Can Publisher Data Play a Role in the Digital Advertising Ecosystem

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

Introduction to the Study

The mission of the Council for Research Excellence (CRE) is to advance the knowledge and practice of methodological research on audience measurement through the active collaboration of Nielsen and its clients. Since its founding, the CRE has identified areas that require exploration and has sought proposals from independent researchers, institutions and research companies to design, execute and report the findings of their research.

The objective of this study, commissioned by the Digital Committee of the Council for Research Excellence (Digital Committee), was to advance the knowledge and practice of methodology explicitly as it pertains to Digital (including cross-platform) Research.

A Request For Proposal to study Publisher Data Collection, Maintenance and Validation practices was sent out in the first quarter of 2011. Four responses were received and the CRE Digital Committee selected Ernst & Young LLP. The study was conducted for the CRE by a team led by Jackson Bazley, Ernst & Young Executive Director, Media & Entertainment Advisory Services. Ernst & Young formally began the effort with the CRE on July 28, 2011, with an internal kickoff meeting, and interviews began in early September, 2011.

The Digital Committee sought to better understand the various data collection and maintenance approaches, and to identify strengths and weaknesses in current approaches to capturing and retaining the various types of user data. As such, the Digital Committee commissioned this effort to study how publisher data might 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.

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.

The study included two primary data collection methodologies: data gathering from the public domain regarding user registration practices and the execution of a survey with entities that maintain a relationship with internet users (typically, publishers and content owners, collectively referred to as publishers herein). 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, the following 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:

Disney/ABC Television Group
Amazon.com
AOL
CBS Interactive
Cox Media Group
Discovery Communications
eBay
ESPN
Facebook
FourSquare
FOX Broadcasting Company
Google
Groupon
Hearst Television
Hulu
Linked-In
Microsoft
MTV Networks
NBC Universal
The New York Times Company
Raycom Media
Scripps Networks
The Washington Post
Turner Broadcasting Corporation
Twitter
Univision Interactive Media
USA TODAY
The Wall Street Journal
Warner Bros. Television
Yahoo! / Yahoo! Right Media

A total of twenty interviews were completed from among this group and the results of those interviews were anonymized, aggregated and presented within this report.

Executive Summary

The study focused on four broad topical areas, related to publishers' data collection practices:

1. What data are you collecting related to your user base?
2. How are you collecting that data?
    - Are any external parties / third-parties utilized?
3. Do you have data quality policies related to this data?
4. What do you do with it / How do you use it?
    - Are any external parties / third-parties provided with any data?

Key Findings

Based on the information collected through the surveys and summarized herein, we identified the following that we considered key findings from this study:

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.

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 utilizing in other media.

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.

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.

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 in 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.

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.

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.

Publishers expressed a lack of 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 interviews with third-party data providers.

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.

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.

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.

Publishers have an inconsistent expectation of future, external use of this 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 clearly common expectation on how their user data may be leveraged externally in the near-term future.

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 Leading Practices

Based on the information collected through the surveys and summarized herein, we identified the following practices or processes that may be considered leading practices, and should be considered when discussing ways to move the industry forward related to leveraging publisher data in audience measurement and targeting:

- 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, 90210 or other common ZIP codes for location,
  - Selection of 867-5309 for telephone number,
  - Selection of the first option in any pre-populated selection field, etc.

  These techniques may not initially identify the individual users whose data is inaccurate, but in total they 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.
- 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.
- A data “Time To Live” (TTL) policy that considers the different data types, association (first-party 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 tended to have more of these potential leading practices currently in place.

Survey Results

As noted previously, the study focused on four broad topical areas, related to publishers' data collection practices:

1. What data are you collecting related to your user base?
2. How are you collecting that data?
    - Are any external parties / third-parties utilized?
3. Do you have data quality policies related to this data?
4. What do you do with it / How do you use it?
    - Are any external parties / third-parties provided with any data?

Data collection and use

Leveraging the Data Segment & Techniques Lexicon published by the Interactive Advertising Bureau Data Council (the "IAB Data Lexicon"), attached as Appendix A to this report, we summarized the responses from the participants related to their data collection practices below.

Across the interviews conducted, we noted the use of first-party declared, first-party inferred and third-party data collection methodologies.

As defined by the IAB Data Lexicon, first-party data collection refers to situations in which the user data is being provided to, or collected by, the entity that owns or controls the website or service that the user is interacting with, whereas third-party data collection refers to data collection from or about users by parties that do not own or control the website or service that the user is interacting with. The IAB Data Lexicon also defines two derivation methods: declared and inferred, where declared data is that which is directly provided by the user, or captured from user actions with no inferences being made, and inferred data is then derived by inferring attributes from observed behaviors.

Among the survey participants, we noted that in many cases, an initial challenge to conducting the interview was determining who in the organization would be appropriate to interact with us in performing the interview, and in several cases, multiple individuals from different functions in the organization participated. As we conducted the interviews, we further noted that there was no common owner of the user data collection and use policies and practices within the organizations. In some cases this might lead to user data collection that does not align with intended use practices, such as making a key attribute optional in the registration form, or omitting it from the registration form altogether. In other cases, it may lead to redundant data collection from the same user, impacting user experience.

First-Party Declared

As noted previously, nearly all of the survey participants are participating in first-party data collection, including collection of declared data. First-party declared data collection techniques

noted across the participant group included user registration processes, newsletter signup, contest entry, user surveys, user agent string processing and analysis of content consumption patterns. The IAB Data Lexicon includes information captured from user actions with no inferences made to be declared data, and includes the example of "users who consume sports content" to illustrate this designation.

In the next section of this report, we have summarized the user data collected through user registration processes across approximately eighty different publishers, including but not limited to study participants.

Among the survey participants, and the website registration processes catalogued, data collection noted in this study tended to focus on more basic or traditional user data points and did not appear 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"), except for a few, limited situations.

In addition to the user data elements collected via registration, we noted certain survey participants that collected first-party declared data in the area of interests, intentions and advertising relevance primarily via user survey methods; user technographic information including operating system, browser type, and device type from user agent strings, and content consumption patterns across the site.

While nearly all of the participants collect declared data from their users using some of all of these techniques, there was variance among the group in terms of the techniques utilized and the data elements collected. The majority of participants utilized some form of user registration, either alone or in conjunction with other declared data collection methods, and very few performed declared data collection without using any form of user registration process.

Participants noted several issues related to user registration processes, including a general unwillingness to require user registration to access content. In general, the participants in the study indicated that requiring user registration either prompted users to abandon the site and seek the content elsewhere, or resulted in a higher level of suspect user data being collected. Related to this, many publishers acknowledged a need to balance between the desire for rich user data with the need to provide users with some value in exchange for providing that information, in order to increase the users' likelihood of providing data, and providing accurate declared data.

Lastly, few of the survey participants indicated having developed techniques for utilizing the first-party declared 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 any of their first-party collected data to external parties.

First-Party Inferred

While the majority of survey participants indicated they were currently participating in collecting declared user data from their users, fewer noted using inferred data collection techniques.

First-party inferred data collection techniques noted across the participant group included IP Address enrichment, TCP/IP handshake data, and analysis of content consumption patterns in ways that do include making inferences about interests or intents.

As with the first-party declared data, few of the survey participants indicated having developed techniques for utilizing the first-party inferred data for audience measurement, advertising targeting, content refinement or other purposes in a meaningful way. Further, very few appear to provide any of their first-party collected data to external parties.

Third-Party

Third-party data collection methodologies included social log-ins, cookie enrichment, email address enrichment, IP Address enrichment, surveys and other techniques. Social log-ins describes a process by which a user can "log in" to a publisher's site using their log-in from another source such as (but not limited to) Facebook, Twitter, Yahoo!, Google, MSN, LinkedIn, AOL or MySpace. We have included as Appendix B to this report a diagram obtained from the Gigya website that outlines the potential user data attributes that a publisher may be able to access, by source, if they utilize the social log-in process. Among the survey participants, we noted some use of this technique; however, very few of the survey participants obtained user data from this process. That is, they permitted users to log-in with these other accounts but did not obtain third-party data from these relationships.

Unlike first-party data described above, 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.

Among the survey participants, a number of third-party data sources were mentioned, including: Audience Science, comScore, Digg, Digital Envoy, eXelate, Experian, Facebook, FourSquare, Google Analytics, MySpace, Nielsen, Omniture, Quantcast, Quova, RapLeaf and Twitter. Other survey participants indicating using third-party sources, but declined to name them, so this listing likely does not encompass all third-party data sources used across the survey participants. The mentioned third-party data sources included a variety of data types, including audience measurement, IP enrichment vendors, cookie enrichment vendors, email enrichment vendors, social log-in partners, and others.

<u>Data quality</u>

In addition to data collection and use, the study also asked survey participants about data quality policies related to user data, as well as their experiences with data quality.

Among the survey participants, we noted that in many cases certain user data was available from or collected in multiple ways. This could include declared versus inferred data collection, first-party versus third-party data collection, and even multiple first-party declared data collection techniques by the same entity (for example, a registration form and contest entry form).

Given that the same user data can be obtained in multiple ways, it was further noted that data conflicts can and do occur, and that certain data elements may have multiple definitions, leading to different but valid values for the same data element.

In the case of geography for example, a user might declare a location by providing a ZIP code in a contest entry form, while their inferred location using an IP enrichment vendor may indicate a different location. Another form of inferred geographic location assignment noted in the study is that of inferring location based upon the context of user activity -- such as on a travel or weather site -- and this variable may constitute a third "location" for the user. Each of these 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 in 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.

Other situations where different values are noted for a single user attribute may indicate situations of suspect data or data inaccuracy. These data conflicts can cause difficulty in leveraging the user profile data in meaningful ways -- unless data quality policies are in place to minimize suspect data, and to address data conflicts. Among the survey participants, we noted very few situations of data validation or resolution techniques or policies.

We noted few situations in which declared user data was validated against external sources at the time of data collection--such as passing a five-digit number given for ZIP code against a listing of valid ZIP codes. Requiring the given response be a five-digit number does not necessarily verify that the five-digit number is a valid ZIP code -- thus, the user declared data could be invalid. The most common data validation technique noted was cross-validation wherein the user data would be compared across multiple sources or means of collection to determine if the result was consistent. However, this technique was used by a minority of the survey participants.

Other data validation techniques include vigilance for illogical distributions such as a user base from ZIP Code 90210 that is significantly larger, on either an absolute or relative basis, than the population of ZIP Code 90210. This could also include a review of responses that represent the first response option in a pre-listed set of values. Methods such as these may not necessarily identify which users have provided fictitious responses. However, when taken together with other data validation methods, they may identify profiles with numerous situations of suspect data that may in turn provide the company with reason to suspect the overall user-declared data.

In situations where the same user data is collected from multiple parties or in multiple ways, the results are not always consistent. Among the survey participants, we noted minimal formalized resolution policies. Data resolution policies would not necessarily be designed to validate the different data sources or external data. Rather, they would establish how data conflicts should be addressed, and would likely vary based on the data element itself (e.g., a different approach may be taken related to geography conflicts than would be taken for gender conflicts). In some

cases, a hierarchy may be established indicating which source would be used in the case of a conflict. In other cases, a policy may indicate that the attribute would be classified as “unknown” – as in the case of conflicting data. Recency of data collection may also be a factor used in a conflict resolution policy.

Our discussions with survey participants also noted that without robust data validation policies, a data conflict policy should not necessarily assume declared data to be superior to inferred data. For example, a user may indicate a fictitious ZIP code of 13579, and thus an inferred location from IP Address enrichment may provide a more accurate geographic assignment for that user.

Summary of Registration Data Collected

User registration forms for eighty-six publishers were identified based on a combination of their audience reach and other judgmental factors. Of the 86 websites inspected, eight had no user registration process and two relied solely on third-party registration (they did not have a user registration process of their own).

The table below presents the data elements collected by the 76 websites where some form of user registration data was noted. For each data point, the total number of websites collecting that data is noted (for example, 76 of 76 collected email address), as is the registration means by which it was collected. Whether the site made it mandatory or optional for users to register in order to meaningfully interact with the site also was considered. Within each form of registration, certain questions were noted as either mandatory (to submit the registration), or optional. While all 76 websites collected email addresses, 31.6% of those collected email as a mandatory data point within a required registration -- whereas 68.4% collected email as a mandatory data point within an optional registration. (In other words, users could elect not to register, but if they did register, email address was a mandatory field to submit the registration.)

Percent of websites collecting data points					
	Websites	Required Registration		Optional Registration	
Data points	collected	Mandatory	Optional	Mandatory	Optional
Email	76	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%

NOTE: Rows may not add to 100.0 due to rounding.

In addition to the above commonly collected data points, a limited number of sites were identified as collecting additional user data, such as: annual income, company size, country of origin, education, industry, internet access, job title, language, presence/number of children, relationship status, and user interests and shopping preferences.

Finally, we noted forty-six of the eighty-six websites that allowed users to sign in (social sign-in) using at least one alternative company account, and as many as seven alternative company accounts.

Following is a list of the social sign-in options noted among these forty-six websites:

- AOL,
- Apple ID,
- Facebook,
- FriendFeed,
- Google,
- Linkedin,
- MySpace,
- Twitter,
- Windows Live, and
- Yahoo!

Appendix A

Data Segments & Techniques Lexicon: Overview

Data can be confusing, especially with differing definitions from media companies and data providers. For example, most companies have a proprietary method for creating an "auto intender" segment. While the definition of an "auto intender" will differ, by using the language below, there can be consistency in how data sets are described. This one-pager is an overview of the longer Data Segments & Techniques Lexicon, a new standard from the IAB Data Council to specify how data segments are defined. The menu below is a tool to help better communicate with media and data partners when enhancing online advertising campaigns with data.

Data Segments & Techniques

There are 4 broad descriptors (Source, Attribution, Derivation, and Modality) to define data segments & techniques and then specific choices within each. Every data segment should be defined by one or more attributes from all four descriptors.

Source: Where is the data acquired?

- Online Commerce – Online purchase history
- Online Action – User-initiated, supplied, and/or completed an action, filled out a form, searched, etc.
- Online View – User-visited content
- Offline Commerce – User transaction data from offline purchase
- Offline Collected – Data from offline databases (public records, surveys)

Attribution: What is the relationship between the data collector and the user?

- 1st Party – Gathered by owner or controller of the web site
- 3rd Party – Gathered by any other party

Derivation: What techniques are used to transform source data into a segment?

- Declared – Derived from information directly provided by the user
- Inferred – Derived from observed behaviors or by analyzing behavior patterns

Modality: What is the temporal state to which the data is being positioned?

- Predictive – Indicative of future action or state
- Descriptive – Indicative of a current or past action or state

For more information, more detailed descriptions, and examples, please visit www.iab.net/data_lexicon

This Data Segments & Techniques Lexicon has been developed by the IAB Data Council.

Appendix B

Below is a diagram obtained from the Gigya website that outlines the potential user data attributes that a publisher may be able to access, by source, if they utilize the social log-in process. Among the participants, we noted some use of this technique;

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