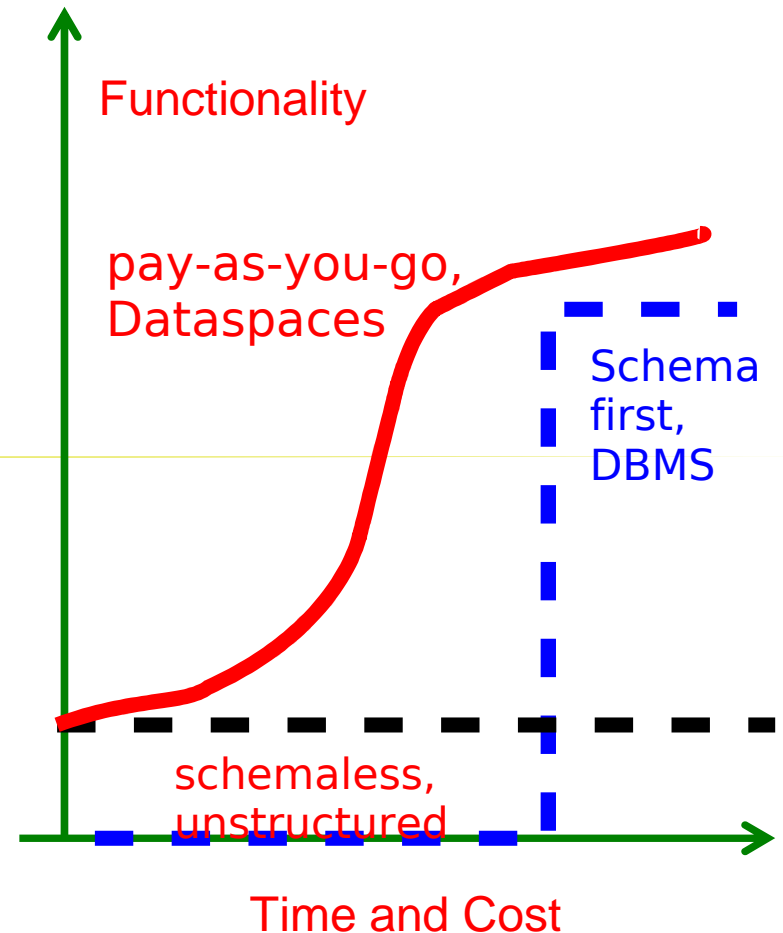
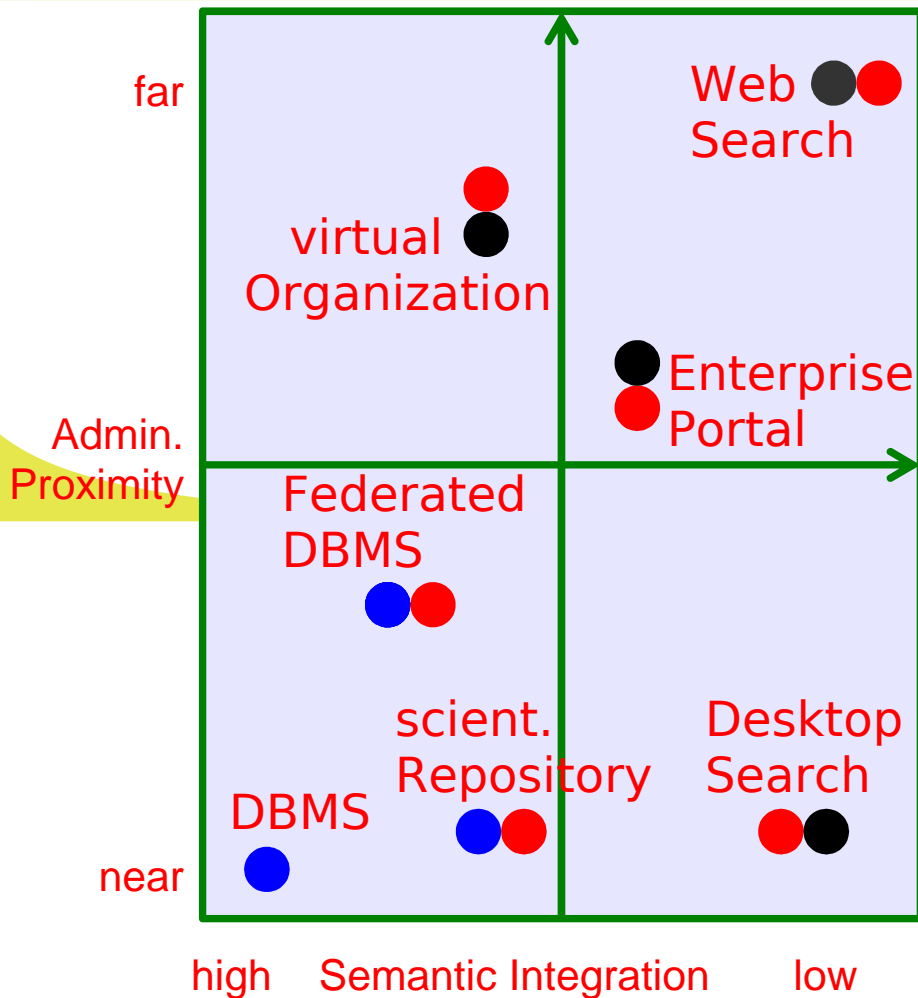
Christopher Ireland, The Open Univ., UK

Fritz Laux, Reutlingen Univ., Germany

The Dataspace Idea

Space of Data Management

Scalable Functionality and Costs



Dataspaces (DS) [Franklin, Halevy, Maier, 2005] is a new abstraction for Information Management

- **DS** are [paraphrasing and **commenting** Franklin, 2009]
 - **Inclusive**
 - Deal with all the data of interest, in whatever form => **but semantics matters**
 - **We need access to the metadata!**
- **derive schema from instances?**
- **Discovering new data sources => The Münchhausen bootstrap problem?**



Theodor Hosemann
(1807-1875)

Dataspaces (DS) [Franklin, Halevy, Maier, 2005] is a new abstraction for Information Management

- **DS** are [paraphrasing and **commenting** Franklin, 2009]

- **Co-existence not Integration**

- No integrated schema, no single warehouse
=> **but ad-hoc matching/mapping required**
- no ownership required
=> **data provenance available?**
=> **availability, reliability?**

- **How to deal with inconsistent data?**

- **Can ontologies help with mapping?**

Dataspaces (DS) [Franklin, Halevy, Maier, 2005] is a new abstraction for Information Management

- **DS** are [paraphrasing and **commenting** Franklin, 2009]
 - **Pay-as-you-go**
 - Keyword search is bare minimum => **how about semantics?**
 - More function and increased consistency as you add work => **interesting: better quality at higher costs?**
 - **How about serious analytics with keyword search?**
 - **What does „better quality“ mean? metrics?**

Statements summary

Kazuko Takahashi: Semantic integration still necessary as basic techniques

Lena Strömbäck: How much can data spaces reduce the need for data integration?

Nipun Agarwal: XML enhanced DBMS technologies will make data integration easier

Chris Ireland: Cost of building a dataspace over time vs up-front cost of integrated data?

Fritz Laux: Dealing with all data of interest, but what is with its semantic?

Will Dataspaces make Data Integration obsolete?

Chris Ireland

The Open University, UK

The Literature

- “Dataspace management is not a data integration approach; rather, it is more of a data co-existence approach” [Halevy]
- “A dataspace must perform operations to reconcile differences in representations of information” [Arnold]
- “How to locate all the relevant data and relationships between them” [Podolecheva]
- “The benefits of classical data integration with reduced up-front costs combined with opportunities for incremental refinement, enabling a pay-as-you-go approach” [Hedeler]

A difference...

- Data integration requires up-front identification of relationships, in a dataspace this is done over time (pay-as-you-go) [Jeffery] [Franklin]
 - But... Initialisation of a dataspace requires up-front work [Hedeler]
 - Techniques for identifying and reconciling differences may be shared?
 - Cost of building a dataspace over time vs up-front cost of integrated data?

DBKDA2011 PANEL

Will Dataspaces Make Data Integration Obsolete?

-- from the viewpoint of AI --

Jan 26, 2011

Kazuko TAKAHASHI

Kwansei Gakuin University

My background

- Artificial Intelligence
 - Knowledge Representation and Reasoning
 - Spatial&temporal representation/DB
 - Not a Database specialist !

My Talk

- My current research
 - A qualitative spatial reasoning
- My opinion on dataspace & data integration

Qualitative Spatial Reasoning (QSR)

- A method that treats images or figures **qualitatively** by extracting the information necessary for a user's purpose
- Useful for the recognition and analysis of physical phenomena, explanation of causality, diagnosis ...

relative size, relative positional relation, ...
~~using coordinates~~

Examples of qualitative data representation

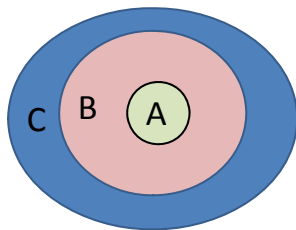
- A and B are connected (relative positional relation)
- A is located in the north-west direction of B (relative direction)
- B is farther than C from A (relative distance)
-

Data abstraction level

- Higher level
 - A and B are connected
- Lower level
 - A and B are connected by a point
 - A and B are connected by two points
 - A and B are connected by a line
 -

Example of qualitative reasoning

- $P(A,B)$ and $P(B,C)$ implies $P(A,C)$



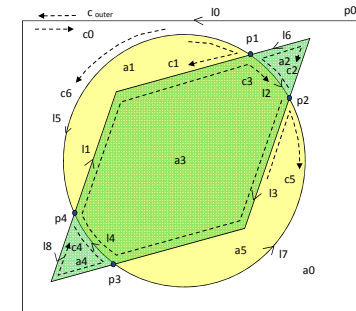
Representation on lower abstraction level

```

e.points = { p0, p1, p2, p3, p4 }
e.lines = { l0, l1, l2, l3, l4, l5, l6, l7, l8 }
e.circuits = { c_outer, c0, c1, c2, c3, c4, c5, c6 }
e.areas = { a0, a1, a2, a3, a4, a5 }
e.outermost = c_outer
l0.points = [ p0, p0 ]
l1.points = [ p4, p1 ]
l2.points = [ p1, p2 ]
l3.points = [ p2, p3 ]
l4.points = [ p3, p4 ]
l5.points = [ p1, p4 ]
l6.points = [ p2, p1 ]
l7.points = [ p3, p2 ]
l8.points = [ p4, p3 ]
c_outer.lines = [ l0+ ]
c0.lines = [ l0- ]
c1.lines = [ l1-, l5- ]
c2.lines = [ l2-, l6- ]
c3.lines = [ l1+, l2+, l3+, l4+ ]
c4.lines = [ l4-, l8- ]
c5.lines = [ l3-, l7- ]
c6.lines = [ l5+, l8+, l7+, l6+ ]
    
```

```

a0.circuits = { c6, c0 }
a1.circuits = { c1 }
a2.circuits = { c2 }
a3.circuits = { c3 }
a4.circuits = { c4 }
a5.circuits = { c5 }
    
```



PLCA expression

Spatial data integration

- It is hard to integrate these data bases
 - with different abstraction level

Dataspaces (WIKI)

An abstraction in data management

An evolved form of data integration

To overcome some of problem in data integration

- to reduce the effort required to set up a data integration system
 - by relying on existing matching and mapping generation techniques,
- to improve the system in pay-as-you-go fashion as it is used

Data integration (WIKI)

- combining data residing in different sources and providing users with a unified view of these data
- Semantic integration is needed
 - Conflict resolution
 - Using ontology

Machine learning in AI

- inductive learning, concept formation, data mining, rule mining
 - find a rule/concept from massive data (generalization)
 - classify a new data or derive a property of a new data by using this rule/concept
 - need much time on the first phase [data integration](#)
- case-based reasoning
 - store a massive data as a set of instances
 - classify a new data or derive a property of a new data directly using some of these data
 - generalization rules are still used [dataspaces](#)

Will Dataspaces Make Data Integration Obsolete?

- Dataspaces:
 - larger amount of data, more changeable
 - reasonable – use only necessary functions only on the time they are required
 - semantic integration is inevitable
 - Conflict resolution
 - Generalization
- These techniques used in data integration is still necessary as basic techniques on handling dataspace



ORACLE[®]

Nipun Agarwal
Director, XML Development

Database Division



Status

- Data Integration
 - Lots of data & sources
 - Schema first approach
 - Requires semantic understanding of various sources
 - Very expensive
 - Very important
 - Very difficult
- Dataspaces
 - Data co-existence approach
 - Provides base functionality over all data sources
 - Best effort result
 - E,g, Search



Vision

- Complementary set of use cases
- Businesses willing to invest upfront if needed
- Need transactional semantics
- Various industry standards promoting data integration
- XML based standards ease the need of a fixed schema
 - XBRL
 - HL7
 - FPML

Will Data Spaces Make Data Integration Obsolete? DBKDA Panel 2011

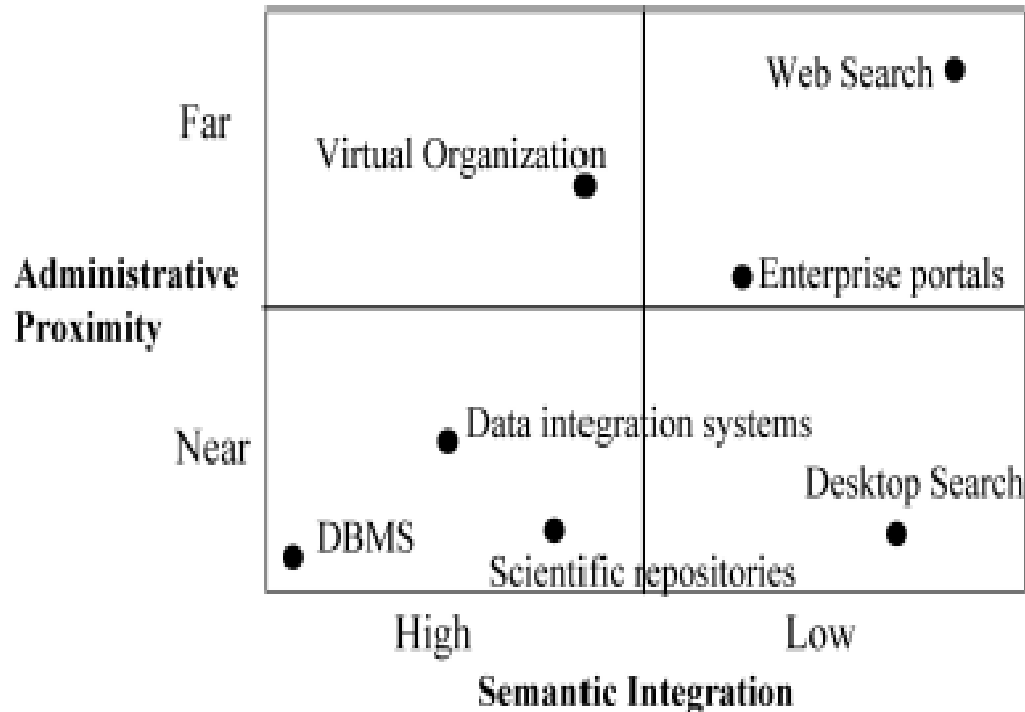
Lena Strömbäck
lena.stromback@liu.se

Linköping University

Why Data Spaces?

Common formats is a prerequisite for efficient data management

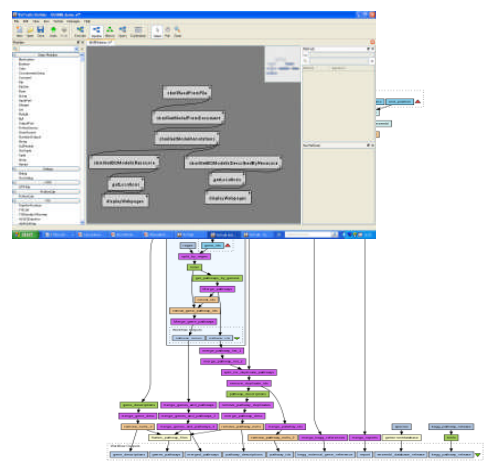
On the web new data formats and actors occurs all the time



From Databases to Dataspaces: A New Abstraction for Information Management
Michael Franklin, Alon Halevy, David Maier (ACM SIGMOD 2005)

Available solutions - Data space technology?

Provenance



Pay as you go

Entity	Action
Species - beta-galactosidase	
Species - allolactose	
Species - lactose_internal	
Species - permease	
Species - ParI-mRNA	
Species - ParII-beta-galactosidase	
Species - ParI-permease	
Species - External_Lactose	
Reaction - basal_mRNA_synthsis	
Reaction - mRNA_Degradation	
Reaction - allolactose_controlled_mRNA_synthsis	
Reaction - allolactose_controlled_ParII_mRNA_synthsis	
Reaction - Beta-galactosidase_Degradation	
Reaction - Beta-galactosidase_synthsis	
Reaction - ParI_beta-galactosidase_synthsis	
Reaction - Beta-galactosidase_reaction	
Reaction - lactose_Degradation	
Reaction - Lactose_transport_Lact	
Reaction - Lactose_transport_Lin	
Reaction - permease_Degradation	

Broker resources

The image shows two web portals. The top one is the MIRIAM (Meta-Interoperable Resources) portal, displaying search results for 'Reaction'. The bottom one is the BioCatalogue portal, which provides a curated catalogue of life science web services. It features a search bar, navigation links, and several service cards for 'DISCOVER', 'REGISTER', 'ANNOTATE', and 'MONITOR'.

Questions

- Are these resources parts of data space technologies?
 - Provenance/lineage
 - Broker resources
 - Pay as you go
- What else is needed?
- How much can data spaces reduce the need for data integration?
 - Many user that work together on smaller problems
 - Technologies that aids and reduces the effort
- Application specific vs. general solutions



Linköping University

expanding reality

www.liu.se