C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 12: Methodologies for Information Quality Assessment and Improvement

Terminologies adopted in chapter sections

Section	Topic	Types of information	Terminology adopted
2.	Methodologies in general	Information in general	Information & Information Quality
3.	Comparison of 13 methodologies	Different types of information	Information & Information Quality
4.	Detailed comparison of three methodologies: TDQM, TIQM, Istat	Different types of information	Information & Information Quality
5.	Assessment methodologies: Description of QAFD	Structured relational data	Data and Data Quality
6.	Assessment & improvement methodologies: the CDQM Methodology	Structured relational data	Data and Data Quality
7.	Case study on CDQM application	Structured relational data	Data and Data Quality
8.	Extension of CDQM	Structured relational data & Semistructured information	Information & Information Quality

Knowledge involved in the IQ measurement and improvement process

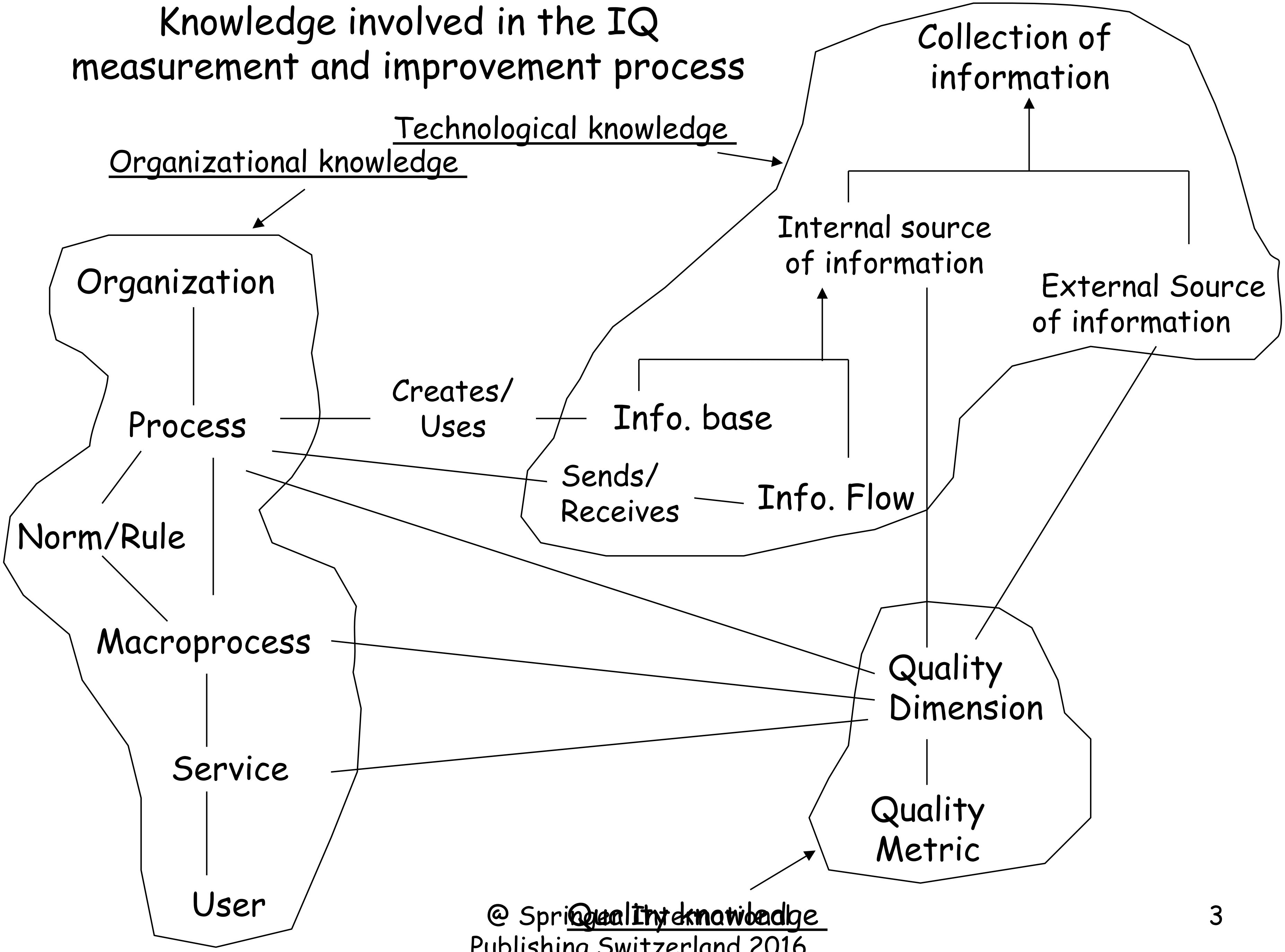
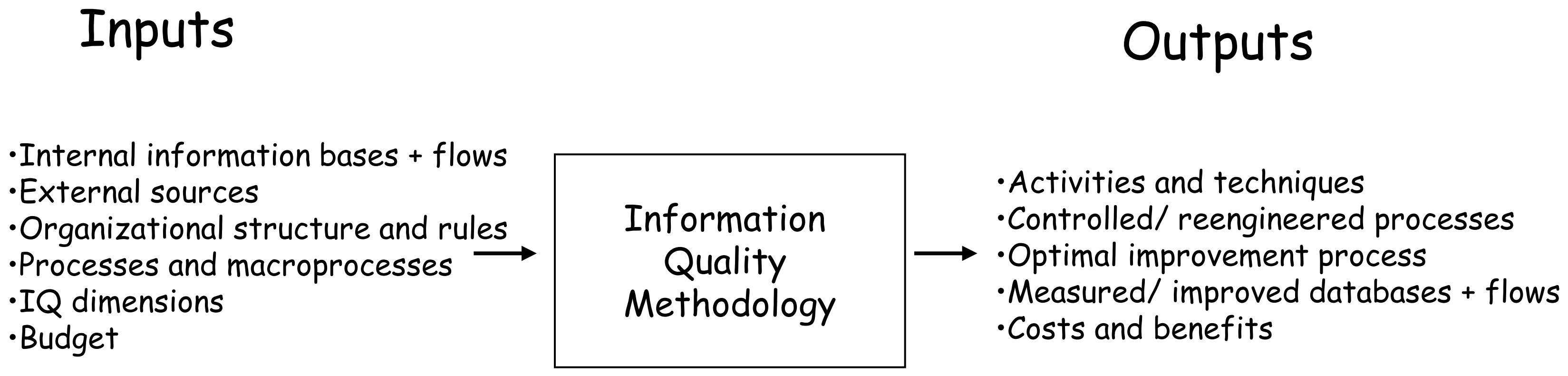


Fig knowledgeinvolved

Inputs and outputs of a IQ measurement and improvement methodology



Inputs

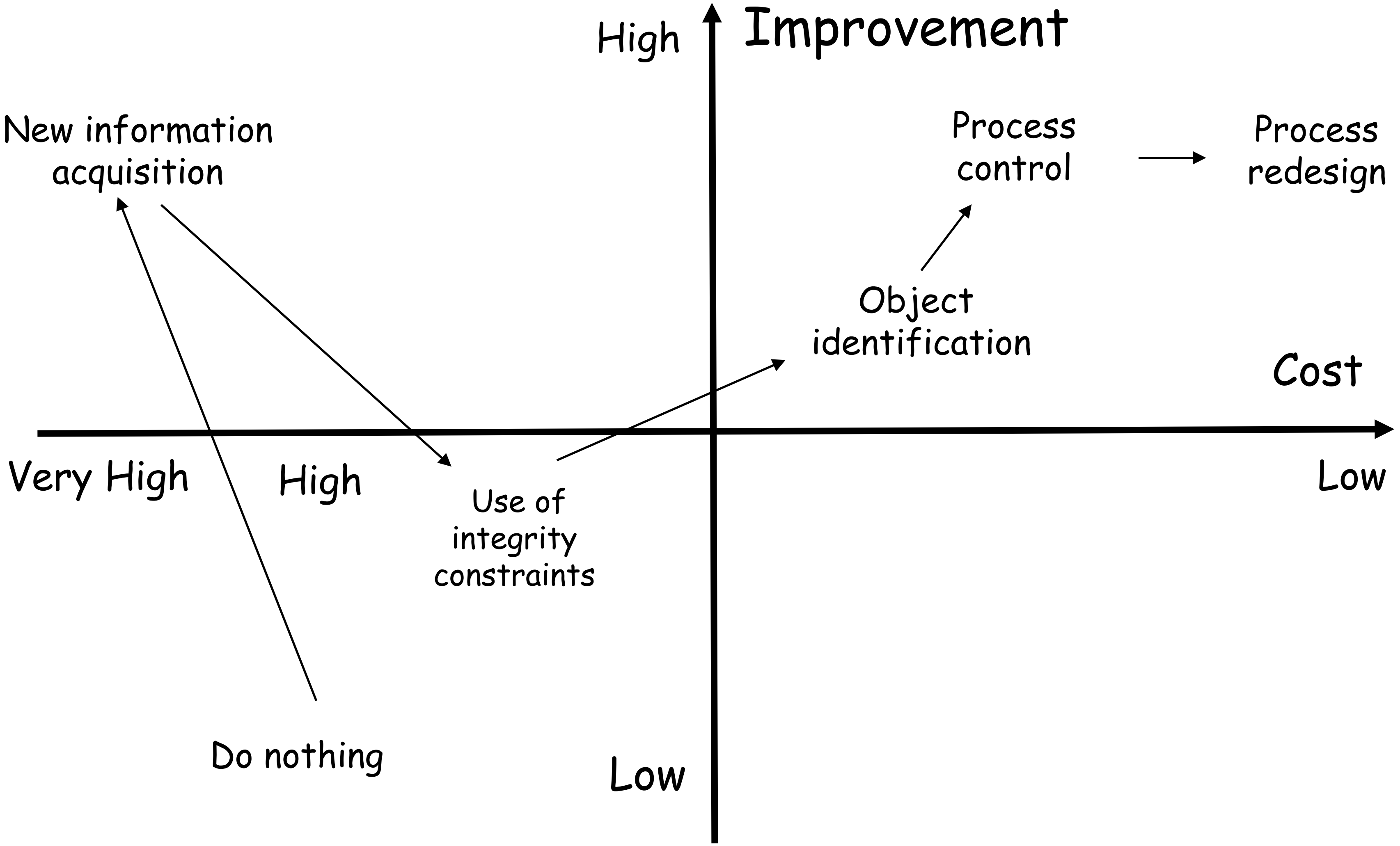
- Internal information bases + flows
- External sources
- Organizational structure and rules
- Processes and macroprocesses
- IQ dimensions
- Budget

Information Quality Methodology

Outputs

- Activities and techniques
- Controlled/ reengineered processes
- Optimal improvement process
- Measured/ improved databases + flows
- Costs and benefits

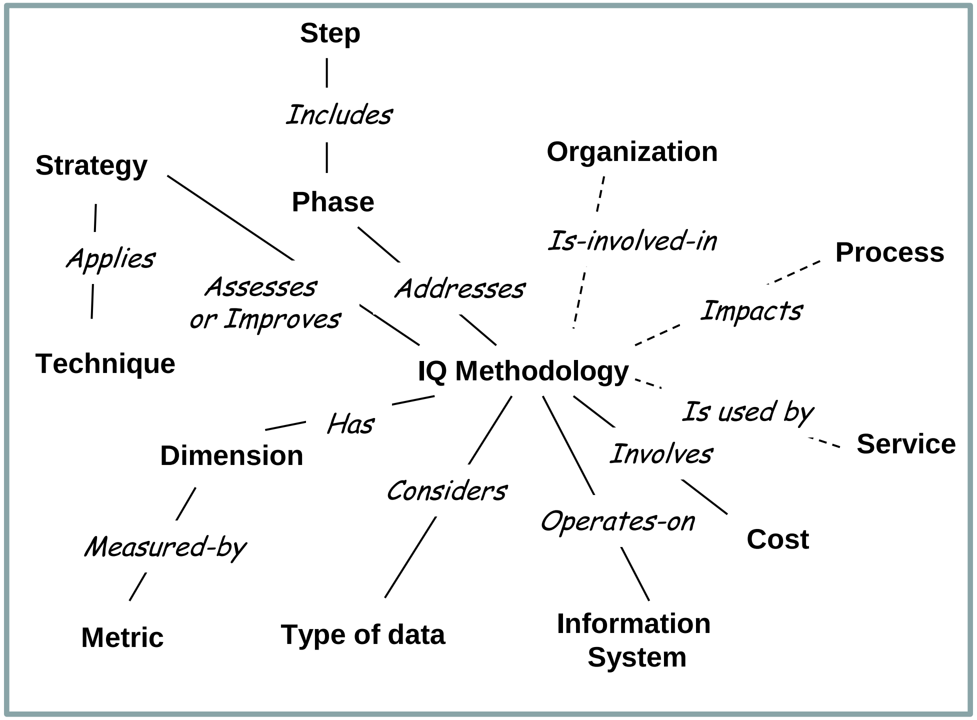
Improvement and cost of information/process-driven strategies: comparison in the long term



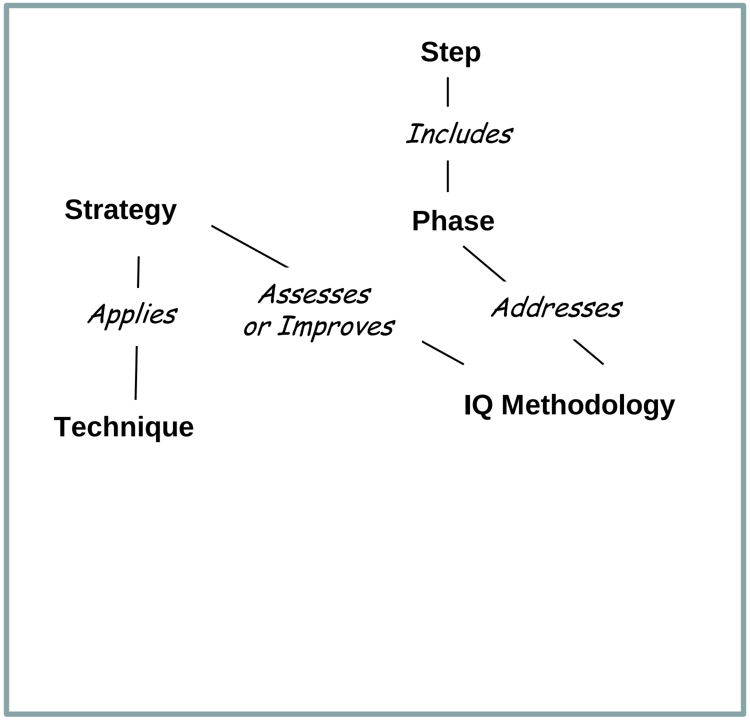
Methodologies compared in this section

Acronym	Extrended name	Main reference
TDQM	Total Data quality Management	Wang 1988
DWQ	The Datawarehouse Quality Methodology	Jarke 1999
TIQM	Total Information Quality Management	English 1999
AIMQ	A Methodology for information quality assessment	Lee 2001
CIHI	Canadian Institute for Health Information Methodology	Long 2005
DQA	Data Quality Assessment	Pipino 2002
IQM	Information Quality Measurement	Eppler 2002
ISTAT	ISTAT Methodology	Falorsi 2003
AMEQ	Activity Based Measuring and Evaluating of Product Information Quality Methodology	Su 2004
COLDQ	Cost Effect of Low Data Quality Methodology	Loshin 2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco 2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis 2004
CDQ	Comprehensive Methodology for Data Quality Management	Batini 2006

(a) Criteria adopted in [41] and (b) criteria considered in this section



(a)



(b)

Methodologies and assessment steps

Step/ Meth Acronym	Analysis	IQ Requirement Analysis	Identification of Critical Areas	Process Modeling	Measurement of quality	Extensible to other dimensions and metrics
TDQM	+		+	+	+	Fixed
DWQ	+	+	+		+	Open
TIQM	+	+	+	+	+	Fixed
AIMQ	+		+		+	Fixed
CIHI	+		+			Fixed
DQA	+		+		+	Open
IQM	+				+	Open
ISTAT	+				+	Fixed
AMEQ	+		+	+	+	Open
COLDQ	+	+	+	+	+	Open
DaQuinCIS	+		+	+	+	Open
QAFD	+	+			+	Fixed
CDQ	+	+	+	+	+	Open

Methodologies and improvement steps - part 1

Step/ Methodology Acronym	Evaluation of costs	Assignment of process responsibilities	Assignment of information responsibilities	Identification of the causes of errors	Selection of strategies and techniques	Design of information improvement solutions
TDQM	+	+	+	+	+	
DWQ	+		+	+	+	+
TIQM	+	+	+	+	+	+
DQA				+		
ISTAT				+	+	+
AMEQ				+		
COLDQ	+			+	+	+
DaQuinCIS				+	+	
CDQ	+	+	+	+	+	+

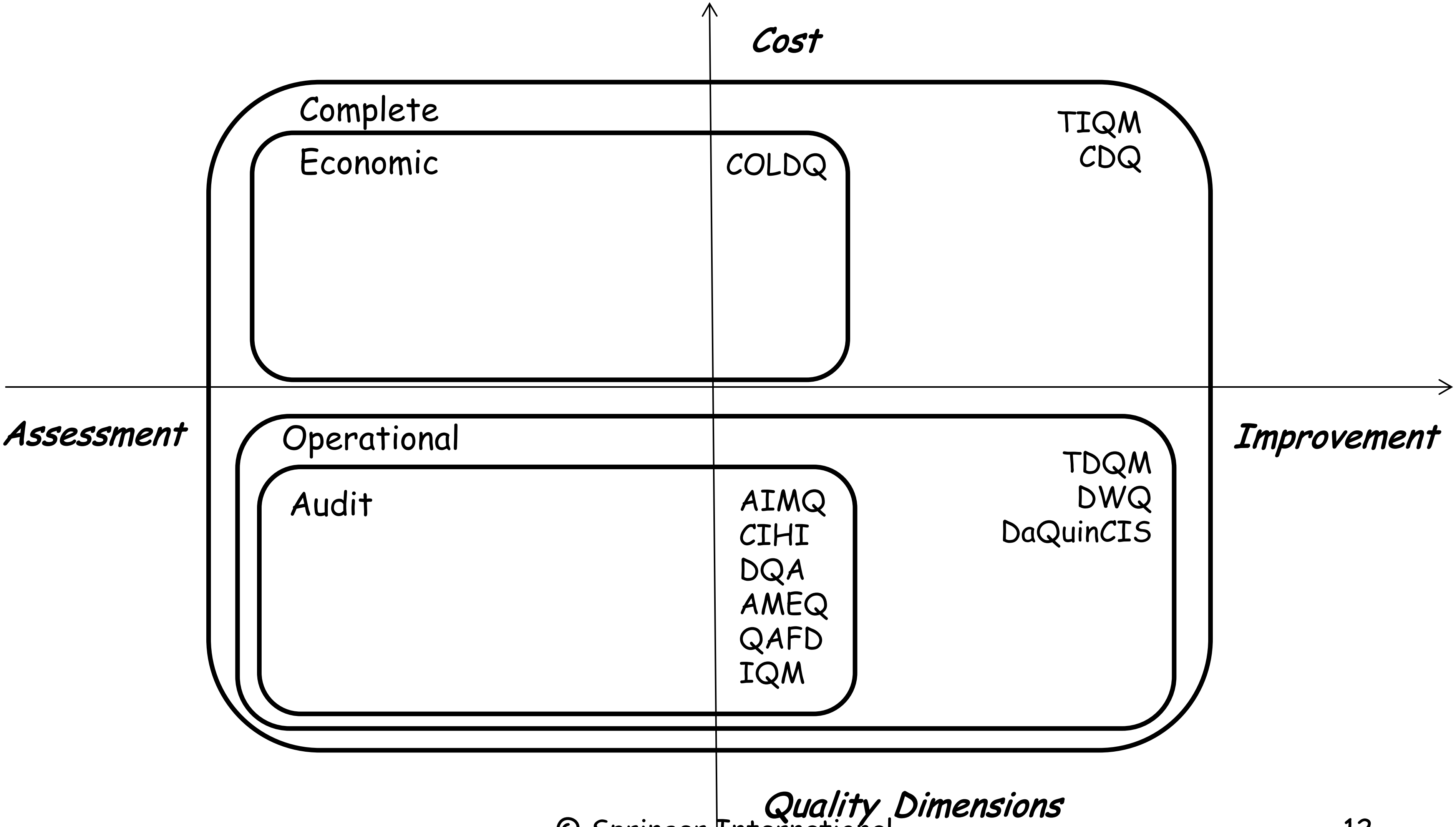
Methodologies and improvement steps - part 2

Step/Meth. Acronym	Process control	Process re-design	Improvement management	Improvement monitoring
TDQM		+	+	+
DWQ			+	
TIQM		+		+
DQA				
ISTAT		+		
AMEQ				+
COLDQ	+	+		+
DaQuinCIS				
CDQ	+	+		

Methodologies, strategies and techniques

Strategy/ Meth. Acronym	Data-driven	Process-driven
TDQM		Process Redesign
DWQ	Data and schema integration	
TIQM	Information cleansing Normalization Error localization and correction	Process Redesign
ISTAT	Standardization Object Identification	Process Redesign
COLDQ	Cost optimization	Process Control Process Redesign
DaQuinCIS	Source trustworthiness Object Identification	
CDQ	Standardization Object Identification Data and schema integration Error localization and correction	Process Control Process Redesign

A classification of methodologies



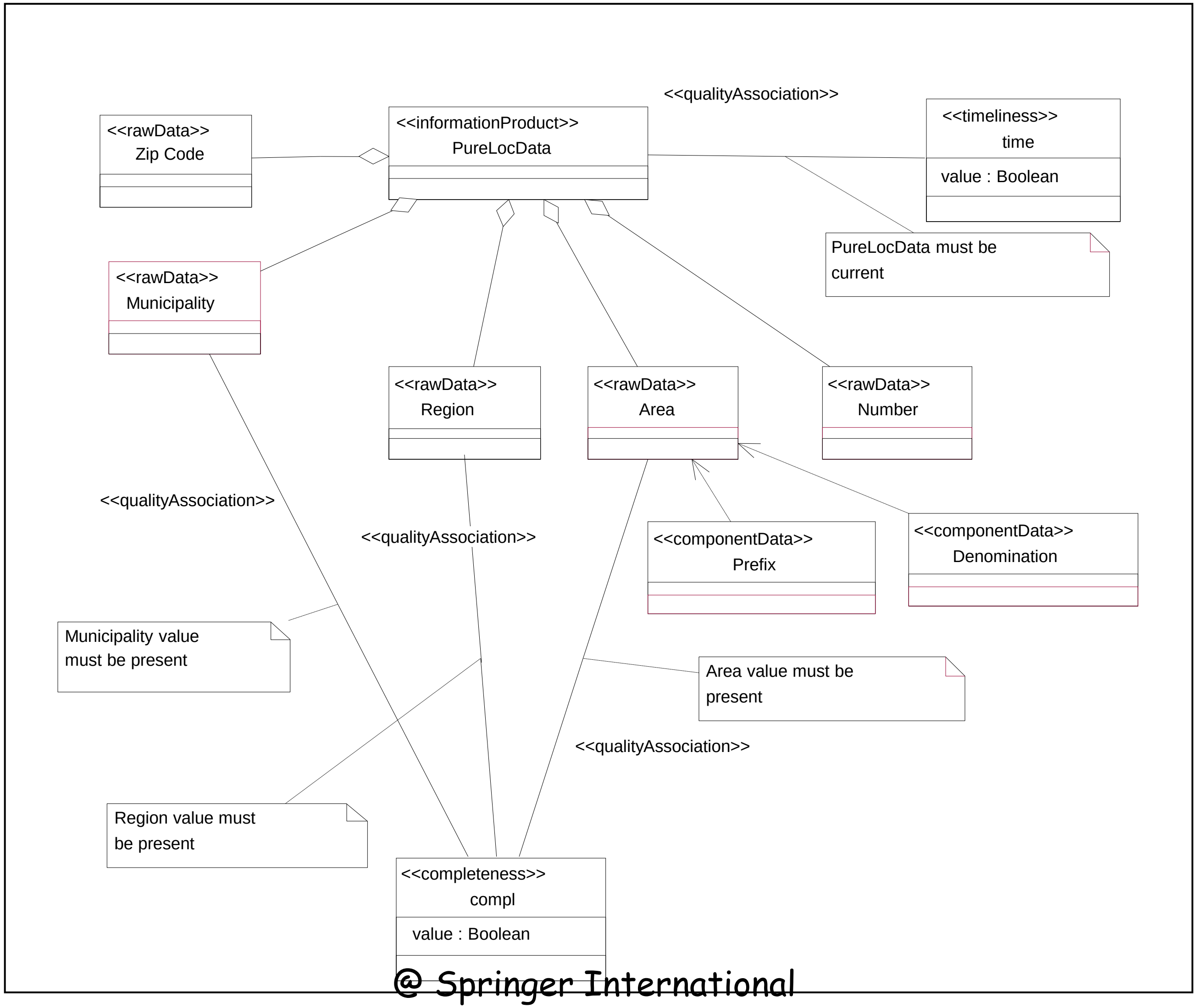
Classification of dimensions in [394] for assessment purposes

	Conforms to specifications	Meets or exceeds consumer expectations
Product quality	Sound Dimensions: Free of error Coincise representation Completeness Consistent representation	Useful Dimensions: Appropriate amount Relevancy Understandability Intepretability Objectivity
Service quality	Dependable Dimensions: Timeliness Security	Usable Dimensions: Believability Accessibility Ease of operation Reputation

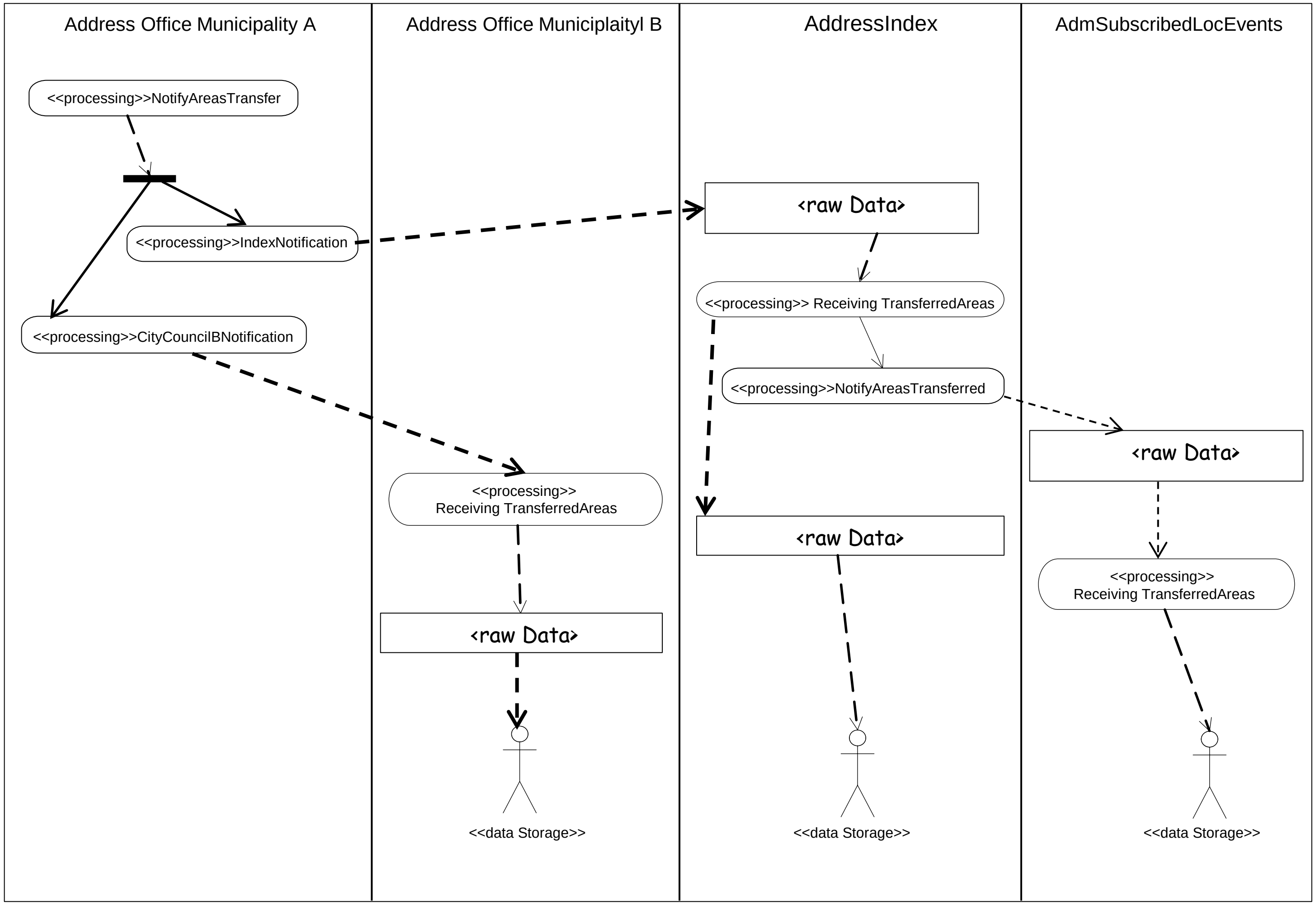
TDQM description

1. Definition
    - Data quality requirements analysis (named Quality Analysis in the IP-UML extension)
2. Measurement
    - Perform measurement (part of Quality Analysis in IP-UML)
3. Analysis
    - Data Analysis (the same name in IP-UML)
    - Model the processes (less relevant in IP-UML)
4. Improvement (Quality improvement in IP-UML)
    - Design improvement solutions on data and processes (Quality verification in IP-UML)
    - Re-design processes (only in IP-UML, named Quality improvement)

An example of quality analysis model in IP-UML



An example of a quality improvement model in IP-UML

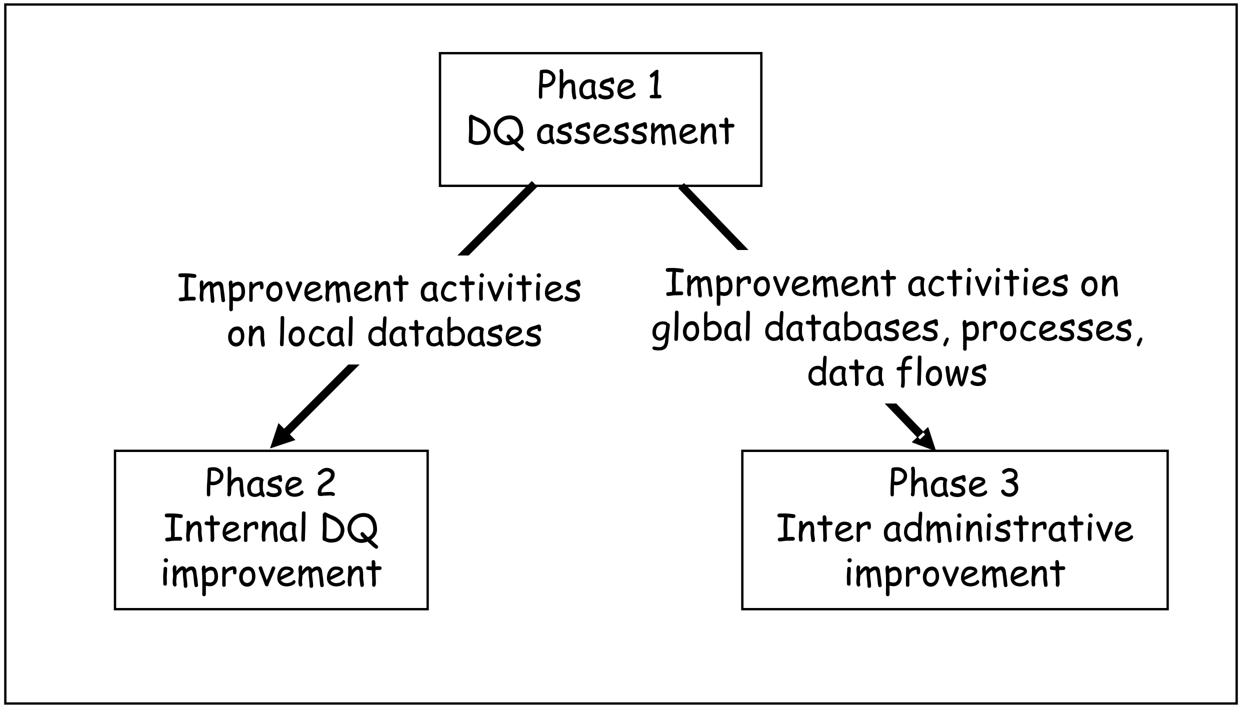

Address Office Municipality A
Address Office Municiplaityl B
AddressIndex
AdmSubscribedLocEvents
<<processing>>NotifyAreasTransfer
<<processing>>IndexNotification
<<processing>>CityCouncilBNotification
<raw Data>
<<processing>> Receiving TransferredAreas
<<processing>>NotifyAreasTransferred
<raw Data>
<<processing>>
Receiving TransferredAreas
<raw Data>
<raw Data>
<<processing>>
Receiving TransferredAreas
<<data Storage>>
<<data Storage>>
<<data Storage>>

@ Springer International
Publishing Switzerland 2016

TIQM description

1. Assessment
    - Data analysis
        - Identify information groups and stakeholders
        - Assess consumer satisfaction
    - DQ requirements analysis
    - Measurement
        - Identify data validation sources
        - Extract random samples of data
        - Measure and intepret data quality
    - Non quality evaluation
        - Identify business performance measures
        - Calculate non quality costs
    - Benefit evaluation
        - Calculate information value
2. Improvement
    - Design solution improvement
        - On data
            - Analyse data defect types
            - Standardize data
            - Correct and complete data
            - Match, transform and consolidate data
        - On processes
            - Check effectiveness of improvement
3. Management of improvement solutions – organizational perspective
    - Assess the organization's readiness
    - Create a vision for information quality improvement
    - Conduct a customer satisfaction survey of the information stakeholders
    - Select a small and payoff area to conduct a pilot project
    - Define the business problem to be solved
    - Define the information value chain
    - Perform a baseline assessment
    - Analyze customer complaints
    - Quantify costs due to quality problems
    - Define information stewardship
    - Analyze the systematic barriers to DQ and recommend changes
    - Establish a regular mechanism of communication and education with senior managers

General view of the Istat methodology



Phase 1
DQ assessment

Improvement activities on local databases

Improvement activities on global databases, processes, data flows

Phase 2
Internal DQ improvement

Phase 3
Inter administrative improvement

Detailed description of the Istat methodology

1. Global assessment and improvemement

1.1 Global assessment

- DQ Requirements analysis – Isolate from a general process analysis relevant qualities for address data: accuracy, completeness.
- Find critical areas, using statistical techniques
    - Choose a national database
    - Choose a representative sample
    - Find critical areas
    - Find potential causes of errors
- Communicate results of assessment to single agencies

1.2 Global improvement

- Design improvement solutions on data
    - Perform record linkage between relevant national databases
    - Establish a national data owner for specific fields
- Design improvement solutions on processes – Use the results of the global assessment to decide specific interventions on processes
- Choose tools and techniques – Make or buy, and adapt, tools for most relevant DQ activities to deliver to agencies

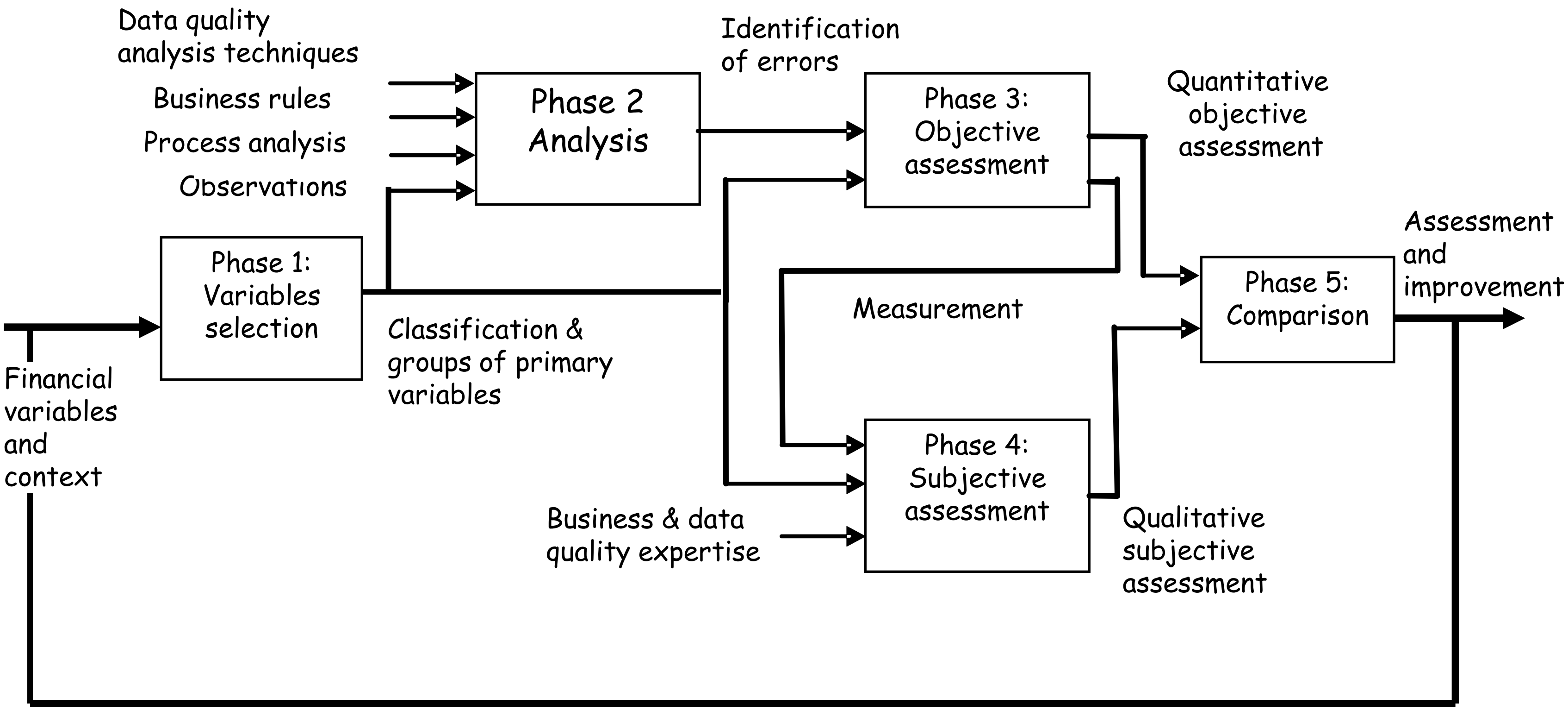
2. Internal DQ improvement (for each agency, autonomous initiative)

- Design improvement solutions on processes
    - Standardize acquisition format
    - Standardize internal exchange format using XML
- Perform specific local assessments
- Design improvement solutions on data and processes in critical areas
    - Use the results of the global assessment and local assessment to decide specific interventions on internal processes
    - Use the results of the global assessment and the acquired tools to decide specific interventions on data, e.g. perform record linkage between internal databases

3. DQ improvement of inter administrative flows

- Standardize inter administrative flows format using XML
- Redesign exchange flows, using a public and subscribe event-driven architecture

The main phases of the assessment methodology described in [168]


Data quality analysis techniques
Business rules
Process analysis
Observations
Phase 2 Analysis
Identification of errors
Phase 3: Objective assessment
Quantitative objective assessment
Phase 1: Variables selection
Classification & groups of primary variables
Financial variables and context
Measurement
Phase 5: Comparison
Assessment and improvement
Phase 4: Subjective assessment
Business & data quality expertise
Qualitative subjective assessment

Example of objective quantitative assessment

	Variables		
Quality dimensions	Moody’s Rating	Standard’s & Poor Rating	Market Currency Code
Syntactic Accuracy	1.7	1.5	2.1
Semantic Accuracy	0	0.1	1.4
Internal Consistency	2.7	3.2	1.3
External Consistency	1.6	1.1	0.1
Incompleteness	3.5	5.5	8.1
Currency	0	0	0
Timeliness	8.6	9.2	2
Uniqueness	4.9	4.9	9.3
Total (average)	3.6	3.2	3.0

Example of subjective quantitative assessment

	Rating Moody's	Rating S&P	Market Currency Code
Syntactic Accuracy	H	H	H
Semantic Accuracy	H	H	M
Internal Consistency	H	H	H
External Consistency	H	H	M
Incompleteness	L	L	L
Currency	H	H	H
Timeliness	M	M	H
Uniqueness	H	H	H
Total	H	H	H

Phases and steps of CDQM

Phase 1: State reconstruction

1. Reconstruct the state and meaning of most relevant databases and data flows exchanged between organizations, and build the *database + dataflow/organization matrixes.*

2. Reconstruct most relevant business processes performed by organizations, and build the *processes /organizations matrix.*

3.For each process or group of processes related in a macroprocess, reconstruct the norms and organizational rules that discipline the macroprocess and the service provided.

Phase 2: Assessment

4. Check the major problems related with the services provided with the internal and final users. Fix these drawbacks in terms of process and service qualities, and identify the causes of the drawbacks due to low data quality.

5. Identify relevant DQ dimensions and metrics, measure data quality of databases and data flows, and identify their critical areas.

Phase 3: Choice of the optimal improvement process

6. For each database and data flow, fix the new DQ levels that improve process quality and reduce costs under a required threshold.

7. Conceive process re-engineering activities and choose DQ activities, that may lead to DQ improvement targets set in step 6, relating them in the data/*activity matrix* to clusters of databases and data flows involved in DQ improvement targets.

8. Choose optimal techniques for the DQ activities.

9. Connect crossings in the *data/activity matrix* in reasonable candidate improvement processes

10. For each improvement process defined in the previous step, compute approximate costs and benefits, and choose the optimal one, checking that the overall cost-benefit balance meets the targets of step 6.

The database/organization matrix

Database/ Organization	Database 1	Database 2	………	Database n
Organization 1	Creates	Uses		Uses
Organization 2		Uses		
…………				
Organziation m		Creates		Creates

The dataflow/organization matrix

Dataflow/ Organization	Dataflow 1	Dataflow 2	………	Dataflow n
Organization 1	Provider	Consumer		Consumer
Organization 2		Consumer		Provider
…………				
Organization m	Consumer	Provider		Consumer

The process/organization matrix

Process/ Organization	Process 1	Process 2	………	Process n
Organization 1	Owner	Participates		
Organization 2		Participates		Owner
…………				
Organization m	Participates	Owner		Participates

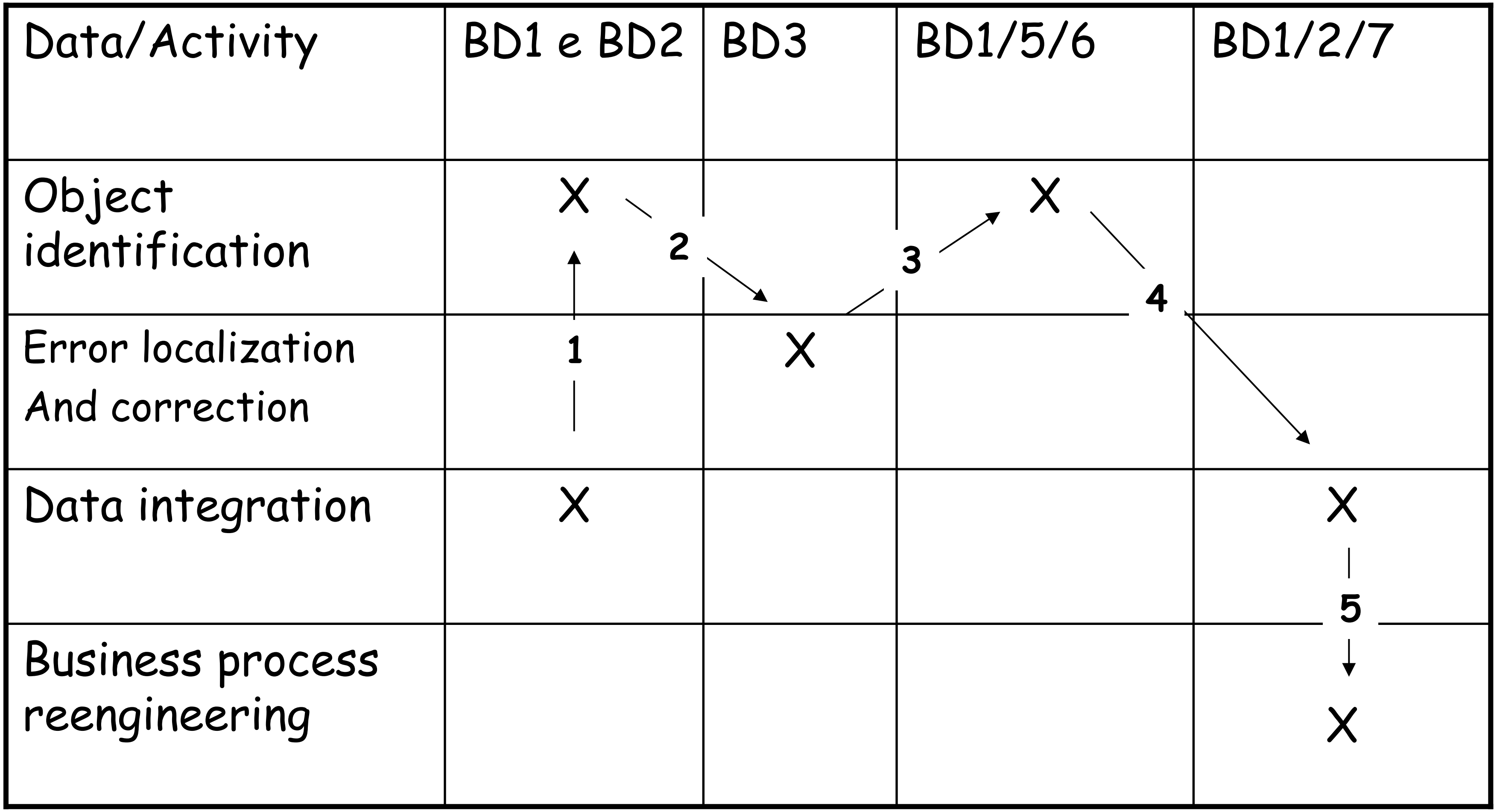
The macroprocess/norm-service-process matrix

Macroprocess	Macroprocess1	Macroprocess2	………	Macroprocess m
Norm/organizational rule	Norm 1	Norm 2		Norm3 and Norm4
Service(s)	S1 and S5	S2 and S5		S3 and S4
Process 1	X			
Process 2		X		
Process 3	X			
Process 4	X			
…				
Process n				X

The data/activity matrix

Data/Activity	DB1+DB2	DB1+DB3	DB4	DB5	DF1+DF2	DF3
DQ Activity 1	X		X			
DQ Activity 2		X				X
DQ Activity 3		X		X	X	
Process Re-engineering Activity 1	X		X			X
Process Re-engineering Activity 1		X	X		X	
Process Re-engineering Activity 1	X	X		X	X	

An example of improvement process



Data/Activity	BD1 e BD2	BD3	BD1/5/6	BD1/2/7
Object identification	X		X	
Error localization And correction		X		
Data integration	X			X
Business process reengineering				X

The database/organization matrix

Database/ Organization	SocialSecurity Registry of businesses	Accident Insurance Registry of businesses	Chambers of Commerce Registry of businesses
SocialSecurity	Creates/Uses		
Accident Insurance		Creates/Uses	
Chambers of Commerce			Creates/Uses

The dataflow/organization matrix

Dataflow/ Organization	Dataflow 1: Information for service request	Dataflow 2: Information related to service provision
SocialSecurity	Consumer	Provider
Accident Insurance	Consumer	Provider
Chambers of Commerce	Consumer	Provider
Businesses	Provider	Consumer

The process/organization matrix

Process/ Organization	Update registered office info	Update branches info	Update main economic activity info
SocialSecurity	X	X	X
Accident Insurance	X	X	X
Chambers of Commerce	X	X	X

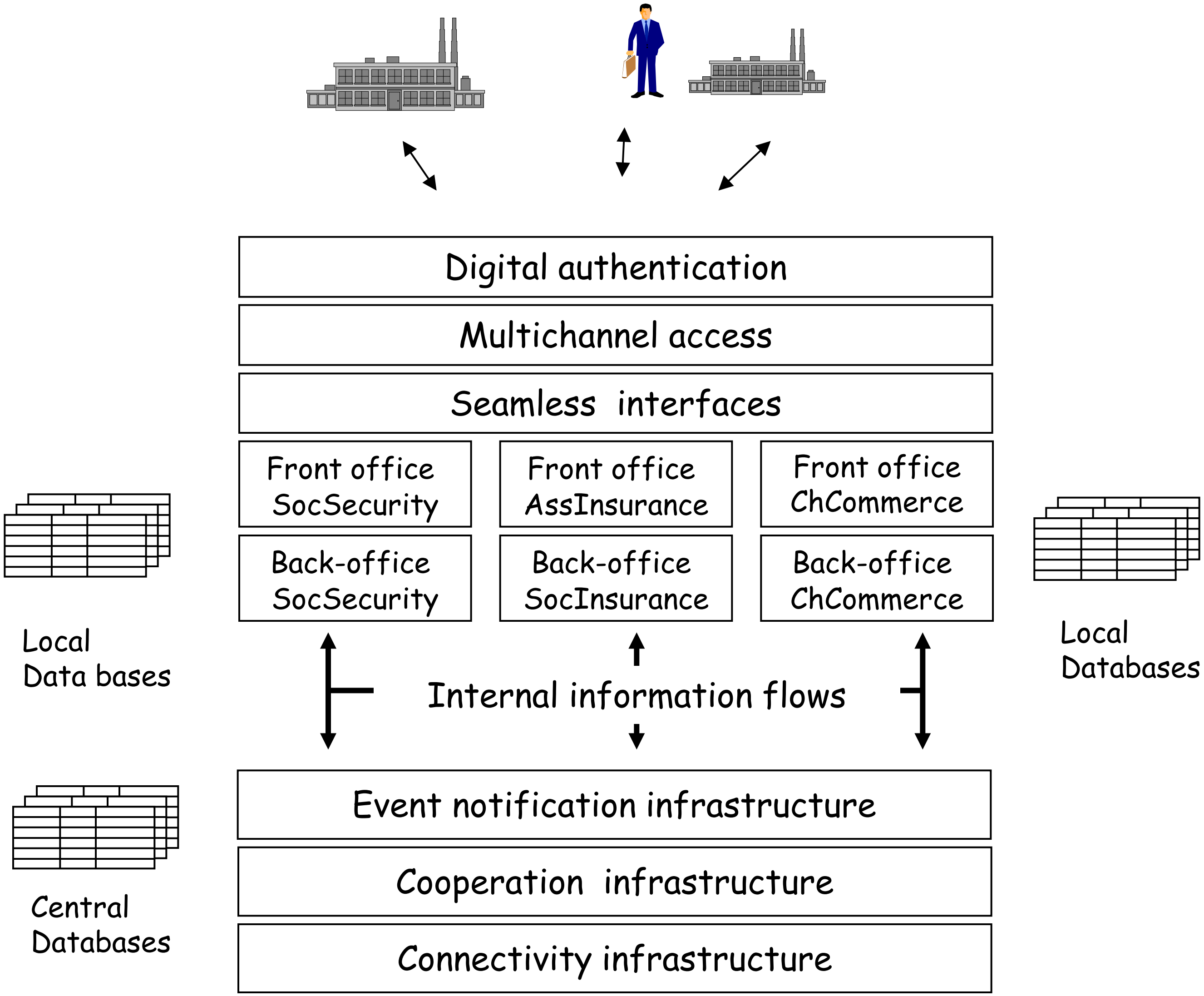
Actual quality levels

Quality dimension/ Database	Duplicate objects	Matching objects	Accuracy of names and addressed	Currency
SocialSecurityDB	5%	--	98%	3 months delay
Accident Insurance DB	8%	--	95%	5 months delay
Chambers of Commerce DB	1%	--	98%	10 days delay
The three databases together	--	80%	--	--

Target quality levels

Quality dimension/ Database matrix	Duplicate objects	Matching objects	Accuracy of names and addressed	Currency
SocialSecurityRegistry	1%	--	99%	3-4 days delay
Accident Insurance Registry	1%	--	99%	3-4 days delay
Chambers of Commerce registry	0.3%	--	99%	2-3 days delay
The three registries together	--	97%	--	--

New technological architecture for Government-to-Business interactions



Digital authentication

Multichannel access

Seamless interfaces

Front office SocSecurity	Front office AssInsurance	Front office ChCommerce
Back-office SocSecurity	Back-office SocInsurance	Back-office ChCommerce

Internal information flows

Event notification infrastructure

Cooperation infrastructure

Connectivity infrastructure

The data/activity matrix

Data/Activity	Type of activity	The three databases together	New flows between agencies	The new Identifiers database
Object identification	Data driven	X		
Process Reengineering on update processes	Process driven	X	X	X

@ Springer International
Publishing Switzerland 2016

An improvement in the example

Data/Activity	The three DataBases together	New flows between agencies	The new Identifiers DB
Object identification	Perform object identification on the stock and consequent deduplication on the three DBs		
Process Reengineering on update processes	Update first the Chambers ofCommerce DB	Use the P&S Infrastructure toUpdate SocSec DB and SocIns DB	Create the DB and use it in the new interagency update process

Costs and savings of the data quality improvement process

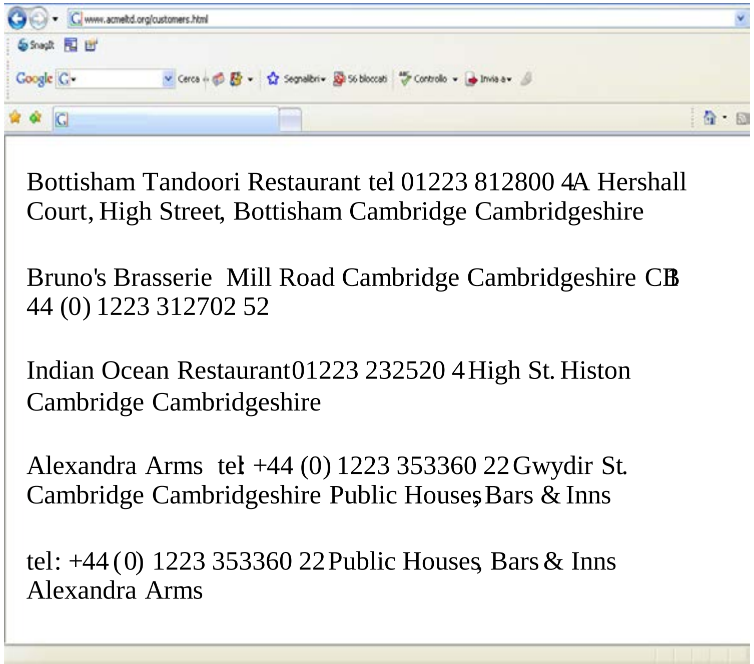
Costs and benefits	Once for all	Yearly
Actual costs due to poor data quality		
Clerical alignement costs		10 Ml
Reduced revenues (prudential)		300 Ml
Other costs		
For businesses		200 Ml
For agencies		100 Ml
Costs of the improvement project		
Object identification - automatic	800.000	
Object identification - clerical	200.000	
Application architecture – set up	5Ml	
Application architecture – maintenance		1Ml
Future costs and savings due to improved data quality		
Increased revenues (prudential)		200Ml
Clerical alignement costs		0
Other savings		
For businesses		130Ml
For agencies		60Ml

Requirements of the case study

- The core business of a private firm is to develop innovative systems for wireless hand-held order entry systems. These systems are used by waiters to collect orders from patrons at their tables and communicate with the kitchen in real time through a wireless connection. As the majority of businesses, the main entities to be managed are those of Customer and Supplier. In this example, we will concentrate on the Customer entity.
- The Marketing Department (MD) and its network of commercial agents are supposed to either seek new customers or propose new solutions and upgrades to old ones. MD agents need to have very precise information on the profile of potential customers as this can be acquired form specific vendors and aggregated along several dimensions, like region, turnover, and cuisine.
- The Technical Department (TD) is supposed to monitor the well running of sold installations and provide both ordinary and extraordinary maintenance upon on it. TD members must then rely on information about customers regarding systems purchased, and where they are located.
- Lastly, the Accounts Department (AD) needs accurate and up-to-date administrative information for invoice drawing and accounting.

The three information collections in input to the process

White page Directory



Bottisham Tandoori Restaurant tel 01223 812800 4A Hershall Court, High Street, Bottisham Cambridge Cambridgeshire

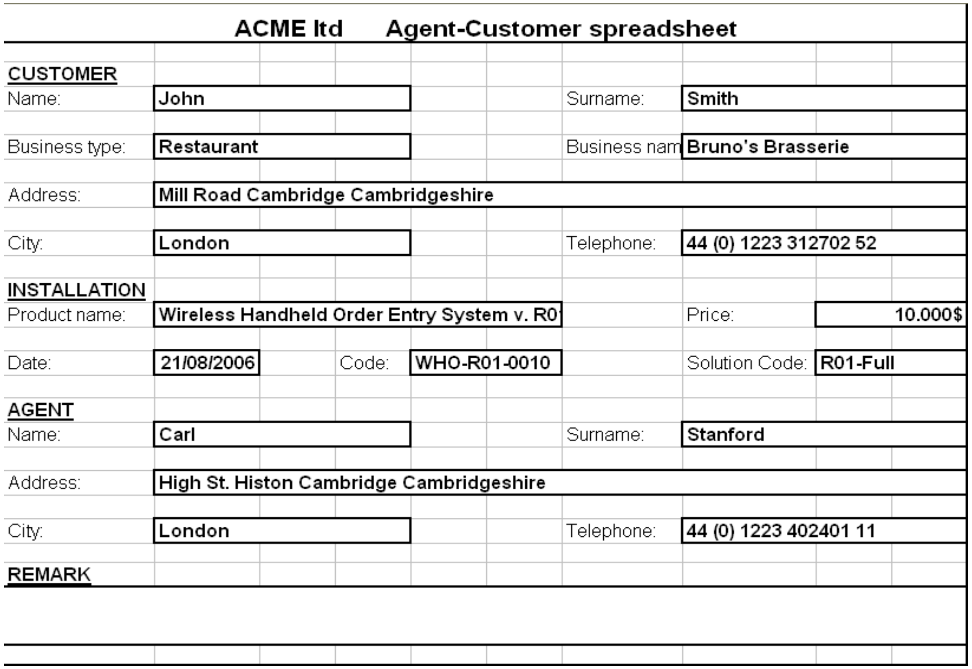
Bruno's Brasserie Mill Road Cambridge Cambridgeshire CB 44 (0) 1223 312702 52

Indian Ocean Restaurant01223 232520 4High St. Histon Cambridge Cambridgeshire

Alexandra Arms tel +44 (0) 1223 353360 22Gwydir St. Cambridge Cambridgeshire Public Houses, Bars & Inns

tel: +44 (0) 1223 353360 22Public Houses, Bars & Inns Alexandra Arms

Agent-Customer spreadsheet



ACME ltd Agent-Customer spreadsheet

CUSTOMER

Name: **John** Surname: **Smith**

Business type: **Restaurant** Business nam **Bruno's Brasserie**

Address: **Mill Road Cambridge Cambridgeshire**

City: **London** Telephone: **44 (0) 1223 312702 52**

INSTALLATION

Product name: **Wireless Handheld Order Entry System v. R0** Price: **10.000$**

Date: **21/08/2006** Code: **WHO-R01-0010** Solution Code: **R01-Full**

AGENT

Name: **Carl** Surname: **Stanford**

Address: **High St. Histon Cambridge Cambridgeshire**

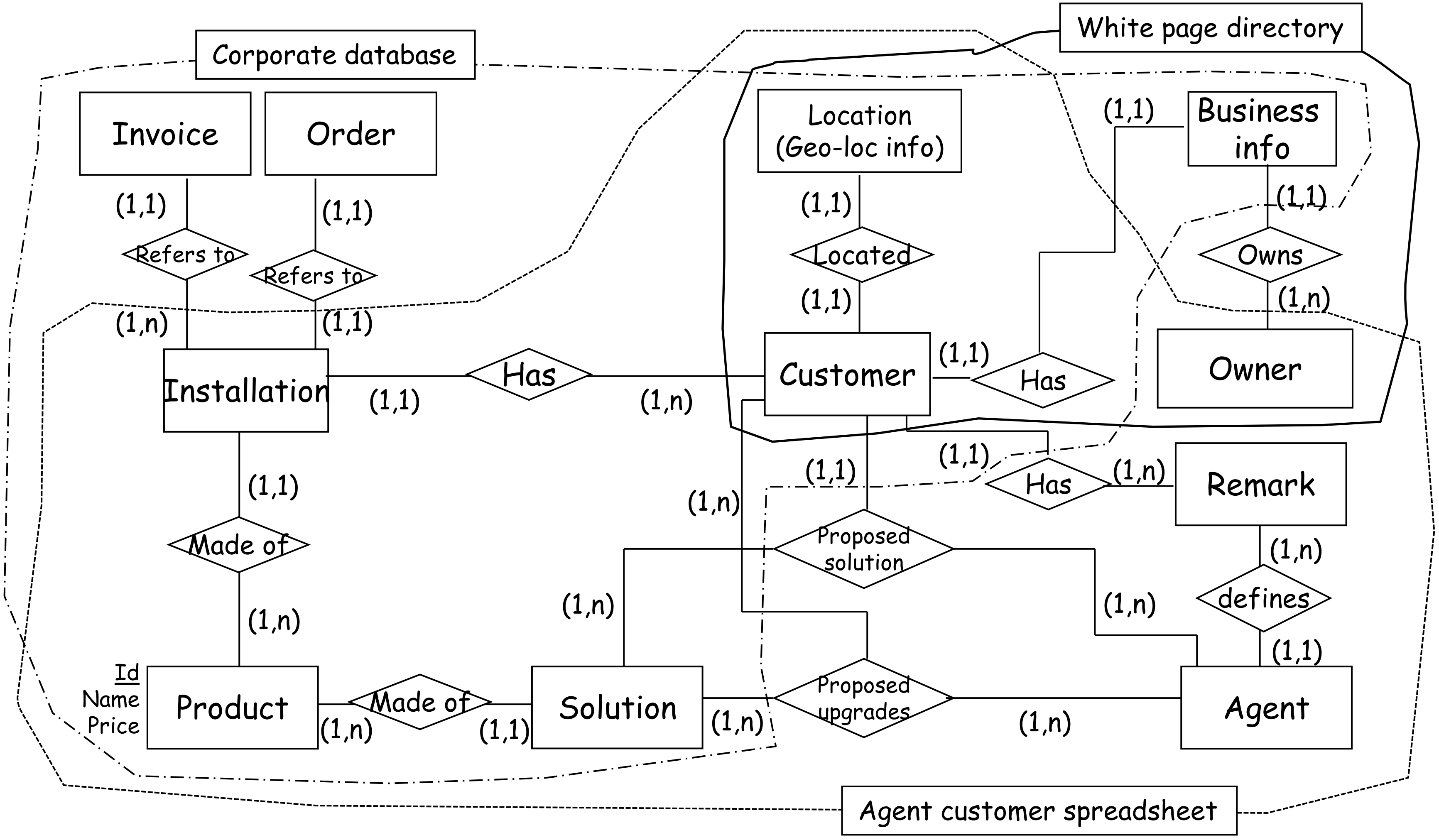
City: **London** Telephone: **44 (0) 1223 402401 11**

REMARK

Corporate database

ID_Customer	Name	Surname	Business Type	Business Name	ID_Installation	ID_Solution
001	John	Smith	Restaurant	Bruno's Brasserie	WHO-R01-0010	R01-Full
002	Simon	Kent	Restaurant	India Ocean	WHO-RO1-0011	R01-Full
003	Paul	Buck	Restaurant	Bottisham Tandoori	WHO-R01-0010	R01-Full

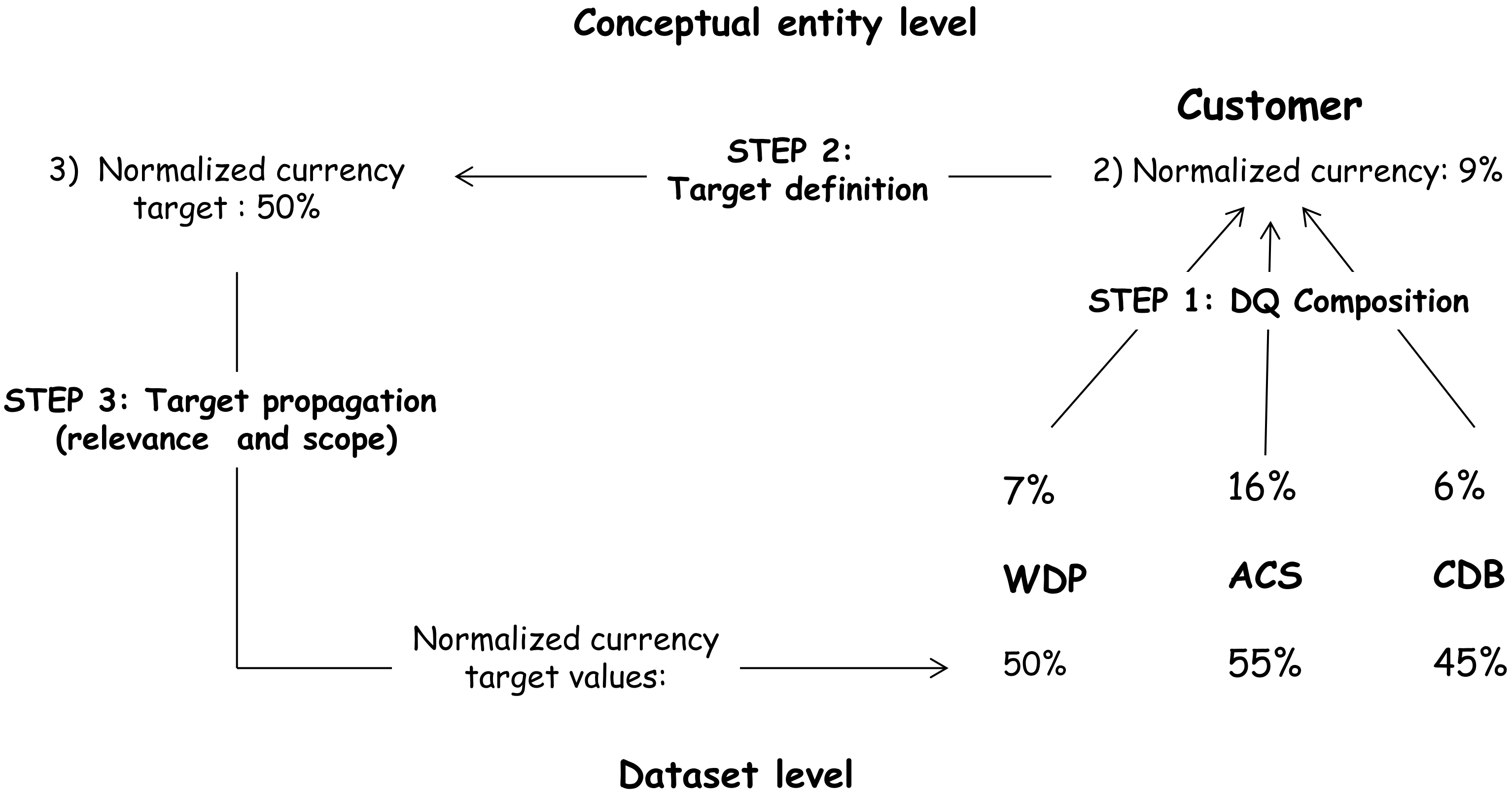
The integrated schema and the three input schemas



Currency assessment

Data set → Dimension	WPD	ACS	CDB
Actual currency	12 days delay	6 days delay	16 days delay
Optimal currency	1 day delay	1 day delay	1 day delay
Normalized currency	7%	16%	6%

Composition of currency values



Conceptual entity level

Customer

3) Normalized currency target : 50%

STEP 2: Target definition

2) Normalized currency: 9%

STEP 1: DQ Composition

STEP 3: Target propagation (relevance and scope)

7% 16% 6%

WDP ACS CDB

Normalized currency target values:

50% 55% 45%

Dataset level