Radio Over IP Measurement

Methodology

CIM

Centre d'Information sur les Médias - Centrum voor Informatie over de Media

Avenue Herrmann Debroux 46 - 1160 Bruxelles

Tel. : 32 2 661 31 50 - Fax: 32 2 661 31 69

Email : info@cim.be

URL : http://www.cim.be

Table of contents

1   Introduction

Since 2019, CIM has started a dedicated measurement of radio streaming (over IP) based on the collection and processing of the streaming servers logfiles.

This new measurement is part of the new design foreseen by the Strategic Committee Audio for the measurement of the audio landscape.

= All Audio	CIM RAM GfK **CURRENCY**	CIM Internet Gemius	*NEW* **CIM Radio Stream Monitor** Neuromedia
Live Radio			
FM/DAB	✓ included	-	-
Over IP	✓ included	✓ *Included**	✓ included
Audio on demand			
Podcast/Time shifted	-	✓ *Included**	-
Other Audio over IP	-	-	-
Method	Radio diaries Sample based	User-side log analysis Traffic on census level Profile panel based	Server-side log analysis Traffic on census level
Limit	Declared behaviour	*Only scripted players	Distribution of audio files (not listening)

In this long-term vision, CIM considers that the future radio measurement will inevitably be a hybrid method including declared data, passive measurement or log file analysis and by combining the advantage of each method.

This methodology concerns the server-side measurement of radio over IP.

1.1   Measurement of radio Over IP

Since several years, two methodologies are available for the measurement of radio over IP:
- The data collection at the level of the player (through a browser or a mobile app), generally called the "client side" measurement.
- Or the data collection at the level of the audio streaming server, called the server-side measurement.

Each method has its advantages but also its limits.

The client-side approach is based on the scripting/tagging of the participating players. All subscribing stations are required to implement a small JavaScript tag into the source code of their players (on website or in app). Each time a stream is requested by a player, a call is sent to one of the measurement servers of the research institute.
- The advantage of the client-side measurement is that it allows the precise measurement of all actions generated by the user with the players and to objectify what has really been listened to.
- The disadvantage of the client side is that the scope of measurement is limited to the scripted players. In other words, it means that streams that are listened outside the proprietary scripted players are not measured (e.g. radio aggregators, wifi radio, …)

The server-side approach uses directly the logfiles generated by the streaming servers. In the logfiles, all events (stream requests by/and sent to a "web service" through IP) are stored by the streaming server.
- The biggest advantage of this server-side measurement is that it allows the measurement of 100% of the delivered streams, independently of the receiving platform.
- The disadvantage is that the method cannot identify the exact consumption of the delivered stream. In the case of AOD for example, it is not possible to assess that a downloaded audio file has been consumed. Even for live stream, most of the players are downloading automatically a small part of the file, called "buffer" before the "real" consumption (play) by a user.

Hereunder you find a summary of strengths and weaknesses of each method.

	Client-Side	Server-Side
Technology	Tagging	Log parsing
Data exhaustivity	⚠ Limited	✔ Complete
User information	✔ Very rich	⚠ (IP + Player)
User behavior	✔ Very rich	⚠ Limited (data chunks)
Data fault tolerance	✖ No recovery	✔ Queued
Data availability	✔ Near real time	✔ Near real time
Setup	✖ Invasive	✔ Non invasive

At CIM, the client-side measurement is done for the players of the broadcaster participating in the Gemius internet measurement.

Since 2019, CIM has also started a server-side measurement in the context of its Radio survey. The measurement is done by Neuromedia, a Belgian Company specialized in streaming measurement (audio and video).

The long-term vision of these different measurement is the following:
- Step 1: Client-side measurement of scripted players
- Step 2: Server-side measurement of participating broadcasters
- Step 3: See how the different radio measurement data flows can be integrated. This integration could allow :
  o An extra validation of server-side data
  o Inclusion AOD measurement
  o Modelling of an exhaustive profiled audience
  o Modelling of integrated audience for offline and online audio

2   Scope of the server-side radio measurement

As described above, the server-side measurement measures all events (stream request and served) at the level of the streaming server.

The scope of the measurement is therefore:
- all channels of broadcasters allowing access to their server logfiles. You can find the list of measured channels in annex 1.
- Streamed through IP
- On all listening platforms, whether it is broadcasted on owned players or external players (radio aggregators, Wi-Fi radio…)
- Audio or video (same audio content) streams of radio stations (excepted DPG stations where the video streams are not yet included).

The measurement is a traffic measurement, not an audience measurement: it measures connections, sessions, unique users (a technical measure based on the IP address x Browser/player ID) but not persons.

FAQ	- Are all stations measured? *No, only the stations for which we have access to the logfiles* - Is the radio listening through TuneIn, Sonos, … measured? *Yes.* - Is the listening through TV measured? *It depends, if an application or a browser is used to listen to radio: yes. If radio is listened through the set-top box: no, the original radio signal is rebroadcasted by the TV channel/operator.*

3   Methodology

3.1   Data collection

Each day, Neuromedia collects the logfiles directly on the stream server of the measured stations to which they have received access.

The logfiles are like text file, where every stream served by the server is stored as a text statement with different parameters such as: IP address, date, time, stream name, browser ID, …

Example of logfile statement:
```
123.123.12.123 - - [02/Oct/2010:11:16:17 -0600] "GET/
Stationname_be_live_64.aac HTTP/1.1" 200 213036 "Referrer"
"NPRRadio/2.2 CFNetwork/485.10.2 Darwin/10.3.1" 103
```

This string contains following information:
- 123.123.12.123 : IP address
- -- : User identifier (usually anonymous)

- [02/Oct/2010:11:16:17 -0600] : connection timestamp
- "GET/ Stationname_be_live_64.aac HTTP/1.1" : Request. Here for the publishing point (also called mount) « stationname_be_live_64.aac »
- 200: the http answer code HTTP. 200 is the standard value.
- 213036 : Number of bytes transferred
- "-" : referer (when available)
- "NPRRadio/2.2 CFNetwork/485.10.2 Darwin/10.3.1": player identifier
- 103: duration of the session in seconds

The format of the logs can vary depending on the stream technology used (Icecast, Shoutcast, HLS, …). NeuroMedia treats all streaming servers commonly used in radio, including IceCast, SHOUTcast and Wowza but also the CDNs using HLS and MPEG-DASH. It is therefore possible to report from several sources with different technologies in a single reporting tool.

3.2   Data processing

3.2.1   Database creation

The logfiles data are read and enriched with additional metrics such as:
- location (based on the IP address)
- device classification
- player type identification

NeuroMedia has built its own player library that allows the identification of all the most used device and players used for radio streams listening (broadcaster players, aggregators, wifi radios, usual standalone players, ...)

The processed and enriched data, called "transactional data", are then temporarily stored in a data store and will be used to create the data models containing the metrics and analysis dimensions used in the reporting tool. These transactional data contain following information:
- IP address
- Timestamp of session start (UTC)
- Timestamp of session end (UTC)
- Duration in seconds
- Number of bytes transferred during the session
- Player / browser identifier
- Content identifier (if AOD)
- Data source identifier
- Source file identifier
- Publication mount identifier (channel stream detail)
- City, region, country
- Referer (if available)

These transactional data are stored during min. 3 days. Older data is deleted on a daily basis.

3.2.2 Editing rules

Filtering of connections

A first rule applied to avoid the measurement of automated traffic is the filtering of micro connections. All connections of less than 2 seconds are excluded. These micro connections are excluded at the stage of the processing of the logfiles and are not stored in the transactional data (see 3.2.1). All other connections are processed and converted.

A second rule is the filtering of sessions with a duration of more than 1.440 minutes, as these sessions are considered as too long to be really listened.

Filtering of non-human traffic

Based on the player identification, traffic coming from specific players (robots, link, ripper) is systematically excluded.

A filtering on excluded IP addresses is also applied (blacklisted IP addresses, broadcasters IP addresses, …).

Sessions fusion

Streaming can be subject to occasional disconnection. In case of disconnection, the session before and after the disconnection will be registered as two separate events in the logfiles.

To avoid an overestimation of the number of sessions, NeuroMedia applies a mechanism of sessions fusion to rebuild the complete sessions. This fusion mechanism is based on a predefined tolerance's threshold of 60 seconds. If two sessions with the same combination IP address/player ID/publishing point have less than 60 seconds between the end time of the first session and the start time next session, they will be merged and considered as one unique session. Beyond this threshold of 60 sec., two sessions are counted. For the HLS technology, a fusion threshold of 180 seconds is applied, due to the specificity of the technology where information of a single session is systematically divided into different events (chunks) in the logfiles. The HLS technology is currently only used for a part of the VRT streams.

This fusion mechanism is applied within a logfile but also between logfiles, allowing also to merge sessions starting at J-1 and ending at J-0.

4   Outputs

Based on the processed transactional data, the publication metrics described hereunder are calculated and stored in the final data models. These data models are stored for a longer period, from some months to several years, depending on the reporting needs defined by the Technical Committee radio.

4.1   Metrics definition

Connections
Sum of started sessions of min. 2 sec, calculated for the chosen period.

```
SELECT SUM (ID) FROM UserSessions WHERE Duration >= 2 AND Start BETWEEN
@PERIODSTART AND @PERIODEND
```

Active sessions
Sum of started sessions of min. 60 sec, calculated for the chosen period.

```
SELECT SUM(ID) FROM UserSessions WHERE DURATION >= 60 AND Start BETWEEN
@PERIODSTART AND @PERIODEND
```

Sessions
Sum of started sessions of min. 600 sec, calculated for the chosen period.

```
SELECT SUM (ID) FROM UserSessions WHERE Duration >= 600 AND Start
BETWEEN @PERIODSTART AND @PERIODEND
```

Average daily unique users
Sum of unique users based on the combination of IP addresses and Browser/Player ID, with a minimum sessions duration of 60 sec., calculated for a chosen period of max. 24h.

The limit of 24h for the calculation of the unique user is set to avoid an overestimation of unique user due to change of IP address (dynamic IP addresses, …) for identical "real" users.

```
SELECT COUNT(DISTINCT(Useragent) FROM UserSessions WHERE Durationn >=
60 AND Start BETWEEN @PERIODSTART AND @PERIODEND
```

Average time spent
Average duration of an active sessions (minimum duration of 60 seconds), calculated by day.

```
SELECT AVG(Duration) FROM UserSessions WHERE Duration >= 60 AND Start
BETWEEN @PERIODSTART AND @PERIODEND
```

Total time spent
Sum of the duration of the active sessions (minimum duration of 60 seconds), calculated by day.

```
SELECT SUM(Duration) FROM UserSessions WHERE Duration >= 60 AND Start
BETWEEN @PERIODSTART AND @PERIODEND
```

Average quarter hour 10' (AQH 10')

Average number of sessions with a duration of 600 or more within an average quarter hour, calculated by day.   E.g. for an hour:
- quarter 1: 25 sessions of more than 600 seconds
- quarter 2: 40 sessions of more than 600 seconds
- quarter 3: 45 sessions of more than 600 seconds
- quarter 4: 38 sessions of more than 600 seconds
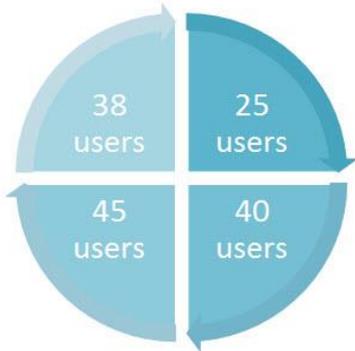
The AQH 10' for that hour is (25 + 40 + 45 + 38) / 4 = 39.5
For an entire day, the AQH 10' is calculated in the same way. In a day, there are 96 quarter hour. Therefore, the AQH will be the sum of sessions of more than 600 seconds in each quarter hour / 96.
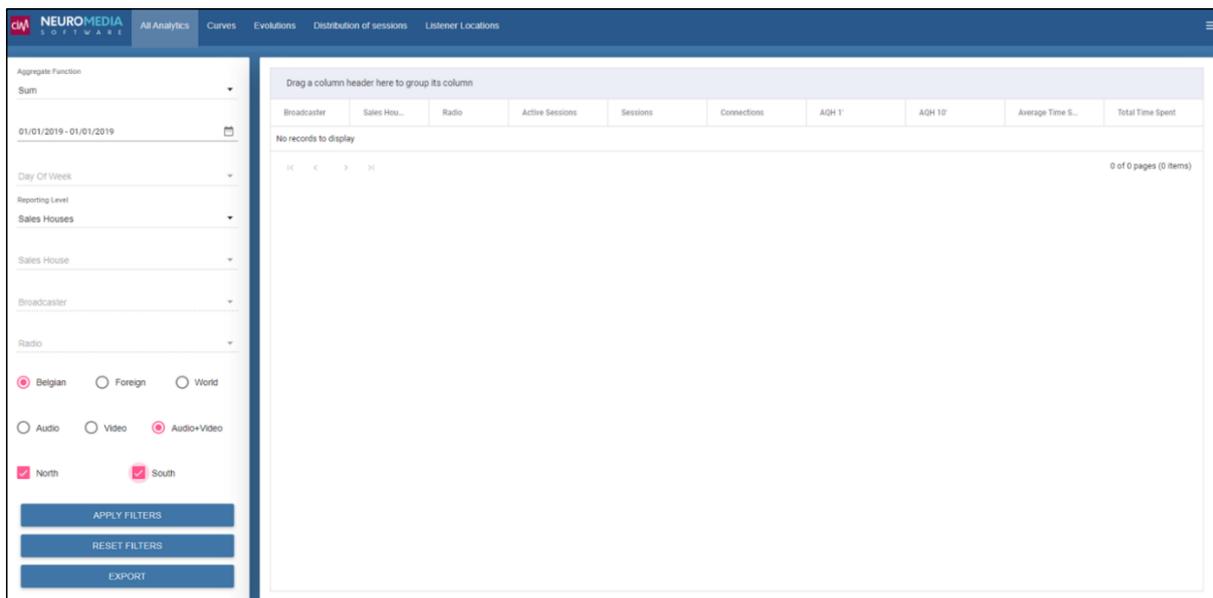
Average quarter hour 1' (AQH 1')
Average number of sessions with a duration of 60 or more within an average quarter hour, calculated by day.

4.2   Publication

4.2.1   Access to the data

Data are accessible for in a dedicated dashboarding tool developed by NeuroMedia.

This tool allows analysis of the results at the level of broadcasters, sales houses, stations.

It contains different report types:
- All analytics: overview of all metrics for the selected period (sum or daily avg. for the selected period).
- Evolution: date by date evolution by metrics.
- Curves: hour by hour results by metrics for an average day of the selected period.
- Distribution of sessions: distribution by active session duration (in min.) for the selected period.
- Listeners location: breakdown of the active sessions by location.

4.2.2 Timing

The results are available at J+5. Technically, it would be possible to publish the data at J+1. The period of 5 days between data collection and publication is used as buffer period:
- To collect missing logfiles due to technical reason.
- Allow the session fusion mechanism between days.
- Controls and correct eventual anomalies before their release in a publication.

4.2.3 Data retention duration

To guarantee the speed of the dashboard, data are stored and accessible through the dashboard for a limited period, defined specifically for each type of report :
- All analytics : 3 years
- Evolution : 3 years
- Curves : 1 year
- Distribution of session : 2 months
- Listener location : 2 months

Older data will still be accessible through backup files available at CIM.

4.2.4 List of published stations

You will find in annex 1. the list of published station as on 31/05/2020.

Annex 1: List of subradios included in the reporting

Sales House	Broadcaster	Radio	# Subradios	Subradios
IP (North)	Nostalgie NL	Nostalgie (NL)	11	Nostalgie WAF 60&70; Nostalgie WAF 80; Nostalgie WAF 90; Nostalgie WAF Belpop; Nostalgie WAF Extra; Nostalgie WAF Lage Landen; Nostalgie WAF New Wave; Nostalgie WAF Party; Nostalgie WAF Relax; Nostalgie WAF Rocks; Nostalgie What A Feeling
		NRJ Vlaanderen	1	NRJ Vlaanderen
	Zen FM	Zen FM	1	Zen FM
DPG Media (North)	DPG Media	Joe	8	Joe; Joe 60-70s; Joe 80s; Joe 90s; Joe Christmas; Joe Easy; Joe Theme; Joe Top 2000
		Qmusic	11	De Foute Radio; Q '00s; Q 10s (new 24/05/2020); Q 90s (new 21/02/2020); Q Theme (stop 13/05/2020); Q-Maximum Hits; Qmusic; Qmusic De Favoriete 100; Qmusic Shut Up And Dance; Qmusic Workalicious; Q-Running
		WILLY	1	WILLY
	Top Radio	Top Radio	1	Top Radio
Var (North)	VRT	Ketnet Radio	1	Ketnet Hits
		Klara	2	Klara; Klara Continuo
		MNM	5	MNM; MNM Back To The 90s & Nillies (new 23/05/2020); MNM Hits; MNM Nineties & Nillies; MNM UrbaNice
		Radio 1	2	Radio 1; Webcam Radio 1
		Radio 2	5	Radio 2 Antwerpen; Radio 2 Limburg; Radio 2 Oost-Vlaanderen; Radio 2 Vlaams Brabant; Radio 2 West-Vlaanderen
		Studio Brussel	5	Studio Brussel; Studio Brussel Bruut (new 25/05/2020); Studio Brussel Hooray (new 25/05/2020); Studio Brussel Tijdloze; The Greatest Switch
IP (South)	Radio H	Radio Contact	9	Contact 2000; Contact Gold; Contact Kids; Contact Lounge; Contact Love; Contact Mix; Contact Plus; Contact Urban; Radio Contact
		Bel RTL	1	Bel RTL
		Mint	1	Mint
	NGroup	Chérie FM	13	Chérie Emotions; Chérie FM; Chérie FM 80; Chérie FM 90; Chérie FM Acoustic; Chérie FM Ballads; Chérie FM Frenchy; Chérie FM Les Plus Belles Voix; Chérie FM Love Songs; Chérie FM Romantic; Chérie FM Zen; Chérie Jazzy; Chérie Sweet Home
		Nostalgie (FR)	25	Le Meilleur de Nostalgie; Nostalgie Jazz; Nostalgie; Nostalgie 60; Nostalgie 70; Nostalgie 80; Nostalgie 90; Nostalgie Camping; Nostalgie Chansons Françaises; Nostalgie Christmas; Nostalgie Cinema; Nostalgie Dance 80; Nostalgie Dance 90; Nostalgie Héros; Nostalgie Italia; Nostalgie Johnny Halliday; Nostalgie Love; Nostalgie New Wave; Nostalgie Pop 80; Nostalgie Pop 90; Nostalgie Rock 80; Nostalgie Rock 90; Nostalgie Rock Classics; Nostalgie SoulParty; Nostalgie+
	Fun Radio	Fun Radio Belgique	1	Fun Radio
RMB (South)	RTBF	Classic 21	10	Classic 21; Classic 21 60's; Classic 21 70's; Classic 21 80's; Classic 21 90's; Classic 21 Blues; Classic 21 Métal; Classic 21 Noir Jaune Rock; Classic 21 Route 66; Classic 21 Soulpower
		Jam.	1	Jam.
		La Première	2	La Première; La Première Vision
		Musiq'3	2	Musiq 3; Musiq 3 Beethoven 250 (new 01/06/2020)
		Ouftivi	1	Ouftivi
		Pure	4	Pure; Pure Lazy; Pure Like; Pure Vision
		Tarmac	2	Tarmac; Tarmac Chillin'
		RTBF Spot	1	RTBF Spot (new 26/04/2020)
		VivaCité	8	Viva+; Vivacité Brabant Wallon; Vivacité Bruxelles; Vivacité Charleroi; Vivacité Liège; Vivacité Luxembourg; Vivacité Mons; Vivacité Vision
	NGroup	NRJ	19	Les Hits de l'été 2019; NRJ; NRJ at Work; NRJ Best Hits Ever; NRJ Dance; NRJ Héro; NRJ Hits; NRJ Hits 2000; NRJ Hits Of The Month (stop 05/04/2020); NRJ Hits of the Week; NRJ Latino; NRJ Love; NRJ Music Awards; NRJ Nouveautés; NRJ pour le sport; NRJ Rap FR; NRJ Relax; NRJ Télétravail (new 24/03/2020); NRJ Top 100 2019 (new 11/01/2020)
	IPM	DH Radio	1	DH Radio