



# Lessons Learned from Data Defects

Rick Rogers, President  
Fenestra Technologies Corp.  
(240) 535-2064  
rick@fenestra.com

“Mistakes are the portals of discovery” – James Joyce

# Outline

- Operational Context
  - Economic Census
  - GIDS Designer
  - GIDS Surveyor
- Data Defect Anecdotes
  - Major defect, needed to be fixed immediately
  - Minor defect, did not need to be fixed
  - Process defect, tricky to find and fix

# Economic Census

- Conducted every five years
- Surveys U.S. businesses on economic activity
- 3.5 million respondents
- ~800 questionnaires, average 12 pages each
- 2002 and 2007 Economic Census developed with “Generalized Instrument Design System”

# GIDS Designer (1)

- Generalized Instrument Design System (GIDS) Designer
  - Connects to content metadata database
  - Designs paper and electronic questionnaire layouts
  - Designs electronic behavior (“edits”)
  - Supports reusable questions (reduced designed pages from ~10,000 to ~1,600)
  - Used for the 2002 and 2007 Economic Censuses
  - Used for annual economic surveys

# GIDS Designer (2)

The screenshot displays the GIDS Designer interface for a custom-formatted question titled "PHYSICAL LOCATION". The question is presented in a table format with the following content:

PHYSICAL LOCATION			
<b>A. Is this establishment's physical location the same as shown in the mailing address? (P.O. box and rural route addresses are not physical locations.)</b>			
0031	<input type="checkbox"/>	Yes	0035 Number and street
0032	<input type="checkbox"/>	No - Enter physical location	0036 City, town, village, etc. 0037 State 0038 ZIP Code
<b>B. Is this establishment physically located inside the legal boundaries of the city, town, village, etc.?</b>			
0041	<input type="checkbox"/>	Yes	0042 <input type="checkbox"/> No 0043 <input type="checkbox"/> No legal boundaries 0044 <input type="checkbox"/> Do not know

The "Content:" pane at the bottom shows the following structure:

- Response [PHYSLOC\_ADDR\_CITY]
  - Keycode [0036]
  - Write-in [WRITE\_IN\_ANSWER\_128]
- Response [PHYSLOC\_ADDR\_ST]
  - Keycode [0037]
  - Segmented answer box [ANSWER\_SEGMENT\_2]
- Response [PHYSLOC\_ADDR\_ZIP]

The status bar at the bottom indicates: USER: QC\_OWNER EMR: EQCDMZ / 6.05.16.13.33.55 GIDS: 2.0.0.479

# GIDS Surveyor (1)

- Generalized Instrument Design System (GIDS) Surveyor
  - Computerized Self-Administered Questionnaire
  - Supports both form-view and spreadsheet-view
  - Validates responses through behavior “scripts”
  - Submits responses to servers through XML and web services
  - Respondents used GIDS Surveyor to provide 877,676 questionnaire responses in 2007

# GIDS Surveyor (2)

Surveyor 2007 (MA-10000(S)) - 2008 Report of Organization and Annual Survey of Manufactures

File Edit View Tools Help

Check for Software Updates Add Location(s) Add/Delete Product Classes Import from Spreadsheet Review All Forms Submit Responses

Welcome Inbox Form Workbook Errors / Warnings

**Form Types:**

- All Locations - Selected Items
- MA-10000(L)
- MA-10000(S)
- NC-99001(L)
- NC-99001(S)
- NC-99007
- NC-99530

Test Example 12  
Street Address 12  
Arlington, VA 22201

EIN: 123456789  
Store/Plant: 12  
CFN: 7000032008

Item 5: Total value of products shipped and other receipts (Report detail in Item 22)

Report in thousands of dollars. Enter "0" for none.

	I	J	K	L	M	N	O	P	Q	R
	ADDR_ST	ADDR_ZIP	EIN_NUM	RCPT_TOT	RCPT_TOT_PY	RCPT_ECO_MM	RCPT_ECO_MM_TOT_PCT	RCPT_ECO_MM_TOT_PCT_PV	EMP_MAR12_PRDWRK	EMP_MAR12_PRDWRK_PV
1	VA	22201	123456789	0	165			5		5
2	VA	22201	123456789		165			5		5

Mailing address plus Items 1-4 / Items 5-7 / Items 8-12 / Items 13-16 / Item 22 / Items 23-29 & Remarks

Review Panel Review Form

Kind	Item #	Explanation
Error	3	Please complete Item 3, <a href="#">Operational Status</a> .
Error	5	In order to submit forms in the manufacturing sector, you must complete Item 5, <a href="#">Total value of products shipped and other receipts</a> .
Error	7	In order to submit forms in the manufacturing sector, you must complete Item 7, line A3, <a href="#">Total number of employees</a> .
Error	7	In order to submit forms in the manufacturing sector, you must complete Item 7, line B1, <a href="#">Annual payroll before deductions</a> .
Warning	4	Please complete Item 4, <a href="#">Months in Operation</a> .



# Major Data Defect (1)

- Defect anecdote
  - We implement a performance enhancement in GIDS Surveyor in October 2008
  - Performance enhancement speeds loading of response data through “pre-fetch” algorithm that assumes a single questionnaire has no more than 10,000 responses areas
  - On December 20<sup>th</sup>, we receive reports of data loss in very large response data sets
  - We release a fix on December 22<sup>nd</sup>

# Major Data Defect (2)

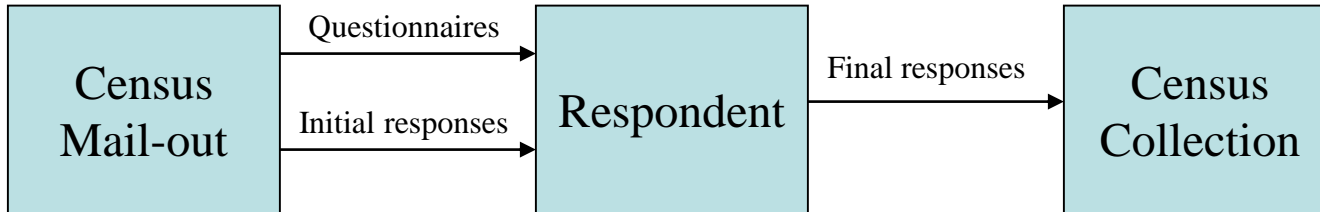
- Cause & Effect
  - Some “unusual” questionnaires had more than 10,000 response data elements
  - Responses “safe” in the local response database, just not included in the XML response file
  - Resubmission with updated software fixes the problem
- Lessons Learned
  - By necessity, performance enhancement often occurs late in the development cycle, leaving less time to uncover defects
  - In our experience, data defects that escape “to the wild” are often caused by performance enhancements

# Minor Data Defect (1)

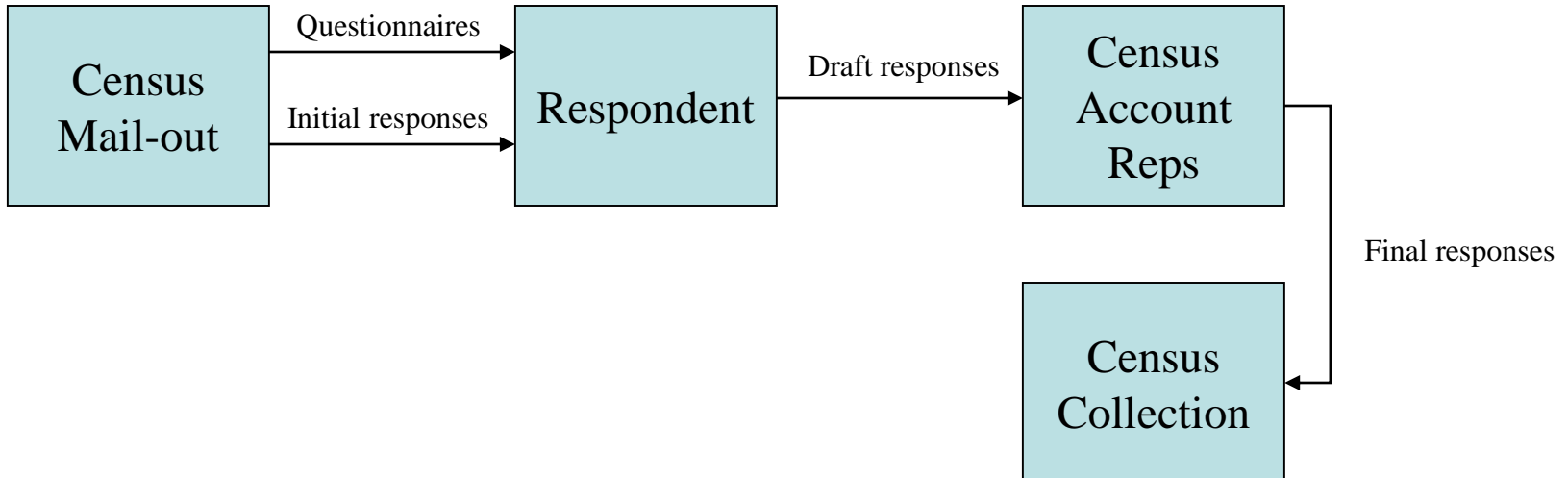
- Defect anecdote
  - On December 20<sup>th</sup>, we receive reports of incorrect metadata for certain cases
  - We determine that the defect occurs when Surveyor is used in an unexpected operational context
  - We release a fix on December 26<sup>th</sup>

# Minor Data Defect (2)

## Initial operational context



## “Real World” operational context



# Minor Data Defect (3)

- Cause & Effect
  - We recorded metadata to indicate responses were “final” too early in the process in the new operational context
  - Downstream data processing could have “massaged” the response data
  - We did release a fix to Surveyor
- Lessons Learned
  - Enumerate and thoroughly test all operational scenarios
  - Human nature focuses on the “normal” cases, ignores rare events (e.g., Nassim Taleb’s “Black Swan” events)

# Process Data Defect (1)

- Defect Anecdote
  - Census sends questionnaires with “pre-listed” responses based on prior period reporting
  - Surveyor maintains a “snapshot” of the original “pre-listed” responses in the response database
  - Surveyor uses several approaches to minimize XML response data set file size for transfer
  - We received report that in some cases, Surveyor was omitting responses incorrectly

# Process Data Defect (2)

<b>Data Element</b>	<b>Original Value</b>	<b>Current Value</b>
Company Name	“Fenestra Technologies”	(empty)
Contact Name	(empty)	“Rick Rogers”
Contact Phone	(empty)	(empty)

- Original (defective) algorithm
  - If Current Value is empty, do not write response data
- Corrected algorithm
  - If both Original Value and Current Value are empty, do not write response data

# Process Data Defect (3)

- Cause & Effect
  - Caused by an optimization in boundary case
  - Discovered through extensive regression testing
  - Rare: 12 cases out of several million
- Lessons Learned
  - Many defects occur in boundary or “edge” cases
  - Process data defects can be subtle
  - Sometimes “no fix” is “best fix”



# Lessons Learned

- Beware performance optimizations
- Foreseeing boundary cases is tough:
  - Rely on experience: use check-lists
  - Consider testing (or simulating) with random data
- Carefully impact of options for fixes:
  - Upstream, in data collection
  - Downstream, in data processing
  - No fix at all