

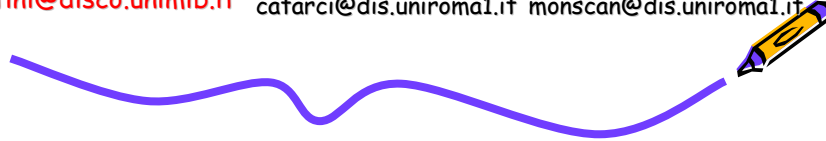


## A Survey of Data Quality Issues in Cooperative Information Systems

Carlo Batini  
Università di Milano  
"Bicocca"  
batini@disco.unimib.it

Tiziana Catarci  
Università di Roma  
"La Sapienza"  
catarci@dis.uniroma1.it

Monica Scannapieco  
Università di Roma  
"La Sapienza"  
monscan@dis.uniroma1.it



### Changes from your slides

- Enlarged some text
- Introduced some examples
- Introduced 3 more techniques (and references)
- Corrected a few errors or missing parts
- Introduced hyperlinks

## General Outline (300 slides)

- 1. Introduction (16 slides)
- 2. Dimensions (34)
- 3. Methodologies (24)
- 3'. Models (52)
- 4. Techniques (135)
  - 4.1 Activities in DQ (9)
  - 4.2 T. for record matching/object identification (88)
    - Basics (24)
    - Short profiles of T. (30)
    - Detailed description of T. (24)
    - Comparison of T. (10)
  - 4.3 T. for data integration (19)
  - 4.4.T. for profiling and editing (19)
- 5. Tools (12)
- 6. Frameworks for CISs (17)
- 7. Open problems (12)
- 8. References (16)

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

3

- Depth first
  - 1. Introduction (16)
  - 2. Dimensions (34)
  - 3. Methodologies (24)
- Breadth first
  - 3'. Models (52)
    - Short profiles of models
    - Detailed profiles (about 20)
  - 4. Techniques (135)
    - 4.1 Activities in DQ (9)
    - 4.2 T. for record matching/object identification (88)
      - Basics (24)
      - Short profiles of T. (30)
      - Detailed description of T. (24)
      - Comparison of T. (10)
    - 4.3 T. for data integration (19)
    - 4.4.T. for profiling and editing (19)
  - 5. Tools (12)
  - 6. Frameworks for CISs (17)
  - 7. Open problems (12)

Details : 85 slides → 30%

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

4

## 1. Introduction (16)

## Introduction: Table of Contents

- Data Quality & CISs
- Basic Choices
- Tutorial organization

## Data Quality: Multidimensional Concept

- **Accuracy**
  - Jhn vs. John
- **Completeness**
- **Currency**
  - Residence (Permanent) Address: out-dated
  - vs. up-to-dated
- **Consistency**
  - ZIP Code and City consistent

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma

Prefix	StreetName	Number	ZipCode	City
	Salaria			Roma

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma
Via	Gracchi	74	00193	Roma

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma

Attribute  
Completeness

Entity  
Completeness

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

7

## Cooperative Information Systems (CISs)

- *"... Distributed and heterogeneous information systems that cooperate requesting and sharing information, constraints, and goals ..."*

[Mylopoulos et al 1997] J. Mylopoulos, M. Papazoglou (eds.):  
Cooperative Information Systems. *IEEE Expert Intelligent Systems & Their Applications*, vol. 12, no. 5,  
September/October 1997

- CISs include:
  - Data integration systems
  - Cross-organization workflow management systems
  - ...
- Examples of contexts requiring CISs
  - Set of public administrations in an e-Government scenario
  - Set of companies of a virtual enterprise
  - ...

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

8

## Data Quality & CISs

- CISs features:
  - Data sharing to accomplish cooperative tasks
  - High data replication

Instance level  
heterogeneities, not  
trusted data → CISs need data quality

High data replication,  
different available  
sources → Data quality needs CISs

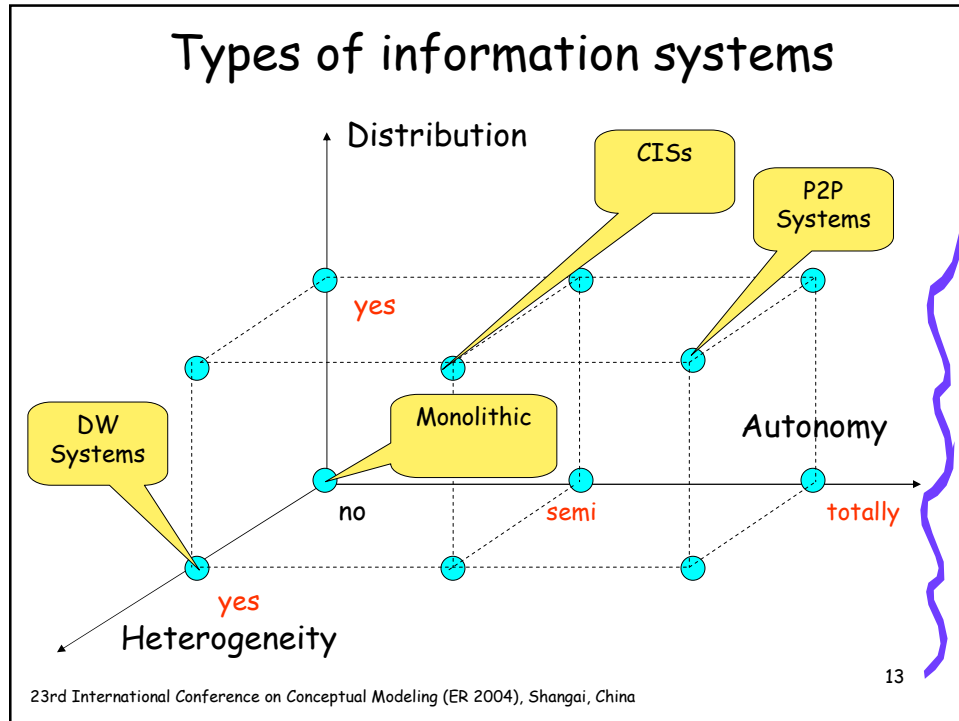
## Basic Choices

## Types of data

- Many authors
  - Structured
  - Semi structured
  - Unstructured
- [Shankaranaian et al 2000]
  - Raw data
  - **Component** data
  - Information products
- [Bouzenghoub et al 2004]
  - Stable,
  - Long term changing
  - Frequently changing
- [Dasu et al 2003]
  - Federated data, that come form different heterogeneous sources
  - Massive high dimensional data
  - Descriptive data
  - Longitudinal data, consisting in time series
  - Web data
- [Batini et al 2003]
  - Elementary data
  - Aggregated data

## Types of data: main focus

- Many authors
  - **Structured**
  - Semi structured
  - Unstructured
- [Shankaranaian et al 2000]
  - raw data
  - component data
  - **Information products**
- [Bouzenghoub et al 2004]
  - stable,
  - long term changing
  - frequently changing
- [Dasu et al 2003]
  - **Federated data, that come from different heterogeneous sources**
  - Massive high dimensional data
  - Descriptive data
  - Longitudinal data, consisting in time series
  - Web data
- [Batini et al 2003]
  - **Elementary data**
  - aggregated data



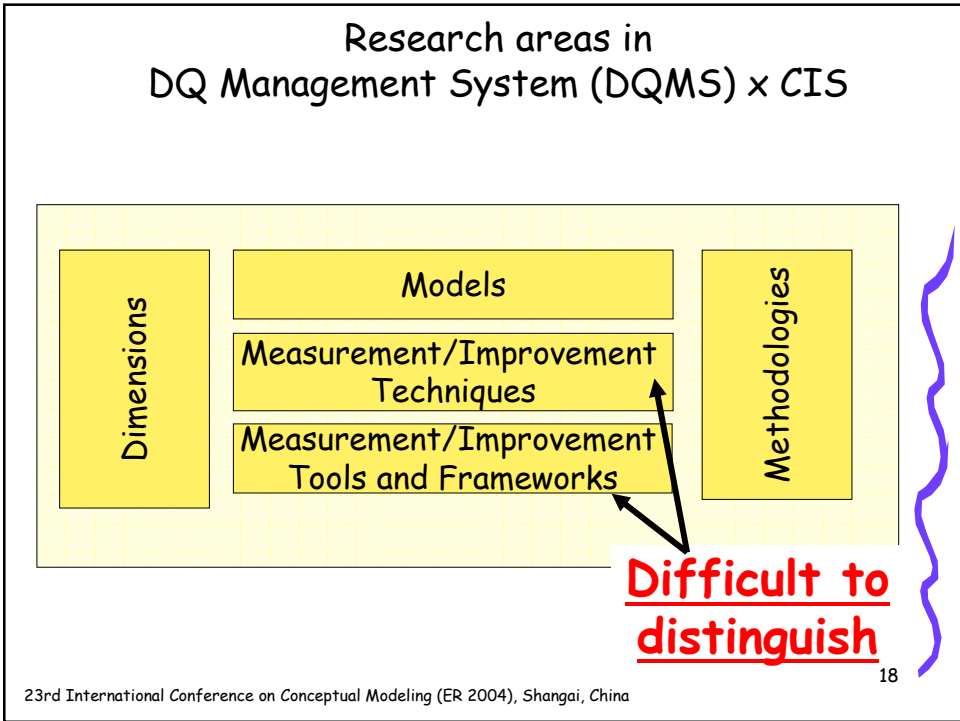
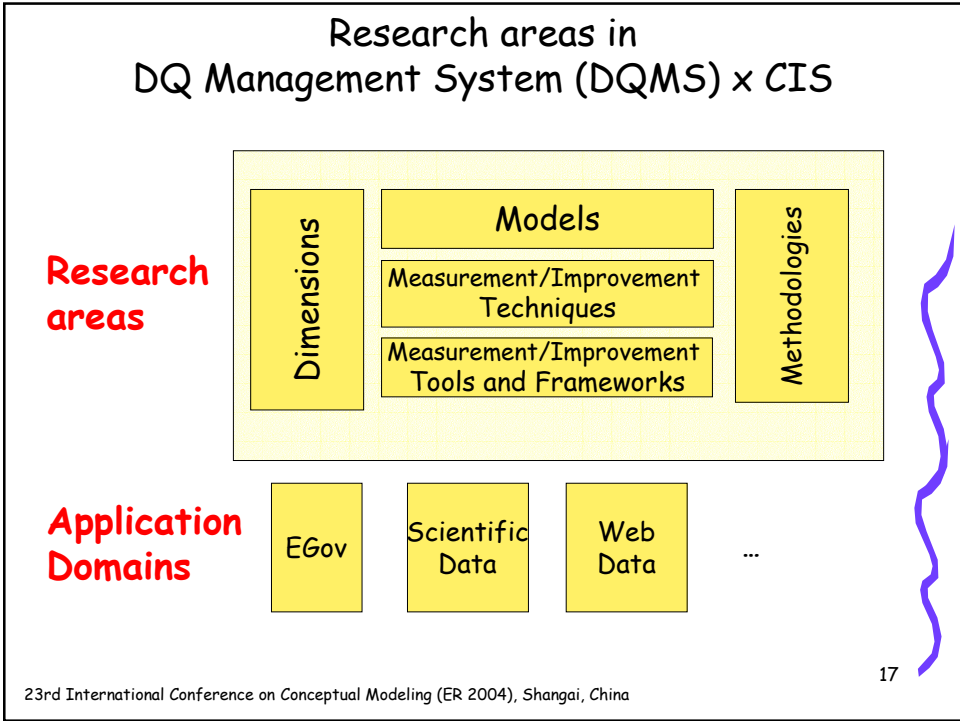
- ## Main focus
- **Type of Data**
    - Structured data
    - Elementary data
    - Information products
    - Federated data
  - **Type of Information Systems**
    - CISs
  - **Quality Dimensions**
    - Complete overview → focus of data values dimensions.
- 23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China
- 14

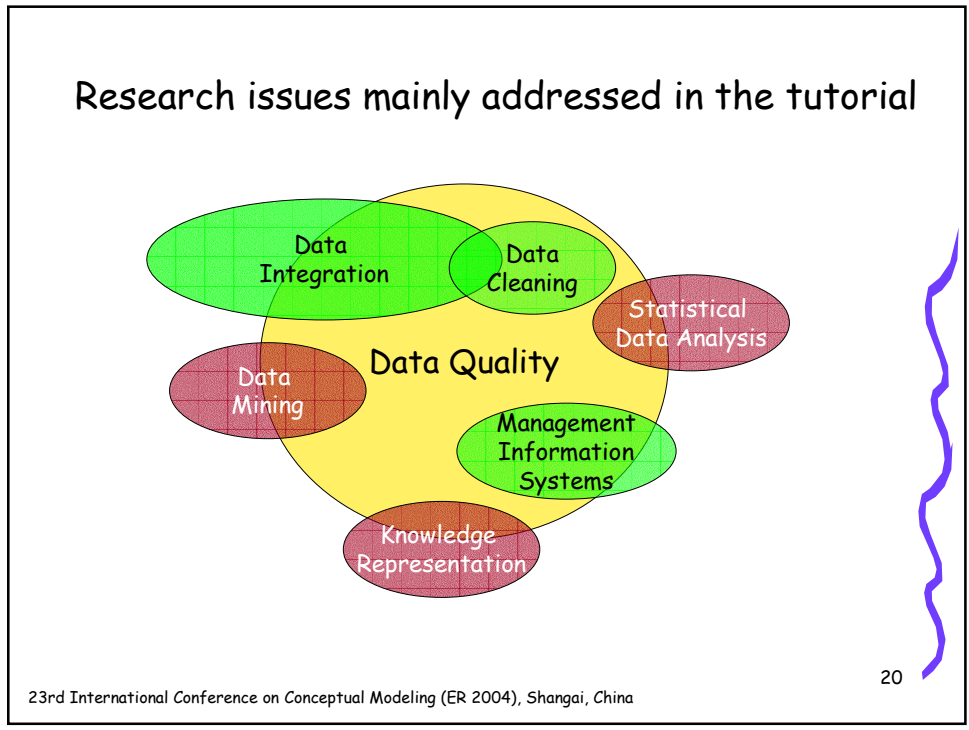
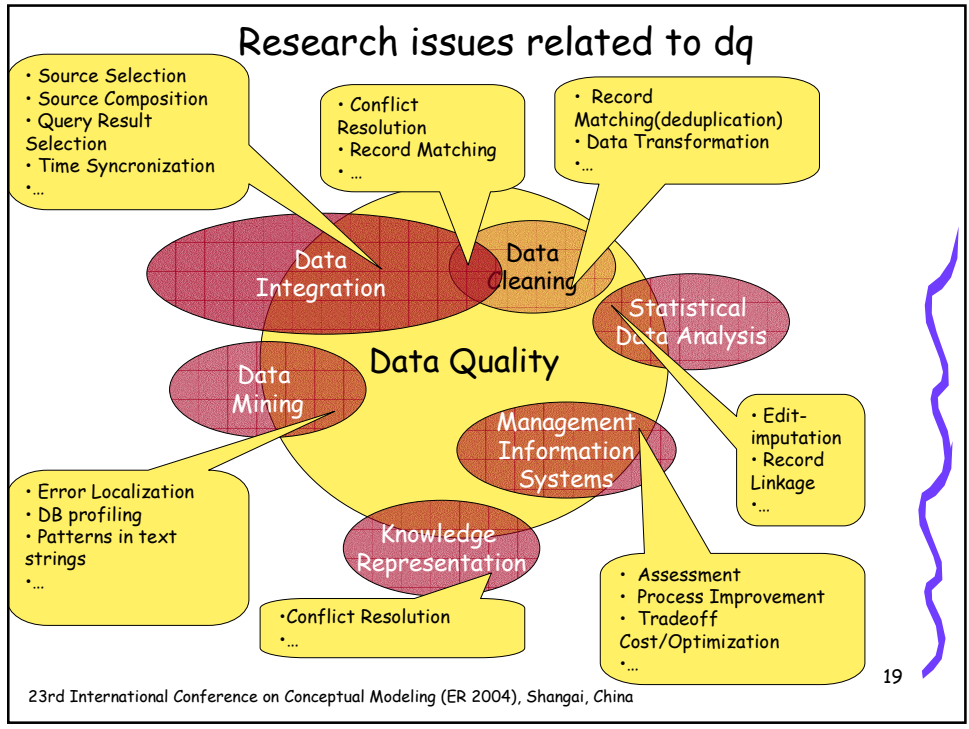
# Tutorial Organization

## Towards a concept of Data Quality Management System (DQMS) x CIS

- DQ literature in CIS can be seen as a great effort to produce a comprehensive set of techniques, services, tools for measuring and improving data quality in one or more cooperating organizations.
- We call such a set of features a

**Data quality management system**





## 2. Dimensions (34)

### Dimensions: table of contents

- Example of DQ dimension classification
- Other classifications
- Theoretical approaches
- Comparison of dimension definitions in different proposals
- The "most important" dimensions
- DQ dimensions vs ISO 9126 software quality dimensions
- Evolution of dimensions and evolution of ICT technologies

Example of DQ Dimension classification:  
[Redman 1996] original definitions

- Conceptual schema **deal with intension**
- Data values **deal with extension**
- Data representation (Format) **deal with data format**

Conceptual Schema - 1

- **Content**
  - **Relevance** - The schema should provide data needed by the application
  - **Obtainability** - Data values should be easily obtainable
  - **Clarity of definition** - Each term in the definition of the schema should be clearly defined

## Conceptual Schema - 2

- **Scope**
  - **Comprehensiveness** - Each needed data item should be included
  - **Essentialness** - No unneeded data item should be included
- **Level of detail**
  - **Attribute granularity** - The attributes should be defined at the right level of detail to support applications
  - **Domain precision** - The domains of possible values should be just large enough to support applications

## Conceptual Schema - 3

- **Composition**
  - **Naturalness** - Each item in the schema should be just large enough to support applications
  - **Occurrence identifiability** - The schema should make identification of individual entities easy
  - **Homogeneity** - entity types should be defined to minimize the occurrence of unnecessary attributes
  - **Minimum redundancy** - Redundancy should be kept to a minimum

## Conceptual Schema - 4

- **Schema consistency**
  - **Semantic consistency** - The schema should be clear and unambiguous and consistent
  - **Structural consistency** - Entity types and attributes should have the same basic structure whenever possible
- **Reaction to change**
  - **Robustness** - The schema should be wide enough so that it does not require change every time applications change
  - **Flexibility** - When necessary the schema should be easily changed

## Data values - 1

- **Accuracy** - accuracy of a datum  $\langle e, a, v \rangle$  refers to the nearness of the value  $v$  to some  $v'$  in the attribute domain, which is considered as the correct one for the entity  $e$  and the attribute  $a$ .
- **Correctness** - if in the accuracy definition  $v$  coincides with  $v'$ , the datum is said to be correct
- **Completeness** - refers to the degree to which values are present in a data collection

## Data values - 2

- **Currency** - refers to the degree to which a datum is up-to-date. A datum is up-to-date if it is correct in spite of a possible discrepancy related to time-related changes due to the correct value, outdated at time  $t$  if it is incorrect at  $t$  but was correct at some time preceding  $t$ .
- **Consistency** - Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules.

## Data Representation -1

- **Appropriateness** - one format is more appropriate than another if it is more suited to user needs.
- **Interpretability** - ability of the user to interpret correctly values from their format
- **Portability** - the format can be applied to as a wide set of situations as possible.
- **Format precision** - ability to distinguish among elements in the domain that must be distinguished by users

## Data Representation -2

- **Format flexibility** - changes in user needs and recording medium can be easily accommodated.
- **Ability to represent null values** - ability to distinguish without ambiguities null as default values from applicable values of the domain.
- **Efficient use of memory** - Efficiency in the physical representation. An icon is less efficient than a code.
- **Representation consistency** - Coherence of physical instances of data with their formats.

## Drawbacks of Redman's Classification

- **Unbalanced:** conceptual more detailed than values and format
- **Heterogeneous:** conceptual and values "different" wrt format
- **Not formal:** several ambiguous definitions
- **Partial:** deals mainly with "intrinsic" dimensions

## Other Classifications - 1

- [Liu et al 2002]
  - Collection quality
  - Organization quality
  - Presentation Quality
  - Application Quality
  - + Theory specific qualities
- [Naumann 2002]
  - Content related, concern the actual information to be retrieved
  - Technical, measure aspects determined by software and hw
  - Intellectual, made of subjective criteria like believability
  - Instantiation related, concern the presentation of info

## Other Classifications - 2

- [Pipino et al 2000]
  - Product/Service
  - Conform to Specification/Consumer Expectation
  - Cross
    - Sound
    - Dependable
    - Useful
    - Usable
- [Wang et al 2001]
  - Intrinsic (quality)
  - Accessibility,
  - Contextual quality,
  - Representational

## Other Classifications 2

- [Bovee 2001]
  - Accessibility
  - Interpretability
  - Relevance
  - Credibility
- [Jarke 1999]
  - Design and Administration Quality
  - Software Implementation Quality
  - Data Usage Quality
  - Data Quality

## Main comments on classifications

Author	Main features
Redman 1996	Good for non technical audience
Naumann 2002	Suitable for Web integrated ISs (data integration systems on the Web)
Pipino 2000	Effective for assessment, where both service and product perspectives need to be evaluated
Jarke 1999	Suitable for data warehouse systems
Bovee 2001	Good both for technical and non-technical audience
Wang 2001	The most general and well founded/validated
Liu 2002	Effective in binding dimensions to the phases of data life cycle. Inherits theory based qualities from other research areas

## Theoretical approaches to dimensions

- Two attempts to provide a theoretical foundation to DQ.
  - [Wand et al 1996] Wand Y., Wang R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. *Communication of the ACM*, vol. 39, no. 11, 1996.
  - [Liu et al 2002] L. Liu, L. Chi Evolutionary Data quality, ICIQ 2002.
- Citation from [Wand et al 1996]:
- **The world is made of things that possess properties, etcetera ...**

## Essentials from [[Wand](#) et al 1996] - 1

- Data should be in an exhaustive **mapping** with the real world. I.e., a real-world state can be mapped into more than one state in an information system but a state in an information system cannot represent two or more states in the real world.
- Data is deemed **incomplete** if there is no state in the information system which corresponds to a real-world state.

## Essentials from [Wand et al 1996] - 2

- Data is deemed **ambiguous** if a state in the information system corresponds to two or more real-world states
- Data is **meaningless** if a state in the information system does not correspond to any real-world states.

## Essentials from [Wand et al 1996] - 3 (new)

- Data is **incorrect** if a real world state is mapped to a wrong state in the information system. Two cases are distinguished:
  1. If there exist meaningless states of the information system, the mapping might be to a meaningless state, **(Carlo → Crlo)** and
  2. the mapping might be to a meaningful, but incorrect information system state. **(Carlo → Ciro)**
- In the first case the user will not be able to map back to a real-world state. In the second case the user will be able to infer back, but to an incorrect state

Essentials from [Wand et al 1996] - 3

1. DQ dimensions are classified in terms of intrinsic classes of dimensions as complete, unambiguous, meaningful, correct (Table 1), according to the nature and source of deficiency.

Table 1: Dimensions and deficiencies

DQ dimension	Nature of associated deficiency	Source of deficiency
Complete	Improper representation: missing IS states	Design failure
Unambiguous	Improper representation: multiple RW states mapped to the same IS state	Design failure
Meaningful	Meaningless IS state	Design failure and operation failure
Correct	Garbling (map to a wrong state)	Operation failure

## Essentials from [Wand et al 1996] - 4

- 2. DQ dimensions mentioned in literature can be classified in terms of internal/external view, data vs system related (Table 2).

Table 2: Dimensions and views

	Data Related	System related
Internal view (design, operation)	Accuracy, Reliability, Timeliness, etc.	Reliability
External view (use, value)	Timeliness, Relevance, Content, etc.	Timeliness, Reliability, Format

## Essentials from [Wand et al 1996] - 5

3. internal dimensions are analyzed from the perspective of the ontological model.

E.g. (syntactic) accuracy refer to cases where it is possible to infer a valid state of the RW but not the correct one.

(Crlo → Carlo)

4. furthermore, the ontological model is used:

- a. to identify the mapping problems and the observed data problems (Table 3), and
- b. to fix generic data deficiency repairs. (This last Table not included).

Table 3: Generic DQ problems

DQ dimension	Mapping problem	Observed data problem
Complete	Certain RW states cannot be represented	Loss of information about the application domain
Unambiguous	A certain IS state can be mapped back into several RW states	Insufficient information: the data can be interpreted in more than one way
Meaningful	It is not possible to map the IS state back to a meaningful RW state	Design failure and operation failure
Correct	The IS state may be mapped back into a meaningful state, but the wrong one	The data derived from the IS do not conform to those used to create these data.

## Essentials from [Liu et al 2002] - 1

- Distinguishes between three approaches to DQ:
  - The **intuitive approach** identifies DQ attributes based on **an expert's personal experience** and intuitive understanding about what attributes are "important"
  - the **empirical approach** lets **data consumers** determine the characteristics to be used to assess whether data fits their tasks
  - the **theoretical approach** emphasizes deriving attributes based on an established theory

## Essentials from [Liu et al 2002] - 2

- Criticism to [Wand et al 1996]: Although the theoretical approach is superior to intuitive and empirical counterparts, existing theoretical approaches are limited in their ability to derive a full-fledged DQ measurement model.
- For example, attributes: complete, unambiguous, meaningful, and correct are only a small sample of the attributes in assessing intrinsic DQ.
- The ontological mapping approach, consequently, leaves many other important attributes unspecified.

### Essentials from [Liu et al 2002] - 3

- The theoretical approach of [Liu et al 2002] is based on the concept of **evolutional data quality**, where the data life cycle is seen as composed of four phases:
  - **Collection**
  - **Organization**
  - **Presentation**
  - **Application**
- Qualities that in other approaches are generically attached to data, here are associated to specific phases. E.g. **accuracy** to **collection**, **consistency** to **organization**.

### Essentials from [Liu et al 2002] - 4

- A "**theory**" is a general designation for any technique, method, approach, or model that is employed during the data life cycle.
- E.g. when data in the Organization phase is stored, a model is chosen, such as a relational or object-oriented model to guide the data organization.

## Essentials from [Liu et al 2002] - 5

- Due to the attachment of data to theories, when defining DQ, we need to consider how data meet the specifications or serve the purposes of a theory. We call such a concept of quality *theory-specific*.
- E.g. in the relational model, theory specific qualities are normal forms, referential integrity, etc.
- So new "inherited" qualities are associated to phases.

## Comparison of dimensions definitions

## Definitions for accuracy - 1

Wand et al 1996	Inaccuracy implies that information system represents real-world state different from the one that should have been represented
Wang et al 1996	The extent to which data are correct, reliable, and certified free of error
Redman 1996	Given the triple $\langle e, a, v \rangle$ , accuracy refers to the nearness of the value $v$ to some value $v'$ in the attribute domain, which is considered as the correct one for the entity $e$ and the attribute $a$ .
Jarke et al 1999	Describes the accuracy of the data entry process which happened at sources

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

53

## Definitions for accuracy - 2

Bovee et al 2001	Accuracy deals with information being true or error free wrt some known, designed or measured value
Naumann 2002	The quotient of the number of correct values in a source and the overall number of values in the source
Liu et al 2002	The extent to which collected data are free of error measurements

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

54

## Definitions for completeness - 2

Wand et al 1996	The ability of an information system to represent every aspect of the represented real world system Different aspects (breadth, depth, scope)
Wang et al 1996	The extent to which data are of sufficient breadth, depth and scope for the task at hand
Redman 1996	The degree to which a data collection More related to attribute completeness
Jarke et al 1999	Percentage of the real world information entered in the sources and the data warehouse
Bovee et al 2001	Entities with information having all required attributes of an entity's information present Provides also a metric
Naumann 2002	It is the quotient of the number of non-null values in a source and the size of the universal relation
Liu et al 2002	All values that are supposed to be collected as per a collection theory

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

55

## Definitions for time dimensions: timeliness, currency, volatility

Wand et al 1996	Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state
Wang et al 1996	Timeliness is the extent to which age of the data is appropriate for the task at hand
Redman 1996	Currency is the degree to which a datum is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value
Jarke et al 1999	Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world.
Bovee et al 2001	Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long ago it was recorded. Volatility is a measure of information instability-the frequency of change of the value for an entity attribute.
Naumann 2002	Timeliness is the average age of the data in a source Related to the utility of information for a task
Liu et al 2002	Timeliness is the extent to which data are sufficiently up-to-date for a task Different!

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

56

## The most important dimensions - 1

Homonyms

Synonyms

	#citations	Coherence of Meaning (Total, High, Marginal)	Cluster Membership
Accuracy	7	total	
Completeness	7	total	
Interpretability	6	total	1
Timeliness	6	high	2
Consistency	5	high	
Relevancy/Relevance	5	high	
Currency	4	high	2
Security/Access Security	4	total	
Appropriate Amount of Data	4	marginal	

We have seen one homonym

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

57

## The most important dimensions - 2

	#citations	Coherence of Meaning (Total, High, Marginal)	Cluster Membership
Volatility	3	total	2
Representational Consistency	3	total	
Reliability	3	low	3
Accessibility	3	total	
Objectivity	3	total	
Ease of Understanding/un derstandability	2	total	1
Credibility	2	total	3
Believability	2	total	3
Reputation	2	total	3

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

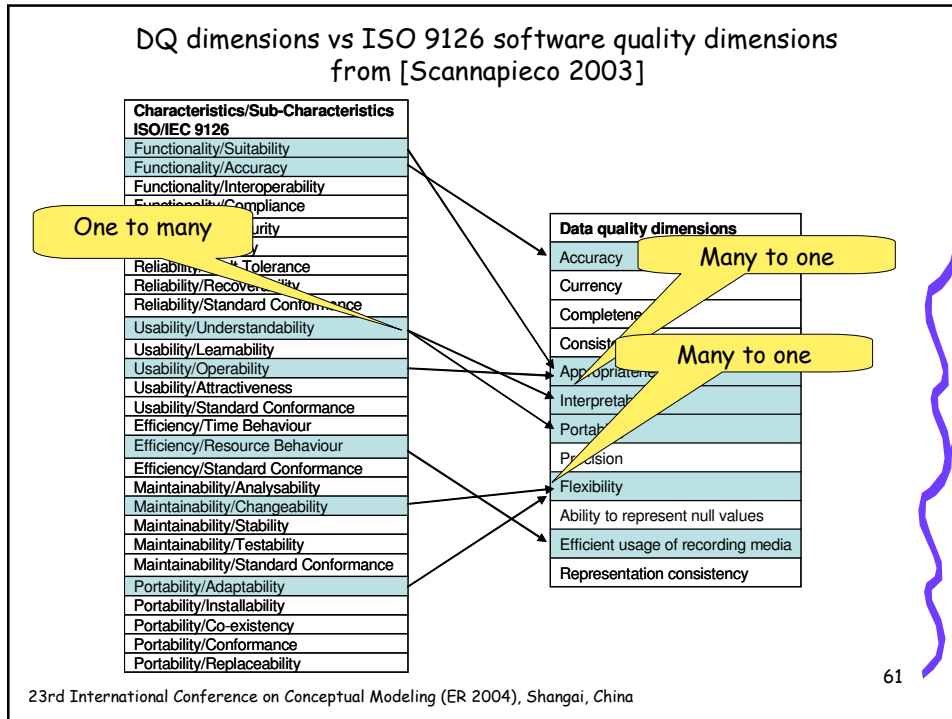
58

### The *most important* dimensions - 3

	#citations	Coherence of Meaning (Total, High, Low)	Cluster Membership
Value-Added	2	total	
Availability	2	total	
Portability	2	low	
Concise representation	2	total	
Responsiveness /Response Time	2	total	

### Clusters of dimensions

Dimension ( <u>major</u> )	Cluster
<u>Interpetability</u>	1
<u>Ease of understanding/ Understandability</u>	1
<u>Timeliness</u>	2
<u>Currency</u>	2
<u>Volatility</u>	2
<u>Reliability</u>	3
<u>Credibility</u>	3
<u>Beleivability</u>	3
<u>Reputation</u>	3



23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

### Evolution of dimensions and evolution of ICT technologies (network/cooperative information systems)

- Traditional dimensions are Accuracy, Completeness, Timeliness, Consistency
- 1. With the advent of networks, sources increase dramatically, and data become often "found data".
- 2. Federated data, where many disparate data are integrated, are highly valued
- 3. Data collection and analysis are frequently disconnected.
- As a consequence we have to revisit the concept of DQ and new dimensions become fundamental.

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

62

## New dimensions - 1

- **Interpretability**, i.e. the documentation and metadata that are available to interpret correctly the meaning and properties of data.
- **Suitability**, in order to determine if the dataset contains relevant and sufficient information to answer the questions we pose.
- **Synchronization among different time series.**
- **Extent of automation**, dynamic operational DQ metric that measures the amount of manual intervention required during the process.
- **Successful completion** of end to end process, dynamic operational metric, that measures the outcome of the process.

## New dimensions - 2

- **Glitches in analysis**, static operational metric which measures the degree to which glitched data causes glitched analysis. **Glitch** means an alteration of data that is not caused by the process.
- **Accessibility**, that measures the ability of the user to access the data with his/her own culture, physical status/functions and technologies available, and can be measured in several ways, e.g. the time between request of access and the actual ability to view the data, and that in general.
- **Conformance to business rules**, dynamic diagnostic metric based on constraints defined on the data set.
- **Conformance to schema**, static diagnostic metric, which measures how well data conform to the metadata in its schema.

## 3. Methodologies (24)

### Methodologies: table of contents

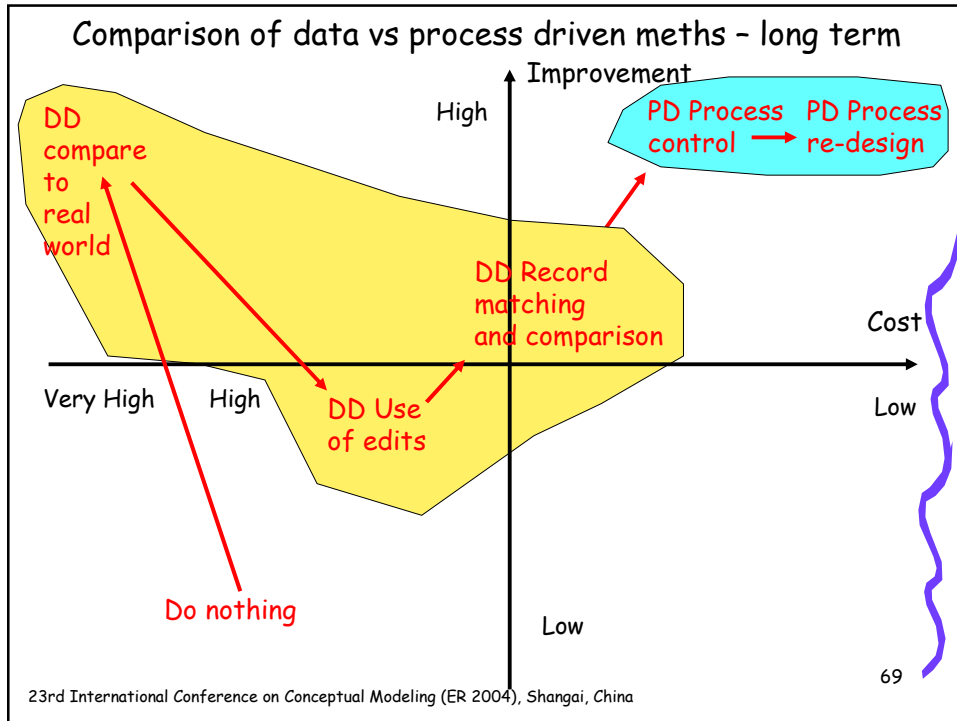
- Types of strategies
- Types of methodologies
- General overview of 4 methodologies
- Relevant steps in methodologies
- Comparison of methodologies

[Redman 1996] classification of  
types of strategies

- **Data driven**
- The strategy is based on using directly data sources
- **Process driven**
- The data production process is analyzed and changed

[Redman 1996] analysis of  
**types of strategies**

- Types of strategies considered
  - **Do nothing**
  - Three data driven strategies
    - Compare to real world
    - Record matching and comparison
    - Use of edits
  - Two process driven strategies
    - Process control
    - Process redesign

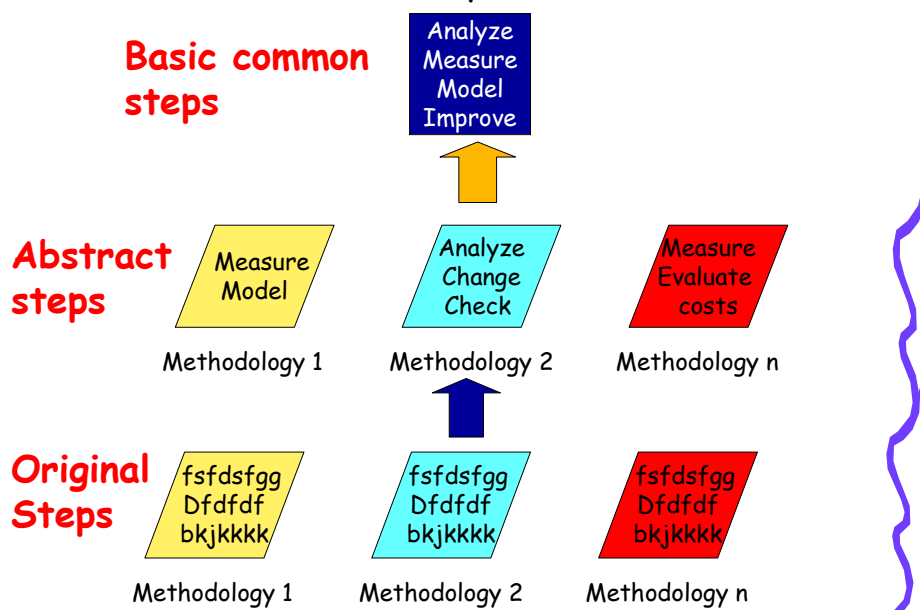


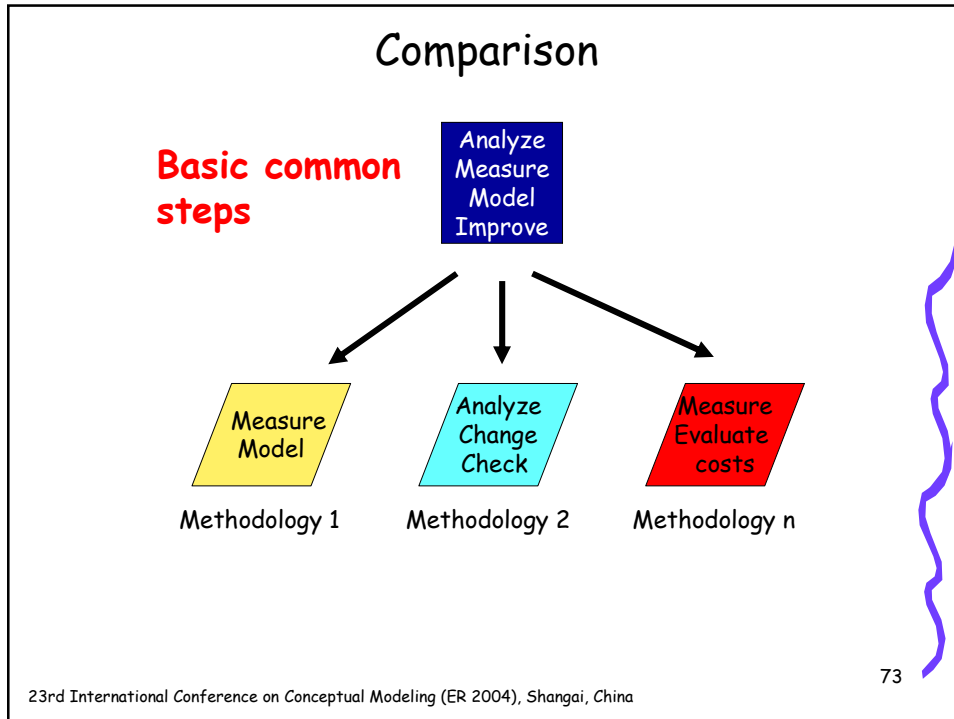
- ## Types of methodologies
- **General purpose (cover a wide set of DQ activities)**
    - For type of information system
      - For intra organizational information systems
      - For inter organizational/cooperative information systems
  - **Specific purpose**
    - For specific phases of DQ
      - DQ assessment
    - For specific types of data
      - Web
- 70
- 23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

### Main references in the literature

Type of methodology	References
General purpose for single organization	[Redman 1996] [English 1998] [Shankaranarayan et al. 2000]
General purpose for cooperative information systems	[Istat 2004] [Bertoletti Batini et al 2004]
Special purpose for assessment	[Wang et al 2001] [Kahn et al] [Pipino et al 2000] [De Amicis Batini 2004]
Special purpose for Web information	[Eppler 2002]

### How basic common steps have been derived





- ### Considered methodologies - 1
- [Redman 1996] Redman T.C.: Data Quality for the Information Age, Artech House, 1996.
    - **The first general purpose methodology proposed for DQ; suitable for managers**
  - [English 1998] Larry P. English Improving Data Warehouse and Business Information Quality, Wiley 1998.
    - **The competing methodology of Redman for professionals.**
- 23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China
- 74

## Considered methodologies- 2

- [Shankaranarayan et al. 2000] (and Wang 1999)  
Shankaranarayan G., Wang R. Y. and Ziad M.: "Modelling the Manufacture of an Information Product with IP-MAP". In Proceedings of the 6th International Conference on Information Quality, Boston, MA, 2000.
  - Originates in a research oriented group
  - Widely applied
- [Istat 2004] Istat and Aipa - A Methodology for improving data quality of address data in Public Administration, 2004 (in Italian).
  - Originates in a public agency whose goal is to provide guidelines to Italian public administrations, conceived specifically for CIS

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

75

## Abstract steps in [Redman 1996]

- **Assessment**
  - Assign responsibilities on processes and data
  - Model the process
  - Quality measurement
- **Improvement**
  - Establish process control
  - Design improvement solutions
  - Check correspondence between requirements and improvement solutions

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

76

## Abstract steps in [English 1998]

- **Assessment**
  - Data analysis
  - Quality requirement analysis
  - DQ measurement
  - **Non quality cost evaluation**
  - **Benefits evaluation**
- **Improvement**
  - Design quality improvement - on data
  - Design quality improvement - on processes
  - Manage improvement solutions - organizational perspective

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

77

## Abstract steps in [Shankaranarayan et al. 2000]

- Steps for one Information Product
  - **Assessment**
    - Data analysis
    - Data quality requirements analysis
    - Define DQ metrics and Perform DQ measurement
    - Model the processes
  - **Improvement**
    - Design improvement solutions
    - **Decide improvement solutions - on processes**

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

78

## Abstract steps in [Istat 2004] - 1

- Global DQ assessment and improvement
  - **Global assessment**
    - DQ requirement analysis
    - **Find critical areas**
      - Define metrics
      - Perform DQ measurement
  - **Global improvement (of data bases)**
    - Design improvement solutions - on data
    - Assign responsibilities - on data
    - Design improvement solutions - on processes
    - Choose tools and techniques

## Steps in [Istat 2004] - 2

- **Internal DQ improvement (for each administration, autonomous initiative)**
  - Design improvement solutions - on processes
  - In critical areas,
    - Assessment - perform DQ measurement
    - Find causes of errors
    - Design improvement solutions
- **Global DQ improvement of inter-administrative flows (characteristic of a Cooperative information System)**
  - Design improvement solutions - on processes
  - Redesign processes

## Basic common steps in methodologies - assessment

- **Global vs specific step**
- Data analysis
- DQ requirements analysis
- Find critical areas
- Model the process
- Perform measurement
- Non quality cost evaluation
- Benefit evaluation
- Assign responsibilities on processes
- Assign responsibilities on data
- Choose tools and techniques

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

81

## Basic common steps in methodologies - improvement

- **Global vs specific step**
- Find causes of errors
- Establish process control
- Design improvement solutions - on data
- Design improvement solutions - on processes
- Redesign processes
- Manage improvement solutions - organizational perspective
- Check effectiveness of improvements

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

82

### Comparison of meths for DQ - assessment phases

Phase	Re96	En98	Wa99	Is04
Global vs specific step			X	X
Data analysis		X	X	
DQ requirements analysis		X		X
Find critical areas				X
Model the process	X		X	
Perform measurement	X	X	X	X
Non quality cost evaluation		X		
Benefit evaluation		X		
Assign responsibilities on processes	X			
Assign responsibilities on data				X
Choose tools and techniques				X

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

83

### Comparison of meths for DQ - improvement phases

Phase	Re96	En98	Wa99	Is04
Global vs specific step				X
Find causes of errors				X
Establish process control	X			
Design improvement solutions - on data	X	X		X
Design improvement solutions - on processes	X	X	X	X
Redesign processes				X
Manage improvement solutions - organizational perspective		X		
Check effectiveness of improvements	X			

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

84

### Comparison of meths for DQ - improvement phases

Phase	Re96	En98	Wa99	Is04
Global vs specific step				X
Find causes of errors				X
Establish process control	X			
Design improvement solutions - on data	X	X		X
Design improvement solutions - on processes	X	X	X	X
Redesign processes				X
Manage improvement solutions - organizational perspective		X		
Check effectiveness of improvements	X			

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

85

### Comparison criteria of methodologies

- Intra/Inter organization
- Data /Process driven
- Level of formalism
- Intended users
- Maturity

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

86

### Criteria based comparison of meths

Criteria	Re96	En98	Wa99	Is04
Intra/inter organization	Single org.	Single org.	Single org, Single process	Cooperative information systems
Data / Process driven	Both	Process	Process	Both
Level of formalism	Very simple	Very simple (e.g.charts)	Highly detailed model	Statistical formulation
Intended users	Managers	Managers	IQ professionals	Managers + Professionals + DBAs
Maturity	High	High	High	Case studies

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

87

3'. Models (54)  
 Short profiles  
 Detailed profiles

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

88

## Models - table of contents

- Use of models
- Extension of DB models
- Models for management information systems
  - Process models
  - Data models
- Cost models

## Use of models in DQ

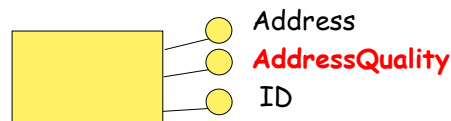
- Extension of DB models
  - Both at the conceptual level and at the logical level there is the need to understand how to enrich models with new structures for defining and analyzing quality
- Models for management information systems
  - New process and data models are needed to help the analyst to understand causes of errors, bottlenecks, perform measures, design improvements
- Cost models
  - Cost models are needed for comparing actual costs of low quality, costs of improvement processes, benefits and savings, and comparing tradeoffs between goals.

## Extensions of DB Models

- **ER Model**: Quality Entity Relation (QER)
- **Relational model** to store and query quality dimensions: Attribute-based Model
- **Relational model** to tag data sources in a data integration setting: Polygen Model
- **XML data model** to represent and query quality dimensions: Data and Data Quality (D<sup>2</sup>Q)

[Storey et al 1998]: Quality Entity Relationship model - 1

- First Hypothesis: quality of an attribute value as another attribute for the same entity
- Consequence: violation of normalization principles in the relational model ("address quality" is dependent upon "address" which is dependent upon "person-id")



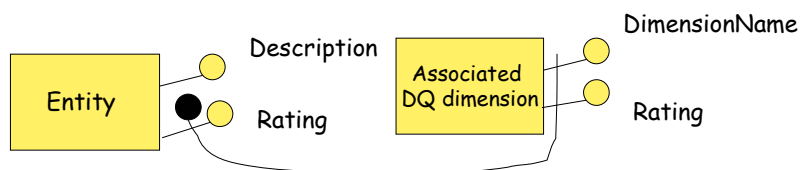
[Storey et al 1998]: Quality Entity Relationship model - 2

### Proposal: Definition of two types of data quality entities

- Data Quality Dimension entity
- Data Quality Measure entity
- **Data Quality Dimension: [Dimension-Name, Rating]**
  - Example: [Accuracy, 1], [Accuracy, 2], [Timeliness, Yes]

[Storey et al 1998]: Quality Entity Relationship model - 3

- **Data Quality Measure: [Rating, description]**
  - Example: [1, excellent] (for accuracy); [Yes, Up-to-date] (for timeliness)



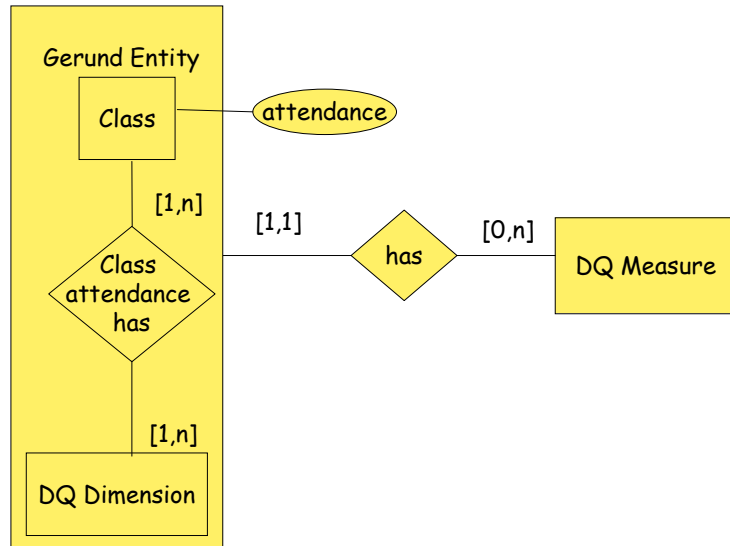
[Storey et al 1998]: Quality Entity Relationship model - 4

- If the dimension metric scale depends upon the attribute, then:
  - Data Quality Dimension: [Dimension-Name, Attribute, Rating]
    - Example: [Accuracy, Name,1], [Accuracy, Cost, good]
- If the accuracy rating for Name has a different interpretation than accuracy for Cost, then:
  - Data Quality Measure: [Dimension-Name, Attribute, Rating, Description]
    - Example: [Accuracy, Name,1, excellent], [Accuracy, Cost, 90%, correct]

[Storey et al 1998]: Quality Entity Relationship model - 5

- In the ER model there is no direct mechanism to associate an attribute (e.g. "attendance") of one entity (e.g. "class") with another entity (a Data Quality Dimension entity or even a Data Quality Measure entity)
- Introduction of the *Attribute Gerund Representation*
  - Creation of the relationship *Class attendance-has Data Quality Dimension*
  - This is represented as unique entity, *Gerund* entity
  - A relationship between the gerund and Data Quality Measure entity provides a mechanism for retrieving the descriptions of the data quality dimension values

## Association of Dq measure to an attribute of an entity



23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

97

## [Wang et al 1995]: Attribute-based model

- Extends the relational model with quality values associated to each attribute value
- The extension regards:
  - Data Structure
  - Data Manipulation
  - Data Integrity

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

98

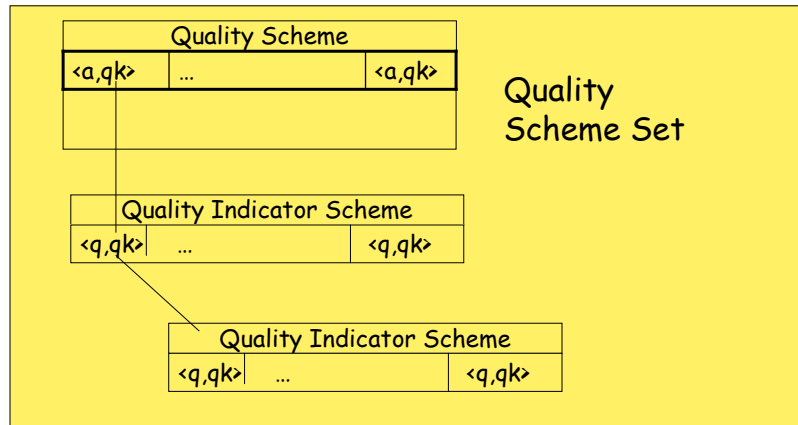
## Data Structure - 1

- An attribute may have an arbitrary number of underlying levels of quality indicators, to which it is linked through a **quality key**
- The couple  $\langle a, qk \rangle$ , where  $a$  is an attribute and  $qk$  is a quality key is called **quality attribute**. Such an expanded scheme is indicated as **quality scheme**
- At instance level, there are **quality tuples** and **quality relations**

## Data Structure - 2

- Each quality key allows to access to a kind of quality tuple called **quality indicator tuple**
- At schema level there are **quality indicator schemes** and at instance level **quality indicator relations**
- A **quality scheme set** is the collection of a quality scheme and all the quality indicator schemes associated to it

## Quality Scheme, Quality Indicator Scheme and Quality Scheme Set



23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

101

## Data Manipulation & Data Integrity

- **Data Manipulation:** Definition of quality relational algebraic operations:
  - Selection, Projection, Union, Difference and Cartesian Product
- Select, from, where clause structure of SQL extended with the clause "**with quality**" to specify the quality requirement of an attribute in a query
- **Data Integrity:** an attribute value and its corresponding quality indicators values (including all descendants) are treated as an *atomic unit*. Integrity rules for insertion, deletion and modification are defined.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

102

[Wang et al 1990]: the Polygen Model - 1

- Proposed for heterogeneous database systems
- Addresses **originating data source tagging** and **intermediate data source tagging**, i.e. "where did data come from" and "which intermediate data sources were used to derive data"
- Extends the relational model in order to trace data sources and intermediate sources

[Wang et al 1990]: the Polygen Model - 2

- A **Polygen (from multiple sources in Greek) domain** is a set of ordered triples:
  - Datum
  - set of local dbs **originating** the datum
  - set of **intermediate** dbs whose data led to the **selection** of the datum
- A **Polygen algebra** is defined in which:
  1. Project, Cartesian product, Union and Difference are extended from the relation model
  2. The restrict operation is introduced in order to select tuples in a polygen relation that satisfy a given condition, and they go to populate intermediate sources

[Wang et al 1990]: the Polygen Model - 3

Polygen algebra cont'd

3. Select and join are defined in terms of the restrict operator, so they also involve intermediate sources
  4. Instead, project, cartesian product and union do not involve intermediate sources
  5. New operators are introduced e.g. Coalesce. Coalesce takes two columns as input and coalesce them into one column (no inconsistency is admitted)
- Observation: a sort of ancestor of provenance (see open issues)

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

105

[Scannapieco et al 2004]

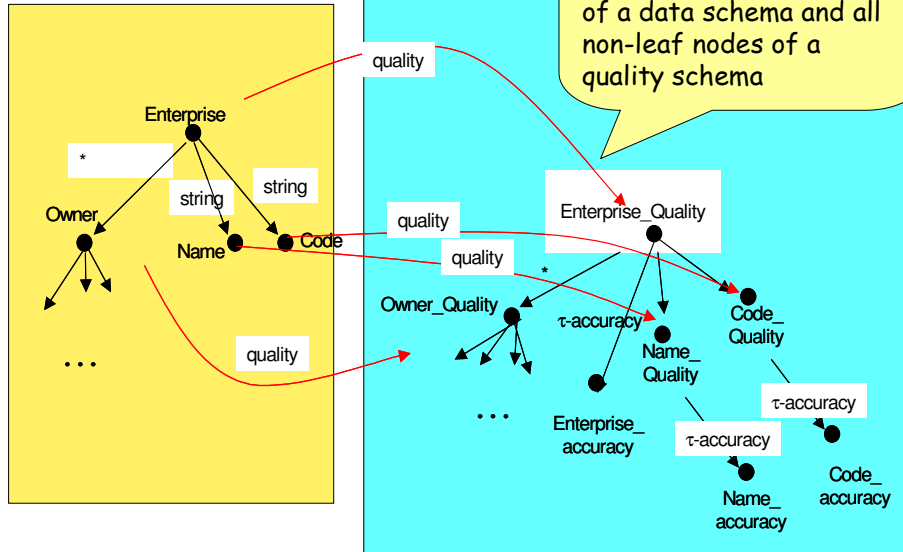
D<sup>2</sup>Q: Data and Data Quality Model

- Graph-based data model, enhancing the semantics of the XML Data Model to represent quality data
- Quality is associated to data in order to:
  - Certify the "correctness" (accuracy, consistency, currency) and completeness of data
    - Benefits for cooperation
  - Support for instance level reconciliation

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

106

## Example of D<sup>2</sup>Q Schema



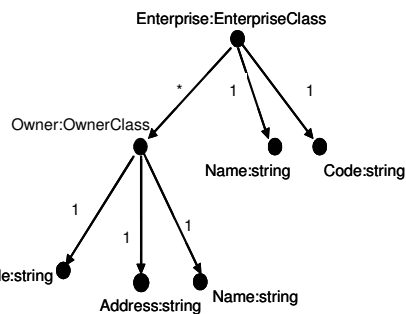
**Quality Association**  
Biunivocal functions among all nodes of a data schema and all non-leaf nodes of a quality schema

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

## Data Schema

- **Data class**
- $\delta(\text{name}_\delta, \pi_1, \dots, \pi_n)$ 
  - Name:  $\text{name}_\delta$
  - Set of properties  $\pi_i$
- =  $\langle \text{name}_i; \text{Type}_i \rangle$  where:
  - $\text{name}_i$  is the name of the property  $\pi_i$
  - $\text{Type}_i$  can be
    - (i) a basic type
    - (ii) a data class
    - or
    - (iii) a type set-of  $\langle X \rangle$ , where  $\langle X \rangle$  can be either a basic type or a data class

- **Data Schema:**  
Node- and Edge-Labelled  
Direct Acyclic Graph of **data classes**

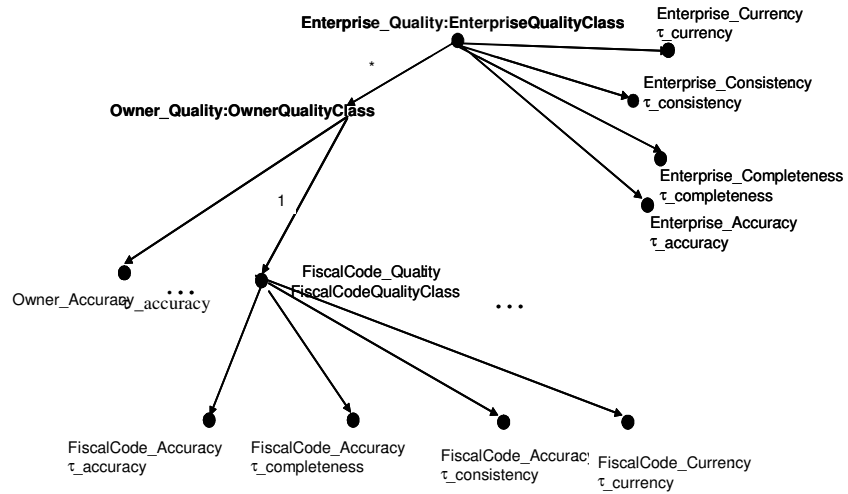


23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

108

## Quality Schema

- **Quality Class**  $\lambda_\delta$  associated to a data class  $\delta$

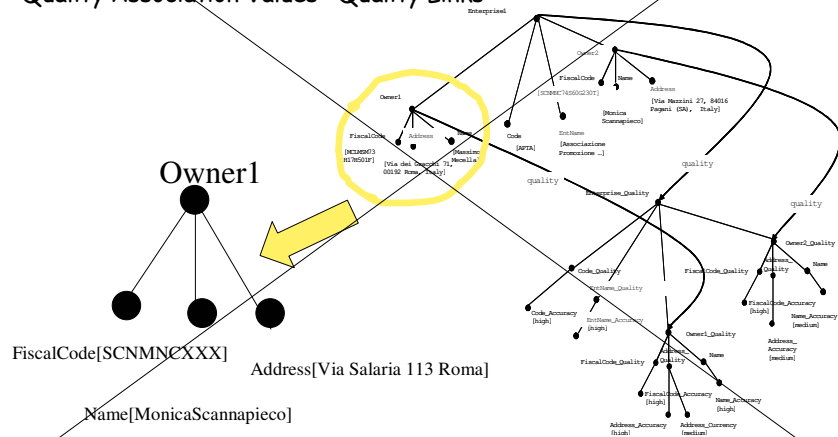


23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

109

## D<sup>2</sup>Q Schema Instances

- Data Classes Instance  $\rightarrow$  Data Objects
- Quality Classes Instance  $\rightarrow$  Quality Objects
- Quality Association Values  $\rightarrow$  Quality Links



23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

110

## Querying D<sup>2</sup>Q

- Usage of Xquery extended with Quality Selectors
- Quality Selectors are user-defined functions that allows to access quality values
- Example:

```
define function accuracy($n as node*) as
node*{
  let $root := searchroot($n),
  qualitydoc:=document(string($root/@qualityfi
le))
  for $q in $n/@quality, $r in
  $qualitydoc//*[ @qOID eq $q]/accuracy
  return $r }
```

## Models for management information systems process and data models

## Evolution of process models for Management Information Systems

Information Manufacturing System (IMS)  
[Ballou et al 1998]

Shankaranarayan et al 2000  
IPMAP

Pierce 2002  
EPC

Scannapieco et al 2002  
IP-UML

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

113

[[Shankaranarayan](#) et al 2000]: IPMAP - 1

- An Information Production Map (IP-Map) is a graphical model designed to help people to comprehend, evaluate, and describe how an **information product** such as an invoice, customer order, or prescription is assembled.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

114

## [Shankaranarayan et al 2000]: IPMAP -2









- IP-Map Objectives

- to visualize the most important phases in the manufacture of an IP and identify the critical phases that affect its quality.
- To identify ownership of the processes and help in implementing quality-at-source.
- to understand the organizational (business units) as well as information system boundaries spanned by the different processes / stages in the IP-Map.
- To measure the quality of the IP at the different stages in the manufacturing process using appropriate quality dimensions.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

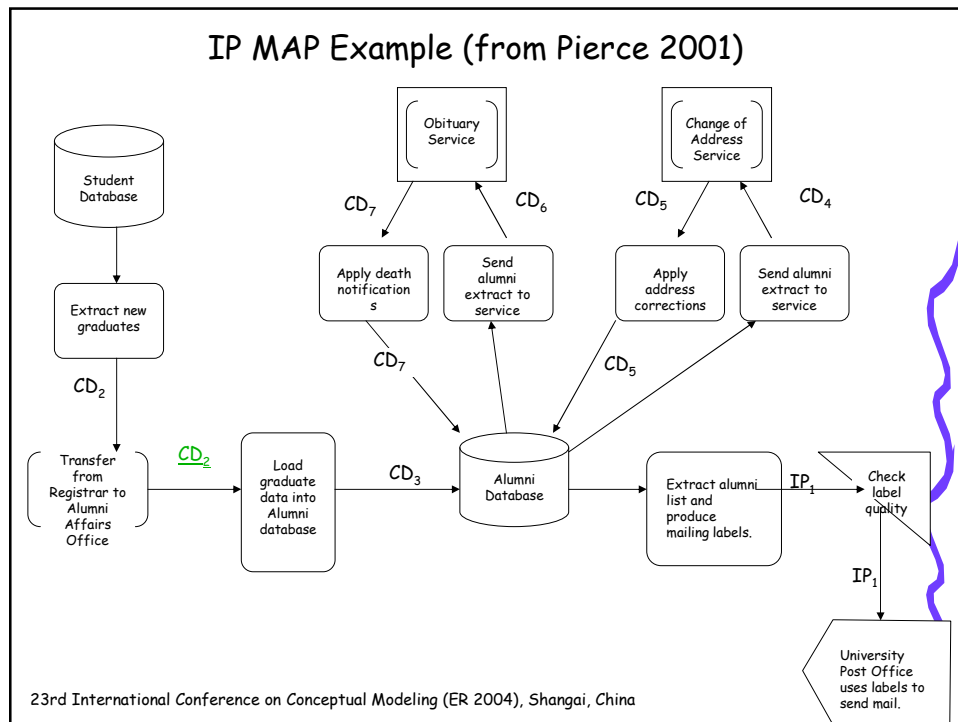
115

## IPMAP blocks

Block name	Symbol
Source (raw input data)	
Customer (output)	
Data Quality function	
Processing	
Data Storage	
Decision	
Business Boundary	
Information System Boundary	

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China







116



## [Pierce 2002]: IP-MAP + Event Process Chain (EPC)

- Extends IP Map in order to represent:
  - The **event** that triggers the use of data by a process
  - The **communication structure** between sources, consumers and organization groups.
  - The **hierarchy of organizational groups/functions**
  - The **relationship between info products, storages, and other data components.**






## Event Process Chain blocks - 1

Construct	Description	Type	Icon
<b>Event</b>	A trigger for a business process	EPC	
<b>Organization</b>	Outline structure of an enterprise	EPC	
<b>Information or resource object</b>	Business object entity	EPC	
<b>Logical operator</b>	Logical relationships between events, processes, etc	EPC	
<b>Control flow</b>	Time dependencies between events and processes	EPC	
<b>Resource/organization unit</b>	Relationships between resource and org.unit	EPC	

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

119

## EPC blocks - 2

Construct	Description	Type	Icon
<b>Information/material flow</b>	Relationship between IP and function that processes it	EPC/ IPMAP	
<b>Function/process block</b>	Function that describe manipulations or transformations performed on products	EPC/ IPMAP	
<b>Source block</b>	Source of each raw input data	IPMAP	
<b>Customer block</b>	The consumer of the information product	IPMAP	
<b>DQ block</b>	Checks on objects for data quality	IPMAP	

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

### [Scannapieco et al 2002]: IP-UML

- The IP-MAP framework already proposes a modeling formalism, which is based on Data Flow Diagrams
- IP-UML proposes a data quality profile of UML based on IP-MAP
- The use of UML instead of IP-MAP formalism has the following advantages:
  - UML is a *standard language* and CASE tools for it have been implemented
  - UML is a language supportive of analysis, design and implementation artefacts, so the *same language can be used in all the phases of the process*
  - the *expressive power* of UML is greater, especially with reference to the process modelling constructs

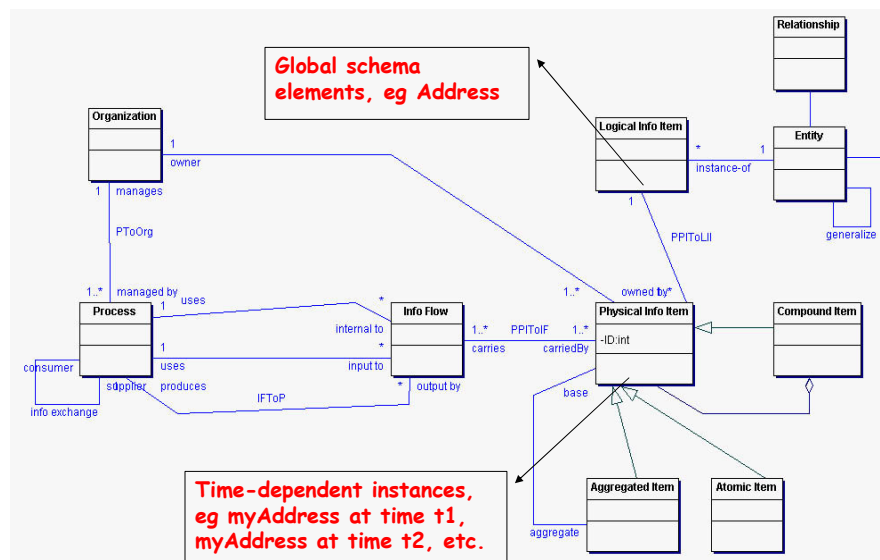
### IP-UML Constructs

Stereotype	Base Class	Description
<<processing>> Processing Activity	activity	It represents IP-MAP processing block
<<quality>> Quality Activity	activity	It represents IP-MAP quality block
<<customer>> Customer Actor	actor	It represents IP-MAP customer block
<<source>> Source Actor	actor	It represents IP-MAP source block
<<dataStorage>> Data Storage Actor	actor	It represents IP-MAP data storage block
<<L/E>> Load/Extract Dependency	dependency	The two elements of the relationship have the role of Loader/Extractor and of the source from which loading/extracting

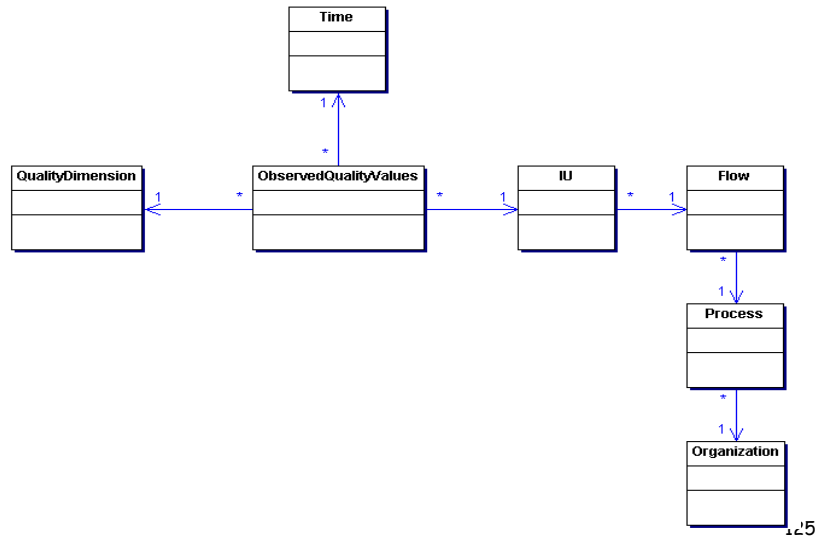
## Extensions of data models

- [Missier Batini 2003] A multidimensional model for Information Quality in Cooperative Information Systems
- Traditional data models and methodologies deal separately
  - Elementary and aggregate data,
  - Data represented in different processes
  - Data managed by different organizations
- [Missier Batini 2003] provides a model to represent in an integrated fashion all these aspects.

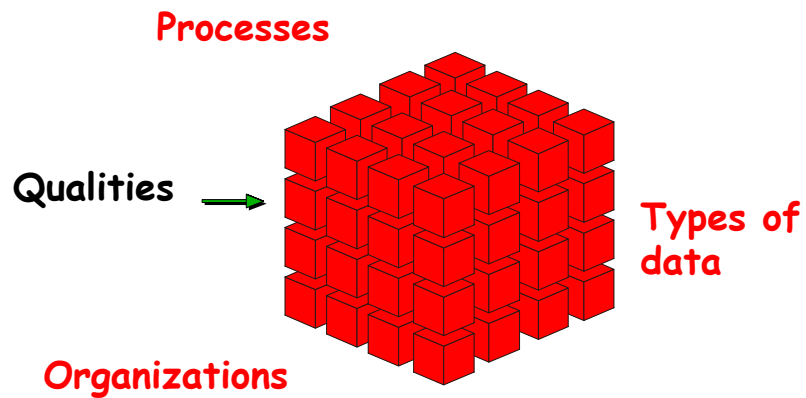
## Data Model



## Quality Model - dimensional modeling for IQ



## Data quality cube



23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

## Cost models - 1

Problem addressed in [Ballou 1999]: maximizing the total value of selected DQ improvement projects related to data warehouse

1. Determine the organizational activities the data warehouse will support
2. Identify all sets of data needed to support the organizational activities;
3. Estimate the quality of each data set on each relevant data quality dimension: **Current quality  $CQ(J, K)$ .**
4. Identify a set of potential projects (and their cost) that could be undertaken for enhancing or affecting data quality;

## Cost models - 2

5. Estimate for each project the likely effect of that project on the quality of the various data sets, by data quality dimension: **Anticipated quality  $AQ(J, K, L)$ .**
6. Determine for each **activity**, data set, and relevant data quality dimension the change in utility should a particular project be undertaken: **Utility(I, J, K, L).**

The problem is formulated as an Integer Linear Programming problem with constraints

## Cost models - 3

Problem Addressed in [[Avenali Batini et al 2004](#)]:  
matching information demand to information  
offer in cooperative information systems, under  
quality constraints, minimizing on the cost

Inputs: global schema + local schemas + mappings; data  
and quality offer, data and quality demand, cost  
model

Output: offer/ demand matching with minimal cost.

Phases:

1. Choice of candidate fragments to be chosen +  
qualities
2. Formulation of a Integer Linear Programming problem  
with constraints

## 4. Techniques (135)

## Techniques - table of contents

- Relevant activities in DQ
- Techniques for record linkage/object identification/record matching
- Techniques for data integration
- Other DQ activities
  - Techniques for profiling
  - Hints on data editing

## Relevant activities in DQ

## Relevant activities in DQ - 0

- Record Linkage/Object identification/Entity identification/Record matching
- Data integration
  - Schema matching
  - Instance conflict resolution
  - Source selection
  - Result merging
  - Quality composition
- Error localization/Data Auditing
  - Data editing-imputation/Deviation detection
- Profiling
  - Structure induction
- Data correction/data cleaning/data scrubbing
- Schema cleaning
- Tradeoff/cost optimization

## Relevant activities in DQ - 1

- Record Linkage/Object identification/  
Entity identification/Record matching
  - Given two tables or two sets of tables, representing two entities/objects of the real world, find and cluster all records in tables referring to the same entity/object instance.

## Relevant activities in DQ - 2

- Data integration
- 1. Schema matching
  - Takes two schemas as input and produces a mapping between semantically correspondent element of the two schemas

## Relevant activities in DQ - 3

- Data Integration
- 2. **Instance** conflicts resolution & Merging
  - Instance level conflicts can be of three types:
    - representation conflicts, e.g. dollar vs. euro
    - key equivalence conflicts, i.e. same real world objects with different identifiers
    - attribute value conflicts, i.e. instances corresponding to same real world objects and sharing an equivalent key, differ on other attributes

## Relevant activities in DQ - 4

- Data integration
- 3. Result merging: *it derives from the combination of individual answers into one single answer returned to the user*
  - For numerical data, it is often called *fused answer* and a single value is returned as an answer, potentially differing from each of the alternatives

## Relevant activities in DQ - 5

- Data integration
- 4. Source selection
- *Querying a multidatabase with different sources characterized by different qualities*
- 5. Quality composition
- *Defines an algebra for composing data quality dimension values*

### Relevant activities in DQ - 6

- Error localization/Data Auditing
- Given one/two/n tables or groups of tables, and a group of integrity constraints/qualities (e.g. completeness, accuracy), find records that do not respect the constraints/qualities.
  - Data editing-imputation
  - Focus on integrity constraints
  - Deviation detection
  - data checking that marks deviations as possible data errors

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

139

### Relevant activities in DQ - 7

- Profiling
  - Evaluating statistical properties and intensional properties of tables and records
  - Structure induction of a structural description, i.e. "any form of regularity that can be found"

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

140

## Relevant activities in DQ - 8

- Data correction/data cleaning/data scrubbing
  - Given one/two/n tables or groups of tables, and a set of identified errors in records wrt to given qualities, generates probable corrections (deviation detection) and correct the records, in such a way that new records respect the qualities.

## Relevant activities in DQ - 9

- Schema cleaning
  - Transform the conceptual schema in order to achieve or optimize a given set of qualities (e.g. Readability, Normalization), while preserving other properties (e.g. equivalence of content)
- Tradeoff/cost optimization
  - Tradeoff - When some desired qualities are conflicting (e.g. completeness and consistency), optimize such properties according to a given target
  - Cost - Given a cost model, optimize a given request of data quality according to a cost objective coherent with the cost model

## Focus in the tutorial

- Record Linkage/Object identification/Entity identification/Record matching
- Data integration
  - ~~Schema matching~~
  - Instance conflict resolution
  - Source selection
  - Result merging
  - Quality composition
- Error localization/Data Auditing
  - Data editing-imputation/Deviation detection
- Profiling
  - Structure induction
- Data correction/data cleaning/data scrubbing
- ~~Schema cleaning~~
- ~~Tradeoff/cost optimization~~

## Techniques for Record Linkage/ Object identification/ Entity identification/Record matching (125)

## Techniques for record linkage/ object identification/record matching

- Introduction to techniques
- General strategies
- Details on specific steps
- Short profiles of techniques
- [Detailed description] (24)
- [Comparison] (10)

## Introduction to techniques

## An example

Id	Name	Type of activity	City	Address
RI BRTBNT40C211891T	CENTRO CARNI DI BARTOLETTI BENITO	52221 COMMERCIO AL DETTAGLIO DI CARNI BOVINE, SVINE, EQUINE, OVINE E CAPRINE	AL NOVI LIGURE	STRADA STATALE 35 BISS DEI GIOV 13/12/1994
INPS 01638630061	BARTOLETTI BENITO CENTRO CARNI	70201 C.H. ALIMENTI, BEVANDE, TABACCHI	AL POZZOLO FORMIGARO	VIA ROMA 9 01/12/1994
INAIL BRTBNT40C211891U	CENTRO PIEMONTE CARNI DI BARTOLETTI B.	0133 MACELLERIE SENZA MATTAZIONE	AL OVADA	PZA MAZZINI 4 28/11/1994

The same business as represented in the three most important business data bases in central agencies in Italy:

- CC Chambers of Commerce
- INPS Social security
- INAIL Accident Insurance

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

147

## The three CC, Inps and Inail records

CNCBTB765SDV	Meat production of Bartoletti Benito
Retail of bovine and ovine meats	National Street dei Giovi
Novi Ligure	
0111232223	Bartoletti Benito meat production
Grocer's shop, beverages	9, Rome Street
Pizzolo Formigaro	
CNCBTR765LDV	Meat production in Piemonte of Bartoletti Benito
Butcher	4, Mazzini Square
Ovada	

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

## Record linkage and its evolution towards object identification in databases ....

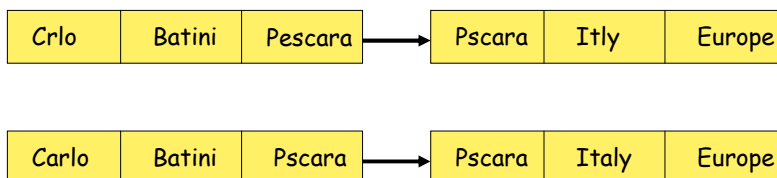
### Record linkage

First record and second record represent the same aspect of reality?

Crlo	Batini	55	Carlo	Btini	54
------	--------	----	-------	-------	----

### Object identification in databases

First group of records and second group of records represent the same aspect of reality?



23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

149

## Type of data considered

- **(Two) formatted tables**
  - Homogeneous in common attributes
  - Heterogeneous
    - Format (Full name vs Acronym)
    - Semantics (e.g. Age vs Date of Birth)
    - Errors
- **(Two) groups of tables**
  - Dimensional hierarchy
- **(Two) XML documents**

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

150

...and object identification in semistructured documents

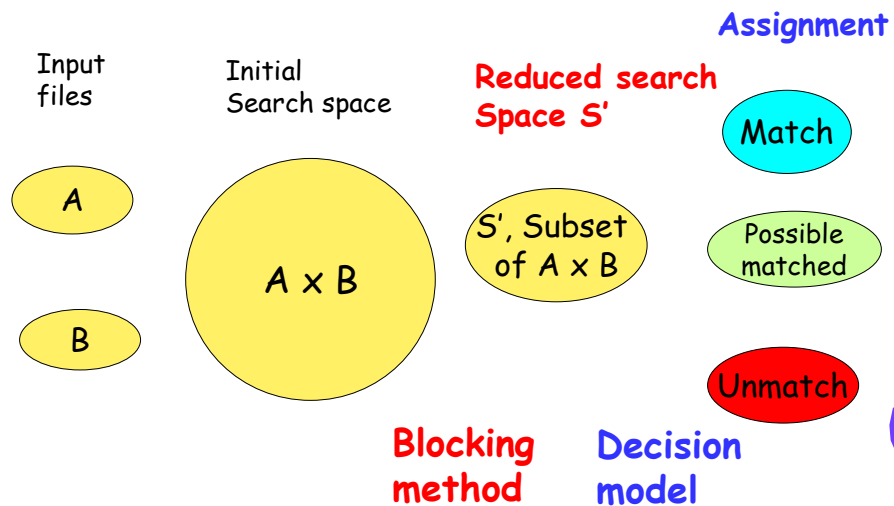
```
<country>
  <name> United States of America </name>
  <cities> New York, Los Angeles, Chicago
  </cities>
  <lakes>
    <name> Lake Michigan </name>
  </lakes>
</country>
```

```
<country>
  United States
  <city> New York </city>
  <city> Los Angeles </city>
  <lakes>
    <lake> Lake Michigan </lake>
  </lakes>
</country>
```

and

are the same object?

## Relevant steps



## General strategy and phases - 1

- 0. Preprocessing
  - Standardize fields to compare and correct simple errors
- 1. Establish blocking method (also called searching method)
  - Given the search space  $S = A \times B$  of the two files, find a new search space  $S'$  contained in  $S$ , to apply further steps.
- 2. Choose comparison function
  - Choose the function/set of rules that express the distance between pairs of records in  $S'$
- 3. Choose decision model
  - Choose the method for assigning pairs in  $S'$  to  $M$ , the set of matching records,  $U$  the set of unmatching records, and  $P$  the set of possible matches
- 4. Check effectiveness of method

## General strategy and phases - 2

- [Preprocessing]
- Establish a blocking method
- Compute comparison function
- Apply decision model
- [Check effectiveness of method]

## Paradigms used in Techniques

- Empirical - **heuristics, algorithms, simple mathematical properties**
- Statistical/probabilistic - **properties and formalisms in the areas of probability theory and Bayesian networks**
- Knowledge based - **formalisms and models in knowledge representation and reasoning**

## General strategy - **probabilistic methods**

- [Preprocessing]
- Establish blocking method
- Compute comparison function
- **Compute probabilities and thresholds**
  - **With training set**
  - **Without training set**
- [Check effectiveness of method]

### General strategy - knowledge based methods

- [Preprocessing]
- Establish blocking method
- **Get domain knowledge**
- **Choose transformations/rules**
- **Choose most relevant rules**
- **Apply rules using a knowledge based engine**
- [Check effectiveness of method]

### General strategy - hierarchical structures/tree structures

- [Preprocessing]
- **Structure traversal**
- Establish blocking/filtering method
- Compute comparison function
- Apply **local/global** decision model
- [Check effectiveness of method]



## Details on specific steps

## Preprocessing

- **Error correction (with the closest value in the domain)**
  - E.g. Crlo → Carlo
- **Elimination of stop words**
  - Of, in, ....
- **Standardization**
  - E.g.1 Bob → Robert
  - E.g.2

Mr George W Bush	Mr	George	W	Bush
George Bush		George		Bush

## Types of Blocking/Searching methods - 1

- **Blocking** - partition of the file into mutually exclusive blocks.
- Comparisons are restricted to records within each block.
- Blocking can be implemented by:
  - **Sorting** the file according to a block key (chosen by an expert).
  - **Hashing** - a record is hashed according to its block key in a hash block. Only records in the same hash block are considered for comparison.

## Type of Blocking/Searching methods - 2

- **Sorted neighbour or Windowing** - moving a window of a specific size  $w$  over the data file, comparing only the records that belong to this window
- **Multiple windowing** - Several scans, each of which uses a different sorting key may be applied to increase the possibility of combining matched records
- **Windowing with choice of key** based on the quality of the key

## Comparison functions

Type	Example
Identity	'Batini' = 'Batini'
Simple distance	'Batini' similar to 'Btini'
Complex distance	'Batini' similar to 'Btaini'
Error driven distance(Soundex)	Hilbert similar to Heilbpr
Frequency weighted distance	1. Smith is more frequent than Zabrinky 2. In IBM research, IBM less frequent than research
Transformation	John Fitzgerald Kennedy Airport Acronym → JFK Airport
Rule (Clue)	Seniors appear to have one residence in Florida or Arizona and another in the Midwest or Northeast regions of the country

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

163

## More on comparison functions

### String comparators 1

1. **Hamming distance** - counts the number of mismatches between two numbers or text strings (→ fixed length strings)
2. **Edit distance** - the minimum cost to convert one of the strings to the other by a sequence of character insertions, deletions and replacements. Each one of these modifications is assigned a cost value.
3. **Jaro's algorithm or distance** finds the number of common characters and the number of transposed characters in the two strings. J distance accounts for insertions, deletions, and transpositions.  
A **common character** is a character that appears in both strings within a distance of half the length of the shorter string. A **transposed character** is a common character that appears in different positions.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

164

## More on comparison functions - String comparators 2

4. **N-grams comparison function** forms the set of all the substrings of length  $n$  for each string. The distance between the two strings is defined as square radix of the sum of differences among the number of occurrences of the substring  $x$  in the two strings  $a$  and  $b$ , respectively.
5. **Soundex code** clusters together names that have similar sounds. For example, the Soundex code of "Hilbert" and "Heilbpr" is similar.
6. **Weighted distance** - considers the frequency of values

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

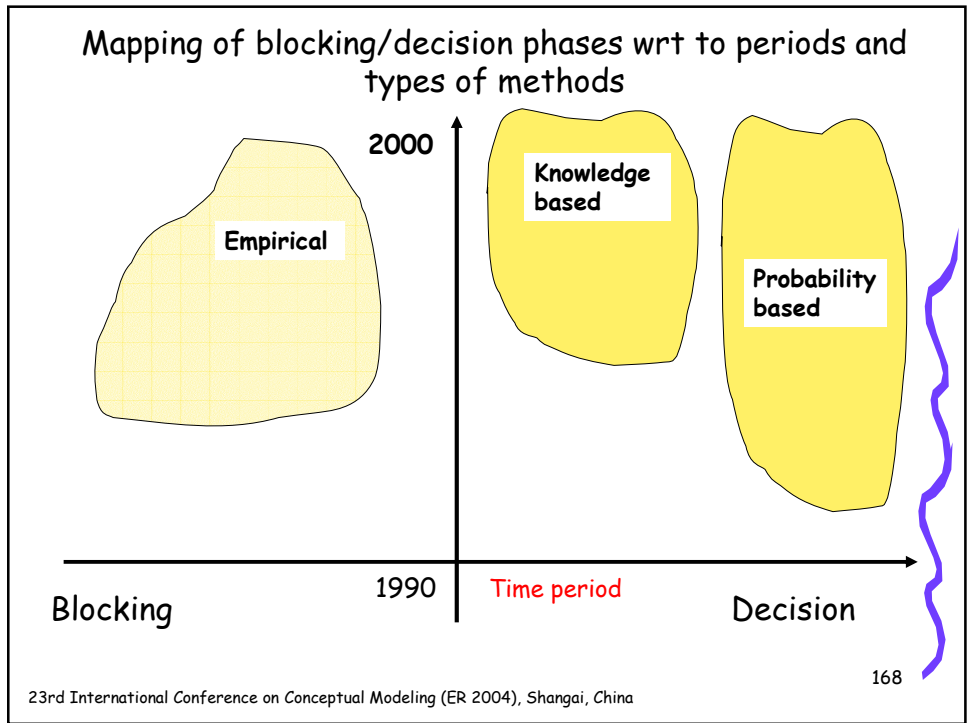
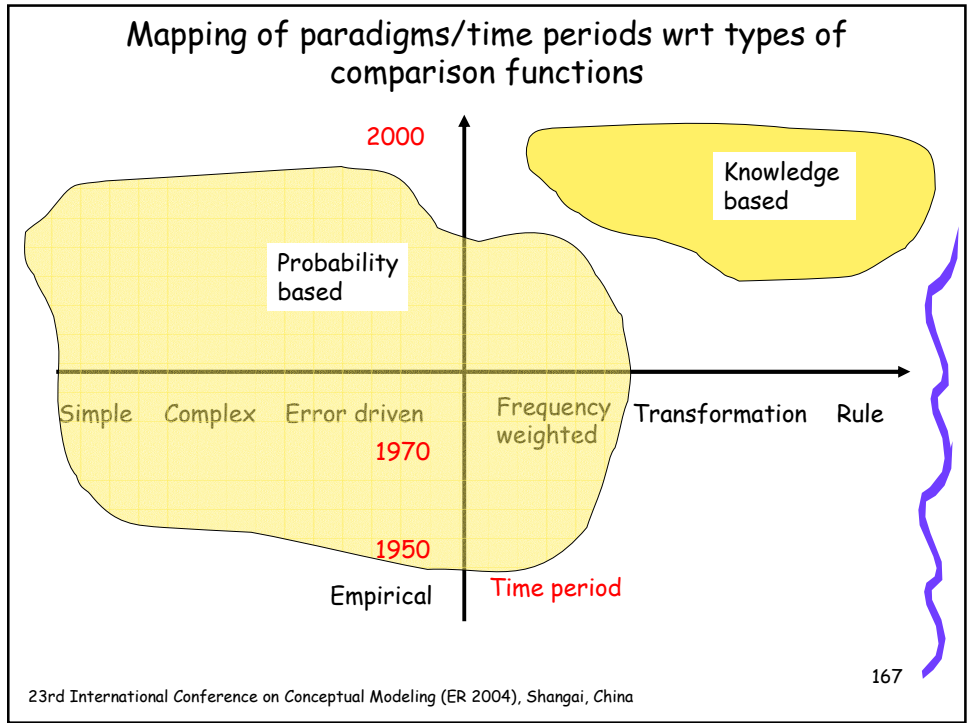
165

## Main criteria of use

Hamming distance	Used primarily for numerical fixed size fields like Zip Code or SSN
Edit distance	Can be applied to variable length fields. To achieve reasonable accuracy, the modifications costs are tuned for each string data set.
Jaro's algorithm and Winkler's improvement	The best one in several experiments
N - grams	Bigrams ( $n = 2$ ) effective with minor typographical errors
Soundex code	Effective for dictation mistakes
Weighted distance	Effective for person's names and surnames (eg Smith vs Zabrinisky)

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

166



## Short profiles of techniques (probabilistic, knowledge based, empirical)

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

169

### List of techniques

Name	Main Reference	Type of strategy
Fellegly and Sunter family	[Fellegi 1969]	probabilistic
[Nigam 2000]	[Nigam 2000]	probabilistic
Cost based	[Elfeky 2002]	probabilistic
Induction	[Elfeky 2002]	probabilistic
Clustering	[Elfeky 2002]	probabilistic
Hybrid	[Elfeky 2002]	probabilistic
1-1 matching	[Winkler 2004]	probabilistic
Bridging file	[Winkler 2004]	probabilistic
Delphi	[Ananthakr. 2002]	empirical
XML object identification	[Weis 2004]	empirical
Sorted Neighbour and variants	[Hernandez 1995] [Bertolazzi 2003]	empirical
Instance functional dependencies	[Lim et al 1993]	Knowledge based
Clue based	[Buechi 2003]	Knowledge based
Active Atlas	[Tejada 2001]	Knowledge based
Intelliclean and SN variant	[Low 2001]	Knowledge based

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

170

## Added techniques

Name	Main Reference	Type of strategy
Approximate string join	[Gravano 2001]	empirical
Text join	[Gravano 2003]	empirical
Flexible string matching	[Koudas 2004]	empirical
Approximate XML joins	[Guha 2002]	empirical

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

171

## For each technique

- Name
- [Main addressed phase]
- **Type of data / paradigm**
- Main idea
- [Pros]
- [Cons]

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

172

## Fellegi and Sunter family

- **Two files/ probabilistic**
- Compute in some way conditional probabilities of matching/not matching based on an agreement pattern (comparison function).
- Establish two thresholds for deciding matching and non matching, based on a priori error bounds on false matches and false nonmatches.
- Also called error based techniques
- Pros: well founded technique; widely applied, tuned and optimized.
- Cons: in most cases, needed training data for good estimates.

## Cost based [Elfeky 2002]

- **Two files/probabilistic**
- Weights are assigned to decision methods, that represent costs of non matching
- Pro: **The costs of incorrect matching are minimized**

## Induction [Elfeky 2002]

- Two files/ probabilistic
- A classifier is employed to predict the matching status of comparison vectors produced by agreement patterns, on the whole set of record pairs
- Cons: training set needed

## Clustering [Elfeky 2002]

- Two files/probabilistic
- A typical k-means clustering method produces n clusters
- A 3-means clustering method is used to produce three clusters, that are assigned to matched, unmatched, and possibly matched
- Pros: does not need training set

## [[Nigam 2000 paper](#)] and Hybrid

- **Two files/probabilistic**
- Exploits a limited set of training data to extract value from a large set of unsupervised data
- 1. Uses a clustering m. to produce a limited training set.
- 2. Uses the induction m. to produce m/nm assignments.
- Pros: in several applications (e.g. censuses) it is costly to build a large training set

## 1-1 [matching](#) [[Winkler 2004](#)]

- **Two files/probabilistic**
- When two files have few internal duplicates, 1-1 matching is forced because many of the second and third best matches might have matching weights that are sufficiently high to necessitate clerical review and not be true matches.
- Pros: **More efficient in time and space than FS**

## Bridging file [Winkler 2004]

- Two files/ probabilistic
- A bridging file, if available, reduces the search space and improves effectiveness
- Pros: The bridging file, if available, is usually of high quality and leads to higher effectiveness

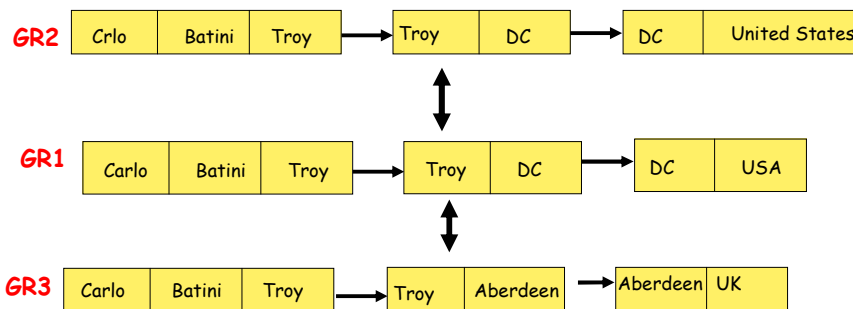
File A				Bridging file			File B			
A11	A12	A13	.....	Name1	Addr1	Zip1	.....	B11	B12	B13
A21	A22	A23	.....	Name1	Addr2	Zip2	.....	B21	B22	B23
A31	A32	A33	.....	Name1	Addr3	Zip3	.....	B31	B32	B33

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

179

## Delphi [Ananthakr. 2002]

- Two dimensional hierarchies/empirical
- Exploits the hierarchical structure of records, using both local (textual) and global (co-occurrence) similarity
- Applies a dynamic tresholding (e.g. names in Argentina are longer than names in USA)

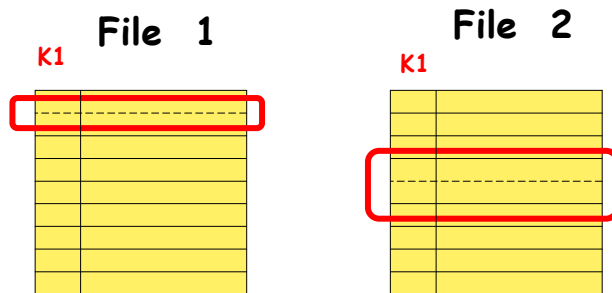


23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

180

## Sorted Neighbour (SN) [Hernandez 1995]

- Two files/Main phase: blocking/empirical
- Improves the blocking method choosing a key and moving a window on sorted files. The key is chosen by an expert.



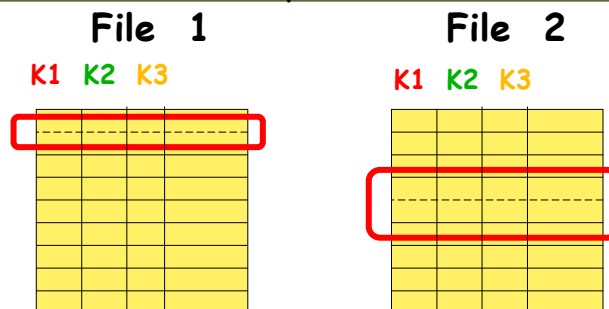
23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

181

## Multi pass SN

- Two files/Main phase: blocking/empirical

Applies SN on several independent keys  
Perform transitive closure on record matched  
in the different steps



23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

182

Variant of Sorted Neighbour [Bertolazzi 2003]

**Two files/Main phase: blocking/empirical**

Optimizes key selection in SN choosing the key according to criteria of **selectivity** and **quality**.

Pros: avoid expert decision; more accurate choice of key and consequent improved effectiveness.

XML object identification [Weis 2004]

- **XML documents/empirical**
- Extends several techniques for searching and decision to XML documents

## KB techniques -Types of knowledge considered

Paper	Name of technique	Type of knowledge
[Lim et al 1993]	Instance level funct. dependencies	Instance level functional dependencies
[Low et al 2001]	Intelliclean	Rules for duplicate identification and for merging records
[Tejadaa 2001]	Active Atlas	Importance of the different attributes for deciding a mapping Transformations among values that are relevant for the mappings
[Buechi et al 2003]	Clue based	Clues, i.e domain dependent properties of data to be linked

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

185

## Instance level functional dependencies

[Lim et al 1993]

- Two files/knowledge based
- Functional dependencies defined at instance level, e.g.  
**Speciality = "Mughalay" → Cuisine = "Indian"**  
are used to decide matching and not matching pairs

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

186

## Intelliclean [Low 2001]

- **Two files/knowledge based**
- Exploits rules (see example) for duplicate identification and for merging records.
- Rules are fed into an expert system engine
- The pattern matching and rule activation mechanisms make use of an efficient method for comparing a large collection of patterns to a large collection of objects.

## Intelliclean Sorted Neighbour

- **Two files/knowledge based**
- Improves multi pass SN by attaching a confidence factor to rules and applying selectively the transitive closure.

## Active Atlas [Tejada 2001]

- Two files / knowledge based
- Extend the "distance concept", by learning the most relevant transformations and mapping rules between attributes

## Clue based [Buechi(222) 2003]

- Two files/ knowledge based
- To find matches make use of clues, i.e. domain dependent relevant properties of data.
- Clues are weighted on a training set, and then used in the decision procedure
- Pro: Claimed to be more effective and efficient than previous methods

## Active Atlas

- **Two files/knowledge based**
- extends the concept of distance between values of attributes into the concept of set of **transformation functions**, e.g. abbreviation, acronym, etc.
- (see later kb techniques for a deeper discussion)

## Short profiles of added techniques

## Approximate string joins in RDBMS [Gravano et al. 2001]

- **DBMS formatted data/empirical**
- Approximate string joins in Web sources within RDBMS
- Based on matching of  $q$ -grams, i.e. short substrings of length  $q$  of the database strings
- Basic intuition: when two strings are within a small edit distance they share many  $q$ -grams
- Auxiliary relational tables to represent  $q$ -grams constructed on the fly
- Pros:
  - Approximate string join functionality built on the top of commercial databases

## Text joins in RDBMS [Gravano et al. 2003] - 1

- **DBMS formatted data/empirical**
- More complex and general way to do approximate string joins within RDBMS
- Application of the tf-idf (term frequency, inverse document frequency) metric used in information retrieval
  - A token (e.g. a word, a  $q$ -gram) is extracted from a db string
  - Each token is associated to a weight that corresponds to the commonality of the token in the db
  - Each string is associated to a weight vector corresponding to extracted tokens

## Text joins in RDBMS [Gravano et al. 2003] - 2

- Definition of a *text join operator* btw two db strings based on computing the cosine similarity metric, i.e. the inner product, of corresponding vectors
- Proposal and validation of a sampling - based strategy for efficiently computing text joins using standard SQL
- Extensions in [Koudas et al 2004] to consider:
  - Multiple string-valued attributes
  - Semantic relationships

## Approximate XML joins [Guha et al 2002]

- XML documents/empirical
- Approximate matching of XML documents based on their tree structure
- Tree edit distance
- Upper and lower bounds for tree edit distance computation
- Pros:
  - Any other metric to compare trees can be used within the proposed framework
- Cons:
  - Only *syntactical* matching, i.e. matching of nodes with same names

## Detailed description of techniques

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

197

## Detailed description of empirical techniques skip

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

198

### Sorted Neighbour searching method

- 1. **Create Keys**: Compute a key K for each record in the list by extracting relevant fields or portions of fields.
- Relevance is decided by experts.
- 2. **Sort Data**: Sort the records in the data list using K
- 3. **Merge**: Move a fixed size window through the sequential list of records limiting the comparisons for matching records to those records in the window. If the size of the window is w records, then every new record entering the window is compared with the previous w - 1 records to find "matching" records

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

199

### Multi pass sorted neighbour method

- In general, no single key will be sufficient to catch all matching records. Attributes that appear first in the key have a higher priority than those appearing after them. To increase the number of similar records merged:
- 1 Execute several independent runs of the sorted neighbourhood method, each time using a different key and a relatively small window.
- 2. Merge set of pairs produced by each independent run
- 3. Apply the transitive closure to those pairs of records

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

200

[[Bertolazzi](#) et al 2003]

In the algorithm the first step **Create key** is improved as follows

- 1. Choose the matching key, based on the optimal value of the formula **quality \* identification power** where
  - A. **quality** is measured according to **accuracy, completeness, consistency**
  - B. the **identification power** is the power of the attribute to discriminate among tuples

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

201

### The [Delphi](#) algorithm

- Perform a top-down tree traversal of the hierarchy of relations. Steps:
  - 1. Process first the top most relation
  - 2. Group relations below into clusters
  - 3. Prune each cluster according to properties of distance functions eliminating tuples that cannot be duplicates.
  - 4. Compare pairs of tuples within each group according to two comparison functions and corresponding thresholds
    - **Textual similarity** between two tuples
    - **Cocurrence** similarity between the children sets of the tuples
  - 5. Dynamically update thresholds

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

202

## XML object identification

- Top down traversal of the XML tree structure,
- 1. Delete from the search space pairs of objects according to four filters:
  - A. length distance
  - B. Triangle inequality
  - C. Bag distance
  - D. object filter
- 2 Measure object similarity using string similarity based on **edit distance** and select matched pairs according to a threshold.
- 3. Cluster selected pairs by computing the **transitive closure**.

## Detailed description of added empirical techniques

## Approximate XML joins [Guha et al 2002] - 1

- Tree edit distance, as the minimum number of operations (node insert, delete, relabel) to transform one tree into another
- Tree edit distance computation is expensive:  $O(n^4)$  for trees of size  $O(n)$

## Approximate XML joins [Guha et al 2002] - 2

- Lower bound for tree edit distance computation

$$\text{Max}(\text{ed}(\text{pre}(T1), \text{pre}(T2)), \text{ed}(\text{post}(T1), \text{post}(T2))) \leq \text{TDist}(T1, T2)$$

Where:

- $\text{pre}(T)$  preorder traversal of  $T$
- $\text{post}(T)$  postorder traversal of  $T$
- $\text{ed}(s_1, s_2)$  string edit distance btw  $s_1$  and  $s_2$

### Approximate XML joins [Guha et al 2002] - 3

- Upper bound for tree edit distance computation
- Basic idea: reduce the search space when performing operations to match two trees by considering:
  - Minimum cost mapping for tree distance computation is sibling and ancestor order preserving
  - For the upper bound: mapping also preserves ancestor order for the *lowest common ancestor of pairs of nodes*. The requirement ensures that distinct subtrees of a tree  $T_1$  are mapped into distinct subtrees of a tree  $T_2$
- Upper bound: proposed algorithm to compute the distinct tree edit distance is  $O(n^2)$

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

207

### Approximate XML joins [Guha et al 2002] - 4

- Reference set to reduce the number of pairwise distance computations to evaluate approximate XML joins
- Given  $S_1$  and  $S_2$  two sets of XML document trees, let  $K \subset S_1 \cup S_2$  a chosen set called *reference set*. Let  $d_1 \in S_1$  and  $d_2 \in S_2$  be two XML documents
  - A vector for each document is constructed, consisting of the distances to the XML docs in  $K$
  - $k_1 \dots k_{|K|}$  is an arbitrary ordering of the reference set  $K$
  - $v_i$  is the vector for  $d_i$  and  $v_j$  is the vector for  $d_j$

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

208

## Approximate XML joins [Guha et al 2002] - 5

### - Cont'd

- $v_{i1} = \text{dist}(d_i, k_1)$  and  $v_{j1} = \text{dist}(d_j, k_1)$ , where  $\text{dist}$  is a generic metric (e.g. the tree distance)
- Since  $\text{dist}$  is a metric, the triangle inequality is valid and thus:

$$|v_{i1} - v_{j1}| \leq \text{dist}(d_i, d_j) \leq v_{i1} + v_{j1}$$

- Assuming we wish to have documents within a certain threshold  $Th$ , i.e.  $\text{dist}(d_i, d_j) \leq Th$  we can say:
  - If  $u_{\dagger} = \min_i v_{i1} + v_{j1} \leq Th$ , then the pair is within  $Th$
  - If  $l_{\dagger} = \max_i |v_{i1} - v_{j1}| > Th$ , then the pair cannot be within  $Th$

## Detailed description of Knowledge based techniques skip

## Intelliclean - 1

- **1. Pre-processing** Data type checks and format standardization
- **2. Processing**
  - **2.1** The conditioned records are next fed into an expert system engine together with a set of rules of the form
    - **If <condition> then <action>**
  - **2.2.** Use the Multi pass Sorted Neighbour searching method
  - **2.3** Check first DI rules and then MP rules using a basic production system to see which ones should fire based on the facts in the database, looping back to the first rule when it has finished.
- **3. Human verification and validation stage:** Human intervention to manipulate the duplicate record groups for which merge/purge rules are not defined.

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China



211

## Example of rule

```
DEFINE RULE COMPANY_RULE
INPUT RECORDS : A B
IF
  (A.currency == B.currency)           AND
  (A.telephone == B.telephone)         AND
  (A.telephone != EMPTY_STRING)       AND
  (SUBSTRING_ANY(A.code,B.code) == TRUE) AND
  (FIELDSIMILARITY(A.address,B.address) >0.85)
THEN
  DUPLICATES (A B), CERTAINTY=0.85
  UPDATE_LOGS
```

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

212

## Relevant types of rules

- **A. Duplicate identification (DI) rules:** specify the conditions and criteria for two records to be classified as duplicates. E.g. text similarity, string manipulation, complex logic for determining record equivalence can be coded into or referenced.
- **B. Merge/purge rules (MP):** specify how duplicate records are to be handled. E.g. only the record with the least number of empty fields is to be kept in a group of duplicate records, and the rest be deleted.
- [back](#)

## Transformations in Intelliclean - 1

### Type I transformations

- \* **Stemming** converts a token into its stem or root.
- \* **Soundex** converts a token into a Soundex code. Tokens that sound similar have the same code.
- \* **Abbreviation** replaces token with corresponding abbreviation (e.g., 3rd or third).

## Transformations in Intelliclean - 1

### Type I transformations

- \* **Stemming** converts a token into its stem or root.
- \* **Soundex** converts a token into a Soundex code. Tokens that sound similar have the same code.
- \* **Abbreviation** replaces token with corresponding abbreviation (e.g., 3rd or third).

## Transformations in Intelliclean - 2

### Type II transformations

- \* **Equality** compares two tokens to determine if each token contains the same characters in the same order.
- \* **Initial** computes if one token is equal to the first character of the other.
- \* **Prefix** computes if one token is equal to a continuous subset of the other starting at the first character.
- \* **Suffix** computes if one token is equal to a continuous subset of the other starting at the last character.

**\*Type II transformations - cont'd**

**Substring** computes if one token is equal to a continuous subset of the other, but does not include the first or last character.

\* **Abbreviation** computes if one token is equal to a subset of the other (e.g., Blvd, Boulevard).

\* **Acronym** computes if all characters of one token are initial letters of all tokens from the other object (e.g., CPK, California Pizza Kitchen).

Complexity in Intelliclean

- The initial complexity
- $O(RF^P)$  where
  - $R$  is the number of rules
  - $P$  is the average number of patterns in the condition part of the rules
  - $F$  is the number of facts in the knowledge base,
- can be dropped to  $O(RFP)$  by remembering what has already been matched from cycle to cycle.

## Intelliclean improvement of Multi pass SN

- Transitive closure tends to increase false positive errors.
- In Intelliclean a certainty factor is applied to each duplicate identification rule, that represents **expert confidence in rule's effectiveness in discovering duplicates**.
- During transitive closure, certainty factors are composed and closure is applied only if the resulting factor is greater than a user defined threshold.

## Active Atlas - 1

- Main idea: extends the concept of distance between values of attributes into the concept of set of **transformation functions**, e.g. abbreviation, acronym, etc.
- The final goal is to find the more accurate **mapping rules**, e.g. **for restaurants addresses**
- **if Name > threshold1 and Street > threshold2 → mapped**
- Two types of knowledge considered:
  - Importance of the different attributes for deciding a mapping
  - **e.g. Name + Telephone number vs Name + Address**
  - Transformations that are relevant for the application (→)

## Transformations in Active Atlas - 1

### Type I transformations (simple)

- \* **Stemming** converts a token into its stem or root.
- \* **Soundex** converts a token into a Soundex code. Tokens that sound similar have the same code.
- \* **Abbreviation** replaces token with corresponding abbreviation (e.g., third → 3rd )

## Transformations in Active Atlas - 1

### Type II transformations (more comp.expensive)

- \* **Equality** compares two tokens to determine if each token contains the same characters in the same order.
- \* **Initial** computes if one token is equal to the first character of the other.
- \* **Prefix** computes if one token is equal to a continuous subset of the other starting at the first character.
- \* **Suffix** computes if one token is equal to a continuous subset of the other starting at the last character.

## Transformations in Active Atlas - 2

### \*Type II transformations - cont'd

**Substring** computes if one token is equal to a continuous subset of the other, but does not include the first or last character.

- **Abbreviation** computes if one token is equal to a subset of the other

- (e.g., Blvd, Boulevard)

- \* **Acronym** computes if all characters of one token are initial letters of all tokens from the other object

- (e.g., CPK, California Pizza Kitchen)

## Active Atlas - Steps

- 1. Compute for every object pair the **similarity scores** based on application of **transformations**, obtaining **candidate mappings**
  - 1.1 First related objects + T1 transformations
  - 1.2 Then T2 transformations to unmatched obj.
  - 1.3 Compute related obj + transf. + frequencies
- 2. Learn the most relevant **mapping rules**, proposing to the **user** as **training examples** the most informative candidate mappings.
  - 2.1 Use decision tree learning
  - 2.2 Involve users through active learning
- 3. Apply mapping rules to candidate mappings to determine the set of mapped objects

## Clue based - process

- 1. Design step
- 1.1 Knowledge extraction
- Find relevant **clues**, i.e relevant properties for records to be matched
- 1.2 Training
- Given a training set, apply clues to the training set and assign a set of weights that maximizes accuracy
- If accuracy is not satisfactory, go to 1.1. knowledge extraction, otherwise proceed to 2.
- 2. Production step
- Use the trained model for matching

Example of clue: predict a match if a senior appears to have one residence in Florida or Arizona and another in the Midwest or Northeast regions of the country

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

Detailed description of  
probabilistic techniques  
skip

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

226

## Fellegi and Sunter RL family of models - 1

For the two data sources A and B the set of pairs  $A \times B$  is mapped in two sets, **M (matched)** where  $a = b$  and **U (unmatched)** otherwise.

Having in mind that it is always better to classify a record pair as a possible match than to falsely decide on its matching status, a third set  $P$ , called **possible matched**, is defined. In the case that a record pair is assigned to  $P$ , a domain expert should manually examine this pair.

The problem, then, is **to determine in which set each record pair belongs to, in presence of errors.**

A **comparison function C** compares values of pairs of records, and produces a comparison vector.

## Fellegi and Sunter RL family of models - 2

We may then assign a weight for each component of the record pair, calculating the composite weight and **comparing the weight against two thresholds T1 and T2, and assigning M, U, or P (not assigned) according to the relative value of the composite weight (CW) wrt the two thresholds. In other words:**

- If  $CW > T2$  then designate pair as a match **M**
- If  $CW < T1$  then designate pair as an unmatched **U**
- Otherwise designate as not assigned.

### Fellegi and Sunter RL family of models - 3

The thresholds can be estimated by a priori error bounds on false matches and false non matches.

This is the reason why the model is called **error-based**

Fellegi and Sunter proved that this decision procedure is optimal, in the sense that, fixed the rates of false matches and false non matches, the clerical review region is minimized.

### Fellegi and Sunter RL family of models - 4

The procedure has been investigated under different assumptions, and for different comparison functions.

This is the reason of the name **family of models**.

E.g. When the comparison function is a simple agree/disagree pattern for three variables, if the variables are independent ( $\rightarrow$  **independence assumption**), we may express the probabilities that the record match/not match as product of elementary probabilities, and solve the system of equations in closed form.

## More on independence assumption

- **Independence assumption** - To make computation more tractable, FS made a conditional independence assumption that corresponds to the naïve Bayesian assumption for BN:
- $P(\text{agree on first name Robert, agree on last name Smith}/C1) =$
- $P(\text{agree on first name George}/C1) P(\text{agree on last name Smith}/C1)$ .
- Under this assumption FS provided a method for computing the probabilities in the 3 variable case. Winkler developed a variant of the EM algorithm for a more general case.
- Nigam 2000 has shown that classification decision rules based on Naïve BN (independence assumption) work well in practice.
- **Dependence assumption** - Varying authors have observed that the computed probabilities do not even remotely correspond to the true underlying probabilities.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

231

## Expectation maximization model

- An **expectation-maximization (EM) algorithm** is an algorithm for finding maximum likelihood estimates of parameters (in our case, probabilities of match and not match) in probabilistic models, where the model depends on unobserved (latent) variables.
- EM alternates between performing an **expectation (E) step**, which computes the expected value of the latent variables, and a **maximization (M) step**, which computes the maximum likelihood estimates of the parameters given the data and setting the latent variables to their expectation.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

232

[Nigam et al 2000]

- Uses the EM algorithm to obtain parameters (probabilities) for the classification decision rules (matching vs non matching)
- This is done combining a moderate amount of training (labeled) data with proper amounts of additional unlabelled data, improving classification decision rules wrt algorithms that use uniquely unlabelled data ~~are improved~~ and algorithms that use the moderate amounts of labelled data alone.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

233

[Winkler 1993] modification of EM method

Proposes a 3-class EM algorithm for household record linkage problems. The three classes considered concern persons:

- a. Within households agreeing on name characteristics
- b. Within households not agreeing on name characteristics
- c. Outside the same household.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

234

### Cost based model

- In the family of FS based models the thresholds are estimated by minimizing the probability of the error of making an incorrect decision for the matching status of a record pair.
- In practice, the minimization of the probability of the error is not the best criterion to use in designing a decision rule as different wrong decisions may have different consequences (costs)
- For example, the incorrect decision to classify an unmatched record pair in the matched set may lead to an undesired action of removing one of the records, whereas the incorrect decision to classify a matched record pair as unmatched may lead to data inconsistencies.
- The cost-based probabilistic record linkage model assigns thresholds based on weights that consider variable costs.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

235

### Induction model - 1

- In supervised machine learning, a training set of patterns in which the exact class of each pattern is known a priori, is used in order to build a classification model that can be used afterwards to predict the class of each unclassified pattern.
- A supervised learning technique can be called a *classifier*, as its goal is to build a classification model.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

236

## Induction model - 2

- In the induction record linkage model the training set consists of instances of the form  $\langle c, f(c) \rangle$  where  $c$  is a comparison vector and  $f(c)$  is its corresponding matching status, i.e., (M or U).  $M$  denotes a matched record pair and  $U$  denotes an unmatched one.
- A classifier is employed to build a classification model that estimates the function  $f$  and is able to predict the matching status of each comparison vector of the whole set of record pairs.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

237

## Clustering model - 1

- Training sets are not readily available for most real-world applications.
- Unsupervised learning tries to approximate the function  $f$  without having any training instances.
- **Clustering is the only known method for unsupervised learning.** The fundamental clustering problem involves grouping together those patterns that are similar to each other.
- Clustering algorithms try to cluster these points into separate groups in the space. A specific technique, called *k-means clustering* tries to cluster the points into  $k$  clusters.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

238

## Clustering model - 2

- The model considers each comparison vector as a point in  $n$ -dimensional space, where  $n$  is the number of components in each record. A clustering algorithm, such as *k-means clustering*, is used to cluster those points into three clusters, one for each possible matching status, *matched*, *unmatched*, and *possibly matched*. After applying the clustering algorithm to the set of comparison vectors, the issue is to determine which cluster represents which matching status.
- Since a perfectly matched record pair is located with the origin  $0,0,..,0$  of the multidimensional space, and a completely unmatched pair is located in  $1,1,..,1$ , the cluster nearest to the origin is considered M, the one nearest to  $1,1,..,1$  is considered U and the third one possibly matched.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

239

## Hybrid model

- Supervised learning gives more accurate results for pattern classification than unsupervised learning. However, supervised learning relies on the presence of a training set, which is not available in practice for many applications.
- Hybrid RL model combines the advantages of both the induction and the clustering record linkage models.
- Unsupervised learning can be used to overcome this limitation by applying the unsupervised learning on a small set of patterns in order to predict the class of each unclassified pattern, i.e., a training set is generated.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

240

### 1-1 matching

- Assumption: **Many applications consist of matching two files that have few internal duplicates.**
- In these situations, it is often efficient to force 1-1 matching because many of the second and third best matches for pairs might have matching weights that are sufficiently high to necessitate clerical review and not be true matches. Forcing 1-1 matching in an efficient manner can greatly reduce clerical review in these situations.

### 1-1 matching

- Jaro provides a **linear sum assignment procedure** to force 1-1 matching. He observed that 1-1 matching via greedy algorithms in earlier record linkage systems could make a higher proportion of erroneous assignments. A greedy algorithm is one in which a record is always associated with the corresponding available record having the highest agreement weight.
- Improvement: Winkler introduced a **modified assignment algorithm** that uses 1/500 as much storage as the original algorithm and has of equivalent speed.

### Bridging file model

- A bridging file is a file that can be used in improving the linkages between two other files. Typically, a bridging file might be an administrative file that is maintained by a governmental unit.

File A				Bridging file			File B			
A11	A12	A13	.....	Name1	Addr1	Zip1	.....	B11	B12	B13
A21	A22	A23	.....	Name1	Addr2	Zip2	.....	B21	B22	B23
A31	A32	A33	.....	Name1	Addr3	Zip3	.....	B31	B32	B33

- **Assumptions: the bridging file is correct, files A and B are almost correct**
- The amount of records respectively between A and BF / B and BF is significantly smaller than records between A and B

## Comparison of techniques

## Types of comparisons

- **Theoretical vs Experimental**
- **For phase**
  - Blocking/Searching method
  - Decision model
- **Single technique evaluation vs Comparative evaluation**
- We will see only some results in the area:
- **Experimental/comparative evaluation**

## Measurement metrics - for the **searching method**

- **Reduction ratio** - measures the relative reduction in the size of the comparison space accomplished by a **searching method**
- **Pairs completeness** - a **searching method** can be evaluated based on the number of actual matched record pairs contained in its reduced comparison space.
- **Pairs Completeness** = the matched record pairs in the reduced comparison space / the total number of matched record pairs in the entire comparison space.

## Measurement metrics - for the **decision model**

- **Accuracy** - tests how accurate a **decision model** is. The accuracy of a decision model is defined to be the percentage of the correctly classified record pairs.
- **Completeness** tests how complete the **decision model** is when considering the matched record pairs.
- **Completeness** = the matched record pairs detected by the d. model / the total number of matched record pairs known in the data.

## Results of comparison of **string** comparators

- For **population** files
  - 1. Weighted distance
- **General** (random)
- Rating:
  - 1. Winkler's variant of Jaro
  - 2. Jaro's algorithm
  - 3. Edit distance
  - 4. Bigram
  - 5. Trigram

Results of comparisons of probabilistic decision models based on the size of the training set

	0 - 20%	> 20%
Accuracy	Better: Induction Clustering Hybrid Worst: Probabilistic	The same
Completeness	1. Induction 2. Clustering 3. Probabilistic	The same
Percentage of record pairs predicted as possible match	Please correct! 1. Probabilistic 2. Clustering and hybrid	Please correct! 1. Clustering and hybrid 2. Probabilistic

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

249

Comparison of blocking methods

- In the **blocking method**, the *pairs completeness* value decreases and the *reduction ratio* increases as the value of the block key length increases
- In the **sorted neighbour method**, the *pairs completeness* value increases and the *reduction ratio* decreases as the value of the window size increases
- Blocking searching method with block key length of the same length of the sorting key has the worst performance, and the other methods have approximately the same performance

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

250

## Comparison of decision models

- The machine learning models (e.g. induction) outperform the probabilistic record linkage models (e.g. Fellegi and Sunter) concerning both the accuracy and the completeness metrics
- The probabilistic record linkage models have lower percentage of possibly matched record pairs.

## Comparisons among Delphi and Sorted Neighbour

- Delphi is compared against:
  - Sorted Neighbour with edit distance
  - Delphi-global, variant of Delphi that uses global fixed thresholds (so ignores dynamic thresholding)
  - Delphi-stripped, variant of Delphi that uses only textual similarity and ignores co-occurrence

## Comparisons among Delphi and Sorted Neighbour: results - 1

- **False Positive Explosion** - lowering thresholds drastically increases false positive percentages for edit distance.
- **Reduction in False Positive Percentages** Because Delphi and Delphi-global have significantly lower false positive percentages, hierarchies and co-occurrence information together significantly reduce false positive percentages.

## Comparisons among Delphi and Sorted Neighbour: results - 2

- **Reduction in False Negative Percentages** - Delphi has the lowest false negative percentages. Therefore, co-occurrence information is useful in reducing false negatives as well.
- **Delphi-Stripped is better than Delphi-Global.** Dynamic thresholding helps reduce false negative percentages.
- However, its impact on false positive reduction seems unpredictable.

## Comparison among Active Atlas and Passive Atlas

- The active learner outperform the passive learner because it is able to choose examples that give it the most information about the domain and guide the learning process.
- The passive learner chooses examples in random manner, independent of the actual data.

## Techniques for data integration (19)

## Techniques for data integration: table of contents

- Instance conflict resolution and result merging
- Source selection
- Composition of quality dimensions

## Conflict resolution and result merging: list of proposals

- [Otzu](#) et al 1999
  - Yan, L. Ozsu, T. (1999). Conflict tolerant queries in AURORA, Proceedings of the Fourth International Conference on Cooperative Information Systems (CoopIS'99), Edinburgh, Scotland, UK.
- [Fan](#) et al 2001
  - Fan, K., Lu, H., Madnick, S., Cheung, D.(2001): Discovering and reconciling value conflicts for numerical data integration, *Information Systems* 26.
- [Naumann](#) et al 2002
  - Felix Naumann, Matthias Häussler: Declarative Data Merging with Conflict Resolution. IQ 2002, Boston, MA, USA, 2002.
- [Sattler](#) et al 2003
  - K. Sattler, S. Conrad, and G. Saake, Interactive Example-Driven Integration and Reconciliation for Accessing Database Integration, *Information systems* 28 (2003)
- [Motro](#) et al 2004
  - Amihai Motro, Philip Anokhin, Aybar C. Acar Utility-based Resolution of Data Inconsistencies IQIS 2004, Paris, France.

Otzu et al 1999: Conflict Tolerant Queries in AURORA

- Problem: Solving instance-level conflicts in a data integration setting
- Solution: Enhance the semantics of query answering by admitting some levels of conflicts
- Pros:
  - dynamic conflict resolution i.e. at query processing time
  - declarative specification, with optimized support
- Cons:
  - higher level of conflicts tolerance, i.e. **high confidence** (all sources agree), **random evidence** (random selection in case of conflicts), **possible at all** (a result meeting selection condition is returned, without conflict detection)
  - need for conflict resolution function (e.g. **AVG**)

Person ID	Age	Salary
001	32	60.000
001	33	50.000
002	34	30.000
002	35	20.000
003	30	40.000
003	31	34.000
003	36	30.000

[Fan et al 2001]: Discovering and reconciling value conflicts for numerical data integration

- Problem: Conflict resolution, where conflicts can be of two types- Context dependent conflicts and context independent conflicts.
  - Context dependent conflicts are related to heterogeneity of data sources (→ more predictable → can be solved systemat.)
  - Context independent conflicts are due to errors
- Solution: Proposal of a methodology and a technique to detect and solve context dependent conflicts
  - Focus on numerical data values
- Pros: - usage of established techniques coming from different areas, e.g. statistics/regression and machine learning/data mining, for conflict detection and resolution
  - proposal of conversion functions to solve conflicts (much work is on conflict detection)
- Cons: - large number of conversion rules if many sources are involved: need for optimized strategy

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

261

[Naumann 2002]: declarative data merging with conflicts resolution

- Problem: Merging data in an integrated database or in a query against multiple sources
- Solution:
  - proposal of a set of resolution functions (e.g. MAX, MIN, RANDOM; GROUP, AVG, LONGEST)
  - declarative merging of relational data sources by common queries through 1. grouping /2. aggregating and partitioning (divide et impera)/3. nested joining (one at a time, hierarchical)
- Pros: the proposal considers the usage of SQL, thus exploiting capabilities of existing database systems, rather than proposing proprietary solutions
- Cons: limited applicability of conflict resolution functions that often need to be domain-specific

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

262

[Sattler et al 2003]: Interactive example-driven integration and reconciliation for accessing database federation

- Problem: Schema and Data integration to be dealt with simultaneously
- Solution: Proposal of an integration approach based on a multidatabase query language that provides mechanisms for conflict resolution
- Proposal of a **process (Data integration: 1. represent. confl., 2. Key eq confl. 3. Attribute confl.)**, and identification of instance-level conflict resolution within such a process

[Sattler et al 2003]: Interactive example-driven integration and reconciliation for accessing database federation

- Pros:
  - Schema and instance level conflicts dealt together in the integration process
- Cons
  - Too much vague in the description of conflict detection
  - Too much general wrt conflict resolution e.g. only representation conflicts are solved by means of mapping tables or conversion functions.

### [Motro et al 2004]: Utility-based Resolution of Data Inconsistencies

- Problem: query answering in a multiple information sources environment in presence of instance level conflicts
- Solution: Proposal of a utility function to fuse/merge numerical data values to be returned to users, when conflicts occur.
- Pros:
  - Proposal of an 'optimal' fusion that maximizes utility
- Cons:
  - Estimation of fusion coefficients to be made by an expert: not an easy task

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

265

### Source Selection: list of proposals

- Naumann et al 1999
  - Felix Naumann, Ulf Leser, Johann Christoph Freytag: Quality-driven Integration of Heterogeneous Information Systems. VLDB 1999, Edinburgh, Scotland, 1999
- Mihaila et al 2000
  - Mihaila, G., Raschid, L., Vidal, M. Using quality of data metadata for source selection and ranking, in: Proceedings of the *Third International Workshop on the Web and Databases (WebDB'00)*, Dallas, Texas, USA 2000.
- De Giacomo et al 2004
  - De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tackling Inconsistencies in Data Integration through Source Preferences, IQIS 2004, Paris France 2004,

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

266

[Mihaila et al 2000]: using quality of data metadata for source selection and ranking

- Define a formalism and a methodology for
  - identifying relevant sources of a query in a distributed environment
  - deciding which source/source is/are best suited for a given task
  - rank these sources
- source content quality descriptor
- a query language to select and rank sources

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

267

[[Naumann](#) et al 1999]: quality-driven integration of heterogeneous information systems - 1

- Provides a model and a process for selecting the "best" sources
- The method extends to DQ known methods for best plan selection in a multidatabase
- Three classes of criteria:
  - Source specific
  - Query Correspondence Assertions (QCA)
  - Attribute specific

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

268

[Naumann et al 1999]: quality-driven integration of heterogeneous information systems - 2

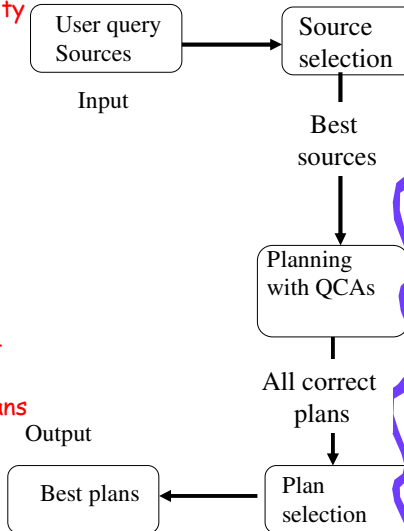
Class	Criterion	Brief explanation
Source specific	Ease of understanding Reputation Reliability Timeliness	User grade based on data presentation User grade based on personal prefs Accuracy of the experimental method Update frequency
Query Corresp. Assertion specific	Availability Price Representational consistency	% of time the data are accessible Monetary price of a query Per query time consumption of the wrapper
Attribute /User query specific	Completeness Amount	Fullness of the relation in each attribute Number of attributes in the response which were not specified in the user query.

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

269

[Naumann et al 1999]: quality-driven integration of heterogeneous information systems - 3

1. **Source selection** - Filtering out low quality sources based on **Source-specific criteria**.
2. **Plan creation** - Finds all combinations of local schemas that obtain semantically correct answers.
3. **Plan selection** - Ranks the plans and ultimately to restrict plan execution to some plans that meet certain cost- or quality-constraints. This is performed in three steps:
  - 3.a **QCA quality evaluation**, in terms of non-source-specific criteria.
  - 3.b **Plan Quality** - The total IQ score of plans is calculated. A weighting vector is adopted, specified by the user/s.
  - 3.c **Plan Ranking**- the previous phase associates an IQ vector to each plan. In this phase vectors are scaled and ranked.



23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

270

[De Giacomo et al 2004]: tackling inconsistencies in data integration through source preferences

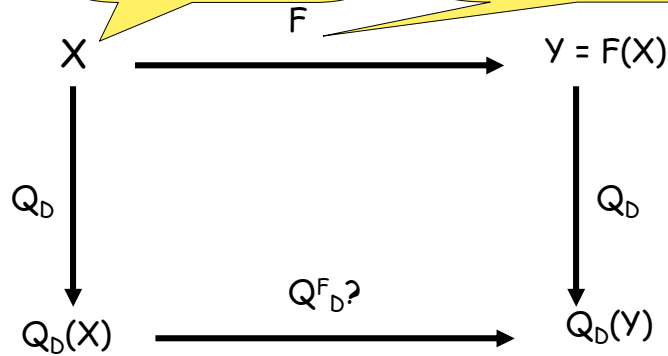
- A declarative approach to deal with inconsistencies in a data integration setting as opposed to a procedural one
- **Inconsistency** arises when:
  - A. there are integrity constraints on the global schema and
  - B. sources can contain data that, combined with other sources can contradict such constraints
- Proposal of a semantics for a **LAV (local as view) integration setting** in which preferences (e.g.  $QD_{s1} > QD_{s2}$ ) can be specified over whole sources

### Comparison of proposals

	Granularity of Selection	Dealing with Inconsistencies	Heterogeneity of Sources (GAV/LAV setting)
<b>Naumann et al 1999</b>	Source Attribute Query Corresp. Assertions	No	Yes (declared)
<b>Mihaila et al 2000</b>	Source	No	No
<b>De Giacomo et al 2004</b>	Source	Yes	Yes

# Composition of Quality Dimensions

Composition e.g. two tables X1 and X2 men e.g. Union (X1, X2)



**X= Data**  
 **$Q_D$ = function that calculates the value of a dimension D for a given data value**

### Comparison of Proposals - 1

	Model	Quality Dimensions	Goal	Algebra Operators	Criteria
<b>Motro 1998</b>	Relational model with OWA (implicit)	Soundness Completeness	-Estimating the quality of query results -Information sources selection	-Cartesian Product -Selection -Projection	-Soundness -Completeness
<b>Wang 2001</b>	Relational model with OWA (implicit)	Accuracy	-Estimating the quality of query results	-Selection -Projection -Cartesian Product	-Deterministic and Probabilistic Tuple Accuracy -Relation Accuracy -Probabilistic Attribute Accuracy -Null Relational Accuracy
<b>Parssian 2002</b>	Relational model with OWA (implicit)	Accuracy Completeness	-Estimating the quality of query results	-Selection -Projection -Cartesian Product -Join	-Relation Accuracy -Relation Inaccuracy -Relation Mismatch -Relation Incompleteness

Open World Assumption

275

### Comparison of Proposals - 2

	Model	Quality Dimensions	Goal	Algebra Operators	Criteria
<b>Naumann 2004</b>	Set of data sources With OWA and CWA (implicit)	Completeness	-Sources composition -Information sources selection	-Join Merge -Full Outerjoin Merge -Left Outerjoin Merge -Right Outerjoin Merge	-Coverage -Attribute Density -Source Density -Query Dependant Density -Completeness
<b>Scannapieco 2004</b>	Relational model with OWA and CWA	Completeness	-Relational sources composition -Relational sources selection	-Union -Intersection -Cartesian Product	-Relation Completeness (OWA without Null) -Value Completeness (with Null) -Attribute completeness (With Null, OWA+CWA, strong+weak) -Tuple completeness (With Null, OWA+CWA, strong+weak) -Relation Completeness (With Null, WA+CWA, strong+weak)

Closed World assumption

276

## Other DQ activities: Profiling and Data Editing skip

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

277

## Techniques for profiling

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

278

## Resemblance of sets

**Resemblance of two sets**, i.e the intersection ratio the union cardinalities, can be calculated from the **signatures** of the sets.

A signature is a function of a hash function that maps uniformly and randomly elements of the sets on integers.

## Finding composite fields and heterogeneous tables

**Sketches** are sample projections of vectors.

The **q-gram sketch** represents the multiset of q-grams of a field, so it is the more discriminating summary of the distribution of field values. They can be used for:

- **Finding composite fields**, by finding candidates with might combine to form a composite field.
- **Finding heterogeneous tables**, where new attributes have been added in time that refer to new heterogeneous entities, resulting in joins that fit only with a portion of a fact table, and so lead to a reduced similarity.

## Transformations - 1

**Fields transformations** - Similar but not equal fields (e.g. with extra text appended), can be discovered through substring resemblance among all possible substrings of a file of length  $k$  (called  $k$ -grams); such resemblance can be measured with the  $k$ -gram signature.

- Two fields with **low similarity** but **large  $k$ -gram resemblance** are related by a small transformation.

## Transformations - 2

- Transformations such as

**Lastname Firstname  $\rightarrow$  Firstname Lastname**

can be discovered with the following strategy:

1. Find textually related fields, computing min hash samples of the  $q$ -grams of the filed values, rather than fields values. Textually similar fields will have a largely overlapping set of  $q$ -grams and therefore a high resemblance.
2. Perform a quick visual inspection, and determine which fields are potentially related and what transformations can be applied.

## Missing values - 1

- Techniques for **imputting missing values** are to be used with caution when values are significant as such, and are not used only through aggregations.
- If the assumption is valid that missing values have the same distribution of non missing, we can use:
  - A. the mean,
  - B. the median, or else
  - c. simulate a distribution using non missing values.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

283

## Missing values - 2

- A more complex method concerns **impute multiple values**, of different attributes.
- Assumption: attributes are missing monotonically, that is  $Y_i$  is not missing only if  $Y_{i-1} \dots Y_1$  are not missing. In this case we may iteratively perform a **regression** filling initially  $Y_1$  and then all other attributes.

$Y_1 \rightarrow Y_2 \rightarrow Y_3$


23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

284

## Missing values - 3

- Another method is Markov Chain Monte Carlo, in which data is assumed to have a multivariate normal distribution. In this case missing values are estimated by
  - a. building the conditional distribution of missing values
  - b. computing the parameters of a multivariate normal distribution using the filled in sample.
- Steps a and b are repeated until the estimates are stable.
- **The method is expensive.**

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

285

## Incomplete data

Can be:

- **truncated** when selective observations (data) are dropped.
- **censored**, when the time interval of observation is truncated, and so in time series interruptions appear, or else overflow produces censored values.
- **Censoring** can be detected with the help of histograms or frequency distribution, where a spike reveal a potential censoring (due to use of a default value). Metadata and context knowledge are crucial for explaining the censoring type.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

286

## Outliers

- In the list 3, 4, 5, 3, 56, 4, 5, 2
- **56** is a potential outlier, i.e. an observation that is suspicious because it is not in line with the rest of the data.
- Outliers can be also explainable true data.
- Outliers can be detected by the departure of data points from what we expect them to be,
- where we have to explain **departure** and **expected**.

## Discovering outliers - 1

- 1. **Control charts** - Compute the distribution of a variable and looking at data over acceptable bounds.
- For pairs of variables, a bivariate control chart looks for outliers based on interrelationships between attributes.
- 2. **Model based** - Investigate the interrelationship between attributes through models such as **linear regression, logistic regression, and others**.
- **Residuals** in the regression are potential outliers.

## Discovering outliers - 2

- 3. **Geometric outliers** - data points on the periphery of the data set, where the periphery is fixed through **geometric methods** such as **convex hulls**.
- 4. **Distribution outliers** - Potential outliers are the points that are in regions of low density: since those points are isolated, there is a good chance that they are outliers.
- They can be found by computing for every point the distance wrt all other points, and select those with  $d > d^*$ .
- **Caution: the outliers could be clustered due to defaults.**

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

289

## Discovering outliers - 3

- 5. **Time series outliers** - Those methods differ from previous ones, since close data tend to be correlated in time. An empirical method is the following:
  - A. partition the attribute space into sections, with a partitioning strategy.
  - B. Treat each class as a state: a given time series is a trajectory of the states, that can be characterized with transition probabilities, and ranked by their likelihood.
  - C. Low likelihood transitions are potential outliers.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

290

## Discovering outliers - 4 Separation of outliers from glitches

- Potential outliers representing abnormal but legitimate behaviour have to be separated from glitches. This can be done with statistical methods, of two types:
  - **Relative deviation**, which represents the movement of a data point relative to another data point. This is **more robust**, since it is difficult to change
  - **Within deviation**, movement wrt its own expected behaviour, that can be defined in several ways. This is **more sensitive** to minor changes, and it is better for capturing long term changes.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

291

## Discovering outliers - 5 Separation of outliers from glitches

- Heuristics for distinguishing between glitches or legitimate values outliers
  - Genuine changes are more persistent in times
  - Data glitches tend to appear randomly without any structure, while glitches with a reason can be "rationalized".
  - The within deviation of a data point can be used to separate differences within structure as opposed to random aberrations.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

292

## Techniques for data editing

### Hints on data editing - 1

- The methods of implementing edit rules essentially have to check the logical consistency of a set of edit rules
- [Fellegi Holt 1976] provide a theoretical model for editing. They had three goals:
  - (1) **Error localization problem** The data in each record should be made to satisfy all edits by changing the fewest possible items of data.
  - (2) Imputation rules should be derived automatically from edit rules.
  - (3) When imputation is necessary, it is desirable to maintain the marginal and joint frequency distributions of variables.

## Hints on data editing - 2

- **Implicit edits**, i.e. edits that can be logically derived from explicitly defined edits, are needed for solving error localization.
- Implicit edits contain information about edits that do not fail initially for a record but may fail as values in fields associated with failing edits are changed.
- The error localization problem is NP-complete, so reducing computation is the most important aspect.

## Hints on data editing - 2

- If implicit edits are not available, then the **Error Localizaton problem** can be solved by direct integer programming methods such as branch-and-bound that are much slower.
- Winkler introduced a greedy heuristic that is more than 100 times as fast as branch-and-bound for error localization.

## 5. Tools skip

### Tools - table of contents

- We concentrate on research tools/prototypes!
- **Names and main features**
- Detailed description
- Comparison of (some) tools

## Names and main features

Reference	Name	Main features
[Raman et al 2001]	<b>Potter's wheel</b>	Tightly integrates transformations and discrepancy/anomaly detection
[Caruso et al 2000]	<b>Telcordia's tool</b>	Record linkage tool parametric wrt to distance functions and matching functions
[Galhardas et al 2001]	<b>Ajax</b>	Declarative language based on five logical transformation <u>operators</u> <u>Separates a logical level and a physical level</u>
[Vassiliadis 2001]	<b>Artkos</b>	Covering all aspects of ETL processes (architecture, activities, quality management) <u>with a unique metamodel</u>
[Elfeky et al 2002]	<b>Tailor</b>	Framework for existing and new techniques analysis and comparison
[Buechi et al 2003]	<b>ClueMaker</b>	Uses clues that allows rich expression of the semantics of data
[Low et al 2001]	<b>Intelliclean</b>	See section on techniques

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

299

## Potter's wheel

- Pros: Highly interactive and user friendly
- Users **specify** or **undo** transformations through a spreadsheet user friendly interface
- Transformation allowed (**add, drop, merge, split, divide, etc.**) are specified
  - a. with a graphical interface and
  - b. through examples.
- In case b, Potter's Wheel automatically infers the structure.
- Values that do not match the inferred structure are flagged as errors, and the user decides whether to modify the transformation or to proceed and clean the data.

23rd International Conference on Conceptual Modeling (ER 2004), Shangai, China

300

## Telcordia's tool

- Development of a training set
- A pre-processing step, with elimination of stop words and standardization
- Parametric tool with customizable parameters being **distance measures** and **descriptions of attributes**.
- Matching rules can be generated, tested and ranked with machine learning and statistical techniques
- High effectiveness documented:
  - Improved matching percentage by 30% over off-the-shelf tools.
  - Improved identification of likely duplicates by a factor of 2
  - Capability of classifying the likely causes of duplication.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

301

## Ajax - operators

- **Mapping** - Expresses generic one to many database mappings
- **View** - Each tuple in the output relation results from some combination of tuples taken from the input relation.
- **Matching** - computes an approximate join between two relations, through a distance value for each pair of tuples.
- **Clustering** - Returns an output relation that groups the elements of the input relation into a set of clusters.
- **Merging** - Partitions the input relations according to some grouping attributes and collapses each partition into a single tuple.
- The five operators can be composed to express all the data transformations for data cleaning found in the research literature.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

302

## Ajax - Levels of the architecture

- Two levels defined:
  - The **logical level** uses a graph of transformations
  - The **physical level** specifies optimized algorithms that can be selected to implement the transformations.
  - E.g. for the matching operator **three algorithms** can be used:
    - "Naïve" Nested loop
    - Multi pass Neighbourhood and a variant
    - Neighbourhood Join

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

303

## Artkos

1. Clear distinction between:
  - a metamodel layer, a metadata layer, an instance layer.
  - conceptual, logical, physical perspectives
3. Management both of the static part of a DW environment, and the dynamic part, i.e. the processes.
4. Representation and management of quality goals, quality dimensions, quality factors.
- A set of improvements can be proposed, and expressed through scenarios, that can be described with three different languages.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

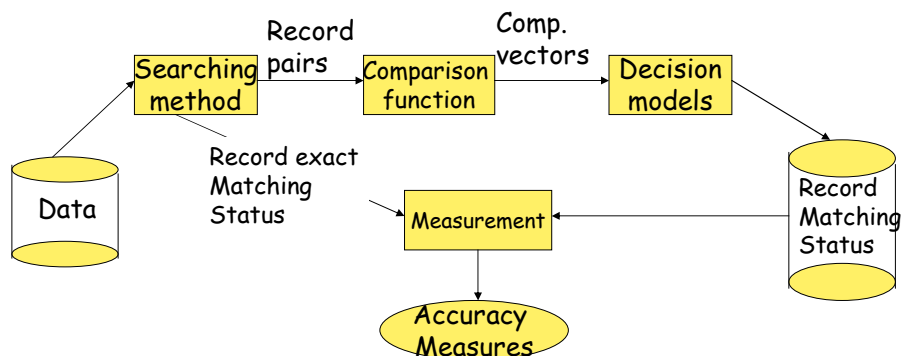
304

## Modules of Artkos

- An activity is described in terms of input tables, error types, policy, output table, quality factors.
- Several primitives can be used in building activities and design processes:
  - **Cleaning primitives:** primary key violation, reference v., Null value check, domain mismatch
  - **Transformation primitives:** transform data according to patterns user defined or else built in.

## Tailor

- Tailor is a framework for comparing techniques. So the emphasis is on the completeness and flexibility of the architecture



## ClueMaker - claimed advantages Greater ...

- **Efficiency**, in terms of time of execution and code optimization
- **Usability**, due to the interface
- **Reliability** and **Correctness** of the code
- **Productivity** - E.g. a clue set with 200 clues for a very complex schema consisting of 60 fields in 10 node types takes two to three person weeks.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

307

## Comparison

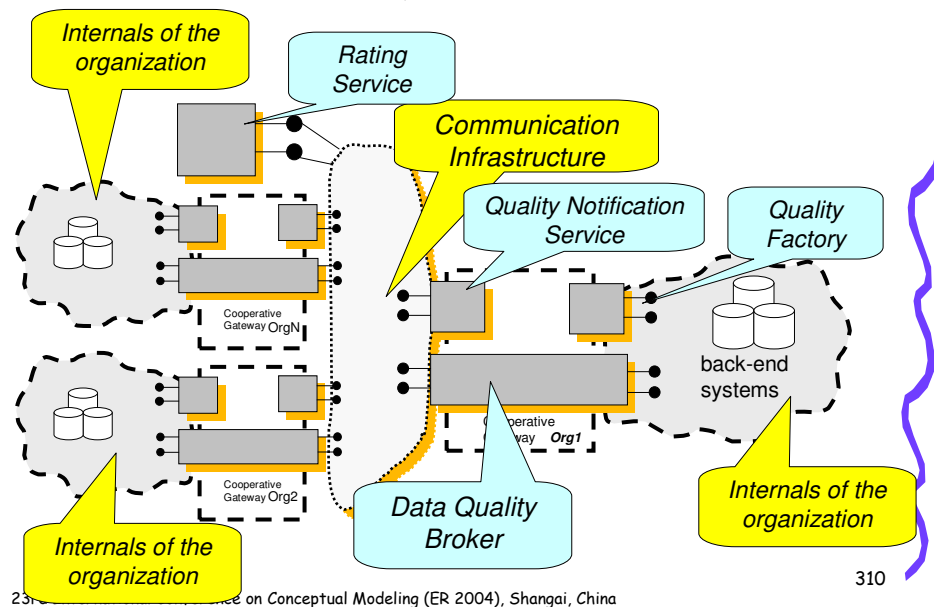
Tool/ characteristics	Potter's wheel	Ajax	Intelliclean
Optimization	Fixed algorithm.	Multiple alg.	Fixed strategy
Anomaly detection	Through visual inspection	No	By rules in knowledge base
Anomaly correction	Yes	Not documented, may be	Yes
Interactivity	Users apply transform. and see the results	Exceptions handled through user interact.	Users can modify merge purge groups
Undo facility	Allowed	Not documented	Not documented Not allowed
Friendliness	High	Low	Not addressed
Supported Operators	Add/ drop/ merge/ split/ divide/ select/ fold/format	Mapping/ View/ Matching/ Clustering/ Merging	Duplicate identification/ merge purge
Scalability	High	High	Low

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

308

## 6. Frameworks and services for cooperative information systems

### [Scannapieco et al 2004]: The DaQuinCIS Platform

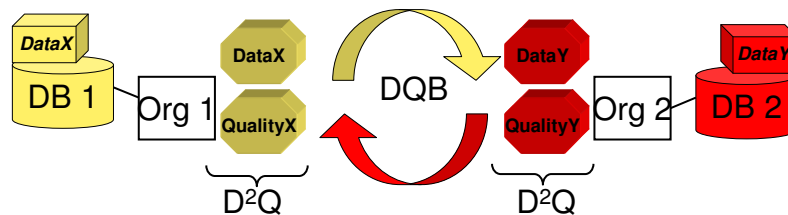


## Frameworks and services : table of contents

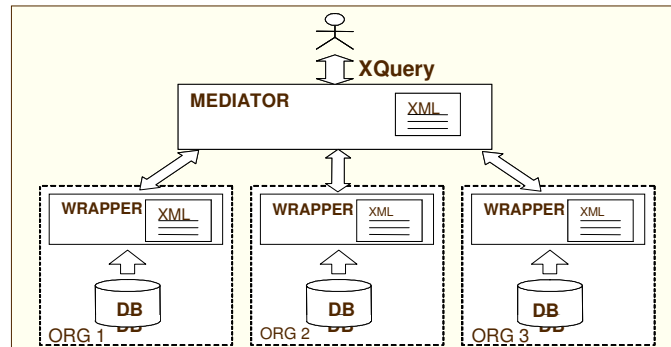
- Data Quality Broker
- Rating Service
- Data Quality Factory
- Quality Notification Service

## Data Quality Broker ...

- **Quality access** functionality: accessing data + quality exported according to the D<sup>2</sup>Q model (previously described)
- **Quality improvement** functionality: comparing different copies of the same data available in the whole CIS



## ... A Quality-driven Data Integration System

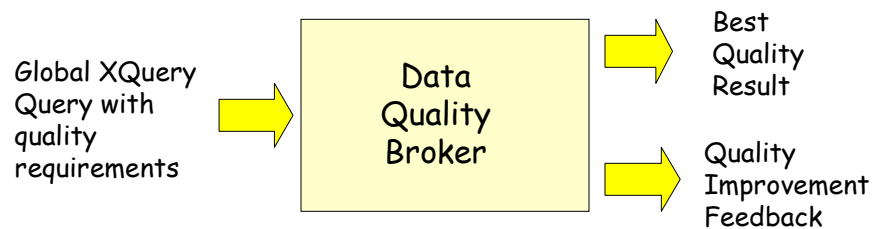


- Wrapper/Mediator Architecture
- Global and Local views expressed as XML Schemas D<sup>2</sup>Q-compliant
- Global as View (GAV) Mapping

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

313

## Semantics



- Best quality copies always available
- Quality Improvement Feature
  - The best quality result is proposed to all data sources that provided lower quality results

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

314

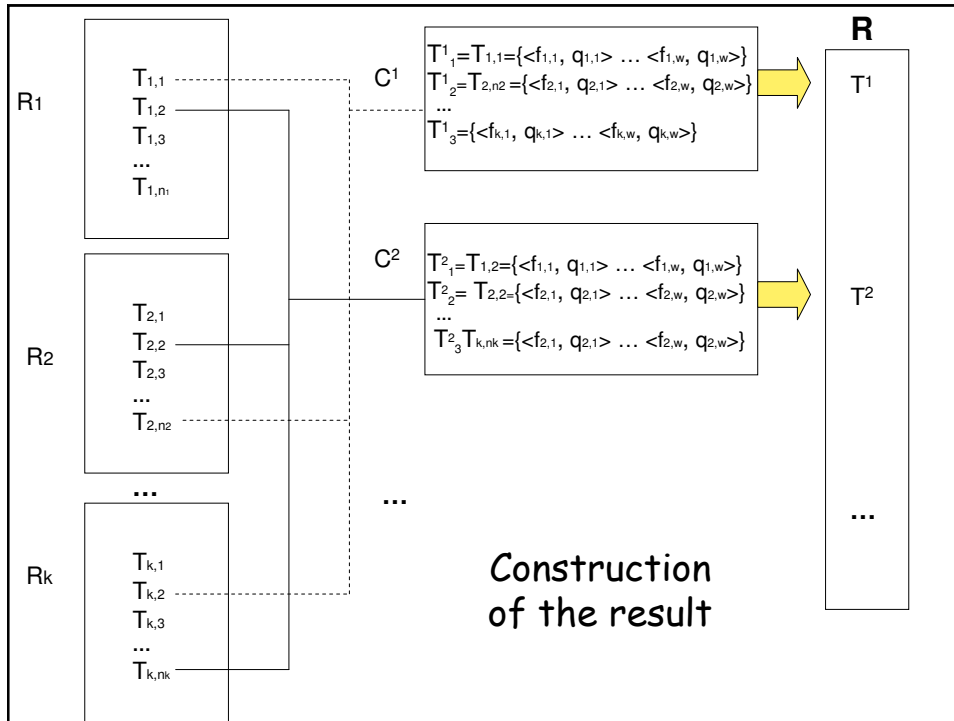
## Query Processing Steps - 1

Given a query  $Q$  on a  $D^2Q$  global schema

1.  $Q$  is unfolded according to a static mapping that retrieves all copies of same data that are available in the CIS
  - Static mapping specified through **path expressions**
  - A path expression allows to locate a concept in a schema
  - XML Schemas are  $D^2Q$  compliant

## Query Processing Steps - 2

3. The execution of local queries returns a set of results, on which a **run-time matching** is performed
  - The result of this step is the construction of set of clusters composed by tuples referring to same real world objects
  - For each cluster, a best quality representative is either selected or constructed.



### Query Processing Steps - 3

4. The result to be returned is built as follows:
  - (i) if no quality requirement is specified, a **best quality default semantics** is adopted. This means that the result is constructed by selecting/constructing the best quality value
  - (ii) if quality requirements are specified, the result is constructed by checking the satisfiability of the requirements on the whole result
5. The result is also *proposed* to organizations having provided lower quality values - **improvement functionality**

[De Santis et al 2003]: Rating Service

- **Trustworthiness** wrt a specific typology of data provided by a specific organization in the CIS
  - Trust level
    - Previous proposals: the whole organization (peer)
    - New proposal: < organization, data type >
- Adaptive "data stewardship" assignment
- Probabilistic model

Rating Service: Trust Parameter

$$R(\langle \text{Org}, D \rangle) = \frac{\sum C_{i,k,D}}{\sum n_{i,k,D}} \quad \forall \text{Org} \in O$$

# of <D, Org<sub>k</sub>> complaints sent by Org<sub>i</sub>

Rating of Org<sub>k</sub> for data type D

# of D-exchanges of Org<sub>k</sub>

## Rating Service: hypotheses

- Data exchanges modeled as a random variable  $X$  such that:
  - $X=1$  if a complaint is fired
  - $X=0$  otherwise
- $X$  binomial probability distribution with
  - $P(X=0) \ll P(X=1)$ , i.e. the number of unsuccessful data exchanges is low wrt the number of successful ones

## Rating service: trustworthiness criterion

Hence the trust parameter  $R(.)$  is a random variable with a **normal probability distribution**

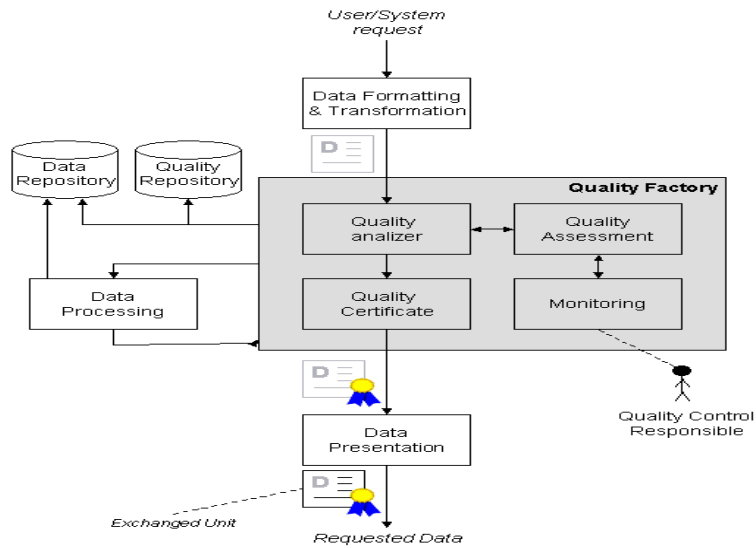
- linear combination of a large number of independent random variables with the same distribution

### Trustworthiness Criterion

***IF  $R(\langle Org^k, D \rangle) \leq m + 2 * \sigma$  THEN  $\langle Org^k, D \rangle$  TRUSTED  
ELSE  $\langle Org^k, D \rangle$  UNTRUSTED***

As there is a normal distribution, the trustworthiness criterion has at least the 95% of success probability

[Cappiello et al 2003] : Data Quality Factory - 1

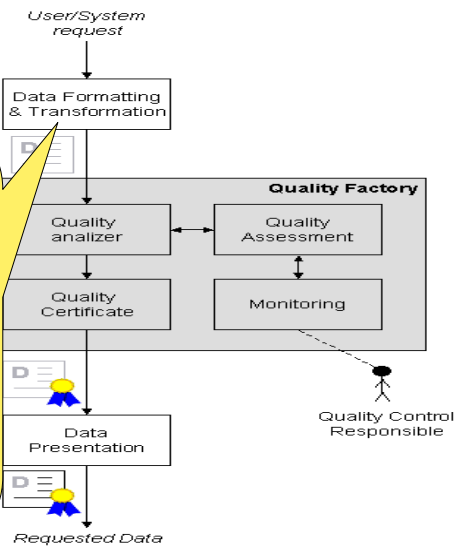


23rd In

23

[Cappiello et al 2003] : Data Quality Factory - 1

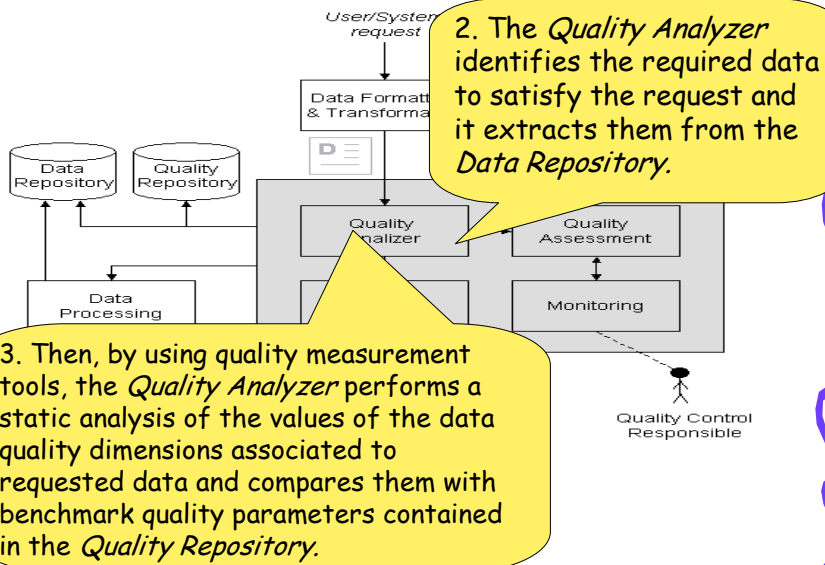
1. An external user sends a request, which is processed by the *Data Formatting & Transformation* module that translates the request into a format that can be understood by the *Quality Analyzer*



23rd In

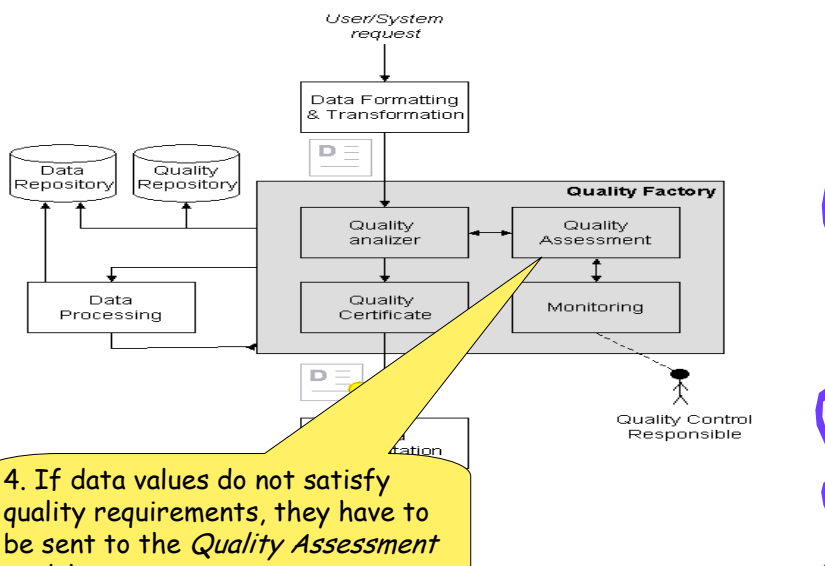
24

[Cappiello et al 2003] : Data Quality Factory - 1



3. Then, by using quality measurement tools, the *Quality Analyzer* performs a static analysis of the values of the data quality dimensions associated to requested data and compares them with benchmark quality parameters contained in the *Quality Repository*.

[Cappiello et al 2003] : Data Quality Factory - 1

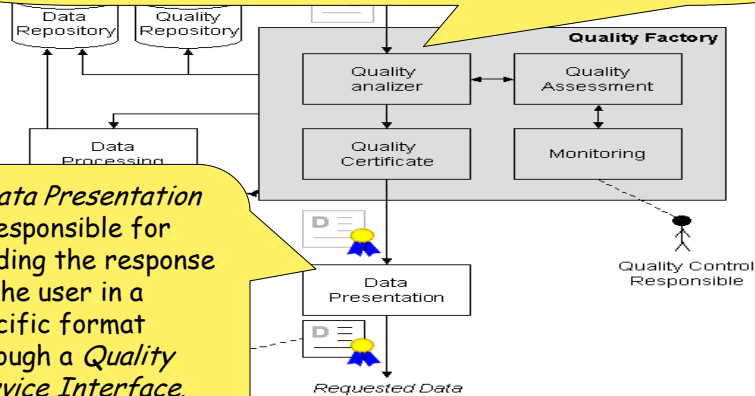


4. If data values do not satisfy quality requirements, they have to be sent to the *Quality Assessment* module

## [Cappiello et al 2003] : Data Quality Factory - 1

5. After quality improvement, data are re-examined by the *Quality Analyzer* and, if values of quality are satisfactory, a quality certificate is associated with data and is sent to the *Data Processing* module.

6. *Data Presentation* is responsible for sending the response to the user in a specific format through a *Quality Service Interface*.



23rd Int

27

## [Cappiello et al 2003] : Data Quality Factory - 2

- 1. An external user sends a request, which is processed by the *Data Formatting & Transformation* module that translates the request into a format that can be understood by the *Quality Analyzer*
- 2. The *Quality Analyzer* identifies the required data to satisfy the request and it extracts them from the *Data Repository*.
- 3. Then, by using quality measurement tools, the *Quality Analyzer* performs a static analysis of the values of the data quality dimensions associated to requested data and compares them with benchmark quality parameters contained in the *Quality Repository*.

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

328

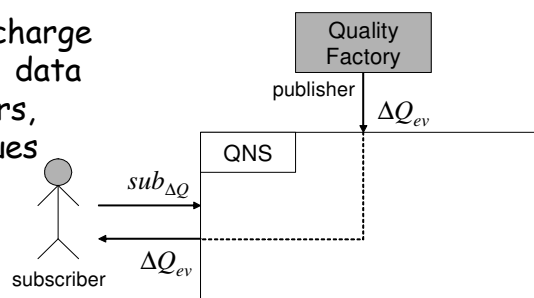
### [Cappiello et al 2003] : Data Quality Factory - 3

- 4. If data values do not satisfy quality requirements, they have to be sent to the *Quality Assessment* module
- 5. After quality improvement, data are re-examined by the *Quality Analyzer* and if values of quality are satisfactory, a quality certificate is associated with data and is sent to the *Data Processing* module.
- 6. *Data Presentation* is responsible for sending the response to the user in a specific format through a *Quality Service Interface*.

### [Scannapieco et al 2004]: Quality Notification Service

- A quality change occurs within an organization
- each time a value of a specific dimension varies.
- Upon each quality change, the Quality Factory
- resorts to the QNS the data object involved in the change along with the associated quality data.

- The QNS is thus in charge
- of notifying the new data
- to all interested users,
- according to the values
- of the quality data



## 7. Open problems

### Open Problems: table of Contents

- Examples in the data integration area
  - Quality of Web data
  - Quality of XML data
  - Quality of data in P2P systems
  - Trust and privacy issues
  - Quality of data for Personal Information Management
- Examples in the management information systems area
- A (final) list of open problems

## Examples in the data integration area

## Quality of Web data - 1

- Most of the current research work regards the assessment of the quality of web sites
- Data published on the Web are different from data published on traditional papers due to different requirements (e.g. time constraints)

## Quality of Web data - 2

- Need of dimensions to characterize quality of web data
  - Example: **completeness** of web data to indicate the level of completeness of a web-available list
- Need of techniques to measure quality of web data
  - How can sites describing the same information content be selected according to the quality of such an information content?
  - How does the frequency according to which web data are published impact on measurement techniques?

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

335

## Quality of XML data

- XML data are more and more widespread
- What is the meaning of "**quality of an XML document**"?
- First steps towards XML object identification in [Naumann et al 2004]
- Need for techniques and tools for data cleaning of XML data
  - More generally there is the need for cleaning semi-structured data, especially in emerging domains, e.g. life sciences

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

336

## Quality of data in P2P systems

- Open systems, like P2P systems, force to consider quality of data as a primary issue
- Need for a **quality-aware semantics** for P2P data integration
- The whole current formalization is ok with no inconsistencies
- The system blows up with a single inconsistency in one peer: unacceptable quality of query answering
- Need to come up with a well-defined semantics for query answering in P2P data integration that deals with inconsistencies

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

337

## Trust and privacy issues

- Trust
  - The role of trust in the context of metadata quality
  - Techniques for feedbacks between data quality and trust
  - Finer level trust models: it is not sufficient to trust an organization as a whole, but an organization should be trusted also wrt provided data typologies
- Privacy
  - How is it possible to guarantee the quality of privacy-protected data?

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

338

## Quality of data for PIM

- Personal Information Management (PIM) is an emerging research area that requires integration of personal data often stored in different devices (cellular phones, PDA, etc.)
- Need for a **device-dependant** quality characterization
- Need for ad-hoc techniques for solving conflicts on data coming from disparate sources, schemaless, incomplete, etc.

## Examples in the management information systems area - 1

- Inter-organizational methodologies
  - Some methodologies are available for dealing with data quality issues inside a single organizational context e.g. IP-MAP
  - Lacking of proposals specific to **inter-organizational** scenarios, such as e-government, e-commerce etc.
    - Need for modeling quality of data inside cooperative processes
    - Need for data quality driven patterns and solutions

Examples in the management information systems area - 2

- Trust Issues:
  - How can **data stewardship** be established in a multi-organizational context?
  - What policies are possible for **certifying quality of data** exported by each single organization?

A list of open problems - 1  
[Data Quality on the Web - [Dagstuhl Seminar](#) 2003]

- **Modeling**
  - General Formal DQ Modeling and Domain-Modeling
  - Evaluate DQ Impact from IS perspective
  - DQ Standards for databases
  - Find a usable DQ format
- **Source Dynamics and DQ assessment**
  - Trace modeling and implementation (?)
  - DQ Monitoring

A list of open problems - 2  
[Data Quality on the Web - Dagstuhl Seminar 2003]

- **Users / Usage**
  - Formal basis to specify DQ expectations
  - GUI / DQ Presentation
  - Trust and DQ
  - Web of Trust
  - Declarative expression of requirements
- **DQ Assessment**
  - Domain-specific DQ Assessment
  - Identify feasible application domains
  - Automation for hard criteria
  - Coverage and Completeness: is it feasible? (?)
  - Metrics
  - Multimedia

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

343

A list of open problems - 3  
[Data Quality on the Web - Dagstuhl Seminar 2003]

- **Query Processing / Integration**
  - Role of Integrity Constraints
  - Integrating DQ values
  - User profiles and requirements and preferences
  - DQ Algebra
    - Data-driven vs. quality-driven
  - Source Selection
- **DQ Improvement**
  - feedback

23rd International Conference on Conceptual Modeling (ER 2004), Shanghai, China

344

## 8. References

## Dimensions - 1

- Introduction
- [Redman 1996] Redman T.C.: Data Quality for the Information Age. Artech House, 1996.
- Theoretical approaches to dimensions
- [Wand Wang 1996] Wand Y., Wang R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. *Communication of the ACM*, vol. 39, no. 11, 1996.
- [Liu 2002] L. Liu, L. Chi Evolutionary Data quality, In *Proceedings of the 6th International Conference on Information Quality*, Boston, MA 2002
- Classifications of dimensions
- [Rahm Hai 2000] E. Rahm and H. Hai Do, Data Cleaning: Problems and Current Approaches, *IEEE Data Engineering Bulletin*, Special Issue on Data Cleaning 23 (2000), no. 4.
- [Wand Strong 1996] Wang R.Y., Strong D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, vol. 12, no. 4, 1996
- [Kahn et al ] Kahn B.K, Strong D.M., , Wang R.Y., Information quality benchmark: product and service performance.
- [Pipino et al 2000] L. Pipino, Y. Lee, R. Wang - Data quality assessment *CACM* 45, 2000
- [Wang et al 2001] Wang R.Y., Strong D.M., Kahn B.K., Lee Y.W. AIMQ A Methodology for information quality assessment 2001 - Information & Management, published by Elsevier Science
- [Dasu 2003] T. Dasu and T. Johnson Exploratory Data Mining and Data cleaning, Wiley 2003.
- [Jarke 1999] Quix C., Jarke M., Jeusfeld M.A., Vassiliadis P.: Architecture and Quality in Data Warehouses: an extended Repository Approach, *Information Systems*, vol.24, no.3, 1999
- [Bovee 2001] Bovee M., Srivastava R. P., Mak B.R.: A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. In *Proceedings of the 6th International Conference on Information Quality*, Boston, MA, 2001
- [Naumann 2002] Naumann F.: *Quality-Driven Query Answering for Integrated Information Systems*; LNCS 2261, 2002

## Dimensions - 2

- Comparison of dimensions
- [Scannapieco 2003] M. Scannapieco PhD Thesis, Dipartimento di Informatica e Sistemistica, Università degli Studi di Roma "La Sapienza", 2003.
- [Ballou Pazer 2003] D. Ballou and Pazer Modeling Completeness versus Consistency Tradeoffs, in Information Decision Context IEEE Trans. on DKE, 2003.
- Time dimensions
- [Bouzeghoub et al 2004] Mokrane Bouzeghoub and Verónica Peralta: A Framework for Analysis of Data Freshness, SIGMOD IQIS workshop, Paris, France 2004.
- [Cappiello et al 2002] Cinzia Cappiello, C. Francalanci, B. Pernici, : A Model of Data Currency in Multi-Channel Financial Architectures. IQ 2002, Boston, USA, 2002
- [Cappiello et al 2004] C. Cappiello, C. Francalanci, B. - Time related Factors of DQ in multichannel information systems Journal of Management Information Systems 2004.

## Methodologies -1

- Complete methodologies
- [Redman 1996] Redman T.C.: Data Quality for the Information Age. Artech House, 1996.
- [English 1998] Larry P. English Improving Data Warehouse and Business Information Quality, Wiley 1998.
- [Shankaranarayan et al. 2000] Shankaranarayan G., Wang R. Y. and Ziad M.: "Modeling the Manufacture of an Information Product with IP-MAP". In Proceedings of the 6th International Conference on Information Quality, Boston, MA, 2000.
- [Pierce 2002] E. Pierce - Extending PI-Maps: incorporating the event driven Process Chain Methodology ICIQ 2002.
- [Pierce 2001] Pierce E. M.: "Using Control Matrices to Evaluate Information Product Maps". In Proceedings of the 7th Conference on Information Quality. Boston, MA, 2001.
- [Scannapieco et al 2004] M. Scannapieco, B. Pernici, E.B. Pierce, IP-UML: A Methodology for Quality Improvement based on IP-MAP and UML. To appear on Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph (Richard Wang, ed.), Sharpe, M.E., 2004.
- [Istat 2004] Istat and Aipa - A Methodology for improving data quality of address data in Public Administration, 2004 (in Italian).
- [Goerk 2004] Manfred Goerk - SAP AG Data Quality@SAP An Enterprise Wide Approach Data quality goals CAISE 2004 Workshop on IQ
- Chapter 4: Economic Framework of data quality, 2004
- [Bertoletti et al 2004] M. Bertoletti, P. Missier, C. Batini, M. Scannapieco, P. Aimetti - Improving Government to Business relationships through data reconciliation and process re-engineering, to be published in MIS Series, 2004.
- [Avenali et al 2004] A. Avenali, C. Batini et al - A formulation of the data quality optimization problem, Caise workshop on Data Quality, Riga 2004.

## Methodologies -2

- Methodologies for specific purposes: methodologies for data assessment
- [Wang et al 2001] Wang R.Y., Strong D.M., Kahn B.K., Lee Y.W. AIMQ: A Methodology for information quality assessment 2001 - Information & Management, published by Elsevier Science
- [Kahn et al] Kahn B.K, Strong D.M., Wang R.Y., Information quality benchmark: product and service performance
- [Pipino et al 2000] L. Pipino, Y. Lee, R. Wang - Data quality assessment CACM 45, 2000
- [De Amicis Batini 2004] F. De Amicis, C. Batini A Methodology for Data Quality Assessment on Financial Data, unpublished, available on request, 2004.
- [Cappiello 2002] C. Cappiello, C. Francalanci, B. Pernici - Process based methods for improving data quality - Daquincis Report 1, 2002.
- Quality of schema assessment
- [Mili 2003] Fatma Mili, Krish Narayanan - Theoretical Framework for Defining validity and quality in modelling ICIQ03.

## Models -1

- Extension of DB models
- [Storey 2001] V. Storey, R. Wang Extending the ER Model to Represent DQ Requirements - in Wang et al DQ, 2001
- [Wang et al 1990] Y. Richard Wang, Stuart E. Madnick: A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. VLDB 1990, Brisbane, Queensland, Australia, 1990.
- [Wang et al 2001] Data Quality, 2001 Extending the Relational model to capture DQ Attributes.
- [Wang et al 1995] Wang R.Y., Reddy M.P., Kon H.B.: Toward quality: an attribute-based approach, Decision Support Systems (DSS), 13, 1995.
- Models for management information systems
- [Ballou et al 1998] D. Ballou, R. Wang, H. Pazer, G.K.Tayi, Modelling Information Manufacturing Systems to Determine Information product Quality, Journal of Management Sciences, Vol.44, No.4, April 1998.
- [Shankaranarayan 2000] Shankaranarayan G., Wang R. Y. and Ziad M.: "Modeling the Manufacture of an Information Product with IP-MAP". In Proceedings of the 6th International Conference on Information Quality, Boston, MA, 2000.
- [Scannapieco et al 2004] M. Scannapieco, B. Pernici, E.B. Pierce, IP-UML: A Methodology for Quality Improvement based on IP-MAP and UML. To appear on Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph (Richard Wang, ed.), Sharpe, M.E., 2004.

## Models -2

- [Pierce 2004] E. Pierce - Assessing Data Quality with control matrices CACM 47, 2, 2004
- [Missier et al 2003] P. Missier, C. Batini - A multidimensional model for information quality in CIS, IQ 2003.
- Cost models
- [Avenali 2004] A. Avenali, C. Batini et al - A formulation of the data quality optimization problem, in Cooperative Information Systems - Caise workshop on Data Quality, Riga 2004.
- [Ballou Tayi 1999] D. Ballou and G. Tayi Enhancing Data Quality in Data Warehouse environments CACM 1999, V.42 N. 1
- [Loshin 2005] David Loshin, Enterprise Knowledge Management - The Data Quality Approach - Chapter 4, Knowledge intelligence incorporated, 2004.

## Record Linkage/Object Identification/Record Matching - 1

- For record linkage/object identification & Based on statistical data analysis and machine learning
- [Fellegi Sunter 1969] I.P. Fellegi and A.B. Sunter, A Theory for Record Linkage, Journal of the American Statistical Association 64 (1969).
- [Fellegi Sunter 1969] I. P. Fellegi and D. Holt, A Systematic Approach to Automatic Edit and Imputation, Journal of the American Statistical Association 71 (1976), 17(35).
- [Winkler a] W. Winkler - The State of Record Linkage and Current Research Problems, Statistical Society of Canada, Proceedings of the Section on Survey Methods, 73-79, 1999
- [Winkler b] W. Winkler - Machine Learning, Information retrieval and Record Linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 20-29, 2000 (also available at <http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf>).
- [Winkler 2004] W.E. Winkler, Methods for Evaluating and Creating Data Quality, Information Systems, vol.29 no.7 2004.
- [Elfeky et al 2002] M. G. Elfeky, V. S. Verykios, Ahmed K. Elmagarmid, Tailor: A Record Linkage Toolbox, ICDE 2002
- [Weis Naumann 2004] Melanie Weis and Felix Naumann: Detecting Duplicate Objects in XML Documents. SIGMOD IQIS workshop, Paris, France 2004.
- M.A. Jaro, Advances in record-linkage methodology applied to matching the 1985 census of Tampa, J. Amer. Statist. Assoc. 89 (1989) 414-420.
- W.E. Winkler, Advanced methods for record linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 1994, pp. 467-472, (longer version report rr94/05 available at <http://www.census.gov/srd/www/byyear.html>).
- [Winkler 1993] Improved decision rules in the Fellegi Sunter Model of Record linkage, Proc. of the section on survey research methods, American Statistical association.
- [Nigam 2000] K. Nigam et al. Text Classification form labelled and unlabelled documents using EM, Machine Learning 39.

## Record Linkage/Object Identification/Record Matching - 2

- Techniques For record linkage/object identification & Based on knowledge representation
- [Low et al 2001] Wai Lup Low, Mong Li Lee, Tok Wang Ling - A knowledge-based approach for duplicate elimination in data cleaning, Information Systems 2001
- [Tejadaa 2001] Sheila Tejadaa, Craig A. Knoblocka, Steven Mintonb Learning object identification rules for information integration Information Systems, 2001
- [Lim et al 1993] Lee-Peng Lim, Jaideep Srivastava, Satya Prabhakar, James Richardson: Entity Identification in Database Integration. ICDE 1993, Wien, Austria 1993.

## Record Linkage/Object Identification/Record Matching - 3

- Techniques For record linkage/ object identification & Based on empirical models
- [Hernandes Stolfo 1995] M.A. Hernandez and S.J. Stolfo, The Merge/Purge Problem for Large Databases, Proceedings of ACM Special Interest Group on Management Of Data International Conference (SIGMOD 1995), San Jose, California, 1995
- [Bertolazzi et al 2003] P. Bertolazzi, Luca De Santis, Monica Scannapieco, Automatic Record Matching in Cooperative Information Systems. Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, Italy, 2003.
- [Ananthakrishna et al 2002] Rohit Ananthakrishna, Surajit Chaudhuri, Venkatesh Ganti: Eliminating Fuzzy Duplicates in Data Warehouses, VLDB 2002, Hong Kong, China, 2002.
- [Gravano 2001] L. Gravano, P. Iperiotis, H. Jagadish, N. Koudas, S. Muthukrishnan, D. Srivastava - Approximate string joins in a database (Almost) for free, Proceedings 27th VLDB Conference, Roma, July, 2001.
- [Gravano 2003] L. Gravano, P. Iperiotis, N. Koudas., D. Srivastava - Text joins in an RDBMS for web data integration, WWW 2003, Budapest, May, 2003.
- [Guha 2002] S. Guha, H. Jagadish, N. Koudas, D. Srivastava, T. Yu - Approximate XML joins, Proceedings ACM Sigmod 2002, Winsconsin, USA, June, 2002
- [Koudas 2004] N. Koudas, A. Marathe, D. Srivastava -Flexiblestring matching against Large Databases in practice - Proceedings 30th VLDB Conference, Toronto, Canada, 2004.
- Comparison of techniques
- [Neiling et al 2003] Mattis Neiling, Steffen Jurk, Hans-J. Lenz, Felix Naumann, Object Identification Quality - Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, Italy, 2003.

## Data integration - 1

- For data integration: general
- [Bouzeghoub Lenzerini 2001] M. Bouzeghoub and M. Lenzerini (editors), Special Issue on Data Extraction, Cleaning, and Reconciliation, Information Systems 26 (2001), no. 8.
- For data integration: Instance Conflict Resolution + result merging
- [Naumann, Häussler 2002] Felix Naumann, Matthias Häussler: Declarative Data Merging with Conflict Resolution. IQ 2002, Boston, MA, USA, 2002.
- [Sattler et al 2003] K. Sattler, S. Conrad, and G. Saake, Interactive Example-Driven Integration and Reconciliation for Accessing Database Integration, Information systems 28 (2003).
- [Fan 2001] Fan, K., Lu, H., Madnick, S., Cheung, D.(2001): Discovering and reconciling value conflicts for numerical data integration, *Information Systems* 26
- [Motro 2004] Amihai Motro, Philip Anokhin, Aybar C. Acar Utility-based Resolution of Data Inconsistencies IQIS 2004.
- [Otzu et al 1999] Yan, L. Ozsu, T. (1999). Conflict tolerant queries in AURORA, Proceedings of the Fourth International Conference on Cooperative Information Systems (CoopIS'99), Edinburgh, Scotland, UK.

## Data integration -2

- For data integration: selection of best sources
- [Naumann 1999] F. Naumann, U. Leser, J. Freytag - Quality-driven Integration of Heterogeneous Information Systems VLDB 1999
- [Mihaila 2000] Mihaila, G., Raschid, L., Vidal, M. Using quality of data metadata for source selection and ranking, in: Proceedings of the *Third International Workshop on the Web and Databases (WebDB'00)*, Dallas, Texas.
- [De Giacomo et Al. 2004] Tackling Inconsistencies in Data Integration through Source Preferences, IQIS 2004, Paris, France
- [Greco et al 2004 ] Data Integration with Preferences among Sources, ER 2004, Shangai China, 2004

## Data integration - 3

- Data Integration: Composition of qualities
- [Motro 1998] A. Motro and I. Rakov, Estimating quality of database, Proceedings of the 3rd International Conference on Flexible Query Answering Systems (FQAS'98), Roskilde, Denmark, 1998.
- [Naumann 2003] F. Naumann, J.C. Freytag, and U. Leser, Completeness of Information Sources, Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, Italy, 2003.
- [Reddy and Wang 2000] M. Reddy and R. Wang - Developing a data quality algebra In R. Wang, et al. Data Quality, 2000.
- [Scannapieco Batini 2004] M. Scannapieco C. Batini Completeness in the relational model, a Comprehensive framework - ICIQ 2004.
- [Parsiann et al 2002] A. Parsiann S. Sarkar and V. S. Jacob, Assessing Information Quality for the Composite Relational Operation Join ICIQ02.

## Other DQ activities

- Data Auditing /Error Localization
- [Mueller et al 2004] Mueller, H., Leser U., Freytag J.C.: Mining for Patterns in Contradictory Data, IQIS 2004, Paris, France, 2004
- [Jarke et al 2003] Dominik Luebbbers, Udo Grimmer, Matthias Jarke Systematic Development of Data Mining-Based Data Quality Tools. VLDB 2003, Berlin Germany.
- [Fellegi Holt 1976] I.P. Fellegi and D. Holt, A systematic approach to automatic edit and imputation, Journal of American Statistical Association, 71 1976, 17-35.
- [Dasu et al 2002] T. Dasu, Theodore Johnson, S. Muthukrishnan, Vladislav Shkapenyuk: Mining database structure; or, how to build a data quality browser. SIGMOD Conference 2002, Madison, Wisconsin, USA, 2002.\*[Dasu 2003] T. Dasu and T. Johnson Explanatory Data Mining and Data cleaning, Wiley 2003
- [Winkler 2004] W.E. Winkler, Methods for Evaluating and Creating Data Quality, Information Systems, vol.29 no.7 2004.
- Profiling
- [Dasu et al 2002] T. Dasu, Theodore Johnson, S. Muthukrishnan, Vladislav Shkapenyuk: Mining database structure; or, how to build a data quality browser. SIGMOD Conference 2002, Madison, Wisconsin, USA, 2002.

## Tools

- [Caruso et al 2000] Francesco Caruso, Munir Cochinwala, Uma Ganapathy, Gail Lalk, Paolo Missier: Telcordia's Database Reconciliation and Data Quality Analysis Tool. VLDB 2000 (Demonstration), Cairo, Egypt, 2000.
- [Raman et al 2001] Vijayshankar Raman, Joseph M. Hellerstein: Potter's Wheel: An Interactive Data Cleaning System. VLDB 2001, Rome, Italy
- [Galhardas et al 2001] Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon, Cristian-Augustin Saita: Declarative Data Cleaning: Language, Model, and Algorithms. VLDB 2001, Rome Italy.
- [Vassiliadis 2001] P. Vassiliadis, Z. Vagena, S. Skiadopoulos, N. Karayannidis, T. Sellis, Arktos: towards the modeling, design, control and execution of ETL processes Information Systems 2001
- [Elfeky et al 2002] M. G. Elfeky, V. S. Verykios, Ahmed K. Elmagarmid, TAILOR: A Record Linkage Toolbox, ICDE 2002.
- [Buechi et al 2003] Martin Buechi, Andrew Borthwick, Adam Winkel, Arthur Goldberg ClueMaker: a language for approximate record matching, ICIQ 2003.
- [Low et al 2001] Wai Lup Low, Mong Li Lee, Tok Wang Ling - A knowledge-based approach for duplicate elimination in data cleaning Information Systems 2001

## Frameworks and services for CISs

- [Cappiello, Francalanci et al 2003] C. Cappiello, C. Francalanci, B. Pernici, P. Plebani, M. Scannapieco: "Data Quality Assurance in Cooperative Information Systems: a Multi-dimension Certificate". In Proceedings of the ICDDT'03 Workshop on Data Quality in Cooperative Information Systems (DQCIS '03), Siena, Italy, 2003.
- [Scannapieco et al 2004] M. Scannapieco, A. Virgillito, M. Marchetti, M. Mecella, R. Baldoni: The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. Information Systems, vol. 29, no. 7, 2004.
- [De Santis 2003] De Santis, M. Scannapieco, T. Catarci: Trusting Data Quality in Cooperative Information Systems. In Proc. of 11th International Conference on Cooperative Information Systems (CoopIS 2003), Catania, Italy, 2003.
- [Cappiello et al 2003] C. Cappiello et al Data Quality assurance in CIS: a multidimension quality certificate IW DQ in CIS, Siena 2003
- [Mecella et al 2003] M. Mecella, M. Scannapieco, A. Virgillitto, R. Baldoni, T. Catarci, C. Batini, Managing Data quality in cooperative Information Systems, Journal of Data Semantics, 2003.

Thank you  
for your attention!