

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/78157>

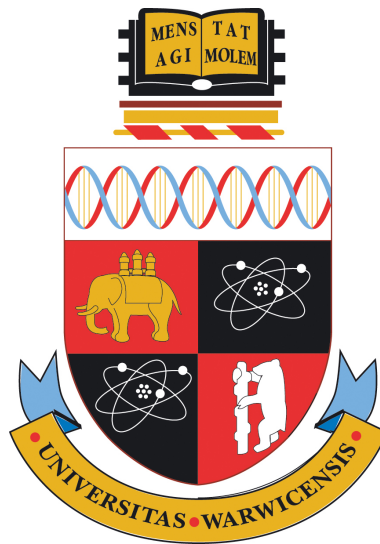
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

INTEREST-BASED SEGMENTATION OF ONLINE VIDEO  
PLATFORMS' VIEWERS USING SEMANTIC TECHNOLOGIES

RADU SORA



A thesis submitted in partial fulfilment of the  
requirements for the degree of

Doctor of Philosophy in Engineering

University of Warwick, WMG  
February 2016

# CONTENTS

---

<b>i</b>	<b>THE PROBLEM AREA</b>	<b>11</b>
<b>1</b>	<b>INTRODUCTION</b>	<b>12</b>
1.1	The Problem Area . . . . .	15
1.2	Previous Research . . . . .	20
1.2.1	Customer Segmentation . . . . .	20
1.2.2	Media Market . . . . .	21
1.2.3	Large Knowledge Bases . . . . .	22
1.3	The Knowledge Gap . . . . .	24
1.4	Structure of the Thesis . . . . .	25
<b>ii</b>	<b>DISCIPLINARY INSIGHTS</b>	<b>28</b>
<b>2</b>	<b>MARKET AND CUSTOMER SEGMENTATION</b>	<b>29</b>
2.1	Origin of the Concept . . . . .	31
2.2	Segmentability of Markets . . . . .	32
2.3	Variables for Segmentation . . . . .	34
2.3.1	Geographic Variables . . . . .	36
2.3.2	Demographic Variables . . . . .	37
2.3.3	Geodemographic Variables . . . . .	40
2.3.4	Behavioural Variables . . . . .	43
2.3.5	Psychographic Variables . . . . .	44
2.4	Customer Relationship Management . . . . .	48
2.4.1	Data mining in CRM . . . . .	49
2.4.2	Hyper-targeting . . . . .	50
2.4.3	Identifying the <i>Right</i> Future Customers . . . . .	52
2.5	Conclusions . . . . .	57
<b>3</b>	<b>THE MEDIA MARKET</b>	<b>58</b>
3.1	Understanding Audiences . . . . .	59
3.1.1	Television Ratings . . . . .	59
3.1.2	Social Media . . . . .	64
3.1.3	Video-on-Demand . . . . .	68
3.2	Audience Segmentation . . . . .	69
3.3	Interests and Recommendations . . . . .	72
3.4	Research Questions . . . . .	76
3.5	Conclusions . . . . .	78
<b>4</b>	<b>LARGE KNOWLEDGE BASES</b>	<b>80</b>
4.1	Emergence of Knowledge Bases . . . . .	81
4.2	Ontologies . . . . .	83
4.2.1	Cyc . . . . .	86
4.2.2	Freebase . . . . .	88
4.2.3	DBPedia . . . . .	89
4.3	Linked Data . . . . .	92

4.3.1	The Web of Data . . . . .	92
4.3.2	Technology Stack . . . . .	95
4.4	Detecting Interests . . . . .	100
4.4.1	Named Entity Recognition (NER) . . . . .	101
4.4.2	Named Entity Disambiguation (NED) . . . . .	102
4.4.3	Web Services . . . . .	105
4.5	Conclusions . . . . .	107
<b>iii</b>	<b>INTEGRATING INSIGHTS</b>	<b>109</b>
5	METHODOLOGY	110
5.1	Research Philosophy . . . . .	112
5.2	Research Strategy . . . . .	113
5.3	Research Process . . . . .	116
5.4	Data Collection and Transformation . . . . .	117
5.4.1	Metadata Stream . . . . .	118
5.4.2	Video Data Stream . . . . .	121
5.5	Data Mining . . . . .	125
5.5.1	Interest Segmentation . . . . .	125
5.5.2	Content Performance . . . . .	132
5.6	Methodological Considerations . . . . .	136
5.6.1	Ethics . . . . .	136
5.6.2	Reliability . . . . .	137
5.6.3	Validity . . . . .	138
5.6.4	Generalisability . . . . .	139
5.6.5	Limitations . . . . .	140
5.7	Conclusions . . . . .	140
6	RESULTS	142
6.1	Interest Segmentation . . . . .	142
6.1.1	Number of Concepts and Relevance . . . . .	142
6.1.2	Choice of Ontology . . . . .	147
6.1.3	Viewer Segmentation . . . . .	150
6.1.4	Viewers Clustering . . . . .	152
6.1.5	Addressing the Research Questions . . . . .	158
6.2	Content Performance . . . . .	163
6.2.1	Temporal Relativity . . . . .	163
6.2.2	Visualising Content Performance . . . . .	164
6.2.3	Visualising Relationships between Concepts . . . . .	166
6.2.4	Tracking Trending Interests . . . . .	168
6.2.5	Addressing the Research Questions . . . . .	169
6.3	Conclusions . . . . .	171
<b>iv</b>	<b>UNDERSTANDING</b>	<b>172</b>
7	DISCUSSION	173
7.1	Interdisciplinary Approach . . . . .	173
7.1.1	Customer Segmentation . . . . .	175
7.1.2	Media Market . . . . .	176
7.1.3	Large Knowledge Bases . . . . .	177

7.2	Contribution to Knowledge . . . . .	177
7.2.1	Main Finding . . . . .	177
7.2.2	Additional Findings . . . . .	182
7.3	Limitations . . . . .	183
7.4	Importance of the Findings . . . . .	186
7.4.1	Broadcasters and Advertising . . . . .	186
7.4.2	Enabling Technologies Growth Prospects . . . . .	187
7.4.3	Human Emotions and the Perception of Value . . . . .	188
7.4.4	Tracking and Privacy Concerns . . . . .	189
7.5	Recommendations for Further Work . . . . .	191
7.5.1	Extracting Concepts from Closed Captions . . . . .	191
7.5.2	Identifying Relationships in the LOD Cloud . . . . .	192
7.5.3	Generalising the Methodology . . . . .	192
8	CONCLUSIONS	194
	BIBLIOGRAPHY	197
A	HIGH RESOLUTION FIGURES	207

## LIST OF FIGURES

---

Figure 1	The Long Tail Phenomenon . . . . .	13
Figure 2	Mass Customisation . . . . .	14
Figure 3	Amazon.com Cross Selling Example . . . . .	16
Figure 4	Google’s Contextual Advertising . . . . .	17
Figure 5	Facebook Adverts Targeting . . . . .	18
Figure 6	TV Advertising Revenues . . . . .	19
Figure 7	Process for Interdisciplinary Research . . . . .	25
Figure 8	Thesis Structure . . . . .	27
Figure 9	Example market segmentation study . . . . .	30
Figure 10	Segmentation Variables . . . . .	35
Figure 11	Working generations in the UK . . . . .	39
Figure 12	Geodemographical clusters example . . . . .	42
Figure 13	VALS Framework . . . . .	47
Figure 14	The Knowledge Gap . . . . .	56
Figure 15	Nielson’s <i>Audience Measurement Process</i> . . . . .	63
Figure 16	Social Media correlation to TV Audiences . . . . .	67
Figure 17	TV Audience Measurement Timeline . . . . .	70
Figure 18	Visual representation of an ontology . . . . .	85
Figure 19	Alan Turing Infobox . . . . .	91
Figure 20	Linked Open Data – 5-star Rating System . . . . .	94
Figure 21	Linked Open Data Cloud . . . . .	95
Figure 22	Semantic Web Stack . . . . .	96
Figure 23	Interdisciplinary nature of the study . . . . .	111
Figure 24	<i>Knowledge Discovery in Databases</i> (KDD) . . . . .	117
Figure 25	Research Process Overview . . . . .	118
Figure 26	Video Player Tracking . . . . .	122
Figure 27	Singular Value Decomposition . . . . .	132
Figure 28	<i>Self-organising Map</i> Example . . . . .	133
Figure 29	Episodes similarity graph . . . . .	144
Figure 30	Words in programme description . . . . .	146
Figure 31	Words to Concepts Relationship . . . . .	146
Figure 32	Programmes and Concepts per Viewer . . . . .	151
Figure 33	Self Organising Maps Training . . . . .	156
Figure 34	Viewers per cell . . . . .	157
Figure 35	Affinity between cell and superconcepts . . . . .	158
Figure 36	Cluster Boundaries . . . . .	159
Figure 37	Travelling Affinity . . . . .	160
Figure 38	Design Affinity . . . . .	160
Figure 39	Generalist Affinity . . . . .	160
Figure 40	Cultural Affinity . . . . .	160
Figure 41	Technology Affinity . . . . .	160

Figure 42	Timeline for the total number of views . . . . .	164
Figure 43	Concepts ranked by relevance . . . . .	165
Figure 44	Episode Performance . . . . .	166
Figure 45	Concepts Co-Occurance . . . . .	167
Figure 46	Google Trends . . . . .	170
Figure 47	Linked Open Data Cloud – Present Version . .	208
Figure 48	Concepts Co-occurance – High Resolution . .	209

## LIST OF TABLES

---

Table 1	Life Style Dimensions . . . . .	45
Table 2	Example <i>Programme - Concept Mapping</i> . . . . .	121
Table 3	Details of Series Analysed . . . . .	124
Table 4	Example <i>User Statistics Data Store</i> . . . . .	125
Table 5	Example <i>Programme Statistics Data Store</i> . . . . .	126
Table 6	Various ontological classifications . . . . .	127
Table 7	Concepts to Genre Relationship . . . . .	145
Table 8	Ontology Comparison . . . . .	148
Table 9	Distribution of programmes and concepts . . .	152

## ACKNOWLEDGMENTS

---

I would like to thank my supervisor, *Professor Jay Bal*, for his essential role not only in coordinating my project, but also developing my critical reasoning, encouraging me to experience a wide range of situations instead of simply focusing on ticking the boxes, and most importantly for being a friend through all these years.

My most sincere gratitude to my wife *Karin*, for her amazing unconditional support which implied changing countries, cities, careers, and an increased threshold in terms of tolerance to accommodate the typical ups and downs of the PhD journey.

I would like to thank my parents for always creating a good environment for studying as well as reiterating the various implications of a solid formal education at different stages of my upbringing.

This research would not have been possible without the financial contribution received from the *Norwegian State Educational Fund*, for which I would like to thank the *Norwegian Government*.

I would also want to thank *Streamhub*, the partner company in this project, for their assistance with data collection, and always keeping an open mind about the ways in which academia can help industry and vice-versa.

I would like to thank *Dr. Sabin-Corneliu Buraga* and *Laurian Gridinoc* for inspiring me to do research in the area of the *Semantic Web* more than ten years ago, when the discipline was at a very early stage.

Finally, I would like to thank the developers and companies that either open sourced their products or made them available for research use, including but not limited to: *Apache Spark*, *Tableau*, *R*, *MatLab*, *IntelliJ*, *AlchemyAPI*, *Stardog*, *Gephi*, *Mendeley*, *MindManager*, *NVivo*. Having a set of great tools available, the researcher can focus solely on the problem they want to solve, and that is invaluable.



## DECLARATION

---

I declare that except where acknowledged, the material contained in this thesis is my own work and that it has neither been previously published nor submitted elsewhere for the purpose of obtaining an academic degree.

---

Radu Sora

## ABSTRACT

---

To better connect supply and demand for various products, marketers needed novel ways to segment and target their customers with relevant adverts. Over the last decade, companies that collected a large amount of psychographic and behavioural data about their customers emerged as the pioneers of *hyper-targeting*. For example, *Google* can infer people's interests based on their search queries, *Facebook* based on their thoughts, and *Amazon* by analysing their shopping cart history. In this context, the traditional channel used for advertising – the media market – saw its revenues plummeting as it failed to infer viewers' interests based on the programmes they are watching, and target them with bespoke adverts.

In order to propose a methodology for inferring viewers' interests, this study adopted an interdisciplinary approach by analysing the problem from the viewpoint of three disciplines: *Customer Segmentation*, *Media Market*, and *Large Knowledge Bases*. Critically assessing and integrating the disciplinary insights was required for a deep understanding of: the reasons for which psychographic variables like interests and values are a better predictor for consumer behaviour as opposed to demographic variables; the various types of data collection and analysis methods used in the media industry; as well as the state of the art in terms of detecting concepts from text and linking them to various ontologies for inferring interests. Building on these insights, a methodology was proposed that can fully automate the process of inferring viewers interests by semantically analysing the description of the

programmes they watch, and correlating it with data about their viewing history.

While the methodology was deemed valid from a theoretical point of view, an extensive empirical validation was also undertaken for a better understanding of its applicability. Programme metadata for 320 programmes from a large broadcaster was analysed together with the viewing history of over 50,000 people during a three-year period. The findings from the validation were eventually used to further refine the methodology and show that it is possible not only to infer individual viewers' interests based on the programmes watched, but also to cluster the audience based on their content consumption habits and track the performance of various topics in terms of attracting new viewers. Having an effective way to infer viewers' interests has various applications for the media market, most notably in the areas of better segmenting and targeting audiences, developing content that matches viewers' interests, or improving existing recommendation engines.

Part I

THE PROBLEM AREA

## INTRODUCTION

---

*Advertising works most effectively when it's in line  
with what people are already trying to do.*

— Mark Zuckerberg, CEO of Facebook

Marketers have long tried to understand consumers' needs and behaviours in order to alter demand for their products. Market segmentation, one of the main strategies for achieving this goal, allows organisations to divide a market into a number of segments with similar needs, interests, or priorities. There are various criteria used for segmentation, including the location of the consumer, demographic traits like age and income level, or behavioural patterns. For example, a hypothetical car manufacturer might choose to divide the market based on disposable income, country of residence, and primary use of the vehicle. Based on this model, potential customers with a high level of disposable income living in warm climates could then be targeted with advertisements for top-end convertible models. While the example is purposefully simplistic, similar strategies had a massive impact on the way organisations segmented and targeted customers for many decades, representing the core of the marketing efforts for many companies.

Following the digital revolution, the overall consumption habits changed dramatically. Traditionally, consumers relied heavily on brick and mortar stores that – given their space limitations – only stocked a small number of mainstream products. However, the

expansion of the Internet allowed consumers to find and purchase products online, from retailers that did not require a high street presence, and could operate without even stocking a product until there is a demand for it. This led to the emergence of the *Long Tail* phenomenon (Figure 1), a term coined by Chris Anderson for describing the shift in the markets from a small number of mainstream products towards a high number of niche products (Anderson 2008).

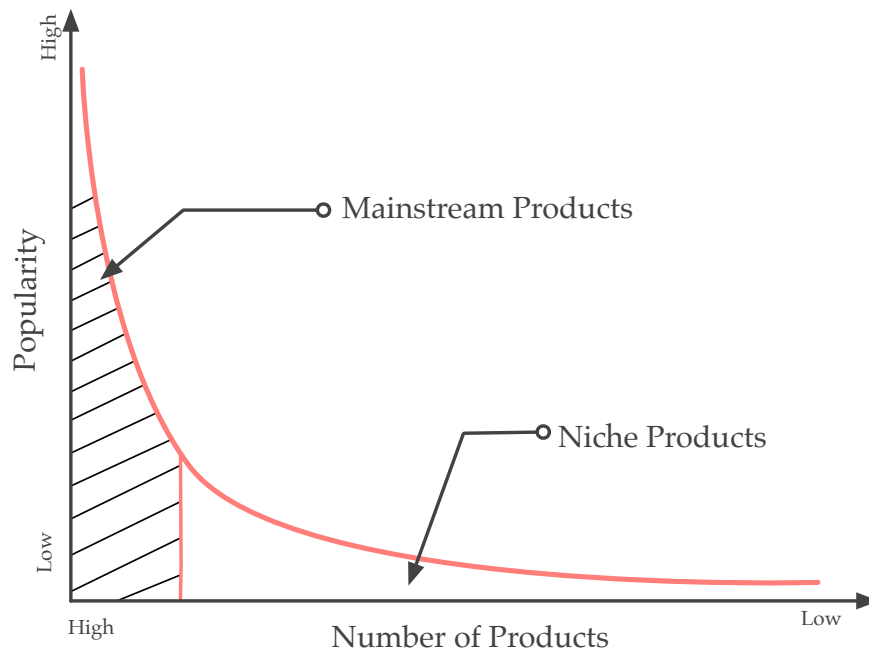


Figure 1: The *Long Tail* Phenomenon describes the shift in the market from a small number of mainstream products to a high number of niche products (Anderson 2008)

In reaction to the changes in the market, companies had to change their strategies in order to ensure that their products can satisfy an increasingly various range of needs, while at the same time being able to market them to the ones interested. On the product front this led to efforts towards mass-customisation of products, a process where one base product is customised in order to appeal to different user segments (Tseng et al. 2007). For example, consumers

can now configure the colours and specifications of their *Mini* cars, choose between different materials and bracelets for the *Apple Watch*, or customise the colours and prints on their *Converse* footwear (Figure 2).

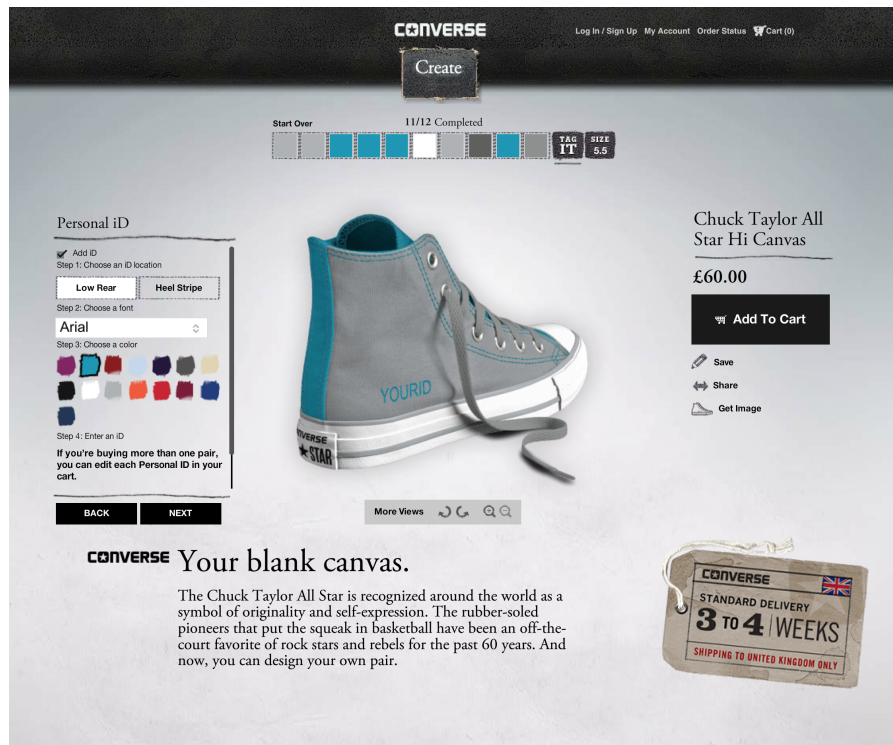


Figure 2: Example of mass customisation for Converse footwear: clients can choose the colour and print for every detail of the product in order to better match them with their taste and preferences

On the marketing side of things, companies started harvesting the vast amounts of data they collect from their clients, and using it to better target them with products and services they are more likely to need. This process, typically referred to as hyper-targeting, moved the focus from wide market segments defined by geography or demographics, to highly specific customer segments, sometimes limited to only one person. To obtain the micro segments, new criteria for segmentation needed to be employed. In this respect, two main categories of variables emerged as most relevant: behavioural ones like usage rate, average transaction amount, or churn probability;

and psychographic ones such as consumers' interests, opinions, and values.

### 1.1 THE PROBLEM AREA

Given the new realities in the market, companies that collected vast amounts of behavioural and psychographical data emerged as pioneers in hyper-targeting (Jovanovic 2014). Being able to define a large number of customer segments, and then targeting them with highly relevant advertising messages, translated into exponential growth. For example, Amazon.com – the largest online retailer in the world – grew from 1.5 million customers in 1994 to more than 270 million in 2014<sup>1</sup>. Having access to shopping cart information for a high number of customers allowed Amazon to understand and exploit the behavioural patterns in the market. This can represent one of the reasons for which the online retailer evolved from an online store to an online marketplace. Currently, any individual or company can sell products via Amazon for a fee. While not directly stated by Amazon, having access to the behavioural information from a wider range of clients, in addition to the ones already served by them, can be deemed more important than the investment required for providing the services. By analysing the large volume of transactions, the online retailer can cross-sell a large number of products based on similarities between shopping carts (Figure 3). Moreover, knowing the temporal nature of various transactions, companies can predict with a high degree of accuracy when a person is most likely willing to make a purchase for a specific item (Kamakura 2008).

---

<sup>1</sup> Number of worldwide active Amazon customer accounts from 1997 to 2014 (in millions) – <http://goo.gl/HUNEN4>



## Customers Who Bought This Item Also Bought

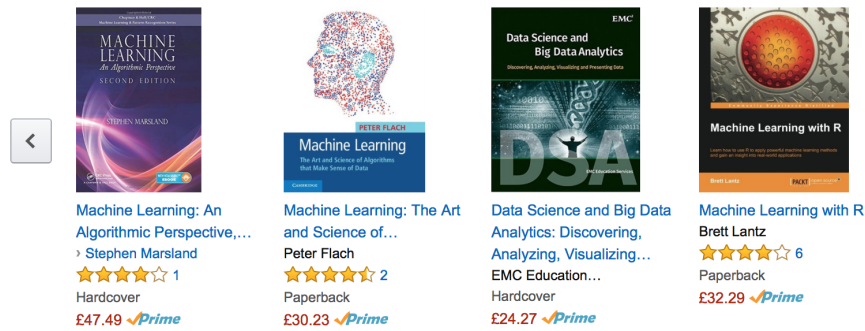


Figure 3: Amazon.com leverages the vast amount of behavioural data collected in order to cross-sell products

Google, a large technology company well known for their search engine, generated a large portion of its income from online advertising. In 2011 reports showed that 96% of the total revenue was generated from advertising only. The main success factor for Google was the ability to match the needs of their users with the advertisements. The patented technology, commonly referred to as contextual advertisement, allowed the company to deliver advertisements based on immediate users needs (Schmitter et al. 2005; Anagnostopoulos et al. 2007). For example, a student searching for an MBA course, would be served advertisements for relevant universities offering such a course (Figure 4). Intuitively, the efficiency of such messages is higher compared to traditional channels like press or television, due to the fact that the message is only displayed to consumers with a certain interest into the area, therefore increasing conversion rate and reducing the amount of resources spent on advertising.

Various social networks, most prominently Facebook, created online environments where their users can create online profiles, and subsequently share various thoughts, photos or videos with their friends or extended social circles. Following a very aggressive growth

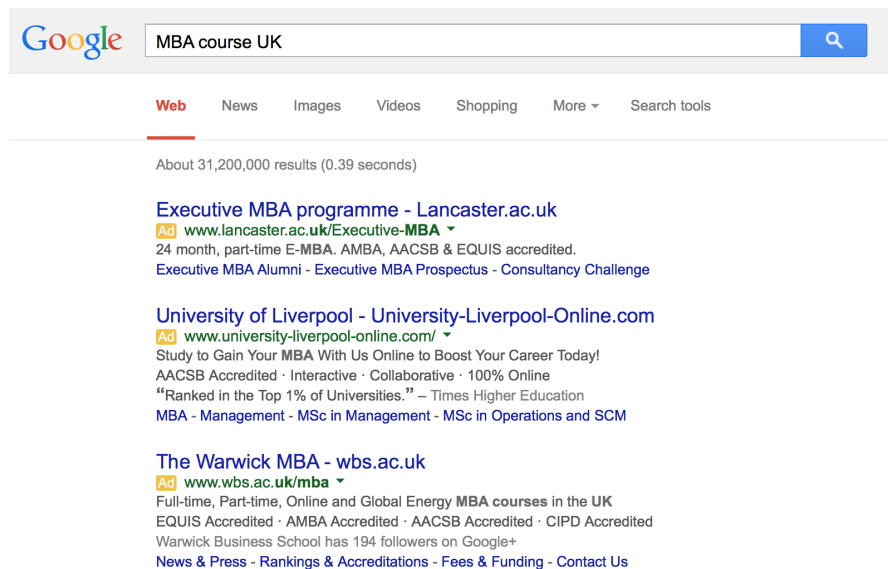


Figure 4: Google infers users' interests based on their search queries, and matches them with the advertising campaigns. When searching for an MBA course, users are presented with relevant messages from various universities, while also taking into consideration the geographic locations from which the query originated

path over the years, Facebook reported over 1.44 billion users in 2015 for their main service (Facebook.com), with another 1 billion shared across some of their secondary services (e.g. Instagram, Whatsapp)<sup>2</sup>. Similarly to Amazon and Google, having access to a high volume of data about their user's thoughts and needs, placed Facebook in a privileged position. The social network can monetise its service by serving highly targeted advertisements base on psychographic criteria, like interests, thoughts and opinions (Figure 5).

While companies that pioneered hyper-targeting grew considerably over the last decade, the more traditional business models that depended on advertising, like press and broadcasters, have been through difficult times. The recent market reports show how the television advertising is constantly shrinking, with both its share of the total advertising revenue and the absolute value

<sup>2</sup> Facebook Q2 2015 Earnings Report – <http://goo.gl/b5fBwP>

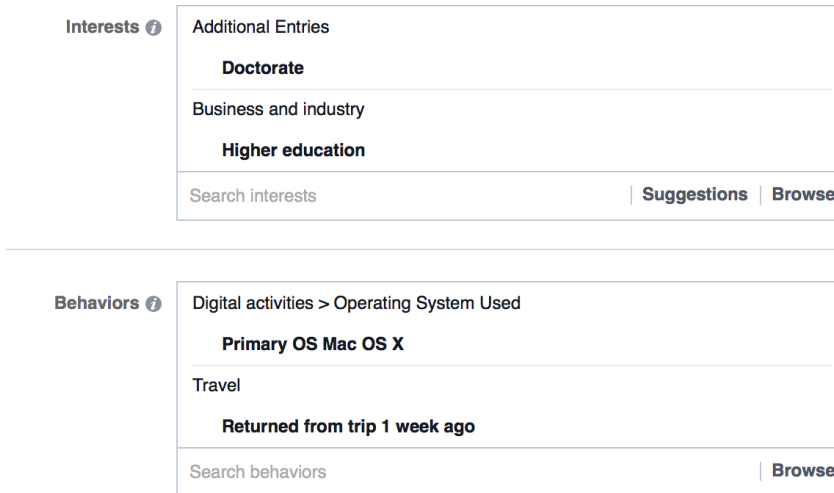


Figure 5: Facebook’s sophisticated targeting mechanism allows companies to define the interests and behaviours of their desired customers. In the example above, one advert is configured to target people with an interest in a Doctorate degree, using an Apple computer, that returned from a trip in the last week

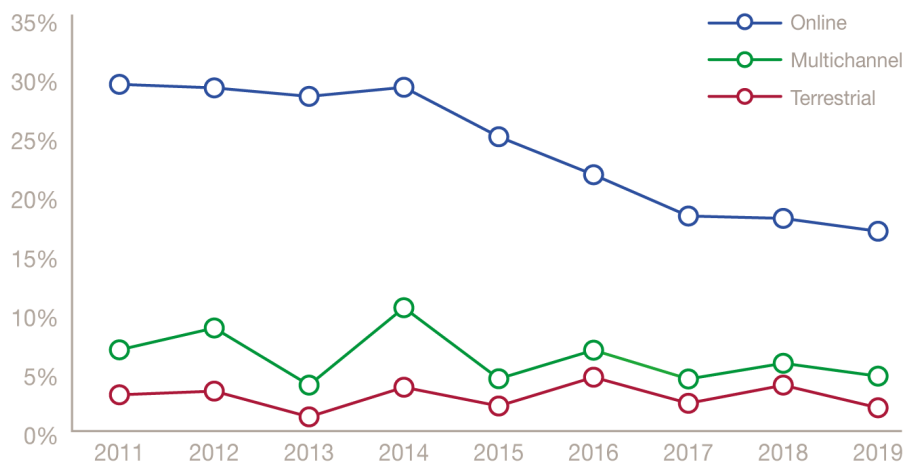
projected to decline over the coming years (Nilsen et al. 2015) (Figure 6). This is counter intuitive, as the media industry could potentially leverage the same hyper-targeting capabilities as the newcomers in the market:

- There is a proven correlation between people’s interests and what they decide to watch online (Taneja et al. 2012). This trend has been accentuated in the recent years as the patterns of consumption changed from watching live content in front of a television set, to watching video on demand on various devices like computers, tablets or smartphones. This means that the viewers interest could potentially be inferred based on the content they watch;
- Following the switch to digital technologies, broadcasters have the opportunity to collect behavioural data about their viewers. This could be used for micro-segmentation, and then

programmatically deciding which advert to display to which viewer based on their inferred interests;

- As opposed to the new players in the advertising market, broadcasters already have a large and typically loyal audience, so further effort needs to be put into monetising it better.

Global terrestrial, multichannel and online TV advertising revenue year-on-year growth (%), 2011–2019



Source: Global entertainment and media outlook 2015–2019, PwC, Ovum

Figure 6: The global revenues from online TV advertising are projected to go down in the coming years at the expense of other type of advertisements; Source: (Nilsen et al. 2015)

However, in the current media landscape, broadcasters still air the same advertisements for all their viewers, without taking advantage of the data they collect about individual viewers. Using behavioural and psychographic variables could help hyper-target viewers with adverts that they are more likely to resonate with. In addition to the ability to monetise their services better, being able to segment audiences by interests has implications also from a content creation angle. Production teams could more easily understand which themes appeal to viewers more than others, which topics need more attention and which need less, etc.

## 1.2 PREVIOUS RESEARCH

The problem of segmenting customers by their interests and the applications of it in the media industry have been studied by different research communities. The most relevant angles to approach the problem, from the area of marketing, media, and knowledge bases will be briefly presented.

### 1.2.1 *Customer Segmentation*

From a marketing perspective, there are various studies that analysed the advantages and disadvantages of using segmentation criteria based on geographic, demographic, geodemographic, behavioural, and psychographic variables. There is also consensus that behavioural and psychographic variables are the most effective, as they can explain consumer decisions where other variables fail to differentiate between segments (Lin et al. 2002). For example, in one study car manufacturers concluded that some of their models need to be marketed to psychologically young people, emphasising that psychographic variables are stronger than demographic traits for predicting consumer behaviour (Chon 2006). Similarly, other studies showed that based on their interests and perception of value, consumers tend to trade-up for a number of items they are interested in, while trading down for others (Silverstein et al. 2003). For example, someone passionate about photographic equipment might choose to spend more money on a high-end camera, while choosing to spend less on household appliances.

Following the wide adoption of Customer Relationship Management (CRM) technology, companies started collecting a large

volume of data about their customers. This allowed them to segment their client base by all the criteria mentioned in the literature, with the exception of psychographic variables like interests and opinions. The methods for doing psychographic segmentation used in marketing discipline involve survey-type research, generally performed on small sample of subjects with low response rates (Wedel 2000; Peltier et al. 1997). The practical difficulties of undertaking this type of research restrict companies in taking full advantage of the micro segmentation and hyper-targeting possibilities. While there were some studies that explored the possibility of inferring consumers interests based on their shopping cart information, most notably Miguéis et al. (2012) who proposed a method for segmenting customers based on their lifestyle, the application of the methodology for grocery shopping represents a sub-optimal fit due to the fact that most of the products purchased are for satisfying a basic need, and therefore do not necessarily reflect people's interests.

#### 1.2.2 *Media Market*

The research undertaken in the media sector for understanding audiences has been mostly centred around television ratings. These are estimates of the audience size for different programmes, channels, and locations based on a sample of around 5,000 households in the UK. Knowing the demographics and viewing habits for this sample, the *Broadcaster's Audience Research Board* (BARB) is extrapolating the data for the whole population in order to predict how many people watched a certain programme, segmented by various demographic traits. In addition to television ratings, other research studies and companies attempted to analyse social media feeds like Facebook and Twitter in order to predict the size and interests of audiences. While

the results have been encouraging, this methodology is restricted by the fact that only a small number of successful shows generate enough social media activity for the analysis to be meaningful.

More recently, the transition from analogue to digital platforms allowed broadcasters to understand what their audiences are watching without relying on sample data from the television ratings. When video content is being watched on set-top boxes (e.g. Sky, Virgin, etc), online players (e.g. iPlayer, ITV Player) or mobile and tablet applications, broadcasters can track the viewing history for every individual viewer. The access to this level of data allowed companies like Netflix to make use of sophisticated recommendation engines in order to predict what people will be interested in watching. However, most of the research done on recommendation engines is either based on collaborative filtering (analysing similarities between users), or on analysing the content of the show in terms of actor or director names. While the choice of actors can influence a viewers decision to watch a programme, it does not provide a holistic view of someone's interests.

### 1.2.3 *Large Knowledge Bases*

In order to infer viewers' interests, the content of a programme needs to be analysed in order to detect the relevant concepts. Each of these concepts could represent an interest, and by mining the data collected from online viewing, one can infer which concepts are of interest to whom. For example, while analysing the programme description for an episode of *Top Gear*, an algorithm should ideally be able to extract concepts like: automotive, the various car brands being showcased, the location where the programme was filmed,

or the name of the guests. There has been a wealth of research in the NLP area (Natural Language Processing) around *Named Entity Recognition* (NER) and *Named Entity Disambiguation* (NED) (Marsh et al. 1998; Carreras et al. 2003; Mihalcea et al. 2008; Milne et al. 2008). These techniques allow the detection of entities in text, and the linking to large knowledge bases to better understand the context. For example, having detected from the textual description of a *Top Gear* episode that the filming location was *Bali*, by linking the entity to the corresponding class in *DBPedia* (a large knowledge base built on top of Wikipedia data), one can infer that the show was actually filmed in Indonesia, on an island known for tourism, tropical weather, and predominantly Hindu.

Being able to infer all the additional facts based on the concepts identified in a programme's description could potentially allow broadcasters to understand the underlying reasons for which certain viewers watch various programmes. Given that *Watson* – IBM's cognitive computing solution – has successfully competed and won in a *Jeopardy* contest against the best ranked human players, there is little doubt that the technology has evolved far enough to allow computers to understand the meaning of text and perform inference for answering questions (Ferrucci et al. 2013). Similarly, the same type of technology has been used in the Finance sector, where large players in the market like *Bloomberg* or *Thomson Reuters*, automatically analyse press and social media content in order to make sub-second decisions for selling or buying stocks<sup>3</sup>.

---

<sup>3</sup> Sentiment Analysis of Financial News and Social Media – <http://goo.gl/WoYYYY>



### 1.3 THE KNOWLEDGE GAP

Based on the literature analysed as well as the realities in the industry, there is a well defined knowledge gap in terms of inferring viewer's interests from the video content watched. Moreover, the existing body of knowledge related to the subject of this study is divided between three different academic disciplines: *Customer Segmentation*, *Media Research*, and *Large Knowledge Bases*. In order to address the knowledge gap, a holistic, interdisciplinary approach was deemed most appropriate. This is rooted in the fact that studying the hypothesis that viewers' interests can be inferred by semantically analysing the programmes' descriptions from the perspective of one discipline alone does not unveil all the limitations of the proposed methodology. For example, from a marketing perspective, psychographic variables like interests and opinions are deemed most relevant in terms of predicting customer behaviour. However, for inferring these variables, a good understanding of the type of data that can be used in the media sector is needed. Similarly, the advantages and disadvantages of the technologies for extracting concepts from text need to be understood and tested in the context of this specific problem.

In order to have a holistic understanding of the problem, this study follows the *Process for Interdisciplinary Research* (Figure 7). First, the various insights from the three disciplines were analysed in relation to the knowledge gap. Secondly, the insights were integrated into a coherent methodology for inferring viewers interests based on the content watched. For validating the methodology, an empirical study was performed, based on the viewing patterns of 57,471 persons. The data was analysed over a period of three years, and

plotted against the 5,264 concepts identified in the 320 programmes analysed. The proposed methodology was deemed successful not only for segmenting the viewers based on their interests, but also for assisting broadcasters in creating better programmes, reacting to the topics trending in social media, or nudging viewers behaviour.

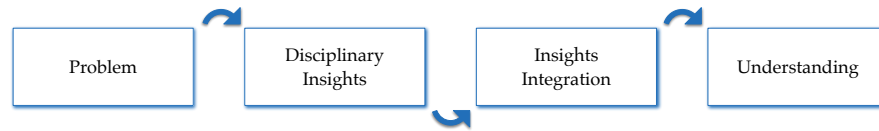


Figure 7: The *Process for Interdisciplinary Research* defined by Repko (2011) provides a theoretical way for analysing the problem, documenting the various disciplinary insights, and then integrating these for creating new knowledge

#### 1.4 STRUCTURE OF THE THESIS

The structure of the thesis can be consulted in [Figure 8](#) and is organised as follows:

- **Chapter 2** reviews the related work in the area of *Market and Customer Segmentation*, with a focus on the variables used for segmentation and their efficiency for predicting consumer behaviour. The advances in technology in terms of *Customer Relationship Management* solutions are analysed, along with the methods for segmenting customers based on behavioural data. A knowledge gap is identified in regards to inferring psychographic variables based on the interaction between customers and products;
- **Chapter 3** analyses the existing methodologies for understanding audiences behaviour in the context of the media market. The various methods for data collection and segmentation are presented with their advantages and disadvantages. Taking into consideration the specificities of

the market, the knowledge gap is translated into four research questions: two address the interest segmentation, while the other two focus on potential applications of the newly defined segments that would further benefit broadcasters;

- **Chapter 4** provides an overview of the various technologies that can be used for identifying concepts in text, and indirectly the interests in programmes' descriptions. The evolution of large knowledge bases, as well as the algorithms for *Named Entity Recognition* and *Named Entity Disambiguation* are presented. Finally, a list of open source and commercial solutions are assessed in the context of this study;
- **Chapter 5** defines the proposed methodology for inferring viewers interests based on the video content watched, detailing the actual steps needed to address the research questions. The choices in terms of research philosophy and strategy are presented, followed by the data collection and analysis methods used for each research question. Finally, the methodological considerations in terms of reliability, validity, generalisability and limitations are detailed;
- **Chapter 6** presents the results of the methodology proposed in the previous chapter when applied to the empirical study. The results are discussed in relation to how they address the research questions, but also in terms of their limitations and potential outcomes in other market sectors;
- **Chapter 7** provides a high level discussion of the findings by changing the focus from the specific to general. The implications of the methodology described are discussed in the context of technological advancements, intelligent agents, the importance of human emotions, and the perception of value;

- **Chapter 8** formulates the conclusions of this study and suggests additional areas of work for improving the existing methodology and applying it in other areas of interest.

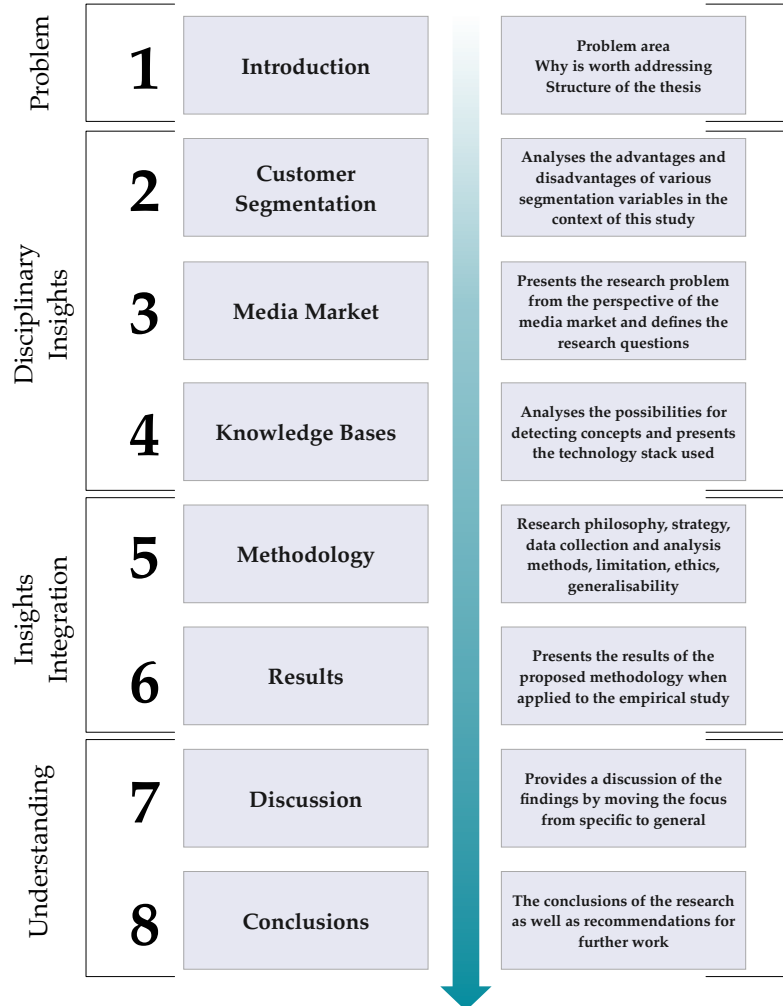


Figure 8: The structure of the thesis

Part II

DISCIPLINARY INSIGHTS

## MARKET AND CUSTOMER SEGMENTATION

---

One of the main subjects of interest for marketers – market segmentation – is not a new concept, having been introduced in the literature back in the 1950s (Smith 1956). It is commonly defined as the process of dividing a broad market into distinct groups of consumers that share the same needs and priorities, and then implementing strategies to target them. For example, airlines could choose to segment their customers based on the reason of travelling between business or leisure, banks make a distinction between people who take credits and those who make deposits, while telecommunication companies might divide the market according to the use of digital devices (Figure 9). In all these instances, marketing and pricing strategies are specifically tailored to each segment. Since its inception, market segmentation was widely adopted, discussed in both industry and academia, often praised for its benefits and criticised for its shortcomings. The high number of assumptions and variables associated with its implementation lead to scholars questioning if it is “too complex for mere mortals to understand” (Green 1977). The topic evolved alongside with the technical methods employed to achieve it, from areas like statistics, data analytics, and customer relationship management. The high volume of consumer data currently collected by companies, as well as the increase in computing power, created new angles to tackle the problem which are now being investigated by the research community.

**Different segments of consumers in China vary widely in their use of digital applications and devices.**

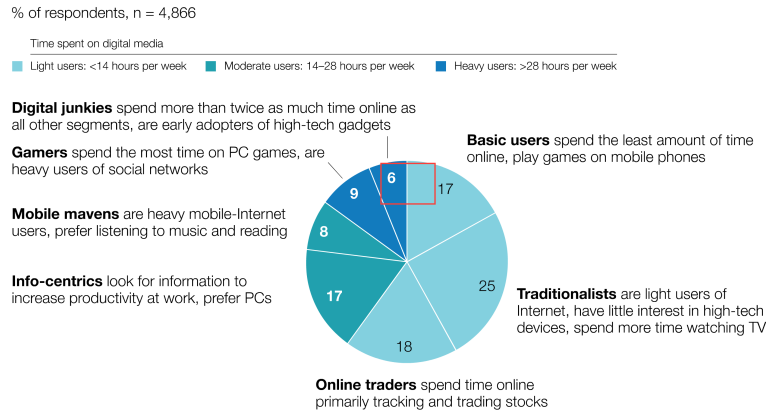


Figure 9: Example of market segmentation study based on adoption and use of digital applications and devices; Source: (D. Lin et al. 2011)

In this chapter, the origin of the concept and the reasons that lead to its existence are presented first. An overview of the segmentability of markets is provided next, while emphasising a set of criteria that can be used to assess it. In the next section, the advantages and disadvantages of using segmentation variables from five different categories are discussed: geographic, demographic, geodemographic, behavioural and psychographic. While most of the segmentation studies used to be performed based on survey data, the development of technologies to track consumer interactions made it possible to segment markets in new ways. An overview of the existing techniques in the customer relationship management field is presented in the last section, along with their perceived effectiveness in the context of hyper-targeting. Finally, based on the literature reviewed, a knowledge gap is identified and the conclusions are formulated.

## 2.1 ORIGIN OF THE CONCEPT

The concept of market segmentation was introduced by Smith (1956) in the 1950s. He defined it as the process of dividing a heterogeneous market into smaller homogeneous markets in regards to their product preferences so that each can be targeted with a different advertising strategy.

Traditionally, marketing strategies would emphasise the differences between the organisation's products and those of the competitors. In its incipient form, market segmentation was proposed as an alternative to product differentiation, which could "bend demand to the will of the supplier" (Smith 1956). From this questionable viewpoint, companies would be able to freely alter demand for certain products only by better targeting individuals with advertising campaigns. This initial angle, possibly rooted in the inability of companies to easily customise their products at the time, lead to some researchers blaming marketing for manipulating consumers (Galbraith 2007). Later on, other studies considered market segmentation to be a complement to product differentiation, with a central role in guiding the product development. The two different angles contributed to an overall confusion about the definitions of *market segmentation* and *product differentiation*. Dickson et al. (1987) performed an analysis of the two conflicting theories, concluding that the two concepts are complementary, as opposed to alternatives. This new perspective constituted a paradigm shift in the field, moving the focus from the product to the customers and the means to serve them in a more bespoke way.



## 2.2 SEGMENTABILITY OF MARKETS

Empirical evidence suggests that segmenting a market might not always be beneficial and that different factors need to be considered beforehand. Fifteen years after the introduction of the concept, Young et al. (1978) argued that too often segmentation studies have had disappointing results, and this is mainly due to the lack of actionability from a marketing standpoint. For example, while it is possible to segment users of a certain product based on their opinion about the brand with the help of surveys, the resulting segments lack actionability because marketing teams cannot target only one of the segments with a specific message, but need to broadcast the same message to everyone. Young et al. (1978) also defined a set of guidelines for identifying cases when segmentation would not be useful: market too small to justify marketing to only a section of it; situations when observation infers that frequent users generate most of the profit; or the brand is dominant in the market. While these circumstances represent the reflection of the market conditions at that time and can be interpreted, there were also attempts to quantify the segmentability of a market.

Dolnicar et al. (2005) introduced a method for simulating market segmentation in the context of competitiveness, with the objective of deriving early insights into the performance of a potential segmentation study. The computer simulation model factored in different variables like price, desire, advertising message, advertising budget, and target segment. Based on these, advertising effect is estimated, and brand choice and brand perception for different market segments can be predicted. While the predictive capabilities of a simulation model are generally relatively weak, the realities

of different markets in terms of price ranges and budgets make this methodology a good place to start for estimating the potential segmentability of a market. In addition to this method, in order to validate any proposed segmentation model, Wind (1981) emphasised the need for management to be able to measure the performance of their products at segment level (e.g. profitability, growth, market share, etc). This ability will enable them to make decisions based on analysing the data, as opposed to relying solely on intuition.

In an attempt to bring more structure to the criteria for market segmentation, Kotler (1980, pp. 291–309) defines three characteristics of useful market segments: *measurable*, *accessible*, and *substantial*. A marketing manager first needs to be able to measure the size of a segment and its individual characteristics, like purchasing power or shopping frequency. Secondly, he or she needs to have a way to access that segment with one of the available marketing vehicles. Finally, the segment must be large enough to justify its existence in the first place. Later on, in order to reflect the changes in the market, the list is complemented by two more characteristics: *differentiable* and *actionable* (Kotler et al. 2011). These highlight the fact that the segments need to respond differently to a marketing stimulus, and at the same time be specific enough so that programmes can be developed for them. While ultimately there are no guarantees for the success of a segmentation study, having a consistent set of criteria to compare the alternatives can improve the quality of the analysis. The ones described above will be used to evaluate the segmentation variables presented in the next section.

### 2.3 VARIABLES FOR SEGMENTATION

The main criteria to decide upon in any market segmentation exercise is the variable or variables to segment on. There is a broad range of options, as any characteristic of a consumer can be considered a variable (Figure 10). In the literature, there are a number of attempts to classify these variables. Most notably, Kotler et al. (2011) proposes a classification based on the type of variable into four main areas: geographic, demographic, psychographic, and behavioural. Alternatively, Raaij et al. (1994) classifies the variables based on two dimensions: level of generality and objective versus subjective character. General variables are related to the consumer's characteristics or behavioural patterns, while domain specific variables are related to a specific product or brand. Similarly, objective variables are the ones that can be assigned without much disagreement between researchers (e.g. income, age, education, frequency of use), while the subjective ones are mental constructs of the consumers, typically identified by survey-type studies (e.g. lifestyle, personality, purchase intention, values). A third classification, also across two dimensions, is proposed by Wedel (2000). In this case, the variables are also divided based on the level of generality (general versus product-specific) while the second dimension concerns the observability factor (observable versus unobservable). While all the classifications models are valid in their own right, the model proposed by Kotler et al. (2011) will be used due to the fact that each category of variables shares the same methods of collecting and analysing the data – therefore making comparisons between categories more meaningful for the purpose of this study.

TABLE 8.1 Major Segmentation Variables for Consumer Markets	
Geographic region	Pacific Mountain, West North Central, West South Central, East North Central, East South Central, South Atlantic, Middle Atlantic, New England
City or metro size	Under 5,000; 5,000–20,000; 20,000–50,000; 50,000–100,000; 100,000–250,000; 250,000–500,000; 500,000–1,000,000; 1,000,000–4,000,000; 4,000,000+
Density	Urban, suburban, rural
Climate	Northern, southern
Demographic age	Under 6, 6–11, 12–17, 18–34, 35–49, 50–64, 64+
Family size	1–2, 3–4, 5+
Family life cycle	Young, single; young, married, no children; young, married, youngest child under 6; young, married, youngest child 6 or older; older, married, with children; older, married, no children under 18; older, single; other
Gender	Male, female
Income	Under \$10,000; \$10,000–\$15,000; \$15,000–\$20,000; \$20,000–\$30,000; \$30,000–\$50,000; \$50,000–\$100,000; \$100,000+
Occupation	Professional and technical; managers, officials, and proprietors; clerical sales; craftspeople; forepersons; operatives; farmers; retired; students; homemakers; unemployed
Education	Grade school or less; some high school; high school graduate; some college; college graduate
Religion	Catholic, Protestant, Jewish, Muslim, Hindu, other
Race	White, Black, Asian, Hispanic
Generation	Silent Generation, Baby boomers, Gen X, Gen Y
Nationality	North American, Latin American, British, French, German, Italian, Chinese, Indian, Japanese
Social class	Lower lowers, upper lowers, working class, middle class, upper middles, lower uppers, upper uppers
Psychographic lifestyle	Culture-oriented, sports-oriented, outdoor-oriented
Personality	Compulsive, gregarious, authoritarian, ambitious
Behavioral occasions	Regular occasion, special occasion
Benefits	Quality, service, economy, speed
User status	Nonuser, ex-user, potential user, first-time user, regular user
Usage rate	Light user, medium user, heavy user
Loyalty status	None, medium, strong, absolute
Readiness stage	Unaware, aware, informed interested, desirous, intending to buy
Attitude toward product	Enthusiastic, positive, indifferent, negative, hostile

Figure 10: Comprehensive list of potential variables that can be used for market and customer segmentation studies; Source: (Kotler et al. 2011)

The choice of a segmentation variable or variables is a subject well analysed in the literature. Arguably, any of the alternatives might be a good choice for a segmentation study, depending on the defined objective. Wind (1981) points out that the primary aspect to decide upon choosing a segmentation variable is the management need or motivation for taking on the study. However, he also stresses that some variables tend to perform better than others. More recently, with the explosion of data sources and computation power available, the management team do not always have a goal in mind before analysing the data. It is quite common to try to analyse the patterns in the data in order to derive insights (Baecke et al. 2009), and therefore considering multiple variables at the same time can increase the chances of spotting a meaningful pattern.

### 2.3.1 *Geographic Variables*

Geographic segmentation is the first type of segmentation successfully used and one that is relatively simple to implement. The process is based on dividing the market into geographical units like countries, regions, neighbourhoods or postcodes. In some cases, the units can be obtained from the values of other variables, like population density or climate. When geographic segmentation is used, the assumption behind it is that the needs of the customers, or the ways to fulfil those needs, vary geographically (Beane et al. 1987).

In some situations, organisations only operate in some areas, because the demand for a type of product only exists over there, as in the case of winter clothing or snorkelling equipment. In other instances, the organisation operates across the geographical boundaries but is tailoring its offering and marketing techniques to the specificities of each area. This scenario usually has to do with the demand for certain products varying between regions as a consequence of local tastes and cultures. In the United States, for example, the effects of local subcultures in the context of consumer behaviour were analysed by Hawkins et al. (1981). The research shows how coffee consumption levels are relatively similar across the United States, but the way in which the coffee is prepared and consumed vary widely between the four main geographical regions (East, Midwest, South, and West). However, even when the product demand is not varying geographically, companies may still personalise their marketing strategies in order to get as close to the customers as possible.

Identifying with the local community and contributing to their causes proved to be an effective marketing technique sometimes labelled as *grassroots marketing* or *astro-turfing* (Kotler et al. 2011). The concept describes the methods that disguise the company message as an authentic grassroots movement. For example *Waitrose* – a large supermarket chain in the United Kingdom – distributes a thousand pounds for each of their branches to three local *good causes* based on their customer's preferences<sup>1</sup>. The effectiveness of these methods along with the ethical practices surrounding them are still a subject of debate.

Geographical variables tend to rate high on the measurable and accessible criteria and proved useful for very specific scenarios. However, they do not provide any substantial and differentiable segments in more complex situations, when there are not clear differences in terms of consumer's needs solely based on their location.

### 2.3.2 Demographic Variables

Demographic segmentation divides the market based on one or more demographical traits. Examples of such traits include gender, occupation, income, race, religion or education level. This method is very well fitted for situations when there is a high probability for a product to fulfil a need that is only present in some demographic. Kotler et al. (2011) argues that demographical traits are often associated with the needs and wants of customers. The same views are echoed by Blattberg et al. (2008) that points out that they correlate well with purchase likelihood and frequency. In addition, they are

---

<sup>1</sup> *Waitrose* – Community Matters – <http://goo.gl/84EJ8e>

easy to understand, and because of that are sometimes used to describe segments obtained based on other segmentation variables.

In regards to gender segmentation, there are many examples of products that became successful while targeting only one gender. However, also when targeting both, having a solid understanding of the gender segmentation allows for better communication strategy. This is particularly important as reports now show that women are driving the global economy (Aguirre et al. 2012) as well as controlling 80% of the decisions about buying consumer goods and services, and 75% of decisions regarding buying new homes (Barletta 2006). Moreover, it was noticed that women tend to value different criteria than men in their purchase decisions. In the context of buying a car, for example, evidence shows that women value interior design and security systems, as opposed to men that take into consideration performance and exterior design (Singh 2014).

Similarly, generations have specificities based on the factors that influenced and shaped their values and beliefs. Various studies identified these differences and provided practical advice for ways of communicating to them. An overview of the generations in the United Kingdom can be consulted in [Figure 11](#). As a general trend, each generation tends to have different aspirations and attitudes towards technology and career. Moreover, the methods of communicating vary from one to another. Being shaped by events, the structure of the generations (or age cohorts) is different from country to country (Schewe et al. 2004). While marketers need to be aware of the differences in order to tune their strategies, the resulting segments are – in general – too broad to be actionable. Typical examples of actions resulting from a generation segmentation are the use of

icons of a specific generation in advertisement or the development of products that fulfil the need for one age cohort only (Kotler et al. 2011).

Characteristics	Maturists (pre-1945)	Baby Boomers (1945-1960)	Generation X (1961-1980)	Generation Y (1981-1995)	Generation Z (born after 1995)
Formative experiences	Second world war Rationing Fixed-gender roles Rock 'n' Roll Nuclear families Defined gender roles – particularly for women	Cold war Post-War boom "Swinging Sixties" Apollo Moon landings Youth culture Woodstock Family-orientated Rise of the teenager	End of Cold War Fall of Berlin Wall Reagan / Gorbachev Live Aid Introduction of first PC Early mobile technology Latch-key kids; rising levels of divorce	9/11 terrorist attacks PlayStation Social media Invasion of Iraq Reality TV Google Earth Glastonbury	Economic downturn Global warming Global focus Mobile devices Energy crisis Arab Spring Produce own media Cloud computing Wiki-leaks
Aspiration	Home ownership	Job security	Work-life balance	Freedom and flexibility	Security and stability
Attitude toward technology	Largely disengaged	Early information technology (IT) adaptors	Digital Immigrants	Digital Natives	"Technoholics" – entirely dependent on IT, limited grasp of alternatives
Attitude toward career	Jobs are for life	Organisational – careers are defined by employers	Early "portfolio" careers – loyal to profession, not necessarily to employer	Digital entrepreneurs – work "with" organisations not "for"	Career multitaskers – will move seamlessly between organisations and "pop-up" businesses
Signature product	Automobile	Television	Personal Computer	Tablet / Smart Phone	Google glass, grapheme, nano-computing, 3D printing, driverless cars
Communication media	Formal letter	Telephone	E-mail and text message	Text of social media	Hand-held (or integrated into clothing) communication devices
Communication preference	Face-to-face	Face-to-face ideally, but telephone or e-mail if required	Text messaging or e-mail	Online and mobile (text messaging)	Facetime
Preference when making financial decisions	Face-to-face meetings	Face-to-face ideally, but increasingly will go online	Online – would prefer face-to-face if time permitting	Face-to-face	Solutions will be digitally crowd-sourced

Figure 11: An overview of the working generations in the United Kingdom (Hilton 2013)

Income is another important demographic trait. While some products target specifically affluent users, with high disposable income, others focus on the lower-income segment while generating profit from a higher volume of sales with a lower margin. Nevertheless, consumers have a lot of mobility between the segments given their interest for a certain product, the overall economical situation, and other influencing factors. Silverstein et al. (2003) observed that more than 80% of the consumers in the United States *trade-down* for up to five product categories, while approximately 60% *trade-up* for two product categories. He defines this category of people *part-martyr and part hedonic*, and argues that the choice is based on the emotional benefits derived from products. For example, someone with an interest in photography can choose to spend more for a better camera model, while spending less for rent or entertainment.



Demographic variables, while incontestably being the most widely used in the industry, have had their relevance debated in many articles. After reviewing the techniques for market segmentation, Hiziroglu (2013) argues that general observable variables (geographic, demographics and socio economics) are easy to collect, but both their validity and reliability are questionable. Additionally, they tend to have low predictable capabilities in terms of consumer behaviour because people with same demographics can have very different preferences in terms of products (Haley 1968). Greenberg et al. (1989) also underline the fact that demographic variables are not generally useful for product development or marketing strategies. He stresses the importance of understanding the underlying psychological motivations. The same concept was noticed by car manufacturers that – in one particular instance – concluded that they have to market one model to *psychologically young* people, segment that actually included a large section of middle-age persons (Chon 2006).

### 2.3.3 Geodemographic Variables

While both geographic and demographical segmentation proved useful for various purposes, different studies identified that there will be added value if they are combined into a single criteria. This idea is based on the assumption that *birds of the same feather flock together*. In other words, people in the same neighbourhood tend to have similar houses, cars and probably read the same media (Mitchell 1995). This type of segmentation moved the emphasis from the consumer as the unit of analysis to the neighbourhood, which is considered to represent a homogenous blend of persons with similar characteristics (Wedel 2000). The first study in this direction was undertaken by Jonathan Robbin in the 1970s in the United States, by correlating

the census data with different surveys. The results of the study were turned into a commercial product called *PRIZM*<sup>2</sup>. In parallel, in the United Kingdom, a similar product – *ACORN*<sup>3</sup> – was developed based on Webber’s *Classification of Wards* Webber (1977), later on followed by Experian’s *Mosaic* (Figure 12).

The method used to obtain the groups for geodemographic segmentation requires the clustering of national census respondents based on their answers. The assumption is that the difference between members of a group needs to be smaller than the difference between the groups as a whole (Wedel 2000). While the number of clusters can vary depending on the product, most of the offerings provide two levels of segmentation with up to a total of 60 segments in the second level (Singleton et al. 2014; Wedel 2000). A comprehensive review of the similarities and differences between the products in this space is provided by Curry (1992).

The use of geodemographic data has proved successful in the industry by many case studies. Also from an academic point of view, the data scores high on the accessibility and identifiability segmentation criteria, as it is easy for marketers to understand and use it (Wedel 2000). The simplicity is facilitated by the fact that all the segments, being geographically defined, are linked to a list of postcodes or census areas. Moreover, the segments tend to have suggestive names, like *city sophisticated*, *career climber* or *lavish lifestyles*.

---

2 *PRIZM* – Potential Rating Index for ZIP Markets – Product offered by *Claritas*, later on acquired by *Neilson*

3 *ACORN* – A Classification of Residential Neighbourhoods – Product offered by *CACI*

Group	Description	% †	% ‡	Type	Description	% †	% ‡
A	Alpha Territory	4.28	3.54	A01	Global Power Brokers	0.32	0.30
				A02	Voices of Authority	1.45	1.18
				A03	Business Class	1.83	1.50
				A04	Serious Money	0.68	0.56
B	Professional Rewards	9.54	8.23	B05	Mid-Career Climbers	2.90	2.30
				B06	Yesterday's Captains	1.80	1.84
				B07	Distinctive Success	0.48	0.48
				B08	Dormitory Villagers	1.81	1.29
				B09	Escape to the Country	1.41	1.31
				B10	Parish Guardians	1.14	1.00
C	Rural Solitude	4.84	4.40	C11	Squires Among Locals	1.01	0.85
				C12	Country Loving Elders	1.32	1.31
				C13	Modern Agribusiness	1.61	1.36
				C14	Farming Today	0.53	0.53
				C15	Upland Struggle	0.36	0.34
D	SmallTown Diversity	9.21	8.75	D16	Side Street Singles	1.21	1.17
				D17	Jacks of AllTrades	2.60	1.99
				D18	Hardworking Families	2.87	2.63
				D19	Innate Conservatives	2.53	2.96
E	Active Retirement	3.41	4.34	E20	Golden Retirement	0.52	0.67
				E21	Bungalow Quietude	1.42	1.79
				E22	Beachcombers	0.57	0.60
				E23	Balcony Downsizers	0.90	1.29
F	Suburban Mindsets	13.16	11.18	F24	Garden Suburbia	2.82	2.14
				F25	Production Managers	2.31	2.63
				F26	Mid-Market Families	3.75	2.70
				F27	Shop Floor Affluence	2.82	2.73
				F28	Asian Attainment	1.45	0.98
G	Careers and Kids	5.34	5.78	G29	Footloose Managers	1.11	1.67
				G30	Soccer Dads and Mums	1.34	1.34
				G31	Domestic Comfort	1.24	1.09
				G32	Childcare Years	1.46	1.52
H	New Homemakers	3.99	5.91	H34	Buy-to-Let Territory	1.08	1.79
				H35	Brownfield Pioneers	1.13	1.38
				H36	Foot on the Ladder	1.48	2.37
				H37	First to Move In	0.30	0.37
I	Ex-Council Community	10.60	8.67	I38	Settled Ex-Tenants	2.08	2.06
				I39	Choice Right to Buy	1.90	1.72
				I40	Legacy of Labour	3.46	2.68
				I41	Stressed Borrowers	3.15	2.20
J	Claimant Cultures	4.52	5.16	J42	Worn-Out Workers	1.82	2.30
				J43	Streetwise Kids	0.90	1.05
				J44	New Parents in Need	1.80	1.80
K	Upper Floor Living	4.30	5.18	K45	Small Block Singles	1.26	1.77
				K46	Tenement Living	0.62	0.80
				K47	Deprived View	0.36	0.50
				K48	Multicultural Towers	1.09	1.11
				K49	Re-Housed Migrants	0.97	0.99
L	Elderly Needs	4.04	5.96	L50	Pensioners in Blocks	0.89	1.31
				L51	Sheltered Seniors	0.67	1.12
				L52	Meals on Wheels	0.51	0.86
				L53	Low Spending Elders	1.98	2.68
M	Industrial Heritage	7.39	7.40	M54	Clocking Off	2.18	2.25
				M55	Backyard Regeneration	2.40	2.06
				M56	Small Wage Owners	2.81	3.09
N	Terraced Melting Pot	6.54	7.02	N57	Back-to-Back Basics	2.50	1.97
				N58	Asian Identities	1.06	0.88
				N59	Low-Key Starters	1.60	2.72
				N60	Global Fusion	1.38	1.44
O	Liberal Opinions	8.84	8.48	O61	Convivial Homeowners	1.74	1.68
				O62	Crash Pad Professionals	1.41	1.09
				O63	Urban Cool	1.25	1.10
				O64	Bright Young Things	1.36	1.52
				O65	Anti-Materialists	1.12	1.03
				O66	University Fringe	1.10	0.93
				O67	Study Buddies	0.87	1.14

Figure 12: Experian’s *Mosaic* product defines 15 consumer segments and associated sub-segments based on the answers in the National Census Exercise with each sub-segment being linked to a list of postcodes – <http://goo.gl/5R4Vii>

The downsides of this method, in addition to the ones related to the demographics variables that also apply in this case, are linked to the methods of collecting the data. In the countries that undertake

national census exercises, the frequency of the studies is typically ten years. For this reason, it takes a long time for changes to be reflected in the segments. In addition, due to the rising costs, it is not yet clear if this kind of data collection exercises will be executed in the future, or at least not in the same form as today (Singleton et al. 2014). Some of the players in the industry already claim to rely less on the census data while using more of the open data made available by the local authorities. This is complemented by a rising number of proprietary data sources, like the price of houses sold or rent levels across the United Kingdom<sup>4</sup>. Another problem related to this type of segmentation is its reliability for companies operating in global markets, where segments obtained from certain key markets are not comparable to others.

#### 2.3.4 *Behavioural Variables*

Behavioural segmentation, as opposed to the previous categories of variables, is based on actual interactions between customers and products. Kotler et al. (2011) identify three subgroups of variables that belong to this group: needs and benefits, decision roles, and user and usage. Analysing the needs or benefits sought from the purchase of the product allows the marketers to better understand why certain individuals buy a given product. This is relevant because, in many instances, different customers purchase the same product for very different reasons. Being able to segment the market based on the underlying reason can lead to a better targeting of consumers. For example, a study that looked at the segments of wine drinkers in the United States identified six distinctive segments: enthusiast, image seeker, savvy shopper, traditionalist, satisfied sippers, and

---

<sup>4</sup> Information obtained from the company that develops the product – <http://google.com/L715RY> – Accessed: 15.02.2015

overwhelmed (Hussain et al. 2007). Similar segments were also identified in other markets including pharma, tourism, and fashion. In the user and usage group, the literature mentions a large number of variables that can potentially be used for segmenting the market: usage rates, brand loyalty, user status (potential, first-time buyer), readiness to buy, occasion or attitude.

There seems to be a broad consensus in the literature that behavioural variables are one of the most valuable sources of information in segmentation studies (Hiziroglu 2013; S.-Y. Kim et al. 2006; Wind 1981). Most of the problems related to it are derived from the data collection methods, usually based on surveys, but advances in the customer relationship management field identified solutions to overcome these problems. An overview of the technologies in this space will be presented in [Section 2.4](#).

#### 2.3.5 *Psychographic Variables*

There are different viewpoints in regards to the definition of psychographics. Kotler et al. (2011) define it as the science of using both psychology and demographics in order to explain the actions of consumers. Wedel (2000) sees it as the operationalisation of the concept of lifestyle that can measure a person in relation to different psychological dimensions like way of living, interests, or opinions on various topics. Psychographic segmentation provides a more granular segmentation when both geographic and demographics variables do not produce any actionable segments.

The initial studies into lifestyle choices back in the 1970s were implemented based on Likert statements. According to Plummer

ACTIVITIES	INTERESTS	OPINIONS	DEMOGRAPHICS
Work	Family	Themselves	Age
Hobbies	Home	Social issues	Education
Social events	Job	Politics	Income
Vacation	Community	Business	Occupation
Entertainment	Recreation	Economics	Family Size
Club membership	Fashion	Education	Dwelling
Community	Food	Products	Geography
Shopping	Media	Future	City Size
Sports	Achievements	Culture	Life cycle

Table 1: Life Style Dimensions; Source: (Plummer 1974)

(1974) most of these questions can be classified into three categories: activities, interests, and opinions (examples of these can be consulted in Table 1). Subsequently, other researchers argued that in order to model lifestyles typologies, values would represent a better choice since they are closely related to motivations as opposed to attitudes (Valette-Florence 1986). Rokeach (1973) defines a value as “enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or state of existence”.

In the literature, there are three main models that use values for psychographic segmentation:

1. The **Rokeach Value Survey (RVS)** identified a set of 18 terminal values (e.g. happiness, self-respect, social recognition, inner harmony) along with 18 instrumental values that can be used to achieve the terminal ones (e.g. polite, forgiving, obedient, loving). A respondent would be asked to rank the values in the order of importance (Rokeach 1973). Different respondents could then be clustered based on the similarity of their answers. The model has proved useful in explaining differences between

individuals that were part of the same group when a different segmentation variable was chosen.

2. The **List of Value Scale** proposed by Kahle (1983) contains a set of eight terminal value: self-respect, self-fulfilment, accomplishment, being well respected, fun and enjoyable, excitement, warm relationship with others, a sense of belonging, and security. The reduced number of values makes data collection easier in comparison to the RVS system, but at the same time offers less granularity.
3. The most used system for psychographic segmentation is the Strategic Business Insights' **VALS™ framework**. The model segments the population into eight groups based on their answers to 35 attitudinal questions and four demographic ones (Kotler et al. 2011). There are four higher resource groups (innovators, thinkers, achievers, and experiencers) and four lower resource groups (believers, strivers, makers, and survivors). When it was introduced in 1978, the model received much criticism from the academic world because of the secrecy around its methodology, making it impossible to validate. An overview of the framework can be consulted in [Figure 13](#).

As opposed to demographics, psychographic variables can provide much deeper insight into consumer behaviour. Early on, Plummer (1974) noticed that lifestyle segmentation – psychographic and demographic – generates much better segments than demographics alone. The conclusion is rooted in the fact that more information about the customers facilitates better communication strategies. Other studies also underline the importance of understanding *why* people buy certain products and their criteria for

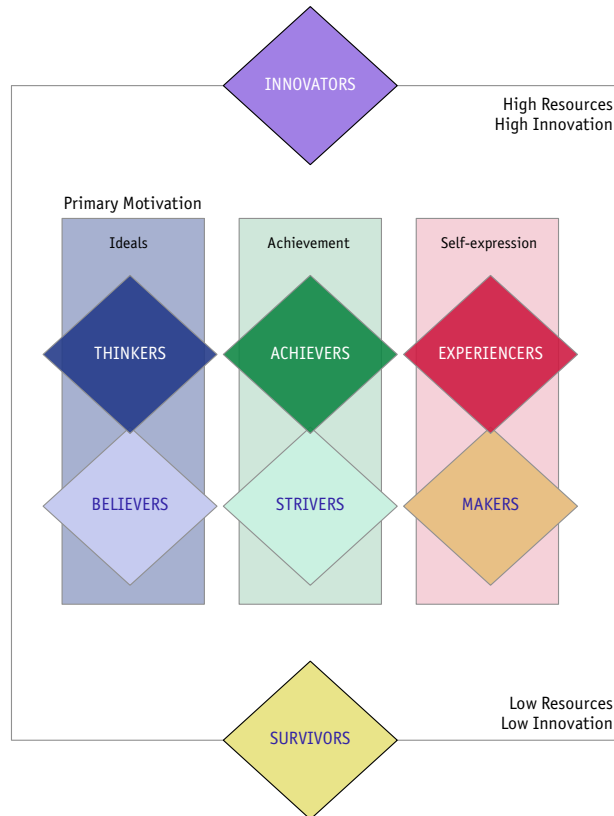


Figure 13: VALS Framework – Adapted from <http://goo.gl/HzzLlc>

evaluating the offers (Peltier et al. 1997; Lin et al. 2002). Wedel (2000) provides a comprehensive review of the psychographic variables, along with their main advantages. In line with the previous scholars, the positive aspects mentioned are: useful in understanding the underlying reasons for consumer behaviour, applicable to a broad range of products and services, can be tailored to specific domains and applications, and are easily implementable since they provide guidance for product development and advertising.

However, the literature also identifies a set of problems related to psychographic variables. These are linked to the methods of collecting and analysing the data. The studies are based on a random selection of people answering Likert-type statements about everyday living. Beane et al. (1987) argues that the way in which the questions are



formulated requires a profound knowledge of the market analysed. For this reason, the results can be heavily influenced by the selection of questions. Another issue, also related to the data collection, is the typical low response rates of such questionnaires. Peltier et al. (1997) mentions that respondents need to be incentivised in order to get their cooperation, adding an extra level of bias in the study. Moreover, because the number of questions in these studies exceeds three hundred, different dimensionality reductions techniques need to be employed for obtaining interpretable results, reducing the desired level of granularity (Wedel 2000).

#### 2.4 CUSTOMER RELATIONSHIP MANAGEMENT

In the recent years, the patterns of consumption became less predictable, as a result of a variety of consumer needs, interests, lifestyles and income levels (Miguéis et al. 2012). In order to accommodate for these changes, organisations needed better ways to segment their markets. While behavioural segmentation of markets proved successful previously, there were inherent difficulties derived from the fact that data needed to be collected via surveys. In an attempt to mitigate for these inaccuracies and also optimise other parts of their businesses, companies started implementing systems to help them track customer interactions.

Initially, these interactions were limited to purchases, but eventually developed to include also complaints, claims, click-stream data from the websites, or even customers movements in the shop. The resulting area – customer relationship management (CRM) – completely changed the way in which enterprises operate and is claimed to be one of the most successful business strategies in

the new millennium (Hwang et al. 2004). While there are many definitions of the term, one of the most used defines CRM as “managerial efforts to manage business interactions with customers by combining business processes and technologies that seek to understand a company’s customers” (J. Kim et al. 2003). The amount of data generated by these systems allowed companies to develop new segmentation models based on user interactions, completely changing the methods to segment a market. The process evolved from a simple statistical exercise on survey data that only covered a small sample of customers, to a more precise data strategy applied to all the interactions between the organisation and consumers (Hwang et al. 2004; Dibb et al. 1997).

#### 2.4.1 *Data mining in CRM*

Based on the existing data collected in the CRM systems, there is a wide variety of behavioural variables that can be measured and then used to segment consumers. Typical examples include purchase type, volume and history for different product categories, call centre interactions, claims, or click streams (Lee et al. 2005). However, since traditionally marketers considered that 80% of the profits are produced by the top 20% profitable customers (Duboff 1992), an application on the Pareto Principle, the most used metrics for behavioural segmentation are derived from customer profitability. The same concept can also be found under the name of lifetime value, customer lifetime value or customer equity. While there are small variations, in most instances the term is defined as the total revenue per customer from which the cost of attracting, selling and servicing is deducted (Jain et al. 2002). Hsieh (2004) mentions that many organisations provide offers and incentives only to customers that

meet a certain profitability criteria. Building on this, S.-Y. Kim et al. (2006) proposed an alternative model, which takes into consideration customer present value, but adds two new perspectives: potential value and loyalty. Based on a case study of the telecommunication market, this model is said to improve cross-selling opportunities while it also helps reduce churn.

In terms of the technology being used, database marketing techniques evolved from models based on recency, frequency and monetary value to more refined statistical techniques like logistic regression (McCarty et al. 2007). Moreover, the large processing power available today also made possible the use of different machine learning algorithms like support vector machines, neural networks or random forests (Baecke et al. 2009; Hyunjung Shin et al. 2006). These techniques, and particularly neural networks, were proved to perform better than the statistical methods in different case-studies. Also, since organisations were able to process more data in the same amount of time, they also started acquiring external databases with transactional data in order to derive deeper insights about their customers (Bult et al. 1995). Another technique frequently used is augmenting internal data with large amount of geographic and demographic data obtained from surveys (Baecke et al. 2009).

#### 2.4.2 *Hyper-targeting*

Another significant aspect to consider in a market or customer segmentation study is the desired size of the segments. Depending on intended users of a product, the potential segments can vary from very broad to extremely narrow. There are four possible levels at which a market can be targeted: mass market, multiple segments, a

single segment, or individuals (Kotler et al. 2011). The choice of the targeting strategy is strongly connected to the segmentation variables, which in turn are related to the marketing vehicles. Sometimes it is efficient to market a product to a very niche market segment. In that case, marketing needs to have the mechanism to reach those individuals at reasonable cost. However, in the recent years the market experienced some changes at both levels: demand for highly specific products (long-tail phenomenon) and better targeting mechanisms (hyper-targeting).

Traditionally, when analysing the distribution of sales in terms of income generated for a company, it was noticed that roughly 80% of the income is produced by 20% of the products (Duboff 1992). However, following the adoption of the Internet as a shopping medium, there was a perceived shift in consumer behaviour. An increasing number of products that were not *hits*, but still very popular with niche segments, started to generate more income. According to Anderson (2008), the distribution of sales is now more balanced, summing up to approximately 50% – 50%. The shift was pioneered by companies like Amazon, eBay, Netflix or iTunes, and proved to be more popular in areas where customers' tastes and interests have a massive impact. Examples of such markets include but are not limited to TV programmes, songs, books, or clothing.

A few decades ago, reflecting on the practicalities of segmentation studies, Dibb et al. (1997) were arguing that no segmentation study would be a success if it is not relying on demographics, so that marketers can identify the target groups. However, in recent years, the tools available for targeting changed dramatically, allowing for much more refined algorithms.

*Hyper-targeting* is the capability of social networking sites to target users with advertisement based on very specific criteria (Shih 2010, pp. 211), which in addition to the usual geographic and demographic criteria also include users' interests. The term was first coined by MySpace back in 2007 when they launched their advertisement solution (myAds). By 2009, the concept was widely accepted in the industry. Nowadays, most large scale web platforms allow interest based targeting (Google, Facebook, Twitter, LinkedIn). While dealing with proprietary solutions, there is not much information in the literature detailing the algorithms behind hyper-targeting. However, according to Gold (2009), based on information derived from the network's privacy policies, the data is inferred from registration information, profile information, and behavioural data.

#### 2.4.3 *Identifying the Right Future Customers*

While the advantages of using CRM technologies for retaining customers and reducing churn are widely acknowledged, some scholars argued that the same technology could be used for identifying the *right* future customers, based on the needs and behaviours observed at existing customers (Payne et al. 2005). Starting from the transactional data available in their systems, organisations typically segment consumers on the basis of behavioural variables. However, these segments are not actionable from a marketing standpoint because the criteria on which they are obtained is only relevant to existing customers as opposed to future ones (e.g. frequency of use, lifetime equity, number of complaints, interactions with products on the website, etc). Nevertheless, after successfully identifying the segment containing the most profitable existing customers, companies can infer their geographic, demographic or

geodemographic traits, in order to provide input for marketing teams and product development. As an example, one organisation can first segment the existing customers based on profitability, and then infer that the majority of these belong to the *city sophisticates* cluster as defined in the ACORN product. Having access to this information, including a list of postcodes with the highest concentration for a particular cluster, a marketing team can more precisely target potential customers with advertising campaigns.

In the instance when geographical data is not already available in the CRM systems, companies can infer it based on the location of shopping till, in the case of brick and mortar stores, or, in the online environment, by using designated libraries that can geolocate the user's IP address (e.g. Maxmind GeoIP). Similarly, in the case when demographic traits for a customer are not already available from their online presence or loyalty system membership, these can be inferred based on the postal address. In the United Kingdom the data is available with high granularity, different demographical traits being presented for each Lower Layer Super Output Area<sup>5</sup> along with their overall ranking, therefore facilitating demographic segmentation. In addition, the geodemographical cluster distribution for each postcode is also available from commercial products offered by Nielsen, CACI, or Experian.

Unfortunately, even with the massive data mining processes in place, many studies pointed out that CRM technology consistently failed to deliver, especially in terms of identifying future customers (Bailey et al. 2009; Rigby et al. 2002; Boulding et al. 2005). This can

---

<sup>5</sup> Lower Layer Super Output Areas (LSOA) are geographical boundaries designed to improve the reporting of census data for small-areas. In England there 32844 such areas, and each has a population between 1,000 and 3,000 inhabitants – Information retrieved from the Office for National Statistics – <http://goo.gl/mS1RJg>

be rooted in the drawbacks of using geographic and demographic variables alone, since these do not always correlate with consumer behaviour patterns. While being able to infer psychographic traits of consumers, like interest and values, based on transactional data could help companies to better understand their customers, the existing body of knowledge in the area does not yet provide us with a methodology to do so. After examining all the articles provided by Web of Science and Google Scholar that scored high on the relevance factor for terms like *Customer Segmentation*, *Psychographic Segmentation*, *Interest Segmentation*, *Lifestyle segmentation* and *Interest Detection*, it was noticed that the vast majority of these are from the social sciences area, and only discuss psychographic segmentation in the context of surveys. The same conclusion was reached by Hiziroglu (2013) that reviewed the state of the art in terms of data mining technologies used for customer segmentation and noticed that only two papers (2.4% of the articles analysed) focus on the general unobservable variables (e.g. interest, values), while the majority are dealing with general observable (geographic, demographic) or product related variables (brand affinity, product preference).

From the papers that discussed psychographic segmentation in the context of data mining, Miguéis et al. (2012) describes a method to segment customers by lifestyle, starting from transactional data. The case study analyses loyalty card data for approximately two million customers of a large European retailer. After clustering the customers based on the similarity of their shopping baskets, the content of the baskets is analysed based on the category of the product (e.g. cheese on the counter, spirit drinks, cereals, soups, body hygiene) and the brand position (e.g. premium, leader, secondary, own-brand, economic). Based on this information, the author infers

the lifestyle of each cluster, for example “the potential buyers of these products seem to have a high economic power, [...] they may have babies, and appreciate practical meal solutions, such as cod-fish meals”. While this methodology provides a practical way to infer some psychographic traits of consumers, it cannot be fully automated as it relies on human assessment of the baskets’ content. Moreover, the method relies on a classification of products that assumes that each product can only belong to one category, and the categories are grouped into business units (e.g. *Beers* and *Current Wines* belong to the *Drinks* business unit). This can limit the algorithm’s ability to understand the consumers, because similarities between products can still exist across the existing categories’ boundaries, as it is in the case of a specific diet, low or high caloric content, various nutrients, festive food, etc.

The same aspect is emphasised by Hsu et al. (2012) that proposes an alternative segmentation model that focuses on the similarity between items based on a hierarchy of concepts. Two items are considered similar based on the length of the path that connects them in the hierarchical tree. For example, the distance between *milk chocolate* and *chocolate* is one, since *milk chocolate* is a subclass of *chocolate*, while the distance between *yoghurt* and *sour cream* is two, since both are subclasses of *milk-based products*. In order to deal with situations where different organisations do not use the same concept hierarchy, as it is usually the case, Hsu et al. (2012) proposes the use of a word semantic similarity measure. While this method is a clear improvement from the one described by Miguéis et al. (2012), it still assumes that every product belongs to one category. Also, while using text semantic similarity can improve the results in some cases, it can also be misleading in others. There are multiple scenarios where



the names of the items have no textual similarity, but are similar in terms of a given feature which might determine consumers with a certain interest to purchase them. For example *avocados* and *raspberries* have no textual similarity, but they both have high fibre content, a potential reason for health conscious buyers to be interested in both.

Not being able to fully automate the process of inferring consumers interests with the help of the transactional data from CRM systems limits organisations in exploiting the full value of the information they collect about their customers. This is particularly relevant in markets where psychographic variables like interests and values play an important role, especially in the context in which the large players in the online advertisement space already support hyper-targeting based on interest. An overview of the overall process described, as well as the knowledge gap in the area of inferring psychographic traits based on transactional data (marked with red background), is provided in [Figure 14](#).

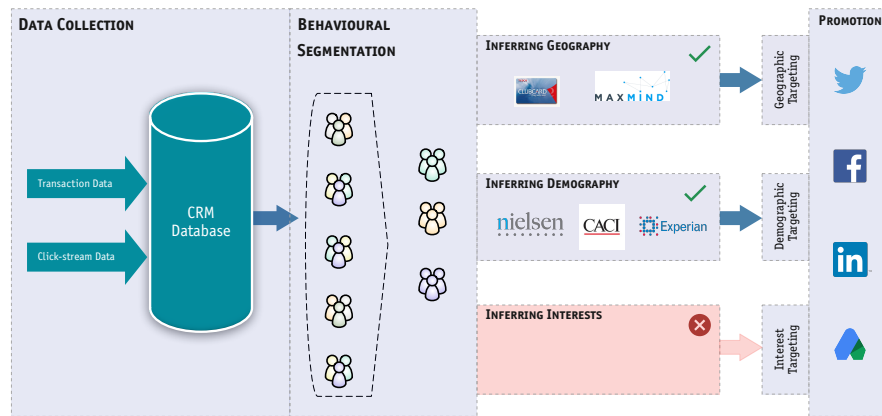


Figure 14: Overview of the typical customer segmentation process based on data mining in the CRM and the identified knowledge gap

## 2.5 CONCLUSIONS

In this chapter an overview of the insights from the customer segmentation discipline was provided. A list of criteria that is relevant in terms of market's segmentability was presented, in order to be able to assess the degree of relevance of different segmentation variables. The most common five categories of segmentation variables (geographic, demographic, geodemographic, behavioural and psychographic) were described and analysed from the perspective of the problem area. While all have advantages and disadvantages, based on the given purpose, it was concluded that psychographic variables are the most effective, but the methods of collecting them, involving complicated surveys on a small sample of subjects, clearly impede their use. The advances in the field of customer relationship management were presented, as they are relevant in understanding which customers are profitable and which are not, but do not provide a way to feed the information back to marketing. While most profitable groups can be described by behavioural or geodemographic features, a knowledge gap was identified in the area of inferring psychographics traits based on the customers' interaction with products. In the next chapter the insights from the media research discipline will be presented.

## THE MEDIA MARKET

---

The previous chapter presented the insights from the customer segmentation discipline, and also identified a knowledge gap in terms of inferring users' interests based on transactional information. While this ability might be useful for various markets, allowing for a better segmentation of customers, there are some markets that could benefit more than others. Intuitively, someone's personal interests and opinions might have a limited role in choosing a banking product or a house appliance. This is rooted in the fact that both items tend to be bought for a particular reason, typically in reaction to a basic need. At the same time, the same person's interests might be highly relevant for deciding which television programme to watch or publication to read. This hypothesis is validated by the existing body of knowledge related to the process of choosing a programme to watch.

Taneja et al. (2012) describes two main theories of media consumption. The first one is based on how different structures affect viewers choices. Examples of structures are the list of available channels, the scheduling of programmes, or the possibility to watch on various devices. The second theory is focused on the influence of psychological factors, like needs and preferences, in choosing a particular programme type or genre (Owen et al. 1992). With more and more content being available for consumption on various devices, the influence of the structural aspects is decreasing at the expense of the psychological ones. Viewers do not have to adjust their schedules in order to view a favourite programme, but can instead pick what

they want to watch based on their interests, and decide on the right timing to do so.

In this context, the media market was selected for the empirical validation of this research, as the connection between an individual's interests and the decision to purchase an item is deemed stronger than in other areas. In the following sections, an overview of the methods used for understanding audiences will be presented, in order to understand how each is relevant for the purpose of this study. Also, the algorithms used for content recommendations in the media space will be briefly reviewed, as some of the work in this area has applicability for psychographic segmentation. Finally, based on the research gap identified in the previous chapter, taking into consideration the specificities of the media market, a set of research questions will be defined.

### 3.1 UNDERSTANDING AUDIENCES

There are three main sources of data that have been used for measuring and understanding audiences in the media market: TV ratings, social media streams, and video-on-demand platforms. The next subsections will present the relevant insights from each of these.

#### 3.1.1 *Television Ratings*

Traditionally, television ratings or public audience shares, represented the primary source of information for understanding audiences. These studies estimate how many people are in the audience at a given time, segmenting them by the channel or programme they are watching, location, and demographics. Given

that advertising prices are negotiated based on these numbers, the television ratings are highly regulated and have not embraced many changes in the methodology over the last seventy years. The evolution of the data collection methods over time will be first presented, followed by an outline of their relevance in the context of inferring viewers' interests.

#### 3.1.1.1 *Phone Surveys*

Since the very beginning of radio and television broadcasting, various solutions were used in order to quantify and understand what is the audience interested in. The most basic one, initially used for measuring the audience of radio shows, was based on a phone survey that questioned listeners about the programmes they listened to the day before. The methodology was later improved by Clark Hooper, who tried to eliminate the bias associated with remembering things from the previous day, by questioning only what listeners are tuned to at the moment of the call, as well as the demographic data for the persons listening (Webster et al. 2013). This method, also known as *telephone coincidentals*, constituted the foundation for the standard measurements used in the industry till the present day (Hill 2014). In the early 1950s, Hooper's company was acquired by Nielsen, which used the methodology for developing the *Nielsen Television Index* (Webster et al. 2013).

#### 3.1.1.2 *The audimeter*

Given the high cost and subjectiveness of phone surveys, there was considerable effort focused on the development of a device that could track what the viewers are watching. The first patent for such a device – the audimeter – was filed by Claude E. Robinson, a Columbia University student, in 1939. In the next few years, the concept

was implemented by two MIT professors, and later on acquired by Nielsen in order to be used for their audience measurements service (Webster et al. 2013). The device could record the frequency or channel that a radio or television set is tuned to and for how long. Initially, the data needed to be collected by Nielsen fieldworkers every week, but subsequent versions of the audimeter stored it on cassettes that were mailable, and later on transmitted it to a central server using a separate phone line (Bjur 2009). One of the main problems with the use of audimeters was the lack of demographic data, since the device could only record what is being viewed, but not who is viewing. In order to tackle this problem, Nielsen asked a subset of the sample population that had audimeters installed to keep viewing diaries, and merged it with the audimeter data in order to predict the size and demographics of the audiences (Hill 2014). The system dominated the market from the 1950s for almost thirty years as it was deemed *good enough* in terms of stability and reliability to constitute the currency for the advertising price (Bjur 2009).

#### 3.1.1.3 *The people meter*

The expansion of the cable television in the United States in the 1980s provided the required capacity for more television networks to broadcast their content. This led to a sudden increase in the number of stations to hundreds, as well as the number of programmes to thousands. Since the existing audimeter and diaries methodology required people to remember what they watched, it was more favourable for popular content, that people were prone to remember. The *Cable Audience Methodological Study*, commissioned by the cable industry to investigate the magnitude of the difference, showed that the existing methods underestimated the numbers of viewers for cable networks by up to 36% (Adams 1994; Bjur 2009).

This eventually led to the introduction of a new device, meant to reduce the contribution of the human factor as much as possible in the process of measuring audiences. The *people meter*, first introduced by *Audits of Great Britain* (AGB), consisted of a piece of equipment that could detect the channel on which a TV is being tuned to, similar to the audimeter. However, in order to replace the need for viewing diaries, the people meter also included a remote control that allowed viewers to signal their presence in front of the screen, as well as some basic demographic traits like age and gender (Hill 2014). Till the present day, people meters are still the de facto standard in the industry, not changing much from the initial model proposed in 1980s. The only adjustments are related to the integration with existing models of set-top boxes<sup>1</sup>. In this way, the people meters do not only track live broadcasting, but also video-on-demand content or shows recorded for viewing at a later moment – an activity also called *time-shifting*.

While television ratings are still the main source of data in the media market, with most of the advertising prices depending on it, they also have a number of shortcomings. Most importantly, they are based on a sample of viewers which is typically relatively small. For example, in the United Kingdom, the number of houses that have a people meter installed is around 5,000 out of the total of 25 million. This leads to very high differences in the audience data between different ratings providers, or whenever the composition of the sample group needs to be changed in order to reflect new realities (e.g. ageing population, immigration, new channels, etc).

---

<sup>1</sup> *Set-top Box* – “Generic name for an electronic interface between a cable television or satellite signal and video display and recording devices. Typically a box that can be placed atop the television set (hence the name), it can have many functions, including acting as a tuner, decoding digital or analog television signals, removing encryption, and allowing the purchase of pay-per-view channels” – Definition from <http://goo.gl/U91Cj7>



Figure 15: Nielsen's *Audience Measurement Process* uses the peplemeter device in order to track what the sample panel is watching; it extrapolates the number to the whole population; Source: <http://goo.gl/fuiWCI>

Typically, there is only one provider of the ratings in each country (e.g. BARB in the United Kingdom, Nielsen in the United States), commissioned to provide the ratings for a given number of years. Whenever the provider has changed, substantial differences in the ratings were noticed, underlining the fragility of the methodology (Taneja 2013). There were also situations when these discrepancies led to the temporary suspension of the ratings for a number of weeks, while the methodologies were being reviewed (Bourdon et al. 2014). Another aspect to consider in regards to the relevance of this data source is that the information is not freely available. Due to the complex and costly process of installing people meters in households, as well as the data processing effort, broadcasters need to pay yearly fees for their audiences to be tracked. Generally, only the large broadcasters afford this service, limiting access to the data for smaller and niche television channels.



While television ratings can provide a good index of popularity for television programmes, segmented by location and demographics, they do not constitute a good data source for inferring viewers interests. This is rooted in the fact that the numbers from the sample are extrapolated to the whole population based on the viewers' demographics. This is in contradiction with the conclusion emphasised in the previous chapter, where it was shown that the demographics have a low impact in predicting consumer behaviour. For example, if the majority of young viewers from affluent households in the sample population watched *House of Cards*, the numbers provided by the TV ratings will be based on the assumption that the same percentage of the whole young and affluent population watched the show – which might not be true, since television watching is more related to people's interests than their demographics. In order to be useful for interest segmentation, a data source needs to provide access to individual interactions with the content, without extrapolating from a sample to the whole population, so that interest-affinity profiles can be built for each viewer.

### 3.1.2 *Social Media*

Even if the methodologies and sample sizes are constantly debated, television ratings are still accepted across the market as the de facto audience measurement tool. However, compiling the data takes time, typically in the range of a few days, so broadcasters cannot react fast enough to the audience's feedback. Moreover, because of the collection methods, the data provides only quantitative measurements. In order to better understand the viewers feedback and be able to react to it fast enough, the media market turned its

attention to the use of social media as a data source for understanding audiences.

Starting in the early 2000s, the large number of users present on social networks generated a wealth of data about their preferences and interests. In the United States alone, 71% of adults use Facebook, while 42% use at least one other social network (e.g. Twitter, Instagram, Pinterest, LinkedIn) (Duggan et al. 2014). While initially the expansion of the Internet, and social networks use in particular, was considered to be a threat for broadcasters, it is now considered to be one of the main drivers for TV consumption. This is rooted in the fact that while people are watching various programmes, they tend to comment online on social networks, a phenomenon also called *second screening* (Proulx et al. 2012). All the activity generated on the networks eventually attracts more viewers for the TV programmes, acting as an uncontrolled marketing channel. In support of this claim, Nielsen estimated in 2008 that approximately a third of the Internet traffic in the United States occurs while people are watching TV<sup>2</sup>. In order to gather more information about the viewers, Sutter et al. (2011) proposed a technical system in which the video signal could be watermarked with special signals that could be recorded and sent back by second screen applications, helping identify who's watching what. This system, designed to replace the existing audimeter, seems relatively easy to implement but privacy concerns would probably deter its use by consumers.

The wealth of data produced by TV viewers online generated a lot of interest both in academia and industry. Many broadcasters established a social media presence for some of their shows, and

---

<sup>2</sup> Nielsen Reports – TV Viewing and Internet Use Are Complementary – <http://goo.gl/NZJRV3>

tried to make use of the data to improve their services. Cheng et al. (2013) developed a model to predict television audience ratings solely based on Facebook data. The model made use of artificial neural networks for estimating the number of viewers for different episodes of popular shows by factoring in number of posts, comments, likes and shares. The model proved to have good to high accuracy, showing strong correlation between the number of viewers of a show and the volume of their activity online. Alternatively, Wakamiya et al. (2011) developed a model to measure the relevancy between tweets and TV programmes popularity. While in the case of Facebook the comments were posted on the official page for the show, and therefore explicitly attributed to a programme, in the case of Twitter this problem needed to be tackled in a different way. The authors developed a methodology to attribute a tweet to a given programme based on a mix between textual relevance (the content of the tweet needs to contain references to the name of the programme or official hash tag), spatial relevance (only activity from certain geographies needs to be tracked for given shows), and temporal relevance (only track tweets in a timespan close to the time of airing a show). While this model performed relatively well on a case study for Japanese broadcasters, the process of assigning tweets to TV programmes is not an exact science, as there is no explicit way for users to refer to a specific show. For example, using the hashtag *#chelsea* has a high probability of referring to the *Chelsea Football Club* in the UK, but also to the neighbourhood or individuals sharing the same name. In order to disambiguate the meaning, algorithms need to understand the context of the message, a process that is relatively hard given the size of a tweet being up to a maximum of 140 characters.

Also in the industry, many companies tried to capitalise on the information available from the social networks and use it in the media space. Most notably, *Bluefin Labs* in the United States and *SecondSync* in the United Kingdom, developed algorithms for correlating social media content with TV programmes (Figure 16). Twitter, one of the main sources of information in this space, noticed the potential of this area and ultimately acquired both companies in 2013 and 2014.

## Olympics 2012 - Opening Ceremony

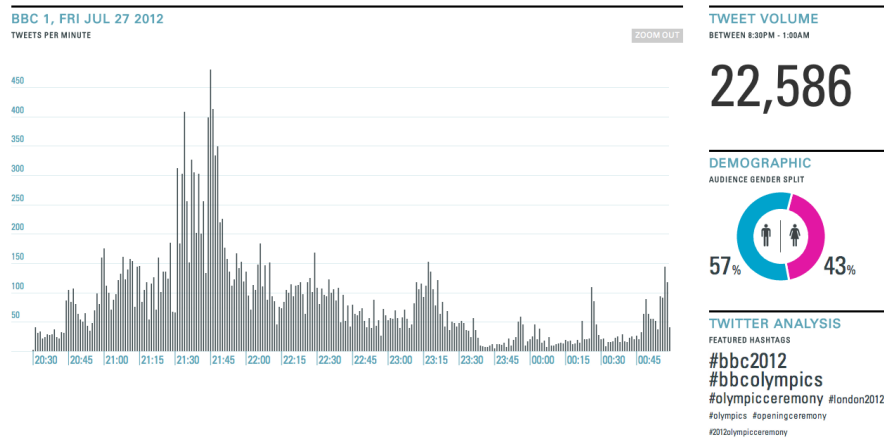


Figure 16: *SecondSync* correlates social media activity to TV programmes, showing the number of tweets as well as the gender distribution and most used hash tags; Source: <http://goo.gl/i5930x>

From the perspective of the research gap, social media does provide a massive amount of data that can help understand audience interests. Moreover, having access to the users' opinions about a certain show is powerful, as it creates a fast feedback loop from the consumer back to the broadcasters. Also, on some social networks most of the content published by users is public, and can be processed by various data mining tools. Taking the example of Twitter, one can not only reason about one user's reaction to a given show, but also scan the whole history of messages that he or she posted, while trying to infer his interests. However, there are also shortcomings for the use of social media. While many studies show correlation between

social data and audience size, in most instances the algorithms were tested on popular shows that attract a lot of attention. This tends to work better for content that is generally discussed on social media – like reality shows or popular series – but might not work for other types of content (e.g. documentaries, travel, culture, etc). More over, inferring interests solely based on social activity could introduce a bias due to the average age of social networks users being lower than the average age of television audiences as a whole.

### 3.1.3 *Video-on-Demand*

Over the last decade, the media scene changed fundamentally. Fuelled by the overall increase of broadband and mobile connection speeds, as well as the proliferation of mobile devices, viewers can consume video content at any time, from anywhere, instead of being restricted to what is broadcasted on TV. This model was pioneered by companies like Netflix, Youtube, Hulu and Amazon, that created platforms that connect content creators with viewers, shortcutting the need for television channels in between. The success of these platforms is backed by numbers that show that the video-on-demand (VOD) market is catching up with TV viewership in terms of volume (Hill 2014). Youtube, a video sharing website owned by Google, reports having one billion unique users that consume approximately six billion videos every month (Winslow 2014). In response to this threat, traditional broadcasters responded with an initiative called *TV Everywhere*, or authenticated video-on-demand. This made it possible for their subscribers to stream content to various devices over the internet. The service was first proposed by Time Warner Cable in 2009, but was rolled out soon after by many other important players in the industry.

These changes in the market triggered by the advent of video-on-demand services also affected the way in which data about viewers is being collected. In a traditional broadcasting context, the content is delivered to the customer over a one-way connection, while the networks do not have any control over how it is consumed and by whom. This is why, as presented in the previous section, estimating the audience's size required special devices to be placed in a number of households. However, in the case of video-on-demand, the content is delivered over the internet. Therefore, the networks can track who is watching what, as they control the software used to stream the content (web players, mobile applications, Smart TV applications, etc). As opposed to the TV ratings, the data can be collected for all viewers, not only a sample, making the statistics much more accurate. While there are still a large number of users using traditional TV services, given the recent increasing trend of online viewing and the expansion of Smart TVs, it is fair to assume that in the next few years most of the content will be streamed via the Internet. From the perspective of viewership segmentation, the data collected by video-on-demand platforms is deemed the most useful of all the sources analysed. The main reason for this is rooted in the fact that the data collected is not sampled, it shows which users are watching which shows, and provides a whole history of their viewing habits. An overview of the timeline of audience measurement events can be consulted in [Figure 17](#).

### 3.2 AUDIENCE SEGMENTATION

The wealth of data being generated by the video-on-demand platforms opened up many opportunities for analysing audience behaviour. If before broadcasters had to rely on statistics about

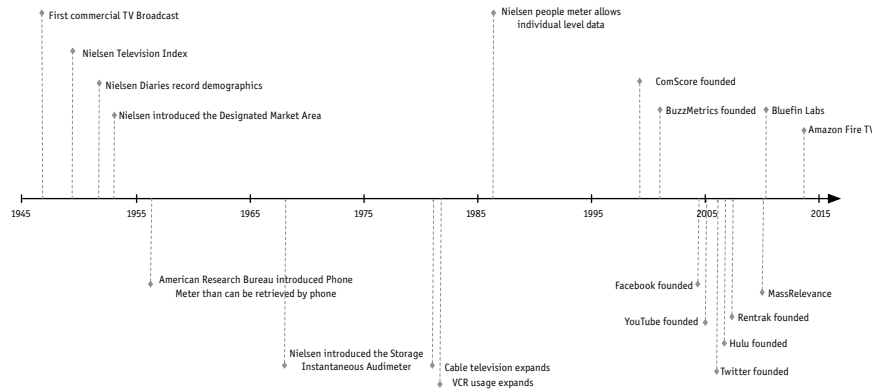


Figure 17: TV Audience Measurement Timeline – Adapted from Hill (2014)

the performance of their programmes based on a small sample of viewers, like the ones offered by TV ratings companies, they can now keep track of every user's actions. This ability allows them to implement programmes similar to the CRM packages used in other industries, most notably retail. The data can be used to improve audience measurement, customer segmentation, and the accuracy of recommender systems.

The following list presents the type of segmentation variables that can be used, based on the categories defined in the previous chapter:

#### 1. GEOGRAPHICAL SEGMENTATION

Can be performed based on the information provided by the user, in the case of video-on-demand services that require viewers to sign-up and authenticate. For services that allow anonymous usage (e.g. BBC's iPlayer, YouTube) the location of the viewer can be inferred based on his IP address. There are a range of companies offering this services, most notably *MaxMind GeoIP*, and the reported accuracy for detecting the country is 90%, while for identifying the correct city is around 60%;

## 2. DEMOGRAPHICAL SEGMENTATION

Given that the geographical location of the viewer can be identified (either by provided information or inferred from the IP address), some demographical traits of the user can typically be inferred based on the national census information. In the case of the United Kingdom for example, this data is provided by the government for each Lower Super Output Area (LSOA), with different features like crime, education, employment, environment, health, housing, income and wellbeing. For example, knowing the postcode of a viewer, based on the national census data and proprietary databases, one can predict the person's income level, religion, ethnicity, education level, or property price;

## 3. BEHAVIOURAL SEGMENTATION

Having access to the viewing history of all users, it is easy to segment them by different behavioural characteristics. Some of the typical measurements include total time watched, average time watched, day part analysis, frequency of views, top programmes watched. Based on the broadcasters specific needs, an ideal viewer could potentially be identified by various variables like his engagement with videos, number of advertisements viewed, click-rate on advertisements, or total time watched;

## 4. PSYCHOGRAPHIC SEGMENTATION

Using social media streams, organisations can analyse the profiles of the viewers that post opinions about various shows online. For example, based on the Twitter profile of someone tweeting about a *Sherlock* episode, analysis can show that the user has a certain profession, or is interested in a specific brand. However, as mentioned before, the use of social media can lack



the desired level of accuracy, as messages cannot be explicitly assigned to a programme, and moreover, the analysis is only possible for programmes that generated a great amount of social media activity, and therefore might not be feasible for all situations.

### 3.3 INTERESTS AND RECOMMENDATIONS

With more content made available on various platforms, viewers face new challenges in being able to find things they are interested in. In order to tackle this issue, companies developed sophisticated recommendation engines for suggesting which content to watch. In the general case, these systems try to predict how likely is for a given user to be interested in an item. In the context of the media industry, they predict the *rating* or preference a viewer would have for a given programme. Based on the approach taken to generate these predictions, there are three main categories of recommender systems: content-based, collaborative filtering, and hybrid (Adomavicius et al. 2007). The content based systems analyse various features of programmes and make predictions based on what the user has viewed before. For example, if a given user watched three movies directed by the same director, a potential recommendation would be another movie from the same director. In a similar way, actors, programme genres, or locations can be factored in the algorithms. By contrast, a collaborative filtering system would generate recommendations based on what other viewers that are considered similar watched before. For example, if two users have a relatively similar viewing history, programmes that have been watched or rated high by one of them could be recommended to the other and vice-versa. Finally, hybrid systems combine both

content-based and collaborative filtering, in an attempt to take advantage of the benefits from both methods.

In one of the most relevant papers in the field from the perspective of the media market, E. Kim et al. (2011) proposed a recommendation engine for TV programmes based on collaborative filtering. While some platforms allow users to rate the content they watched, using that as an explicit rating for recommendation, it is not the general case. Therefore, the proposed model implicitly infers the viewers ratings based on how long they watched a given programme (the longer the viewing time, the higher the rating). Given the high number of programmes available on video on demand platforms, often up to tens of thousands, the chances of two users seeing the same programme is relatively low. Therefore, collaborative filtering systems tend to recommend only a small subset of the available content, deemed more popular, while ignoring the majority of the titles. In order to cater for this scenario, E. Kim et al. (2011) factored in the viewers preference for a certain channel or genre, instead of programme alone. For the experiment they used a classification of eight genres (sports, news, information, drama & movie, child, education, amusement, others). Soares et al. (2014) proposed an alternative system relying on hybrid recommendation. The collaborative filtering part is based on the similarity between users computed as cosine similarity of their ratings for watched programmes. In addition to that, the system computed the degree of similarity between two programmes as the degree of overlapping between genres, actors and directors.

One apparent problem with the previously described systems is that they both rely on a very limited number of programme

genres, which can also be different for various broadcasters, or simply users' perception. For example the movie *Eternal Sunshine of the Spotless Mind* is rated by reviewers on the ratings website IMDB consistently in the top 50 for three different genres: drama, romance, and Sci-Fi. In order to bring more consistency to the genres, as well as eliminate some of the ambiguity, Hyoseop Shin et al. (2009) proposed a different hybrid recommender system that relies on the TV-Anytime genre taxonomy. In this case, the similarity between two different programmes is calculated using a new metric called *Descendant Nesting Similarity Measure* (DNSM). This factors in the common ancestor's depth, as well as the level of node nesting. In addition to that, given the relative nature of genre classifications, a keyword based content rating method is proposed. This is computing the similarity between two content types by using the *inverse document frequency* formula for the words used in each programme's description. Finally, collaborative based algorithms are used in addition to the content-based ones (both DNSM and keyword based) in order to generate recommendations. The results show a 83.5% accuracy on the training data. Without relying on viewing data at all, Peleja et al. (2013) proposed a system that can produce recommendations solely by analysing ratings and users comments on social platforms. By using sentiment analysis on the users comments, the algorithm, trained on over 50,000 reviews on IMDB and more than 600,000 comments on Amazon entertainment media, reported good accuracy compared to other alternatives which rely on explicit ratings and the associated biases.

While recommendation engines have an important role in the media industry, facilitating access to the vast library of content, they have a relatively limited use for audience segmentation. This is rooted

in the fact that they do not explain the reasons for which a large group of users would enjoy a certain programme. Without understanding the underlying reasons for user's preferences, content developers and marketing teams have insufficient information to work with. However, the work done in the area of content-based recommender systems is relevant for this study, as it is trying to explain the degree of similarity between two programmes – which is presumably related to the viewers' interests and topics of each show.

The existing research currently relies on comparing genres or keywords. However, genres are very subjective, and while there is an existing taxonomy available, it is not widely adopted. Moreover, the taxonomy lacks enough granularity, and requires each programme to fit into a given category (e.g. sport, football, etc). However, a programme can touch on various topics, and trying to fit into only one genre would severely affect the reliability of the calculation. For example, the *Top Gear* series could be of interest for automotive fans, but also for viewers interested in travel and foreign cultures. In addition to this, a study done by Japan's public broadcaster (NHK), identified that programme genres are not static in nature. The perception about the genre of a programme tends to change over the years, and what previously was classified into a given genre can be classified as something else in a few years time (Hara et al. 2004). A typical example is the *reality show* genre, that existed in different forms from the 1980s, but really became a genre of its own after the commercial success of *Big Brother* and *Survival* in the early 2000s. This means that costly surveys need to be done periodically in order to ensure that the taxonomy is right. This also has implications with local cultures, making one general taxonomy that would work for all countries and cultures highly unlikely.

Similarly, the keyword based approach, while arguably better than the genre based one, adds a lot of subjectivity based on the wording of a programme description. For example, a tennis match could be described in one instance as the *Australian Open*, and in another one as *Wimbledon Finals*. While the degree of similarity between the two should be considered high, a keyword-based system would not be able to find any similarity points. Therefore, in this study, the various methods that can be used to identify the topics that are touched upon by a programme will be analysed, since any topic can be considered to be a viewer's interest. Moreover, it could also be relevant to vary the granularity level of the topics. For example, one could infer that a show about *Wimbledon Finals* is about *tennis*, which is also about *sports*, *sports with a racquet* and *outdoor activities*. This information can then be used for segmenting users based on their interests, and also understand which programmes gather most attention based on the topics discussed.

### 3.4 RESEARCH QUESTIONS

In the previous chapter a research gap was identified in terms of inferring users interests based on their interaction with products. This capability would allow companies to segment their customers by psychographic traits, which are more relevant for understanding consumer behaviour compared to geographic and demographic variables. Based on the specificities of the media market, two research questions were identified that would address the knowledge gap. First, the programme metadata needs to be analysed for each programme in order to detect which interests / topics are present. Ideally, these could also be ranked by their perceived relevance for each programme (RQ1). Having this information available, the

viewing data for a given channel and video on demand service could be analysed. This would allow broadcasters to understand what the interest affinities are for each user, and then segment the entire user base by their perceived interests (RQ2). However, being able to better understand the relationship between viewers, programmes, and interests has implications not only for segmentation, but also for content development. Having analysed the set of interests / topics for each programme, along with the number of viewers that watched each show, could be used for better understanding what attracts more viewers in terms of content (RQ3). Also, information extracted from the social media and search queries could help broadcasters visualise which topics / interests are trending, and schedule content accordingly, or develop new programmes in order to capitalise on that trend (RQ4). The two main areas of interest along with the related research questions are listed below:

#### INTEREST SEGMENTATION

RQ1 How can the list of topics / interests for a given programme, along with their perceived relevance, be identified?

RQ2 How can the interest affinities for a viewer be determined and used for segmenting viewers by interest?

#### CONTENT PERFORMANCE

RQ3 How can it be quantified what attracts more viewers in terms of content and what does not based on the interests / topics of each programme?

RQ4 How can content be better scheduled or promoted by understanding which are the trending interests in social media?

### 3.5 CONCLUSIONS

This chapter presented the relevant insights from the media market in the context of interest segmentation. First, the existing methods for understanding audiences were reviewed, with a focus on the data sources used for each. The use of TV ratings, social media streams, and data from video-on-demand services were compared as potential data sources for audiences segmentation. It was concluded that TV ratings do not provide enough data for segmentation because they only analyse a small sample of the viewers and extrapolate the numbers for the whole population. It was also noticed that social media has predictive capabilities for TV audiences, and can be used for inferring viewers' interests, but in many cases there is not enough activity online for all programmes in order to draw meaningful conclusions. By contrast, the video-on-demand solutions are able to track all the viewing history for individual users and therefore can be used for the purpose of the study.

In accordance with the research gap presented in the previous chapter, it was noticed that also in the media space, there are no methodologies in place to automatically segment viewers by their psychographic traits, most notably interests. The work done in the field of recommender systems was briefly presented, since there are some overlapping areas of interest between content-based recommendations and interest segmentation. After reviewing the advantages and disadvantages of different methods to quantify the similarity between various programmes, a list of research questions was structured on two main areas: interest segmentation and content performance. In the next chapter will describe the relevant insights

from the *Large Knowledge Bases* discipline, as it represents one viable alternative for identifying concepts in programmes.



## LARGE KNOWLEDGE BASES

---

In the previous chapter it was shown that in order to segment viewers by their interests, the topics that each programme is alluding to need to be analysed, as each can represent an interest. The existing methodologies for doing so are based on the comparison of programme genres or the similarity of the keywords used in the show's description. Programme genres are typically generic and subjective, and therefore do not represent a good fit. Keyword-based algorithms constitute a step forward, but make the comparison dependent on the wording of the description, as they do not detect semantic similarities between related concepts. For example, one person might have an interest in *South East Asia*, but if the keywords used in programs' descriptions are compared, a hypothetical show called *Exploring Bali* and one called *Island Life in Thailand* will not be deemed similar, since different keywords are used.

An alternative approach – in this case – could be based on linking the entities identified in the descriptions to the large knowledge bases available, like *Wikipedia* or *Freebase*. These sources of structured information contain millions of entities and hundreds of millions of facts about them, making it easier to identify various relationships. The following sections will present an overview of the area of large knowledge bases along with the formats used for the representation and linking of data, as well as the state of art in the field of named entity disambiguation. Finally, a set of open source and commercial tools for detecting entities in text are presented, as they provide

a practical means for creating the links between a programme's metadata and the relevant knowledge base entities.

#### 4.1 EMERGENCE OF KNOWLEDGE BASES

The emergence of knowledge bases has its roots in *Artificial Intelligence*, a term coined by McCarthy in the 1950s and defined as "the science and engineering of making intelligent machines" (McCarthy 2007). The associated field of study is interdisciplinary, combining knowledge from various disciplines like computer science, mathematics, linguistics, and philosophy. In order to accomplish the long term overall goal, Russel et al. (2009) identified a series of topics on which research is currently focused, including but not being limited to:

1. *Problem Solving* – an attempt to mimic the human decision-making process;
2. *Knowledge Representation* – the ability to collect and store a large number of facts about the world;
3. *Machine Learning* – the process of improving an algorithm based on previous experiences; and
4. *Natural Language Processing* – the technology for understanding what humans communicate via text or speech.

Given the promise of an intelligent machine that could perform the same tasks as a human, in the early 1950s there was a wealth of funding coming from governmental bodies, most notably the *Department of Defence* in the United States. However, given the reduced computational power available, correlated with difficulties in acquiring large bodies of knowledge, the results did not match the

expectations. As a consequence, both the United States and United Kingdom decided to funnel research funds into other directions, deemed more productive at the time (Russel et al. 2009). However, research into the area still progressed, and the funding situation changed abruptly during the 1980s, after the success of *expert systems*.

Expert systems are computer systems that try to emulate the decision-making capabilities of a human expert. They are typically divided into two main parts: a *knowledge base*, consisting of a set of facts and rules, and an *inference engine*, that can apply the rules in order to create new facts. For example, a knowledge base can contain the fact that *Bali is an island in South East Asia*, as well as a rule stating that *All islands in South East Asia have a tropical climate*. Using this information, an inference engine can create a new fact that states that *Bali has a tropical climate*. The first expert systems were developed by Feigenbaum (1980) – a Stanford University professor – and were aimed at solving problems in highly specific areas like diagnosing diseases (Shortliffe et al. 1975), or finding new molecules (Lindsay et al. 1993). While in other, more generic projects, building the knowledge base was seen as a multi-decade exercise, the focused approach taken by the expert systems community facilitated this process and lead to positive results early on. These accomplishments made expert systems the first successful form of artificial intelligence, and again attracted more interest and funding for the area. Over the next years, the artificial intelligence research went through additional cycles of excitement and increased funding followed by disappointment, most notably after the collapse of the market for Lisp machines (1987), the fall of expert systems (1993), and the abandonment of the *Fifth generation computer* project in Japan – all

these intervals being generically labelled *AI winters* (Russel et al. 2009, 24) .

Nevertheless, the emergence of expert systems had implications in terms of the technology stack being used. For developing the inference engines, a complete new set of languages emerged that were more expressive in the context of the problem being solved. This included systems based on IF-THEN statements, Lisp, or Prolog. In regards to the knowledge base repositories, the database systems being used at the time in the enterprise systems were not designed for this type of task. In principle, storage systems had to deal with flat tabular data, while knowledge bases needed support for structured data, typically in a format similar to the object model in computer programming: classes, subclasses, and instances. This allowed the knowledge to be easily classified, and at the same time the creation of symbolic links from one entity to another. This alternative way of structuring data is often called an *ontology* in the context of knowledge repositories.

#### 4.2 ONTOLOGIES

The origin of the term *ontology* is derived from a Greek word that translates to *being*, and in the context of information science represents a model that describes the world. Initially the term was used in the field of philosophy, where is referred to the study of being, becoming, or in other words, the systematic explanation of existence (Gómez-Pérez 1999). In order to make a clear distinction between the use of the term in different domains, Gruber (1995) introduced a definition specifically for the technical areas, like information science and computer science: “An ontology is a description (like a formal

specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general. And it is a different sense of the word than its use in philosophy”.

While there is no standard accepted way of building an ontology, the information is generally structured and formalised using a number of predefined components. Some of these, as described by Gruber (1995) and Gómez-Pérez (1999), are presented in the following list:

- **Classes** are sets of concepts, where a concept can be anything, abstract or concrete; they have the same meaning as classes in object oriented programming, describing types; example of classes could be *Person*, *Car*, *Programme*, or *Island*;
- **Instances** or **Individuals** are the concrete objects in an ontology, which are classified according to the classes defined previously. For example, if *Person* is a given class, *Alan Turing* or *Winston Churchill* are instances of that type;
- **Attributes** or **Properties** describe different features that instances can have, typically in the form of a text or number. For example the object *Alan Turing* can have attributes like *hasBirthdate: 23 June 1912*, or *hasNationality: British*;
- **Relationships** define links between concepts, describing how they are related to each other. For example, *Alan Turing* can have a *hasAlmaMater* relationship with *King's College, Cambridge*. While there are some disagreements in terms of terminology being used in the community, relationships between entities are generally considered to be of two main types: IS-A

relationship, and semantic relationships. The IS-A relationships are differentiating themselves from others because they form a hierarchy of concepts that can be navigated from specific to generic and vice-versa (e.g. *Car isA Vehicle*, *Dog isA Animal*, etc). The hierarchy is modelled as a *Rooted Directed Acyclic Graph (RDAG)*, a type of graph that has a single top-node and no cycles. The semantic relationships are also highly relevant, as they make the difference between a *taxonomy* (hierarchical classification of concepts) and the more general term of an *ontology*, which also includes other types of links in addition to the subsumption relation (IS-A);

- **Rules** can be IF-THEN statements or more complex structures expressed in ruled-based languages, that describe logical inferences that can be derived from facts. A visual representation of an ontology along with some of the components described above can be consulted in [Figure 18](#).

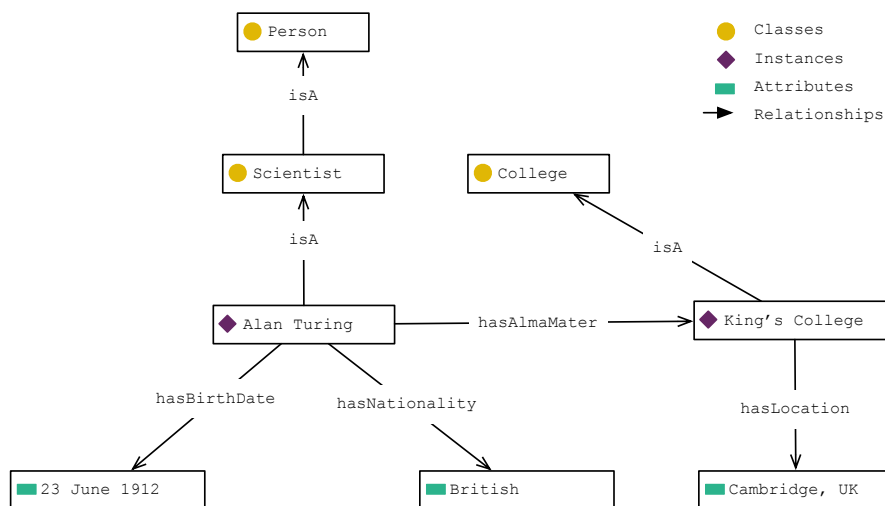


Figure 18: Visual representation of an ontology

Based on the degree of generality, ontologies are divided in two main types: domain ontologies and upper ontologies. Domain

ontologies describe only the concepts related to a particular field, like automotive or e-commerce, and are therefore smaller in size and easier to develop. They are relatively common in various areas where they help power expert systems. Typical examples include BioPAX (exchange and interoperability of biological pathway), BMO (e-Business Model Ontology), or PRO (Protein Ontology of the Protein Information Resource, Georgetown University) (Gupta et al. 2013). By contrast, upper ontologies attempt to describe general concepts and are hence reusable across different domains. Given their generality, they are not easy to build, and the process is automated in some instances. For the purpose of this research, upper ontologies are more relevant, as viewers' interest can point to a wide variety of subjects, therefore making the process of building a lower level ontology for describing all the options challenging. Instead, by using an upper level ontology, an interest can refer to any concept (e.g. persons, locations, activities, hobbies, organisations, etc). In the following subsections the most common large knowledge bases will be briefly described along with their corresponding ontologies.

#### 4.2.1 *Cyc*

*Cyc* is one of the oldest attempts to create a large knowledge base that describes common-sense knowledge. The project started in 1984, and at the time was estimated to require 350 years of man-effort in order to manually define 250,000 rules, the number estimated at the time as appropriate to achieve the overall goal. With an initial focus on artificial intelligence applications, the purpose of *Cyc* was to enable applications to reason in a similar way as a human, with the help of the extensive knowledge base. More than thirty years after, the project contains more than 500,000 concepts

and five million assertions, and it is released under three different versions: *EnterpriseCyc*, *ResearchCyc*, and *OpenCyc*. The *OpenCyc* version is a subset of the original knowledge base offered as open source, while the other two provide access to the entire data, with different licensing models for research and commercial applications. In addition to the knowledge base, Cyc also provides an inference engine based on logical deduction.

The format used for defining the Cyc knowledge base is called *CycL* and has a syntax similar to the Lisp programming language. The concepts are called *constants*, prefixed with `#$` and are used for individuals (e.g. `#$Dog`, `#$Plant`), collections of individuals (e.g. `#$Mammals` – contains all mammals), truth functions (given an argument consisting of one or more concepts, return a value of true or false – e.g. `#$siblings` returns true if the provided arguments have a sibling relationship), and functions (derive new terms based on arguments – e.g. `#$Fruits` returns only the fruits from a collection of items provided as an argument). In addition to the constants, Cyc contains a large number of predicates, that are written before their arguments, in parenthesis (e.g. `(#$capitalCity #$UnitedKingdom #$London)` – specifies that London is the capital city of the United Kingdom). The predicates can also contain variables instead of constants, in which case they are called *rules*. The most important predicates used in the platform are for describing the hierarchy of concepts, namely `#$isa` and `#$genls`, that refer to the fact that one item is an instance of a collection, or that one collection is a sub-collection of another one.



#### 4.2.2 Freebase

*Freebase* is another widely used knowledge base developed by a company called *Metaweb*. The objective of the project was to be a public repository of the world knowledge, hence supporting a large body of diverse and heterogenous data. According to Bollacker et al. (2008), *Freebase* was inspired from the model of online information communities like *Wikipedia*. While traditional database schemas are rigid, and do not accommodate structural diversity, wiki-type websites are a much better fit for the task, but do not provide enough methods to query the data. Therefore, the *Freebase* project is an attempt to combine the benefits of both worlds while merging the scalability and querying capabilities of databases, while adding support for diverse data, more specific to collaborative sites like *Wikipedia*. In contrast with the *Cyc* project, that developed its knowledge base entirely in-house, *Freebase* took a different approach and stored information contributed by its community members, or harvested from other sources like *Wikipedia*, *Notable Names Database (NNDB)*, *Fashion Model Directory (FMD)*, or *MusicBrainz*. As of April 2015, *Freebase* stores approximately 50 million topics and three billion facts about them.

In terms of structuring the data, the entities in *Freebase* are called *topics*, and each topic can have one or more *types*. For example, *Alan Turing* is a topic, and some of the types associated to it are *Person*, *Computer Scientist*, *Academic*, or *Film Subject*. The facts for each topic depend on the type of the entity, for example persons can have a birthdate and a birth place, or film subjects contain a list of movies about the subject, etc. Another important difference compared to *Cyc* is rooted in the fact that *Freebase* does not enforce a certain

ontology for classifying the topics, but instead adopts a folksonomy – a collaborative approach in which users can add their own categories (O’Reilly 2008). Therefore, there is no standard way of classifying the data, and contradictory or overlapping types or properties will reflect different opinions or understandings of the users (Bollacker et al. 2008).

Metaweb, the company that developed Freebase, was acquired by Google in 2010 for the inclusion of its data into the *Knowledge Graph* project<sup>1</sup>. The idea behind it was to enhance the search engine results with the help of semantic information from different sources. The size of the Knowledge Graph is estimated to be around 500 million topics with 18 billion facts, collected from Freebase, Wikipedia, CIA World Factbook and others. In December 2014 Google announced that it will shutdown the Freebase service over the next six months, and transfer all the data to the Wikidata initiative<sup>2</sup>.

#### 4.2.3 *DBPedia*

While the Freebase initiative attempted to build a knowledge base on the model in which wikis work, *DBPedia* approach was based on parsing the data already available in Wikipedia. The project started in 2007, around the same time as Freebase, and was developed as a collaborative effort between two academic partners (*University of Leipzig* and *University of Mannheim*) and one technical partner (*OpenLink Software*). While the information available in most Wikipedia pages is in the form of free text, many articles also have an *infobox* attached, usually displayed on the top right of the page. These infoboxes contain structured information, typically related to

<sup>1</sup> Introducing the Knowledge Graph: things, not strings – <http://goo.gl/GLArfX>

<sup>2</sup> Freebase Shutdown Announcement – <http://goo.gl/1lgpf6>

the category of entity described, geographic coordinates, or links to external pages. An overview of the infobox displayed in the article about *Alan Turing* can be consulted in [Figure 19](#). The latest English version of DBPedia, extracted from a 2014 snapshot of Wikipedia, contains information about four million entities, most of them classified according to the DBPedia ontology, and approximately 70 million facts. The complete version, including topics from all languages, describes around 40 million entities along with three billion facts.

While the Cyc project adopted an ontology and Freebase a folksonomy for classifying their data, DBPedia adopted a hybrid approach and classifies the data in relation to a number of ontologies and folksonomies. Bizer, Lehmann, et al. (2009) described the different classifications schemes as well as their intended applications:

- **Wikipedia Categories** are similar to the types used in Freebase. The concept is used for grouping together articles that have similar subjects. For example, the article about *Alan Turing* on Wikipedia is currently linked to 47 categories, including *English Computer Scientists*, *People from Maida Vale*, or *Fellows of Royal Society*. The advantage of using categories is related to their dynamic nature, as they are being user contributed and evolve along with Wikipedia; and their large numbers, currently in the order of hundreds of thousands. However, the Wikipedia categories do not provide any hierarchical topic tree, and the strength of the link between two members of the same category can be rather weak (Bizer, Lehmann, et al. 2009). Arguably, based on the example mentioned, the link between two members of the *British Cryptographers* list might be deemed

## Alan Turing



From Wikipedia, the free encyclopedia

"Turing" redirects here. For other uses, see *Turing (disambiguation)*.

**Alan Mathison Turing**, *OBE*, *FRS* (/tɪʝərn/ *TEWR-ing*; 23 June 1912 – 7 June 1954) was a British pioneering **computer scientist**, **mathematician**, **logician**, **cryptanalyst**, philosopher, mathematical biologist, and marathon and ultra distance runner. He was highly influential in the development of **computer science**, providing a formalisation of the concepts of "**algorithm**" and "**computation**" with the **Turing machine**, which can be considered a model of a general purpose computer.<sup>[2][3][4]</sup> Turing is widely considered to be the father of theoretical computer science and **artificial intelligence**.<sup>[5]</sup>

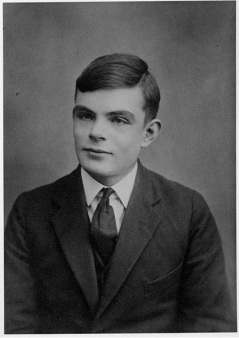
During the **Second World War**, Turing worked for the **Government Code and Cypher School** (GC&CS) at **Bletchley Park**, Britain's **codebreaking** centre. For a time he led **Hut 8**, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German **ciphers**, including improvements to the pre-war Polish **bombe** method, an **electromechanical** machine that could find settings for the **Enigma machine**. Turing's pivotal role in cracking intercepted coded messages enabled the Allies to defeat the Nazis in many crucial engagements, including the **Battle of the Atlantic**; it has been estimated that the work at Bletchley Park shortened the war in Europe by as many as two to four years.<sup>[6]</sup>

After the war, he worked at the **National Physical Laboratory**, where he designed the **ACE**, among the first designs for a stored-program computer. In 1948 Turing joined **Max Newman**'s Computing Laboratory at **Manchester University**, where he helped develop the **Manchester computers**<sup>[7]</sup> and became interested in **mathematical biology**. He wrote a paper on the chemical basis of **morphogenesis**, and predicted **oscillating chemical reactions** such as the **Belousov–Zhabotinsky reaction**, first observed in the 1960s.

Turing was prosecuted in 1952 for **homosexual acts**, when such behaviour was still **criminalised in the UK**. He accepted treatment with **oestrogen** injections (**chemical castration**) as an alternative to prison. Turing died in 1954, 16 days before his 42nd birthday, from **cyanide** poisoning. An inquest determined his death a suicide, but it has been noted that the known evidence is equally consistent with accidental poisoning.<sup>[8]</sup> In 2009, following an **Internet campaign**, **British Prime Minister Gordon Brown** made an **official public apology** on behalf of the British government for "the appalling way he was treated". **Queen Elizabeth II** granted him a posthumous **pardon** in 2013.<sup>[9][10][11]</sup>

<b>Contents</b> <span>[hide]</span>
1 Early life and family
2 Education
2.1 University and work on computability
3 Cryptanalysis
3.1 Bombe
3.2 Hut 8 and Naval Enigma
3.3 Turingery
3.4 Delilah

**Alan Turing**  
**OBE, FRS**



Turing at the age of 16

**Born** Alan Mathison Turing  
23 June 1912  
Maida Vale, London, England

**Died** 7 June 1954 (aged 41)  
Wilmslow, Cheshire, England

**Residence** Wilmslow, Cheshire, England

**Nationality** British

**Fields** Mathematics, cryptanalysis, computer science, biology

**Institutions** University of Manchester  
Government Code and Cypher School  
National Physical Laboratory  
University of Cambridge

**Alma mater** Sherborne School  
King's College, Cambridge  
Princeton University

**Thesis** *Systems of Logic based on Ordinals* ⓘ (1938)

**Doctoral advisor** Alonzo Church<sup>[1]</sup>

**Doctoral students** Robin Gandy<sup>[1]</sup>

**Known for** Cryptanalysis of the Enigma, Turing machine, Turing test

**Notable awards** Smith's Prize (1936)  
OBE  
FRS<sup>[2]</sup>

Figure 19: The infobox from Alan Turing's page on Wikipedia, marked with a red rectangle, from which structured information is being extracted by the DBpedia project

stronger than the one between two *Residents of Maida Vale*, but the assessment is highly subjective and varies based on its intended use;

- **YAGO** is an alternative taxonomy obtained from the linking of the previously described Wikipedia categories with WordNet, a lexical database for the English language. Because the linking

is done automatically, the resulting taxonomy is not entirely correct, but according to its creators it was humanly verified to be 95% correct. Following this process, the entities in YAGO form a deep subsumption hierarchy, so it is possible to navigate from a general concept to a more specific one and vice versa. The latest version of the YAGO taxonomy contains approximately 350,000 classes. In addition to the taxonomy, the project also contains the YAGO Ontology, a collection of ten million entities with around 120 million facts. The structured knowledge in YAGO has been used by Watson, an artificial intelligence system developed by IBM that managed to win in a *Jeopardy!* contest against the two most successful players in the history of the game;

- **DBPedia Ontology** is a shallow, manually built ontology, derived from the most common Wikipedia infoboxes types. It currently contains around 700 classes with approximately 3,000 properties that describe them. Similarly to YAGO, the classes in the ontology form a hierarchical classification under the form of a directed acyclical graph.

## 4.3 LINKED DATA

### 4.3.1 *The Web of Data*

The previous sections showed how the emergence of ontologies provided a new way of structuring knowledge. The process started with domain ontologies, used by expert systems for making decisions, and continued with the development of upper ontologies that powered systems describing general knowledge like DBPedia and Freebase. However, all these knowledge bases created silos of

information, with no connections between them, making it hard for applications to make sense of the whole data being made available on the Internet. Moreover, they all used different ontologies to structure the information, making it hard to reason when using data from more than one system. Ideally, if all the information published on the internet would be described in a machine readable format and interlinked, a complete new set of opportunities will arise for intelligent agents to take advantage of.

The solution for this problem came from the *Semantic Web* community, that had the vision of a *Web of Data*, in which information is published in a format that machines can naturally understand and navigate. While there are different terms used in the community, based on the aspect emphasised (e.g. *Semantic Web*, *Web of Data*, *Linked Data*), they all try to achieve the same overarching goal. A more formal definition of the *Linked Data* concept is provided by Bizer, Heath, et al. (2009) that defines it as “data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets”. In order to achieve this goal, Berners-Lee (2006) defined a set of rules for publishing data in a format that would make it compliant with the vision, usually referred to as the *Linked Data Principles*:

- Use URIs<sup>3</sup> for identifying things;
- Use HTTP URIs so that machines can look-up / access things;
- Provide useful information when someone looks-up things using a standard format;

<sup>3</sup> “Uniform Resource Identifier (URI) are string of characters used to identify a name of a resource. Uniform Resource Locator (URL) refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism (e.g. its network location)” - Definition from <http://tools.ietf.org/html/rfc3986>

- Create links to other URIs so that more things can be discovered;

Four years after the publishing of the *Linked Data Principles*, in order to encourage stakeholders – especially government data owners – to follow them when publishing data, a 5-star rating system complemented the initial criteria. An overview of the rating system can be consulted in [Figure 20](#).

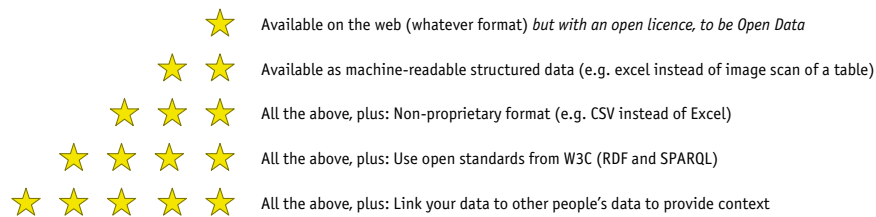


Figure 20: Linked Open Data – 5-star Rating System – Adapted from Berners-Lee (2006)

The ability to easily connect different data sources on the web created new opportunities by moving from multiples silos of knowledge on the web to one web-scale knowledge base. A number of *World Wide Web Consortium* efforts along with *European Union* projects (*Linking Data Around the Clock*<sup>4</sup>, *PlanetData*<sup>5</sup>, *Data-and-Platform-as-a-Service*<sup>6</sup>, *Linked Open Data* <sup>7</sup>) led to the creation of a *Linked Open Data Cloud*. The size of the cloud grew fast from an initial set of only twelve datasets in 2007 to approximately 1,014 currently. The visualisation of the cloud at the end of 2007 can be consulted in [Figure 21](#), while the the present version is displayed in [Appendix A](#). Interestingly, from the very beginning of the project, the DBPedia dataset emerged as a crystallisation point for the Linked

<sup>4</sup> *Linking Data Around the Clock* – <http://goo.gl/4iSM9E>

<sup>5</sup> *Planet Data* – <http://goo.gl/ar5G6D>

<sup>6</sup> *Data-and-Platform-as-a-Service* – <http://goo.gl/ARjNc2>

<sup>7</sup> *Linked Open Data 2* – <http://goo.gl/1v2wtR>

Open Data Cloud, as many of the datasets linked their own entities to DBpedia (Bizer, Lehmann, et al. 2009).

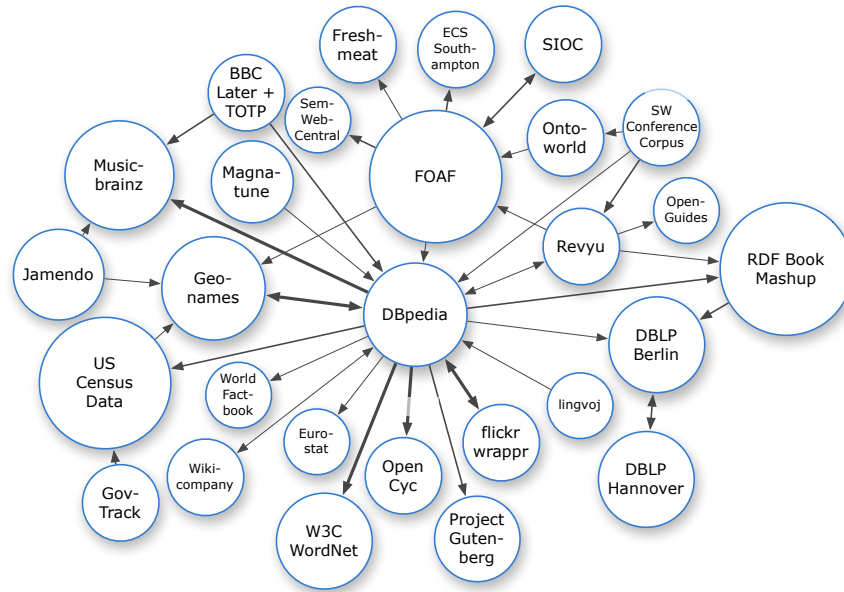


Figure 21: The state of the *Linked Open Data Cloud* at the end of 2007

#### 4.3.2 Technology Stack

In a similar way in which the developing of knowledge bases in the early 1970s revolutionised the technology stack, making the transition from traditional databases to ontology based systems, the large scale adoption of the *Web of Data* required a standardisation of the technologies used. The objective was to make the various systems that publish data on the web adhere to the linked data principles, and therefore be interoperable. The first proposal for a technology stack was created by Berners-Lee (2000) and consisted of a layered architecture, where each layer depends on functionality provided by the previous ones, namely: identifiers (URI), syntax (XML), data interchange (RDF), taxonomies, ontologies, rules, querying, unifying logic, proof, and trust. While the base levels have been standardised,



and are already used in various applications, is it not the same case for the upper layers, and therefore the overall vision of the Semantic Web is not yet accomplished (Figure 22).

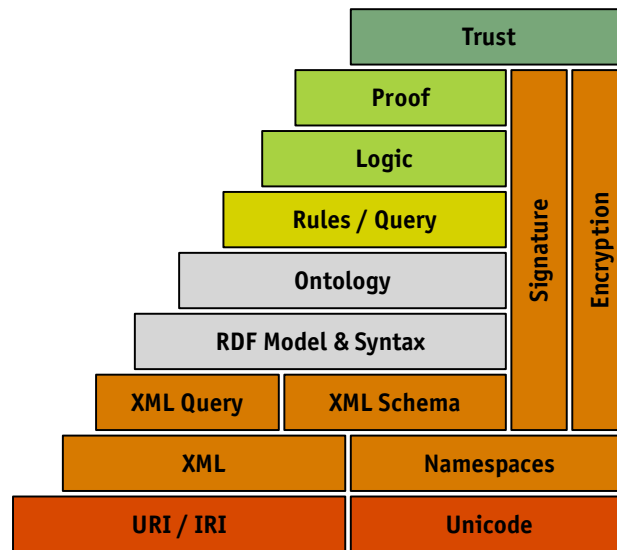


Figure 22: The Semantic Web Stack describes the various layer of technologies being standardised (Berners-Lee 2000)

In order to cater for the level of expressivity and generality required for describing knowledge, the World Wide Web Consortium provided a specification for a model called *Resource Description Framework (RDF)*. This is similar in approach to the entity-relationship model, but more general, as it can express any statement in terms of *triples*. A triple is a construct in the form of *Subject – Predicate – Object*, where the *Subject* and *Object* denote entities (URIs in the context of Linked Data), while the *Predicate* describes either a relationship between the entities or the existence of an attribute. For example, the *Alan Turing is a British scientist* sentence can be decomposed into the following two RDF triples, as shown in Listing 1.

Since all the nodes have unique identifiers (the dbPedia short form used in the previous example is replaced by the full URI

```

@prefix dbpedia-owl: <http://dbpedia.org/ontology/>.
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbprop: <http://dbpedia.org/property/>.

<dbpedia:AlanTuring> <rdf:type> <dbpedia-owl:Scientist> .
<dbpedia:AlanTuring> <dbprop:nationality> "British"@en .

```

Listing 1: RDF Serialisation Example

described by its associated prefix), any collections of triples can be represented as a labelled directed multi-graph, the same format used by the ontologies described earlier.

Being a text-based format, RDF can be serialised in various ways, most notably Turtle (a compact human-friendly format), N-Triples (optimised for the serialisation and deserialisation of large datasets), or RDF/XML (a XML-based syntax initially used for serialising RDF). Because of its simplicity, RDF can be easily queried, and there is also a language called SPARQL designed for that purpose. The name is a recursive acronym for *SPARQL Protocol and RDF Query Language*, and was introduced as an official recommendation by the *World Wide Web Consortium* in 2008, with an updated version (1.1) being released in 2013. According to Pérez et al. (2006), SPARQL is essentially a graph matching query language, where for a given graph used as a data source, a query consists of a pattern that is matched against the graph. Moreover, the results can be preprocessed into the required format before being streamed back to the user. The code snippet below describes a simple query that uses the Friend-of-a-Friend Ontology (FOAF) to return a list of all the persons in the data source along with their email addresses. In the pattern matching section of the query, defined under the *where* clause, all the conditions are specified in terms of triples (subject – predicate – object), and all the variables are prefixed with the question mark sign. The first condition filters

all the nodes in the graph that have a the type `foaf:Person`, and then the next two conditions filter the nodes that are linked to the nodes selected in the first condition with `foaf:name` and `foaf:mbox` type of relationships. All the other nodes in the graph are ignored. The describe query can be consulted in [Listing 2](#).

```

PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person rdf:type foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}

```

Listing 2: SPARQL Query Example

In addition to the ability to query data from one source at the time, SPARQL also has support for federated queries, where data from multiple sources (usually called SPARQL endpoints in this context) can be queried in parallel, and the results aggregated. The need to store a large volume of triples as well as the ability to query them attracted a lot of attention from both commercial and open source players. Currently there are a number of triple stores that implement the SPARQL standard and can store massive amounts of data: *Oracle Spacial Graph*<sup>8</sup> (>1 trillion triples), *Allegro Graph*<sup>9</sup> (>1 trillion triples), *Stardog*<sup>10</sup> (50 billion triples), *Open Link Virtuoso*<sup>11</sup> (15 billion triples), *Apache Jena TDB*<sup>12</sup> (1.7 billion triples). In addition to these systems, any other graph storage solution could potentially be used, even if the syntax for querying could be different. One example is *neo4j*<sup>13</sup>, a graph database that does not support SPARQL,

<sup>8</sup> *Oracle Spacial Graph* – <http://goo.gl/zZWMfu>

<sup>9</sup> *Allegro Graph* – <http://goo.gl/B1Ta00>

<sup>10</sup> *Stardog* – <http://goo.gl/01Bx1T>

<sup>11</sup> *Open Link Virtuoso* – <http://goo.gl/LQAdNv>

<sup>12</sup> *Apache Jena TDB* – <http://goo.gl/pqEcVY>

<sup>13</sup> *neo4j* – <http://goo.gl/yIRhRg>

but includes an alternative graph matching query language called *Cypher*.

While RDF represents a format for data interchange, being able to model a graph for describing any type of knowledge, it lacks some of the expressivity of languages used for defining ontologies (e.g. *CycL* used by the *Cyc* ontology). In order to cater for that, the Semantic Web stack includes additional layers of standards that add further expressivity. *RDF Schema* (RDFS) builds on top of RDF by defining a set of classes for describing subjects, objects, or predicates and their relationships and properties. While very generic, RDFS can be used for defining basic restrictions on the data (e.g. the *hasFather* property can only be defined between two nodes of *Person* type), or that one class or property is a subtype of a more general one (e.g. *hasSister* is a subtype of *hasSibling*) and validate a dataset against that criteria. For more complex requirements, the *World Wide Web Consortium* also proposed a specification for OWL (Web Ontology Language). OWL adds support for simple rules, for example *if A is a sibling of B, B is a sibling of A, or if A is a sibling of B, and B is a sibling of C, A is a sibling of C*. Moreover, OWL introduced a concept that defines a relation of similarity between items (*owl:sameAs*). This is a powerful technique in the context of the *Web of Data*, since multiple data sources can describe the same entity using different vocabularies. Being able to collect all the data into one place that contains the entire set of properties and assertions about that specific entity, opens the way to new algorithms than can make better use of the continuously evolving knowledge bases.

The remaining layers in the *Semantic Web* stack have not been standardised yet. Most notably, for the rules level, OWL provides

some functionality to address it, but it can also be complemented by other languages that were proposed for defining more expressive rules. SWRL is one such language, that adds the benefits of logical programming on top of the description logic provided by OWL-DL. Other alternatives include *TopSpin* and *RIF*, however, none of these proposals have been standardised and provided as a recommendation by the *World Wide Web Consortium*. There are also a wide range of reasoners that can make use of OWL and SWRL in order to infer new facts based on the existing knowledge and a set of rules. The remaining layers, *Proof* and *Trust*, haven't been explored enough just yet, but they would be required for the vision of the *Semantic Web* to be fully implemented.

#### 4.4 DETECTING INTERESTS

Being able to understand the relationship between various entities, a representation of viewers interests for the purpose of this study, generates a new set of possibilities for identifying patterns that can be then used for interest segmentation. For example, if two viewers watch a number of programmes, some about *Maldives* and some about *Seychelles*, using knowledge bases one could infer that both viewers might have an interest in *Indian Ocean Islands*. However, for accomplishing such a task, a reliable way of identifying the relevant entities for each show and linking them to the knowledge bases is needed. This type of algorithm could analyse the programmes' descriptions or closed captions for extracting the concepts. The following sections will provide an assesment of the state of art in terms of *Named Entity Recognition* (NER) and *Named Entity Disambiguation* (NED). These two technologies make the identification of concepts in text corpus and linking them to

knowledge bases possible. Finally, a number of implementations available for the two technologies, both from the open source and commercial environments, will be presented.

#### 4.4.1 Named Entity Recognition (NER)

*Named entity recognition* is the process of identifying various entities in text and classifying them according to a list of predefined categories. Typical examples of entities recognised include persons, organisations and locations, along with time related, quantity related, and monetary related entities. For example the following input text:

*“The atmosphere at King’s College, Cambridge, was highly conducive to free-ranging thought. As an undergraduate there, Alan Turing developed the inspiration he had received from Christopher Morcom, and combined it with the newest ideas in mathematics.”*

can be processed and annotated in the following form:

*“The atmosphere at [King’s College, Cambridge]<sub>Organisation</sub>, was highly conducive to free-ranging thought. As an undergraduate there, [Alan Turing]<sub>Person</sub> developed the inspiration he had received from [Christopher Morcom]<sub>Person</sub>, and combined it with the newest ideas in [mathematics]<sub>Discipline</sub>.”*

The algorithms implementing NER are typically divided into two main parts: detecting the entities and classifying them (Carreras et al. 2003). The first step, usually referred to as *entity chunking*, needs to take into consideration that names can be defined as a sequence of tokens (e.g. *Bank of England* is considered one entity,

even if *England* is an entity of its own). Moreover, entities do not necessarily have to be represented by a noun, but can also be represented by a phrase (e.g. *the Apple founder*). In regards to the classification of entities, this is usually performed in relation to a provided hierarchy. Some of the categories commonly used include *BBN Categories for Answer Types*<sup>14</sup> and *Freebase* extracted categories. The techniques used are either based on grammar (linguistic) or machine learning (statistical) approaches. The first category usually provides better accuracy, but at the same time involves a lot of human effort in the process. The machine learning approaches reduce the effort, while only slightly reducing precision. Most implementations, including the *Stanford NER*<sup>15</sup> are based on *Conditional Random Fields (CRF)*. In terms of accuracy, state of the art algorithms achieve close to human performance for recognising entities, with experiments constantly resulting in approximately 93% accuracy, compared to a human accuracy of 97% (Carreras et al. 2003; Marsh et al. 1998).

#### 4.4.2 *Named Entity Disambiguation (NED)*

While the process of Named Entity Recognition can identify and categorise the various entities present in text, it does not make any difference between ambiguous constructs. For example, the noun *Jobs* could refer to either the common noun, a person with that last name, or the *Apple* founder *Steve Jobs*. *Named Entity Disambiguation (NED)* attempts to make this distinction, by linking the recognised entities to their references into a common knowledge base, a process also referred to as *Entity Linking*. This is highly relevant for the overall goal of detecting interest for two main reasons:

<sup>14</sup> *BBN Categories for Answer Types* is used for question answering and contains 29 types and 64 subtypes – <http://goo.gl/pl09ke>

<sup>15</sup> *Stanford Named Entity Recogniser (NER)* – <http://goo.gl/QvGNs0>

- it can more precisely identify a given interest when the word sense is ambiguous;
- it links the entities to a knowledge base, and therefore unlocks new possibilities to analyse data based on the facts found in the knowledge base (e.g. *Seychelles* is in the *Indian Ocean*, *Apple* is known for manufacturing the *iPhone*, etc).

Most of the algorithms for entity linking have been developed for annotating corpus with links pointing to their corresponding DBpedia entities, a process also called *Wikification*. The first attempt to do so was *Wikify!*, a system proposed by Mihalcea et al. (2008). The process involves two main steps, namely identifying keywords in the text, and then disambiguating them. For identifying the keywords, the authors computed the average number of links from a Wikipedia article to another. This is relevant given Wikipedia author's guidelines, which recommend users to only create links to other content pages that are strongly related to the current one. Having this number computed, the next step is ranking the candidate keywords by a metric called *keyphraseness*. This is the ratio between the number of times the term was selected as keyword (in the form of a link to another Wikipedia page), divided by the number of times the term has been used. The precision of this method has been estimated to approximately 53%, ranking higher than the alternatives, namely *tf.idf* and  $x^2$ . For the second phase – disambiguation – the method employed is based on a machine learning classifier trained with the context of a given term (3 words to the left and right and their parts of speech) and the resulting link. For example, most of the mentions of the word *bar* that include in the context terms like *drink* or *bill* contain a link for the *Bar - Establishment* page on Wikipedia, while most references where the term *legal* appears in the context will link the



the *Bar - Law* page. These two methods combined achieve a precision of over 90% compared to a human performing the same task.

Milne et al. (2008) propose an alternative machine learning algorithm for disambiguation, that factors in *commonness* as well as *semantic relatedness*. *Commonness* is a metric derived from the number of times a term is linked to on Wikipedia. For example, according to Milne et al. (2008), the term *tree* refers in 93% of the cases to the plant, and in only 3% of the cases to the computer science data structure. *Semantic relatedness* is computed as the ratio between the number of articles that mention two terms together (e.g. *tree* and *flower*), divided by the number of articles that mention any of the two terms. While *semantic relatedness* is clearly a more relevant metric, it is highly dependent on the number of existing links for a given term. Therefore, when the context is not rich enough, the *commonness* metric can be used instead. In order to overcome this problem, Kulkarni et al. (2009) proposed a collective optimisation approach based on a combination of local content (*node potential*) with global context (*clique potential*). However, the approach is NP-hard so it is computationally expensive to use for large corpus, even with the proposed domain specific heuristics (Moro et al. 2014). While the foundations of the *Named Entity Disambiguation* are clearly understood, the area retains a high academic interest, especially in the context of social media messages, where the size of the typical message – in some cases as short as 140 characters (Twitter) – generates new challenges given the lack of context words.

#### 4.4.3 Web Services

Following the success of NER and NED in the academic environment, the technology has been commoditised, being made available to the wider community under the form of open source libraries and web service APIs. Some of the most used solutions in the area are described in the following list:

- **Stanford NER**<sup>16</sup> is a widely used open source Java library for NER, that recognises and classifies locations, persons, and organisations. While the precision of the system is reported to be higher compared to the alternatives (Rodriquez et al. 2012), it does not support disambiguation of the entities by linking the results to a knowledge base, and therefore has limited use for the purpose of this study;
- **DBPedia Spotlight**<sup>17</sup> has been made available by the DBPedia project, and consists of a web application complemented by a web service that disambiguates text corpus in relation to DBPedia entities. Users have the possibility of restricting the entities that are recognised, having filters in place for the DBPedia Ontology and Freebase Categories. The algorithms behind it are described by Daiber et al. (2013), while the source code is also freely available;
- **AIDA**<sup>18</sup> is another open source project, also available as a web service, that links entities from text to the YAGO knowledge base. For the named entity recognition the project relies on the *Stanford NER Tagger*, while for disambiguation there are three different options: *Prior Only* (the entity is linked to

<sup>16</sup> Stanford NER – <http://goo.gl/QvGNs0>

<sup>17</sup> DBPedia Spotlight – <https://goo.gl/af7uVq>

<sup>18</sup> AIDA – <https://goo.gl/8s50U0>

the most common sense in the knowledge base), *Local* (each entity is disambiguated individually), and *Cocktail Party* (all entities are disambiguated collectively using a graph-based approach)(Cornolti et al. 2013; Hoffart et al. 2011).

- **AlchemyAPI**<sup>19</sup> is a commercial web service that provides support for NED amongst other options (sentiment analysis, image tagging, language detection). While Spotlight and AIDA link the entities to their corresponding ontologies, AlchemyAPI provides more options including DBPedia, YAGO, Freebase, OpenCyc, and others. Also, in addition to identifying named entities, there is also support for identifying concepts in the text in the same way a human would do by analysing the relationship between entities. For example, a text that mentions the *CERN* and *Higgs boson* would be linked to the *Large Hadron Collider*, even though the term itself was not in the analysed text (Hooland et al. 2013). While the service requires a subscription, it is available at no cost for up to 1,000 transactions a day, limit which is raised up to 30,000 a day for research purposes. The company was acquired in 2015 by IBM in order to improve Watson’s deep learning capabilities<sup>20</sup>.
- **Open Calais**<sup>21</sup> is another commercial product, part of the Thomson Reuters’ offering. It can detect entities, relationships, facts, and events from free text, using various machine learning algorithms trained by Reuters’ editorial team. In addition to the disambiguation service, Open Calais can also identify the topics in a given text. However, one of the drawbacks compared to the other solutions is that the results are structured using

<sup>19</sup> AlchemyAPI – <http://goo.gl/n9MSdM>

<sup>20</sup> IBM Acquires AlchemyAPI, Enhancing Watson’s Deep Learning Capabilities – <http://goo.gl/5MqyZt>

<sup>21</sup> Open Calais – <http://goo.gl/aG4tI0>

a proprietary ontology, and only few of the classes are linked to their corresponding entities in the Linked Open Data Cloud (Hooland et al. 2013). The service can be used for up to 5,000 transactions a day, but requires a subscription over that limit.

#### 4.5 CONCLUSIONS

In the previous chapters a knowledge gap was identified in the area of inferring users' interest based on their interaction with products. This is especially relevant in markets like media, where the decision to view a certain programme is highly influenced by a person's interests. However, having a limited capability to analyse the topics that each programme is touching upon, companies cannot effectively segment users by their interests or improve their content to fulfil a need in the market. This chapter presented the relevant insights from the area of large knowledge bases, including relevant technologies that represent a viable alternative for automating the process of identifying the topics in each programme. Moreover, by leveraging the use of large knowledge bases for understanding the relationship between concepts, companies might also be able to understand the underlying reasons for which viewers watch certain programmes. While these technologies have been commoditised in the recent years, their applicability to the topic of this study is not yet determined. For example, one needs to be able to understand if the concepts detected are relevant from the perspective of viewers' interests, which ontologies might be better fitted for detecting related concepts, or what type of programme descriptions could work better.

Having concluded the *Disciplinary Insights* part of this research, the hypothesis of this study is that it might be possible to

fully automate the process of segmenting viewers based on their interests. Drawing on these insights, the next chapter will present a methodology for doing so, which will be empirically validated.

Part III

INTEGRATING INSIGHTS

## METHODOLOGY

---

The previous three chapters identified a series of insights by approaching the interest segmentation problem from three different disciplines (Figure 23). First, from a customer segmentation point of view, scholars identified that psychographic segmentation is one of the most useful techniques for predicting consumer behaviour. However, the research methods used in marketing, relying heavily on surveys, make large scale psychographic segmentation studies difficult to perform. From a media market perspective, it was noticed that there is a wealth of data collected about individuals' viewing habits, but the segmentation methods currently employed focus solely on demographics, since psychographic data cannot be inferred. Finally, the advances in the field of large knowledge bases were presented, with a focus on the techniques that identify concepts in text and disambiguate them by creating links to large knowledge bases.

Given that any concept can represent a viewer's interest, the hypothesis of this study is that by analysing the patterns of online viewing, and collating this with information extracted about the concepts identified in the programmes' descriptions, a methodology for automatically segmenting customers based on their interests can be created. While this is the primary goal of this study (identified by the research questions RQ<sub>1</sub> and RQ<sub>2</sub>), this capability also opens new possibilities for media companies in terms of understanding the performance of their programmes (RQ<sub>3</sub>) or reacting to social

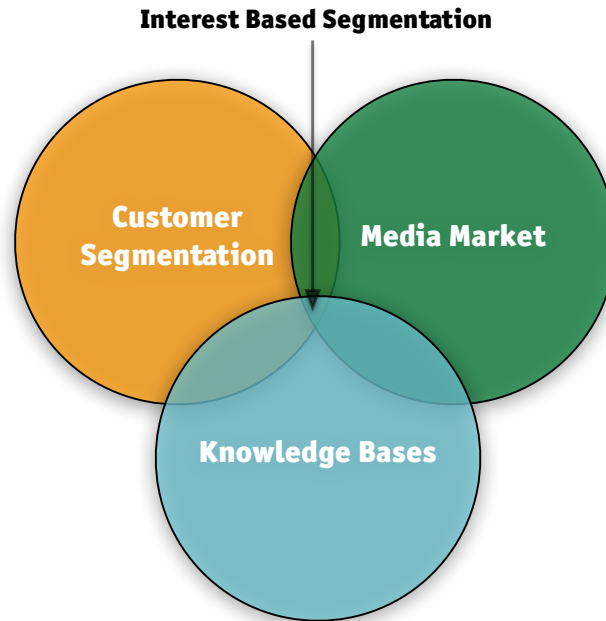


Figure 23: Interest-based segmentation of viewers is considered to be an interdisciplinary area, where insights from *Customer Segmentation*, *Media Market*, and *Large Knowledge Bases* need to be integrated in order to define a coherent methodology

media trends (RQ4). Drawing on the disciplinary insights, this chapter will present the proposed methodology for inferring viewers interests based on the video content they watched. The methodology presented will be empirically validated, and the results presented in [Chapter 6](#) used to further refine it.

The following sections will first describe the scientific inquiry's philosophy, followed by the research design, data collection, and data analysis methods employed. Finally, the ethical implications along with considerations on generalisability, validity, and limitations will be presented.



## 5.1 RESEARCH PHILOSOPHY

From a philosophy point of view, this study adopts *post-positivism*. This particular stance is considered to be a good fit for interdisciplinary work, as it requires the scientific rigour of positivism, but at the same time accepts the effects of the researcher's biases and values. Translated in terms of ontology and epistemology, the philosophy implies the existence of an objective world, but underlines the fact that the knowledge about the world is only partial, and filtered through individuals' experiences. While ideally, from a scientific point of view, the success of a segmentation study should be quantified, the practicality of this approach is relatively low from a research point of view, as it would require a company to expose itself to a considerable amount of risk. For example, having a methodology for interest segmentation defined, a company could attempt to use it for better targeting their customers and reducing churn. However, given the multitude of factors that are relevant in such a scenario, isolating the effect of segmentation in the overall success of the company could prove challenging, and pursuing a business to undergo through an extensive experiment has low chances of success.

This is particularly relevant for the nature of the study due to the fact that some of the research questions can be addressed exclusively through the use of quantitative methods, but others require a qualitative interpretation of the statistical results. For example, the size of the customers segments generated can be precisely computed, but a certain degree of subjectivism is required in understanding the relevance of the resulting segments for the organisation. More specifically, some of the identified segments can pass all the relevant

quantitative criteria (measurable, accessible, substantial), but using this type of segmentation might not lead to a better targeting of the customers. In this context, the disciplinary insights from the previous chapters will also be used as a basis for critically assessing the results.

## 5.2 RESEARCH STRATEGY

In terms of the research strategy an *empirical validation* approach was selected. This decision is rooted in the fact that the interdisciplinary character of this work requires an in-depth exploration of the problem, and the boundaries between the different disciplines are not clearly defined. For example, while there are clear methodologies for segmenting customers, for extracting concepts from text, and for collecting data about which programmes people watch, combining these three into a coherent methodology has not been explored previously. This can lead to practical problems that need to be analysed in detail for a given empirical study. While this type of research can be biased by the context of the study selected and difficult to generalise, the fact that most broadcasters collect data about their viewers in the same way, and store similar levels of programme metadata would suggest that the same methodology could be applied for many cases. For example, the BBC could use all the viewing statistics collected from iPlayer and collate it with the semantic analysis of the programmes' descriptions in order to infer viewers' interests. Similarly, Netflix might be able to improve their recommender engine by considering the interests extracted from the programme's descriptions as additional features. Nevertheless, further details will be provided in [Section 5.6.4](#) where the generalisability of the solution is presented, as well as in [Chapter 7](#), which provides a higher level discussion of the results.

When selecting the subject of the empirical study for this research, the following set of factors were considered:

- **Online Service:** As previously shown in [Chapter 3](#), there are three main channels through which media companies collect data about their audiences: television ratings, social media, and online video platforms. However, television ratings do not provide a relevant source of information as they are based on demographic sampling, and therefore cannot be used for viewer segmentation. Social media can be used for predicting audiences' size and the general opinion about programmes, but only some categories of content generate enough social media activity in order to be able to derive insights from it. Finally, online video platforms provide a clear picture of what every user is watching online, and therefore constitute a good source of data for interest segmentation. Based on these characteristics of the data collection methods, the subject of the study should be an organisation that collects or has access to detailed information about individual viewers for their online video service;
- **Type of Content:** From a theoretical standpoint, the techniques for extracting concepts from the description of a programme can work for all types of content. However, the method is arguably more effective for broadcasters that air a high proportion of documentaries, as these tend to be watched by viewers interested in that specific topic (Rosa et al. 2014). By comparison, news stations would not constitute an optimal fit, as news programmes are generally watched without knowing the content of the programme upfront. Similarly, dramas or comedies are considered to have lower predictive value compared to documentaries in terms of interests, but could

unveil the viewers' preferences for actors, directors, action types, or even filming locations. Based on this considerations, a broadcaster that focuses on documentaries is considered to be a better fit as subject of this study, but additional considerations related to the limitations of this approach and the further work required to generalise it will be provided in [Chapter 7](#);

- **Quality of Metadata:** Given the importance of being able to extract concepts from the programmes' descriptions, the quality of metadata available for the broadcasters analysed is highly relevant. As a general rule, as shown in [Chapter 4](#), the size of the corpus can affect the ability of algorithms to disambiguate text, since more context words improve the chances of detecting the right entity in the cases when there is more than one meaning. While there are no rules for the minimum number of words in such a situation, using richer corpus is considered advisable. In the scenario where no programme descriptions are available, the use of closed captions for the programme can lead to similar results, but this specific alternative and its implications will not be explored by this study;
- **Pragmatic Factors:** Given that the type of data required for this study is not publicly available and considerably difficult to access, the decision was also influenced by the set of available options that the researcher had with *Streamhub*<sup>1</sup>, the partner company in this research that provided some of the data analysed;

Based on the previously defined criteria, the online service of a large broadcaster was selected as subject for the study.

---

<sup>1</sup> *Streamhub* - Company providing a scalable video analytics service that delivers actionable insights for understanding and increasing audiences – <http://goo.gl/0IxADZ>

The organisation targets foreign audiences with programmes that promote the local culture and lifestyle. Given the high number of documentaries broadcast, the fact that data about individual viewers is collected by the partner company, and that some of the programmes have ample descriptions for each episode on the corporate website, the selected broadcaster was considered to represent a good fit in the context of this research.

### 5.3 RESEARCH PROCESS

In order to analyse the patterns of viewing and then segment viewers by their interests, the proposed methodology will rely on a data mining exercise. The de-facto approach for this type of work is the *Knowledge Discovery in Databases (KDD)* process (Fayyad et al. 1996), so the methodology will be structured in line with the same steps.

The process starts with understanding the objectives of the data analysis exercise and identifying or creating a dataset that has the potential to address those goals. Having the dataset available, the selection process identifies a target dataset, which can be a sample from the initial one, or contain just a subset of the total variables which are deemed most relevant for the objective. The next step involves the preprocessing of data, a process that varies from case to case, but generally involves eliminating noise from data, dealing with missing entries, or any other generic step that needs to be performed before the actual data analysis. The next two steps transform the data from the initial format into one that is adapted to the context of the analysis, and then analyse the data for patterns using appropriate

algorithms. Having identified the patterns, the last step of the process evaluates the results and emphasises the new knowledge obtained.

An overview of the described steps can be consulted in Figure 24. All the steps up to the evaluation phase will be described in this chapter, and represent the proposed methodology; while the interpretation and evaluation of results, as well as the knowledge generated, will be covered in Chapter 6 and Chapter 7. The results will be analysed for validating the proposed methodology, and the findings incorporated in order to further refine it.

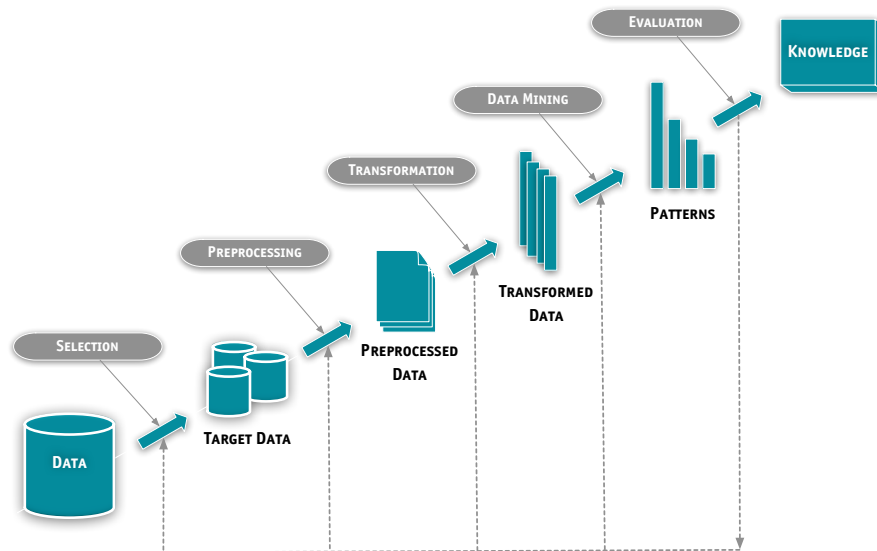


Figure 24: The methodology proposed by this study is based on a data mining exercise, and therefore structured according to the *Knowledge Discovery in Databases (KDD)* process; Source: Fayyad et al. (1996)

#### 5.4 DATA COLLECTION AND TRANSFORMATION

The first four steps in the KDD process – data collection, selection, preprocessing, and transformation – are common for all the research questions, while the rest of the steps (data mining and

evaluation) are specific to each. A detailed outline of the first four steps, grouped on two different streams (*Video Data* and *Programme Metadata*) can be consulted in Figure 25. The next subsections will describe all the relevant tasks (marked with T) and data stores (marked with DS) corresponding to each of the two streams.

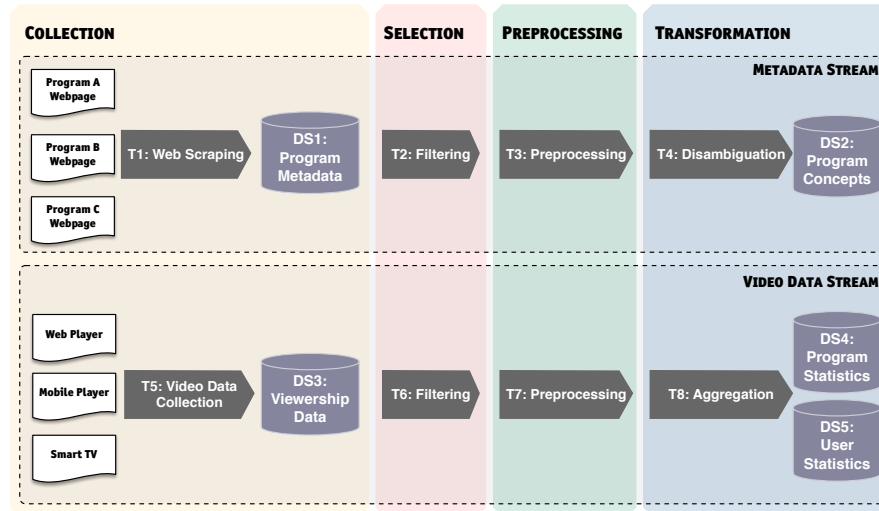


Figure 25: Overview of the *Data Collection*, *Selection*, *Preprocessing*, and *Transformation* steps of the proposed methodology. The data is altered by a sequence of *Tasks* (T) and eventually persisted in *Data Stores* (DS)

#### 5.4.1 Metadata Stream

The *Metadata Stream* contains all the relevant steps that must be taken in order to analyse programme’s descriptions and extract the relevant concepts.

##### 5.4.1.1 Data Collection (T1, DS1)

In order to detect the topics corresponding to each programme, metadata for all the programmes has to be collected and analysed. Ideally, structured data provided by the broadcaster can be used. However – in this specific case – the broadcaster could only provide short descriptions for all the programmes in structured

form. As previously mentioned, concept detection algorithms tend to perform better on large corpus, and therefore using short descriptions, typically one sentence length, would not generate the desirable output. However, the broadcaster's website contains ample descriptions for some of the programmes, and was therefore considered an appropriate data source. In order to collect this data, a basic *web scraper* tool was built on top of *GNU Wget*<sup>2</sup> – an open-source file retriever – and then used to create an offline version of all the pages on the broadcaster's site that describe programmes. Alternatively, if structured data can be obtained from the broadcaster, the following two steps can be skipped.

#### 5.4.1.2 *Data Filtering (T<sub>2</sub>)*

With the content being stored locally, the structure of the HTML code can be analysed, and the relevant tags that point to the programme's name, series' name, air date, and programme's description identified. By eliminating all the other content on the webpages, as well as scrapping the HTML tags in the text, structured data can be generated, containing only the relevant details about the programme and its description. For each resulting programme the number of words used in the description can be computed, and only the series that contain rich descriptions should be analysed.

#### 5.4.1.3 *Data Preprocessing (T<sub>3</sub>)*

Due to the fact that the data for the metadata stream is collected via web scraping – a method relying on the number of assumptions about the page's HTML structure – a certain level of manual validation and corrections need to be made. For example, while all the pages analysed describe the same type of content (a programme in this case), it is highly likely that the structure of the HTML page

<sup>2</sup> *GNU Wget* – <http://goo.gl/gd5dem>



varies from series to series. Some pages tend to use tabs to present information from different topics of each show, while others present it all in one page. All these discrepancies need to be assessed and the data extraction mechanism adapted till the required level of precision is achieved.

#### 5.4.1.4 Data Transformation ( $T_4$ , $DS_2$ )

As previously mentioned in [Chapter 4](#), over the recent years the technology for extracting entities from text has been commoditised. A large number of open source and commercial solutions are now available. Drawing on these insights, the criteria used for selecting the most appropriate one for the objective of this study was based on the following factors:

- performance of the algorithm in terms of precision and recall;
- ability to disambiguate the concepts identified in text by linking them to *DBPedia*, the crystallisation point of the *Web of Data*;
- type of license required;

Based on these factors, *AlchemyAPI* was selected as the preferred solution for detecting concepts in the programme metadata. The service is considered to be one of the most precise in the area (Hooland et al. 2013), and has the possibility to link the detected concepts to a number of ontologies including *DBPedia*, *OpenCyc*, *Freebase* and *YAGO*. Moreover, it is free to use for research purposes up to 30,000 transactions per day, and has the backing of a large company for support and maintenance, as it was recently acquired by *IBM*. For the purpose of this study, a software component was developed that uses *AlchemyAPI* to detect the list of disambiguated concepts for each programme belonging to the series selected in

the previous step. However, similar studies can rely on other components that operate in the same manner, or implement some of the algorithms described in the literature, most notably by Mihalcea et al. (2008) or Moro et al. (2014). An example of the type of data stored in DS<sub>5</sub> can be consulted in Table 2.

	CONCEPT	RELEVANCE	DBPEDIA ENTITY
PROGRAMME A	<i>19th century</i>	0.92	dbpedia.org/19th_century
	<i>Clothing</i>	0.84	dbpedia.org/Clothing
	<i>Design</i>	0.80	dbpedia.org/Design
	<i>Neo-Victorian</i>	0.73	dbpedia.org/Neo-Victorian
	<i>Science fiction</i>	0.66	dbpedia.org/Science_fiction
	<i>Fashion</i>	0.62	dbpedia.org/Fashion
PROGRAMME B	<i>Carbon dioxide</i>	0.94	dbpedia.org/Carbon_dioxide
	<i>Global warming</i>	0.54	dbpedia.org/Global_warming
	<i>Bento</i>	0.49	dbpedia.org/Bento
	<i>Japanese cuisine</i>	0.48	dbpedia.org/Japanese_cuisine
	<i>Emission standard</i>	0.46	dbpedia.org/Emission_standard
	<i>Kyoto Protocol</i>	0.37	dbpedia.org/Kyoto_Protocol

Table 2: Example rows from DS<sub>5</sub>: Each programme is associated with a list of concepts detected in the metadata along with their perceived relevance and the corresponding *DBPedia* entity

#### 5.4.2 Video Data Stream

The *Video Data Stream* contains all the tasks and data stores that are required for analysing the tracking messages received from the video players.

##### 5.4.2.1 Data Collection (T<sub>5</sub>, DS<sub>3</sub>)

In order to be able to measure how many viewers each programme has, the broadcaster selected as a subject for the empirical study works with the partner company for data collection and analysis services. For all the online platforms in use (web browsers,

mobile devices, SmartTVs), for the whole duration during which a viewer is using a video player, a tracking message is being sent every minute to the data collection web service (Figure 26). The data contained in this message includes:

- the date and time when the programme is watched (e.g. 01.06.2015 12:00);
- the unique identifier for the programme that the user is currently watching (e.g. 8647);
- the unique identifier for the player that the user is using for watching the programme (e.g. *WebFlash*, *SmartTV*, *iPad*, etc);
- a randomly generated 128 bit identifier for the viewing session (re-generated every time the player is used);
- a randomly generated 128 bit identifier for the viewer (generated the first time a viewer is using the web player and stored in a web cookie for re-use on any subsequent sessions);
- the IP address of the viewer (e.g. 10.34.24.49);

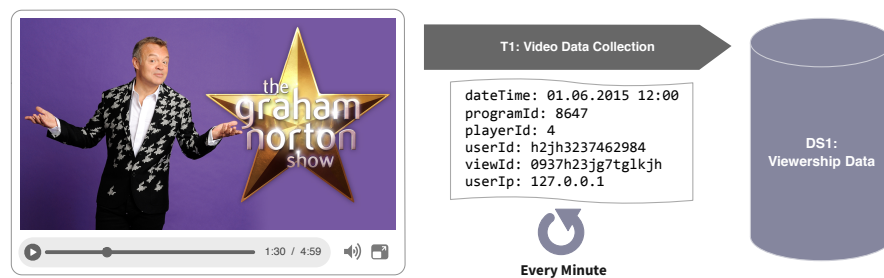


Figure 26: As long as a viewer is watching a programme on an online video platform, tracking messages are being sent every minute to the data collection service

The data was collected by the partner company for a period of 38 months, spanning from *September 2011* to *October 2014*, and it is used in this study as secondary data. The dataset is used in

raw form, has not been sampled, pre-aggregated, or tampered in any way by the partner company before it was handed off for the purpose of this research, with the exception of anonymisation. The methods for collecting viewing data using periodically sent tracking messages represent the standard in the video analytics industry, and similar systems are used by other leading market players including *ComScore*<sup>3</sup>, *Ooyala*<sup>4</sup>, *Brightcove*<sup>5</sup>, and others.

#### 5.4.2.2 *Data Filtering (T6)*

While analysing all the data collected is considered ideal, given the realities of the media market it was noticed that it is sometime the case that some programmes lack the required level of metadata. Moreover, in order to understand viewers' interests, and be able to see how these change over time, it is also deemed important that a long period of time is analysed. Combining these two criteria, only the viewing data for the series that have proper metadata and span over a considerable amount of time should be kept. In the case of this particular study, a number of four series totalling 332 episodes were selected. The details of these series can be consulted in [Table 3](#). In order to protect the identity of the broadcaster analysed, the name of the series were anonymised. While the series selected have no overlapping genres, this is not considered to introduce any bias, as the presentation of the results in [Chapter 6](#) will show that the a large number of concepts are actually shared between multiple series, irrespective of their genre.

#### 5.4.2.3 *Data Preprocessing (T7)*

While the data being used by this stream is secondary data and expected to be already cleansed, additional steps should be taken in

<sup>3</sup> *ComScore* – <http://goo.gl/0i4ZIL>

<sup>4</sup> *Ooyala* – <http://goo.gl/ViB66h>

<sup>5</sup> *Brightcove* – <https://goo.gl/0B7EKv>

SERIES	EPISODES	TIMESPAN	GENRE
<i>Series A</i>	146	36 months	Travel
<i>Series B</i>	37	14 months	Fashion
<i>Series C</i>	46	29 months	Culture
<i>Series D</i>	103	31 months	Technology

Table 3: High level details about the four series that were filtered for analysis. The selection was based on the number of episodes aired as well as the volume and quality of the metadata

order to validate the quality of the dataset. All the messages stored should be analysed for duplicates, as the same message can be sent to the data collection service twice in isolated circumstances (e.g. network partitions, faulty players, etc). All the duplicate messages should be removed, along with the ones lacking certain fields, or containing a programme or player identifier that is not matching the list provided by the broadcaster. Moreover, unusually high levels of traffic from the same IP address in a short amount of time should also be excluded from the analysis, as these can be the result of a *denial-of-service* attack, and generally do not reflect genuine traffic.

#### 5.4.2.4 Data Transformation ( $T8$ , $DS_4$ , $DS_5$ )

The objective of the transformation phase is to prepare the data so that it is in a format suitable for use by the data mining methods. Since the dataset provided is under the form of individual tracking messages – sent periodically during every user’s viewing sessions – aggregation is needed before the information can be effectively used. Given the large volume of data analysed, summing up to over 100 GB for this specific study, a *map-reduce* algorithm is recommended for running the aggregation in a distributed manner. The implementation of the algorithm used in the empirical validation was created on top of *Spark*<sup>6</sup>, a widely used large-scale data processing framework, but a

<sup>6</sup> *Spark* - a fast and general engine for large-scale data processing – <http://goo.gl/jBk6cs>

number of alternatives can be used including *Apache Hadoop* or *Apache Flink*.

In order to generate the results for the *User Statistics* data store (DS4), the messages will first be aggregated by their user identifier, and then for each user a list of programmes watched is generated, while for each programme the number of minutes watched by the user is also computed. Similarly, in order to aggregate information for the *Programme Statistics* data store (DS5), the messages will be aggregated by the programme identifiers, and then for each programme the number of unique view identifiers and viewer identifiers is computed, along with the average time watched. The structure of the data stores along with some example rows are presented in [Table 4](#) and [Table 5](#).

USER	PROGRAMME	TIME WATCHED
User 1	<i>Programme 2</i>	10 minutes
User 1	<i>Programme 5</i>	30 minutes
User 1	<i>Programme 5</i>	5 minutes
User 2	<i>Programme 38</i>	13 minutes
User 2	<i>Programme 25</i>	29 minutes
User 2	<i>Programme 242</i>	3 minutes
User 2	<i>Programme 25</i>	26 minutes

Table 4: Example rows from the *User Statistics* data store (DS4) – The tracking messages are aggregated so that the time watched for each programme by every user is calculated and persisted

## 5.5 DATA MINING

### 5.5.1 Interest Segmentation

The data mining phase for the interest segmentation related research questions (RQ1 and RQ2) is based on the data stored

	VIEWS	VIEWERS	AVG. TIME WATCHED
<i>Programme 2</i>	28,748	22,948	27 minutes
<i>Programme 5</i>	53,836	46,948	22 minutes
<i>Programme 38</i>	32,876	31,097	29 minutes
<i>Programme 242</i>	14,003	13,099	12 minutes

Table 5: Example rows from the *Programme Statistics* data store (DS5) – The tracking messages are aggregated so that the number of views, unique viewers, and average time watched can be computed for each programme

in the previous steps in DS2 (*Programme Concepts*) and DS5 (*User Statistics*). Every item in DS2 contains the details about a programme, and the list of concepts identified in the programme’s description. The concepts are disambiguated with links to the corresponding *DBPedia* entity, so that the difference between similar words can be emphasised (e.g. *bar* as establishment, or *bar* as legal association). However, the advantage of using large knowledge bases is not only limited to the role of disambiguation, but also for understanding the relationship between various concepts. For example, if one programme contains *Bali* in the list of concepts, while another programme contains *Sumatra*, using the *DBPedia Categories Ontology* one can infer that both are provinces of *Indonesia*. This is particularly relevant for the purpose of the study, as the relationships between concepts can be used to infer further interests of viewers which might not be directly apparent.

As previously shown in [Chapter 4](#), the entities in *DBPedia* are linked to three different ontologies: *DBPedia Ontology*, *YAGO Ontology*, and *DBPedia Categories*. Given their different structure and purpose, each of these could be used for expanding the list of concepts for each programme from the initial list of entities to a broader list of concepts that includes the classes related to each entity.

An example of the categories linked to the *Bali* entity in *DBPedia* can be consulted in [Table 6](#).

DBPedia Entity – Bali	
<i>DPPedia Ontology</i>	<i>dbpedia-owl:Place</i> <i>dbpedia-owl:Location</i> <i>dbpedia-owl:PopulatedPlace</i> <i>dbpedia-owl:Settlement</i>
<i>YAGO Ontology</i>	<i>yago:District</i> <i>yago:Island</i> <i>yago:Region</i> <i>yago:LesserSundaIslands</i>
<i>DBPedia Categories</i>	<i>category:Bali</i> <i>category:Islands_of_Indonesia</i> <i>category:Lesser_Sunda_Islands</i> <i>category:Provinces_of_Indonesia</i>

Table 6: All the *DBPedia* entities identified as concepts can be categorised according to three different ontologies

In order to augment the data with the additional concepts – *Stardog*<sup>7</sup> – a semantic graph database preloaded with the *DBPedia* dataset was used. For each concept associated to a programme, the *SPARQL* queries presented in [Listing 3](#) were used for identifying all the related classes according to each ontology. The relevance of these additional concepts might be different for various studies, so a certain degree of validation and exploration is required on a case by case basis. In order to validate the relevance of the concepts, this methodology adopted two different strategies:

- **Manual Validation:** While clearly impeding the efforts to fully automate the interest segmentation process, analysing a certain percentage of the detected concepts can be a very precise way of understanding which ontologies provide better relationships

<sup>7</sup> *Stardog* – Cross-platform semantic graph database – <http://stardog.com>



in the context of the study. Moreover, the emergence of crowdsourcing platforms simplified this process, by connecting companies that require this sort of repetitive work with persons willing to do it (e.g. *Amazon Mechanical Turk*, *Crowdfunder*, etc).

- **Visualisation Based:** Given that additional concepts are inferred for each programme, the expectation is that programmes from the same series, and therefore sharing the same genre, will have more concepts in common than the programmes belonging to different series. In order to visualise this relationship, the *cosine similarity* was computed between all programmes, based on the concepts they share and their relevance. With the data generated, a graph visualisation can display all the programmes as nodes, coloured based on their corresponding series, and the *cosine distance* between them being inversely proportional to their *cosine similarity*. Therefore, two nodes that are placed closer in the graph are deemed more similar based on the concepts identified. This visualisation can be helpful in understanding in which way various ontologies can expand the initial list of concepts.

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX purl: <http://purl.org/dc/terms/>

# Fetch list of DBpedia & YAGO Ontology
# classes for each concept
SELECT * where {<insert-entity> rdf:type ?o}

# Fetch list of DBpedia Categories
# for each concept
SELECT * where {<insert-entity> purl:subject ?o}
}

```

Listing 3: The SPARQL queries to be executed for each programme in order to retrieve the related classes for each ontology

Having identified the extended list of concepts for all programmes (RQ<sub>1</sub>), in order to be able to segment users by their interests, the DS<sub>5</sub> (*User Statistics*) data store generated in the previous step will be used. Every item in the table contains the unique identifier of one user and the list of the programmes he or she watched along with the time watched for each programme. In order to eliminate noise from the data, only the users that have watched at least five programmes from the total number of 332 episodes will be considered, with the additional condition that every viewing session for an episode should be of at least five minutes. This decision is rooted in the fact that analysing users that watched few programmes might not produce a good indication of their interests. In the case of the chosen broadcaster five minutes was considered relevant since all the programs had the same duration, but in the general case it is preferable for the chosen duration to be defined as a fraction of the program duration. In order to infer the affinity of viewers to concepts, all the programmes watched are considered, and the score for each concept associated to a programme is multiplied with the total time watched and the relevance of that programme. By using this formula it is possible to create an affinity matrix between viewers and concepts. The more minutes a viewer watched and the higher the relevance of the concept is for a given programme, the higher the affinity will be between the viewer and the associated concepts. The resulting structure of the matrix along with the formula used can be consulted below:

$$\begin{array}{c}
 \text{Viewer}_1 \\
 \text{Viewer}_2 \\
 \vdots \\
 \text{Viewer}_x \\
 \vdots \\
 \text{Viewer}_n
 \end{array}
 \begin{pmatrix}
 \text{Concept}_1 & \text{Concept}_2 & \dots & \text{Concept}_y & \dots & \text{Concept}_m \\
 & & & \vdots & & \\
 & & & \vdots & & \\
 & & & \vdots & & \\
 \dots & \dots & \dots & \alpha_{xy} & & \\
 & & & & & \\
 & & & & & 
 \end{pmatrix}$$

$$\alpha_{xy} = \sum_z R^{(y,z)} * T^{(x,z)}$$

where

$\alpha_{xy}$  Affinity between  $\text{Viewer}_x$  and  $\text{Concept}_y$

$z$  Identifier of programme watched by  $\text{Viewer}_x$

$R^{(y,z)}$  Relevance of  $\text{Concept}_y$  to  $\text{Programme}_z$

$T^{(x,z)}$  Number of minutes of  $\text{Programme}_z$  watched by  $\text{Viewer}_x$

After the affinity matrix between viewers and concepts has been generated, any stakeholder can easily quantify how strong is the relationship between a viewer and a concept, which can be highly predictive of his interests. There is a wide range of applications that can use this type of matrix, going from programmatic advert placing to recommendation engines. In order to understand which concepts are most relevant for the current viewer base, a ranking of the concepts sorted by the number of viewers that have a high affinity for them will be generated, and is expected that the concepts will be ranked from general to specific.

Given that the number of concepts will probably be in the order of thousands, a method to vary the granularity of such concepts is desirable for being able to fully visualise, understand, and action on

the segments. Therefore, a dimensionality reduction algorithm was used in order to reduce the affinity matrix from the size of  $UXC$ , where  $U$  is the number of users and  $C$  is a number of concepts, to a smaller  $UXS$  matrix – where  $S$  denotes the number of desired *super-concepts*. While there are various algorithms that can be used for dimensionality reduction, one of the most commonly employed is *Singular Value Decomposition* (Sarwar et al. 2000). It is based on a matrix factorisation technique that can factor an initial matrix  $M$  with  $m$  rows and  $n$  columns, into three different matrices  $U$ ,  $\Sigma$ , and  $V$ , with the following properties:

- $U$  has  $m$  rows and  $r$  columns, where  $r$  is the rank of matrix  $M$ ;
- $V$  has  $n$  rows and  $r$  columns;
- $\Sigma$  is a diagonal matrix of size  $r$ ;

From the three resulting matrices,  $U$  is the matrix that contains the affinities between all the users and the newly inferred *super-concepts*, while  $V^T$  shows the affinities between the initial concepts and the inferred *super-concepts* (Rajaraman et al. 2011). The latter will be used for understanding which are the main themes present in the newly inferred *super-concepts*. The diagonal matrix,  $\Sigma$ , contains the strength of each *super-concept*, and can be used to derive how many *super-concepts* need to be taken into consideration in order to retain most of the energy in the original matrix. An overview of the SVD factorisation can be consulted in [Figure 27](#).

Finally, the last step required for obtaining the viewers segments based on their interests is clustering. There are a series of algorithms that can be used for clustering the viewers, most notably *k-means*, but the required number of clusters is needed as an input variable. In

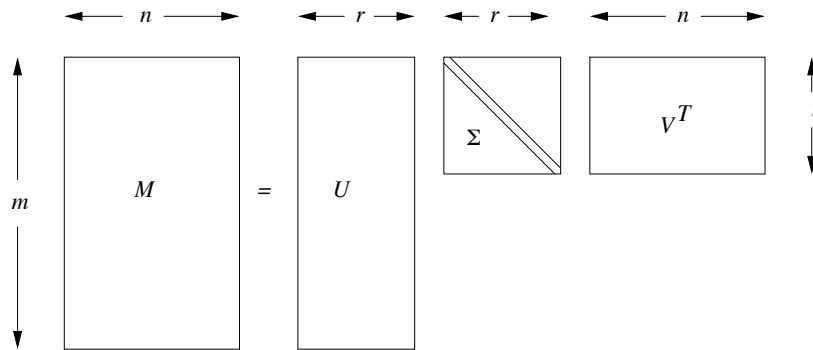


Figure 27: Overview of the *Singular Value Decomposition* factorisation;  
Source: Rajaraman et al. (2011)

order to mitigate for this problem, one of the alternatives commonly used is based on *Self-organising Maps (SOM)*. These are a type of artificial neural networks trained using unsupervised learning in order to produce a two dimensional visualisation of a multi dimensional space (Kohonen 1982; Kaski 1997). This method is typically used in similar situations, most notably in the case of census data, for segmenting a population into a smaller number of groups based on the similarity of their answers (Skupin et al. 2005). The differences between the persons in the same group (identified as cells in the visualisation) are supposed to be smaller than the differences between the groups themselves. Having generated a SOM for the input data, the information can be used to visualise and understand the different segments in the audience, and eventually cluster the cells into broader groups using *k-means*. One example visualisation of a SOM can be consulted in Figure 28, while the actual result for the empirical study will be presented in Chapter 6.

### 5.5.2 Content Performance

While the interest-based segmentation is the main research area explored by this study, having access to this level of information,

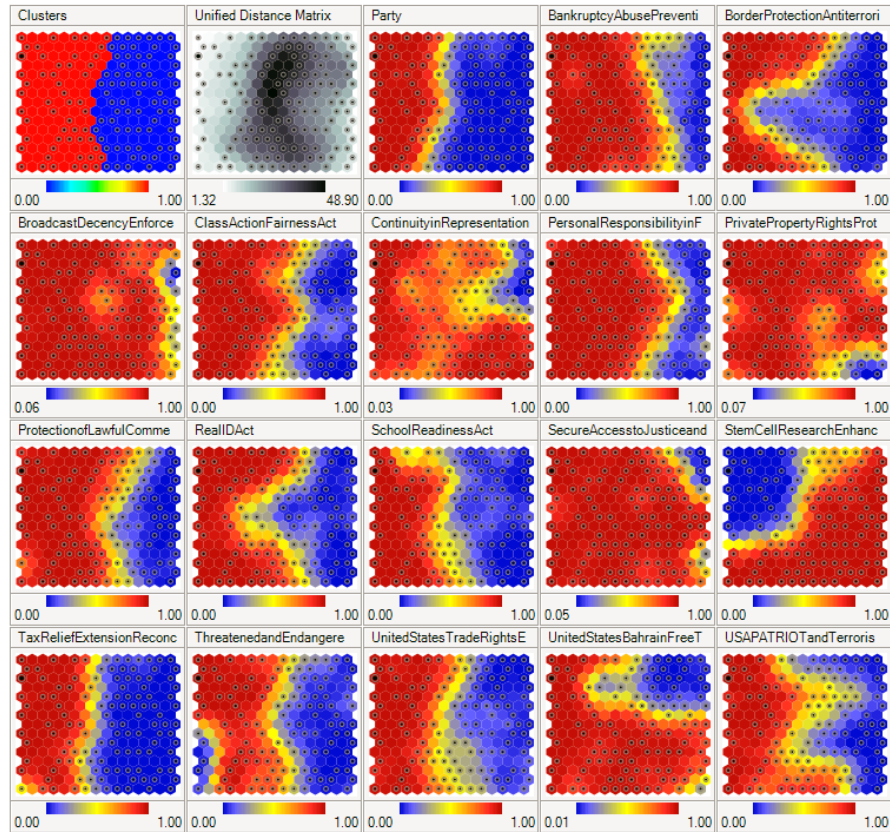


Figure 28: *Self-organising Map* example visualisation for the voting patterns in the US Congress; Source: Royall (2014)

which was previously not possible, opens the way for a new set of use cases. Understanding content performance, as discussed in [Chapter 3](#), has previously been done only at series or programme level. However, given the high correlation between viewers' interests and the decision to watch a certain programme, it would be relevant to understand what is the relation between the concepts linked to a programme and the number of views that the programme generates. In terms of the data mining step for RQ<sub>3</sub>, the data stored in DS<sub>2</sub> (*Programme Concepts*) and DS<sub>4</sub> (*Programme Statistics*) can be used for understanding this relationship.

Given the long period of time being analysed, the programme statistics data needs to be normalised so that it reflects more accurately the performance for one programme, while levelling out

the differences from the number of airings for each programme or the time of the day when it was broadcast. Moreover, in the case of many broadcasters, the number of views tends not to be constant over the years, are therefore relative values might be better for comparing the performance of programmes as opposed to absolute ones. In order to cater for that, instead of using the actual number of views generated by a programme, the performance metric used will be the percentage of views generated by a specific programme out of the total number of views for all the programmes in a two week period, spanning from one week before a programme was aired to one week after. This value will be averaged out between all the airings of a given programme using the following formula:

$$\lambda_p = \frac{\sum_{a=1}^n \frac{Views_a * 100}{\sum_{d=Day_{a-7}}^{Day_{a+7}} Views_d}}{n}$$

where

$\lambda_p$  Performance indicator for Programme P

$n$  Number of airings of Programme P

$Views_a$  Number of views of Airing a

$Day_a$  The day when a was aired

$Views_d$  The total number of views of Day d for all programmes

Having computed the performance indicator for each programme, the value can be multiplied with the relevance of the concepts associated to it, and therefore obtaining a relation between the various concepts present in programmes and their impact on the number of views. This can be helpful for identifying the underlying reasons for which any given programme had a performance above or below the average. Moreover, a ranking of the

most successful concepts can be derived and then subsequently used for creating better programmes, which are more aligned with the viewers' interests.

While being able to understand the performance impact of individual concepts is a powerful technique, there could also be added value for understanding the impact of a combination of concepts in one programme. This is relevant since some viewers might be interested in a certain topic only in a given context (e.g. audience interested in programmes about *India*, but only in regards to *culture* and *travel*, but not *cuisine*). In order to model this relationship a graph visualisation can be built where every node represents a concept, the size of the node is directly proportional to its impact factor, and the connections between the nodes are based on the number of times they co-occur in the same viewer's viewing history. The proposed visualisation will reflect the most relevant topics that have an increased number of viewers, the relationship between the topics, and the series and programmes which are related to this topics. While there are multiple solutions for implementing this visualisation, the one presented in [Chapter 6](#) is generated with *Gephi*<sup>8</sup>, an open-source graph visualisation platform.

Having a good overview of the viewers' interests and which of these generate most views, in line with the *Media Market* insights identified in [Chapter 3](#), the next step for broadcasters would be to react to the trends identified with the help of search engines or social media. For example, if at some point in time the term *San Francisco* is trending, broadcasters could understand for which of the related concepts they already have content available, and alternatively, if

---

<sup>8</sup> *Gephi* – interactive visualisation and exploration platform networks and complex systems – <http://goo.gl/SmFWmu>



it would be useful to develop new programmes related to those interests in order to satisfy a demand in the market. In order to assess this possibility, example trends coming from *Google Trends* and *Twitter Trending Topics* will be disambiguated using *Alchemy API*, and then matched to the existing concepts detected in the broadcaster's programmes.

## 5.6 METHODOLOGICAL CONSIDERATIONS

This section will present the methodological considerations of this study in terms of ethics, reliability, validity, generalisability, and limitations.

### 5.6.1 *Ethics*

From an ethics point of view, there are two main dimensions that are relevant for this study: protecting the identity and confidentiality of the persons that watched the programmes analysed, and protecting the business interests of the broadcaster that provided the data. In regards to the tracking of viewers, the online service being analysed does not require users to authenticate or provide any information in order to watch programmes online. Moreover, the cookie-based mechanism used for tracking individual users is based on a randomly generated sequence of letters and numbers that cannot be used to identify a person. While the data collection service initially stores the IP address of the users, and uses that for identifying the country from which the service is being used, the information is then removed since is considered *personally identifiable information*<sup>9</sup>. In addition to this, if a user does not want to be tracked, he has an option to do so, by either

<sup>9</sup> *Personal Identifiable Information* – <https://goo.gl/eFZYI1>

explicitly refusing to accept cookies from the broadcaster, or by using a browser in *incognito* or *private* mode, and implicitly not accepting any cookies.

In order to protect the business interests of the broadcaster used for the empirical validation, the name of the broadcaster, the series, and the programmes analysed were anonymised. Moreover, all the relevant charts displaying potentially sensitive data about the number of users do not display values on the charts' axes. In this way the general trend can be observed, while the specific values are kept confidential. Moreover, additional steps were taken so that any concepts that could uniquely identify the programme or the broadcaster being analysed were removed in the data filtering step.

Finally, in order to ensure data confidentiality, the dataset has been anonymised on the partner company's infrastructure, and only then transferred for the purpose of this research. The analysis of the data was performed on secure cloud infrastructure provided by *Amazon Web Services*. The service is compliant with the *EU Data Protection Directive (95/46/EC)* that regulates the protection of individuals in regards to the processing of personal data and the free movement of such data.

### 5.6.2 *Reliability*

For maximising the reliability of this study, the methods used for data collection and analysis were thoroughly documented, so that the same steps can be repeated by other researchers while expecting similar results. From a data collection point of view, the data has not been sampled, so the process is considered to be highly

reliable. With regards to the data analysis steps, the reliance on external tools can introduce a small variation in results. If a different study chooses to use the same web service for detecting concepts in text (*AlchemyAPI*), similar results can be expected, while a different web service can produce slightly different results. Regarding the segmentation algorithms used – given the nature of machine learning – some slight variation can be expected on a different dataset, but the objectives of the study are considered to be achievable by using the same methods as described in this methodology.

### 5.6.3 *Validity*

In regards to the validity of this study, the methodology proposed for inferring viewers' interests based on the content they watch online relies on well established and validated methods in their specific disciplines. The methods used for tracking video content are standard in the industry, being employed in the media by the major video-on-demand platforms but also for tracking the use of websites (e.g. *Google Analytics*). Similarly, the methods used for extracting concepts from text and disambiguating it by linking to an existing knowledge base are widely used, and their precision and recall documented (Hooland et al. 2013). Finally, the algorithms used for the data mining step are established methods in the machine learning field that were successfully used for similar use cases, most notably segmenting respondents of national censuses.

Since all the methods employed in this study are valid from an academic point of view, applying them in a sequence as proposed by this methodology is assumed to generate a valid result. However, in order to further validate this assumption and uncover any potential

problems, an empirical study was undertaken and the results are presented in [Chapter 6](#).

#### 5.6.4 *Generalisability*

From the point of view of generalisability, this study is focused on the case of broadcasters that air their programmes via an online video platform. This is rooted in the ability to track individual users' viewing habits, without relying on aggregated, sampled, or extrapolated data, as the assumptions required in the process might not be valid. However, in the recent years, most of the content being watched on TV can be tracked via the use of set-top boxes, SmartTVs, or mobile devices. While there are no available predictions in regards to the trends in this area, the overall direction seems to point to the fact that more video content will be tracked, with the ultimate goal of ensuring a better experience for the viewer. Given this assumption, the conclusions of this study can be applied to a wide selection of broadcasters, the only requirement being related to their ability to track individual viewers' habits.

From a content point of view, this study is deemed to be most useful for documentaries, as the link between the topic of a given programme and the actual viewers' interest is considered stronger. However, all content types can be analysed using the same methods, and different type of entities could be detected (e.g. actors names for dramas and comedies, filming locations for reality shows, topics discussed in news sections, etc). A similar technology could also be used in sectors like retail by analysing the description of products. While positive outcomes can be expected, this study did not explore this angle and chose to focus on the media sector solely. Additional

aspects related to the generalisability of this study in a broader context will be presented in [Chapter 7](#).

#### 5.6.5 *Limitations*

The limitations of this research are highly related to the generality aspects mentioned before. Ideally, different broadcasters would be analysed in order to understand how the methodology can be applied to different content types and common realities in the industry. However, accessing data from a large number of broadcasters over a long period of time was deemed too complex given the time limitations for this study. Instead, by focusing on an empirical study where the methodology is considered most useful, the relevance of the findings can be assessed and further on refined with the use of additional studies.

Another limitation is related to the use of external services for the detection and disambiguation of concepts from text. While *AlchemyAPI*, the service used in this study, is a proprietary solution and therefore used as a *black box*, alternative open source solutions performing the same task are widely available. Moreover, the algorithms behind these services are thoroughly documented in various academic papers. Further limitations identified after the empirical validation will be presented in [Chapter 7](#).

## 5.7 CONCLUSIONS

While the previous three chapters presented the state of the art from three disciplines in terms of interest-based segmentation, the methodology chapter defined the actual steps that have to

be taken in order to provide answers for the research questions. In order to do this, the choices in terms of research philosophy and strategy were presented, along with the arguments that recommended them. For each research question, all the relevant methods used for the collection and analysis of data were detailed. Finally, the methodological considerations were formulated, so that the reliability, validity and generalisability are clearly described. The next chapter will present the results of the empirical study along with considerations on how to further refine the methodology.

## RESULTS

---

The current chapter presents the results of the empirical validation of the methodology proposed in [Chapter 5](#). The results are presented according to the two major themes on which the research questions were structured: interest segmentation and content performance.

### 6.1 INTEREST SEGMENTATION

#### 6.1.1 *Number of Concepts and Relevance*

The main prerequisite for being able to infer viewers' interests is the ability to extract relevant concepts from the programme descriptions. As previously shown in [Chapter 4](#), the algorithms for *Name Entity Recognition* (NER) and *Named Entity Disambiguation* (NED) have advanced considerably in terms of precision and recall in the recent years, to the degree where they have been commoditised. However, since the majority of the studies assessing their performance have been done on standard training corpus<sup>1</sup>, additional steps were undertaken in order to understand their efficiency in the context of this study. The two main directions for the assessment were: a) the relevance of the identified concepts; and b) the number of concepts linked to each programme.

---

<sup>1</sup> *British National Corpus* (BNC) – 100 million word corpus available at <http://goo.gl/L1a024>

In terms of relevance, given the large number of programmes analysed, a randomly chosen sample of 75 episodes out of the total of 320 were manually assessed. All the identified concepts were compared with the programme description, and from a total of 520 entities linked to the episodes, 453 were considered to be highly relevant, approximately 87%. The relevance was established by assessing if a human reading the programme description would identify the concepts detected by the algorithms as important in the context of its topic – this means that the concepts have to be present in the text but also relevant to its core message. There were two main sources of errors: concepts that were misidentified (approximately 30% from the total number of errors), and concepts that were linked to entities that can have different meanings accounting for the rest of 70%. For example, the entity *Kobe* was linked to *Kobe Bryant* – a famous basketball player – instead of the one that was correct in the given context, namely *Kobe Beef*.

In addition to the manual assessment of the concepts' relevance, the similarity between all the episodes analysed was computed based on the *cosine similarity* of the concepts identified for each. If two episodes share a high number of concepts they will be deemed closer to each other. Due to the fact that all the episodes are part of four series that have different genres (*Travel*, *Culture*, *Fashion* and *Technology*), the expectation is that the distance between episodes from the same series will be shorter than between episodes from different series. In order to assess this, a graph visualisation was generated, with the nodes representing episodes, and the distance between them being directly proportional to their *cosine distance* (as described in [Chapter 5](#)). As it can be seen in [Figure 29](#), there are relatively three distinct clusters for *Fashion*, *Technology*, and *Travel*,



while the overlapping between *Travel* and *Culture* series is substantial. When analysing the actual concepts, it was noticed that there is indeed a genre overlap between the two series, as the travelling show explored the local culture and vice versa.

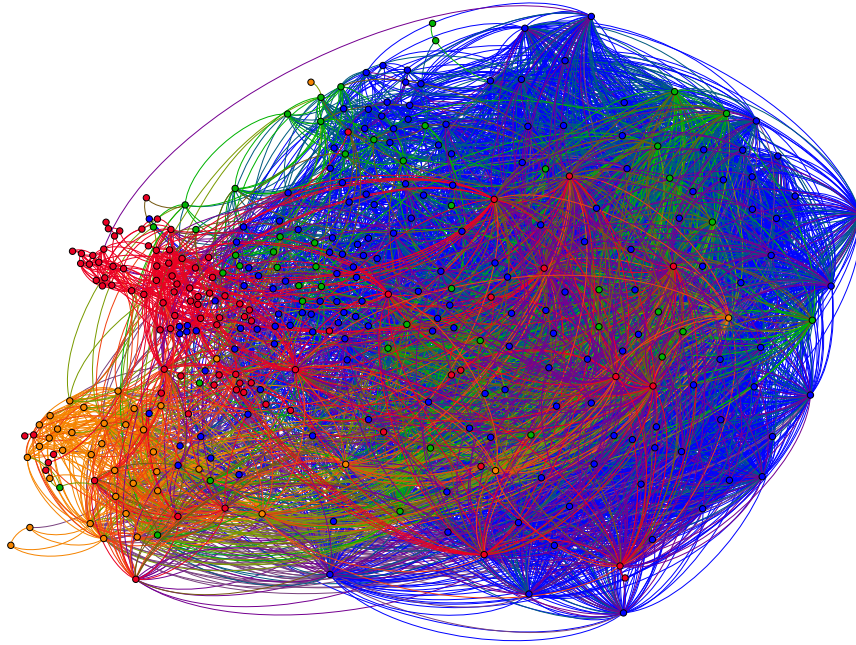


Figure 29: Episodes plotted based on their perceived similarity. Every node represents an episode, the edges are directly proportional to the cosine distance between the episodes, while the nodes are coloured based on the series they belong to; Legend: Blue - *Travel*, Orange - *Fashion*, Green - *Culture*, Red - *Technology*

This study acknowledges the fact that different persons can have different perceptions when asked to identify the relevant concepts in a textual description, and therefore it is hard to quantify the precision of such algorithms. Various studies employed similar validation techniques, involving human assessment with the help of *Turing Tests* (Mihalcea et al. 2008). However, by collating the results of the manual validation with the visualisation that resembles the four different clusters based on the series genres, it is considered that the concepts identified are precise enough to be used for predicting viewers' interests.

In regards to the number of concepts linked to each programme, based on the literature reviewed, the expectation was that a higher number of words in the description would increase the relevance and number of identified concepts. This is rooted in the fact that a low number of words would not provide enough context for the analysis. As it can be observed in [Figure 30](#), the number of words in the programmes' descriptions varied widely between 47 and 1800, with *Series A* and *Series D* having a higher number of words, while *Series B* and *Series C* are under the average value. However, when plotting the number of concepts detected against the number of words in the description, as it can be seen in [Figure 31](#) and [Table 7](#), *Series D* has the lowest average number of concepts detected and the highest standard deviation across all series. The observation is backed by the analysis of the correlation between the number of words in a programme's description and the number of identified concepts using the *Spearman Rank Correlation Coefficient*. For a total number of 332 data points, the computed value for R is -0.07865 and the two-tailed value of P is 0.17425, therefore suggesting that by normal standards the association between the two variables would not be considered statistically significant.

	Concepts (Avg.)	Concepts (St. Dev)	Genre
<i>Series A</i>	7.945	0.367	Travel
<i>Series B</i>	7.595	1.092	Fashion
<i>Series C</i>	7.674	1.175	Culture
<i>Series D</i>	6.563	2.008	Technology

Table 7: The relationship between the number of concepts detected and the genre of each series

Given the fact that text analysis algorithms were used, one possible cause for this difference could be related to the various writing styles in the programmes' descriptions. Typically, a more

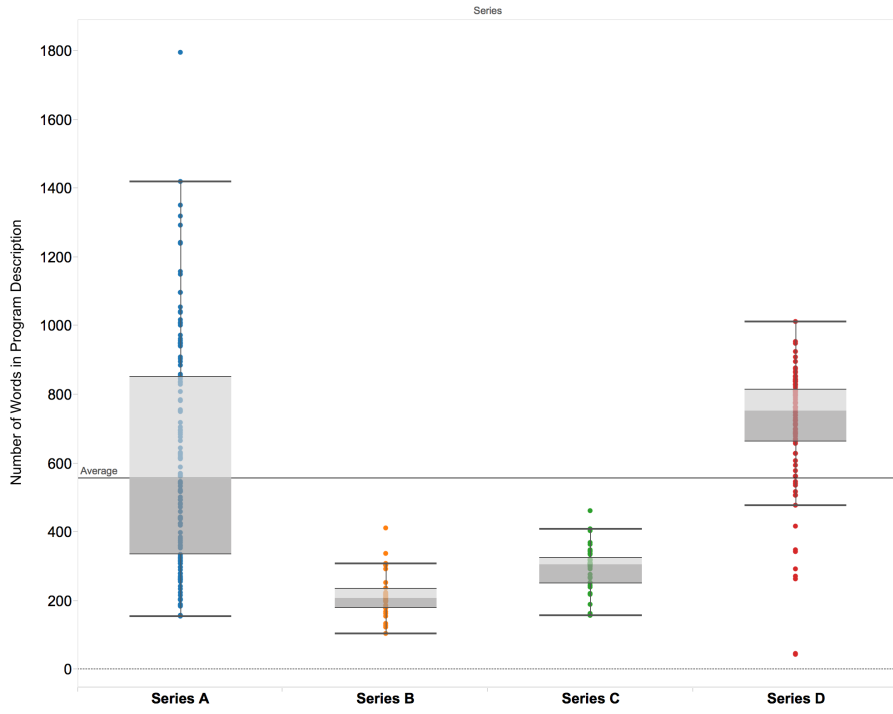


Figure 30: Number of words in the description of programmes, grouped by series

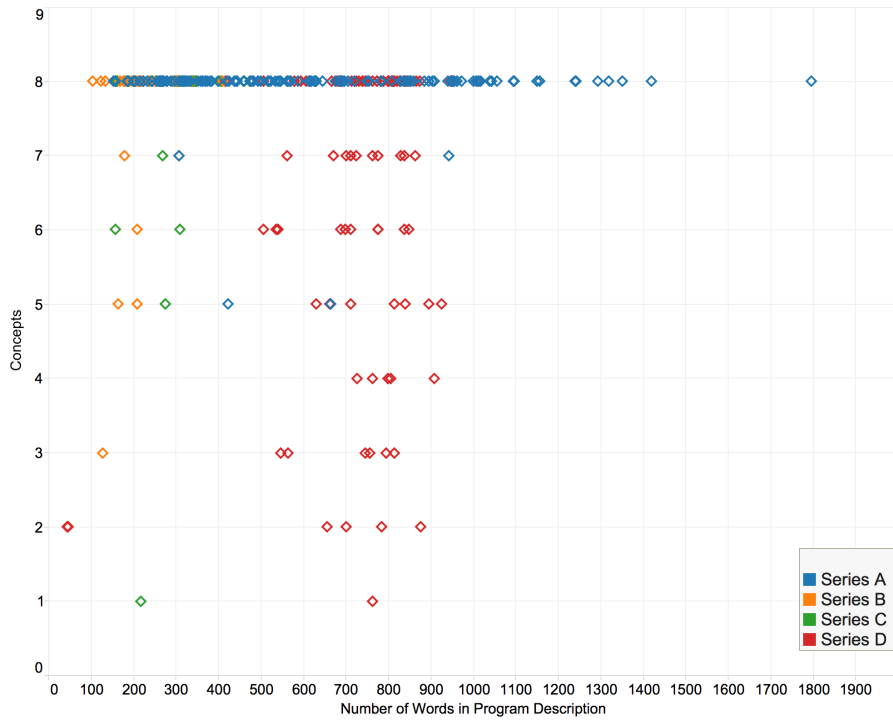


Figure 31: Number of concepts identified for each programme in relationship with the number of words in the description, coloured by series

general description would tend to generate a lower number of concepts, while a richer, more specific description would lead to the opposite effect. However, in the case of this study, all the descriptions were provided by a single source, and were expected to be relatively similar in style. A more plausible cause for the difference was linked to the genre of each series. It was noticed that the series about *traveling* and *culture* would tend to generate more concepts, as many geographic locations have corresponding entities on *DBPedia*. In the case of *Series D*, which focused on emerging technologies and gadgets, many of the entities being presented were too new to have an article already created in *Wikipedia*, and therefore lacked a corresponding *DBPedia* entity.

While this particular problem is specific to this study only, it is relevant to acknowledge that the quality of the concepts extracted from the programme descriptions can vary with the style of the writing (abstract vs. detailed) and with the genre or domain area of the series analysed. This is consistent with other authors that tested entity disambiguation techniques for various domains and concluded that their relevance varies compared to the standard training corpus (Hooland et al. 2013). Nevertheless, for the purpose of identifying viewers interests for documentary-type content, the results suggest that the technology is more accurate than the level required for statistical significance, and therefore represents a good fit.

#### 6.1.2 Choice of Ontology

Having validated that the concepts identified in the programmes' description are relevant for the purpose of this study, the next step involves the evaluation of the three different ontologies for classifying

the concepts. While all the concepts are *DBPedia* entities (e.g. *Japan*, *Sushi*, etc), using the classifications derived from the ontologies could unveil additional connections between the programmes. For example, if one programme is about *Sushi* while another one is about *Bento*, ideally the algorithm would consider the two programmes to have a degree of similarity since they both describe *Japanese cuisine* dishes. All the entities in the *DBPedia* knowledge base are classified using three distinct ontologies: *DBPedia Ontology*, *YAGO Ontology*, and *DBPedia Categories*. The number of concepts associated to each episode in [Figure 29](#) has been increased by the addition of the parent entities from the three different ontologies (method described in [Chapter 5](#)). The resulting statistics are presented in [Table 8](#).

	Concepts	Edges	Graph Density
<i>DBPedia Entities</i>	1,233	7,819	0.118
<i>DBPedia Ontology</i>	128	31,887	0.480
<i>YAGO Ontology</i>	1,895	58,810	0.885
<i>DBPedia Categories</i>	3,654	20,920	0.315

Table 8: Comparison between the number of concepts, edges in the programme similarity graph, and graph density (number of edges divided by the total edges that are possible given the number of nodes) when using different ontologies

Compared to the case when only the initial *DBPedia* entities are used, by augmenting the data with the superclasses from the ontology’s classifications increased the number of connections in the programme’s graph substantially. The number of concepts derived from each of the three ontologies varies from 128 for *DBPedia Ontology* to 3,653 for *DBPedia Categories*, while *YAGO* is in between the two values. This is probably caused by the different structure of the ontologies, as described in [Chapter 4](#). While the *DBPedia Ontology* is a high level classification derived from the types of info boxes on

*Wikipedia*, *DBPedia Categories* is a folksonomy. This means that every user of *Wikipedia* can create new categories and assign pages to them. Ideally, in order to infer viewers interests, the programmes should be highly connected in the graph based on relevant entities. The graph density measure reflects the percentage of edges between the nodes, as a proportion of the maximum number of edges possible in the given graph<sup>2</sup>. It can be noticed that in the case of the *YAGO Ontology* the percentage is almost 90%, meaning that most programmes are connected to all the other programmes based on at least one concept. In order to assess the relevance of the additional concepts derived with the three ontologies, the actual classes were compared.

Analysing the additional entities identified based on the *DBPedia Ontology*, it is clear that they are highly abstract (e.g. *Place*, *Administrative Region*, *Body of Water*, *Railway Line*, *Volcano* etc). Therefore, there is a limited value in using this classification, as given their abstract nature they define only high level concepts. *YAGO* entities by contrast are much more numerous, and provide a mix of abstract entities (e.g. *Physical Entity*, *Geographical Area*, *Building*) and more specific concepts (e.g. *Hot Spring in Japan*, *Ryukyu Islands*, *Host Cities of the Summer Olympic Games*, etc). Finally, *DBPedia Categories* seem to be the best fit for the purpose, as being user contributed they have not been automatically extracted from text, and they only contain concepts with a high semantic value (e.g. *Naturalised Citizens of Japan*, *Japanese Society*, *World Heritage Sites*, *Garden Plants of Asia*, etc). Also, the number of concepts derived using this method is high (3,654) and the density of the graph is clearly improved compared to using just the initial entities. Moreover, being user contributed, the list of categories evolves as the same pace with *Wikipedia*, reflecting

<sup>2</sup> The calculation is for 365 nodes, as there were a total of 365 episodes for which metadata was analysed. However, out of these, only 320 episodes had viewing data collected, and were then used in the segmentation

the changes in the society and perceptions. For example, in the hypothetical event of one of the countries deciding to leave the *European Union*, the *DBPedia* entity for that country will not be part of the *Member States of the European Union* category. Finally, an additional benefit of using a folksonomy is that the concepts detected are indirectly mapped to human values, as each category is reflecting someone's perception about an entity, and multiple perceptions can exist at the same time. For example, *Alan Turing* entity is both part of *British Scientists* category, but also *Residents of Maida Vale*.

Having identified the right ontology for deriving additional concepts for the programmes, a balance needs to be found for the weights of the initial concepts and the derived ones in the formulas. Building on the previous example, if two programmes are about *Sushi* their degree of similarity should be higher compared to the case when one is about *Sushi* and the other one about *Bento*. While there is no unique solution for this problem, as it mainly depends on the use case, for this study the initial concepts were weighted double in the formulas compared to the ones derived from *DBPedia Categories*. This is based on the assumption that a concept included in a programme description has a higher importance than one that is being inferred with the help of the ontology, and therefore its overall weight in the formula has to reflect that.

### 6.1.3 Viewer Segmentation

In order to infer the viewers interest for certain concepts, the data obtained in the previous section about the programmes is correlated with the viewing data already generated in DS5 (*User Statistics*). For all the viewers that watched at least five of the total 320 programmes

analysed, the affinity to each concept is computed by multiplying the number of minutes watched for a programme with the relevance of the topics associated with the programmes watched. The result of this method is a matrix where every row represents one viewer, and every column one concept. The values in the matrix are directly proportional with the strength of the affinity between a user and a concept. The computed matrix contains a total of 57,471 rows and 5,264 columns. While on average the number of programmes watched per viewer was approximately 14, the average number of associated concepts per user is around 400. The exact distributions of concepts and programmes per viewer can be consulted in [Figure 32](#) and [Table 9](#).

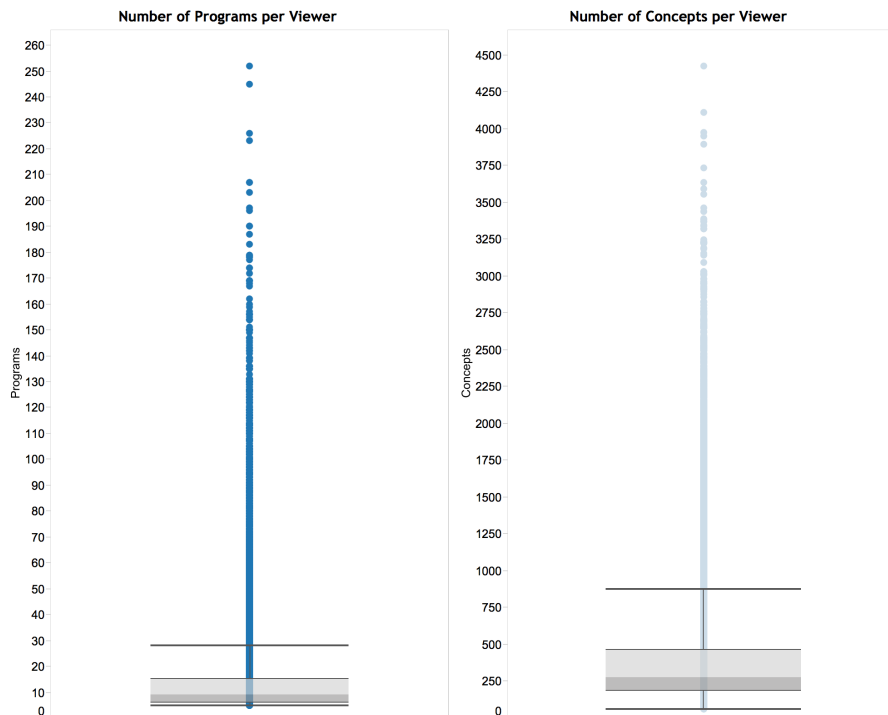


Figure 32: Number of programmes and concepts associated to each viewer

Being able to compute the matrix that relates viewers to concepts is considered a powerful technique for understanding the patterns of content consumption. For example, assuming a broadcaster wishes to



	<i>Programmes per Viewer</i>	<i>Concepts per Viewer</i>
<i>Upper Whisker</i>	28	874
<i>Upper Quartile</i>	15	462
<i>Median</i>	9	278
<i>Lower Quartile</i>	6	187
<i>Lower Whisker</i>	5	61
Average	13.80	395.60

Table 9: The statistical distribution of the number of programmes and concepts associated to each viewer

target the viewers interested in *Japanese Cuisine* with an advertising campaign, just by analysing the column in the matrix for that particular concept, all the users that previously watched content related to that topic can be identified, as well as ranked by the strength of the affinity. Similarly, assuming that a broadcaster needs to choose which advertising to display in an online video player, having identified the user currently watching, the advertisement closest to the viewers interest can be selected from a pool of options.

#### 6.1.4 Viewers Clustering

While having access to high granularity data about viewers preferences can be useful for broadcasters in some areas, like programmatic advertisement, it is relatively hard to spot the overall viewing patterns when comparing more than 5,000 concepts. Therefore, in order to be able to cluster and visualise the main segments of viewers in terms of their interests, a dimensionality reduction technique was applied in order to reduce the number of concepts. Starting from the initial matrix of 57,471 rows and 5,264 columns, the outputs of the *Singular Value Decomposition* algorithm consisted of three matrixes:

- Matrix U – containing 57,471 rows and 100 columns, reflecting the affinity of all users to 100 super-concepts;
- Matrix S – a diagonal matrix containing 100 values reflecting the strength of each super-concept;
- Matrix V – containing 5,264 rows and 100 columns, reflecting the strength of the relationship between the original concepts and the inferred super-concepts;

To further reduce the number of super-concepts, the values in Matrix S were analysed in order to compute the energy contained in the values, and only the highest five were retained, as those super-concepts generated more than 80% of the total energy. As a result of this method two matrices were obtained, one describing the relationship between the viewers and the five super-concepts, and one describing the five super-concepts in relation to the original concepts.

While the initial concepts have descriptive names that were easy to understand, the resulting super concepts reflect the general patterns in the data, but lack names or descriptions. In order to understand what they represent, the data in Matrix V was analysed, so that for each super-concept a ranking of the concepts based on the strength of the relationship can be generated. By taking into consideration the distribution of values, especially in terms of highest and lowest ranked concepts, the following descriptions were assigned to the super-concepts:

- **Travel Oriented** – This segment has a high percentage of travel related programmes. The highest ranked concepts include: *Prefectures of Japan, Populated Coastal Places in Japan, Port Settlements in Japan, Capitals in Asia, Shikoku Region, and Honshu Region*;

- **Design Oriented** – There is a high proportion of fashion, architecture and design concepts that rank high for this super concept. Some of the most notable ones include: *Arts, Design, Tokyo, Architecture Design, Clothing, Fashion, Graphic Design, and Textiles*;
- **Generalist** – While analysing the concepts associated to this segment, there is no obvious trend that emerges. This suggests that the viewers in this segment tend to be generally interested in the programmes broadcast, and do not have certain specific interest. However, when analysing the lowest ranked concepts, the data suggests that programmes related to travelling generate on average less interest than others;
- **Culture Oriented** – This segment is relatively similar to the *Generalist* one, due to the large variety of themes in highly ranked concepts. However, as opposed to the previous segment, some concepts like *Buddhism, Japanese Words and Phrases, Bathing in Japan, or Craftsmanship* suggest that the audience is slightly more focused on cultural matters;
- **Technology Oriented** - The last segment is considerably skewed towards technology and business related concepts. Some of the highly ranked entities include: *Automotive Companies, Airlines in Japan, Marine Engine Manufacturers, Wheeled Vehicles, Alternative Energy Sources*.

While four out of five super-concepts have similar names to the genres of the series, therefore suggesting a potential source of bias in the process of selecting the series, this information is derived solely from the program descriptions, while the genres of the series are not factored in. Similarly, the number of super-concepts is computed

by analysing the values in the  $S$  diagonal matrix as generated by the *Singular Value Decomposition* algorithm, in line with the standard recommendation to conserve 80% of the energy in the original matrix, and is therefore considered not to be biased. Moreover, the boundaries of the super-concepts overlap in between series, as for example some programmes in the *Travel* genre have a strong *Culture* dimension, etc.

The last step required for segmenting the viewers is the clustering in terms of affinities for super concepts. As presented in [Chapter 5](#), *Self Organising Maps* are one of the methods traditionally used for visualising high dimensional data sets into a two dimensional representation. The algorithm is first trained over a number of 100 iterations and the median distance to the closest unit assessed. As it can be observed in [Figure 33](#), after approximately 60 iterations, the performance metric stays relatively constant, therefore suggesting that there is no need for further iterations for the algorithm to converge.

In [Figure 34](#) the number of viewers assigned to each cell is presented. The distribution is relatively constant, with the exception of the central area of the map, which has higher variation. While theoretically it is desirable for the number of viewers to be as equally distributed as possible, the variation obtained in the central area is not very relevant since, as presented in the following figures, all the cells are part of the same cluster. In [Figure 35](#) the affinity between each cell and the super concepts can be visualised. Every cell contained a small chart with the components for each super concepts being directly proportional to the affinity. For example, with the *Travelling* oriented super concept being represented with green, it can be seen that only

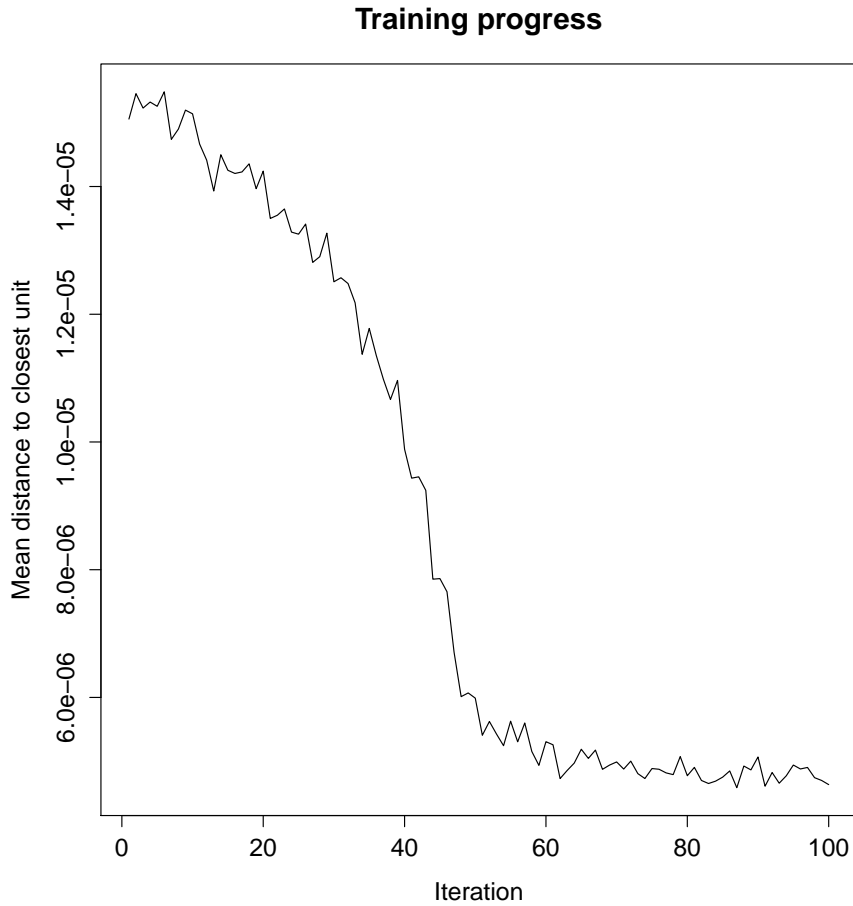


Figure 33: The relationship between the number of iterations and the median distance to the closest unit for the *Self Organising Maps* algorithm training

the last node on the second row from the top contains viewers with a high affinity to that type of content. Nevertheless, since it is relatively hard to discern the patterns between individual nodes using this chart, the next step involves clustering the nodes according to their similarity into nine main clusters, a number that was considered a good balance between minimising the squared distance, while also keeping the number of segments low. The resulting clusters can be visualised in [Figure 36](#).

Analysing the resulting clusters of nodes it is clear that most of the viewers, over 90% of the total number, are part of one cluster,

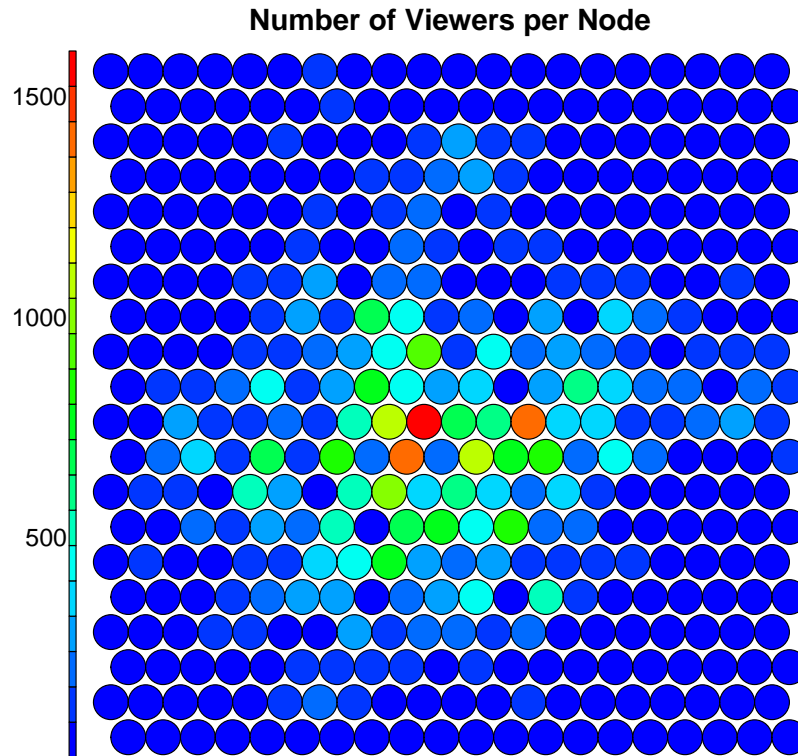


Figure 34: The number of viewers associated with each cell in the Self Organising Map

marked in blue in [Figure 36](#). They tend to watch what is being broadcast, with no specific affinities to any of the identified super concepts. This is an expected outcome given the subject of the empirical study, as the programmes analysed were broadcast live. As we previously shown in [Chapter 3](#), the correlation between interest and decision to watch a programme is deemed higher for video on demand platforms. In that case, viewers decide specifically what they want to watch, and tend not to consume content for the sole purpose of having access to it. Nevertheless, even in the case of a live broadcaster, it is noticeable that there are eight other distinctive clusters with highly specific interest areas scattered towards the edges of the map. For example the *technology* oriented cluster is marked

## Affinity Between Nodes and Super Concepts

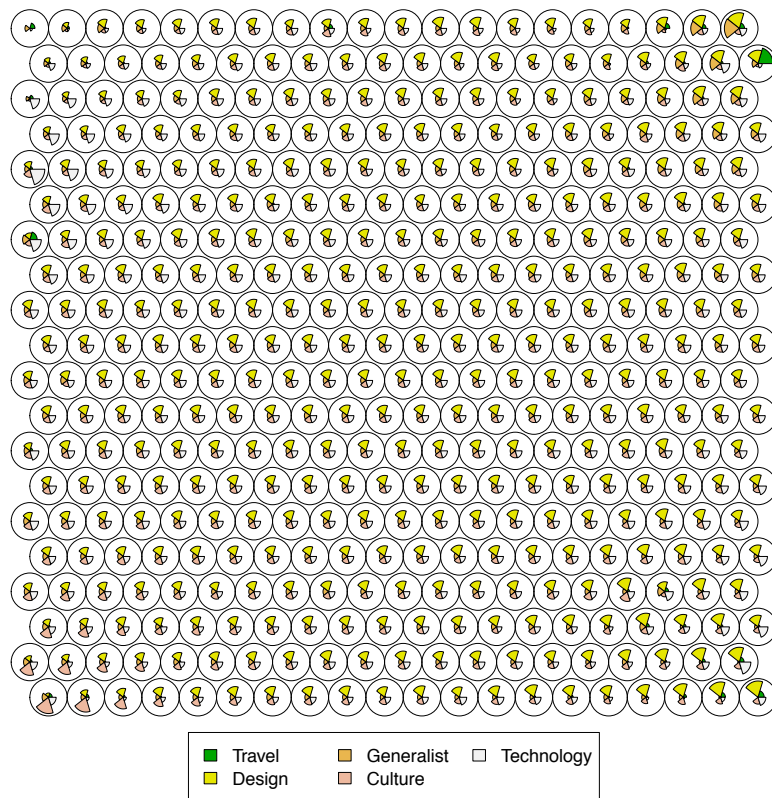


Figure 35: The affinity between each cell and the super concepts identified in the previous sections

in pink to the upper left side, the ones interested in *design* to the lower right, *travel* in the upper right section, etc. With the help of *Self Organising Maps* it is easy to visualise the existing clusters of viewers, based on their similarities in terms of interests towards a number of concepts. Moreover, for all the clusters, the skew towards one or many concepts can also be easily assessed.

### 6.1.5 Addressing the Research Questions

There were two research questions identified in [Chapter 3](#) in relation to *Interest Segmentation*:

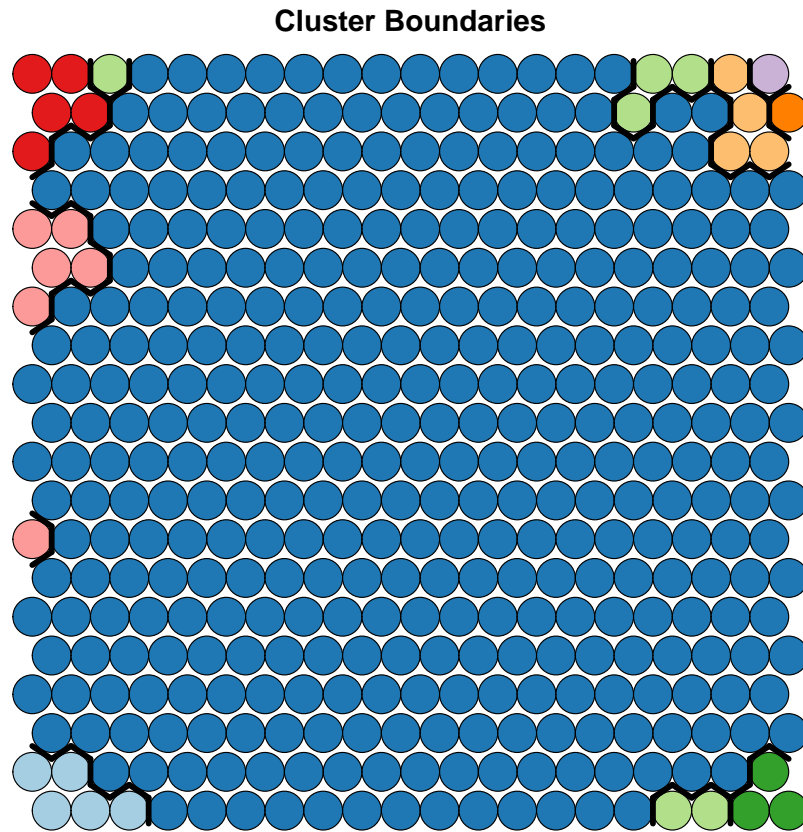


Figure 36: Cluster Boundaries – Each colour represents a distinct cluster of viewers in terms of interest affinities. The various affinities to each super-concept for individual cluster boundaries can be consulted in the subsequent diagrams

**RQ1** How can we identify a list of topics / interests for a given programme, along with their perceived relevance?

**RQ2** How can we understand the interest affinities for a viewer, and use this information to segment viewers by interest?

In regards to RQ1, the results of the study have shown that the current technologies for *named entity disambiguation* can be successfully used for extracting concepts from text and rank them according to their relevance in the given context. In order to use this approach for inferring viewers interests, a methodology was



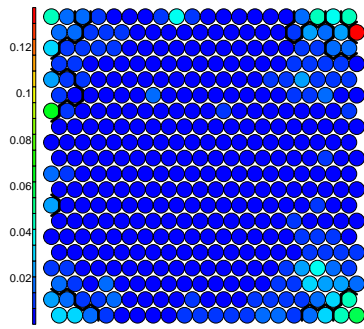


Figure 37: Travelling Affinity

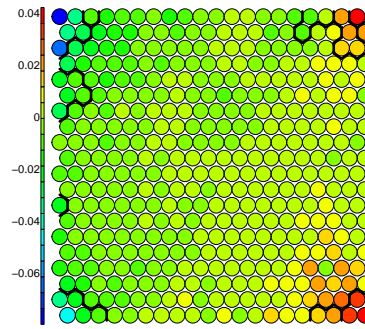


Figure 38: Design Affinity

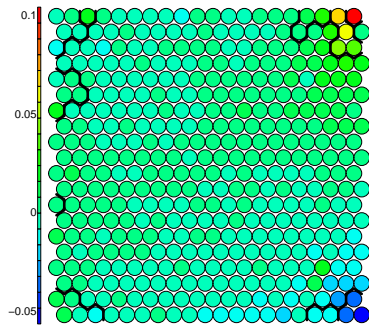


Figure 39: Generalist Affinity

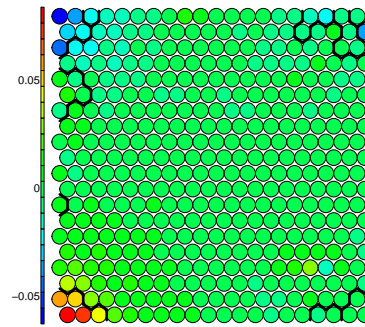


Figure 40: Cultural Affinity

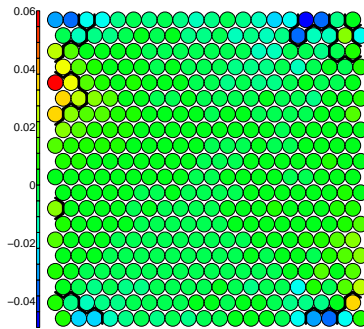


Figure 41: Technology Affinity

developed that takes into consideration the classification of the entities detected in the previous step in relation to the *DBpedia Categories*. By factoring in the categories, a constantly growing folksonomy linked to the *Wikipedia Categories*, the process is improved

by uncovering various relations between interests that are not always visible. While the process is deemed successful for the given empirical study, it is acknowledged that the quality of the concepts identified is related to the quality of the metadata about the programmes. This has implications in terms of the size of the corpus, the level of abstraction, and most importantly, the area in which the broadcaster operates. One potential alternative for reducing the dependency on the programme metadata being available is the use of closed captions, which broadcasters have available for most of their programmes due to compliance regulations. In the case where these can't be analysed, there is a wide variety of audio-to-text technologies, that could be used for obtaining the whole transcript of a show. Similarly, in regards to other domain areas where the concept detection might be less precise than in this case study, specialised ontologies for the given domain can be used in addition to *DBPedia*, a solution proposed in the context of mining forums for health related data by Issa (2015).

Building on this foundation, the answer to RQ2 provided by this study represents the main contribution to knowledge. As previously mentioned in the introductory chapters, it is widely acknowledged that the concept of *value* has different meanings to different people. For example, one product or service can be very useful to one individual, while irrelevant to another. Being able to understand to whom is a product most useful is directly related to understanding people's interests, values, and opinions. There is consensus in the marketing literature that, when choosing a segmentation criteria, interest and opinions have the highest predictive capability. However, the traditional methods for obtaining this sort of data involve complicated surveys with hundreds of questions, which are prone to be biased based on the sample chosen. However, in the media market,

the technology for collecting data about the viewers interaction with content and their viewing habits is already in place. By using the viewing data collected by broadcasters, and correlating it with the concepts identified in the textual description of the programmes, companies can now infer their viewers' interests.

Using the proposed methodology two different outcomes are achieved: being able to precisely segment viewers based on a very high number of interests, and being able to cluster them in bigger groups based on their content consumption habits. Segmenting is mainly important from the perspective of online advertising. Having the ability to understand the interests of one viewer can help a company display the most relevant advertisement from a pool of options. The success of the contextual adverts pioneered by *Google*, can now potentially be offered in the media market with the help of the proposed methodology. While the details about *Google's* algorithms are not public, it is a fair assumption that also in that space there is a shift from solely using keywords towards inferring additional information using ontologies, fact underlined by the acquisition of *Freebase* as well as increasing the *Page Rank* for pages that tag their content using *schema.org*<sup>3</sup>. Secondly, being able to cluster users into major groups based on their interests can help companies understand the overall trends, and adjust their content production or scheduling strategies in order to increase the number or quality of the viewers. The following section will explore these new angles which were previously difficult to implement without being able to infer viewers interests.

---

<sup>3</sup> Details about using *Structured Data* are available on the *Google Developers* website – <https://goo.gl/ebXBNK>

## 6.2 CONTENT PERFORMANCE

### 6.2.1 *Temporal Relativity*

As shown in [Chapter 3](#), broadcasters currently track the performance of their service at series or episode level. For example, by analysing the numbers of views, they can conclude that one series (e.g. *Top Gear*, *Newsnight*, etc) has had more viewers than another one, or that one episode was more popular compared to the rest in the same series. This information is valuable, as understanding the popularity of the show has impact in terms of advertising costs and also content creation. However, the problem with this approach is that reporting at series or programme level ignores the fact that viewers choose what to watch based on their interests. For example, one person might not be constantly watching *Top Gear*, but having an interest in *Patagonia*, he or she might decide to watch one episode that has been filmed in the region.

In order to understand which concepts generated more views, the number of viewers for all the episodes that were analysed in this study was computed. One preliminary finding was that, in some situations, simply comparing the performance of episodes based on the number of views might be biased. This is rooted in the fact that over long periods of time, given the broadcasters popularity, there can be an inflation or deflation of views. As it can be seen in [Figure 42](#), there was a considerable increase in the number of views over time for the subject of the study. In order to mitigate for this, instead of comparing the absolute values of views for a programme, the percentage of views that the programmes generated out of the total number of views for a period of one week before and after the air date

was used instead. This approach is considered a better fit because it measures the impact of various interests relative to a temporally defined baseline instead of an absolute one. For example it is more relevant to know that 30% more viewers watched a programme about the *South Pacific* in a given month compared to the average number of views in that month, instead of simply computing that the number of viewers was 12,000.

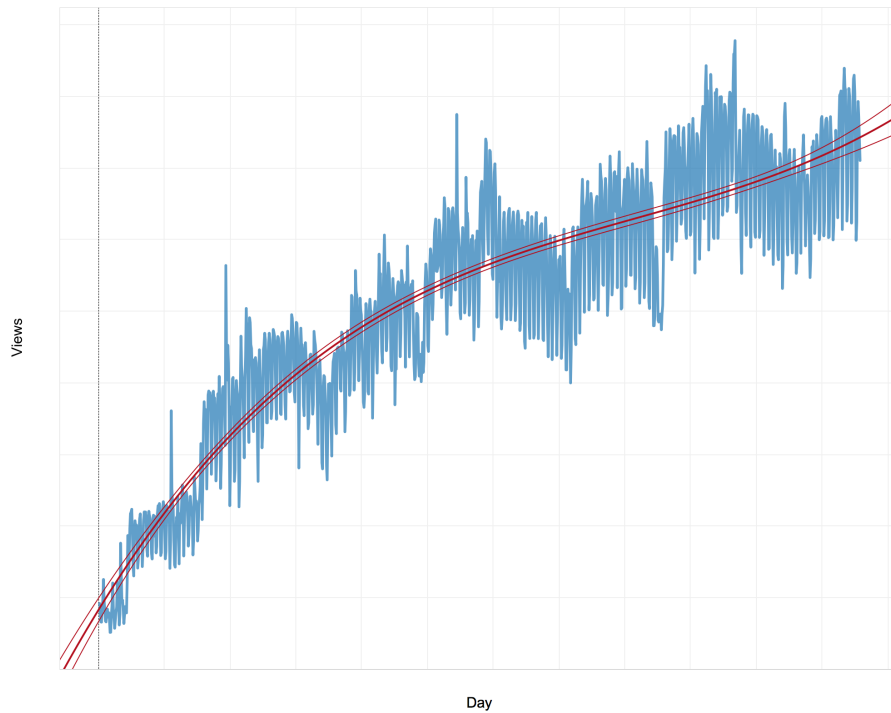


Figure 42: The number of views generated over time. The trend-line and associated confidence intervals are marked in red. In this specific case the number of views was increasing due to the deployment to multiple digital platforms, and therefore the ability to access more viewers. The re-occurring short term drops are due to seasonal changes in demand.

### 6.2.2 Visualising Content Performance

Using this calculation makes the percentages relative to the performance of the broadcaster at the given time, and the two week range helps in evening out local spikes of traffic due to unforeseen

circumstances. Having the performance indices calculated, the concepts can be plotted in a chart similar to the one in [Figure 43](#). While the chart contains just a random sample of 30 concepts, it is easy to understand which of the concepts attracted more or less viewers compared to the average. For stakeholders involved in the development or scheduling of programmes, understanding which concepts generated more viewers is highly relevant. The advantage of using this chart compared to the one showing just the performance of each episode, is that it contains insights about what actually triggered users to view a certain episode, and tracks this metric across programmes and series.

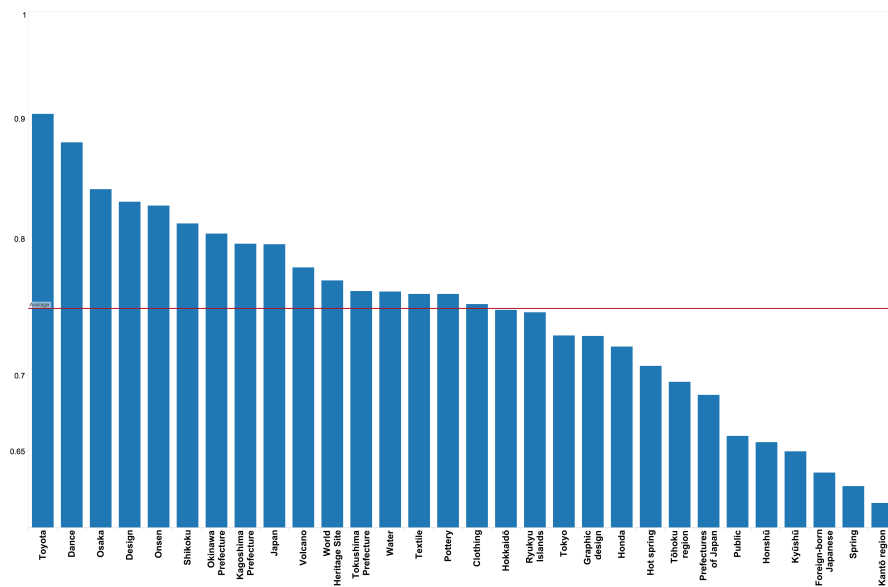


Figure 43: Concepts ranked by relevance - Random sample of 30 chosen

While [Figure 43](#) ranks the overall performance of concepts, using the same data as before, it is also possible to plot the performance by episode. However, instead of only plotting the number of views for each episode, each of the bars that represents an episode is divided based on the relevance of the concepts identified in that specific episode. For example, in [Figure 44](#), it can be observed that the first episode was mainly about *education* and also that it had less viewers



concepts appear in the same programme. For example, if in most episodes that discuss *Japanese cuisine* there is also a mention of *Sushi*, the two concepts will be displayed closer to each other compared to other unrelated concepts. Moreover, the size of the nodes is directly proportional to the performance of each concept. A subset of the described visualisation can be consulted in [Figure 45](#), while a high resolution version is included in [Appendix A](#). By analysing it, a media company stakeholder can quickly understand which are the concepts that generated most viewers, and how are these concepts related to each other. By identifying these key connections, better content can be developed, while existing content can be marketed in a better way.

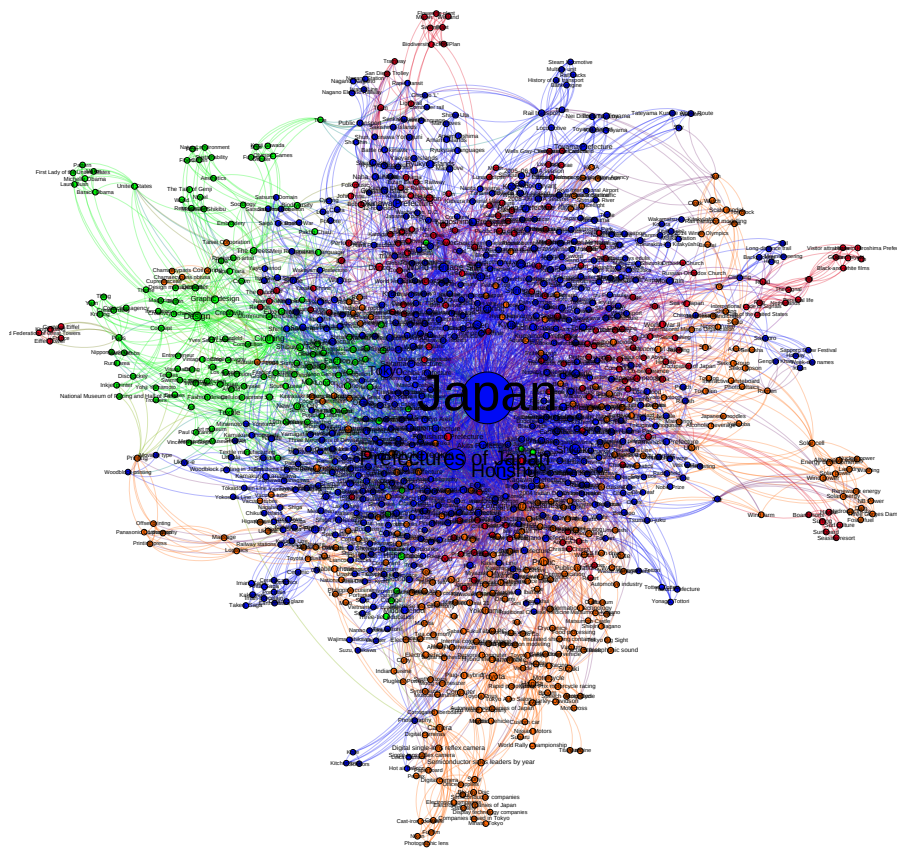


Figure 45: Concepts Performance and Co-occurrence



#### 6.2.4 *Tracking Trending Interests*

The methods previously described can help a broadcaster better understand which are the interests that their audiences have. However, in order to develop better content in the future, stakeholders need to have a good understanding of the general shift in interests over a period of time, outside the scope of their own audience. This can help to attract new viewers as well as to reduce churn. There are a variety of methods available for understanding the interests of people from different markets. Most notably, the large social networks like *Facebook* or *Twitter* provide access to *Trending Topics*, which are short descriptions of the topics mostly commented by their users. The list of topics can be accessed either in the global format, reflecting the views of all users, or for a particular region or country. Similarly, *Google* publishes *Google Trends*, a list of the most common searches for a particular country and time.

In order to match the topics trending on social media platforms or search engines with the ones present in the programmes broadcast, *Trending Tweets* and *Google Trends* were analysed for a number of days. For identifying the topics related to the trends, the same methodology proposed for detecting concepts in programmes' descriptions was used. However, both in the case of *tweets* and *Google Trends*, the results were not satisfactory, as the relevant concepts were either not identified or pointing to the wrong *DBPedia* entity. The cause of the problem was related to the insufficient context in terms of words, as in both cases the trend contained up to a maximum of 10-15 words. While there are multiple possible solutions for this problem, a *heuristic* was used in this case, as for each *Google Trend* the first page of *Google Search* results was used for detecting concepts. While not

guaranteed to work well in all situations, it was observed that the first page of results generally contains only webpages that are highly relevant for the specified trend. Given that every webpage returned contains also a small excerpt from its content, the algorithms used for detecting concepts have more context words and therefore increased precision.

Using this automated way of detecting the relevant concepts trending in social media, and then matching these with the programmes that contain the same concepts, broadcasters can more easily promote some of their videos when there is an increased interest in the market. Moreover, by analysing the trends over a long period of time, producers can understand what is likely to be demanded by the audiences in the future, and create programmes according to that.

### 6.2.5 *Addressing the Research Questions*

There were two research questions identified in [Chapter 3](#) in relation to *Content Performance*:

**RQ3** How can it be quantified what attracts more viewers in terms of content and what does not based on the interests / topics of each programme?

**RQ4** How can we develop better content by understanding which are the trending interests in social media?

In regards to RQ3, the results showed that it is possible to measure the performance at concept level. This is relevant for broadcasters due to the fact that audiences might watch a given programme not necessarily because they usually follow that series,

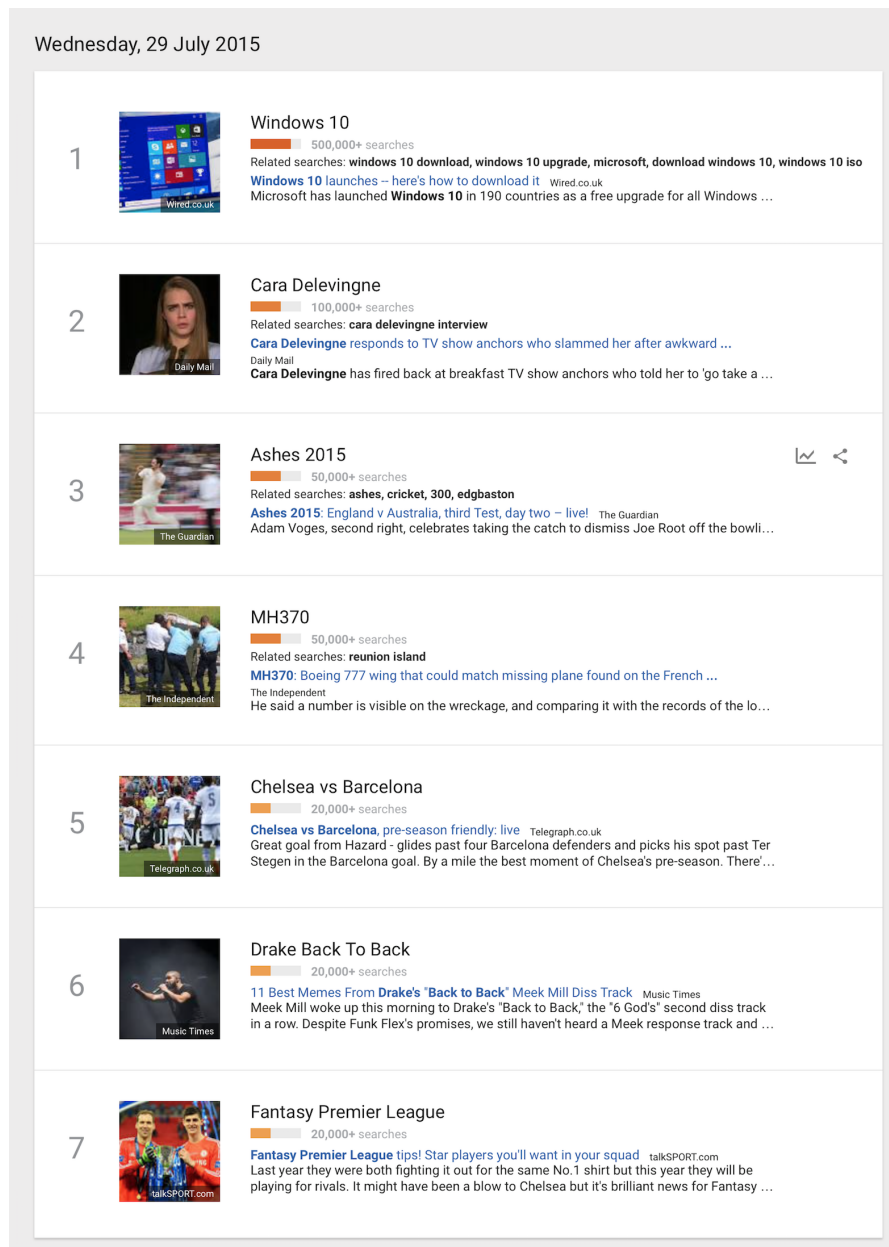


Figure 46: Example of *Google Trends* results for 29th of July 2015

but due to their interest in something present in the programme (e.g. the theme, one character, the filming location, etc). In order to simplify the process of understanding which interests are performing better than others for attracting viewers, the methodology proposed three different visualisations: *a*) a ranking of the most successful concepts based on the relative number of the views that the programmes linked to them had; *b*) a bar chart representation of

programmes where each section of the bar represents a concept linked to the programme, and is proportional to its relevance; c) a graph where all the nodes represent concepts, the size of the nodes is proportional to the concept performance, while the position of the nodes is based on the number of times the two concepts co-occur in the same programme. The methodology also proposed a heuristic for RQ4 that will allow broadcasters to detect concepts in the trending searches being published by *Google*. While the methods for doing so are not deemed precise in all instances, having a simple method to detect the concepts and link them to the existing programmes can help broadcasters to better market their existing content or develop programmes.

### 6.3 CONCLUSIONS

This chapter presented the findings from the empirical validation of the methodology. Based on the data analysed it was concluded that the number of concepts identified is not necessarily correlated to the size of the corpus, but mostly dependent on the genre of the programme. Similarly, it was shown that the *DBPedia Categories* ontology provides a good foundation for inferring additional interests based on the concepts detected, and that other folksonomies are expected to provide better value compared to machine generated ontologies. Building on these initial findings that validated the relevance of the concepts identified, the following sections validated the methodology proposed for inferring viewers interests, visualising the main audience groups, and finally tracking performance of programmes at interest level. The next chapter will provide a discussion of the findings in the wider context, while emphasising the contribution to knowledge.

Part IV

UNDERSTANDING

## DISCUSSION

---

While the previous part of the thesis focused on the integration of the disciplinary insights into a coherent methodology for interest segmentation, the *Understanding* part will provide a discussion of the results. This chapter will first present the relevance of an interdisciplinary approach in the context of this study, followed by a brief reiteration of the contribution to knowledge in contrast with existing research. The limitations of the methodology as well as its applicability to other markets will also be discussed. Finally, the importance of the findings and suggestions for further work will be presented.

### 7.1 INTERDISCIPLINARY APPROACH

While traditionally most of the research projects relied on the framework provided by individual disciplines, finding solutions to a number of complex problems required a new approach that crosses these predefined boundaries. While exploring this phenomenon, Palmer (2001) concluded that *"The real-world research problems that scientists address rarely arise within the orderly disciplinary categories, and neither do their solutions"*. This is also the case of interest-based segmentation of audiences, as the existing body of knowledge related to the area is divided between three disciplines: *Customer Segmentation, Media Market, and Large Knowledge Bases*.

However, simply gathering input from the various disciplines is generally not enough for deriving a coherent methodology. This is rooted in the fact that the assumptions, methods, and jargon of the various disciplines are in many situations not aligned, or even contradictory. This requires the researcher to first develop adequacy in each of the disciplines, derive the relevant insights, critically assess them, and finally integrate them into a new methodology. A more metaphoric description of this process is provided by Nissani (1995) that suggests that the difference between multidisciplinary research and interdisciplinary research is similar to the one between a *bowl of fruits* and a *smoothie*. While in the case of the *bowl of fruits* each fruit can be clearly identified, in the case of the *smoothie*, the individual fruits are blended in such a way that does not resemble any individual fruit, but is a result of combining the flavours in an unique way.

In order to be able to propose an interdisciplinary methodology, this study first presented the insights from the three different disciplines. These insights are not equivalent to the literature reviews in more traditional disciplinary studies, but offer a combination of basic concepts and schools of thought, relevant literature, and various insights critically assessed in the context of the problem area, all of which presented from general to specific. This structure provides the relevant foundation for understanding the ways in which each discipline operates, which are the knowledge gaps that can be filled by other disciplines, as well as the underlying reasons for the choice of research methods. Building on this foundation, a novel methodology is proposed and validated with the help of an empirical study.

Arguably, an alternative approach could have been to position the existing research into the media market discipline alone. However, the reliability of the study in this case would have been undermined, as it would lack a deep understanding of the reasons for which psychographic segmentation is superior to other types of segmentation, as well as the comprehension of the technologies used and their implications. Therefore, in line with the interdisciplinary research process, it is acknowledged that the knowledge gap in the case of this study is located in between the three disciplines. The most relevant insights from each discipline are discussed below.

#### 7.1.1 *Customer Segmentation*

From the marketing discipline angle there is a large body of knowledge related to the variables used for customer segmentation and their effectiveness. For example geographic variables are relevant for products where the demand only exists in certain areas, demographic variables perform well when a certain product is only relevant for an age or income group, etc. Following the development of CRM systems, companies were able to segment their customers based on behavioural variables, and use this information to infer the characteristics of their most profitable groups in order to target others with similar traits. However, while psychographic variables like interests, opinions and values, were consistently considered the most relevant for understanding consumer behaviour (Lin et al. 2002; Wedel 2000), the existing systems can not infer these and instead rely on small scale surveys that are both expensive and inaccurate. Alternative methods to automate this process were proposed in the retail area, most notably by Miguéis et al. (2012) that developed a method for inferring life-styles of the consumers based on the



category of the products purchased. This approach will be compared to the one undertaken by this study in the next section.

### 7.1.2 *Media Market*

The relevant research from the media market discipline focuses on the methods of collecting and analysing data for understanding audiences. Traditionally, broadcasters used to rely on *TV Ratings* measurements, which analysed the patterns of viewing for a small sample of households and extrapolated them based on demographics to the whole population. Given the inability of these ratings to precisely measure the audiences size (Taneja 2013), other studies in academia and the industry focused on analysing social media in order to predict the audience size and understand the underlying reasons for which people are watching a programme. However, while social media is a relevant data source, there is a clear skew in terms of volume for some categories of programmes, while others lack enough activity for the analysis to be able to infer any insights. While the proposed methodology is also skewed towards documentaries, these tend to be watched by people with a specific interest into a given topic, while highly popular reality shows or sports are assumed to be less predictive of people's specific interests. Moreover, using social media alone would add an additional bias in terms of the demographics of the people using various networks.

More recently, the change in video content consumption from analogue to digital platforms allowed broadcasters to track with precision the entire viewing history of the viewers on various platforms: set-top boxes, Smart TVs, mobile devices, desktop computers, etc. This technology can precisely measure the audiences'

size, but does not provide the means to segment audiences based on interests. Some attempts to do that were based on the genre of the programme, or the actors and directors associated with a programme (Hyoseop Shin et al. 2009; E. Kim et al. 2011; Soares et al. 2014), all of which will be compared to the proposed methodology in the next section.

### 7.1.3 *Large Knowledge Bases*

There is a wealth of research in the large knowledge bases area around information representation and natural language processing. Vast amounts of information can be now stored and queried easily, spanning hundreds of millions of facts about the world. Moreover, the advances in the fields of *Named Entity Recognition* and *Named Entity Disambiguation* made it possible to extract the relevant concepts from a given text and link them to an existing knowledge base like *DBPedia*. While these methods have been proven to be accurate when tested on generic corpus, their relevance in the context of this study was not established yet. Moreover, there are no details in the literature around which ontologies are more relevant for inferring viewers' interests and the methods to identify relationships between entities.

## 7.2 CONTRIBUTION TO KNOWLEDGE

### 7.2.1 *Main Finding*

The main contribution to knowledge derived from this study is a methodology for inferring viewers interests based on their consumption of video content. This can be used to segment viewers according to their interests and cluster them into a variable number of

groups, based on the level of granularity required. The methodology is built by integrating the disciplinary insights, and then validated and further refined with the help of an empirical study. In addition, a number of novel ways of presenting this information for commercial decision making have been derived. The following subsections will provide an analysis of how the proposed methodology is different when compared to the alternative methods identified in the literature.

#### 7.2.1.1 *Comparison to Survey-based Psychographic Segmentation*

There are a number of methodologies described in the literature and reviewed in [Section 2.3.5](#) for segmenting customers by their psychographic traits like interests, opinions and values. However, due to their reliance on surveys, the process of segmenting viewers is costly and can not be automated by leveraging the data that broadcasters already collect about their audiences. In addition to this, the number of interests that can be listed in a survey is limited, especially since the respondents are usually asked to rank various concepts based on their perceived importance. This represents a problem given the changes in the markets specific to the *long tail* phenomenon, as now customers tend to have a much more diverse range of specific interests. Moreover, surveys would reflect the opinions of a sample of the viewers which needs to be extrapolated to the whole user base, and therefore reducing the accuracy of the study.

By contrast, the methodology proposed by this study is leveraging the data that broadcasters already collect about their viewers. Moreover, the process can be fully automated, since it does not require any interaction with the viewers, and relevant interest affinities can be derived at individual viewer's level. This can be

useful for a diverse range of applications like programmatic advert buying, content recommendation engines or content development processes. In addition to this, the number of concepts that was derived in the empirical study was higher than 5,000, a level that offers much higher granularity than what would be possible with a survey-based study.

#### 7.2.1.2 *Comparison to Content-based Recommendation Engines*

While targeting a different overall goal, the work done in the area of content recommendation engines is also relevant in the interest segmentation field, as it usually relies on a matrix that stores the affinities between users or items and a number of features, similar to the approach undertaken by this study. The body of knowledge related to recommender systems is generally split in between collaborative filtering and content based recommendations. While the first alternative, based on the similarities between the way viewers consume content, is generally considered more precise, it lacks the ability to recommend content for new users, a problem usually referred to as the *cold start problem* (Schein et al. 2002). The content based recommenders, by contrast, analyse various aspects of the programmes and suggest similar shows. The computed similarity is in most instances either based on actors and directors, or on the genre of the programme (e.g. comedy, drama, reality show, etc).

The use of actors, directors, or filming locations is considered an improvement compared to the baseline of randomly suggesting content, but does not necessarily reflect viewers interests. For example, given that a particular viewer has rated highly a movie starring *Anthony Hopkins*, one hypothesis could be that he or she would enjoy watching any of the movies in which the actor

plays. However, this strategy severely limits the number of options considered for recommendations, and does not work for all types of content. In the case of documentaries, the importance of the actors and directors is assumed to be considerably lower when compared to the actual topic of the programme. Similarly, the use of genres for content recommendations limits the number of concepts considered to only 20 - 40, depending on the classification used. Moreover, research has shown that genres are not universally accepted, and that they tend to change over time (Hara et al. 2004). For example, *reality show* was not a genre of its own, but became one following the success of programmes like *Big Brother* or *Survival*. Given the constantly changing nature of genres as well as their specific differences from country to country, they are not considered to be good predictors of viewers' interests.

As opposed to limiting itself to genres, actors, or directors, the proposed methodology provides a way to infer the viewers affinity to a wide range of concepts. Moreover, using the relationships defined in the large knowledge bases like *DBPedia*, various connections between apparently unrelated interests can be identified. For example, while there might not be any apparent similarity between *Alan Turing* and *David Gilmour*, using *DBPedia Categories* it can be inferred that both were residents of *Maida Vale*. While some of the connections between concepts present in a viewer's history might indeed be accidental, given that a cumulative weighting technique is used, it is relatively easy to derive which are relevant and which are not. Moreover, since some of the most widely used large knowledge bases are constantly updated with the help of *crowdsourcing*, the concepts and the relationships between them are based on recent data as opposed

to the genres described before, so the process of inferring interests can be fully automated.

### 7.2.1.3 *Comparison to Keyword-based Segmentation*

There were multiple attempts to infer people's interests based on their interactions with products. The main technique utilised for this purpose was the analysis of the similarity between the keywords in the product name or description. The work done in this area is present both in market segmentation discipline and media research. However, simply using a keyword based metric is considered a simplistic approach compared to the *named entity disambiguation* alternative used by this study.

Analysing the keywords similarity results in highly specific concepts that are dependent on the wording of the descriptions. For example, two products called *milk chocolate* and *vanilla eclair* would not be deemed similar, while both are desserts. Similarly, a *pear* and an *avocado* have high fibre content, *Seychelles* and *Maldives* are both in the *Indian Ocean*, and *Anthony Hopkins* and *Sean Connery* are British-born actors. Using a combination of *Named Entity Disambiguation* and inferring additional concepts based on the *DBPedia Categories* can help uncover many additional connections that are not detected when just analysing descriptions at keyword level. Moreover, the system can also easily understand combinations of words and detect the underlying entity based on spatial and temporal factors. For example, a description of a movie that contains the words *Apple Founder* will be linked to the entity corresponding for *Steve Jobs*, while the words *Apple CEO* might direct to *Steve Jobs* or *Tim Cook* based on the time when the description was written.

Some of the other attempts to use keywords also factored in a hierarchy of concepts. In the retail sector, Hsu et al. (2012) suggested that a similarity measure could include the distance between the two products in the tree that defines the categories of products in a supermarket. For example, the similarity between *milk* and *yoghurt* is deemed stronger than the one between *milk* and *sugar*, since the first two are part of the same branch of *dairy* products. While this study also compared the use of different ontologies for classifying the data and inferring additional concepts, it was concluded that hierarchical ontologies like the ones used in the previous example are not the best fitted for deriving similarities. This is rooted in the fact that the hierarchical structures are artificial constructs with low predictive capabilities of people's interests. By contrast, based on the result of the empirical validation, it was concluded that *folksonomies* are better fitted for this task, as they map the human perception of an entity or concept to a number of alternatives. For example, the entity *Bali* can be perceived by different individuals either as a province of Indonesia, a tourist destination in the *Indian Ocean*, or an island that is predominantly *Hindu*. This ability is highly relevant in the context of understanding viewers' interests, and is therefore considered a better fit compared to rigid hierarchical structures.

### 7.2.2 Additional Findings

While the methodology for inferring viewers' interests based on the programmes they watch is the main contribution to knowledge of this study, having an affinity matrix between viewers and interests can contribute to the development of a complete new set of algorithms. For example, measuring the performance in the media sector was typically done at series or programme level. While this

can show that some series are more popular than others, it does not provide a good understanding of why that is the case. Being able to measure and visualise the performance of various programmes at concept level, as well as the relation between various concepts, as described in [Chapter 5](#), offers broadcasters a better source of insights for creating and monetising content. Similarly, the same technology for inferring viewers' interests based on the programme descriptions can also be leveraged to measure and track the popularity of various themes based on social media or search engine use. Using the methodology described by this study, broadcasters can more easily market their shows based on real time tracking of *Google* search trends, or understand the high level picture of trending topics in order to create programmes that would appeal to new viewers' interests. Finally, while not described in this study, the affinity matrix between viewers and concepts could also be used to increase the precision of recommendation engines by providing additional features for the algorithms to take into consideration, decide which adverts to display to which viewers, or even nudge consumer behaviour.

### 7.3 LIMITATIONS

In addition to the methodology's limitations described in [Chapter 5](#), following the empirical validation study, a couple more potential limitations were identified. One of these is related to the way in which broadcasters track their viewers. In the case of the platforms that require users to authenticate themselves in order to access the service, every user is assigned an unique identifier that can be then used in order to track that person across various platforms. However, in the case of platforms that do not require users to authenticate,



the method of tracking is generally based on web cookies containing unique random identifiers that are generated the first time the user is watching a programme on a given platform. However, these unique identifiers will be generated on each platform, so if a user is watching content on both a desktop computer and a mobile device for example, the tracking system will consider the traffic as originating from two distinct users. This limitation, while not specific to this methodology alone, affects all the other internet services that track their users. However, some solutions to this problem are already being actively used. Most notably *Google* offers a cross-device tracking solution that leverages the fact that many users of the *World Wide Web* authenticate with the search engine provider for using its services. Having precise information of which users use the services from which devices, *Google* can automatically link a number of devices to the same person. Other solutions that rely on heuristics based on the IP address of the user, patterns of accessing the service, and *user agents* of the devices are also available from a number of companies like *Adbrain*<sup>1</sup> or *Tapstream*<sup>2</sup>.

Another limitation identified is related to the fact that the interest affinities are computed from the perspective of one broadcaster only. For example, a broadcaster that airs documentaries about *Asian Culture* can track its users and infer their interests based on their affinity to the concepts related to the available programmes. However, while the affinities might be correctly inferred, they are based on a limited perspective of one broadcaster and might not reflect the overall reality in terms of viewers' interests. It might as well be the case that the same audience interested in *Asian Culture*

---

<sup>1</sup> *Adbrain* – Global Provider of Intelligent Cross-Device Technology Solutions – <http://goo.gl/7SqUut>

<sup>2</sup> *Tapstream* – Attribution and Cross-Device Tracking Solutions – <https://goo.gl/3AytPY>

could also be passionate about *Winter Sports*, but without being given the option to watch such a programme, it will be impossible for the broadcaster to infer this additional interest. While viewers are generally watching a certain programme or broadcaster with a preconceived expectation about the overall theme, this limitation could potentially be addressed by sharing data between various broadcasters. While this could be simpler for companies that have shared ownership, it might also be possible for TV channels that have competing business interests. In a similar way various online businesses decided to join *cookie sharing schemes* in order to be able to understand more about their users than they would normally do<sup>3</sup>. While striking the right balance between sharing critical business information and being able to analyse more data about the customers might be difficult, it is definitely a possibility that requires further consideration.

While the proposed methodology was proved successful in the case of documentaries, it is not conclusive based on the data analysed if the same range of technologies can provide a similar level of information for other types of content. Nevertheless, the documentaries market share in the UK is constantly reported in the top three genres, totalling on average only 3-5% less viewers than *dramas* and *entertainment*.<sup>4</sup> This suggests that there are a large number of viewers that consume this type of content and therefore there is enough data to be able to infer viewers interests. While the interactions between the viewers and other types of content might not generate any insights, this is also the case in the *contextual advertising* industry where not every user interaction is successfully matched with an advert. Future studies could explore the accuracy with which top

<sup>3</sup> *How Advertisers Use Internet Cookies to Track You* – “It’s rarely a coincidence when you see Web ads for products that match your interests” – <http://goo.gl/LRK1D0>

<sup>4</sup> BARB Audience Data by Genre – <http://goo.gl/uHnm7U>

level or specialised ontologies can identify the viewers' interests for certain actors, filming locations, or sport personalities, as well as the relevance of this information for targeting viewers with adverts and developing better content.

## 7.4 IMPORTANCE OF THE FINDINGS

### 7.4.1 *Broadcasters and Advertising*

As previously emphasised in [Chapter 1](#), following the emergence of the *long-tail phenomenon* and the transition from a small number of broad market segments to a wide variety of niche ones, the media market failed to adapt to the new realities. By contrast, companies in other sectors successfully leveraged the behavioural and psychographic data they collect about their customers, in order to better target them with adverts. For example *Google* is using information derived from search queries and email messages in order to understand their users' interests and needs and show them relevant adverts, using a technology called *contextual advertising*. Similarly, *Amazon* emerged as a leader in the online retail sector by being able to cross-sell additional products due to their holistic understanding of their clients' online shopping behaviour. In the social media space, *Facebook* is displaying highly relevant adverts by being able to infer people's interests from the messages they and their extended social circles post online.

The methodology proposed by this study is considered highly significant in the context described above, as it can represent an enabling technology for the media market. By leveraging the ability to segment the viewers based on their interests, broadcasters could

potentially bridge the widening gap between internet advertising and traditional media in terms of revenues. Moreover, given the large amount of behavioural data they collect about their audiences and the high correlation between personal interests and video content consumption, players in the media market could be in a better position to compete in the advertising space compared to other businesses. In addition to this, being able to infer people's interests based on the content they watch has a wide number of secondary applications like improved recommendation engines, better intelligence for content development teams, or the ability to react to social media trends.

#### 7.4.2 *Enabling Technologies Growth Prospects*

While the empirical validation of the methodology suggests that it is already accurate enough for being used by broadcasters, the prospects of its precision increasing over the coming years are highly likely. This is rooted in the fact that the methodology relies on two technologies that are undergoing an extensive growth phase: video data collection and large knowledge bases.

While traditionally broadcasters relied on *TV ratings* for understanding their audiences, the flaws in the methodology related to the sample size and extrapolation of the results based on demographics became clear over the years (Taneja et al. 2012). Following the transition from analogue to digital signal, more of the video consumption is being actively tracked. While in the beginning only the videos watched on online platforms were being tracked, more recently a lot of the programmes being watched on a TV set are also monitored. This is done either through the use of *SmartTVs*,

or with the help of the *set-top boxes* used by cable providers like *Sky*, *Virgin*, or *BT Vision*. Similarly, programmes watched on mobile devices are also tracked, either when watched online or offline during commute, etc. Being able to track all the viewers' interactions with the content, and generating accurate interest based profiles, offers broadcasters a holistic picture of their audiences. Given the recent growth of these emerging technologies, it is highly likely that the trend is going to continue, unless there will be an increased concern about viewers' privacy.

The area of large knowledge bases, which is at the very centre of the proposed methodology, is also experiencing high growth over the recent years. After many decades of incremental progress in the attempt to create a structural representation of the world's knowledge (e.g. *Cyc Project*), the emergence of the Internet represented a catalyst for the area. Individual knowledge bases like *Wikipedia*, and its associated semantic version *DBPedia*, expanded from only 500,000 articles in 2005 to almost 5 million in 2015. Moreover, the *Linked Open Data Cloud* that connects various knowledge bases grew from 12 datasets in 2007 to more than a thousand. Given the global movement of opening access to various data sources, correlated with the general trend of *crowd-sourcing* information, it is expected that the process is going to continue and more information will be made available freely online, further improving the number of concepts and relationships between concepts that can be accurately identified.

#### 7.4.3 *Human Emotions and the Perception of Value*

One of the main challenges when segmenting viewers by their interests is the relative nature of the process. While there

were some attempts to create taxonomies for people's interests, like the ones being used by *Facebook* or *Twitter* for their advert targeting mechanisms, it is considered highly unlikely to achieve wide consensus for such a structure. This is strongly connected to the ways human emotions influence the perception of value. For example, the same product or experience can have a different value to different individuals given their interests or personal context. At the same time, the perception of the same entity (e.g. a product, a location, a person, etc) can be influenced by someone's interests. The proposed methodology is considered to be a step forward in this direction, as it infers viewers interest based on a *folksonomy* (*Wikipedia Categories*), a collaborative approach where people can classify various entities according to their own perception of it. In this case, contradictory or overlapping classifications simply represent different perceptions, but are both deemed valid points of view. This method is considered to produce better results compared to the use of rigid classifications, as it maps viewers behaviour to a highly comprehensive and continuously evolving list of human generated values and interests.

#### 7.4.4 *Tracking and Privacy Concerns*

Finally, the importance of the findings of this study need to be analysed in the context of their implications in terms of privacy and the way markets function. While similar tracking mechanisms are used in other markets, specifically for contextual advertising, in some instances it might not always be clear for the individuals how the data collected about themselves is being used. For example, some companies choose to offer certain products to the consumer at no cost (e.g. *Gmail* from *Google*, *Free-to-air* TV channels, etc) in

order to collect information about individuals which is subsequently used for displaying adverts. Other companies choose to charge the use of their product, while still collecting information in order to optimise their services (e.g. *Netflix* collects behavioural data in order to improve their recommendation engine). In any instance, as long as there is complete transparency regarding the ways the data is collected, analysed, used, or passed to 3rd parties, the contractual relation between consumers and the service provider is deemed to be fair.

The findings of this study are considered to be a novel way to better connect the supply and demand for products and services. This is considered to be a win-win situation for the viewer, the broadcaster, and the companies that purchase advertising space. From the viewers perspective, watching adverts can be tolerated as it is subsidising the costs of producing video content. In this context, being targeted with adverts that are relevant to their interests should create an overall better experience when compared to seeing random adverts. However, as a prerequisite for this whole ecosystem to work properly, is the sharing of personal data. This data can indeed be exploited in various way by the companies, not always clearly defined, and is at risk of being exposed if the security of the systems is compromised by an attack. While the philosophical implications of a data sharing economy are not part of this study, it does represent an important area to look further into in order to ensure that strong processes are in place wherever personal data is being collected.

## 7.5 RECOMMENDATIONS FOR FURTHER WORK

Following the empirical validation of the proposed methodology, two distinct areas were identified as potential recommendations for further work: extracting concepts from the closed captions of programmes, and inferring additional connections between interests using the *Linked Open Data Cloud*.

### 7.5.1 *Extracting Concepts from Closed Captions*

In the methodology described in [Chapter 5](#), the entities that are mapped to potential viewers interests are extracted from the programmes' description. While in the empirical validation the method is considered viable, as the number and relevance of the detected entities is significant, relying on a summary of the programme introduces a layer of subjectivity in the analysis. For example, some series might have more ample and detailed descriptions, while others brief ones. Similarly, the style of writing might be different from case to case, in some instance being highly abstract while in others more concrete. In order to mitigate for these situations, it would be desirable if further work would assess the possibility of analysing the whole content of the programme as opposed to the description only.

One potential way of doing this is by using closed captions. In many countries this data is already available given compliance rules, as broadcasters need to cater for hearing-impaired people. If this is not the case in some markets, one of the many available algorithms or commercial solutions for *Speech to Text* processing could alternatively be used. In addition to the potential improvements in



terms of precision and objectivity for extracting concepts, having a better understanding of the moment in time where certain entities appear in a programme could then be used to derive further metrics. For example, during a news programme, by being able to map certain time ranges with the relevant concepts, broadcasters could infer which concepts determine viewers to stop watching or increased the number of viewers.

### 7.5.2 *Identifying Relationships in the Linked Open Data Cloud*

In addition to the entities directly identified in the programmes' descriptions, the proposed methodology is also inferring additional concepts based on the *DBPedia Categories* ontology. As shown previously, this is considered useful for identifying the underlying similarities between entities. However, while the subsumption relation (IS – A) used in the methodology is probably the most useful, further work could analyse how other type of relations could identify connections between interests. Given the large number of knowledge bases that are interconnected in the *Linked Open Data Cloud*, starting from a *DBPedia* entity, it is possible to retrieve additional facts about it from other datasets. By increasing the number of facts known about an entity, further similarities between the viewers interests could then be mapped.

### 7.5.3 *Generalising the Methodology*

While the proposed methodology for inferring people's interests has been framed and validated only in the media space, the same set of technologies can potentially be used to infer interests in any scenario where there is a trackable interaction between an individual

and a product or service that has a textual description. For example, companies might be able to infer newspapers' readers interests by analysing who's reading what and which are there comments. Similarly, the retail sector might be able to better understand their customers behaviour by analysing their shopping history and inferring facts from the products descriptions (e.g. tendency to buy organic products, high chances of the customer having a newborn, etc). Ideally, by collating data from a number of studies from different sectors, an ontology describing people's interests, both specific for a given industry and a general one, could be built.

The applications of being able to infer people's interests are not only limited to the business space. Understanding how public opinion is formed, how a certain message is actually perceived by the recipient, and what can determine an individual to become interested in a given topic has major implications in politics. For example political parties could more easily understand which are the key messages in an election campaign and what their electorate is interested in. Similarly, government bodies could try to use the same technology for nudging behaviour, as it is the case with the campaigns set up to reduce the number of smokers or increase the level of tax collection.

## CONCLUSIONS

---

This study proposed a novel methodology for inferring viewers' interests based on the video content watched. This can be used for a better segmentation and targeting of audiences, improved recommendation engines, and also a more advanced source of information for developing content. In summary, the conclusions drawn from this research are as follows:

1. Some of the complex problems and their solutions are sometime not situated in the boundaries of one discipline. While various insights and research methods from individual disciplines have applicability in other areas, a solid understanding of each viewpoint and rigorous process to integrate the insights is required in order to mitigate for the potential unconscious bias;
2. From the wide range of variables that can be used for segmenting and targeting audiences, psychographic variables like interests and opinions are the most effective as they are strongly related to the underlying factors that shape consumer behaviour;
3. While the media market has been in a privileged position to monetise their large audiences by better targeting them with relevant adverts, the failure to do so translated into reduced revenues from advertising at the expense of other players in the field. While the causes for this trend are complex, the cognitive inertia related to the broadcasters' use of TV Ratings, a system

that did not experience much change during the last 70 years, is considered to be a deciding factor;

4. Most of the digital platforms for video content consumption (e.g. SmartTVs, set-top boxes, mobile devices, computers) are now able to track their viewers. At the same time, the advances in the field of *Large Knowledge Bases* make it possible to detect concepts from the programmes' descriptions. Using a combination of these two emerging technologies it is possible to infer individual viewers' interests with a high level of accuracy;
5. While the concepts detected in text are useful for inferring viewer's interests, the use of the *Wikipedia Categories Ontology* can improve the process, as it represents a novel way to translate from one concept into a range of human perceptions of it, and indirectly identifying the underlying cause for a viewers' interest into a number of apparently disparate concepts;
6. The interest segmentation data that can be directly used by algorithms (e.g. programmatic advert buying) is fundamentally different than the one that media executives can adopt in their processes. Therefore, various techniques for reducing the number of concepts detected and visualising the main segments identified in the audiences have to be employed. This research adopted a combination of *Singular Value Decomposition*, *Self Organising Maps*, and graph visualisations for achieving this objective, but multiple alternative methods for dimensionality reduction, clustering, and visualisation can be used. When compared to alternative clustering algorithms like *K-means* or *Power Iteration Clustering*, *Self Organising Maps* is considered a better fit due to the fact that in addition to clustering it also

provides a method for easily visualising highly dimensional spaces;

7. Finally, while the methodology proposed by this research is considered to generate more value for the viewers, broadcasters, and companies advertising, it is paramount that strong processes are in place in order to warrant for complete transparency regarding the ways in which the personal information of the viewers is collected, analysed, and transferred between the relevant parties.

## BIBLIOGRAPHY

---

- Adams, William Jenson. 1994. "Changes in Ratings Patterns for Prime Time Before, During, and After the Introduction of the People Meter." *Journal of Media Economics* 7 (2): 15–28. ISSN: 08997764.
- Adomavicius, Gediminas, and Youngok Kwon. 2007. "New Recommendation Techniques for Multicriteria Rating Systems." *IEEE Intelligent Systems* 22 (3): 48–55. ISSN: 15411672.
- Aguirre, DeAnne, Leila Hoteit, Christine Rupp, and Karim Sabbagh. 2012. *Empowering the Third Billion: Women and the World of Work in 2012*. Technical report. Booz & Company.
- Anagnostopoulos, Aris, Andrei Z. Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. 2007. "Just-in-time Contextual Advertising." In *Proceedings of the sixteenth ACM conference on Conference on Information and Knowledge Management - CIKM '07*, 331. New York: ACM Press.
- Anderson, Chris. 2008. *The Long Tail: Why the Future of Business is Selling Less of More*. 267. Hachette Books. ISBN: 9781401309664.
- Baecke, Philippe, and Dirk Van den Poel. 2009. "Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data." *Journal of Intelligent Information Systems* 36 (3): 367–383. ISSN: 09259902.
- Bailey, Christine, Paul R. Baines, Hugh Wilson, and Moira Clark. 2009. "Segmentation and Customer Insight in Contemporary Services Marketing Practice: Why Grouping Customers Is No Longer Enough." *Journal of Marketing Management* 25 (3-4): 227–252. ISSN: 0267257X.
- Barletta, Marti. 2006. *Marketing to Women: How to Increase Your Share of the World's Largest Market*. 2nd ed. 325. Kaplan Business. ISBN: 9781419520198.
- Beane, T.P., and D.M. Ennis. 1987. "Market Segmentation: A Review." *European Journal of Marketing* 21 (5): 20–42. ISSN: 0309-0566.
- Berners-Lee, Tim. 2000. *Semantic Web on XML*. Accessed September 25, 2015. <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html>.
- . 2006. *Linked Data - Design Issues*. Accessed September 25, 2015. <http://www.w3.org/DesignIssues/LinkedData.html>.
- Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. "Linked Data - The Story So Far." *International Journal on Semantic Web and Information Systems* 5 (3): 1–22. ISSN: 1552-6283.

- Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. "DBpedia - A crystallization point for the Web of Data." *Journal of Web Semantics* 7 (3): 154–165. ISSN: 15708268.
- Bjur, Jakob. 2009. "Transforming Audiences: Patterns of Individualization in Television Viewing." PhD diss., University of Gothenburg. <http://gupea.ub.gu.se/handle/2077/21544>.
- Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. 2008. "Why Database Marketing." In *Database Marketing*, 18:13–46. International Series in Quantitative Marketing 4. New York, NY: Springer New York. ISBN: 978-0-387-72578-9.
- Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. "Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, 1247. New York, NY: ACM Press. ISBN: 9781605581026.
- Boulding, William, Richard Staelin, Michael Ehret, and Wesley J. Johnston. 2005. "A Customer Relationship Management Roadmap: What Is Known, Potential Pitfalls, and Where to Go." *Journal of Marketing* 69 (4): 155–166. ISSN: 0022-2429.
- Bourdon, Jérôme, and Cécile Méadel. 2014. *Television Audiences Across the World*. Palgrave Macmillan. ISBN: 9781137345103.
- Bult, J. R., and T. Wansbeek. 1995. "Optimal Selection for Direct Mail." *Marketing Science* 14 (4): 378–394. ISSN: 0732-2399.
- Carreras, Xavier, Lluís Màrquez, and Lluís Padró. 2003. "A Simple Named Entity Extractor Using AdaBoost." In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, 4:152–155. Edmonton, Canada: Association for Computational Linguistics.
- Cheng, Yu Hsuan, Chen Ming Wu, Tsun Ku, and Gwo Dong Chen. 2013. "A predicting model of TV audience rating based on the Facebook." In *International Conference on Social Computing (SocialCom)*, 1034–1037. Alexandria, VA: IEEE. ISBN: 9780769551371.
- Chon, Gina. 2006. "Car Makers Court Two Generations." *Wall Street Journal* (New York). <http://www.wsj.com/articles/SB114712690858247084>.
- Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita. 2013. "A Framework for Benchmarking Entity-Annotation Systems." In *International World Wide Web Conference*, 249–259. ISBN: 9781450320351.

- Curry, David J. 1992. *The New Marketing Research Systems: How to Use Strategic Database Information for Better Marketing Decisions*. 1st. 432. Wiley. ISBN: 978-0471530589.
- Daiber, Joachim, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. "Improving Efficiency and Accuracy in Multilingual Entity Extraction." In *Proceedings of the 9th International Conference on Semantic Systems I-SEMANTICS '13*, 121–124. ISBN: 9781450319720.
- Dibb, Sally, and Lyndon Simkin. 1997. "A program for implementing market segmentation." *Journal of Business & Industrial Marketing* 12 (1): 51–65. ISSN: 0885-8624.
- Dickson, Peter R, and James L Ginter. 1987. "Market Segmentation, Product Differentiation, and Marketing Strategy." *The Journal of Marketing*, 51 (2): 1–10.
- Dolnicar, Sara, Roman Freitag, and Melanie Randle. 2005. "To Segment or Not to Segment? An Investigation of Segmentation Strategy Success Under Varying Market Conditions." *Australasian Marketing Journal* 13 (1): 20–35. ISSN: 14413582.
- Duboff, R S. 1992. "Marketing to maximize profitability." *The Journal of business strategy* 13:10–13. ISSN: 0275-6668.
- Duggan, Maeve, and Aaron Smith. 2014. *Social Media Update 2013*. Technical report. Pew Research Center. <http://pewinternet.org/Reports/2013/Social-Media-Update.aspx>.
- Fayyad, Usama, and Ramasamy Uthurusamy. 1996. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17 (3): 37–54. ISSN: 00010782.
- Feigenbaum, Edward. 1980. *Expert systems in the 1980s*. Technical report. Stanford University. <https://saltworks.stanford.edu/assets/vf069sz9374.pdf>.
- Ferrucci, David, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T. Mueller. 2013. "Watson: Beyond Jeopardy!" *Artificial Intelligence* 199-200:93–105. ISSN: 00043702.
- Galbraith, John Kenneth. 2007. *The New Industrial State*. 1st ed. 576. Princeton University Press. ISBN: 9780691131412.
- Gold, Harry. 2009. "Hypertargeting Registered Users." *Marketing News and Expert Advice*. <http://www.clickz.com/clickz/column/1710063/hypertargeting-registered-users>.
- Gómez-Pérez, A. 1999. "Ontological engineering: A state of the art." *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence* 2 (3): 33–43.
- Green, Paul E. 1977. "A new approach to market segmentation." *Business Horizons* 20 (1): 61–73. ISSN: 00076813.



- Greenberg, Marshall, and Susan Schwartz McDonald. 1989. "Successful Needs/Benefits Segmentation: A User's Guide." *Journal of Consumer Marketing* 6:29–36. ISSN: 0736-3761.
- Gruber, Thomas R. 1995. "Toward Principles for the Design of Ontologies." *International Journal of Human-Computer Studies* 43 (5-6): 907–928. ISSN: 10715819.
- Gupta, Piyush, and O. P. Gandhi. 2013. "Ontological modeling of spatial shaft-position knowledge for steam turbine rotor." *International Journal of System Assurance Engineering and Management* 4 (3): 284–292. ISSN: 0975-6809.
- Haley, Russell I. 1968. "Benefit Segmentation: A Decision-Oriented Research Tool." *Journal of Marketing* 32 (3): 30–35.
- Hara, Yumiko, Yumiko Tomomune, and Maki Shigemori. 2004. "Categorization of Japanese TV viewers based on program genres they watch." In *User Modelling and User-Adapted Interaction*, 14:87–117.
- Hawkins, Del I, Don Roupe, and Kenneth A Coney. 1981. "The influence of geographic subcultures in the United States." *Advances in Consumer Research* 8 (1): 713–717.
- Hill, Shawndra. 2014. "TV Audience Measurement with Big Data." *Big Data* 2:76–86. ISSN: 2167-6461.
- Hilton, Richard. 2013. *Talking About My Generation: Exploring the Benefits Engagement Challenge*. Technical report. Barclays. [https://wealth.barclays.com/employer-solutions/en%7B%5C\\_%7Dgb/home/research-centre/news/talking-about-my-generation.html](https://wealth.barclays.com/employer-solutions/en%7B%5C_%7Dgb/home/research-centre/news/talking-about-my-generation.html).
- Hiziroglu, Abdulkadir. 2013. "Soft computing applications in customer segmentation: State-of-art review and critique." *Expert Systems with Applications* 40 (16): 6491–6507. ISSN: 09574174.
- Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenu, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. "Robust disambiguation of named entities in text." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 782–792. Edinburgh, Scotland. ISBN: 978-1-937284-11-4.
- Hooland, S. van, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. 2013. "Exploring entity recognition and disambiguation for cultural heritage collections." *Literary and Linguistic Computing*: 1–18. ISSN: 0268-1145.
- Hsieh, N C. 2004. "An integrated data mining and behavioral scoring model for analyzing bank customers." *Expert Systems with Applications* 27:623–633. ISSN: 0957-4174.

- Hsu, Fang Ming, Li Pang Lu, and Chun Min Lin. 2012. "Segmenting customers by transaction data with concept hierarchy." *Expert Systems with Applications* 39 (6): 6221–6228. ISSN: 09574174.
- Hussain, Mahmood, Susan Cholette, and Richard Castaldi. 2007. "Determinants of wine consumption of US consumers: an econometric analysis." *International Journal of Wine Business Research* 19:49–62. ISSN: 1751-1062.
- Hwang, H, T Jung, and E Suh. 2004. "An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry." *Expert Systems with Applications* 26:181–188. ISSN: 0957-4174.
- Issa, Ahmad. 2015. "A method for ontology and knowledge base assisted text mining for diabetes discussion forum." PhD diss., University of Warwick. <http://webcat.warwick.ac.uk/record=b2812650%7B~%7DS1>.
- Jain, Dipak, and Siddhartha S. Singh. 2002. "Customer lifetime value research in marketing: A review and future directions." *Journal of Interactive Marketing* 16 (2): 34–46. ISSN: 10949968.
- Jovanovic, Dragana. 2014. "Age of Hyper, Micro and Nanotargeting." In *Economic and Social Development: Book of Proceedings*, 408–417. Varazdinâ: Varazdin Development / Entrepreneurship Agency.
- Kahle, Lynn R. 1983. *Social values and social change: adaptation to life in America*. 324. Praeger. ISBN: 9780030639098.
- Kamakura, Wagner A. 2008. "Cross-Selling: Offering the Right Product to the Right Customer at the Right Time." *Journal of Relationship Marketing* 6 (3-4): 41–58. ISSN: 1533-2667.
- Kaski, Samuel. 1997. "Data exploration using self-organizing maps." PhD diss., Helsinki University of Technology. <http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.30.4343%7B%5C%7Drep=rep1%7B%5C%7Dtype=pdf>.
- Kim, Eunhui, Shinjee Pyo, Eunkyung Park, and Munchurl Kim. 2011. "An automatic recommendation scheme of TV program contents for (IP)TV personalization." *IEEE Transactions on Broadcasting* 57 (3): 674–684. ISSN: 00189316.
- Kim, Jonghyeok, Euiho Suh, and Hyunseok Hwang. 2003. "A model for evaluating the effectiveness of crm using the balanced scorecard." *Journal of Interactive Marketing* 17 (2): 5–19. ISSN: 10949968.
- Kim, Su-Yeon, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. 2006. "Customer segmentation and strategy development based on customer lifetime value: A case study." *Expert Systems with Applications* 31:101–107. ISSN: 09574174.

- Kohonen, Teuvo. 1982. "Self-organized formation of topologically correct feature maps." *Biological Cybernetics* 43 (1): 59–69. ISSN: 0340-1200.
- Kotler, Philip. 1980. *Principles of Marketing*. Edited by Englewood Cliffs. 291–309. New Jersey.
- Kotler, Philip, and Kevin Lane Keller. 2011. *Marketing management*. 14th. 816. Prentice Hall. ISBN: 9780132102926.
- Kulkarni, Sayali, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. "Collective annotation of Wikipedia entities in web text." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 457–466. New York. ISBN: 9781605584959.
- Lee, J, and S Park. 2005. "Intelligent profitable customers segmentation system based on business intelligence tools." *Expert Systems with Applications* 29 (1): 145–152. ISSN: 09574174.
- Lin, Davis, Laxman Narasimhan, and Jun He. 2011. *Understanding China's Digital Consumers*. Technical report. McKinsey China. [http://www.mckinseyonmarketingandsales.com/sites/default/files/pdf/understand%7B%5C\\_%7Dchina%7B%5C\\_%7Ddigital%7B%5C\\_%7Dconsumers.pdf](http://www.mckinseyonmarketingandsales.com/sites/default/files/pdf/understand%7B%5C_%7Dchina%7B%5C_%7Ddigital%7B%5C_%7Dconsumers.pdf).
- Lin and Chin-Feng. 2002. "Segmenting customer brand preference: demographic or psychographic." *Journal of Product & Brand Management* 11 (4): 249–268. ISSN: 1061-0421.
- Lindsay, Robert K, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. 1993. "Dendral: A Case Study of the First Expert System for Scientific Hypothesis Formation." *Artificial Intelligence*, no. 61: 209–261.
- Marsh, Elaine, and Dennis Perzanowski. 1998. "MUC-7 Evaluation of IE Technology: Overview of Results." In *Proceedings of the seventh message understanding conference (MUC-7)*. 20.
- McCarthy, John. 2007. *What is Artificial Intelligence?* Accessed April 13, 2015. <http://www-formal.stanford.edu/jmc/whatisai/whatisai.html>.
- McCarty, John A, and Manoj Hastak. 2007. "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression." *Journal of Business Research* 60:656–662. ISSN: 01482963.
- Miguéis, V.L., A.S. Camanho, and João Falcão e Cunha. 2012. "Customer data mining for lifestyle segmentation." *Expert Systems with Applications* 39 (10): 9359–9366. ISSN: 09574174.
- Mihalcea, Radu, and Andras Csomai. 2008. "Linking Documents to Encyclopedic Knowledge." *IEEE Intelligent Systems* 23 (5): 233–241. ISSN: 1541-1672.

- Milne, David, and Ian H. Witten. 2008. "Learning to link with Wikipedia." In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, 509–518. ISBN: 9781595939913.
- Mitchell, Susan. 1995. "Birds of a Feather Flock Together." *American Demographics* (February): 40–48.
- Moro, Andrea, Alessandro Raganato, Roberto Navigli, Dipartimento Informatica, and Viale Regina Elena. 2014. "Entity Linking Meets Word Sense Disambiguation: a Unified Approach." *Transactions of the Association for Computational Linguistics* 2:231–244.
- Nilsen, Eivind, and Øystein Sandvik. 2015. *Global Entertainment and Media Outlook*. Technical report. PricewaterhouseCoopers. <http://www.pwc.com/gx/en/industries/entertainment-media/outlook.html>.
- Nissani, Moti. 1995. "Fruits, Salads, and Smoothies: A Working Definition of Interdisciplinarity." *The Journal of Educational Thought* 29 (2): 121–128.
- O'Reilly, Tim. 2008. *Freebase Will Prove Addictive*. Accessed September 25, 2015. <http://radar.oreilly.com/2007/03/freebase-will-prove-addictive.html>.
- Owen, Bruce, and Steven Wildman. 1992. *Video Economics*. 384. Harvard University Press. ISBN: 9780674937161.
- Palmer, Carole L. 2001. *Work at the Boundaries of Science: Information and the Interdisciplinary Research Process*. Boston: Kluwer Academic Publishers. ISBN: 1402001509.
- Payne, Adrian, and Pennie Frow. 2005. "A Strategic Framework for Customer Relationship Management." *Journal of Marketing* 69 (4): 167–176.
- Peleja, Filipa, Pedro Dias, Flávio Martins, and João Magalhães. 2013. "A recommender system for the TV on the web: integrating unrated reviews and movie ratings." *Multimedia Systems* 19 (6): 543–558. ISSN: 0942-4962.
- Peltier, James W., and John A. Schribrowsky. 1997. "The use of need-based segmentation for developing segment-specific direct marketing strategies." *Journal of Direct Marketing* 11 (4): 53–62. ISSN: 0892-0591.
- Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. 2006. "Semantics and Complexity of SPARQL." In *The Semantic Web - ISWC 2006*, edited by Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora M Aroyo, 4273:30–43. ISBN: 978-3-540-49029-6.
- Plummer, Joseph T. 1974. "The Concept and Application of Life Style Segmentation." *Journal of Marketing* 38 (1): 33–37.

- Proulx, Mike, and Stacey Shepatin. 2012. *Social TV: How Marketers Can Reach and Engage Audiences by Connecting Television to the Web, Social Media, and Mobile*. 272. Wiley. ISBN: 9781118167465.
- Raaij, W. Fred van, and Theo M.M. Verhallen. 1994. "Domain-specific Market Segmentation." *European Journal of Marketing* 28:49–66. ISSN: 0309-0566.
- Rajaraman, Anand, and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. 67:328. Cambridge: Cambridge University Press. ISBN: 9781139058452.
- Repko, Allen F. 2011. *Interdisciplinary Research: Process and Theory*. 2nd. 544. SAGE Publications. ISBN: 978-1412988773.
- Rigby, Darrell K, Frederick F Reichheld, and Phil Schefter. 2002. "Avoid the four pitfalls of CRM." *Harvard Business Review* 80 (0202): 101–109.
- Rodriquez, Kepa Joseba, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. "Comparison of Named Entity Recognition tools for raw OCR text." In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, 2012:410–414. Vienna. ISBN: 385027005X.
- Rokeach, Milton. 1973. *The nature of human values*. 438. New York: The Free Press.
- Rosa, Maria De, and Marilyn Burgess. 2014. *Learning from Documentary Audiences: A Market Research Study*. Technical report September. Ontario, CA: Hot Docs. <http://www.hotdocs.ca/i/learning-from-documentary-audiences>.
- Royall, Emily. 2014. *City Science & Spatial Planning: Is the City Alive?* Accessed September 25, 2015. <http://colabradio.mit.edu/city-science-spatial-planning-is-the-city-alive/>.
- Russel, Stuart, and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach*. 3rd. 1152. Prentice Hall. ISBN: 9780136042594.
- Sarwar, Badrul, George Karypis, Joseph Konstan, and John Reidl. 2000. *Application of Dimensionality Reduction in Recommender Systems - A Case Study*. Technical report. Minneapolis: Army HPC Research Center. <http://files.grouplens.org/papers/webKDD00.pdf>.
- Schein, Andrew I., Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. "Methods and Metrics for Cold-Start Recommendations." In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, 253. Sigir. New York, New York, USA: ACM Press. ISBN: 1581135610.
- Schewe, Charles D., and Geoffrey Meredith. 2004. "Segmenting global markets by generational cohorts : Determining motivations by age." *Journal of Consumer Behaviour* 4 (1): 51–63. ISSN: 1472-0817.

- Schmitter, Thomas, and James Rosen. 2005. *Contextual Advertising System*. Google. <https://www.google.com/patents/US20050033771>.
- Shih, Clara. 2010. *The Facebook Era: Tapping Online Social Networks to Market, Sell, and Innovate*. 2nd. 368. Boston: Pearson Education. ISBN: 9780137085125.
- Shin, Hyoseop, Minsoo Lee, and Eun Y. Kim. 2009. "Personalized digital TV content recommendation with integration of user behavior profiling and multimodal content rating." *IEEE Transactions on Consumer Electronics* 55:1417–1423. ISSN: 00983063.
- Shin, Hyunjung, and Sungzoon Cho. 2006. "Response modeling with support vector machines." *Expert Systems with Applications* 30:746–760. ISSN: 09574174.
- Shortliffe, Edward H, Randall Davis, Stanton G Axline, Bruce G Buchanan, C Cordell Green, and Stanley N Cohen. 1975. "Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system." *Computers and Biomedical Research* 8 (4): 303–320.
- Silverstein, Michael J, and Neil Fiske. 2003. "Luxury for the Masses." *Harvard Business Review* 81 (4): 48–57.
- Singh, Sarwant. 2014. "Women in Cars: Overtaking Men on the Fast Lane." *Forbes* (May). <http://www.forbes.com/sites/sarwantsingh/2014/05/23/women-in-cars-overtaking-men-on-the-fast-lane/>.
- Singleton, Alexander D., and Seth E. Spielman. 2014. "The Past, Present and Future of Geodemographic Research in the United States and United Kingdom." *The Professional Geographer* 66 (4): 558–567. ISSN: 0033-0124.
- Skupin, Andre, and Ron Hagelman. 2005. "Visualizing Demographic Trajectories with Self Organizing Maps." *GeoInformatica* 9 (2): 159–179.
- Smith, Wendell R. 1956. "Product differentiation and market segmentation as alternative marketing strategies." *The Journal of Marketing* 21 (1): 3–8.
- Soares, M., and P. Viana. 2014. "TV Recommendation and Personalization Systems: Integrating Broadcast and Video On demand Services." *Advances in Electrical and Computer Engineering* 14 (1): 115–120. ISSN: 1582-7445.

- Sutter, Robbie De, Mike Matton, Niels Laukens, Dieter Van Rijsselbergen, and Rik Van De Walle. 2011. "Establishing a Customer Relationship Management between the Broadcaster and the Digital User." In *The 7th International Conference on Digital Content, Multimedia Technology and its Applications*, 185–187. ISBN: 978-1-4577-0473-4.
- Taneja, Harsh. 2013. "Audience Measurement and Media Fragmentation: Revisiting the Monopoly Question." *Journal of Media Economics* 26:203–219. ISSN: 0899-7764.
- Taneja, Harsh, J. G. Webster, E. C. Malthouse, and T. B. Ksiazek. 2012. "Media consumption across platforms: Identifying user-defined repertoires." *New Media & Society* 14 (6): 951–968. ISSN: 1461-4448.
- Tseng, Mitchell M., and Jianxin Jiao. 2007. "Mass Customization." In *Handbook of Industrial Engineering*, 684–709. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Valette-Florence, Pierre. 1986. "Les démarches de styles de vie : concepts, champs d'investigation et problèmes actuels." *Recherche et Applications en Marketing* 1 (1): 93–110.
- Wakamiya, Shoko, Ryong Lee, and Kazutoshi Sumiya. 2011. "Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs." In *Database Systems for Adanced Applications*, edited by Jianliang Xu, Ge Yu, Shuigeng Zhou, and Rainer Unland, 6637:390–401. Hong Kong: Springer. ISBN: 978-3-642-20243-8.
- Webber, Richard. 1977. *An Introduction to the National Classification of Wards and Parishes*. Technical report. London: Centre for Environmental Studies. <https://www.iser.essex.ac.uk/research/publications/509345>.
- Webster, James G., Patricia F. Phalen, and Lawrence W. Lichty. 2013. *Ratings Analysis: Audience Measurement and Analytics*. 4th. 344. Routledge. ISBN: 9780415526524.
- Wedel, Michael. 2000. *Market segmentation: Conceptual and methodological foundations*. 2nd. 382. ISBN: 0792386353.
- Wind, Yoram. 1981. "Issues and Advances in Segmentation Research." *Journal of Marketing Research* 15 (3): 317–337.
- Winslow, George. 2014. "The Measurement Mess." *Broadcasting & Cable* (New York). <http://www.broadcastingcable.com/news/news-articles/measurement-mess/114643>.
- Young, Shirley, Leland Ott, and Barbara Feigin. 1978. "Some Practical Considerations in Market Segmentation." *Journal of Marketing Research* 15 (3): 405–412.

A

HIGH RESOLUTION FIGURES

---



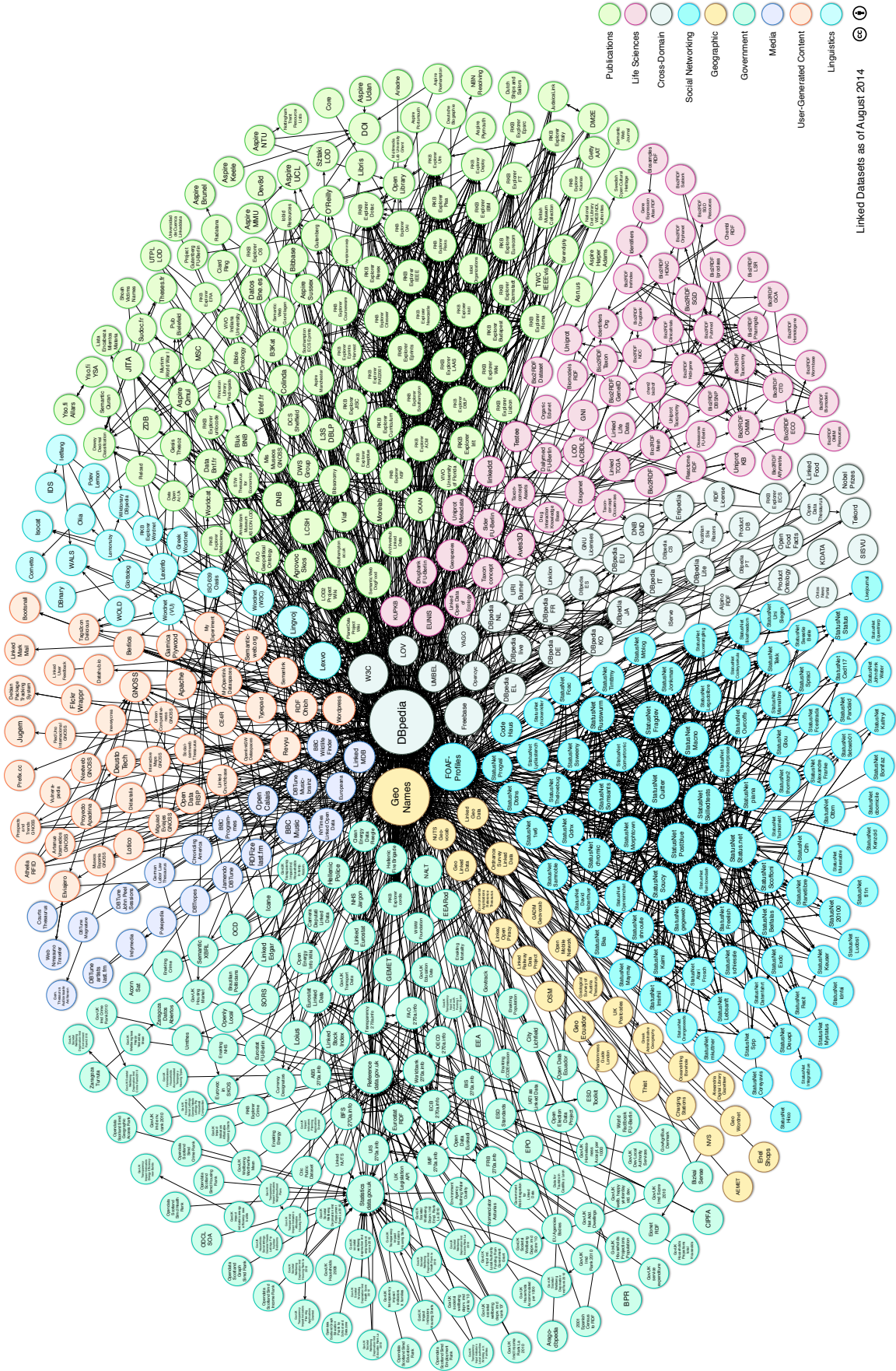


Figure 47: Linked Open Data Cloud – Present Version

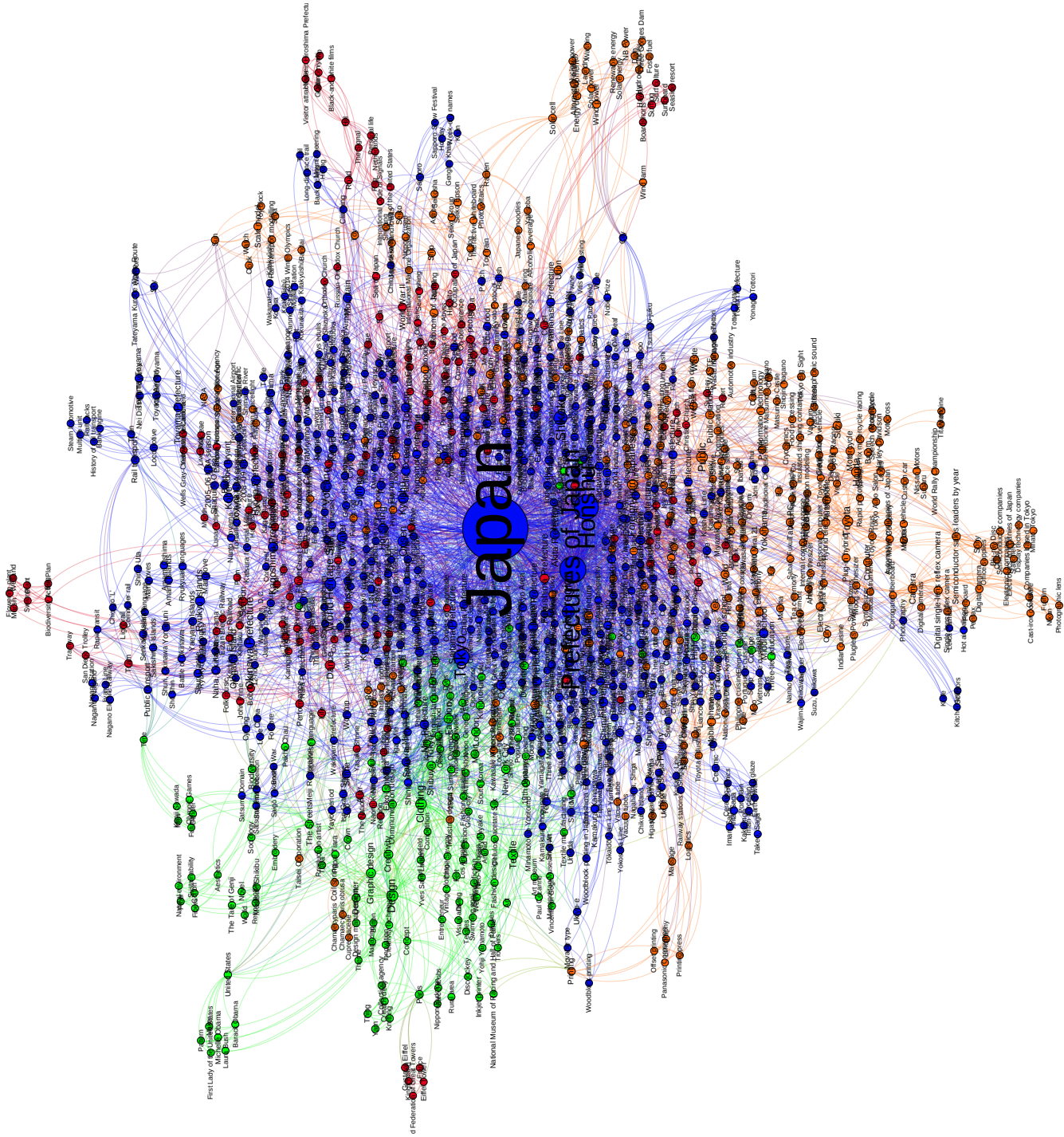


Figure 48: Concepts Co-occurrence – High Resolution