discussion on data quality

China National Institute of Standardization

Nov, 2009

outline

* 1.Background
* 2.Current status of standards development
* 3.future work

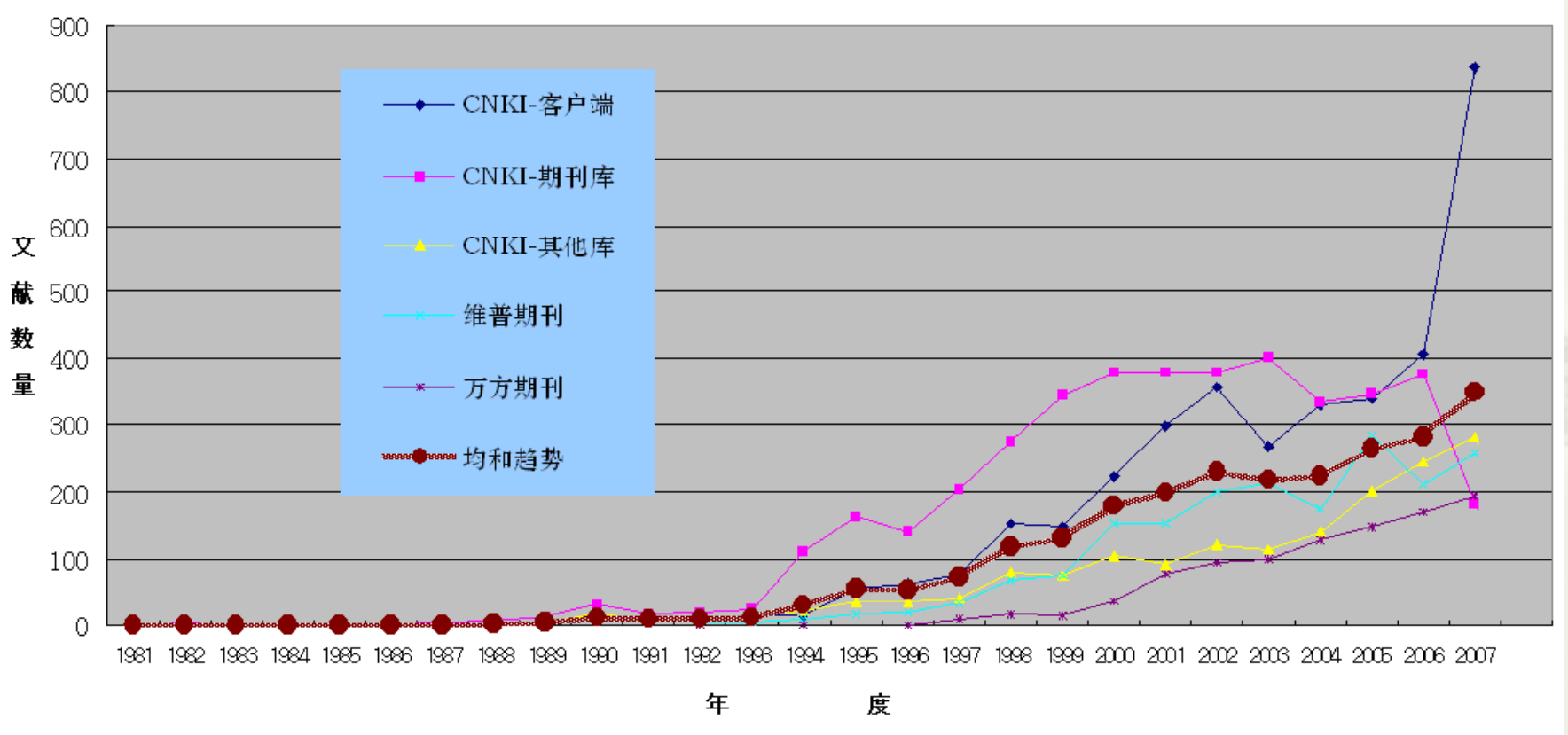
1.Background

1.1 In general

- Electronic data play a crucial role in the information and communication technology society, and the “quality” of such data and its related effects on every kind of activity of the ICT society are more and more critical.
- web search of the term “data quality” through Google：
  - ✓ 3 millions of pages in 2006
  - ✓ more than 7 millions in 2009

1.Background

In China， We did an investigation in 2008 about the papers related to the topic of data quality published in academic journals since 1981, and the brown line in the graph shows the tendency.



1.Background

There are several international conferences that have data quality as their main topic:

* The international conference on information quality(ICIQ), started in 1996
* The international workshop on information quality in information systems(IQIS), held since 2004
* The international workshop on data and information quality(DIQ) held since 2004
* The international workshop on quality of information systems(QoIS) held since 2005
* ……

1.Background

1.2 The benefits of data of high quality

- ✓ Data of high quality is a valuable asset
- ✓ Data of high quality can increase customer satisfaction
- ✓ Data of high quality can improve revenue and profits
- ✓ Data of high quality can be a strategic competitive advantage

1.Background

1.3 The consequences of data quality problems

- Data Warehousing Institute(in a 2002 report):there is a significant gap between perception and reality regarding the quality of data in many organizations, and the data quality problem cost US business more than 600 billion dollars a year.
- The explosion of the space shuttle Challenger was due to the data quality problems.
- The “Year 2000 problem”:the cost to modify such software applications and databases have been estimated to be around 1.5 trillion US dollars.

2.Current status

2.1. Standards developed and under development

- ISO/TS 8000-100:2009 Data quality -- Part 100: Master data: Overview
- ISO/TS 8000-110:2009 Data quality -- Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification.
- ISO/TS 8000-120:2009 Data quality -- Part 120: Master data: Exchange of characteristic data: Provenance
- ISO/PAS 26183-2006 SASIG Product data quality guidelines for the global automotive industry
- ISO/TR 21707-2008 Intelligent transport systems - Integrated transport information, management and control - Data quality in ITS systems
- ISO/IEC 25012:2008 Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model

2.Current status

2.1. Standards developed and under development

- BS 7986-2005 Data quality metrics for industrial measurement and control systems – Specification
- ISO 19113:2002 Geographic information — Quality principles
- ISO 19114:2003 Geographic information — Quality evaluation procedures
- ISO TS 19138-2006 Geographic information — Data quality measures
- ISO/NP 19157 Geographic information -- Data quality
- ISO/NP TS 19158 Geographic information - Quality assurance of data supply
- International Monetary Fund, IMF: Data Quality Assessment Framework, DQAF

2.Current status

2.2. Lack of a unified definition of data quality

- 1. Data Quality refers to the degree of excellence exhibited by the data in relation to the portrayal of the actual phenomena. GIS Glossary
- 2. The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. Government of British Columbia
- 3. The totality of features and characteristics of data that bears on their ability to satisfy a given purpose; the sum of the degrees of excellence for factors related to data. Glossary of Quality Assurance Terms
- 4. Information Quality : the fitness for use of information; information that meets the requirements of its authors, users, and administrators. (Martin Eppler)
- 5. Data quality: The processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria
- 6.ISO/PAS 26183:2006 defines product data quality as a measure of the accuracy and appropriateness of product data, combined with the timeliness with which those data are provided to all the people who need them.
- And more……

2.Current status

2.3. Focus on specific domains

All the standards developed and under development focus on some specific domains, e.g. manufacturing, intelligent transportation, geography, and so on.

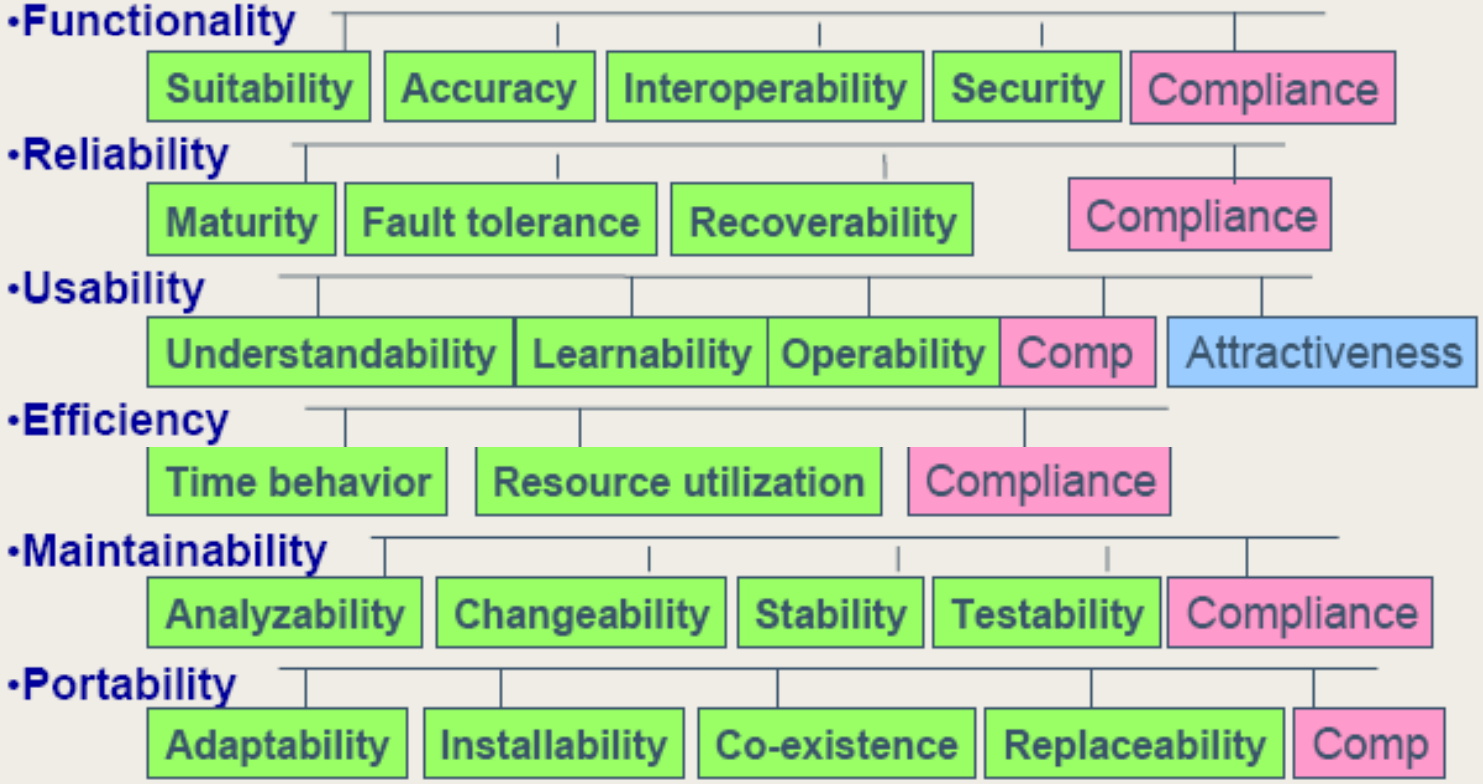
2.Current status

2.4. no general model/framework of data quality

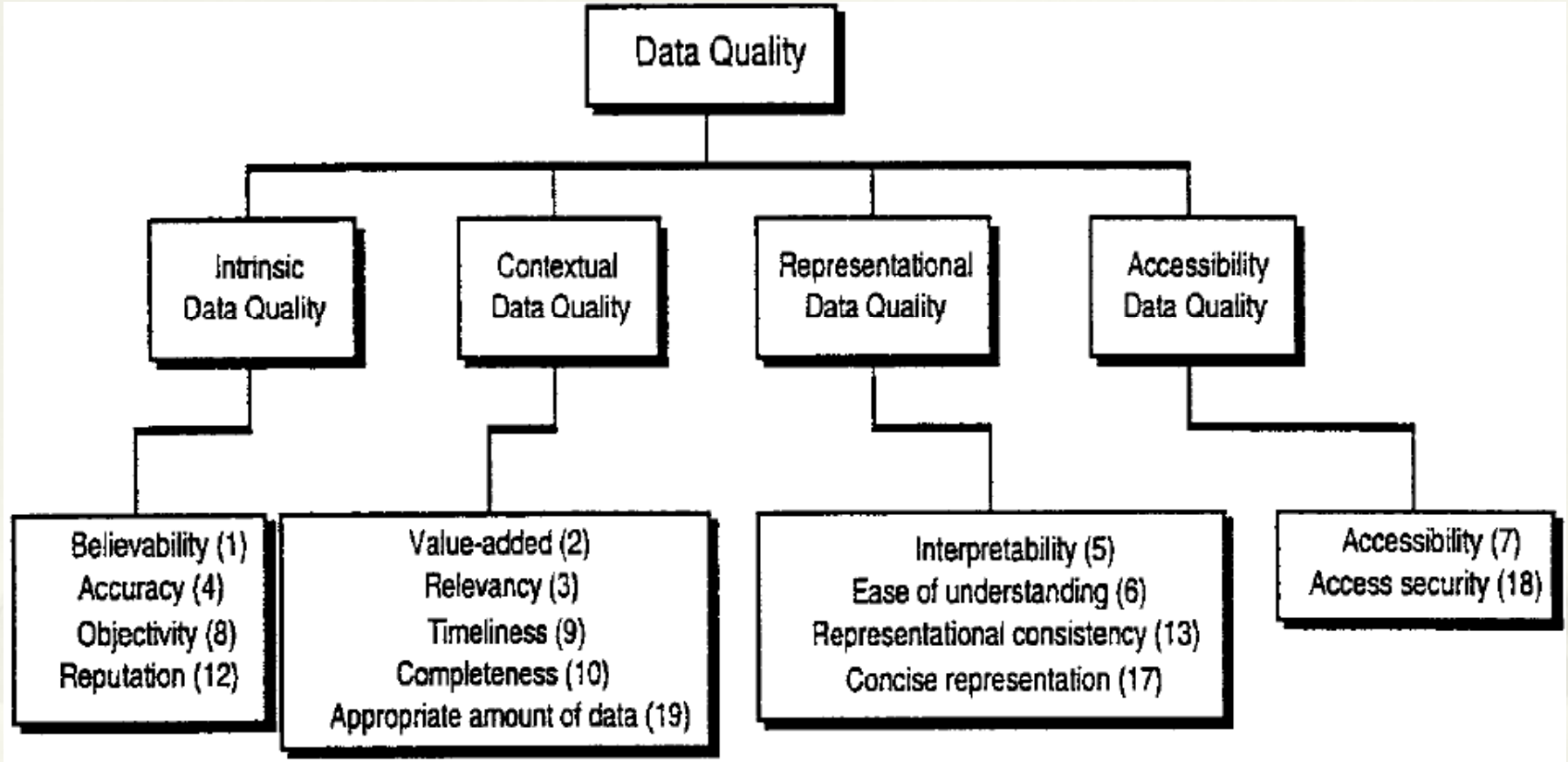
There is still not an general model/framework of data quality independent of any particular domain or application， and the existing domain models of data quality differ on dimensions, attributes, measurements， evaluations and so on.

2.Current status

Internal and External Software Quality Model (ISO/IEC 9126)



2.Current status



A conceptual framework of data quality, (RICHARD Y. WANG AND DIANE M. STRONG .Beyond Accuracy: What Data Quality Means to Data Consumers)

2.Current status

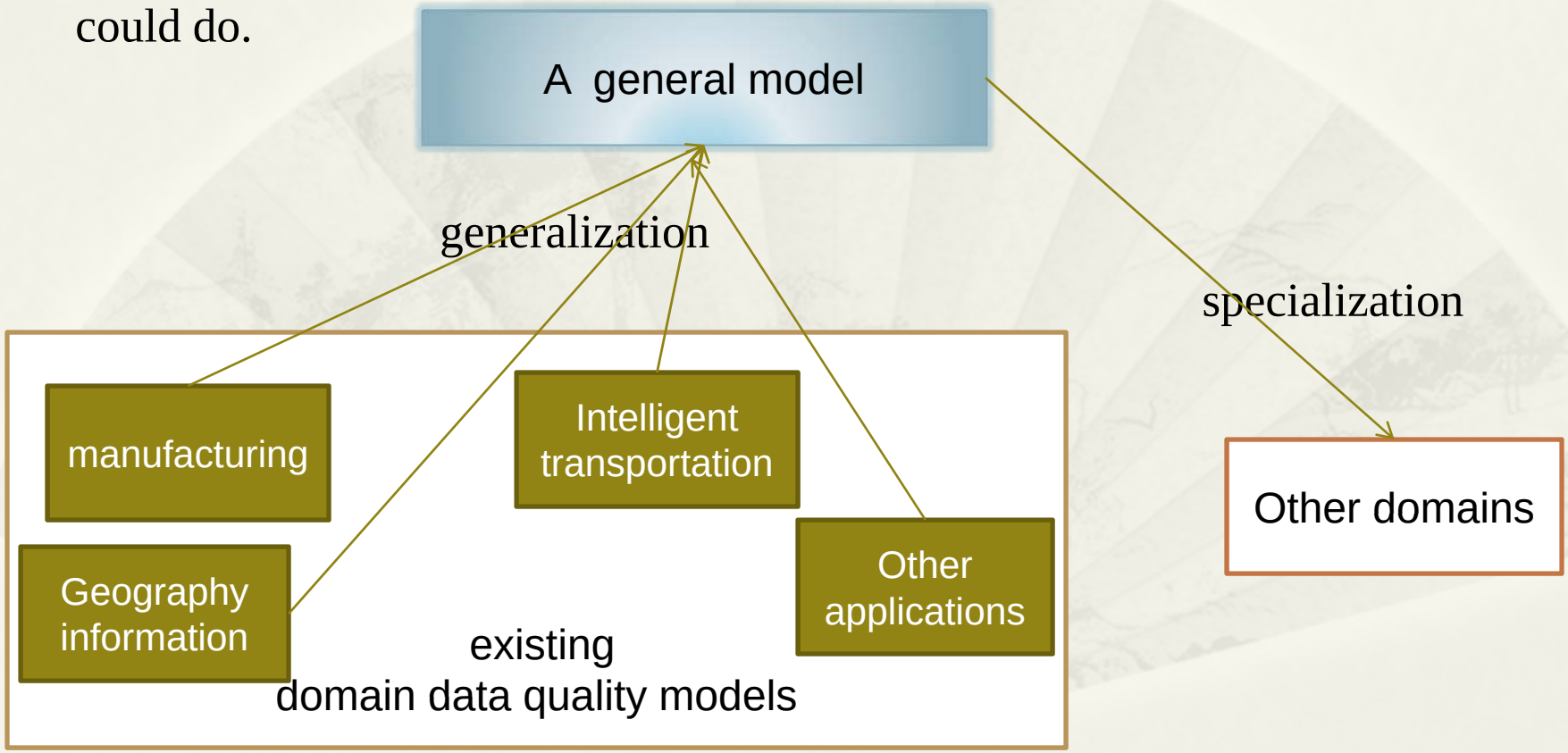
- ISO/TR 21707:2008 identifies a set of parameters or meta-data such as accuracy, precision and timeliness etc.
- Five dimensions--assurances of integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility--of data quality and a set of prerequisites for data quality are the center of the IMF Data Quality Assessment Framework (DQAF).

2.Current status

- 2.5Summary of current status:
  - Some standards on data quality have been developed and applied in specific domains;
  - There is No unified definition of data quality across different domains;
  - There is No general model/framework of data quality independent of any particular domain or application and the existing domain models of data quality differ in many aspects.

3.Future work

There should be something in common on data quality and that is what we could do.



3.Future work

- **To construct a general model/framework (**independent of any specific domain**)of data quality, specifies:**
  - Definition of data quality
  - Dimensions /categories of data quality
  - Attributes in each dimension
  - How to measure these attributes
  - How to control and improve data quality