**Why is audience measurement research important?**

The discussion in W3C DNT is naturally focused on the technical specifications for enabling people to control the extent to which their surfing behaviour is tracked for the purpose of delivering tailored content to them as an individual. The audience measurement research (AMR) industry has been concerned that a specification designed to deliver effective user control of tracking could inadvertently render accurate audience measurement of online content impossible.

Accurate audience measurement is important because the pricing of all media is determined by the characteristics and size of the audience it reaches. The online medium does not operate in a vacuum, it competes for revenue with other media. AMR allows publishers to understand their audience and optimise their content to meet the needs of their consumers; advertisers use the data to choose media channels and brands to reach their potential customers in the most cost-effective way.

Reliable, independent, audience measurement data enables media markets to function effectively. It stimulates diversity of provision of content and also effective competition to hold down prices. It enables niche players to compete on a level playing field with major providers and it enables cross media campaign planning, allowing advertisers to optimise their expenditure across different media e.g. online, TV, radio etc.

**How audience measurement research is used**

Advertisers, whether global brands, government or local trades people, usually want to reach particular sectors of the population; a car manufacturer like Mercedes or BMW may want to remind people who are thinking of buying a new car to consider buying their latest model. They will want to reach people who are sufficiently wealthy to be able to afford a luxury car. A government agency wanting to advise people on healthy eating may wish to reach people who are heavy consumers of junk food and a local firm providing household services like plumbing may want to reach people who own their own property.

In the non-digital world, AMR makes it possible to identify the media brands where there are likely to be users who are the advertiser’s potential customer; an obvious example for luxury cars would be magazines about cars, but also these people probably consume content from media which specialise in sailing, high end fashion, golf etc. In the non-digital environment this information is collected by research surveys and panels of people who have consented to provide information about their media consumption and also about their behaviour, expenditure and interests. The output of the surveys and panels is usually presented as a number of viewers, listeners or readers.

This is done by scaling up each survey respondent by a factor which represents their proportion of the total potential audience. For example, a panel of 50,000 in a universe of 50 million would involve each individual representing 1,000 people. Using this data, a campaign can be designed and evaluated by adding together the audiences of a number of different media brands. Diversity in the range of media brands and programmes produced is encouraged, because more closely focused programmes or specialist magazines attract particular audiences to an environment where relevant marketing content is more likely to get the attention and engagement of the user.

Publishers are interested in analysing their aggregated audiences, so that they can explain its size and character to people who might pay to insert content. However, advertisers **not only wish to know the size of the audience** who will be exposed to their content but also **the number of unique viewers** and **the number of times they will be exposed to understand the effectiveness of their content. I**n the offline world this exposure will be obtained via placement in different publications and programmes over what can be quite a long period of time; too much exposure of the same content to a single individual is a waste of money, but too little may mean they do not notice and potentially absorb the message.

In the online world Online Behavioural Advertising (OBA) provides the real time facility to do this type of targeting on an individual basis but, **in the future, this will only be possible if an individual agrees to provide their data for tracking purposes**.

AMR has a different function as it provides reliable data for media buying and selling and campaign evaluation. It does not need to identify specific individuals in order to do this, but it does need to be able to include in the counts the part of the audience which has opted out of tracking in order to ensure the data for specific sites and content is accurate.

**How AMR works**

Online audience measurement, as conducted by all the major providers worldwide, aims to provide accurate information on the size of the audience for a specific campaign and the frequency of exposure of the audience to its content using measures which are consistent with offline media. The data is collected from panels of individuals who have consented to have their media consumption and other behaviour measured. However, online is unlike most offline media in that there are millions of sites. In offline media, with the exception of out-of-home (billboards in the street or mall etc.), the number of different media entities (programmes or publications etc.) is sufficiently small (in the hundreds or thousands) to be accurately measured by a sample survey or panel of users. In the online world the number of sites and biases introduced by problems sampling certain categories of users mean that the pure panel data need adjustment to be correct.

For this reason it is necessary to **calibrate** the data coming from an online panel to correct the estimate of the number of hits on specific content by using census counts. At the level of the total sample this corrects for non-demographic biases in the panel sample. This is the data which publishers and media buyers use when evaluating a specific site. However, the data may still be slightly “wrong” when looking at a single content item which may appear on a number of sites.

That number is then further corrected by **calculating through** the panel derived data and applying the census number to the information derived from the consenting panel members who saw that content. It is important to be able to do this cross-site to ensure that the correct figure for unique exposures and frequency of exposure are obtained. However, the census counts are only used for correcting the aggregated data collected via the consenting panel. Any demographic or behavioural information used to describe the audience for the content comes from the panel, not from the census data and cannot be used to target an individual or tailor their experience.

**Verification of the quality of the data**

All media are valued by the information derived from AMR. The amounts of money involved are very large and therefore confidence in the reliability of the data is very important. For this reason most major media surveys are subjected to independent audit and oversight to ensure that they conform to the agreed specification. In the USA this is done by the Media Ratings Council, in Germany by the AGMA, in France by the CESP and in the UK by the Joint Industry Committees (JIC’s) and there are similar bodies in most other developed countries worldwide. There is a need for the auditors to be able to **validate** the data which is why census counts and raw panel data need to be retained for the length of the campaign or 53 weeks, whichever is shorter.

These auditing bodies are national and do not have a consumer responsibility for explaining the role of AMR to the public, which is why it has been proposed to establish an independent certification and consumer information organisation which can ensure compliance with the process and consistent application worldwide. This organisation would ensure that members would comply with the W3C requirements.

**Specific questions about Issue 25 text:**

The data collected by the third party: Must be pseudonymized before statistical analysis begins, such that unique key-coded data are used to distinguish one individual from another without identifying them".

**Ed Felton’s questions:**

1. What does "identifying" mean in this text? (One might read "without identifying" as requiring that data be "de-identified" according to the definition that appears elsewhere in the spec. But if the data qualifies as de-identified then no permitted use is required here because the general safe harbor for de-identified data already applies. Alternatively, if "identifying" means something different here, then that should be spelled out.)
2. What does "unique key-coded data" mean? Is the text about "unique key-coded data ..." meant to serve as a definition of "pseudonymized"? If so, it seems overly prescriptive, requiring one particular method that (purportedly) qualifies as pseudonymized. Alternatively, this text might be read as requiring a particular (purported) pseudonymization method. If so, why require this particular method?

**Answer:** The controls regarding the census data include assigning a random number to the record and obfuscating the last three digits of the IP address. These are the current minimum requirements. Different companies may adopt further pseudonymization practices for technical reasons and these may change with technology or with national law eg in Germany it is required that the IP address is hashed as well.

If there is future agreement at international level on pseudonymization standards or definition, we will adhere to these if they are higher than our standards as they become available. The census data is held securely, as is all audience research data, and deleted within the maximum time period for validation and auditing.

We note that the wording is open to misinterpretation because the data is pseudonomized during processing, and then aggregated (ie de-identified ) data is provided to clients as statistical reports. Therefore without specifying the method used for pseudonomization, alternative wording could describe a testable outcome:

**CURRENT TEXT:** The data collected by the third party:

*Must be pseudonymized before statistical analysis begins, such that unique key-coded data are used to distinguish one individual from another without identifying them.*

**NEW PROPOSAL:** The data collected by the third party:

*Must be pseudonymized before statistical analysis begins, such that it is possible to distinguish one individual from another but the data by itself, cannot be attributed to a specific device.*

**Ed Felton’s question**

1. Why allow pseudonymization to be delayed until "statistical analysis begins"? Why not require pseudonymization to be done promptly when data is initially collected?

**Answer:** This data first needs to be filtered on a continuous basis to detect fraudulent activity such as web bots. As the campaign progresses, you may detect additional doubtful elements and then need to re-process the data again to check that they are removed. Once it is certain that the data is clean, it is pseudonymized before analysis.

**Ed Felton Question** the "independent certification process under the oversight of a generally-accepted market research industry organization that maintains a web platform providing user information about audience measurement research. This web platform lists the parties eligible to collect information under DNT standards and the audience measurement research permitted use ..."

1. The authors appear to have a specific organization in mind. Which organization is that, and who runs it?
2. What is the rationale for giving a particular organization control over the certification process and the ability to declare who is eligible to exercise this permitted use?

**Answer:** The proposal for Issue 25 has been developed by the major global providers of AMR. This paper is intended to provide clarity about why the proposal has been written in the way it has and to help people who are not familiar with this kind of market research understand how our industry works to protect consumers’ personal information, ensure that advertising money is spent efficiently and encourage effective competition and good innovation by media publishers. We have tried to incorporate sufficient protections in the specification to provide reassurance to the members of W3C that this is in fact the case, but we remain willing to discuss further issues of clarification or amendment which will provide additional clarity and reassurance.

As noted, explanations and opt-outs are currently offered by AMR providers separately and there are various self-regulatory mechanisms already in place. The intention in Issue 25 is to provide an additional level of transparency and education for users, noting that this use case is not immediately apparent even for experts in this W3C group. We think that a common AMR explanation and opt-out will help users understand the purpose, and ensure that this permitted use remains with the boundaries specified by the W3C standard. The body would be set up with the participating research companies as founder members with expert oversight and all companies operating in this field are welcome to join. We remain open to moving this into the non normative section of Issue 25 and further discussion as the standard evolves in practice.