Expanding Your Notion of Data Quality Challenges
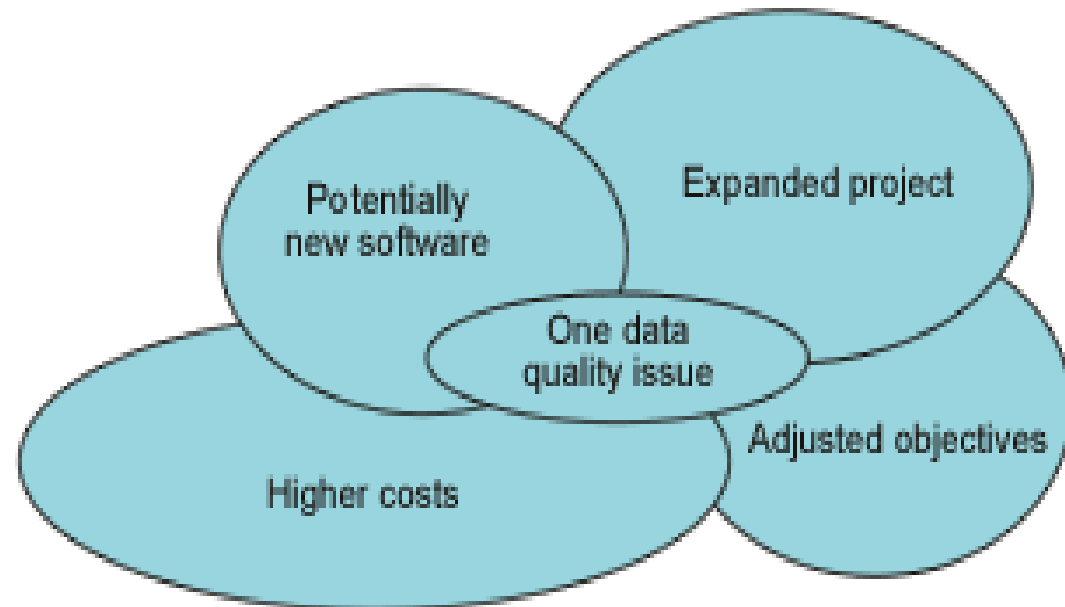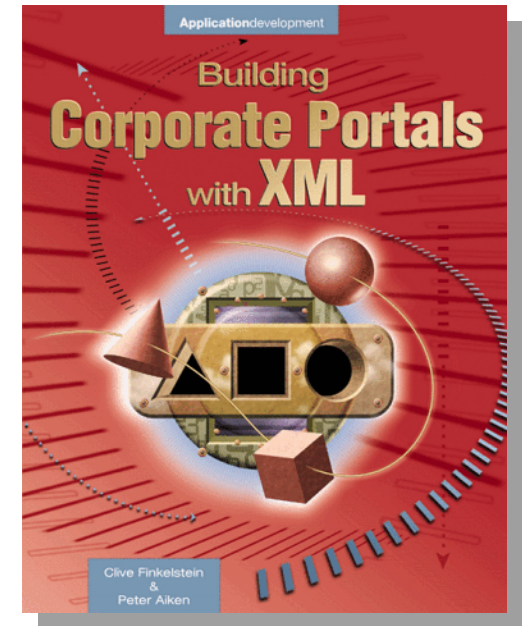
John Sells
Data Blueprint

jsells@datablueprint.com

VCU School of Business
Virginia Commonwealth University
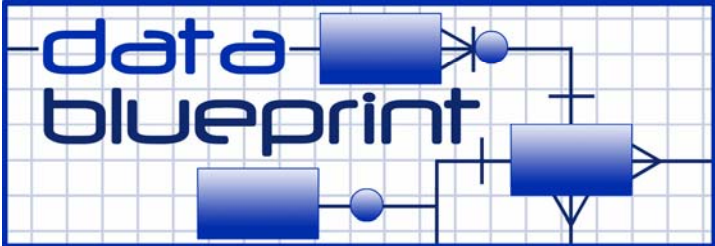
Famous Words?

- Question:
  - Why haven't organizations taken a more proactive approach to data quality?
- Answer:
  - Fixing data quality problems is not easy
  - It is dangerous -- they'll come after you
  - Your efforts are likely to be misunderstood
  - You could make things worse
  - Now you get to fix it
- A single data quality issue can grow into a significant, unexpected investment

Building **Corporate Portals** with **XML**

Application development

Clive Finkelstein & Peter Aiken

Potentially new software

Expanded project

One data quality issue

Higher costs

Adjusted objectives

Oct 2004 IRS Accomplishment

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- Unified five definitions of "child"
- Reduce 5 definitions to 1 for tax return preparations such as:
  - Dependent
  - Earned income tax credit
  - Child credit
- Different reasons, either it
  - "Was developed to carry out social policy objective(s), or
  - Someone perceived it was going to save revenue"
- "Is it easier for (customers) to understand and it is easier for IRS to audit and there a lots of things like that we can do"
- Initiative started in 1991 - it took **13** years including 2.5 years moving as legislation!

Source: Pamela F. Olson former Assistant Secretary for Tax Policy (quote from the Diane Rehm Show • 11/29/04 • http://www.wamu.org/programs/dr/04/11/29.php)

data blueprint

How to solve this data quality problem using just tools?

Retail price for the unit was $40

Letter from the Bank

BANK ONE.   **CHASE**

*************** AUTOCR ** R-004
Dr Peter Aiken
13155 Country Garden Ln
Montpelier, VA 23192-3028

Important information about your Bank One account ending in 7814.

Good news about the Bank One and Chase merger.

Dear Dr Peter Aiken:

You may have heard about the recent merger between Bank One and JPMorgan Chase. As a result of this event, your credit card information will be sent to you under the Chase name. However, you may still receive an occasional letter from Bank One, so please continue to open your mail from either Chase or Bank One.

This change will be positive for you. All the card benefits you've come to rely on are still there for you including 24-hour customer service - online or on the phone. Your account number, recurring bill payments, and any scheduled payment of your credit card account remain unchanged. You can customize your account by choosing your own billing date or by adding additional cards. Soon, you will have the opportunity to explore a broader range of products than ever before, including:
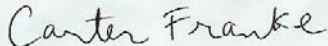
Retail Banking: Enjoy the power and convenience of checking, savings and loan products at more than 2,300 branches and 6,500 ATMs at Chase and Bank One nationwide. If you make your credit card payments at our branches, please continue to do so at the same locations you use now until further notice.

Chase Auto Finance: We offer competitive rates, a fast and easy application process and flexible financing options.

Online Brokerage Services: Through BrownCo, a brokerage service of J. P. Morgan Invest, LLC, we offer discount brokerage for experienced investors and active traders.

Visit us online at www.BankOne.com/information to learn more about these products or the merger. At Chase, we've assembled the right resources, products and expertise to help you do more with your money. We're committed to doing the little things right, the big things faster and providing you with the very best customer service possible.
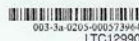
Sincerely,

Carter Franke

Carter Franke
Chief Marketing Officer
Chase Card Services

P.S. Please be on the lookout for any upcoming communications from either Chase or Bank One regarding your Bank One credit card and any other Bank One product you may have.

Bank One is a division of JPMorgan Chase Bank, N.A. Member FDIC.
Credit card accounts are owned by Chase Manhattan Bank USA, N.A. © JPMorgan Chase & Co. 2004.

CVC1

003-3a-0205-000573964
LTC12990

… so please continue to open your mail from either Chase or Bank One

P.S. Please be on the lookout for any upcoming communications from either Chase or Bank One regarding your Bank One credit card and any other Bank One product you may have.
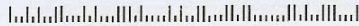
Problems

- I initially discarded the letter!
- I became upset after reading it
- It proclaimed that Chase has data quality challenges

A congratulations letter from another bank

Problems

- Bank did not know it made an error
- Tools alone could not have prevented this error
- Lost confidence in the ability of the bank to manage customer funds

Perfect (adjective)

- Lacking nothing essential to the whole; complete of its nature or kind."

- Metadata quality goal is to be accurate and lacking nothing essential

- Lack of anything required to respond to the customer's request is considered imperfect.

- Imperfections are either practice-oriented or structure-oriented

[American Heritage English Dictionary 1993]

Metadata Quality Dimensions

Figure 9-1

Refined dimensions of perfect Enterprise Portal data.

Quality of Data Value

Practice related

Quality of Data Representation

Data Quality

Quality of Data Models

Structure related

Quality of Data Architecture
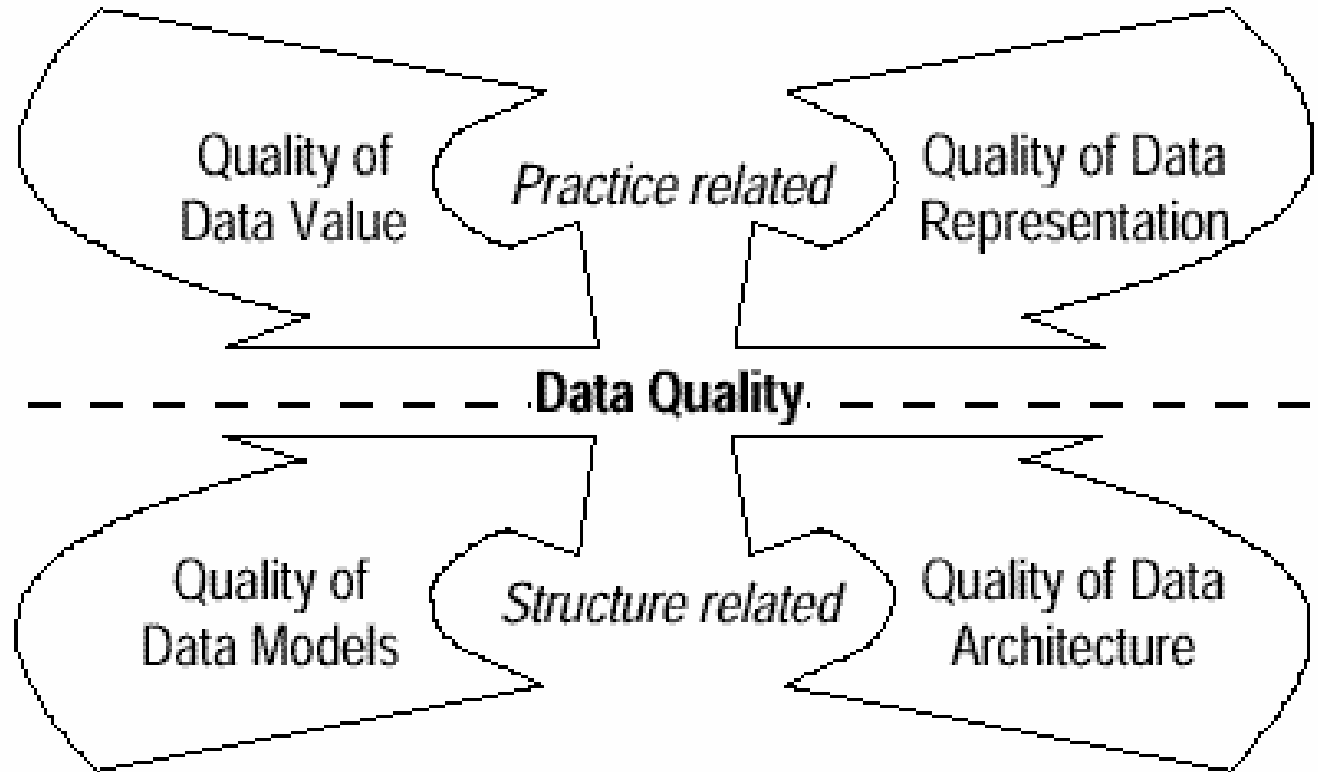
Quality Attributes

(closer to the user) ←—————————————————————————————————→ (closer to the architect)

Data Representation Quality	Data Value Quality	Data Model Quality	Data Architecture Quality
as presented to the user	as maintained in the system	as understood by developers	as an organizational asset
Completeness	Completeness	Completeness	Completeness
Correctness	Correctness	Correctness	Correctness
Timeliness	Currency	Conceptual Correctness	Enterprise Model Utility
Conciseness	Frequency	Conceptual Completeness	Data Management Quality
Clarity	Time Period	Syntactic Correctness	Data Sharing Ability
Detail	Precision	Syntactic Completeness	Data Engineering Quality
Order	Reliability		Data Operation Quality
Presentation	Relevance		Data Evolvability
Media	Scope		Data Self Awareness
Unambiguous	Granularity		
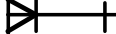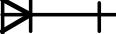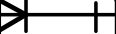
Traditional Quality Life Cycle

Figure 9-8

Levitan and Redman's Data Acquisition and Usage Cycles [Levitin and Redman 1993].

Data acquisition activities

Data storage

Data usage activities

Data Life Cycle Model

Metadata Creation

Metadata Refinement

Metadata Structuring

Data Refinement

Data Assessment

Data Creation

Data Storage

Data Utilization

Data Manipulation

Data Life Cycle Model

With inputs and outputs added

Data Architecture Quality

Data Architecture and Model Quality

Metadata Creation

Metadata Refinement

Primary Foci of Data Life Cycle Model

Data Model Quality

Metadata Structuring

Data Refinement

Data Assessment

Data Representation Quality

Data Value Quality

Data Creation

Data Storage

Data Utilization

Data Manipulation

What is DM3?

- **D**ata **M**anagement **M**aturity **M**easurement
- You are currently managing your data,
  - But, If you can't measure it,
  - How can you manage it effectively?
- How do you know where to put time, money, and energy so that data management best supports the business?
- DM3 is an adaptation of the SEI-CMM® to the discipline of Data Management

  Carnegie Mellon
  Software Engineering Institute

- An assessment of the relative development of organizational data management practices using a CMM framework

data
blueprint
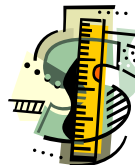
SEI CMMI Capability Maturity Model Levels

"Self" Improving

We have a process for **improving** our DM capabilities

Optimizing (5)

Unsustainable We **manage** our DM processes so that the whole organization can follow our standard DM guidance

Unpredictable

Managed (4)

We have experience that we have **standardized** so that all in the organization can follow it

Inconsistent

Defined (3)

We have DM experience and have the ability to implement **disciplined** processes

Repeatable (2)

One concept for process improvement, others include:
- Norton Stage Theory
- TQM
- TQdM
- TDQM
- ISO 9000

And focus on understanding current processes and determining where improvements can be made.

Out of control

Our DM practices are **ad hoc**

Initial (1)

data blueprint

Enterprise DM Functions and their Inter-relationships

Organizational Strategies

Guidance

Data Program
Coordination

Goals

Implementation

Enterprise
Data Integration

*Integrated
Models*

Data
Stewardship

Standard Data

Data
Development

*Application Models
& Designs*

Direction

Data Support
Operations

*Business
Data*

Data
Asset Use

Feedback

Business Value
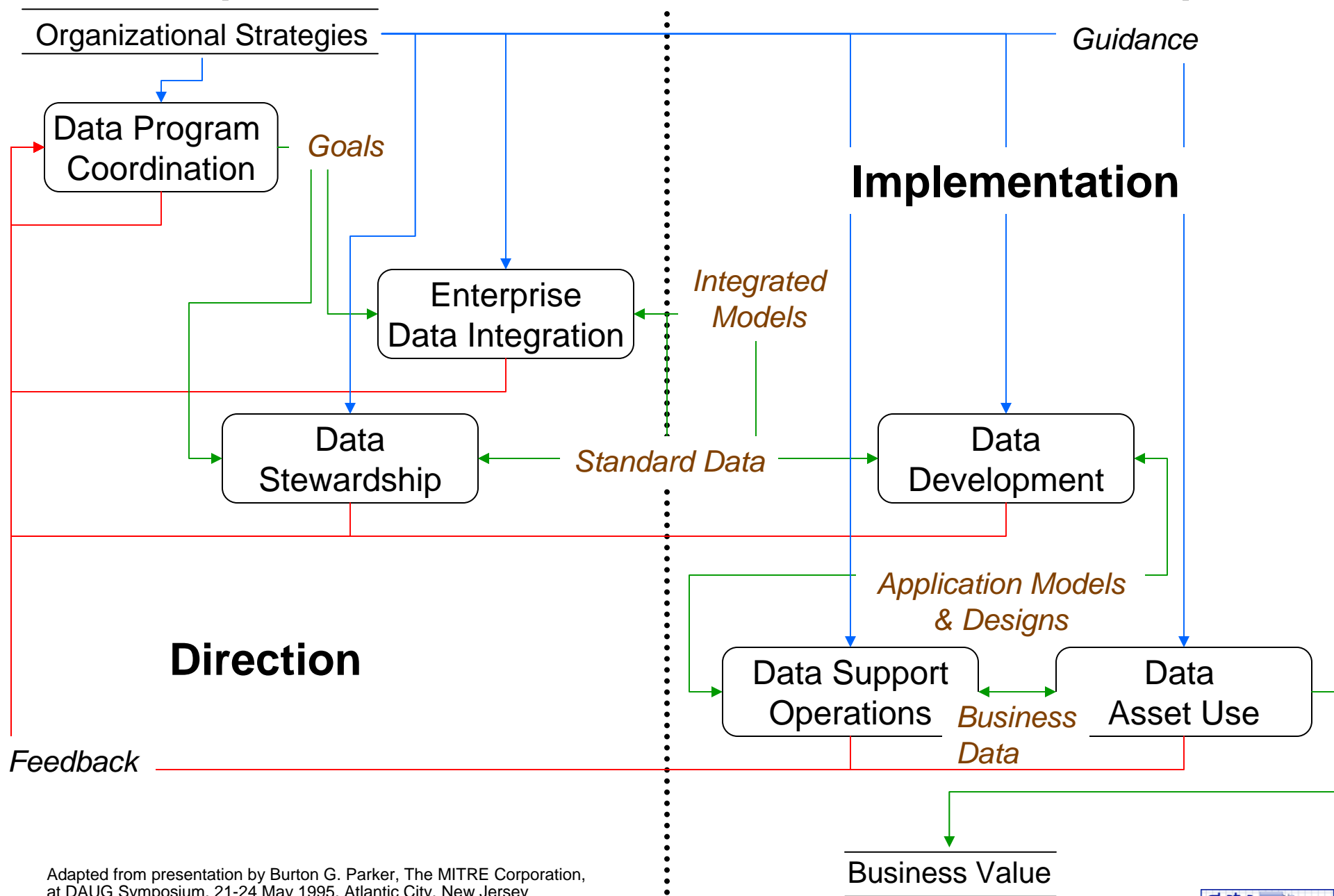
Adapted from presentation by Burton G. Parker, The MITRE Corporation,
at DAUG Symposium, 21-24 May 1995, Atlantic City, New Jersey

data
blueprint

After more than a decade …

Question   How many software practices (surveyed) are above level 1 on the CMM?

Answer   By far most organizations surveyed are producing software using informal processes

Question   How many organizations have demonstrated at least some proficiency according to the DM3? (i.e., scored above level 1)

Answer   One in ten organizations has scored above level 1 in the DM3 according to our surveys

Phase	Problem at Hand	Approach	Benefits
Initial Test	SAMMS stored data in free-text fields.	Evaluated automation of data extraction from free text fields for one DLA Supply Center for Option B.	Documented Data Audit Business Case that outlined potential feasibility and savings.
I		Extract text data from Option B and retrieve associated data from other data sources for one Supply Center.	Substantial time and monetary savings as a result of Data Quality Audit versus a complete manual approach.
II		Audit extracted text data against defined quality standards for one Supply Center for Options B and E.	A repeatable process was created for extracting data out of free text fields. Documented results of Business Case.
III	Initial Test - Phase II was only implemented at Richmond Supply Center.	Expanded Phase II to all Centers and added business rules to harmonize differences between Centers.	Successfully provided cyclic audit results and expanded Initial Benefits DLA-wide.
IV	Required method to ensure that production data remained clean.	Provided web-based data cleansing environment to correct non-textual fields and continue data auditing.	Post-audit data remained clean; knowledge worker-friendly web-based front-end enforced business and data quality rules.
V	Needed to address other textual Options and lack of a real-time Data Audits.	Updated web application for all Options and a pseudo-real time feature to conduct Data Audit.	Fulfilled desired Data Quality objective for business process and technology with a terminal interface.

data blueprint

Sample Free Text

1	Manufacturer Accel Systems
2	CAGEC 44910
3	Aircraft frame aluminum MIL-STD-339184
4	P/N 33919340-44491
5	SEE ALSO REF 331018

**Manufacturer Accel Systems
CAGEC 44910
 P/N 33919340-44491
SEE ALSO REF 331018**

data blueprint

Sample Free Text (cont'd.)

Manufacturer Accel Systems
CAGEC 44910
P/N 33919340-44491
SEE ALSO REF 331018

NSN	CAGE	PART
1234567890123	44910	33919340-44491

- Data here is extractable – we can tell by looking at the metadata markers
- These two fields naturally go together by order

Three Categories of Data

- Extractable Data

  – The markers are found, and the data is pulled.

- Ignoreable Data

  – There is no data in the field, and we can prove it.

- Unextractable Data

  – We cannot be sure if there is or is not data.

Solution Goals

- Maximize the extracted.
  - These are where the actual results come from
  - The more accurate the approach, the bigger this set.
- Maximize the ignored.
  - Size the problem. Which records are not worth worrying about?
  - This set may have its own set of interesting characteristics
- Minimize the unextractable.
  - These records ultimately must be addressed manually
  - Only the most unpredictable in this category

Rev#	Unmatched Items	Unmatched Items (% Total)	Ignoreable NSNs	Ignoreable Items (% Total)	Items Matched	Avg Extracted Per Item	Items Matched (% Total)	Items Extracted
1	329948	31.47%	14034	1.34%	N/A	N/A	N/A	264703
2	222474	21.22%	73069	6.97%	N/A	N/A	N/A	286675
3	216552	20.66%	78520	7.49%	N/A	N/A	N/A	287196
4	340514	32.48%	125708	11.99%	582101	1.100022161	55.53%	640324
…	…	…	…	…	…	…	…	…
14	94542	9.02%	237113	22.62%	716668	1.114291415	68.36%	798577
15	94929	9.06%	237118	22.62%	716276	1.113928151	68.33%	797880
16	99890	9.53%	237128	22.62%	711305	1.11530075	67.85%	793319
17	99591	9.50%	237128	22.62%	711604	1.115439205	67.88%	793751
18	78213	7.46%	237130	22.62%	732980	1.207281236	69.92%	884913

An Iterative Process…

data blueprint

Efficiencies

- 1,000,000 items were run, comprising over 6,000,000 lines of text.

- About 70% had all data extracted.

- About 23% provably had no data.

- The remaining 7% could not be handled.

- The scope of the manual effort was reduced from 1,000,000 records to 70,000
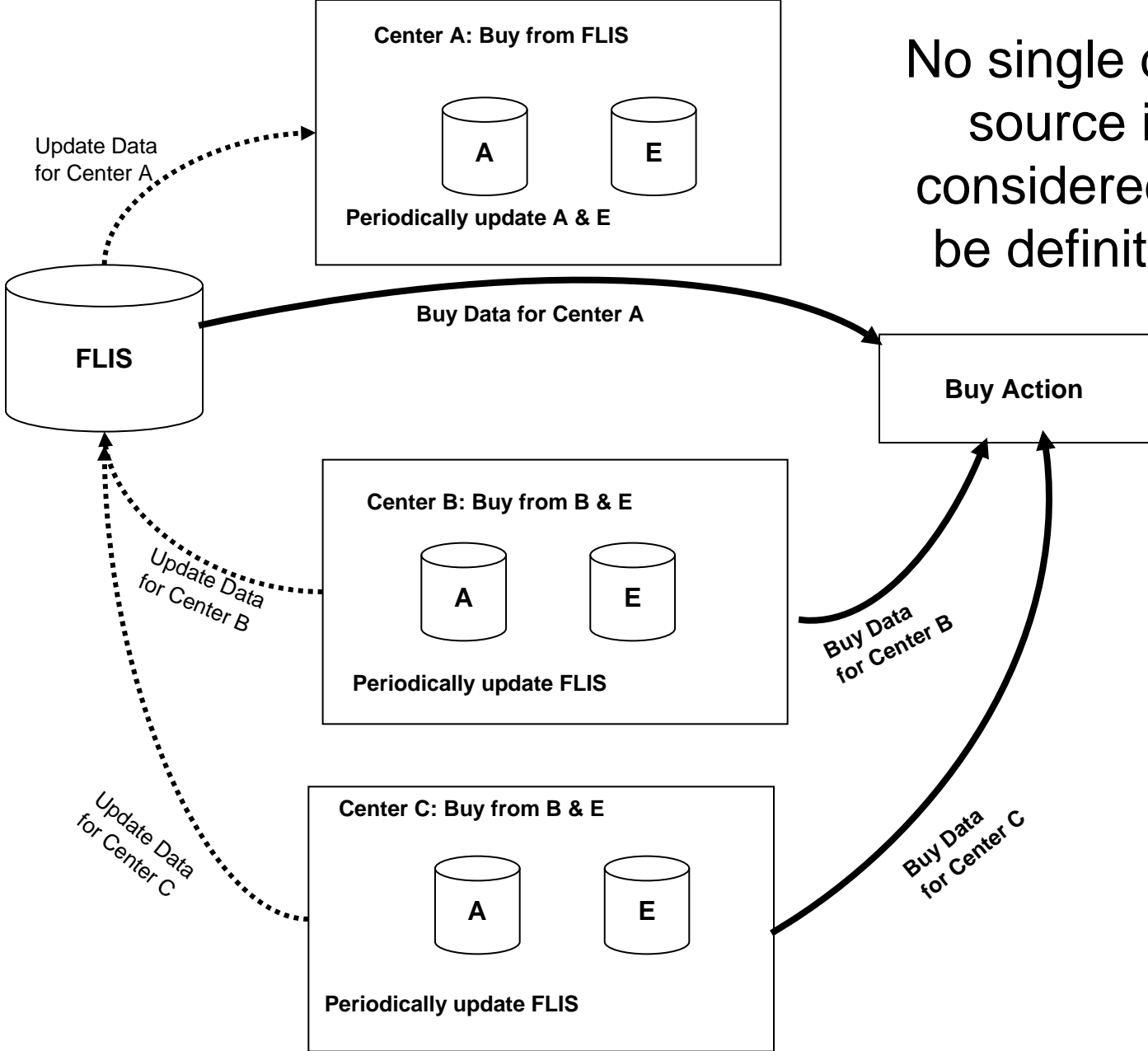
data blueprint

Savings

- Average manual rate was 5,000 recs/person/month. This varies by person and by data.

- Problem size was 1,000,000 records.

- On this project, automated savings amounted to 12.75 person years.

- With FTE costs of $60,000+/year, this is over $750,000 in savings.

DLA Information Coordination

Center A: Buy from FLIS

A   E

Periodically update A & E

Update Data for Center A

No single data source is considered to be definitive

FLIS

Buy Data for Center A

Buy Action

Center B: Buy from B & E

A   E

Periodically update FLIS

Update Data for Center B

Buy Data for Center B

Center C: Buy from B & E

A   E

Periodically update FLIS

Update Data for Center C

Buy Data for Center C

data blueprint

Sample Analysis Audit Flow
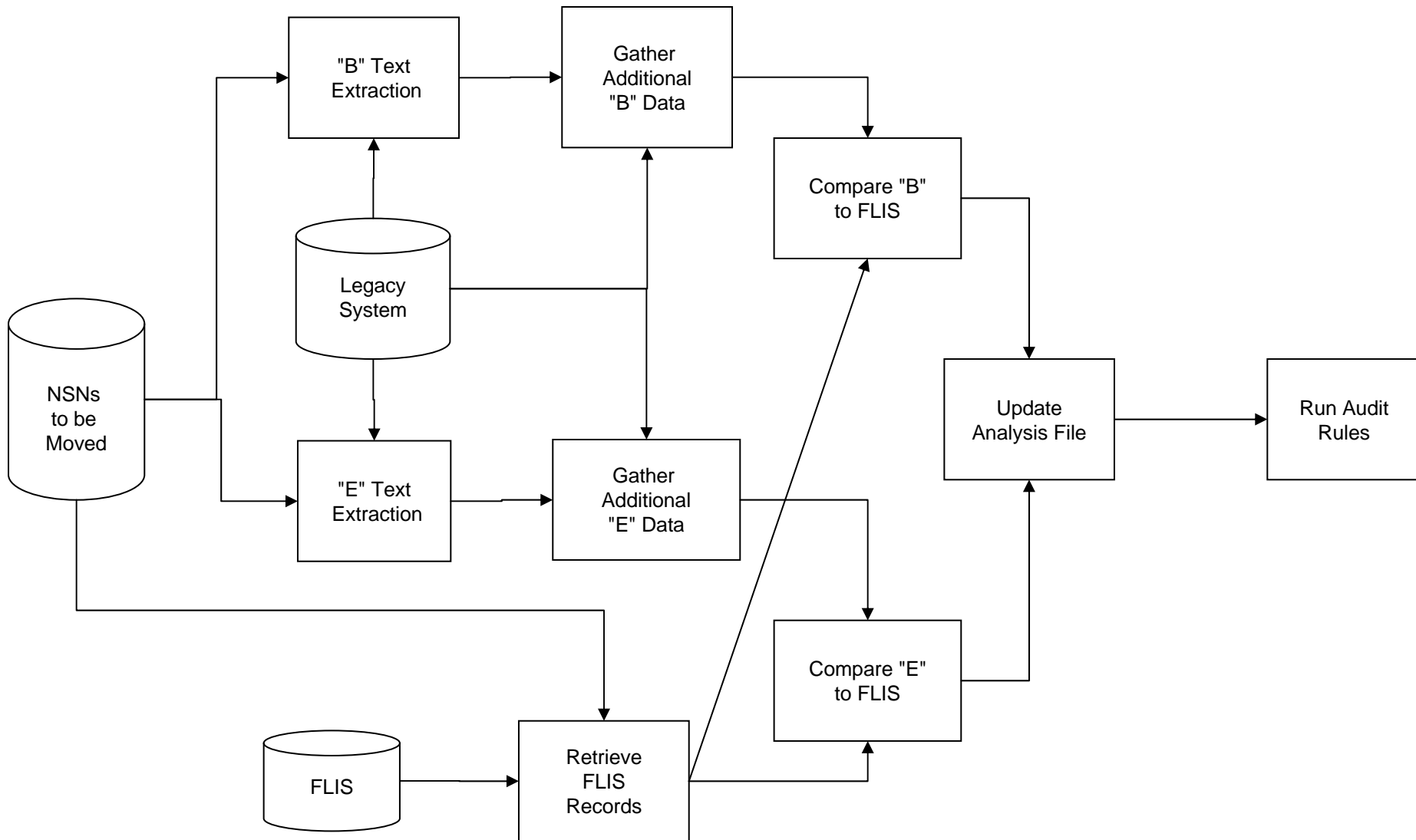
SAMMS Terminal Interface Option A

Corresponding Web-based Front-end

WebCTDF Architecture

End
User

Front End

Back End

Application API

Application Web Pages

Application Cache Tables

Legacy Data Mirror

Daily Update

Daily Cleanup

MQ-based Data Pull

SAMMS Cards

Transactions Table

SAMMS Cards

Legacy Application

data blueprint
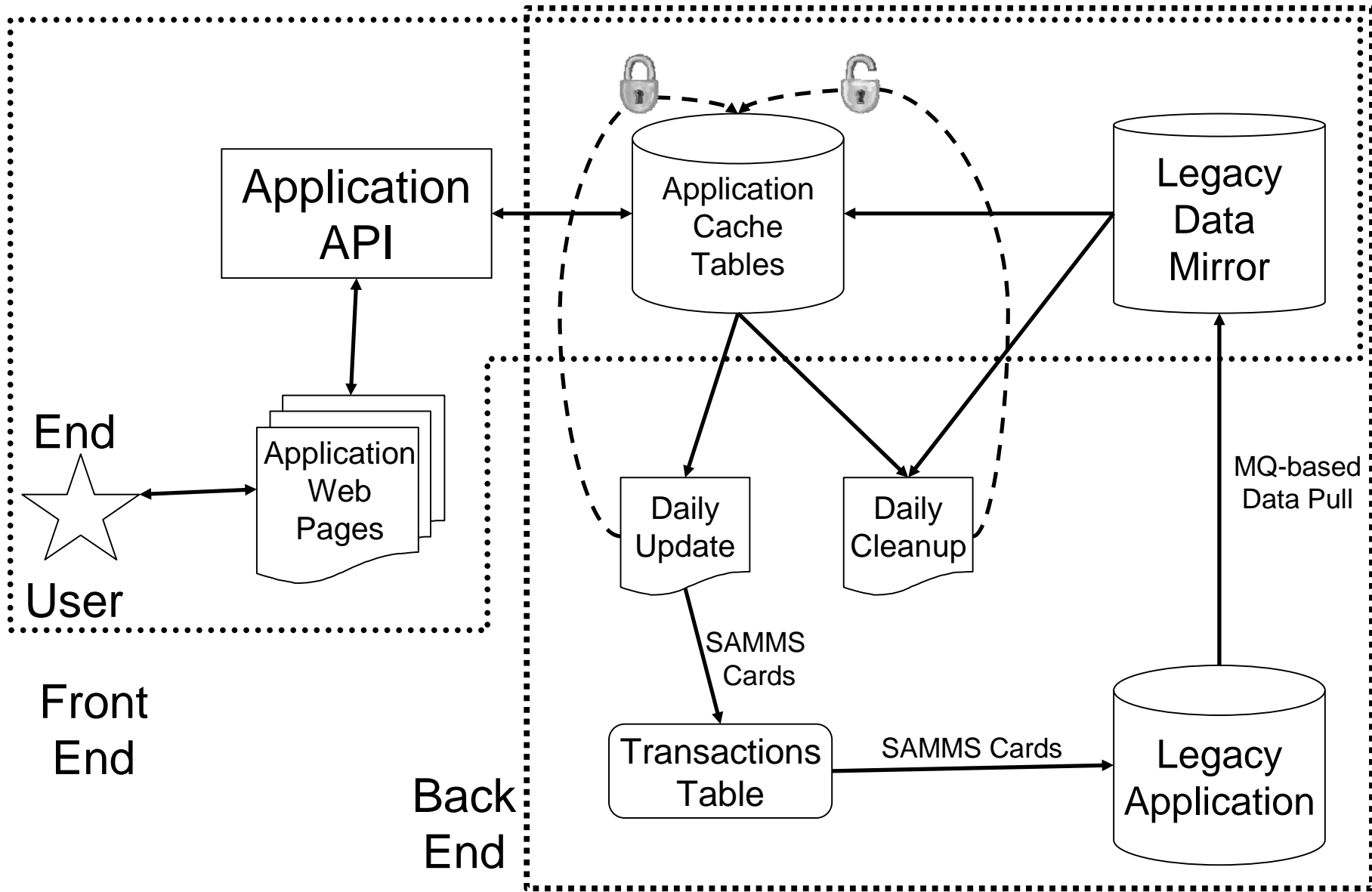
Quantitative Benefits

Time needed to review all NSNs once over the life of the project:	
NSNs	2,000,000
Average time to review & cleanse (in minutes)	5
Total Time (in minutes)	10,000,000

Time available per resource over a one year period of time:	
Work weeks in a year	48
Work days in a week	5
Work hours in a day	7.5
Work minutes in a day	450
Total Work minutes/year	108,000

Person years required to cleanse each NSN once prior to migration:	
Minutes needed	10,000,000
Minutes available person/year	108,000
Total Person-Years	92.6

Resource Cost to cleanse NSN's prior to migration:	
Avg Salary for SME year (not including overhead)	$60,000.00
Projected Years Required to Cleanse/Total DLA Person Year Saved	93
Total Cost to Cleanse/Total DLA Savings to Cleanse NSN's:	**$5.5 million**

data blueprint