OMeBa
Online Media Behaviour analytics

Authors
Theodore Koutmeridis
University of Glasgow
Theodore.Koutmeridis@glasgow.ac.uk

Kristofer Erickson
University of Glasgow
Kristofer.Erickson@glasgow.ac.uk

Martin Kretschmer
University of Glasgow
Martin.Kretschmer@glasgow.ac.uk

Jesus Rodriguez Perez
University of Glasgow
Jesus.RodriguezPerez@glasgow.ac.uk

This release was supported by the RCUK funded Centre for Copyright and New Business Models in the Creative Economy (CREATe), AHRC Grant Number AH/K000179/1.
1. Introducing OMeBa

Online behaviour is changing rapidly due to technological progress. The legal framework, and copyright policy specifically, needs to keep pace with digital innovation and new business models. These changes in consumption of digital goods are also challenging existing theoretical propositions, requiring new academic attention. The use of appropriate data is crucial for the understanding of the perplexing patterns of online behaviour.

Ofcom and the Intellectual Property Office (IPO) have invested substantial resources to develop (with the market research company Kantar) a unique dataset with rich information about online behaviour in the UK. The survey of a representative sample of the UK population aged 12+ was run consistently in 5 waves from May 2012 to May 2015, providing an exceptional source to track the evolution of online behaviour for various creative industries. Surprisingly, researchers and policy makers have not yet engaged in a detailed longitudinal analysis of this data.

CREATe, the RCUK copyright centre, identified a need in the academic, policy and industry communities for more direct access and has developed a user-friendly website, OMeBa (Online Media Behaviour analytics) offering easy access to this database. OMeBa provides access to all 5 waves allowing effortless comparison both across variables and across time. OMeBa is available here: [http://copyrightcentral.arts.gla.ac.uk/omeba/](http://copyrightcentral.arts.gla.ac.uk/omeba/)

There are two main benefits to using the OMeBa tool. First, it provides a simple way to access and download the raw data obtained through the survey research. These data have been linked with their original survey questions for ease-of-use, and have been validated in replication exercises. Secondly, the OMeBa tool allows researchers to quickly select and visualise statistics from key questions contained in any of the waves, and compare the results over time. For instance, the data suggest that the proportion of individuals who acquired access to books solely through infringing behaviour increased from 4% in 2013 to 6% in 2015. This is just an illustrative example of the possibilities offered by the interface. The presence of this data visualisation and extraction web tool is anticipated to inform future policy interventions based on evidence.

This document is organised as follows: Section 2 describes the Ofcom-Kantar surveys. Section 3 demonstrates how the OMeBa interface can be used. Section 4 offers some illustrative examples, while the last section discusses potential policy implications and concludes.

2. The Ofcom-Kantar surveys

Under the Digital Economy Act (DEA 2010, now largely defunct), the communications regulator Ofcom acquired an obligation to monitor copyright infringement. Sections 11 and 12 provided for a code under which Internet Service Providers would have to limit Internet access to infringers which would have been enacted by a Statutory Instrument (this was laid before Parliament twice but never passed). Section 17 and 18 provided for blocking injunctions of websites that made infringing materials available (sections 17 and 18 were repealed in 2011, since right owners had
used successfully Section 97 of the Copyright, Designs and Patents Act 1988 to obtain blocking injunctions).

The Hargreaves Review of Intellectual Property and Growth recommended in 2011 that Ofcom should not wait until its formal reporting duties arising from the Digital Economy Act began, and start immediately to gather independent data so that trends could be established. The Government adopted this recommendation and tasked Ofcom and IPO to work together.

Ofcom commissioned market research company Kantar to develop a survey instrument that would gauge the consumption patterns of UK consumers with respect to both lawful and unlawful content accessed through the web. The first 4 waves of the survey were run in 4 subsequent quarters in 2012 and 2013 with a large sample size (4400 individuals), and a hybrid online and face-to-face survey methodology.

Wave 5 was commissioned by the IPO in 2015, again from Kantar. It was run from March-May 2015, and allows backward comparison to Wave 4 which was conducted in the same quarter (March-May 2013).

Quoting from the official 5th Wave report:

“Researching copyright infringement and digital behaviours is complex. The ways in which consumers access and share copyright material online change regularly, and infringement levels, in particular, are notoriously difficult to measure. ... Rather than focusing on one industry, the study looks at six main types of online content – music, film, TV programmes, books, video games and computer software – and for each of these assesses levels of infringement. These are then assessed within wider patterns of consumer behaviour and content consumption.”

The surveys are very extensive and contain 520+ different questions with 5000+ data points each. Furthermore, each of the questions in the raw dataset is expressed as a code, and a separate file is in place to decipher each of these codes to the questions expressed in natural language. OMeBa helps to simplify the process of browsing the survey data by linking variable codes with plain questions in English language taken from each of the surveys.

Currently, the raw values for certain questions with numerical answers seem not to be available or are very difficult to find within the downloadable datasets. Raw numerical values are considered as any positive integer, instead of ranges such as 1 to 10; 11 to 20; designed to group different respondents’ answers. It is possible that a change in the naming convention is the reason why we could not identify the raw values in many cases.

For example: Wave 1 contains question codes q10ext01, q10ext02 ... q10ex120 whereas wave 2 contains question codes q10a01, q10a02 ... q10b20, which are believed to refer to the same questions, as both waves have exactly 46 question codes with the q10 prefix. Whilst this particular problem was not overly difficult to solve with some human intervention, inconsistencies in naming convention across waves do become problematic when dealing with 500+ different question codes in an automatic way.
Furthermore, some essential data has to be computed from the raw results in order to be able to replicate results reported within officially published documents. This is the case for the following fields:

- Infringement volume: i.e. how many items have been subject to infringement for a particular respondent.
- Whether the respondent belongs to the top 10%, 20%, or bottom 80% infringer group.
- Type of infringer (or non-infringer): Computed by clustering with K-means. (K=4)

Moreover, the official reports often do not specify which of the questions in the dataset were utilised to compute each of the results reported. A particular example of this can be found in the “deep dive” analysis report. Chapter 9 introduces a formula to estimate the volume of infringement for each subject, which is given by:

\[
\text{Number of illegal files} = \text{files downloaded} + \text{files streamed} - \text{files paid for} - \text{legal files}
\]

The report does not specify which of the questions refer to each of the variables used, which requires estimating the most appropriate combination for purposes of replication. For example, in terms of e-Books for the downloaded and streamed content, it is not clear if one should use:

\begin{align*}
\text{qb6}_3\text{a} & \rightarrow \text{Q.B6}_3\text{A Number - e-Books downloaded in the last 3 months} \\
\text{qb6}_3\text{b} & \rightarrow \text{Q.B6}_3\text{B Number - e-Books accessed in the last 3 months} \\
\text{or} \\
\text{qb6}_3\text{ab} & \rightarrow \text{Q.B6}_3\text{A/B Number - e-Books downloaded or accessed in the last 3 months}
\end{align*}

Additionally, wave 4 does contain a file which specifies how the question codes in the dataset refer to each of the questions in the questionnaire. As a result, we can reverse-engineer this information for a considerable number of questions within the other waves. Moreover, it also contains some of the information that was computed and included in the reports such as the membership of a respondent to the top 10% infringers.

As a result (although widely alleviated in wave 4) this and other issues make it problematic to explore and work with this dataset thus perplexing the replication of the results that can be found in published reports. These challenges limit the effective use of this outstanding resource by the public and the research community. Addressing this gap is the purpose of the OMeBa online tool, which takes into account such inconsistencies and makes it easier to extract meaningful results from the publically available dataset.

In order to address some of these issues and provide a more intuitive access to the data we have produced this documentation on how to use OMeBa.

---

1 This report is available here: [http://stakeholders.ofcom.org.uk/binaries/research/telecoms-research/online-copyright/deep-dive.pdf](http://stakeholders.ofcom.org.uk/binaries/research/telecoms-research/online-copyright/deep-dive.pdf)
3. How to use the OMeBa interface

The Online Media Behaviour analytics (OMeBa) tool has been designed to allow users, such as policymakers, entrepreneurs, researchers and students, to explore this useful resource. To this end OMeBa develops the following functionalities:

- Exploration of available questions
- Subset selection
- Data export

**Figure 1: Main OMeBa interface screen**

![OMeBa Interface Screen](image)

Note: The main OMeBa interface consists (from top left to bottom right): a dropdown menu to select survey wave and a search bar to select variables based on plain-English survey questions; buttons to select questions based on medium, such as Software, Music, etc.; candidate questions appear below the search bar based on search criteria. Selecting questions moves them to the right-hand window. Clicking on the Analysis button (bottom-centre of the screen, represented by the funnel icon) brings up the data analysis window (see Figure 2).

### 3.1 Exploration of available questions

The exploration of questions is enabled by a search box, a question-type filter and a tag-cloud of common terms (bottom right). The search box can be used to filter questions containing a particular set of words. The question has to include all specified words, as partial matching is not allowed. OMeBa also includes a set of medium-specific filter buttons. The types refer to the types of media included in the dataset. The user is able to show only questions that relate to music, TV, films, software, etc. Last but not least, a tag-cloud is used to show the most occurring bigrams within the currently selected questions, which gives an idea of what are the predominant subgroups of information contained in the current questions.
Note: Clicking on the funnel icon will activate the analysis pane for selected questions. The right-hand window contains a list of the selected questions and their category values. Green and red icons enable the user to select and deselect specific values for sets of questions.

OMeBa provides basic statistical figures extracted from questions that the user may want to explore further. Figure 3 illustrates the three ways to look at the statistical data. First, a histogram shows the distribution of answers for each of the selected questions as an overview (selected field). Second, a table for the selected question displays numerically the distribution of answers included in the survey and their weighted average, offering the evolution of trends across waves (field evolution). Third, a figure illustrates the values for the selected question across all waves in a line diagram, which allows a different form of comparison across time (field comparison).
3.2 Subset selection

OMeBa includes a mechanism to extract a subset of the dataset. This approach allows the extraction of records for selected questions through the use of relational filters. Records are filtered using a logical “AND” between each of the questions (see Figure 4). For instance, it is possible to extract a subset of individuals who are “males” and “have internet connection at home” and “have downloaded 10 to 21 things”.
More information can be found here: http://copyrightcentral.arts.gla.ac.uk/omeba/filters%20explained.pdf

3.3 Data export

Ultimately, OMeBa offers the possibility to export the selected subset of data as a comma separated file (.CSV). This popular format allows the user to easily import the selected data to most of the existing statistical analysis packages. OMeBa provides a useful introductory tool to browse and learn about the dataset, allows the filtering of variables, and offers the extraction of the desirable data in a popular format, before the users commence any further analysis within the statistical analysis packages of their choice (see Figure 5).

4. Online behaviour: some examples and descriptive statistics

This section offers some illustrative examples, which are derived using the OMeBa interface. For example, the overall level of infringement, as this is measured by the fraction of individuals who access content only illegally (100% illegal activity) increased from 7.70% in 2013 to 9.43% in 2015. As far as books are concerned, the percentage of users who access digital content weekly increased from 31.89% in 2013 to 36.94% in 2015, while the percentage of users who download
them increased from 15.74% in 2013 to 19.55% in 2015. Infringement levels for books, as these are measured by any illegal activity, increased from 9.45% in 2013 to 11.91 in 2015, whereas the fraction of individuals who access books only illegally (100% illegal activity) increased from 4.60% in 2013 to 6.69% in 2015.

**Case example: Online infringement of film from 2013-2015**

With the OMeBa tool, one can determine the change, if any, in levels of infringement for specific mediums across the period covered by the survey. This example compares rates of online infringement of film measured in two separate waves of the survey.

To determine the level of infringement activity measured for a previous period such as 2013, first select the appropriate wave using the “Select Dataset” dropdown menu in the top-left of the screen. In this example we select March-May 2013 as the reference. Next, select the appropriate set of questions from within the survey. To filter only those questions that correspond to the medium of interest, enter ‘film’ into the search bar. In order to compare rates of infringement activities between waves, select Question 4_7: “You indicated that you have downloaded or streamed/accessed [n] films for free in the past 3 months. How many of these do you think were done so legally?”

Select sub-question qb4_7_pro_b2,qb4_7_pro_b2 --> “Proportion - Total films downloaded or streamed legally in the last three months”. The answers to this question will enable us to determine how many consumers of film in March-May 2013 who answered the survey were doing so legally and illegally. Click on the sub-question to make it appear in the right-hand window. Next, click on the funnel icon at the bottom centre of the screen to bring up the analysis pane.

According to the initial results, 82.59% of survey takers did not answer this question (because they chose not to, or because they did not report consuming film at all). Taking the survey sample as a whole, 11.63% of all survey respondents reported that all of their downloading was done from legal sources, while 3.7% of respondents admitted that all of their online consumption of film was illegal (see Figure 6).

**Figure 6:** The legal consumption of film content (all)
If we are interested in the proportion of film consumers who download legally vs. illegally, then we must filter out the non-respondents from the results. This is done by clicking the red icon next to the ‘not answered’ response group and then clicking ‘update this’ button. Now, we see that 66.84% of respondents claimed their downloads of film were legal in this period, leaving a combined 33.16% who engaged in some illegal downloading behaviour. This is consistent with the Kantar results published in July 2015 (see Figure 7).

**Figure 7:** The legal consumption of film content (respondents)

To compare these results to a later wave, simply select the wave March 2015 – May 2015 from the dropdown menu, and re-select qb4_7_pro_b2. This updates the analysis pane with the new results (see Figure 8).

**Figure 8:** The legal consumption of film content (comparison across time/waves)

One can see that for March-May 2015, the proportion of illegal activity among film consumers has decreased. 77% of consumers now claim that their downloads were always legal, leaving a combined 23% of consumers who engaged in some illegal downloading of film during the period. In 2015 there is a reduction of approximately 10 percentage points in reported illegal downloading behaviour compared with the same period in 2013.

5. Online Behaviour, Evidence and Policy

This paper describes the OMeBa tool, which allows easy access to the richest dataset regarding online behaviour in the UK. The Ofcom/Kantar surveys provide detailed information about infringement levels and allow the comparison of results across time. They also allow detailed insights into online behaviour that may be valuable to media firms beyond the assessment of infringement. It is surprising that a more detailed analytical approach has not yet been developed by researchers or policymakers, exploiting the potential of this uniquely rich dataset.

There seems to be a consensus that policy interventions require more and better data for evidence-based reforms (Gowers, 2006; Hargreaves, 2011; IPO 2013); CREATe’s research has made a concerted effort to advance this agenda (Kretschmer and Towse 2013; Koutmeridis, Erickson and Kretschmer, 2015). OMeBa attempts to contribute to evidence-based policy making by bringing an important dataset closer to policymakers, researchers and entrepreneurs, allowing a better examination of online behaviour and the design of informed policies for the creative industries that would promote innovation, economic prosperity and social welfare in an increasingly globalised digital economy.

References:


This document has been written by Theodore Koutmeridis, Kris Erickson, Martin Kretschmer and Jesus Rodriguez Perez on behalf of the CREATe research centre.

OMeBa has been developed and is maintained by CREATe, the RCUK Centre for Copyright and New Business Models in the Creative Economy, at the University of Glasgow (www.create.ac.uk).

OMeBa is available here: http://copyrightcentral.arts.gla.ac.uk/omeba/