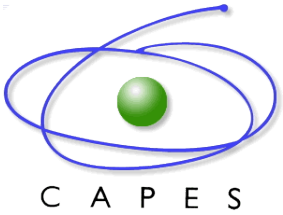
Data Integration and Data Provenance

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

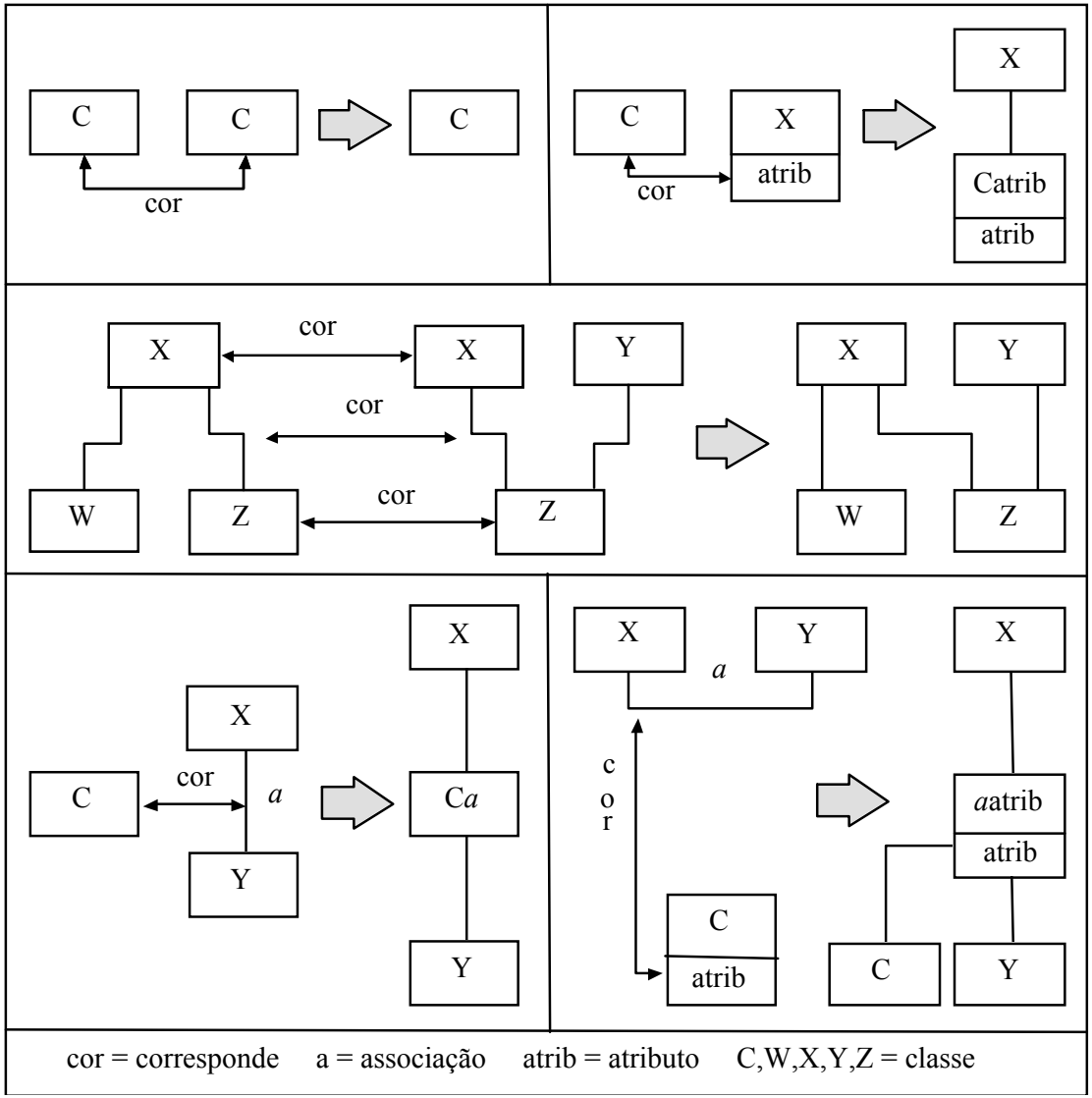
Schema Level Integration

- 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 = ({“Distributed query processing in a relational database system”}, {“169-180”}, {p_1; p_2; p_3}; {c_1})

a_2 = ({“Distributed query processing in a relational database system”}, {“169-180”},{p_4; p_5; p_6}; {c_2})

Examples of entities from the class person (p)

p_1 = ({“Robert S. Epstein”}, null, {p_2, p_3}, null)

p_2 = ({“Michael Stonebraker”}, null, {p_1, p_3}, null)

p_3 = ({“Eugene Wong”}, null, {p_1, p_2}, null)

p_4 = ({“Epstein, R.S.”}, null, {p_5, p_6}, null)

p_5 = ({“Stonebraker, M.”}, null, {p_4, p_6}, null)

p_6 = ({“Wong, E.”}, null, {p_4, p_5}, null)

p_7 = ({“Eugene Wong”}, {“eugene@berkeley.edu"}, null, {p_8})

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

p_9 = ({“mike”}, {“stonebraker@csail.mit.edu"}, null, null)

Examples of entities from the class conference (c)

c_1 = ({“ACM Conference on Management of Data”}, {“1978”}, {“Austin, Texas”})

c_2 = ({“ACM SIGMOD”}, {“1978”}, 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)

$grouping_1 = \{a_1, a_2\}$

Grouping from entities of the class person (p)

$grouping_2 = \{p_1, p_4\}$

$grouping_3 = \{p_2, p_5, p_8, p_9\}$

$grouping_4 = \{p_3, p_6, p_7\}$

Grouping from entities of the class conference (c)

$grouping_5 = \{c_1, c_2\}$

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

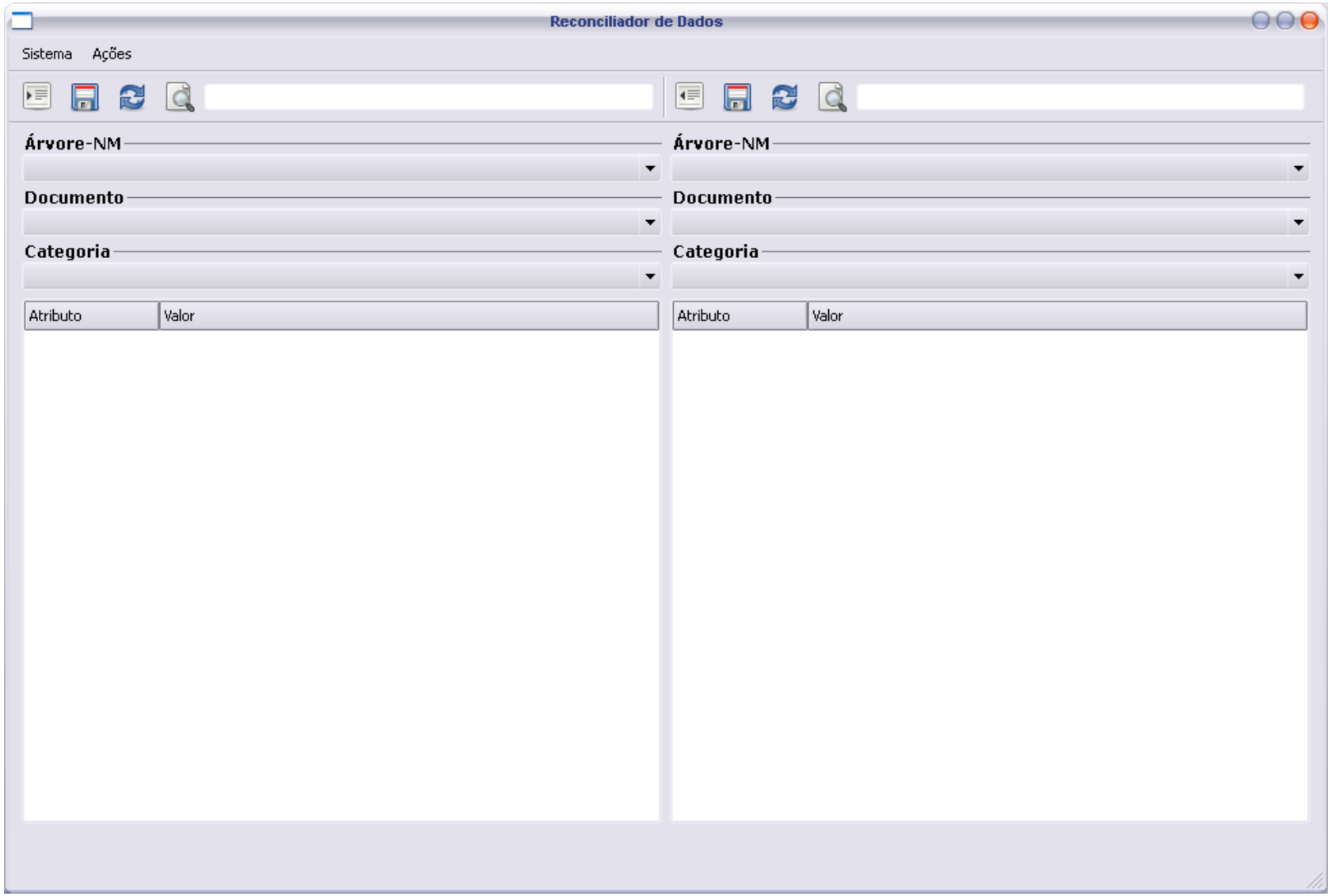
Our Work

- 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

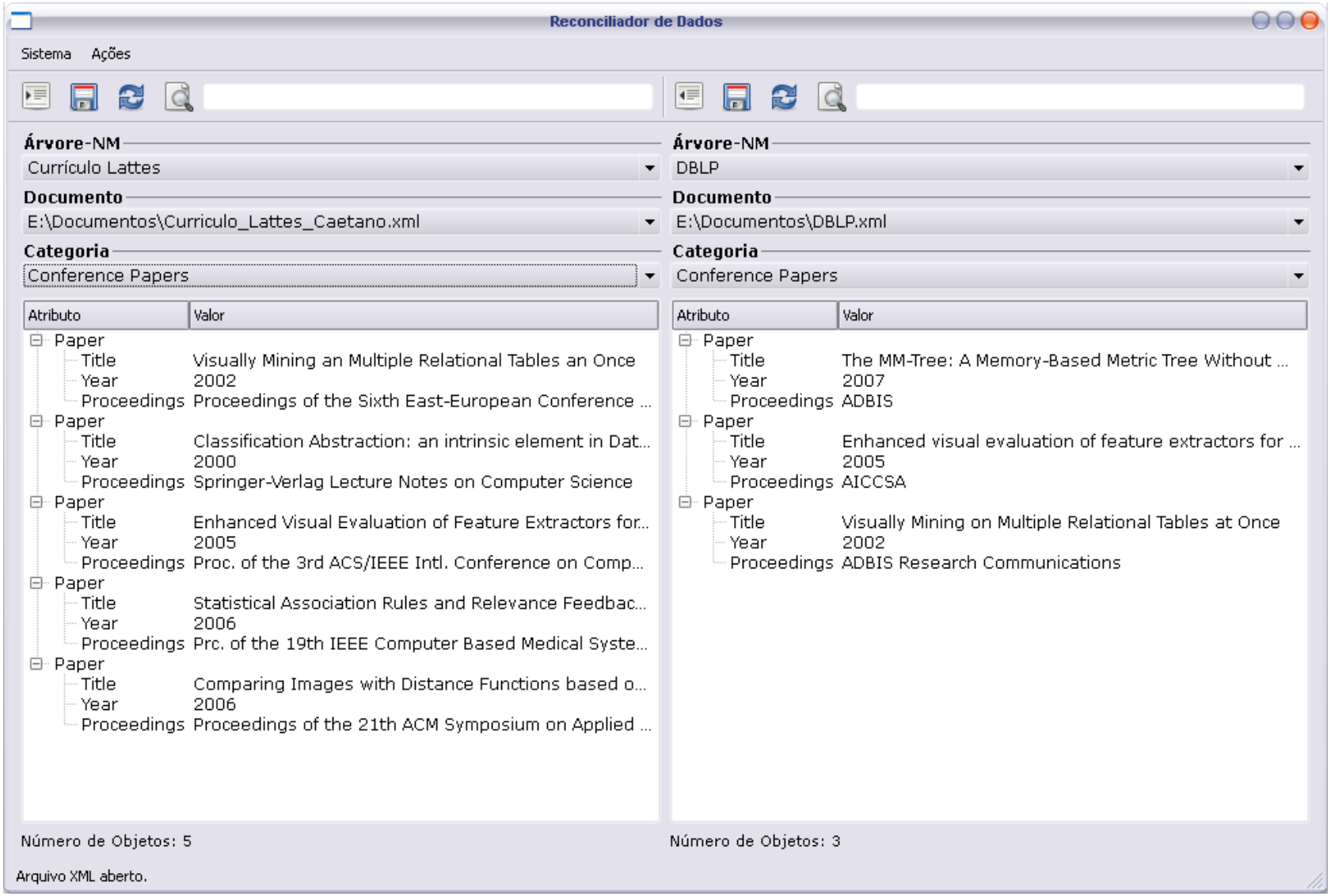
Academic Data Reconciler Tool (ADR)

- 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

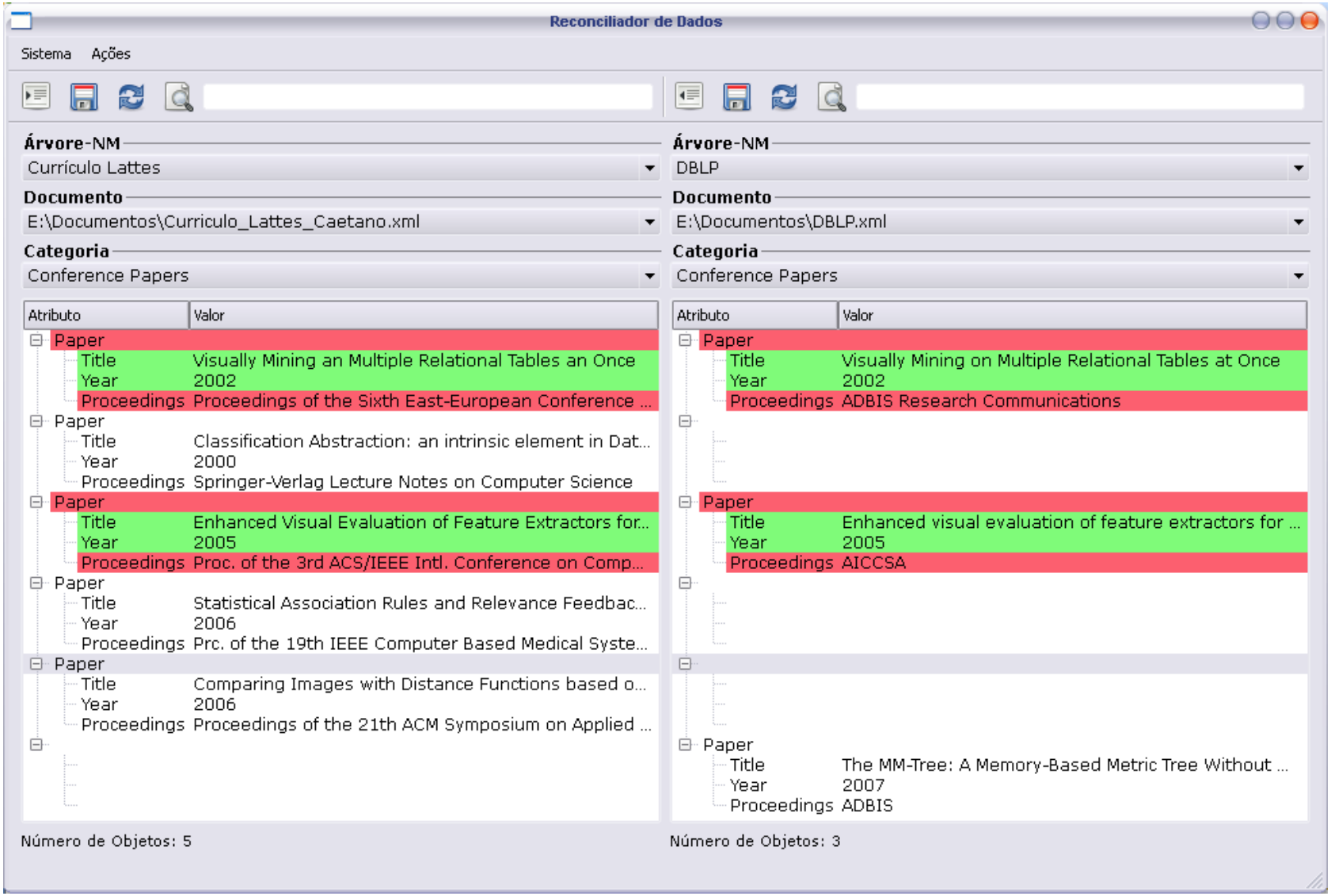
Interface


Reconciliador de Dados
Sistema
Ações
Árvore-NM
Documento
Categoria
Atributo
Valor
Árvore-NM
Documento
Categoria
Atributo
Valor

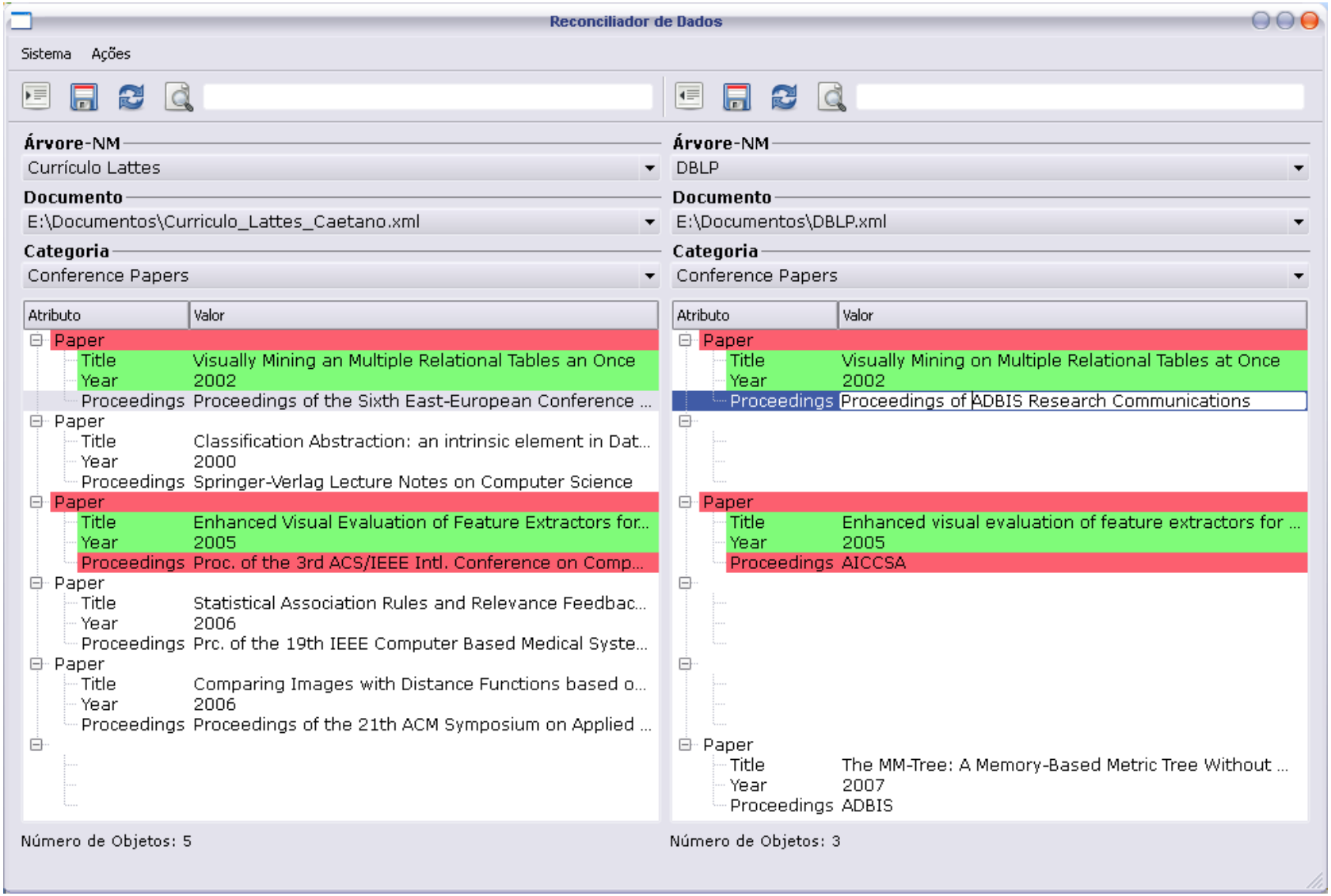
Visualization



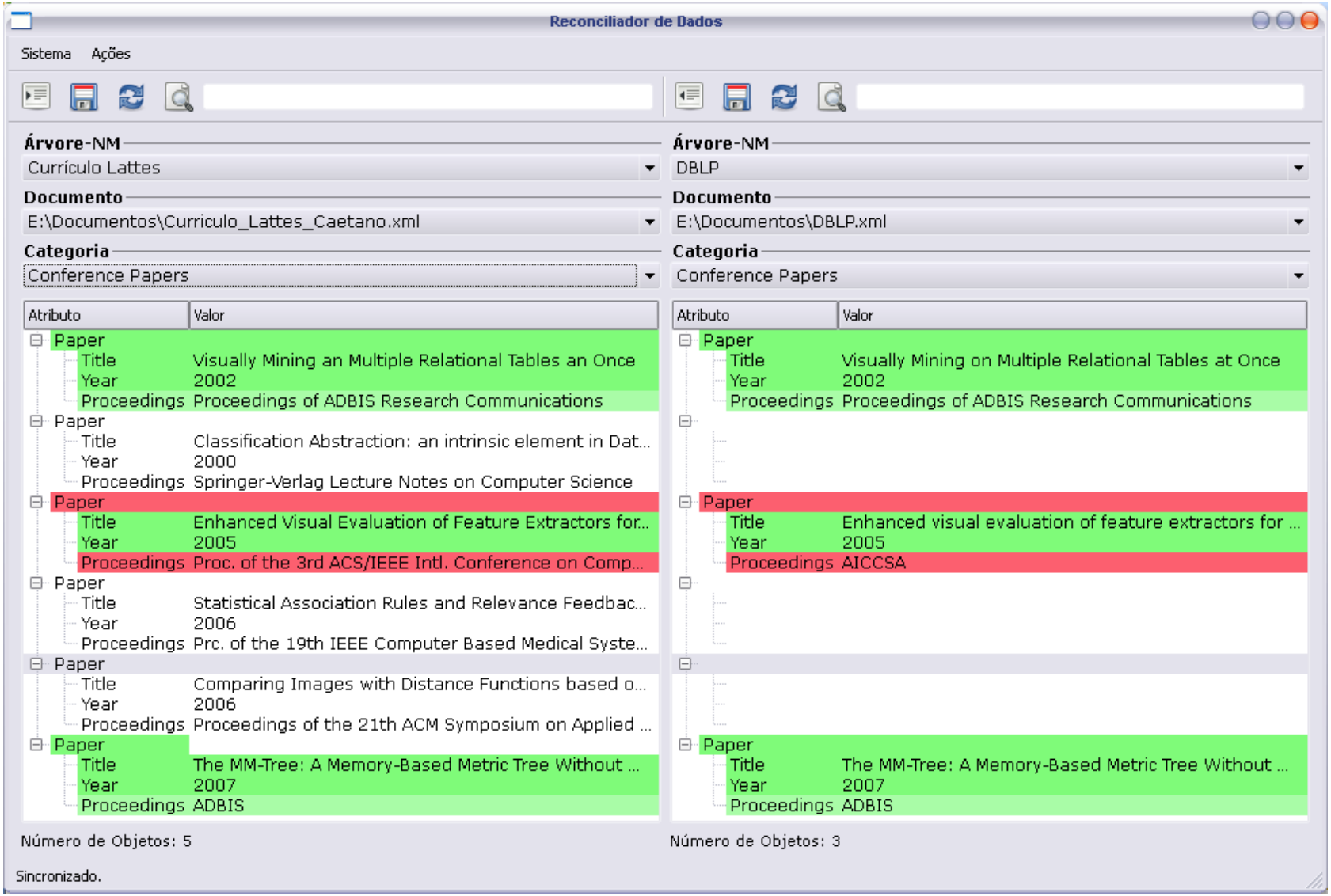
Synchronization



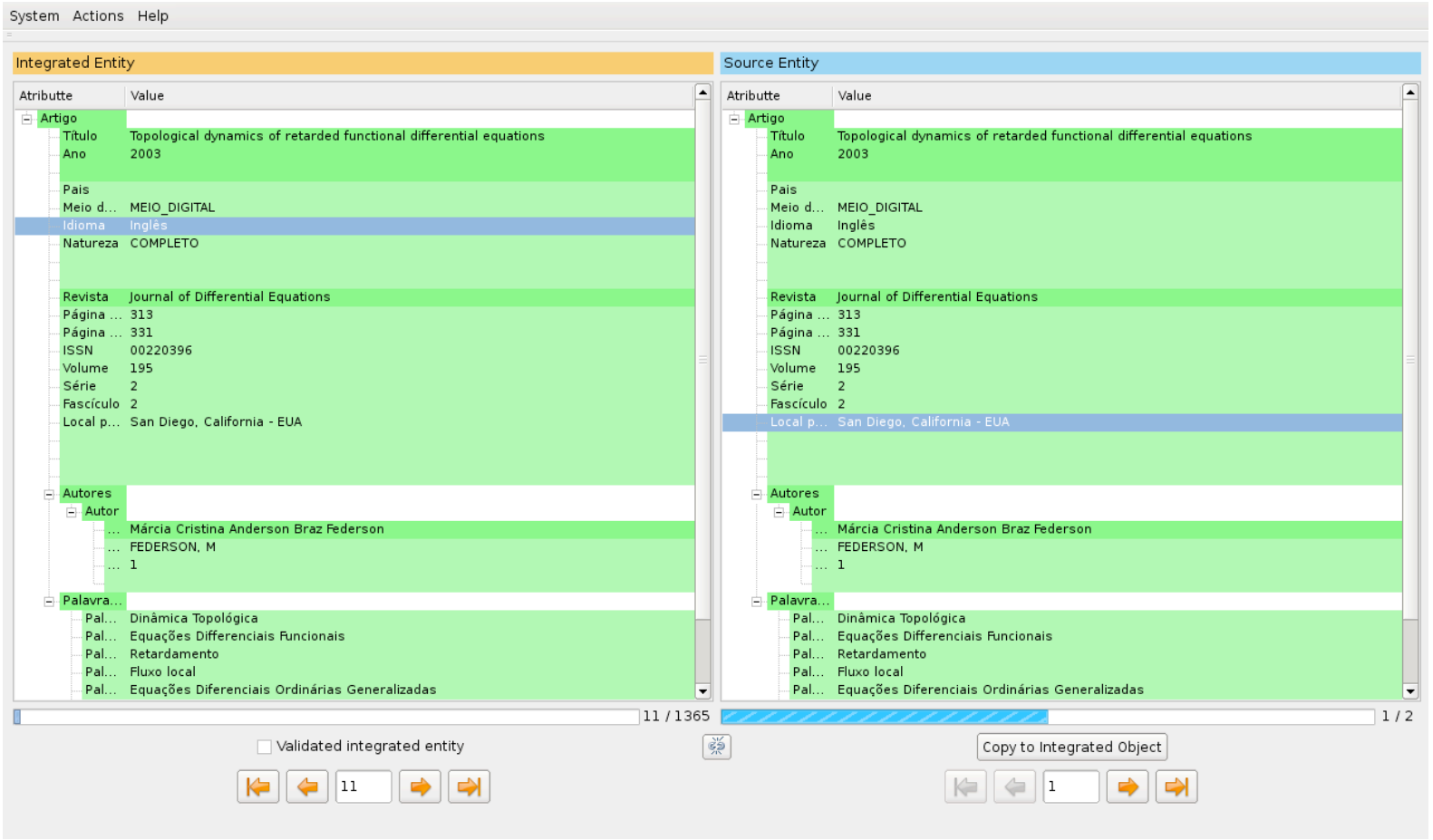
Edition


Reconciliador de Dados
Sistema
Ações
Árvore-NM
Currículo Lattes
Documento
E:\Documentos\Curriculo_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
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



ADR for Reference Reconciliation



integrated entity

similar entities that belong to the same grouping