Information Quality and Entity Resolution-Identification

Planning Conference on Cyber Transportation Logistics
April 3-5, 2008, Winthrop Rockefeller Center

Dr. John R. Talburt, Professor of Information Science and Acxiom Chair of Information Quality
Presentation Agenda

- Why Information Quality
- What is Information Quality
- What is Entity Resolution/Identification
- Overview of the IQ Programs
  - MS, PhD, Certificate
  - Web-Based Distance Learning
Why Information Quality

- The information age depends on information. Everyone has a story to tell about their experiences with poor quality information.

- Information Quality is becoming a more prominent issue in our Transportation and Logistics systems.
Southwest grounds 44 planes

☐ After an "ambiguity related to required testing" was found during a review of records, the airline said.

_CNN.com, March 12, 2008_

☐ The Federal Aviation Administration initiated actions to seek a $10.2 million civil penalty
Warhead fuses mistakenly sent to Taiwan

☐ Fuses for Minuteman nuclear warheads were sent to Taiwan instead of helicopter batteries

*Boston Globe, March 26, 2008*

☐ Mistake discovered only last week - a year and a half after the erroneous shipment
Spinach contaminated with E.coli

☐ Twenty-nine people have been hospitalized, 14 of them with kidney failure.

*USA Today, Sept 15, 2006*

☐ Natural Selection Foods LLC said in a statement that it was cooperating with federal and state health officials to identify the source of the contamination
Information Quality

- The quality of data or information evaluated by its “fitness for purpose” in a particular application
  - The quality of data and the way it is measured vary by application
  - Information quality must be evaluated in context, starting with user requirement and working back to data sources
Big IQ

- Founded on the principles of Total Quality Management (TQM) that has proved successful in manufacturing
  - Deming, Crosby, Juran, et al
  - Culture of quality in the entire enterprise

- Concept of Information Product
  - Information is a product, not by-product
  - \( E = MC^2 \), but IQM = MC^3
  - More than accuracy (multi-dimensional)
Little IQ

- Conformance to Data Model
  - Attribute and relational integrity
  - State transition validation
  - Entity resolution and identity

- Adequacy of the Data Model and Architecture
  - Are problems due to poor data or poor model?
More than Accuracy*

<table>
<thead>
<tr>
<th>DQ Category</th>
<th>DQ Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic DQ</td>
<td>Accuracy, Objectivity, Believability, Reputation</td>
</tr>
<tr>
<td>Accessibility DQ</td>
<td>Access, Security</td>
</tr>
<tr>
<td>Contextual DQ</td>
<td>Relevancy, Value-Added, Timeliness, Completeness, Amount of data</td>
</tr>
<tr>
<td>Representational DQ</td>
<td>Interpretability, Ease of understanding, Concise representation, Consistent representation</td>
</tr>
</tbody>
</table>

*Wang – Strong Framework*
Access

A 2004 survey shows that employees in an average business environment spend 30% of their time searching for information that is relevant to the project they are working on.

IBM Industry Study, Customer Interviews
Consistent Representation

NASA’s Mars Climate Orbiter was lost in space last week because engineers failed to make a simple conversion from English units to metric (loss of $125 Million)

Washington Post, October 1, 1999
Completeness

In analyzing the relation between temperature and O-ring failures prior to launch, NASA excluded observations where no O-ring failed believing they were unimportant.

The Challenger Space Shuttle Disaster of January 1986
Reputation

Chase Card Services send a credit card offer to a man, Sami Habbas, addressed to him as “Palestinian Bomber”

AP Wire Service, 8/24/2005
Accuracy

A trader with Mizuho Financial Group in Japan entered an order to sell one share of J-Com at 610,000 Yen as an order to sell 610,000 shares at 1 Yen causing a loss of $224 million

Entity & Identity Resolution
What is Entity Resolution?

- The process in which records determined to represent the same real-world entity are successively located and merged
- Special form of data integration
- “Real-World” database join
- No primary keys
Which belong together?

J. Talburt
UALR T00090575

J Talburt
White Dr, Batesville, AR

John Talburt
700 Pierce, LR, AR

John Talburt
Hazelton Rd, Pea Ridge, AR

J R Talburt
Old Country Ln, Cabot, AR
Which belong together?

- J. Talburt
  UALR T00090575

- J Talburt
  White Dr, Batesville, AR

- John Talburt
  700 Pierce, LR, AR

- John Talburt
  Hazelton Rd, Pea Ridge, AR

- J R Talburt
  Old Country Ln, Cabot, AR

- John Talburt
  Wood Sorrell Pt, LR, AR
  xxx-xx-1519

- Jerry Talburt

- John Talburt
  DOB 3/1961
Entity Identification

- Identity
  - A cluster of attributes sufficient to distinguish one entity from all others

- Entity Identification
  - Entity Resolution where records are associated with a particular identity

- Entity Disambiguation
  - Differentiation of entity references – does not require identification
  - Like Crime Scene Investigation
Related to Information Quality

- Deals with data
  - Containing inaccuracies
  - That is incomplete
  - Inconsistent in format
  - Inconsistent in content
  - From different time periods
  - Not easily understood
  - Varying degrees of believability
Locating Entities

- Entity extraction from unstructured sources is one of the fastest growing areas of research and practice in the field of Entity Resolution.
- As much as 80% of enterprise data is stored in unstructured sources.
  - Additional relevant information
  - Risk management
UALR Master of Science in Information Quality (MSIQ)
MSIQ Objective

- To meet the growing demands by government and industry for professionals with graduate-level education that covers:
  - Skills essential for defining, measuring, analyzing, and improving information quality
  - Principles for developing information quality policies, strategies, and programs
  - Relationships between information quality and other issues like data privacy, protection, data governance, architecture...
History

- ADHE Approval – April 28, 2006
- First Classes Offered – Fall 2006
  - 25 Students enrolled in 4 Courses
- First Graduate (Chris Heien) August 2007
- Distance Education via Internet Begins
  - 45 Students enrolled (5 remote)
- Five Graduates for December 2007
  - 48 Students enrolled (8 remote)
<table>
<thead>
<tr>
<th>Information Quality</th>
<th>Information Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro to IQ</td>
<td>Data Visualization</td>
</tr>
<tr>
<td>TQM &amp; Statistical QC</td>
<td>Info Systems Analysis</td>
</tr>
<tr>
<td>IQ Tools &amp; Industry</td>
<td>Database Systems</td>
</tr>
<tr>
<td>IQ Theory</td>
<td>IS Elective</td>
</tr>
<tr>
<td>IQ Elective</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thesis or Project (6 hrs)</td>
</tr>
</tbody>
</table>
Example Projects

- Showing Information Quality Improvement within Medical Records by Establishing an Electronic Storage System (*Birch Tree Communities*)
- Every Child by Name: Applying Information Quality Principles to Average Daily Membership (*Ark Dept Ed*)
- Improving Information Quality for the Grants and People Database for the Arkansas Arts Council (*Ark Arts Council*)
Example Projects

- Impact of Quality of Major Source Replacement in a Consumer Data Product (Acxiom Corp)
- International Survey on Information Quality and Data Governance Practices (co-sponsored by the International Association for Information and Data Quality (IAIDQ))
- Fusion Center Data Integration (Ark State Police)
UALR Applied Science PhD Information Quality Emphasis
Started Spring 2008

- First students admitted Jan 2008

- Curriculum
  - 72 hours beyond baccalaureate degree
  - Min 18 hours of course-work
  - Min 42 hours doctoral dissertation
UALR Graduate Certificate in Information Quality

Approved by the Ark Dept of Higher Education February 12, 2008
Will Begin Fall 2008

Features

- GRE/GMAT not required for admission
- 12 hours of coursework
- All courses can be directly applied toward the MSIQ or Applied Science PhD program
- Gateway to the Master of Science degree
Program Sponsors

- Acxiom Corporation (Little Rock, AR)
- Citizant, Inc. (Chantilly, VA)
- Flake-Wilkerson (Little Rock, AR)
- FUZZY! Informatik (Ludwigsburg, Germany)
- MIT Information Quality Program (Cambridge, MA)
- SAS Institute (Cary, NC)
Blended Approach to Distance Education
Blended Web Experience

- Programs are open to either full or part-time students
- Classes accessible on the UALR campus in Little Rock, Arkansas or in remote locations via broad-band Internet
- Classes are scheduled in the late afternoon and evening, one day per week to accommodate working professionals
Online Support

- Support and resources for all students available online
  - Admission
  - Registration
  - Bookstore
  - Class materials
  - Technical support...
Same Standards, Local/Remote

- All students - whether learning via the web or in the classroom, are held to the same performance standards
- Local students have the option to attend classes in-person or remotely – accommodates travel, illness, inclement weather,...
Proctored Examinations

- Local students take exams in class
- Remote students must arrange an approved proctor for scheduled exams
- Proctors may include librarians, testing center personnel, university faculty, corporate/military education officers
Advantages: Blended Experience

- Students can better organize their lecture and learning activities around their schedules.
- Students whether local or remote have access to state-of-the-art web instruction and online campus services.
Asynchronous and Synchronous

The WebCT classroom is an online learning environment that offers students and the instructor:

- Ability to distribute online course content including syllabi, student readings, learning modules, class calendar, and announcements
- Ability to distribute, collect, and grade assignments and quizzes online
INFQ 7318 - Total Quality Management - Fall 2006 section 02

Your location: Home Page

Syllabus
The Syllabus for this course contains general information about the instructor and the course as well as a description of the policies governing the course this semester.

Student Readings
This folder contains a list of articles on information quality.

Information Quality Web Links
This folder contains a list of websites dedicated to Information Quality.

Exams and Quizzes
This folder contains links to the exams and quizzes that will be due this semester.

How can we make this course (and program) better for you?
Under this topic please feel free to post your ideas and suggestions for making this course and the MSIQ program a better fit for your needs.

Learning Modules
This folder contains the links to all the learning modules that will be covered this semester.

Student Exercises
This folder contains the solutions for each of the chapter exercises.

Assignments
This folder contains links to the assignments that will be due this semester.
Wimba Live Classroom

- Allows students both online and physically in the classroom to
  - Participate in two-way audio exchanges in the virtual/physical classroom
  - See all materials presented on the instructor’s desktop through the Online Virtual Classroom
- Accommodates remote lectures and demonstrations
Frequently Asked Questions

- What is this course about?
- What resources will I need for this course?
- How will I be graded?
- How will the course be delivered?
- Where can I get help with the course?
- What's next?
- Any other questions?
To Find Out More

- To learn more about what the MSIQ has to offer, check out our website at [http://technologize.ualr.edu/msiq/](http://technologize.ualr.edu/msiq/)
- Applications for the MSIQ program can be done online today!
- Questions? Contact Dr. John Talburt, Acxiom Chair of Information Quality: [jrtalburt@ualr.edu](mailto:jrtalburt@ualr.edu) or (501) 371-7616