# Standardizing Web Paradata Tools at the U.S. Census Bureau

Renee Ellis, US Census Bureau
Joanna Fane Lineback, US Census Bureau

Acknowledgements: A special thanks to the Sabin Lakhe, Alina Kline, Luke Larsen, and the Web Paradata Group whose work has contributed to this presentation.

# What are paradata?

- Data collected as a byproduct of the survey data collection process
- Often distinguished by mode
  - Field
    - Contact history information
    - Neighborhood observation information
    - Interviewer characteristics
  - Paper
    - Image captured from paper
    - Mail barcode tracking
  - Phone
    - Interview disposition information
  - Internet
    - Browser information
    - Time in survey or on particular questions
    - Patterns of movement through instrument (backing up, changing answers, clicking links, using help files)

# How are web paradata used?

To monitor data collection:

- Failed logins
- Completion rates
- Breakoff rates
- Browser counts
- Operating system counts
- Device type counts

For improving the survey instrument and user experience:

- Last action counts
- Last page counts
- Help access counts
- Response times
- Error/warning info
- Answer changes

# Challenges to the use of web paradata

- Policies
- Customization requests
  - Variables
  - Timing
- Unstructured data
- Comparability
  - Definitions
  - Formulas

# Why standardize?

- Standardization is not always the right idea, but in some cases it make sense for:
  - Reducing duplication of effort
  - Ensuring data quality
  - Comparability
  - Encouraging collaboration
- Issues facing the use of web paradata meet many of these criteria
  - Many surveys have an internet component using the same software (Centurion) therefore have the same underlying structure
  - Standardized programs can be adjusted for other software once they are created

# Goals

- Structured data
  1. Generic XML parser
  2. Useragent string parser
- Increased comparability across programs
  3. Define common terms
  4. Define common statistics
  5. Create generic programs and reports
  6. Develop other tools

# Unstructured nature of web paradata

- Web paradata are often in an unstructured, non-rectangular format
- May look something like-->

```
<event time="1443025340" type="login">
    <environment useragent="Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/45.0.2454.99 Safari/537.36" />
  </event>
  <event page="main_menu" time="1443025341" type="entry" />
  <event page="main_menu" time="1443025351" type="field_change">
   <field name="undefined" value="" />
  </event>
  <event page="main_menu" time="1443025385" type="exit" />
  <event page="main_menu" time="1443025385" type="hyperlink">
   <link url="https://respond.census.gov/ntpsrsc/main_menu" />
  </event>
```

Data source: Suzanne Dorinski's completely fake XML paradata file

# Parsed XML

- Want it to look something like→

| Obs | id | page | time | type | name | value | formatted_time |
|---|---|---|---|---|---|---|---|
| 1 | 99971 | login | 1340752727 | failed_login | | | 26JUN2012:19:18:47 |
| 2 | 99971 | login | 1340752747 | failed_login | | | 26JUN2012:19:19:07 |
| 3 | 99971 | login | 1340752822 | failed_login | | | 26JUN2012:19:20:22 |
| 4 | 99971 | login | 1340752946 | failed_login | | | 26JUN2012:19:22:26 |
| 5 | 99971 | login | 1340796017 | failed_login | | | 27JUN2012:07:20:17 |
| 6 | 99971 | login | 1340796400 | login | | | 27JUN2012:07:26:40 |
| 7 | 99971 | supplied_ | 1340796401 | entry | | | 27JUN2012:07:26:41 |
| 8 | 99971 | | 1340796405 | logout | | | 27JUN2012:07:26:45 |
| 9 | 99971 | supplied_ | 1340796405 | exit | | | 27JUN2012:07:26:45 |
| 10 | 99971 | supplied_ | 1340796405 | hyperlink | | https://asj-sherw010.centurion-qa.ssd.census.gov/asj /logout | 27JUN2012:07:26:45 |
| 11 | 99971 | login | 1340796491 | login | | | 27JUN2012:07:28:11 |
| 12 | 99971 | supplied_ | 1340796492 | entry | | | 27JUN2012:07:28:12 |
| 13 | 99971 | | 1340796494 | logout | | | 27JUN2012:07:28:14 |
| 14 | 99971 | supplied_ | 1340796495 | exit | | | 27JUN2012:07:28:15 |
| 15 | 99971 | supplied_ | 1340796495 | hyperlink | | https://asj-sherw010.centurion-qa.ssd.census.gov/asj /logout | 27JUN2012:07:28:15 |
| 16 | 99971 | login | 1340839936 | failed_login | | | 27JUN2012:19:32:16 |
| 17 | 99971 | login | 1340839957 | login | | | 27JUN2012:19:32:37 |
| 18 | 99971 | supplied_ | 1340839958 | entry | | | 27JUN2012:19:32:38 |
| 19 | 99971 | supplied_ | 1340840006 | field_change | supplied_ | Andy Taylor | 27JUN2012:19:33:26 |
| 20 | 99971 | supplied_ | 1340840014 | field_change | supplied_ | Sherri | 27JUN2012:19:33:34 |
| 21 | 99971 | supplied_ | 1340840038 | field_change | supplied_ | Sheriff | 27JUN2012:19:33:58 |
| 22 | 99971 | supplied_ | 1340840053 | field_change | supplied_ | andy@mayberry.org | 27JUN2012:19:34:13 |
| 23 | 99971 | supplied_ | 1340840116 | next_action | | | 27JUN2012:19:35:16 |
| 24 | 99971 | supplied_ | 1340840116 | entry | | | 27JUN2012:19:35:16 |
| 25 | 99971 | supplied_ | 1340840116 | error_trigger | | Please provide a valid phone number | 27JUN2012:19:35:16 |
| 26 | 99971 | supplied_ | 1340840146 | field_change | supplied_ | 301 | 27JUN2012:19:35:46 |
| 27 | 99971 | supplied_ | 1340840148 | field_change | supplied_ | 555 | 27JUN2012:19:35:48 |
| 28 | 99971 | dashboarc | 1340840154 | entry | | | 27JUN2012:19:35:54 |
| 29 | 99971 | supplied_ | 1340840154 | next_action | | | 27JUN2012:19:35:54 |
| 30 | 99971 | supplied_ | 1340840154 | field_change | supplied_ | 0 | 27JUN2012:19:35:54 |
| 31 | 99971 | dashboarc | 1340840533 | exit | | | 27JUN2012:19:42:13 |
| 32 | 99971 | dashboarc | 1340840533 | hyperlink | | | 27JUN2012:19:42:13 |
| 33 | 99971 | sec1/s1q1 | 1340840534 | entry | | | 27JUN2012:19:42:14 |
| 34 | 99971 | sec1/s1q1 | 1340840569 | field_change | s1q1_1 | 8 | 27JUN2012:19:42:49 |
| 35 | 99971 | sec1/s1q1 | 1340840572 | field_change | s1q1_3 | 2 | 27JUN2012:19:42:52 |
| 36 | 99971 | sec1/s1q1 | 1340840583 | field_change | s1q1_5 | 11 | 27JUN2012:19:43:03 |
| 37 | 99971 | sec1/s1q1 | 1340840586 | field_change | s1q2_1 | 1 | 27JUN2012:19:43:06 |

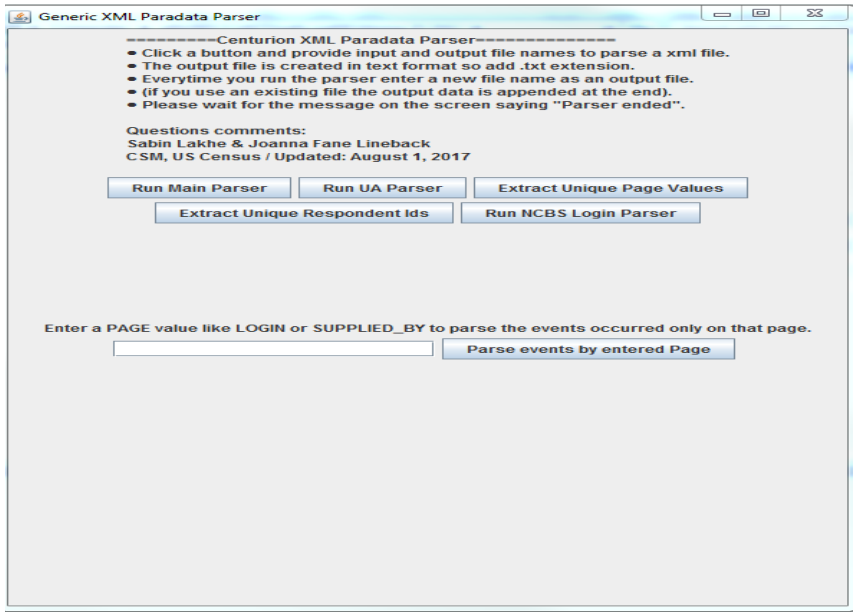
Data source: Suzanne's completely fake paradata file

# Why this is easier said than done

- For the XML file as a whole:
  - Conversion programs differ by software
  - People code differently
- For just the useragent string:
  - Tricky because most free software requires that we push this outside of our firewall for parsing, BUT the string is protected information, in most cases so we should not be doing this
  - Need software that can parse with the internet turned off
  - software requires regular updating to account for changes in technology

# XML parser

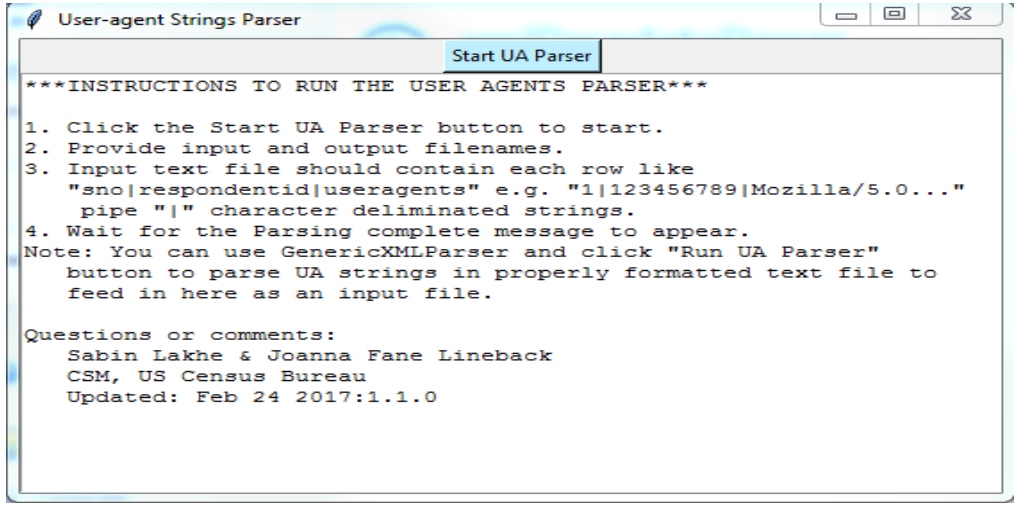
- Using Java, we created a "Generic Parser" with GUI interface
- Assumes XML output with predetermined attributes
- Creates two files: TXT and SAS




- There is also a version that can create more customized output, ingenuously called the "Customized Parser"

# Useragent string parser

- Developed a useragent parsing GUI using Python parsing tool




- Python tool being updated regularly to account for technology changes
- Go in once every 6 months or so to get the latest version
- Secure method because nothing is being pushed to the internet for parsing

useragent="Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/45.0.2454.99 Safari/537.36"

# Parsed Useragent String

| Useragent | BrowserFamily | BrowserVersion | OSFamily | OSVersion | DeviceFamily | DeviceBrand | DeviceModel | IsMobile | IsTablet | IsTouchCapable | IsPC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.86 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.80 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.80 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Linux; Android 5.0; M-G900V Build/LRX21T) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.76 Mobile afari/537.36 | Chrome Mobile | 46.0.2490 | Android | 5 | Samsung SM-G900V | Samsung | SM-G900V | True | False | True | False |
| lozilla/5.0 (Windows NT 6.1) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.86 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.86 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (iPad; CPU OS 8_4 like Mac S X) AppleWebKit/600.1.4 (KHTML, ke Gecko) Version/8.0 Mobile/12H143 afari/600.1.4 | Mobile Safari | 8 | iOS | 8.4 | iPad | Apple | iPad | False | True | True | False |
| lozilla/5.0 (Macintosh; Intel Mac OS X 0_7_5) AppleWebKit/537.78.2 KHTML, like Gecko) Version/6.1.6 afari/537.78.2 | Safari | 6.1.6 | Mac OS X | 10.7.5 | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/42.0.2311.135 afari/537.36 | Chrome | 42.0.2311 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like ecko) Chrome/46.0.2490.80 afari/537.36 | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |
| lozilla/5.0 (Windows NT 6.1; WOW64) ppleWebKit/537.36 (KHTML, like | Chrome | 46.0.2490 | Windows 7 | | Other | None | None | False | False | False | True |

Data source: Suzanne's completely fake paradata file

# Inconsistent use of terms

- Examples:
  - Error v. warning
  - Action v. event
  - Login v. session
  - Page v. screen
  - Mobile device v. Personal computer

# Mobile device v. personal computer

- Tablet - This would be part of the user-agent string; A mobile device that is a computer without a keyboard (defined further using industry standards reflected in the user agent string)

- Laptop - This would be part of the user-agent string; A portable personal computer (defined further using industry standards reflected in the user agent string)

# Formulas

- When calculating statistics, consistent use of terms is especially important
- In the previous example, it is an important distinction that we're treating a tablet as a mobile device and a laptop as a personal computer

Percentage using a personal computer=
[Count(PC)/Count(MD)+Count(PC)]*100

Percentage using a mobile device=
[Count(MD)/Count(MD)+Count(PC)]*100

# Taxonomy of formulas

- Information for monitoring
  - Useragent information
  - Login information
  - Completion and breakoff rates
  - Logout and breakoff information
  - Response time
- Information for research (includes info for monitoring, plus all below)
  - Detailed login information
  - Detailed logout and breakoff information
  - Detailed response time information
  - Error and warning messages
  - Accessing HELP
  - Answer changes

# Programs and reports

- Using the standardized definitions and formulas we can create programs that can be run on any parsed data
- This will allow a standard set of reports  to be generated
- Programs will be customizable for particular data needs.

| Device Information | | |
|---|---|---|
| | All respondents (850) | |
| | number | percent |
| **Device type** | | |
| **Mobile** | 200 | 24% |
| Mobile phone | 140 | 70% |
| Tablet | 60 | 30% |
| **PC** | 650 | 76% |
| Desktop | 200 | 31% |
| Desktop (touchscreen) | 25 | 4% |
| Laptop | 400 | 61% |
| Laptop (touchscreen) | 25 | 4% |
| | | |
| **Device Operating System** | | |
| Android | 10 | 1% |
| Chrome OS | 15 | 2% |
| IOS | 20 | 2% |
| Mac OS | 180 | 21% |
| Windows | 625 | 74% |
| | | |
| **Device Browser** | | |
| Chrome | 425 | 50% |
| Chrome Mobile | 45 | 5% |
| Firefox | 90 | 11% |
| Internet Explorer | 150 | 18% |
| Safari | 35 | 4% |
| Safari mobile | 105 | 12% |

Fake data by Renee Ellis

# Tools for other modes

- Contact History Users' Guide
- Metadata for other tables in the Census Bureau's research paradata database

# Future work

- Finalize definitions and formulas
- Provide central place for access to paradata information
- Provide programs that create reports based on finalized definitions and formulas.
- Other analysis tools

# For more information about Census Bureau paradata

- Census Paradata Users Group (CPUG) -

https://collab.ecm.census.gov/teamsites/paradata/SitePages/Census%20Paradata%20Users%20Group.aspx

- CSM paradata sharepoint (forthcoming)

- Contact: renee.ellis@census.gov