CONSTRUCTING A NEW DATA STRATEGY TO DRIVE PERSONALISATION AND RECOMMENDATION

Dr. Pierre-Nicolas Schwab, *Big Data/CRM Manager* – **RTBF**

RTBF, the public service broadcaster serving Belgium's French-speaking population, has embarked on a new strategy that harnesses data and technology to deliver a more personalised and relevant experience to its audience. With people, especially in younger age groups, increasingly turning to social media as their primary – or indeed only – source of news and information, RTBF has recognised that its competitors are now the likes of Facebook and Twitter, rather than only traditional media as in the past. As a source of entertainment, its competitors now include Netflix and on-demand music streaming services, and all of these are currently far more advanced in personalisation and recommendation than legacy television and radio broadcasters.

egta spoke to Dr. Pierre-Nicolas Schwab, Big Data/CRM Manager at RTBF, who designed and directs the broadcaster's new data strategy, bringing together internal resources with external partner organisations.

The starting point

RTBF has a large audience and a great deal of unique, high quality content; however, it knows relatively little about the former and the characteristics of the latter cannot currently be identified automatically. Together, these two limitations prevent the broadcaster from recommending content to users or delivering a personalised experience.

As a broadcaster with television, radio and digital assets, RTBF has multiple touchpoints and interactions with its audience, both online and offline. However, very little information is systematically collected from these interactions, and while the broadcaster has about a million user ID's in its databases, for the most part this data is too limited to be actionable for personalisation.

"It can be when people call a radio station, send an email, attend an event or go to our websites, those are all sources of interaction", explains Schwab. "The problem was, for most of them we capture no data. When someone was calling a radio station, there was no track of it. For some of them we have databases. What is interesting is that the name of the database is the name of the supplier, and what you suddenly understand is that through the years, everything was very fragmented. We had to reconcile everything."

The second main challenge is the lack of robust metadata associated with RTBF's content. In theory, any piece of content, be it an article on the broadcaster's websites, a segment of audio or video, can be identified by a series of metadata points, such as its title, a description, associated tags and so on. In practice, this metadata is usually very patchy and, again, too incomplete to allow personalised content to be delivered to RTBF's audience. As noted by Schwab, "How can I do a recommendation based on metadata that is very poor? I have to recreate everything. I

could hire an army of people to write metadata, but that won't be scalable."

Building out the new data capabilities

At the start of the process, in 2015, the team at RTBF set about investigating the broadcaster's various audience touchpoints and identified more than 50 such sources of interaction. To be actionable, for example for targeted advertising, a person needs to be identifiable, for instance through an email address, a unique ID, or a combination of several other identifiers. However, for content recommendation, it is necessary to know something more about these people, what content they like to consume, their interests and their social environments.

SSO and social login: building richer user profiles

At the heart of the new data collection infrastructure is RTBF's single sign-on (SSO), launched in December 2016, which allows a larger proportion of the audience to be turned from anonymous visitors into known audience. Gigya was selected to deliver this service following a public tender. RTBF chose to outsource the implementation and ongoing management of the SSO, rather than build the technology in-house, as identity and access management (IAM) systems are complicated to develop and a number of providers already exist on the market. Furthermore, these providers take care of storing, processing and dashboarding the data collected through SSO, as well as managing the complex and often changing environment of logins using social media (social login).

RTBF offers users the possibility of registering either using their email address (first name, surname, email address, protected with a password) or through social login. "We recognised very early in the SSO process that we wanted to take advantage of Facebook Connect," explains Schwab. "Why? Because we want to understand what people like outside of the RTBF world. If I observe only what people do on my websites, it becomes tautological. So, I can recommend only what I am producing. And soon, I am trapped in my production bubble."

The broadcaster will request information about each user's network of Facebook friends, in addition to data about the user themselves, in order to build a more comprehensive understanding about its logged-in audience, which in turn allows better recommendation. At launch, RTBF chose to offer social login only through Facebook, as it delivers richer and more relevant data than alternatives such as Twitter and other social networks. However, as the system matures, it may be more appropriate to allow users to log in using alternative social media where consumers of particular types of content might have a better affinity with networks such as Instagram, for example.

The data from both social and non-social login is consolidated by Gigya, with which RTBF can also leverage progressive completion strategies to update users' profiles over time, for example by periodically asking them to add or confirm additional information. It should be noted that while the SSO has only been live for a few weeks at time of writing, registering with name and email is currently preferred by a majority of RTBF users over social login.

Connecting consumption with people

Complementing the Gigya database of people is a separate database to collect and store raw data about *consumption*, described by RTBF as a *datalake*. Additional infrastructure and technology of one of RTBF's partner organisations uses Big Data technologies to convert raw data on website interactions so that

it can be combined with the data held in Gigya to power the recommendation engine.

"We are thinking about non-relational databases to store transformed data and start visualising it and using it. When people visit the website, we drop a cookie. And when people register on Gigya, that cookie – or ID – is being replaced with a Gigya ID, which allows us to collect back all of the history of that person before he or she registered."

A reliance on first-party data

RTBF's recommendation strategy is dependent entirely on the broadcaster's own first-party data, either collected though the SSO or from anonymous digital visitors. "Very early, we decided that we will never buy data. We will not be selling data and we will not be buying data."

This reflects the broadcaster's priority, which is to develop a data strategy and architecture to enable targeted content to be delivered to users, rather than to deliver targeted advertising, which typically requires integration of third-party data sources, for example through a DMP.

Collecting data in the mobile environment

The collection of data on consumption through web browsers on desktop and laptops is enabled largely through the use of cookies, which are of more limited use in the mobile environment. In 2017, RTBF plans to launch a series of new mobile apps for its websites, Auvio video and audio platform and a new multimedia platform for young people currently going by the working title Média Z, all of which will be connected to the single sign-on.

The organisation is seeking to lead people to use the RTBF apps, and motivating them to log in through incentives, rather than to force all consumption to follow this path. This is a different strategy to some broadcasters, which now only allow their video – and in some cases online radio – to be consumed by logged-in listeners.

Preparing content for the recommendation engine

RTBF's recommendation engine is designed for its journalistic output, where metadata can be extracted based on what is *said*, rather than for music or television shows, for example, where metadata based on factors such as *emotions* can only be delivered by humans rather than algorithms.

"We have developed a prototype where you can inject a video or an audio file, which will go through a speech-to-text and a sequencing algorithm, then through an ontology algorithm that extracts semantic fields around what is said and attach them back to the content." Yet there are still some technical challenges to solve to make it a reliable product.

This prototype will be refined in 2017 to allow RTBF to build up a richer understanding of its own content using an automated, algorithmic process, which together with the user profiling described above will power the recommendation engine.

Two philosophies of recommendation and the conundrum of the filter bubble

Broadly speaking, recommendation engines can propose content following two models: *exploitation* and *exploration*. The former presents content that is similar to a user's past consumption, meaning that exposure to new ideas or concepts is limited. Amazon and Netflix are both examples of platforms that base their recommendations on exploitation. The latter, on the other hand, seeks to present content that – while it continues to be relevant – can expose the user to materials that they

might otherwise not have consumed. As part of its public service commitment, RTBF will design its recommendation engine to promote exploration while at the same time retaining the flexibility to adapt to meet the needs of users who may prefer to receive a narrower range of suggestions.

The nature of recommendation strategies has become increasingly relevant in discussions around the so-called *filter bubble*, in which people are exposed to information in the digital realm in a way that serves to reinforce their existing opinions and isolate them from alternative viewpoints. The subject of considerable debate in media, publishing, broadcasting and academic circles, there is as yet little consensus on how the filter bubble effect operates, whether it actually exists and how news publishers should respond to it. However, this effect is an important consideration when designing the operational parameters of a broadcaster's recommendation engine.

Implementation of the RTBF recommendation engine

Having established the data collection, storage and processing elements of its architecture, including the staging area and single sign-on, the next stage is to start recommending content to RTBF's users. The technology behind the recommendation engine is handled by the Liege part of the consortium, and it is linked to the RTBF's online platforms via API integrations.

RTBF refers to the first iteration of this system as its *similarity engine*, which will make the same recommendations to all people watching or viewing a piece of content, and this is due to launch in February 2017. "This content-based recommendation is not new, it is not innovative, but we had to rebuild the foundations in order to grow from there and have some more refinement and technological changes."

In December 2017, under specifications that are currently being developed, the system will become more sophisticated, allowing RTBF to expose people to content that they are not used to consuming. "The purpose being to expose people to the diversity of the world, to the diversity of opinions, to new pieces of content." However, if users choose not to receive this type of experience, "the system will just learn to not serve more of those unexpected types of content, using machine learning."

Data security and the GDPR

RTBF has overhauled its internal data privacy documentation and developed new communication materials about how and why it collects data for its users. Inspired in part by the Channel 4 Viewer Promise (see page 25), a set of straightforward texts and explanatory videos have been produced to increase transparency and help educate users on this topic. It is also important to recognise that recommendation itself works better when people understand the basis on which content is being selected and placed before them.

As a public organisation, RTBF will be required to have an in-house Data Protection Officer (DPO) when the new EU General Data Protection Regulations (GDPR, see page 22) come into force in May 2018. This raises a number of questions for the internal organisation of RTBF, such as where within the broadcaster's structure the DPO should be located in order to function properly.

Challenges and learnings

Almost 2 years into the RTBF's data project, Schwab reflects on the various challenges and learnings that have been made along the way. "There are two worlds. There are the people who believe in big data and correlation as being the sole predictor of future behaviour and the world of more sociology-oriented people who

think that you have to understand causation to make good recommendations. And I'm just in the middle!"

"It's so important that people do not believe that correlations between variables can predict the future of individuals. We need to understand *why* people are doing things. We must reflect the role of RTBF – also of broadcasting in general – in terms of what is the job that people hire us to do. When you are on public transport and you are watching the RTBF app, are you watching because you want to consume news or because you just want to spend time? These are two very different jobs.

"What are the consumption moments, why are people using RTBF? Hence also the SSO – I have to make the link between the user on their desktop, in their home, on their mobile or tablet. Yet the data I get will get will not be sufficient to infer *why* someone is doing what they are doing. And the problem with data scientists is that they just believe in *what*, they don't try to understand *why*.

"In terms of human resources, the challenge is currently between the business people here in the RTBF and the data scientists in the other organisations involved in the project. They don't always understand each other, and I have to bridge this gap. Making that link is the most crucial thing."

Looking back at the process so far, Schwab explained that "it is not a technological challenge, it is more of a human challenge. That's something that I did not believe two years ago, but now I believe it."

CASE LEARNINGS:

- As a public service broadcaster and news source, the RTBF's competitors now include the likes of Facebook and Twitter, Netflix and other on-demand video and music streaming services
- Challenge 1: RTBF historically had several touchpoints with its audience, but little data was systematically collected
- Challenge 2: the broadcaster had patchy and incomplete metadata about its content
- Personalisation and recommendation requires both of these challenges to be resolved
- User identification is now possible through a new single sign-on for all RTBF platforms, with social login also available (through Facebook Connect)
- The broadcaster is developing a recommendation engine based on *exploration*, rather than *exploitation*
- The General Data Protection Regulation requires RTBF to install a Data Protection Officer; this raises questions over where in the organisation this person will sit
- The biggest challenge of the project is human, rather than technical