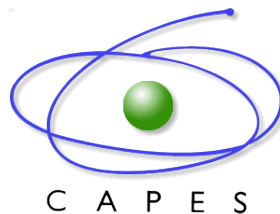


Data Integration and Data Provenance

Profa. Dra. Cristina Dutra de Aguiar Ciferri
cdac@icmc.usp.br



Outline

- Data Integration
 - Schema Integration
 - Instance Integration
 - Our Work
- Data Provenance
 - Basic Concepts
 - The PrInt Model

Outline

- Index Structures
 - Biological databases
 - Similarity search of complex data
 - Spatial data warehouses
 - Similarity search over data warehouses of images
- Mining of Medical Data
- **Data Integration and Data Provenance**

Data Integration

- Schema Level Integration
 - specification of mappings that describe the semantic relationships among schemas from heterogeneous data sources
- Instance Level Integration
 - identification of which entities from heterogeneous sources refer to the same entity in the real-world
 - resolution of value conflicts

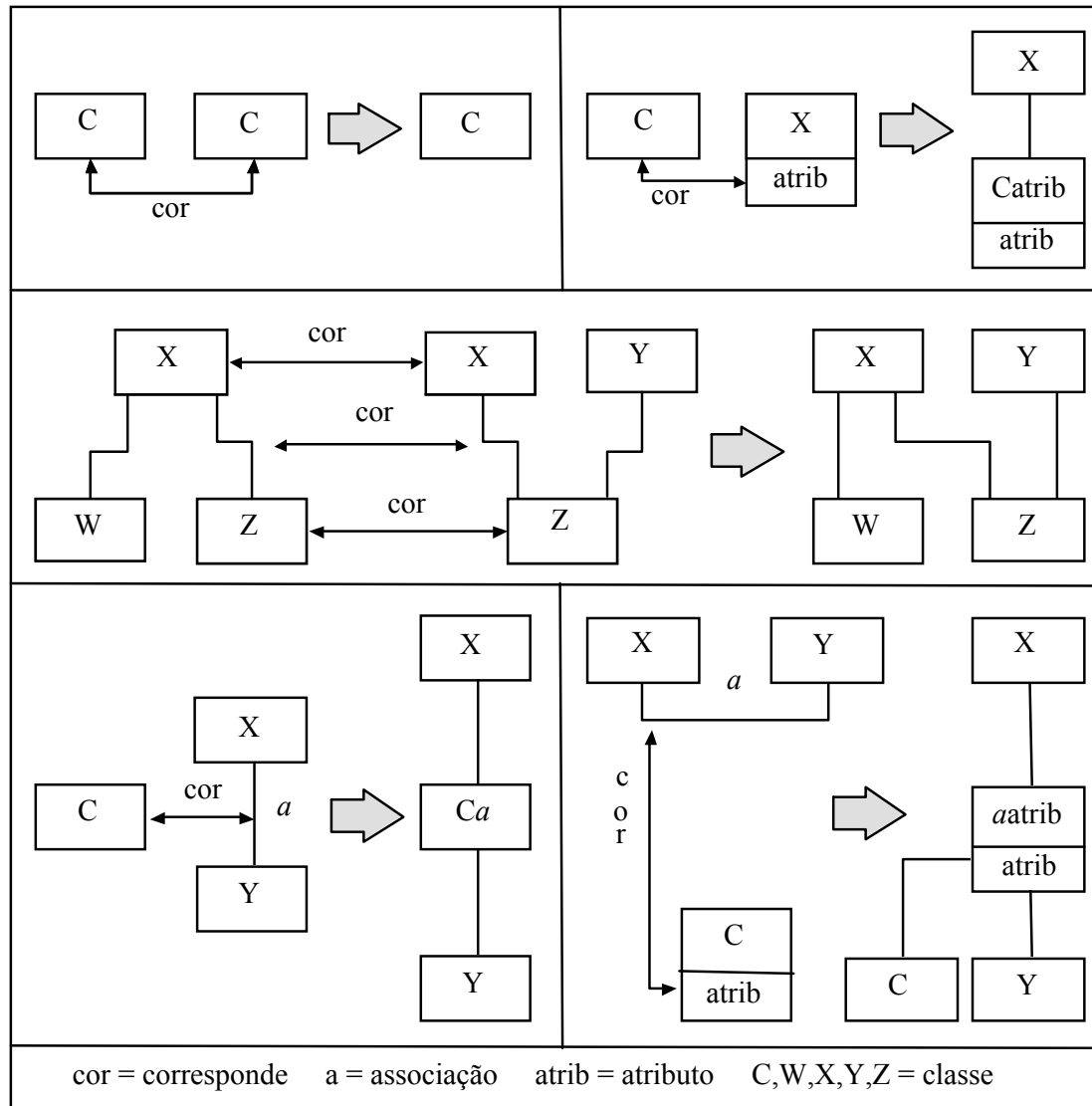
Schema Level Integration

- Semantic Relativism
 - conflicts between two or more representations is related to the fact that different users model the same piece of the real-world in different ways according to their perceptions
 - Types of Conflict Identification
 - name, including homonymous and synonymous
 - semantic
 - structural

Types of Schema

- Global Schema (or Mediated Schema)
 - integration of several heterogeneous local schemas into a homogeneous global schema
 - development of mappings that describe the semantic relationships between the mediated schema and the schemas of the sources
- Federated Schema
 - there is not a homogeneous global schema
 - there are several heterogeneous local schemas, each one related to a given source

Example of Structural Mappings



SPACCAPIETRA, S., PARENT, C. View Integration: A Step Forward in Solving Structural Conflicts. *IEEE Transactions on Knowledge and Data Engineering*, v.6, n.2, p.258-274, 1994

Instance Level Integration

- Reference Reconciliation (or Entity Resolution)
 - automatically detect references to the same entity of the real-world and group them in a cluster of similar entities
- Value Conflict Resolution
 - solve the differences among values of attributes of the entities that refer to the same entity of the real-world

Reference Reconciliation

Examples of entities from the class article (a)

$a_1 = (\{\text{"Distributed query processing in a relational database system"}\}, \{\text{"169-180"}\}, \{p_1; p_2; p_3\}; \{c_1\})$

$a_2 = (\{\text{"Distributed query processing in a relational database system"}\}, \{\text{"169-180"}\}, \{p_4; p_5; p_6\}; \{c_2\})$

Examples of entities from the class person (p)

$p_1 = (\{\text{"Robert S. Epstein"}\}, \text{null}, \{p_2, p_3\}, \text{null})$

$p_2 = (\{\text{"Michael Stonebraker"}\}, \text{null}, \{p_1, p_3\}, \text{null})$

$p_3 = (\{\text{"Eugene Wong"}\}, \text{null}, \{p_1, p_2\}, \text{null})$

$p_4 = (\{\text{"Epstein, R.S."}\}, \text{null}, \{p_5, p_6\}, \text{null})$

$p_5 = (\{\text{"Stonebraker, M."}\}, \text{null}, \{p_4, p_6\}, \text{null})$

$p_6 = (\{\text{"Wong, E."}\}, \text{null}, \{p_4, p_5\}, \text{null})$

$p_7 = (\{\text{"Eugene Wong"}\}, \{\text{"eugene@berkeley.edu"}\}, \text{null}, \{p_8\})$

$p_8 = (\text{null}, \{\text{"stonebraker@csail.mit.edu"}\}, \text{null}, \{p_7\})$

$p_9 = (\{\text{"mike"}\}, \{\text{"stonebraker@csail.mit.edu"}\}, \text{null}, \text{null})$

Examples of entities from the class conference (c)

$c_1 = (\{\text{"ACM Conference on Management of Data"}\}, \{\text{"1978"}\}, \{\text{"Austin, Texas"}\})$

$c_2 = (\{\text{"ACM SIGMOD"}\}, \{\text{"1978"}\}, \text{null})$

article: title, pages, *authors, *conference
person: name, email, *authors, *emailContact
conference: name, year, local

Reference Reconciliation

Grouping from entities of the class article (a)

grouping₁ = {a₁, a₂}

Grouping from entities of the class person (p)

grouping₂ = {p₁, p₄}

grouping₃ = {p₂, p₅, p₈, p₉}

grouping₄ = {p₃, p₆, p₇}

Grouping from entities of the class conference (c)

grouping₅ = {c₁, c₂}

DONG, X.; HALEVY, A.; MADHAVAN, J. Reference Reconciliation in Complex Information Spaces.
In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, p.85-96, 2005

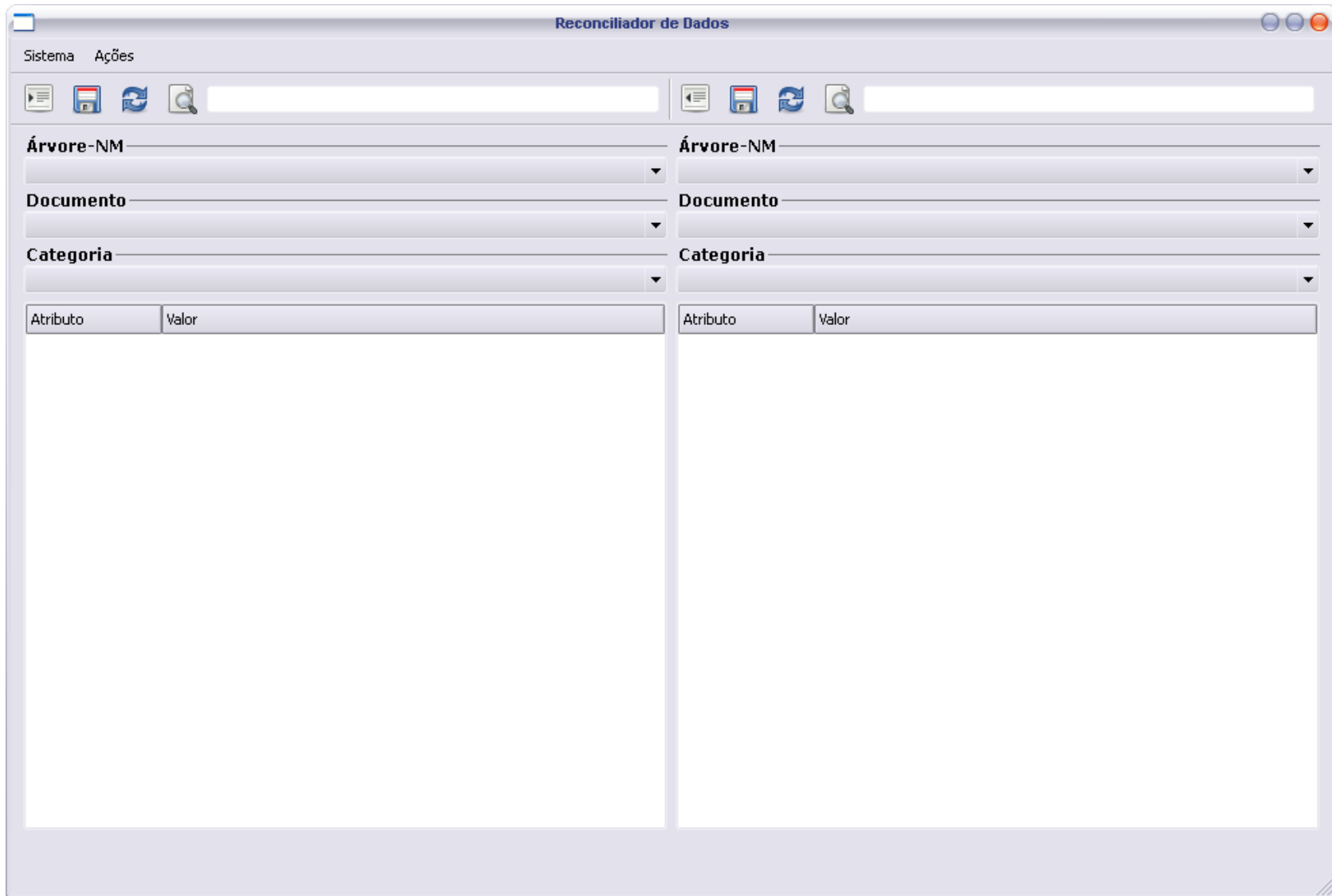
Our Work

- The Academic Data Reconciler Tool
 - semi-automates the identification of
 - correspondent objects
 - inconsistencies
 - helps user to
 - eliminate inconsistencies
 - complete data
 - exchange data
- The Academic Data Reconciler Tool for Reference Reconciliation

Academic Data Reconciler Tool (ADR)

- Functional aspects
 - visualization of objects from two documents
 - side-by-side visualization
 - synchronization of objects
 - edition of incomplete and erroneous data
 - data exchange between objects from different documents

Interface



Visualization

Reconciliador de Dados

Sistema Ações

Árvore-NM: Currículo Lattes | Documento: E:\Documentos\Currículo_Lattes_Caetano.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5
Arquivo XML aberto.

Árvore-NM: DBLP | Documento: E:\Documentos\DBLP.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	ADBIS Research Communications

Número de Objetos: 3

Synchronization

Reconciliador de Dados

Sistema Ações

Árvore-NM: Currículo Lattes | Documento: E:\Documentos\Currículo_Lattes_Caetano.xml | Categoria: Conference Papers

Árvore-NM: DBLP | Documento: E:\Documentos\DBLP.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5

Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	ADBIS Research Communications
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3

Edition

Reconciliador de Dados

Sistema Ações

Árvore-NM
Currículo Lattes

Documento
E:\Documentos\Currículo_Lattes_Caetano.xml

Categoria
Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Pr. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5

Árvore-NM
DBLP

Documento
E:\Documentos\DBLP.xml

Categoria
Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3

Data Exchange

The screenshot shows the 'Reconciliador de Dados' application window. The interface is split into two panes, each displaying a tree view of data objects. The left pane is titled 'Currículo Lattes' and the right pane is titled 'DBLP'. Both panes show a list of papers with their attributes (Title, Year, Proceedings) and values. The papers are highlighted in green, and one paper in each pane is highlighted in red, indicating a match or discrepancy. The status bar at the bottom of each pane shows the number of objects: 'Número de Objetos: 5' on the left and 'Número de Objetos: 3' on the right. The status bar also indicates 'Sincronizado.'

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 5
Sincronizado.

Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3

ADR for Reference Reconciliation

The screenshot displays a software interface for reference reconciliation, comparing an 'Integrated Entity' (left) and a 'Source Entity' (right). Both entities represent an article with the following attributes and values:

- Artigo**
 - Título: Topological dynamics of retarded functional differential equations
 - Ano: 2003
 - Pais: [empty]
 - Meio d...: MEIO_DIGITAL
 - Idioma: Inglês
 - Natureza: COMPLETO
 - Revista: Journal of Differential Equations
 - Página ...: 313
 - Página ...: 331
 - ISSN: 00220396
 - Volume: 195
 - Série: 2
 - Fascículo: 2
 - Local p...: San Diego, California - EUA
- Autores**
 - Autor
 - Márcia Cristina Anderson Braz Federson
 - FEDERSON, M
 - 1
- Palavra...**
 - Pal...: Dinâmica Topológica
 - Pal...: Equações Diferenciais Funcionais
 - Pal...: Retardamento
 - Pal...: Fluxo local
 - Pal...: Equações Diferenciais Ordinárias Generalizadas

The interface includes a menu bar (System, Actions, Help), a status bar (11 / 1365), and navigation controls. A 'Copy to Integrated Object' button is visible at the bottom right.

integrated entity

**similar entities that belong
to the same grouping**