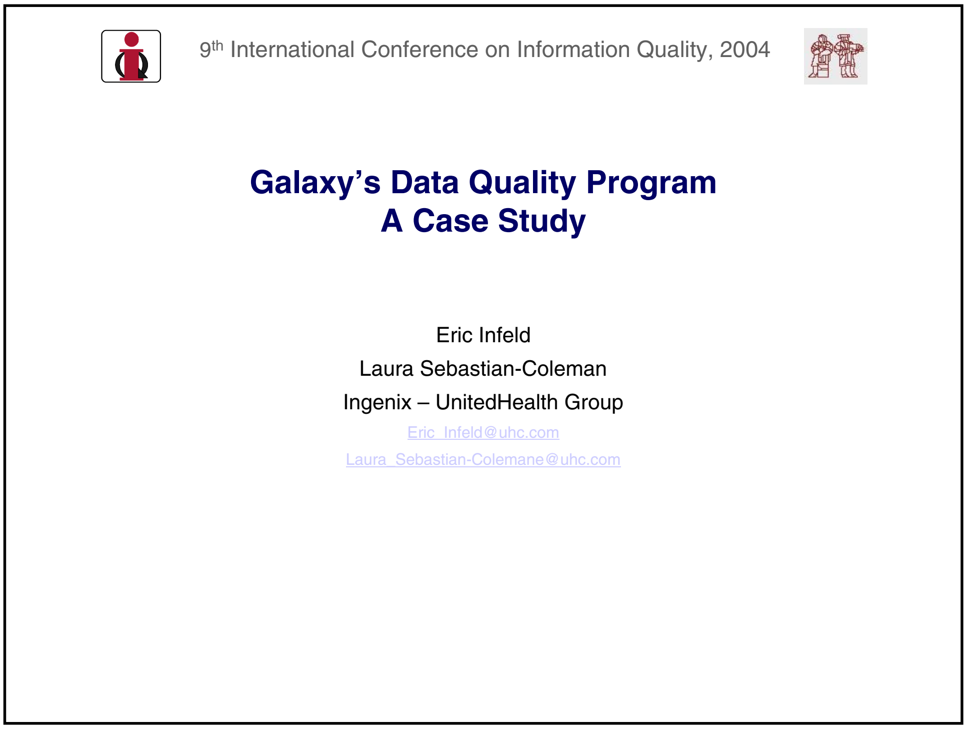
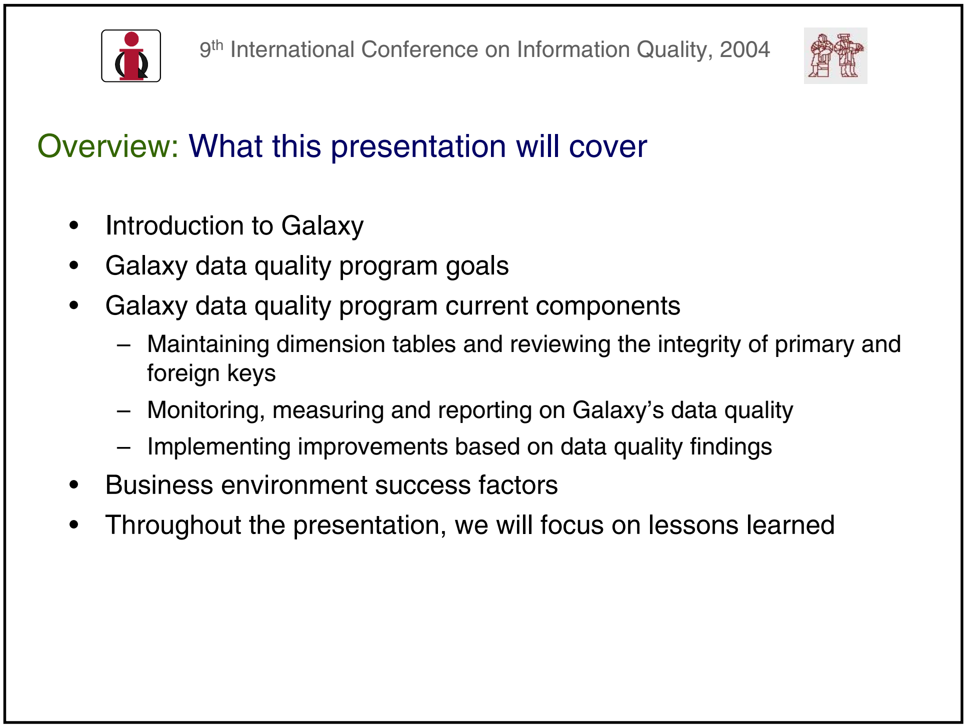


9th International Conference on Information Quality, 2004

Galaxy's Data Quality Program
A Case Study

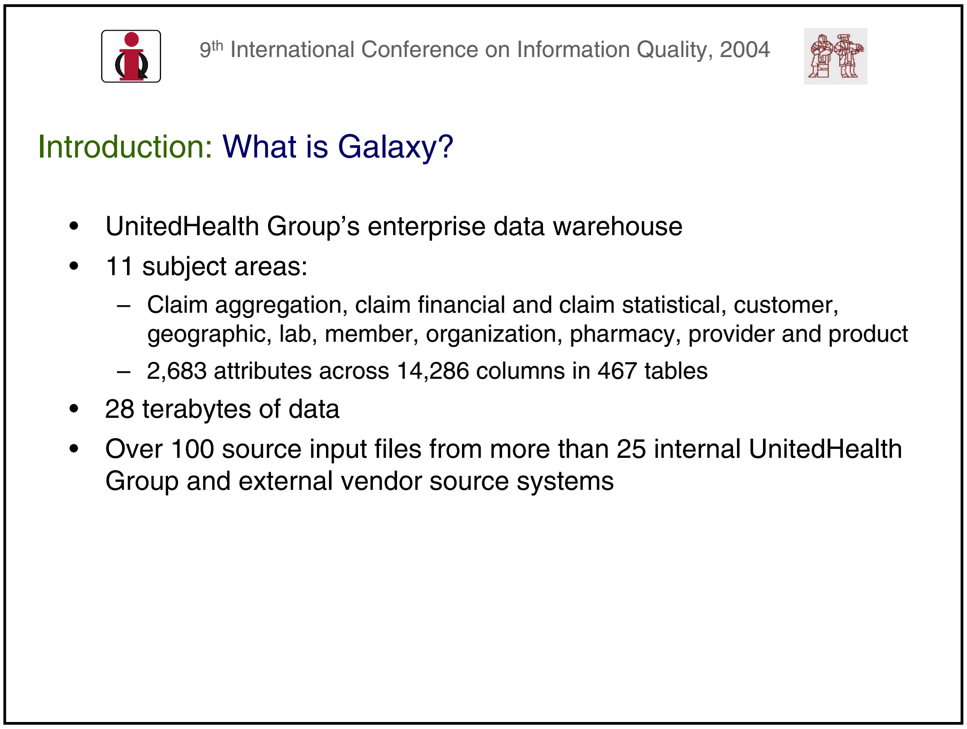
Eric Infeld
Laura Sebastian-Coleman
Ingenix – UnitedHealth Group
Eric_Infeld@uhc.com
Laura_Sebastian-Colemane@uhc.com



9th International Conference on Information Quality, 2004

Overview: What this presentation will cover

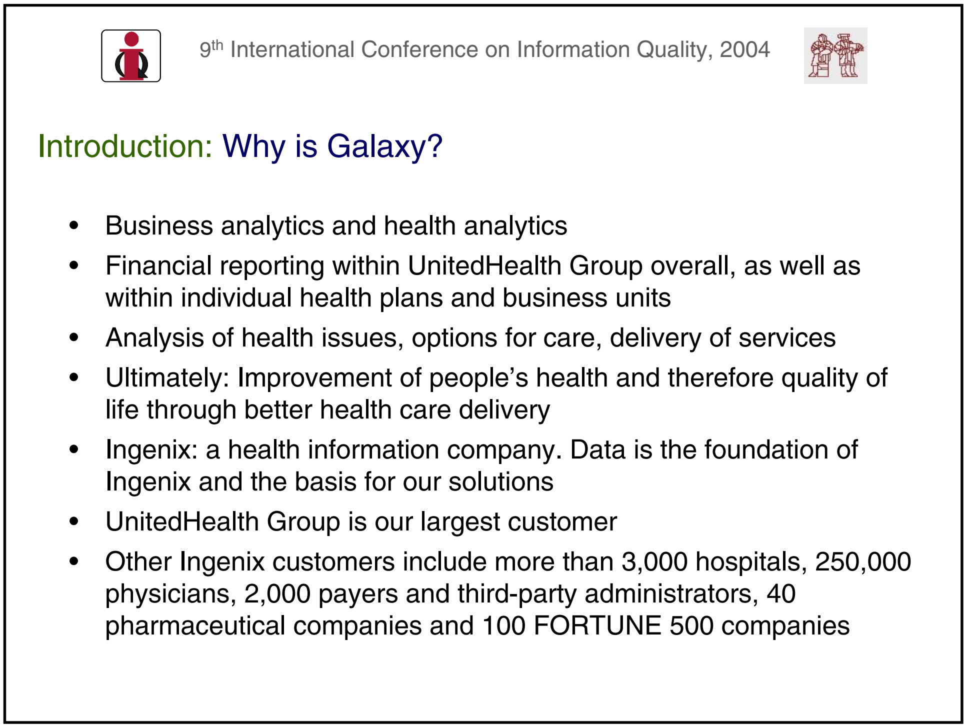
- Introduction to Galaxy
- Galaxy data quality program goals
- Galaxy data quality program current components
  - Maintaining dimension tables and reviewing the integrity of primary and foreign keys
  - Monitoring, measuring and reporting on Galaxy's data quality
  - Implementing improvements based on data quality findings
- Business environment success factors
- Throughout the presentation, we will focus on lessons learned



9th International Conference on Information Quality, 2004

Introduction: What is Galaxy?

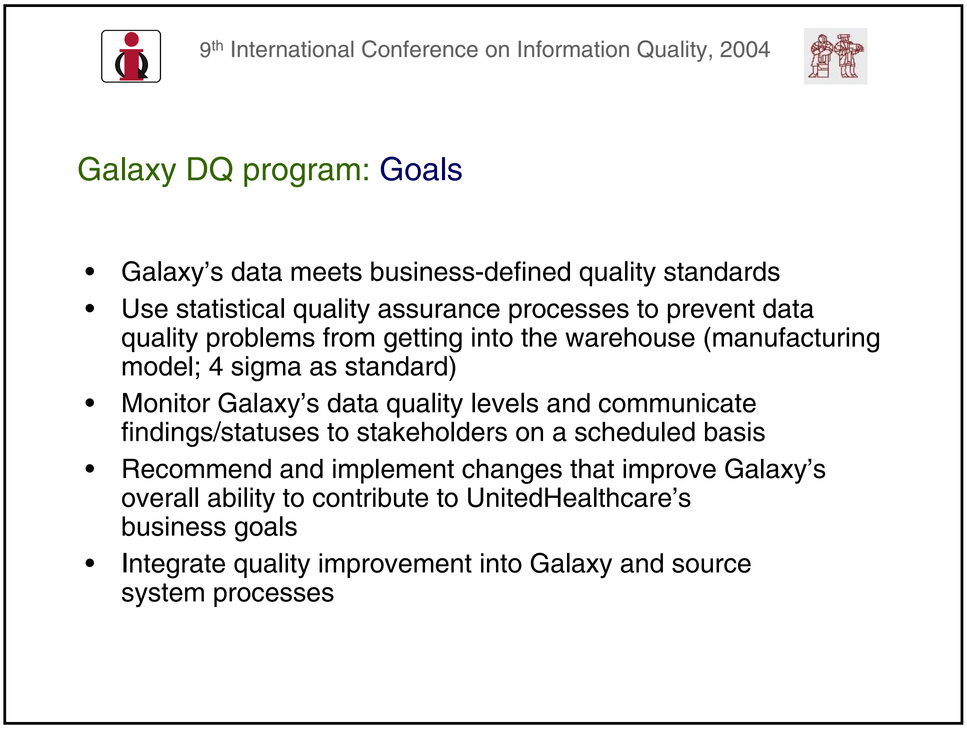
- UnitedHealth Group’s enterprise data warehouse
- 11 subject areas:
  - Claim aggregation, claim financial and claim statistical, customer, geographic, lab, member, organization, pharmacy, provider and product
  - 2,683 attributes across 14,286 columns in 467 tables
- 28 terabytes of data
- Over 100 source input files from more than 25 internal UnitedHealth Group and external vendor source systems



9th International Conference on Information Quality, 2004

Introduction: Why is Galaxy?

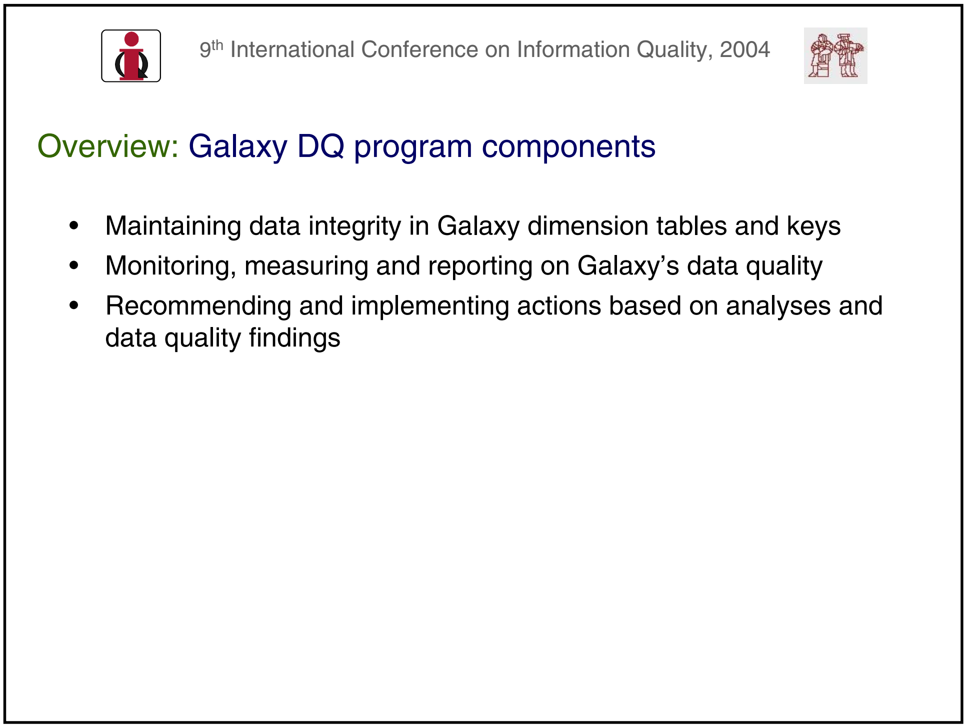
- Business analytics and health analytics
- Financial reporting within UnitedHealth Group overall, as well as within individual health plans and business units
- Analysis of health issues, options for care, delivery of services
- Ultimately: Improvement of people’s health and therefore quality of life through better health care delivery
- Ingenix: a health information company. Data is the foundation of Ingenix and the basis for our solutions
- UnitedHealth Group is our largest customer
- Other Ingenix customers include more than 3,000 hospitals, 250,000 physicians, 2,000 payers and third-party administrators, 40 pharmaceutical companies and 100 FORTUNE 500 companies



9th International Conference on Information Quality, 2004

Galaxy DQ program: Goals

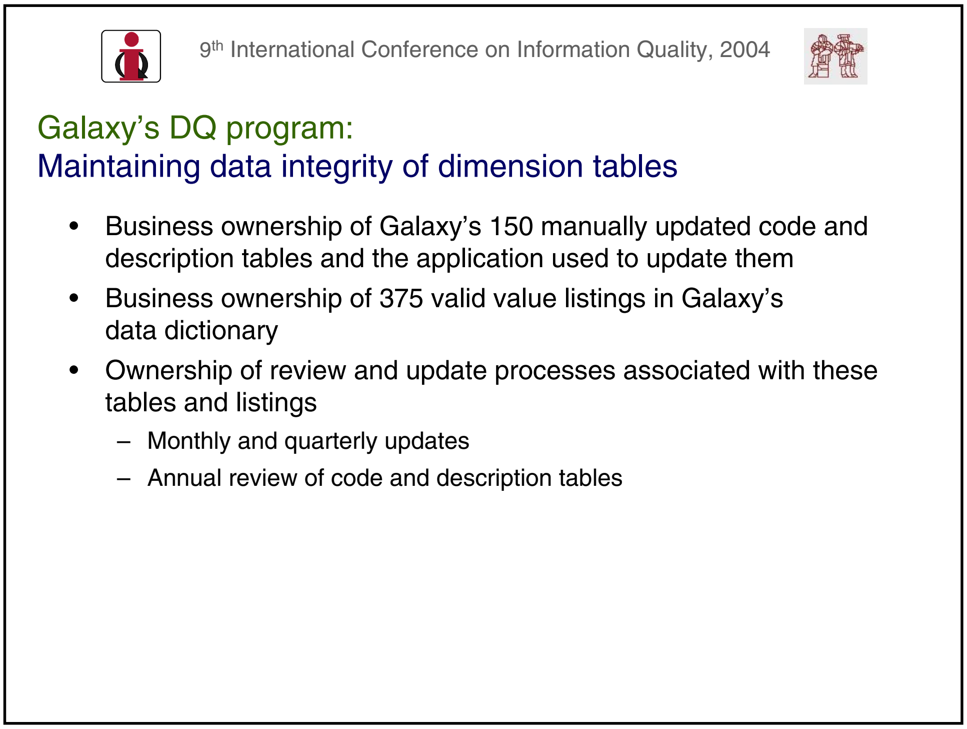
- Galaxy’s data meets business-defined quality standards
- Use statistical quality assurance processes to prevent data quality problems from getting into the warehouse (manufacturing model; 4 sigma as standard)
- Monitor Galaxy’s data quality levels and communicate findings/statuses to stakeholders on a scheduled basis
- Recommend and implement changes that improve Galaxy’s overall ability to contribute to UnitedHealthcare’s business goals
- Integrate quality improvement into Galaxy and source system processes



9th International Conference on Information Quality, 2004

Overview: Galaxy DQ program components

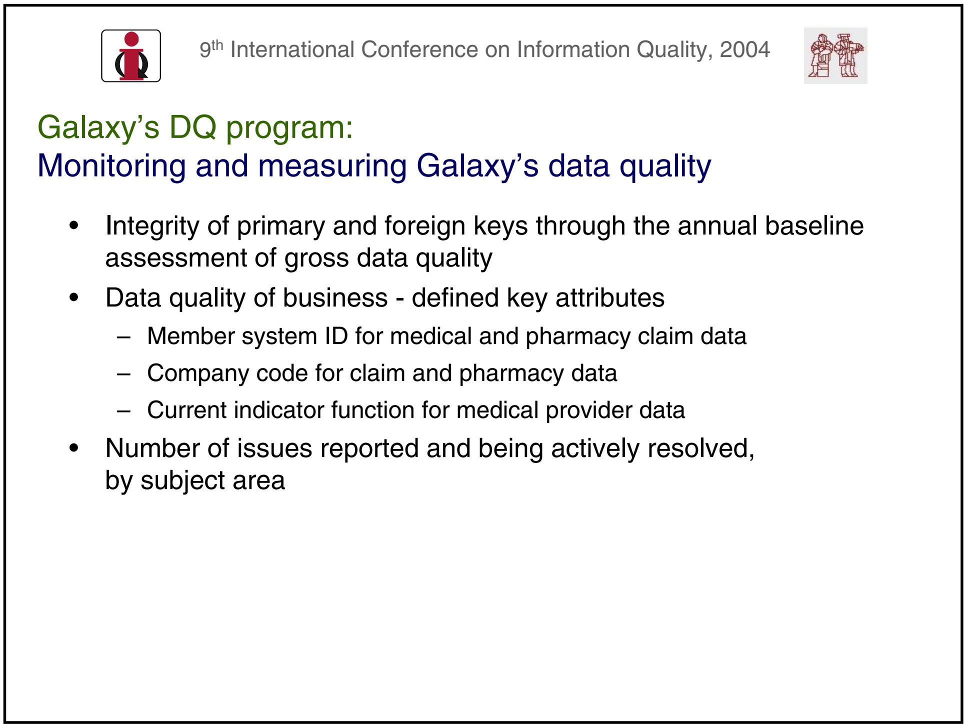
- Maintaining data integrity in Galaxy dimension tables and keys
- Monitoring, measuring and reporting on Galaxy’s data quality
- Recommending and implementing actions based on analyses and data quality findings



9th International Conference on Information Quality, 2004

Galaxy's DQ program:
Maintaining data integrity of dimension tables

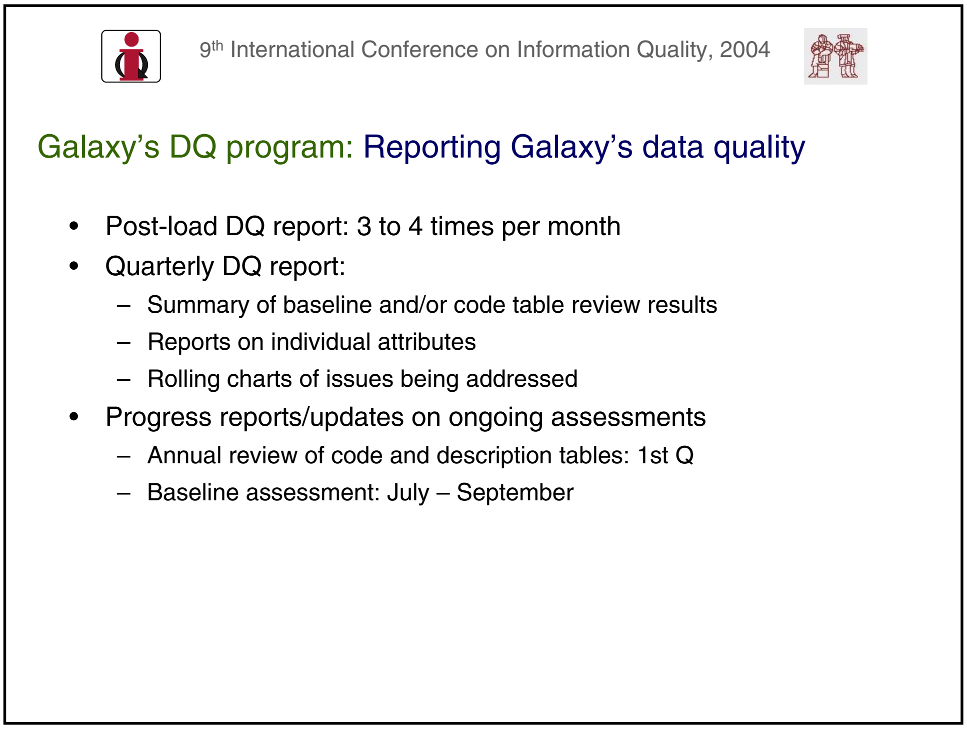
- Business ownership of Galaxy's 150 manually updated code and description tables and the application used to update them
- Business ownership of 375 valid value listings in Galaxy's data dictionary
- Ownership of review and update processes associated with these tables and listings
  - Monthly and quarterly updates
  - Annual review of code and description tables



9th International Conference on Information Quality, 2004

Galaxy's DQ program:
Monitoring and measuring Galaxy's data quality

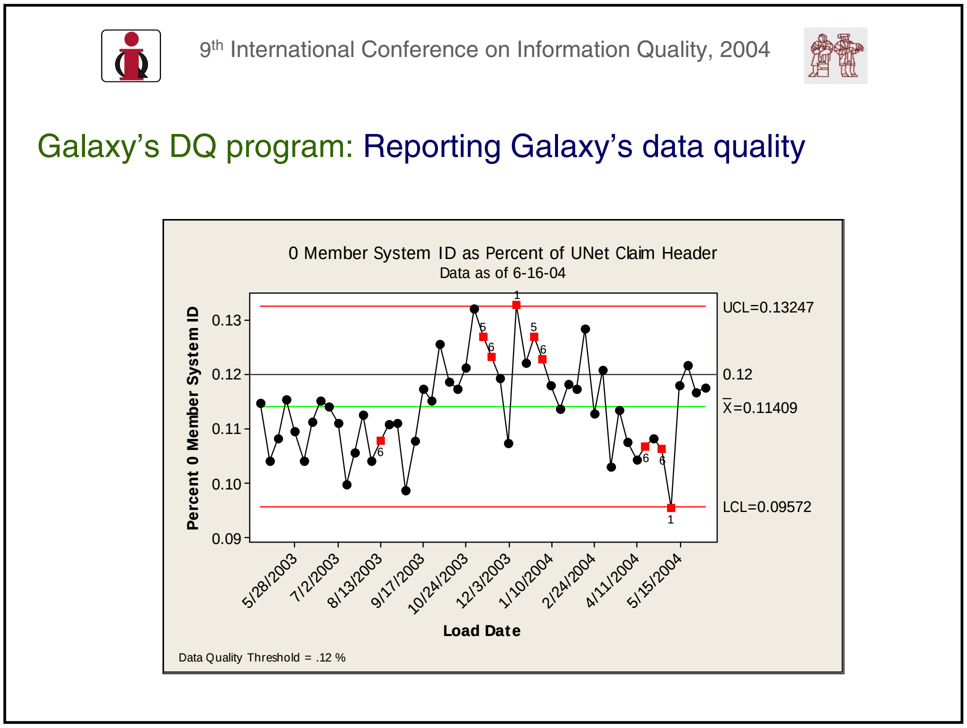
- Integrity of primary and foreign keys through the annual baseline assessment of gross data quality
- Data quality of business - defined key attributes
  - Member system ID for medical and pharmacy claim data
  - Company code for claim and pharmacy data
  - Current indicator function for medical provider data
- Number of issues reported and being actively resolved, by subject area



9th International Conference on Information Quality, 2004

Galaxy’s DQ program: Reporting Galaxy’s data quality

- Post-load DQ report: 3 to 4 times per month
- Quarterly DQ report:
  - Summary of baseline and/or code table review results
  - Reports on individual attributes
  - Rolling charts of issues being addressed
- Progress reports/updates on ongoing assessments
  - Annual review of code and description tables: 1st Q
  - Baseline assessment: July – September

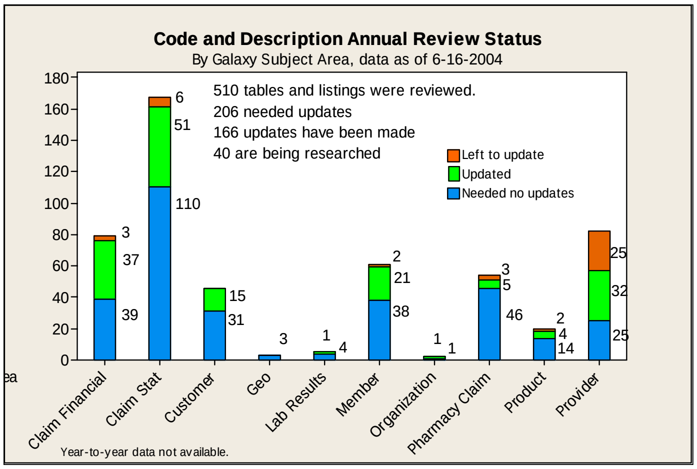


9th International Conference on Information Quality, 2004

Galaxy’s DQ program: Reporting Galaxy’s data quality

9th International Conference on Information Quality, 2004

Galaxy's DQ program: Reporting Galaxy's data quality



Code and Description Annual Review Status

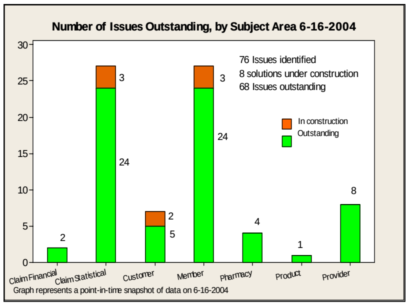
By Galaxy Subject Area, data as of 6-16-2004

510 tables and listings were reviewed.
206 needed updates
166 updates have been made
40 are being researched

Year-to-year data not available.

9th International Conference on Information Quality, 2004

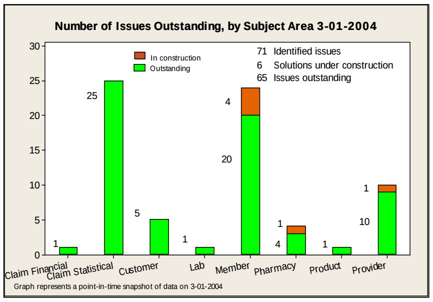
Galaxy's DQ program: Reporting Galaxy's data quality



Number of Issues Outstanding, by Subject Area 6-16-2004

76 Issues identified
8 solutions under construction
68 Issues outstanding

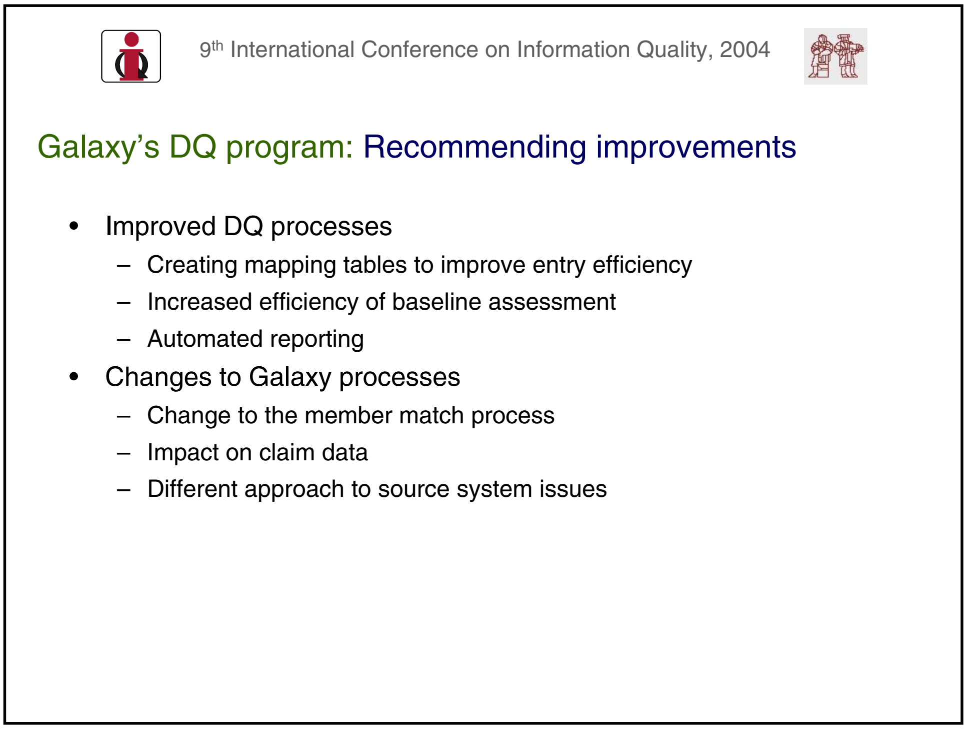
Graph represents a point-in-time snapshot of data on 6-16-2004



Number of Issues Outstanding, by Subject Area 3-01-2004

71 Identified issues
6 Solutions under construction
65 Issues outstanding

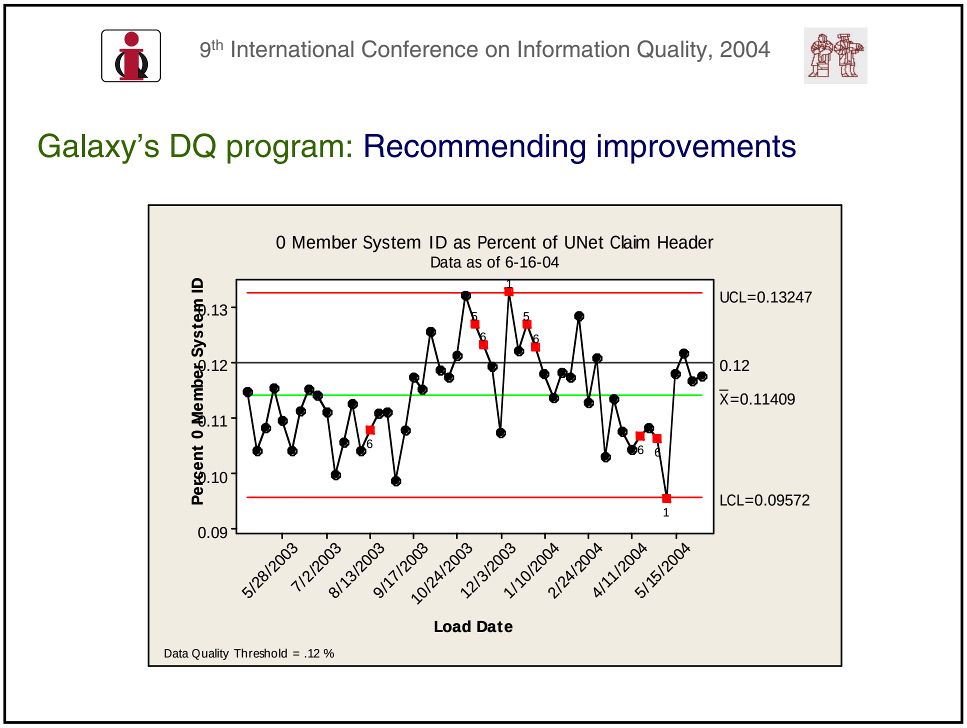
Graph represents a point-in-time snapshot of data on 3-01-2004



9th International Conference on Information Quality, 2004

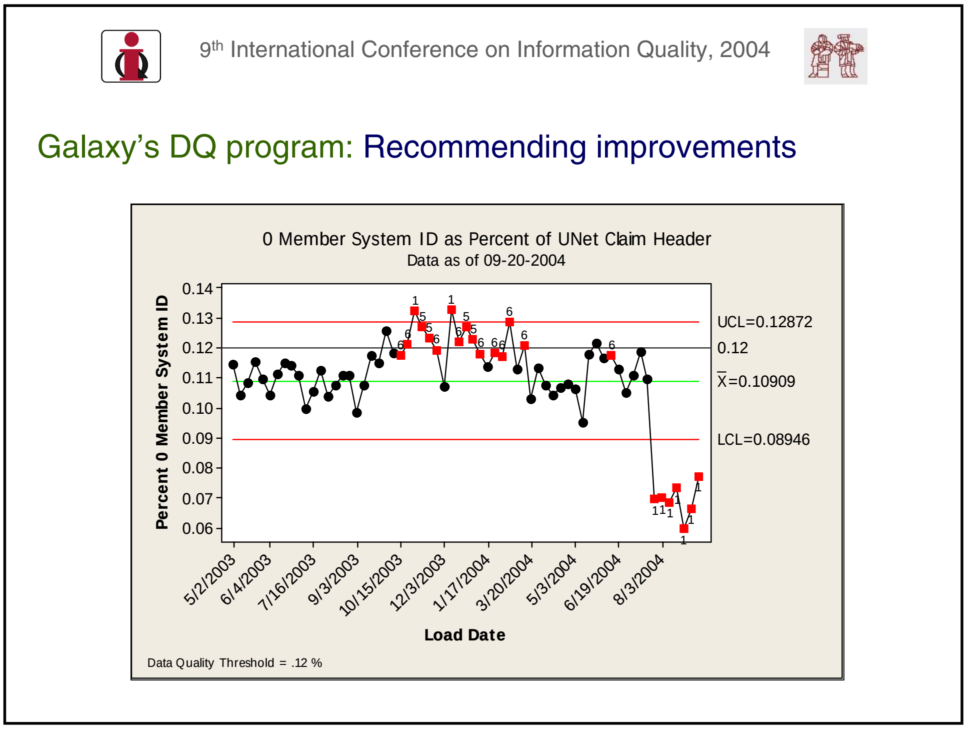
Galaxy's DQ program: Recommending improvements

- Improved DQ processes
  - Creating mapping tables to improve entry efficiency
  - Increased efficiency of baseline assessment
  - Automated reporting
- Changes to Galaxy processes
  - Change to the member match process
  - Impact on claim data
  - Different approach to source system issues



9th International Conference on Information Quality, 2004

Galaxy's DQ program: Recommending improvements



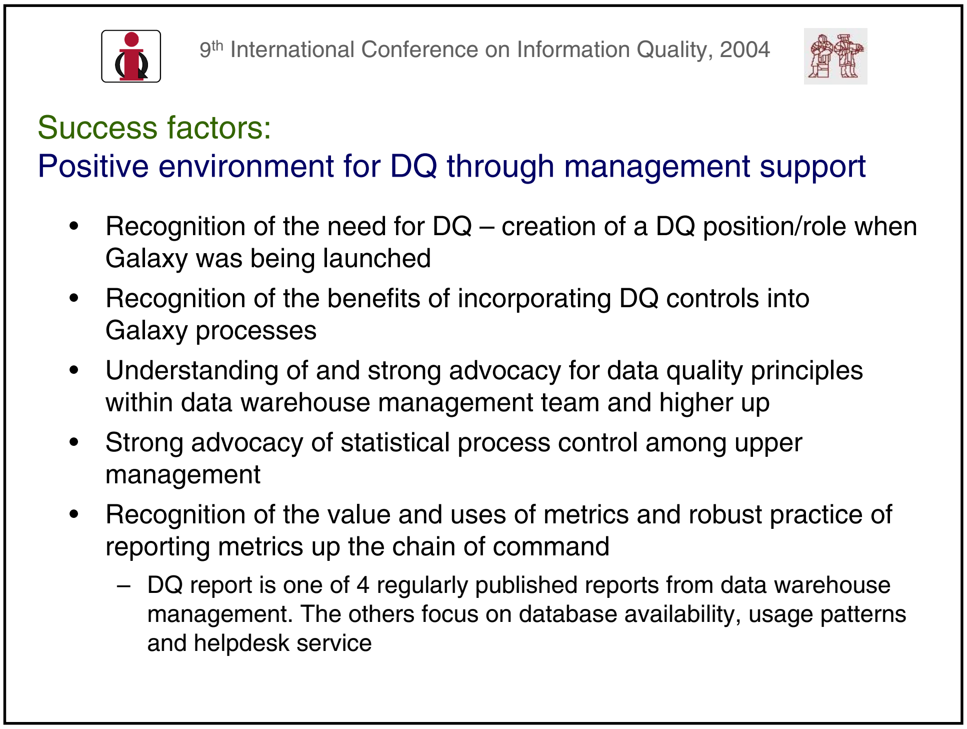
9th International Conference on Information Quality, 2004

Galaxy's DQ program: Recommending improvements

9th International Conference on Information Quality, 2004

Overview: Success factors in the business environment

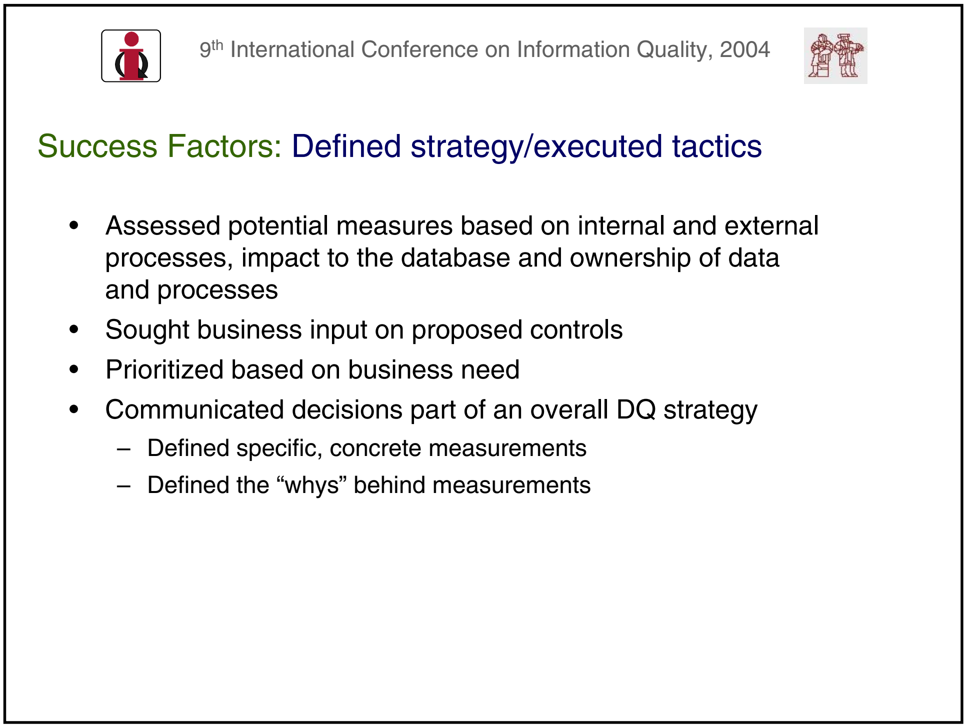
- Management support and team knowledge
- Defining strategy/executing tactics
- Building credibility



9th International Conference on Information Quality, 2004

Success factors: Positive environment for DQ through management support

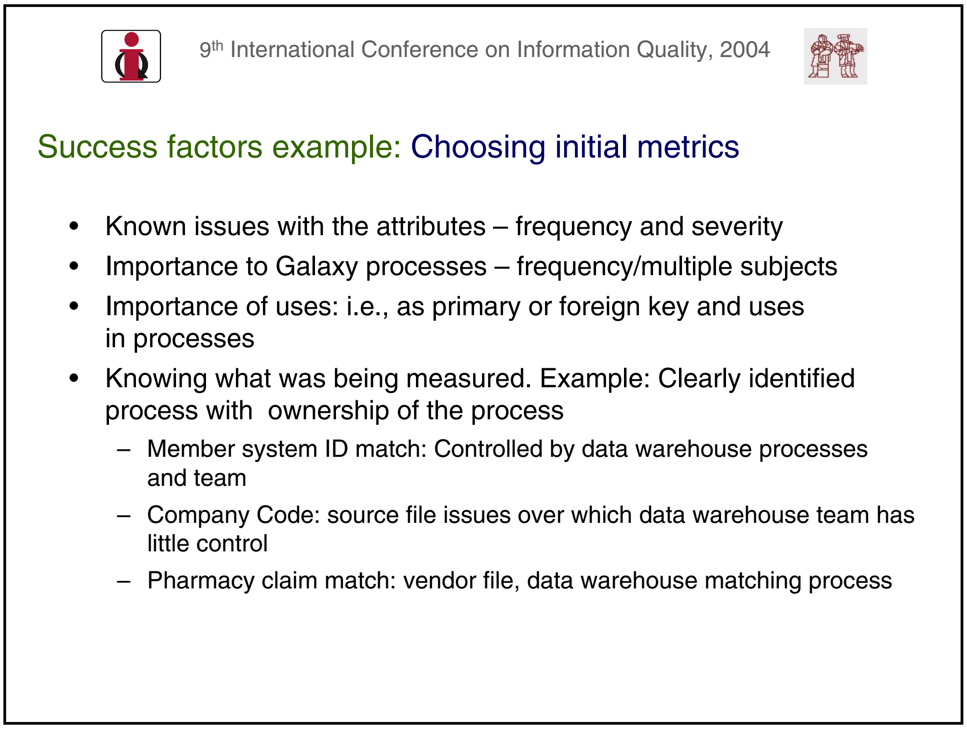
- Recognition of the need for DQ – creation of a DQ position/role when Galaxy was being launched
- Recognition of the benefits of incorporating DQ controls into Galaxy processes
- Understanding of and strong advocacy for data quality principles within data warehouse management team and higher up
- Strong advocacy of statistical process control among upper management
- Recognition of the value and uses of metrics and robust practice of reporting metrics up the chain of command
  - DQ report is one of 4 regularly published reports from data warehouse management. The others focus on database availability, usage patterns and helpdesk service



9th International Conference on Information Quality, 2004

Success Factors: Defined strategy/executed tactics

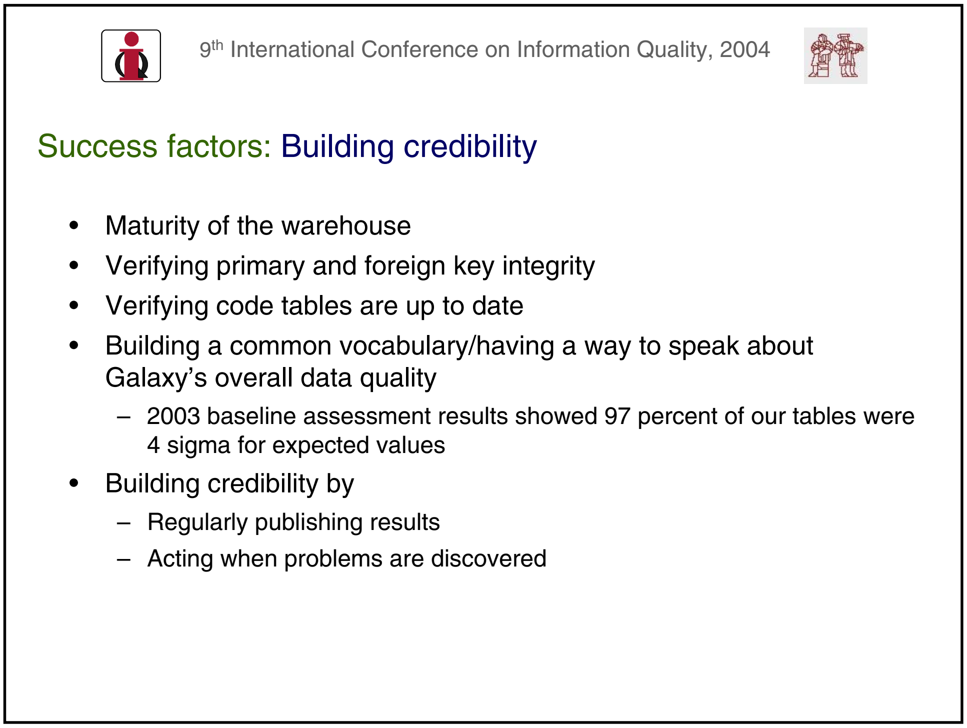
- Assessed potential measures based on internal and external processes, impact to the database and ownership of data and processes
- Sought business input on proposed controls
- Prioritized based on business need
- Communicated decisions part of an overall DQ strategy
  - Defined specific, concrete measurements
  - Defined the “whys” behind measurements



9th International Conference on Information Quality, 2004

Success factors example: Choosing initial metrics

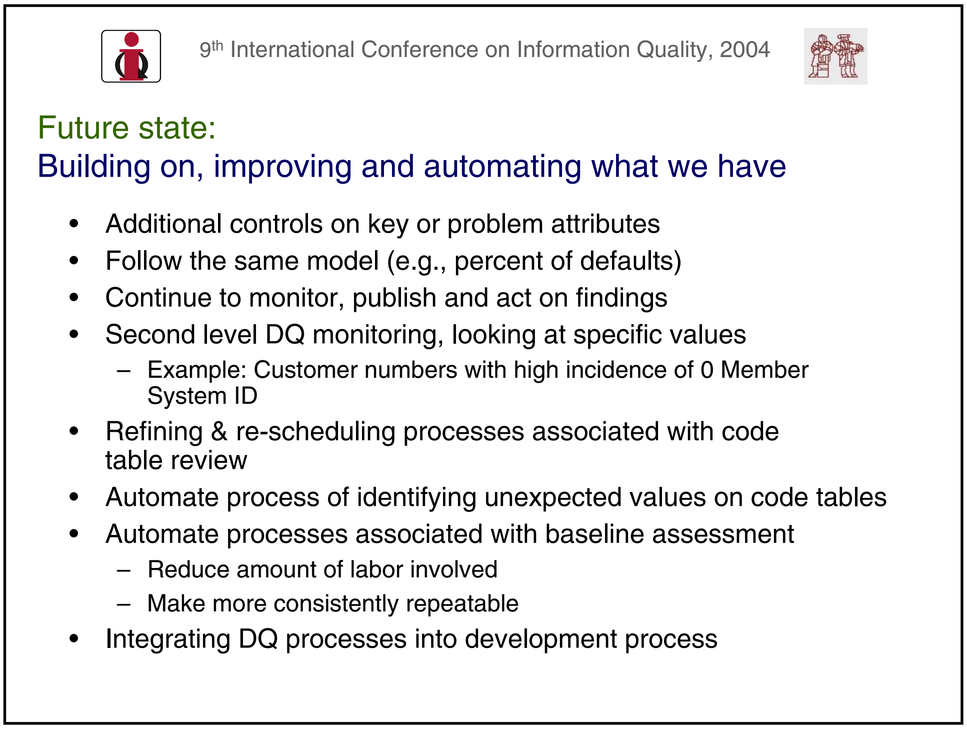
- Known issues with the attributes – frequency and severity
- Importance to Galaxy processes – frequency/multiple subjects
- Importance of uses: i.e., as primary or foreign key and uses in processes
- Knowing what was being measured. Example: Clearly identified process with ownership of the process
  - Member system ID match: Controlled by data warehouse processes and team
  - Company Code: source file issues over which data warehouse team has little control
  - Pharmacy claim match: vendor file, data warehouse matching process



9th International Conference on Information Quality, 2004

Success factors: Building credibility

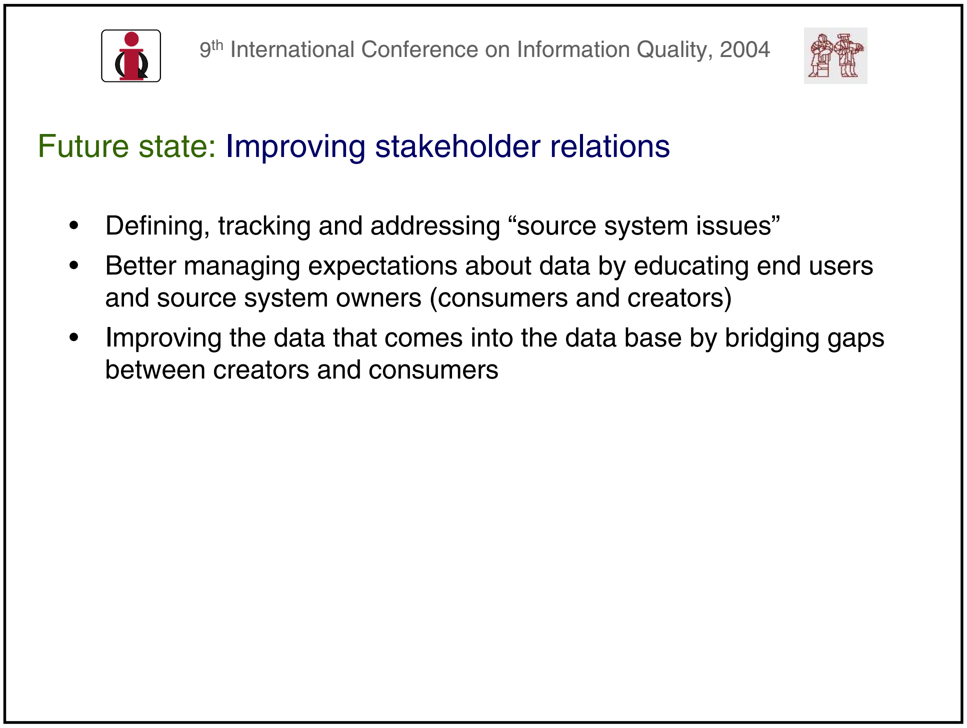
- Maturity of the warehouse
- Verifying primary and foreign key integrity
- Verifying code tables are up to date
- Building a common vocabulary/having a way to speak about Galaxy’s overall data quality
  - 2003 baseline assessment results showed 97 percent of our tables were 4 sigma for expected values
- Building credibility by
  - Regularly publishing results
  - Acting when problems are discovered



9th International Conference on Information Quality, 2004

Future state:
Building on, improving and automating what we have

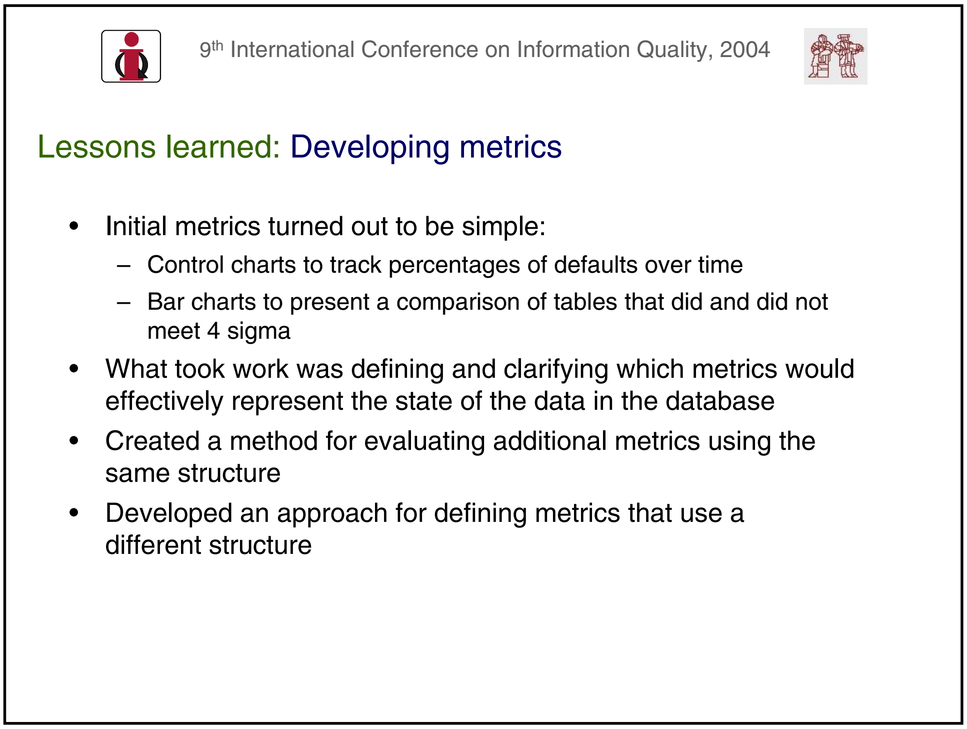
- Additional controls on key or problem attributes
- Follow the same model (e.g., percent of defaults)
- Continue to monitor, publish and act on findings
- Second level DQ monitoring, looking at specific values
  - Example: Customer numbers with high incidence of 0 Member System ID
- Refining & re-scheduling processes associated with code table review
- Automate process of identifying unexpected values on code tables
- Automate processes associated with baseline assessment
  - Reduce amount of labor involved
  - Make more consistently repeatable
- Integrating DQ processes into development process



9th International Conference on Information Quality, 2004

Future state: Improving stakeholder relations

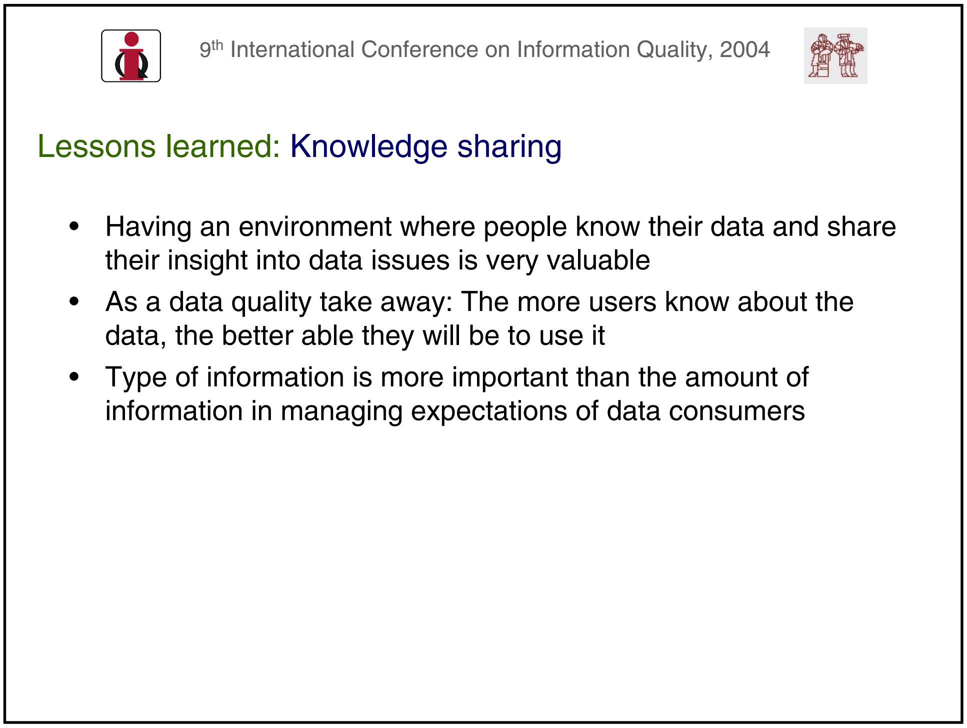
- Defining, tracking and addressing “source system issues”
- Better managing expectations about data by educating end users and source system owners (consumers and creators)
- Improving the data that comes into the data base by bridging gaps between creators and consumers



9th International Conference on Information Quality, 2004

Lessons learned: Developing metrics

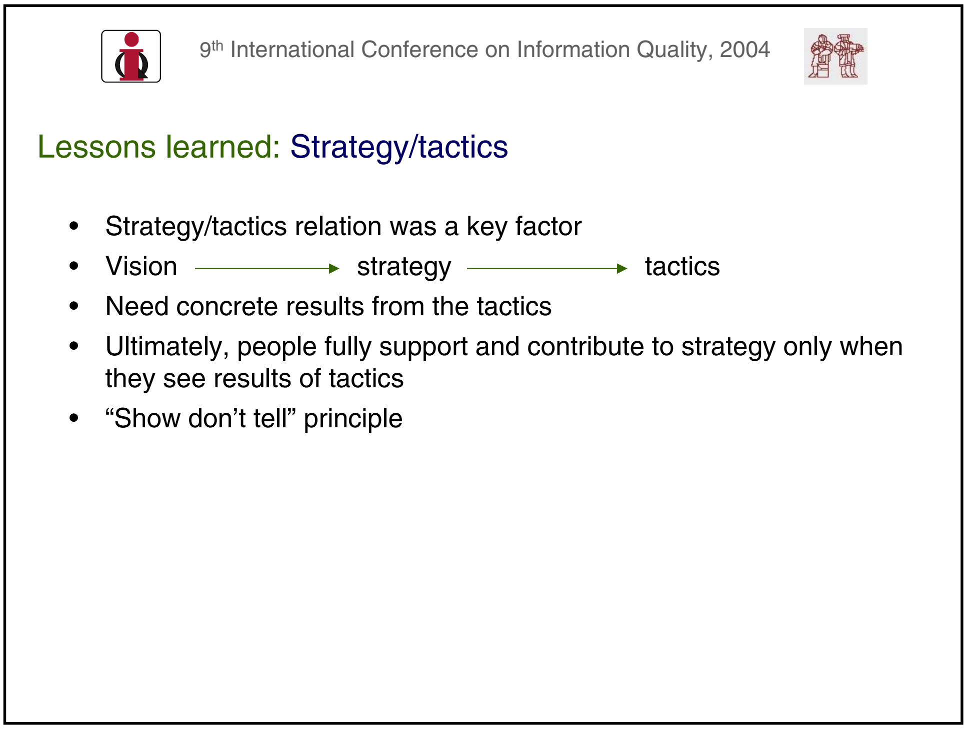
- Initial metrics turned out to be simple:
  - Control charts to track percentages of defaults over time
  - Bar charts to present a comparison of tables that did and did not meet 4 sigma
- What took work was defining and clarifying which metrics would effectively represent the state of the data in the database
- Created a method for evaluating additional metrics using the same structure
- Developed an approach for defining metrics that use a different structure



9th International Conference on Information Quality, 2004

Lessons learned: Knowledge sharing

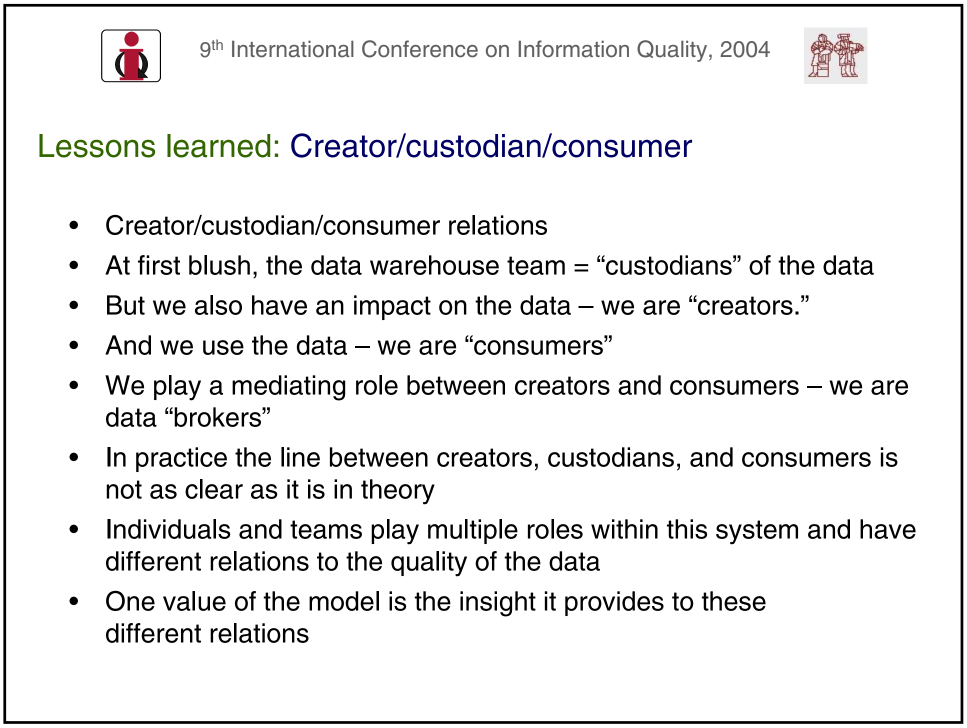
- Having an environment where people know their data and share their insight into data issues is very valuable
- As a data quality take away: The more users know about the data, the better able they will be to use it
- Type of information is more important than the amount of information in managing expectations of data consumers



9th International Conference on Information Quality, 2004

Lessons learned: Strategy/tactics

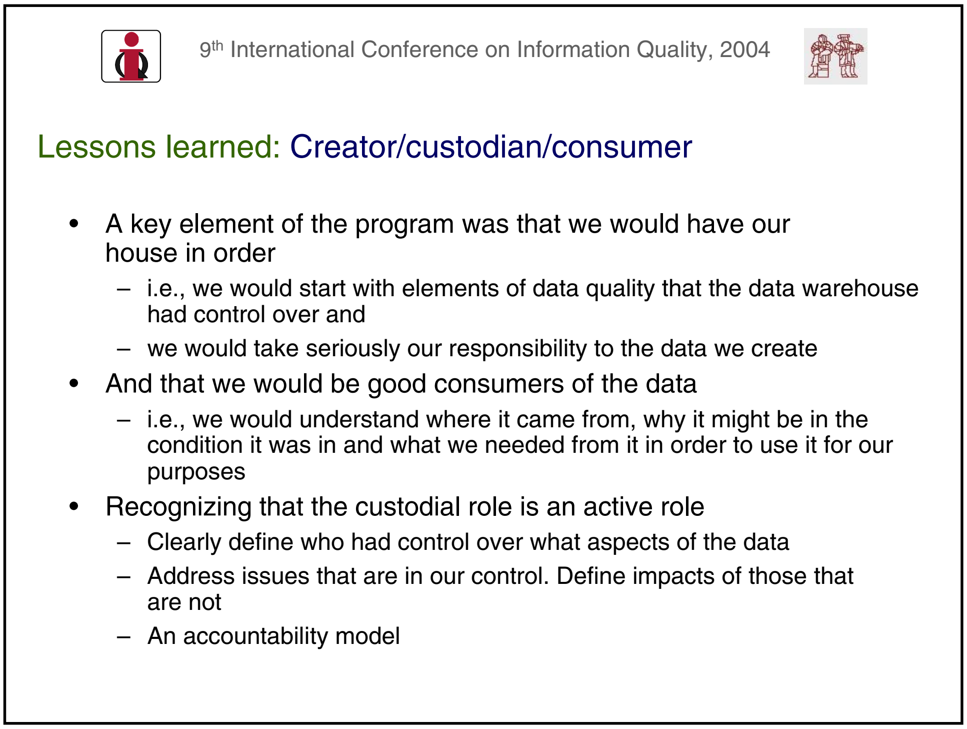
- Strategy/tactics relation was a key factor
- Vision → strategy → tactics
- Need concrete results from the tactics
- Ultimately, people fully support and contribute to strategy only when they see results of tactics
- “Show don’t tell” principle



9th International Conference on Information Quality, 2004

Lessons learned: Creator/custodian/consumer

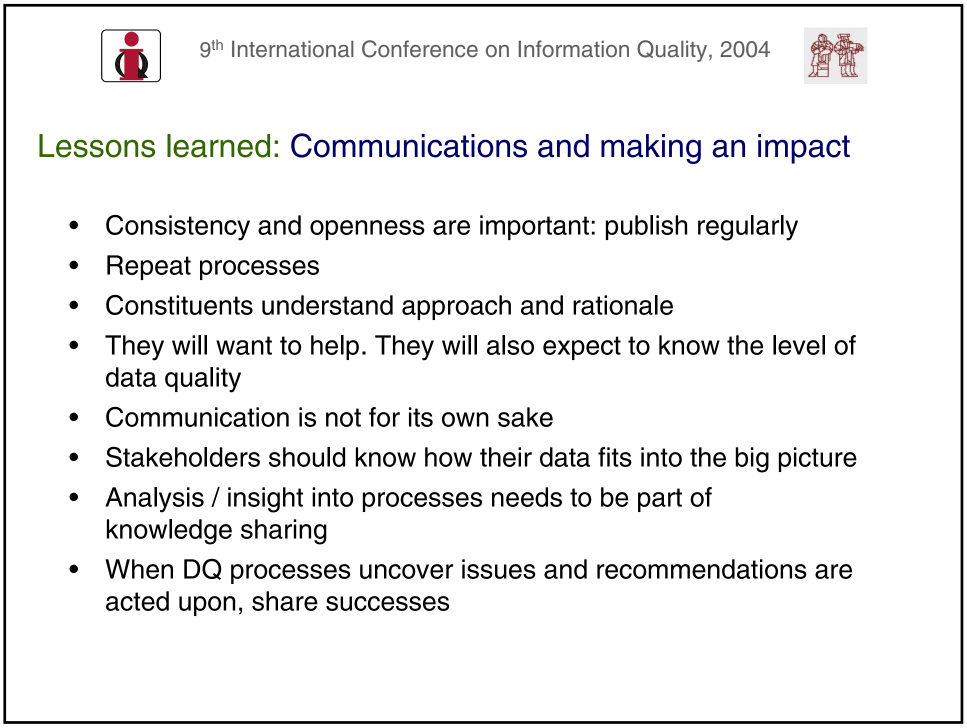
- Creator/custodian/consumer relations
- At first blush, the data warehouse team = “custodians” of the data
- But we also have an impact on the data – we are “creators.”
- And we use the data – we are “consumers”
- We play a mediating role between creators and consumers – we are data “brokers”
- In practice the line between creators, custodians, and consumers is not as clear as it is in theory
- Individuals and teams play multiple roles within this system and have different relations to the quality of the data
- One value of the model is the insight it provides to these different relations



9th International Conference on Information Quality, 2004

Lessons learned: Creator/custodian/consumer

- A key element of the program was that we would have our house in order
  - i.e., we would start with elements of data quality that the data warehouse had control over and
  - we would take seriously our responsibility to the data we create
- And that we would be good consumers of the data
  - i.e., we would understand where it came from, why it might be in the condition it was in and what we needed from it in order to use it for our purposes
- Recognizing that the custodial role is an active role
  - Clearly define who had control over what aspects of the data
  - Address issues that are in our control. Define impacts of those that are not
  - An accountability model



9th International Conference on Information Quality, 2004

Lessons learned: Communications and making an impact

- Consistency and openness are important: publish regularly
- Repeat processes
- Constituents understand approach and rationale
- They will want to help. They will also expect to know the level of data quality
- Communication is not for its own sake
- Stakeholders should know how their data fits into the big picture
- Analysis / insight into processes needs to be part of knowledge sharing
- When DQ processes uncover issues and recommendations are acted upon, share successes