Can Publisher Data Play a Role in the Digital Advertising Ecosystem?

A Report to the Council for Research Excellence by Ernst & Young

April 4, 2012

cre
council for research excellence

CRE Mission

- To advance the knowledge and practice of methodological research on audience measurement through the active collaboration of Nielsen and its clients.

CRE Digital Committee

Chair:

Dan Murphy, Univision

Committee Members:

Brad Adgate, ShariAnne Brill, Cheryl Brink,
Michele Buslik, Nancy Gallagher, George Ivie,
Sherrill Mane, Daria Nachman, Carrie Nicholson,
Beth Rockwood, Bryon Schafer, Ceril Shagrin,
Beth Uyenco Shatto, Kate Sirkin, Debbie Solomon

CRE Digital Committee

- **Objective:**

  Advance the knowledge and practice of methodology explicitly as it pertains to Digital (including cross-platform) Research

- **Phase I:**

  Creation of an educational framework to serve as guidance to CRE investment

- **Phase II:**

  Study of Publisher Data Collection, Maintenance and Validation practices

Introduction to the study

- RFP to study Publisher Data Collection, Maintenance and Validation practices sent out in the first quarter of 2011.
- We had four responses and selected Ernst & Young.

Introduction to the study

- **Purpose and aims**
  - The goal of this study, commissioned by the Digital Committee of the Council for Research Excellence (Digital Committee), was to study how publisher data can play a role in supplementing panel demographic information to augment audience measurement.
  - In short, this is a "current state" study. The Digital Committee wanted to assess current data collection practices, commonalities, areas of opportunity and potential leading practices to enhance hybrid Digital Audience Measurement methodologies.

Introduction to the study

- **Privacy cautions**
  - At the outset, the Digital Committee sought input, through consultations with industry representatives, on privacy regulations (current and proposed) and the potential impact of those on this study. This was done to understand the current and near-future landscape and to confirm that the aims of this study and proposed topics and questions to be discussed would not be likely to place study organizers or participants at risk with respect to these regulations. Additionally, this consultative effort was taken to provide potential input into the project findings and recommendations, and to minimize the likelihood of study recommendations that conflict with current or foreseeable near-term privacy regulations.

Introduction to the study

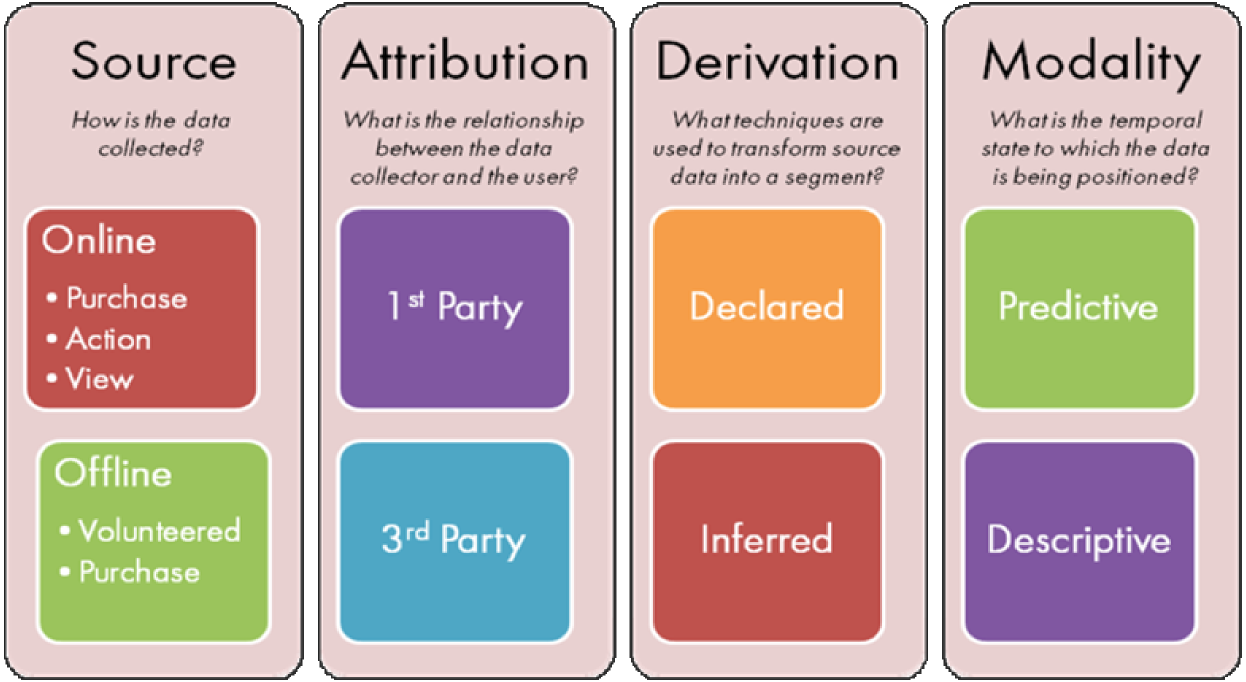
- **Ernst & Young team**
  - Ernst & Young's Media Research Assurance Services practice was selected to perform the study, and the following team worked with the Digital Committee throughout the process:
    - Jackson Bazley, Executive Director
    - Joe Bailey, Senior Manager
    - Nicole Kuntz, Manager
    - Shazad Muneer, Manager
  - EY formally began the effort with the CRE on July 28, 2011, with an internal kickoff meeting, and interviews began in early September, 2011.

Introduction to the study

- **Fielding the study**
  - User registration forms for approximately eighty publishers were identified based on a combination of their audience reach and other judgmental factors, and the data elements collected within each were noted.
  - Separately, working with the CRE Digital Committee working groups, thirty publishers were identified as digital leaders in content including Tech Media, Digital Video, Social Media, News and eCommerce, and were invited to participate in the study.
  - A total of twenty interviews were completed and the results of those were anonymized and aggregated and presented.

Introduction to the study

- IAB Data Segments & Techniques Lexicon



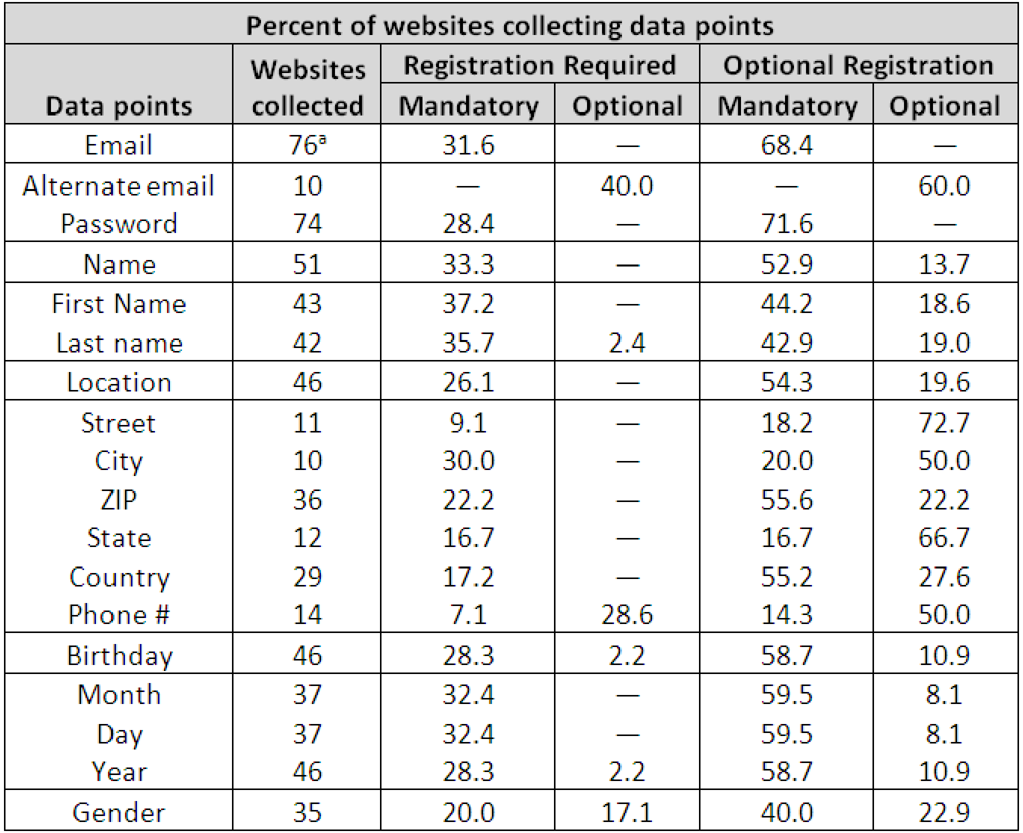
Source	Attribution	Derivation	Modality
How is the data collected?	*What is the relationship between the data collector and the user?*	*What techniques are used to transform source data into a segment?*	*What is the temporal state to which the data is being positioned?*
Online • Purchase • Action • View	1st Party	Declared	Predictive
Offline • Volunteered • Purchase	3rd Party	Inferred	Descriptive

Source: Interactive Advertising Bureau (http://www.iab.net/data_lexicon)

Key findings

- **Publishers are participating in data collection**
  - Nearly all of the participants in this study are engaged in some form of data collection with respect to their users. Many use some form of user registration, whether required or optional, through which their users provide declared data.
  - *User registration forms for approximately eighty publishers were identified and the data elements collected within each were noted.*

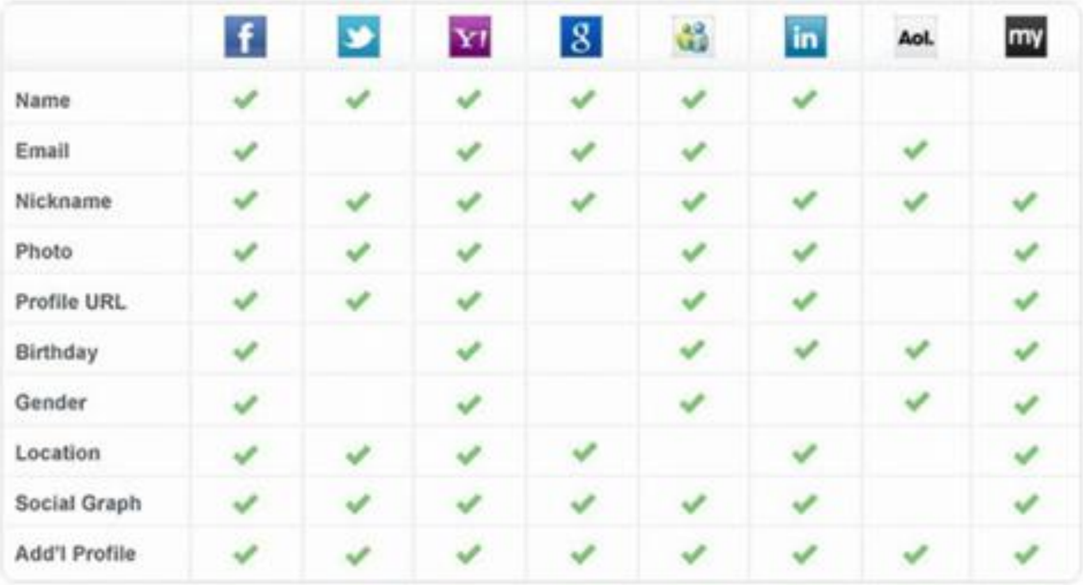
Summary of registration data



Percent of websites collecting data points					
Data points	Websites collected	Registration Required		Optional Registration	
		Mandatory	Optional	Mandatory	Optional
Email	76[a]	31.6	—	68.4	—
Alternate email	10	—	40.0	—	60.0
Password	74	28.4	—	71.6	—
Name	51	33.3	—	52.9	13.7
First Name	43	37.2	—	44.2	18.6
Last name	42	35.7	2.4	42.9	19.0
Location	46	26.1	—	54.3	19.6
Street	11	9.1	—	18.2	72.7
City	10	30.0	—	20.0	50.0
ZIP	36	22.2	—	55.6	22.2
State	12	16.7	—	16.7	66.7
Country	29	17.2	—	55.2	27.6
Phone #	14	7.1	28.6	14.3	50.0
Birthday	46	28.3	2.2	58.7	10.9
Month	37	32.4	—	59.5	8.1
Day	37	32.4	—	59.5	8.1
Year	46	28.3	2.2	58.7	10.9
Gender	35	20.0	17.1	40.0	22.9

Summary of registration data

- **User data collection**
  - May be via site registration or through a link to another company with a login process (e.g., “log in with”).



	f		Y!	g		in	Aol.	my
Name	✓	✓	✓	✓	✓	✓		
Email	✓		✓	✓	✓		✓	
Nickname	✓	✓	✓	✓	✓	✓	✓	✓
Photo	✓	✓	✓		✓	✓		✓
Profile URL	✓	✓	✓		✓	✓		✓
Birthday	✓		✓		✓	✓	✓	✓
Gender	✓		✓		✓		✓	✓
Location	✓	✓	✓	✓		✓		✓
Social Graph	✓	✓	✓	✓	✓	✓		✓
Add'l Profile	✓	✓	✓	✓	✓	✓	✓	✓

Summary of registration data

- **User data collection**
  - Data collection noted in this study tended not to include richer user data (e.g., employment, education, marital status, income) or more complex target definitions (e.g., interests, purchase intent, "Do-It-Yourself") that advertisers may be accustomed to in other media.

Key findings

- **The level of data collection varies greatly among publishers**
  - Very few of the participants require users to register and provide declared data; however nearly all employed optional user registration. Within the user registration, a few data elements were nearly universal, but beyond that the amount and type of user data collected varied greatly. A number of participants noted that there is a need to provide the user with a reason or value to provide user data in order to obtain quality information.

Key findings

- **Few publishers are currently utilizing this data in a meaningful way**
  - Only a few of those surveyed have developed techniques for utilizing the data collected from their users for audience measurement, advertising targeting, content refinement or other purposes in a meaningful way. Further, very few appear to provide this information to external parties.

Key findings

- **Certain data elements have multiple definitions**
  - Geography can be collected via declaration such as in a registration form, inferred from IP Address or inferred from the context of user activity such as on a travel or weather site. Each represents a different and potentially accurate and valuable geographic association for the user, such as “home” (declared), “current” (inferred from IP Address), and “interested” (inferred from context). While a restaurant may look to target advertising based on a user’s current or interested location, a car dealer may only be interested a user’s home location. Additionally, media companies and pharmaceutical companies may only be interested in the user’s current location for blackout or legal compliance reasons.

Key findings

- **Email address and ZIP Code are key user data elements**
  - Where publishers are utilizing external data enrichment sources to obtain or provide additional user profile data, email address and ZIP Code appear to be two key variables utilized for this enrichment purpose.

Key findings

- **Data flow from third-parties is generally unidirectional**
  - Some publishers noted the use of third-parties as data sources through processes such as social log-ins, cookie or email enrichment, or other techniques, but very few indicated providing first-party collected user data to external third-party data sources.

Key findings

Third Party Data Sources Mentioned

Audience Science
comScore
Digg
Digital Envoy
eXelate
Experian
Facebook
FourSquare
Google Analytics
MySpace
Nielsen
Omniture
Quantcast
Quova
RapLeaf
Twitter

Key findings

- **Publishers expressed a lack confidence in third-party data**
  - A number of participants indicated they have concerns with the quality and accuracy of user profile data acquired from third-party data sources.
  - Unlike first-party data collection practices, we were unable to determine the derivation method, declared or inferred, of third-party data as this study did not include interviewing third-party data providers.
  - *Potential action: conduct a similar or related survey of the major third-party data suppliers to understand their data collection processes, data validation practices, and level of transparence provided to their users.*

Key findings

- **Data conflicts can and do occur, however very few publishers have resolution policies**
  - A number of the data elements may be collected in different ways (first party vs. third party and declared vs. inferred) and may conflict. Additionally, as noted in the geography example discussed above, these differences may also identify different valid characteristics for the same user. Very few of the participant companies had established resolution policies or methodologies to address these situations.
  - *Potential leading practices noted related to data conflict resolution policies.*

Key findings

- **Publishers currently maintain minimal data quality practices**
  - The majority of participants noted minimal or no formal data quality and validation practices with respect to user data. The most common technique cited was cross-validation with additional sources, but this process was used by a minority of the participants. Without robust data validation practices, declared data cannot be assumed to be more or less accurate than inferred data. For example, without robust data quality procedures, declared geography may be fictitious, thus IP enrichment may identify a more accurate geographic assignment for that user.
  - *Potential leading practices noted related to data quality practices.*

Key findings

- **There is no common "data owner"**
  - Among the participants, there was no common department or function within the company that "owned" or controlled the user data. In some cases, these functions were decentralized with different departments or functions owning different components of the overall user data profile.
  - *Potential leading practices noted related to centralized ownership.*

Key findings

- **Publishers have an inconsistent expectation of future, external use of these data**
  - Among the participants, several indicated they thought there was future potential in expanding the use of geographic information in targeting. Others indicated an expectation of future use of combined data elements in a targeting profile. Still others noted a future use of behavioral, interest or intent information in targeting. Very few identified external audience measurement systems as a potential future use of their user data. However, there was no clear common expectation on how their user data may be leveraged externally in the near-term future.

Key findings

- **Publishers do not see advertisers seeking to target on these richer data profiles**
  - A number of participants indicated that the potential to leverage their user data externally was inhibited by a lack of interest or sophistication on the buy-side related to targeting on these richer targets. Publishers appear to be reluctant to develop their processes until they have a better sense of what the buy-side can and wants to buy.
  - *Potential action: conduct a similar or related survey of the major buy-side parties to understand their views on user data and the use of such for targeting or other purposes.*

Potential leading practices

- Data edits/validations at the time of collection to determine if the response is valid in the context (e.g., a valid ZIP code based on reference to a USPS database).
- Review of declared data for illogical or suspect responses. For example:
  - Selection of January 1 for birth date,
  - Selection of 12345, 90201, or other common ZIP codes for location,
  - Selection of the first option in any drop-down selection field, etc.

  These techniques may not initially identify the individual users whose data is inaccurate, but in total may highlight specific response data as suspect allowing for additional validation processes focused on those users and data elements, such as consideration of the preponderance of IP Address inferred locations as compared to declared location.

Potential leading practices

- Data validation techniques initiated by and focused on user changes to their profile data.
- Cross-validation techniques employing external or alternate data sources.
- Defined process to address data conflicts across collection methodologies and parties.
  - Data quality procedures that pre-identify potential conflicts among multiple sources, and a policy such as a data hierarchy.
- Ability of users to review their collected user data, so they can update or correct it, if necessary or possibly remove it from their profile.

Potential leading practices

- A data “Time To Live” (TTL) policy that considers the different data types, association (first or third party sources) and derivation (declared or inferred) for each element and establishes a TTL for that data, at which point the data must either be refreshed or discarded.

- Centralized function to oversee data collection, quality and use across the organization, such as a research or CRM function.
  - Among the survey participants, those that indicated they currently have some level of centralized function such as these, also tended to have more of these potential leading practices currently in place.

Questions and comments

Thank you to
Participating Publishers