PURSUING A CAREER IN INFORMATION QUALITY: THE JOB OF THE DATA QUALITY ANALYST

(Research-in-Progress)

Elizabeth M. Pierce
Indiana University of Pennsylvania
Eberly School of Business and Information Technology
empierce@iup.edu

Abstract Previous research on the skills needed by data quality professionals have focused on Information Systems (IS) curriculum standards, survey input from Information Quality (IQ) professionals, and the training required by related disciplines such as database and quality management. In this paper, Internet-posted job descriptions for the position of data quality analysts are examined. The purpose of this analysis is to determine the requirements and responsibilities expected by organizations seeking to fill this position. Such knowledge should prove useful to IQ educators looking to enhance the content of their courses so as to meet the needs of individuals interested in pursuing an IQ career as a data quality analyst.

Key Words: Data Quality Analyst, Information Quality Education

1 INTRODUCTION

1.1 The Role of The Data Quality Analyst

The data quality analyst job emerged in many organizations as a result of the implementation of data warehouses. Many organizations found that even if the data extracts, mappings, transformations, and loads were executed as planned, poor quality data still entered the data warehouse environment. Thus a new role, the data quality analyst (DQA), was created to help ensure that data loaded into the data warehouse met the informational requirements of the organization. William Inmon, J.D. Welch, and Katherine Glassey in their 1997 book, *Managing the Data Warehouse*, described the five major responsibilities of the data quality analyst [1]:

- Review data loaded into the data warehouse for accuracy
- Recommend maintenance enhancements to data acquisition processes to improve accuracy of data warehouse data
- Make recommendations to operational support for enhancements to systems of record to improve accuracy of operational data
- Review referential integrity of data warehouse data
- Review historical integrity of data warehouse data

What Inmon et al. [1] did not mention; however, was what types of skills would best prepare the data quality analyst for carrying out these responsibilities. In addition as the Information Quality field has developed, it is possible that organizations have found new uses for the data quality analyst beyond

working in the data warehouse environment. Therefore the focus of this paper will be to examine what is the current role of the data quality analyst in organizations and what types of skills do organizations feel that data quality analysts should have to carry out their duties.

1.2 Skills Required for Data Quality Professionals

Although previous studies have not specifically addressed the skills required for the job of the data quality analyst, there have been several works that have addressed data quality education for IS and management students at the undergraduate, graduate and executive education level.

Mathieu and Khalil [2] observed that while the IS'97 Model Curriculum for undergraduate students in IS addressed a variety of topics that impact the quality of data such as "EDP auditing", "data dictionaries", and "software development procedures", there were no specific units to guide students in measuring, tracking and improving data quality in an organization. Their survey of five representative database system textbooks revealed a similar pattern. No explicit mention of data quality; however, issues related to data quality such as database design were covered. They recommended incorporating classroom exercises into the undergraduate database course to help students to better understand the relationships between data quality control, business process management, data acquisition, and data use.

Fisher [3] wrote of his experiences in offering a *Data Quality and Information Systems* course to Information Systems seniors at Marist College. Like Mathieu and Khalil [2], Fisher found that although data quality was mentioned tangentially in many IS courses at Marist College; there was no one course that directly dealt with data quality problems. He chose the exercises and readings in his course to help students to understand data quality concepts, to recognize data quality problems and to suggest solutions, and to appreciate the role and importance of data quality in an organization. Instructors interested in developing a similar course on data quality will find this paper an extremely useful resource.

Khalil, Strong, Kahn, and Pipino [4] used a Product and Service Performance Model for Information Quality as the base for comparing two IS curriculum models (IS'97 and IRMCM). The PSP/IQ Model consists of a 2 x 2 matrix that captures how well organizations produce sound and useful information products and deliver dependable and usable information services to information consumers. The rows of the PSP/IQ matrix reflect information product and service quality while the two columns depicts two views of quality performance goals: conforms to specifications and meets or exceeds customer expectation. Their findings echoed the theme that while bits and pieces of the PSP/IQ model are addressed by a variety of courses (marketing, operations, accounting, programming, database, systems analysis and design), there was no integration of data quality issues for IS and management students. Their study suggested a paradigm shift is needed to change educators from a system focus to an information product focus and to promote information quality concepts throughout the IS and management curriculum.

Chung, Fisher, and Wang [5] used General Systems Theory to derive an educational framework for data quality professionals based on three areas: Technical Capabilities, Adaptive Capabilities, and Interpretive Capabilities. Their framework is described in Table 1.

To test their framework, they surveyed participants at ICIQ-2001 and found preliminary evidence for supporting the empirical measurement model of this framework. In addition they found evidence that different groups of data quality professionals value the different capability areas differently. Executives and managers value Interpretative Capabilities more while consultants, project managers, and analysts rate Adaptive Capabilities as more desirable. Rather than simply incorporating their model into existing

curriculum standards, their study suggested that a data quality education should be tailored to support where individuals are in their careers.

This last study raises an interesting issue for this paper to address about the skills required for DQA's. Do the skills currently being sought by organizations for the job of DQA match the data quality focused IS curriculum being advocated by many members of the IQ community?

Capability Area	Description	Skills Applied to Data Quality
Technical	Ability to understand and use computational theories and practical skills.	Data Mining / Knowledge Discovery Analytic Models (ex. Regression) Data Warehouse Setup Relational Algebra (ex. SQL) Statistical Techniques (ex. SPC) Programming Languages
Adaptive	Ability and knowledge for effectively interacting with data users, managers, and other stakeholders.	Data Quality Measurements Total Quality Management Data Entry Improvement Data Quality User Requirements
Interpretative	Ability to identify and describe the complex interplay between technologies and organizational structure.	Change Process DQ Implications & Potential Impacts DQ Cost/Benefit Analysis Database Error Detection

Table 1. General Systems Theory Framework for Data Quality Education

2. AN ANALYSIS OF INTERNET POSTED JOB DESCRIPTIONS

2.1 How the Job Descriptions were Obtained

To collect the job descriptions, the search engine, www.google.com, was employed. Using the advanced search option, a list of 53 web sites containing the exact phrase "data quality analyst" along with the words "job description" was obtained. A more detailed examination of these 53 web sites revealed that 21 contained meaningful job descriptions. The rest were either sites containing duplicate listings of the same job, sites advertising software tools, or sites pertaining to non-related subject matter. A second search using the exact phrase "information quality analyst" was also employed but it did not supply any additional useful hits. A similar search of the monster.com web site picked up an additional 4 job postings for data quality analyst. Table 2 in the appendix lists the job descriptions for data quality analyst positions that were used for this study. All web postings were identified on or about June 24, 2003 with the exception of NetRatings, Inc., which was obtained during an earlier search in February 2003.

2.2 Results of the Analysis

Because this is a convenience sample of 26 job postings for the position of DQA, one must be cautious in interpreting the results. It does appear that companies from a variety of different industries are employing DQA's to assist them in improving their information products and services. Many of the job postings are not dated so it is difficult to determine the currency of the job descriptions. While some job advertisements have dates of June 2003, others have dates in 2002, 2001, and 2000. Some of the non-

dated postings may be even older. By far, the biggest challenge in working with this type of study is the lack of uniform format or content among the job postings. A few job descriptions specify salary, benefits, and to whom the DQA reports; however, most do not. Information about the DQA position range greatly. Some job postings go into great detail about the role and requirements of the DQA while others do not go much farther than indicating a position is available. It is important to keep in mind that an organization may require skills and responsibilities that are not cited in the job postings. The result is that while this type of study is useful for getting a sense of what it means to be a DQA, a more detailed study involving classical survey techniques combined with a formal methodology like Content Analysis is necessary in order to draw definitive conclusions. Content analysis [6] is a research tool helps one to quantify and analyze the presence, meanings, and relationships of specified words and concepts and then make inferences about the messages within the texts. Although this first research effort does not incorporate this methodology, it may prove useful in future studies seeking to improve IQ education and career planning.

2.3 DQA Responsibilities Cited by Job Postings

As shown in Table 3, the two most common DQA responsibilities cited by the majority of job postings involve activities dealing with resolving data quality problems and providing quality assurance oversight of data flows and stores. Typical phrasing for resolving data quality problems are words like "Assist in determining root cause of quality issues and recommending short term and long term solutions" (Bluebell, PA), "Investigate, trouble-shoot, and document source-to-target and other data issues from internal and external customers" (DataQuick), and "Evaluate current quality control processes and make recommendations for process improvements" (IRI). Typical terminology for the responsibility of providing data quality assurance include phrases such as "Establish, monitor, and report on Service Level Agreements for financial instrument data quality with the outsourcing vendor" (NY/Jersey City), "Perform audits as required to ensure quality processes exist and are being adhered to within Operations" (Instill), and "Perform routine data verification procedures" (Unicare).

The third most common DQA responsibility cited involve activities related to providing data quality advisory services. For example, ComScore expects their DQA's to maintain relationships with vertical business managers to ensure validity of business rules and data quality processes. Boehringer Ingelheim specifies that their DQA's should actively communicate database issues with multiple internal and external customers.

Roughly half of the job postings also contain activities that could be roughly grouped into two additional task areas. Some organizations assign their DQA's the task of measuring data quality. For instance, Genscape cites that one responsibility of their DQA's is to "develop and implement a data quality monitoring system". Other organizations also request their DQA's to manage documentation related to data quality. CareSource requests that their DQA's create flowcharts of all processes that lead to the input of data to the organization or output of data to other organizations.

Finally, a few organizations specify some additional responsibilities for their DQA's. EarthLink is one of four companies that charge their DQA's with providing data quality training to others in the organization. Two companies make reference to activities that speak to documenting the return on investment of data quality activities. Bioinsource requires information about the DQA's contribution to other projects be provided upon request while Experian asks that their DQA's prepare reports to improve workforce focus on maximizing the return on the data used by Experian. Finally, Target assigns its DQA's the task of ensuring compliance with the company's data privacy policy.

DQA Responsibilities	**Job Postings Citing these DQA Responsibilities**
1.Resolve data quality problems. Activities include: Identify, assess, fix, document, and communicate potential quality issues in the way data is collected, stored, processed, or used. This may involve interacting with business users to resolve data quality problems using techniques such as root cause analysis as well as collecting user requirements. Fixing the data problem may require the DQA to recommend, design, implement and monitor quality procedures for use in the data production process. It may also include data cleansing operations like removing duplicate records.	**20 Job Postings** (A, C, D, F, G, H, I, J, K, L, M, N, O, P, S, V, W, X, Y, Z)
2. Provide quality assurance oversight of data flows and stores. Activities include: Review quality of data entry, loads, transformations, extractions, merges, or other production jobs. This may also include the establishment and monitoring of service level agreements, communication protocols with data suppliers, and data quality assurance policies.	**20 Job Postings** (A, C, D, E, F, G, H, I, J, K, L, P, R, S, T, U, V, W, X, Z)
3. Provide data quality advisory services to the organization. Activities include: Participate in cross-functional teams and development or improvement task forces. Assist other groups in database setup, design of edit checks, test plans, DQ policy decisions, etc. to ensure data quality objectives are met.	**16 Job Postings** (C, E, F, G, H, I, J, K, N, P, R, T, U, V, X, Z)
4. Measure data quality. Activities include: Design, collect, analyze, and report on data quality assurance / production performance metrics.	**12 Job Postings** (A, C, G, H, I, J, L, N, P, U, Y, Z)
5. Manage data quality documentation. Activities include: Define and maintain data standards, definitions, and models (e.g. data dictionary, organizational data model. This may also include any DQA documents such as checklists, guidelines, manuals, templates, forms, etc.) .	**11 Job Postings** (A, C, E, G, P, T, U, V, W, X, Z)
6. Provide data quality training. Activities include: Provide data quality training and presentations to members of organization.	**4 Job Postings** (E, J, K, U)
7. Document the return on investment of data quality activities Activities include: Document DQA's contribution to other projects. Analyze the return on the use of quality data.	**2 Job Postings** (E, K)
8. Ensure enforcement of organization's data privacy policy	**1 Job Posting** (V)
** No specific responsibilities cited in job posting*	**2 Job Postings** (B, Q)

Table 3: DQA Responsibilities Cited by Job Postings

2.4 DQA Skill Requirements Cited by Job Postings

Table 4 summarizes the requirements for the DQA as mentioned in the job postings. The requirements are organized along the framework suggested by Chung, Fisher, and Wang [5]. In terms of technical skills, over half of the organizations expect their DQA's to be familiar with relational databases and/or SQL. Only four job postings explicitly list data warehousing experience, which may indicate that the role of the data quality analyst extends beyond data warehousing to any business process where data is collected, stored, transformed, or used. In addition, only four job postings explicitly mention that programming and application development skills are desirable while eleven job postings list knowledge of statistics or data analysis as an important requirement. By far, the most popular software tool cited is the Microsoft Office Suite and in particular, the Excel and Access components for use in the manipulation and analysis of data. Taken as a whole, the technical requirements for a DQA appear consistent with the technical skills identified by Chung, Fisher, and Wang [5] for data quality professionals.

Since adaptive skills deal with the ability and knowledge for effectively interacting with other stakeholders, this paper places communication skills, project/time management skills, and teamwork skills in this category along with the other adaptive skills (DQ Measurements, TQM, Data Entry Improvements, User Requirements) previously identified by Chung, Fisher, and Wang [5]. The DQA job descriptions request good communication skills the most in this category followed by the ability to organize one's work and to multitask, and then by the ability to work in teams. Interestingly, only eight job advertisements explicitly cite TQM, QA, Six Sigma, or other data quality experience as a requirement and only one job posting (NY/Jersey City) mentions establishing user requirements.

This is rather striking considering that most job postings list data quality assurance activities as a primary DQA responsibility. Further research is necessary to determine if organizations are aware of these skills and if so, what is the reason for not listing them as a job requirement. Perhaps too few people have training in data quality assurance so rather than look for experienced people, organizations instead rely on individuals learning these data quality skills on the job. This may be an area where DQA's might benefit from additional training on how to measure data quality, how to improve the quality of data entry, how to obtain data quality requirements from users, and how to apply TQM principles to data quality management.

Under interpretative skills, over half of the job postings list phrases like "well developed conceptual and analytical skills" (HDSS), "strong cognitive skills (DataQuick), and "strong analytical / problem solving skills (Yoh IT) as a requirement for the DQA position. This is consistent with the role of the DQA as a data error detector and corrector. Chung, Fisher, Wang [5] also listed understanding pervasiveness of data quality problems and their potential impacts, conducting cost/benefit analysis of data quality management, and managing the change process resulting from data quality management under interpretative skills. Although no job posting explicitly lists these skills, this may be a similar situation to that of the overlooked adaptive skills. Further research is needed to determine how much the work of the DQA's would be enhanced if they possessed a greater set of data quality interpretative skills.

Finally, it is interesting to note that most job postings for DQA's look for people with prior experience in the industry. It appears that the DQA is typically not an entry-level position, but requires at least a few years working in an application area to gain experience with how the data is defined and used. In addition to the field experience, a college education is the most frequent educational level cited by the job descriptions. However, six organizations would consider candidates with high school or associate degrees while two organizations prefer individuals with advanced degrees like an MBA. This may suggest an individual interested in pursuing a career as a DQA should be prepared to combine data quality training with work experience in a particular industry.

Job Postings Citing These Skill Requirements for DQA	
Technical Skills	**Number of : (specific companies citing skill)**
Design, develop and maintain software applications (SDLC knowledge) (Programming Languages).	**4:** (B, C, J, L)
Relational DBMS skills across multiple platforms. (SQL, relational algebra and other query tools)	**15:** (A, B, C, E, F, H, I, J, K, O, Q, R, U, X, Y)
Data Warehouse Knowledge / Experience	**4:** (A, H, J, V)
Knowledge of research methods, statistical or data analysis. This may also include a familiarity with statistical packages (SAS, SPSS) and data manipulation software (FOCUS, EZ-Trieve).	**11:** (E, G, H, I, M, N, O, P, V, W, Z)
MS Access	**11:** (A, E, G, I, K, P, Q, R, U, W, Y)
MS Excel	**16:** (A, D, E, F, G, I, K, L, M, P, Q, R, U, W, X, Y)
MS Word	**10:** (D, E, F, G, I, Q, U, W, X, Y)
MS Outlook	**9:** (E, F, G, I, Q, U, W, X, Y)
MS PowerPoint	**8:** (E, G, I, Q, U, W, X, Y)
Miscellaneous Computer Skills (MS Project, Visio, MS Work, Internet Browser, Basic Computer Skills, specialty packages like Clintrial4, etc.)	**10:** (D, E, G, I, J, M, T, X, Y, Z)
Adaptive Skills	
Knowledge of TQM / QA practices / Six Sigma / Process Improvement (As Applied to Data Quality)	**8:** (A, B, C, E, I, J, P, X)
Good verbal and written communication skills (good interpersonal skills, presentation skills) (able to consult, liaise, and negotiate)	**15:** (A, C, D, E, F, G, H, I, J, M, N, O, U, X, Y)
Good project / time management, and organizational skills. Good attention to details. Able to multi-task, work under pressure.	**13:** (A, C, F, J, L, M, P, Q, R, V, U, X, Z)
Ability to work in teams as well as work independently.	**10:** (A, F, G, I, J, M, O, P, Y, Z)
Other Adaptive Skills: Experience in DQ Measurements, Data Entry Improvement, User Requirements.	**1:** (A)
Interpretative Skills	
Good problem solving (error detection, root cause analysis, error correction), conceptual, analytical and decision making skills.	**14:** (A, D, F, G, I, J, M, N, O, P, R, V, X, Y)
Other Interpretive Skills: Experience in Change Process, DQ Implications, DQ Cost/Benefit Analysis.	**0:** ( )
Application Specific Knowledge (For F, H, L, N, Q, no specific field cited.)	
Financial / Credit / Securities / Insurance	**4:** (A, D, K, X)
Healthcare / Medical Insurance / Pharmaceutical	**7:** (B, C, E, G, S, T, W)
Data Analysis / Market Research	**3:** (M, O, Z)
Education	**1:** (U)
Food Service	**1:** (P)
Production Support	**1:** (J)
Real Estate	**2:** (I, R)
Retail, Distribution/Inventory of Retail Goods	**2:** (V, Y)
Minimum Preferred Education Background (For D, N, P, R, Y, Z, no minimum education requirement cited.)	
High School Degree	**2:** (B, S)
Associate Degree	**4:** (G, I, Q, W)
Bachelor's Degree	**12:** (C, E, F, J, K, L, M, O, T, U, V, X)
MBA / Advanced Degree	**2:** (A, H)

Table 4: DQA Skill Requirements Cited by Job Postings

3. CONCLUSIONS AND FUTURE RESEARCH

It appears from the job postings for data quality analysts that this position extends beyond the data warehouse environment to any process where data quality is an issue. It seems that many organizations expect their data quality analysts to take the lead in resolving data quality problems and providing quality assurance oversight of the data stores and flows. Other duties of the data quality analyst may involve providing data quality advisory services to the rest of the organization, measuring data quality, and managing documentation related to the data quality efforts. A few organizations also expect their data quality analysts to provide training, to document the return on investment of data quality activities and to ensure enforcement of the corporate data privacy policy; however, more research is need to determine if these are aberrations from the norm or emerging trends.

In terms of the skills required for the data quality analyst, it appears that many of the technical skills expected of someone with an IS education focused on data quality carry over well to this position. Good communication, project management, and teamwork skills are also important. Many of the adaptive and interpretative skills cited in previous work on data quality are not being listed as requirements in the job postings. However, further research must be done to verify if organizations are aware of these skills and if they are, what is the reason for not listing these skills as a requirement. Perhaps too few people have these skills or perhaps some of these skills are not required for the data quality analyst position. More research is also needed to see if data quality analysts with these skills perform their responsibilities better than individuals who lack these skills.

There are several other questions that further research on this topic could explore. For instance, is there any relationship between the role of the data quality analyst and the characteristics of the organization such as its size or structure? How well do job listings coincide with the actual work being done by the data quality analysts in the organization? If the job of the data quality analyst is not an entry-level position, when is the best time to schedule training and what should be the content and format of this training? More formal, investigative efforts are needed to address these important IQ educational issues.

REFERENCES

[1] Inmon, W.H., J.D. Welch, K.L. Glassey., *Managing the Data Warehouse*. 1997: John Wiley & Sons.

[2] Mathieu R. and O. Khalil., *Teaching Data Quality in the Undergraduate Database Course* in *Second International Conference on Information Quality*. 1997. Cambridge, MA.

[3] Khalil, O.E. M, D. M. Strong, B.K. Kahn, and L.L. Pipino., "Teaching Information Quality in Information Systems Undergraduate Education", *Informing Science,* 1999. 2(3): p. 53-59.

[4] Fisher, C., *A College Course: Data Quality in Information Systems.* in *Sixth International Conference on Information Quality.* 2001. Cambridge, MA.

[5] Chung, W., C. Fisher and R. Wang., *What Skills Matter in Data Quality?* in *Seventh International Conference on Information Quality*. 2002. Cambridge, MA.

[6] writing.colostate.edu/references/research/content/, Accessed on August 13, 2003.

APPENDIX

Company / Location	Industry	Web Address
A. Unknown Client NY/Jersey City	Financial	www.careersonthemove.com/Tech%20-%20NY.htm
B. Unknown Client Philadelphia, PA	Healthcare	www.jobvertise.com/job/1271957
C. Unknown Client Bluebell, PA	Communication Campaigns	www.jobvertise.com/job/1222337
D. A-T Financial Info New York, NY	Financial (Software Dev.)	www.job-list.com/listings/X08025.htm
E. Bioinsource East Bay, CA	Pharmaceutical Staffing	www.medzilla.com/jobs/wjbioinsource_10218.htm
F. Boehringer Ingelheim Ridgefield, CT	Pharmaceutical	jobsearch.monster.com (search for data quality analyst, select from job list)
G. CareSource Dayton, OH	Healthcare Information	ohio.computerwork.com/cfm-bin/jobsearch/1117539.cfm
H. ComScore Reston, VA	Consumer Research	www.comscore.com/about/jobs/tech_data_qa.asp
I. DataQuick San Diego, CA	Real Estate Information	www.dataquick.com/careers.asp
J. EarthLink Atlanta, GA	Internet Service Provider	jobsearch.monster.com/jobsearch.asp?co=xearthlinkx (select data quality analyst from job list)
K. Experian Orange, CA	Credit Information	jobsearch.monster.com (search for data quality analyst, select from job list)
L. Genscape Louisville, KY	Energy Information	www.genscape.com/about_openpos_analyst.shtml
M. GlobeSoft Chicago, IL	IT Consulting	www.globesoft.com.sg/jobsdetail.asp?JOBID=SG0151#
N. HDSS Victoria, Australia	Healthcare	hdss.health.vic.gov.au/bulletin/20-5401.htm
O. IRI Chicago, IL	Marketing Research	www.hotjobs.com/cgi-bin/company-job-list?G__GID=2643 (select data quality analyst from job list)
P. Instill Redwood City, CA	Food Service	www.instill.com/company_lm_car_job2.shtml
Q. Integrated DNA Tech. Coralville, IA	Manufacturing	www3.iwd.state.ia.us/iowajobs2.nsf/0/75678df71ebcb71686256d26004abf3c?OpenDocument
R. LoopNet Monrovia, CA	Real Estate Information	www.loopnet.com/about.asp?LNSection=Careers
S. Mammoth Hospital Mammoth Lakes, CA	Healthcare Provider	www.mammothhospital.com/html/data_quality_analyst.html
T. Novo Nordisk Princeton, NJ	Pharmaceutical	jobsearch.monster.com (search for data quality analyst, select from job list)
U. SAISD San Antonio, TX	Education (School District)	www.saisd.net/admin/personnel/admnprof.shtm (select data quality analyst from job list)
V. Target Minneapolis, MN	Retail	jobsearch.monster.com (search for data quality analyst, select from job list)
W. Unicare Hartford CT	Healthcare	www.sconsig.com/jb199809.htm (search for data quality analyst, select from job list)
X. Wells New Business Dev. Atlanta, GA	Real Estate Funds	www.wellsref.com/jobfair/job-fair_body_0603.asp (select data quality analyst from job list)
Y. Yoh IT St. Louis, MO	Technology Staffing	jobsearch.monster.com (search for data quality analyst, select from job list)
Z. NetRatings, Inc. Milpitas, CA	Marketing Research	www.netratings.com/corporate/careers.htm

Table 3: List of data quality analyst job descriptions used in study.