2F21

# Identifying Applications Emerging from the Knowledge Convergence of Big Data and Broadcasting :The Case of NHK and BBC

○Kumiko Miyazaki, Santiago Ruiz-Navas (Ritsumeikan Asia Pacific University)

## I INTRODUCTION

The competitive environment of public broadcasters has been changing rapidly because of emerging technologies. On the one hand, viewers' style of content viewing has changed, for example, the average TV viewing hours is becoming shorter and there is an increased preference for Video on Demand (VoD) services and personalization of content delivery over TV. Users are now viewing content on various devices, rather than on fixed platforms (TV, PC, movie theater). Users want to be more empowered on how they enjoy a video, when, where and with whom. Moreover, big data technologies are being implemented by new entrants such as Amazon Video, Youtube, Netflix to gain competitive advantage over incumbents, resulting in an intensified competition over users' program viewing time.

For a long time, public broadcasters have enjoyed a unique position, given the investment costs behind the infrastructure to broadcast programs (Metcalfe and Miles 2000) and the access to radio spectrum required to provide broadcasting services. Public broadcasters have been affected by the changing environment as they have to provide relevant content and services to the public viewers. BBC and NHK are examples of leading public broadcasters taking measures to adapt to the new competitive environment. These two broadcasters are pace-setters, possessing enough resources and accumulated technological capabilities to play a leading role in the broadcasting sector. BBC and NHK have shown themselves as champions of their sector (BBC 2013) by exploring big-data technologies to create new broadcasting services. Their R&D initiatives are activities to adapt to the changing competitive environment, in order to adopt new technological capabilities (Miyazaki 1995; Teece and Pisano 1994).

The main objective of this paper is to analyze the technological trajectories emerging from the knowledge convergence of big data and broadcasting technologies and hence identify the emerging applications by two leading public broadcasters, NHK and BBC.

## II BIG-DATA AND PUBLIC BROADCASTERS' VALUE CHAIN

Big-data is commonly defined by using the Vs framework, Volume, Velocity, Variety. We select the 3Vs framework following the recommendations of the OECD.

- Volume explains the size of big-data sets, like whole amount of tweets in a month or HD video content broadcasted in a year.
- Variety provides differentiation of structured data that comes with defined format and meta-data which can be easily queried, and non traditional form of data, such as video, graphics on social media
- Velocity is a measure of how fast the data is coming in. For example, data that has to be processed in real-time or streams of data, e.g., financial transactions, as well as live-broadcast data.

Big-data technologies are understood as the methods and technologies required to get value from the data possessing a high value of 3Vs (De Mauro et al. 2015). Big-data technologies have gained a stronghold in some sectors, e.g., science, healthcare, business intelligence. In this context, big-data technologies are understood as emerging technologies (Huang et al. 2015).

The broadcasting sector organizations produce and use data related to the media content. They create, transmit, and manage audio, text, video in different formats. NHK's R&D lab presented an overview of the potential of big-data for broadcasters, defining data sources for potential services such as Social Networks (SNS), access logs of web services (Nakagawa 2014).

There are different opportunities, applications, technologies, knowledge, data sources involved with the use of big-data in the broadcasting sector. Therefore, we introduce in Fig 1 the broadcasting value chain (Ruiz-Navas and Miyazaki 2015) as a conceptual tool to analyze the potential applications of big data for public broadcasters' services.

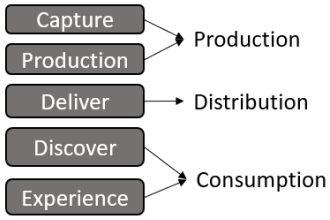


Fig. 1. Broadcasting's value chain. Source (Ruiz-Navas and Miyazaki 2015)

- Capture: The creative process of raw content generation
- Produce: Processing part of the content to meet specific objectives.
- Deliver: Refers to the activities that are needed to transport the content to the users/viewers.
- Discover: This sub-stage describes broadcasters' services that allow the users to explore broadcasters' content.
- Experience: Explains the service that broadcasters make available so users can enjoy the content

Big data may have an impact on several parts of public broadcaster's value chain, for example in the case of NHK– "Opinion analysis on programs, analyzing the natural language of viewers contained in social network posts related to NHK's programs; Program recommendation technology and Multimedia data processing, such as speech recognition" (Nakagawa 2014).

## Ⅲ RELATED LIERATURE

Organizations can adapt to paradigm changes by modifying their technological strategies, adopting new capabilities and competencies or exploring and selecting new technological trajectories. (Miyazaki 1995) analyzed the dynamics of competence building of several firms pursuing activities in the optoelectronics sector.

(Oltra and Saint Jean 2009), explored the technological trajectories of Low Emission Vehicles by analyzing the patent data of car makers (Yuan and Miyazaki 2013), identified the technological trajectories, in which the EV technologies developed.

In the research on systematic technology identification using patents and scientific papers, words or keywords are used as a proxy for knowledge to study science and technology. It is possible to use keywords to model the knowledge behind a given technology or defined topic and study its related phenomena, among them, technological change (Callon et al. 1992; Chen and Xiao 2016; Leydesdorff 1989). Words provide the advantage of detail but using them is challenging. Some of these challenges are the issues of vocabulary diversity used by researchers across disciplines or sectors (Leydesdorff 1997), additionally, homonymy and polysemy.

## Ⅳ OBJECTIVE AND RESEARCH QUESTIONS

Technological trajectories are connected to an organization's technological competencies, history, strategies and their environment. Technologies covered under the term big-data are constantly changing and are difficult to pin down (Huang et al. 2015). Thereby, to study technological trajectories related to big-data, first, we need to identify them.

If the technological trajectories are unknown beforehand. proxies of technological development can be used to identify them (Yuan and Miyazaki 2013). We propose the use of words and keywords to identify technological trajectories in big-data. The properties of keywords in texts or networks, makes it is possible to assess their relevance to a field and provide quantitative measures of their development and further study knowledge dynamics. (Ávila-Robinson and Miyazaki 2014) used centrality measures to detect importance of knowledge topics in MEMS. We propose to use keywords to identify the big-data technological trajectories, "hidden," in the publications available in scientific database accredited to public broadcasters

We propose two research objectives

- To identify the big-data technology trajectories of public broadcasters
- To explore the link between the big data related technology trajectories and the position in the value chain, in the case of BBC and NHK

## Ⅴ METHOD

The method consists of a combination of analysis of a keyword co-occurrence of network, finding similarities between two knowledge topics using keywords and finally a visualization and analysis of technological trajectories defined by the accumulation of frequency of keywords over time. It consists of four steps, a) Dataset acquisition b) Creation of keyword co-occurrence network for big-data and prioritization of keywords. c) Identify the shared keywords between prioritized big-data dataset and public broadcaster's dataset, over the years (2008-2016). d) Identify and analyze the technological trajectories.

### *A. Dataset acquisition*

The objective of this step is to obtain the dataset of publications available in scientific databases, to describe the knowledge related to big-data and that of public broadcasters.

Two lexical queries are used on the scientific databases to obtain the document datasets for big-data and public broadcasters, respectively. We borrowed the core lexical query proposed by (Huang et al. 2015) and extended the core and expanded lexical queries and obtained an updated lexical query for big-data which covers the topic over the years 2008-2016 (Ruiz-Navas, K. Miyazaki 2018a). The lexical query is run on Web of Science to obtain the dataset for big-data, The public broadcaster's lexical query was run on Web of Science and Scopus, and the documents obtained from these two databases were fused.

### *B. Creation of keyword co-occurrence network for big-data and priorization of keywords*

We use the results from a previous study done by the authors in (Navas and Miyazaki 2018b). In that work, we identified knowledge converging into big-data in two steps. First, we created a keyword co-occurrence network to describe the knowledge related to big-data and second, we selected the Keyword Proxy of Knowledge Convergence (KPKC) from the keyword co-occurrence network. Additionally, we implemented a disambiguation process based on the use of Wikipedia as an established knowledge source[1], to account for word polysemy and homonymy. The process to detect KPKC consists in detecting two keyword characteristics, Keyword's newness to the knowledge base (NKB) and keyword growing importance to the Knowledge base (GIKB). The list of KPKC for the big-data set is composed of 642 keywords which represent the knowledge converging into big-data.

### *C. Identify big data applications*

The objective of this step is to find the document published by NHK and BBC that describe big data applications.

We identify the common keywords of the two datasets (big-data and NHK; big-data and BBC) over time (2008-2016).-Such connection of keywords will allow us to identify the documents describing possible big data applications developed by NHK and BBC. Furthermore, we can analyze these papers to describe how the broadcasters have adopted knowledge from the big data.

The analysis of the papers describing big data applications consists of two stages, verify that the documents describe big data applications and 2) identify how the public broadcasters are using the big data application.

- Verify if the document describes big data applications: we read the document's abstracts searching for the three Vs of big data. If we found at least one of them, the paper was accepted as big data application, for example if we identified a paper describing a technique to identify viewers' preferences in a survey of 100 persons, we removed that document form the list, because a survey is not proceed in real time (Velocity), does not require big-data size storage (Volume) and requires a single data type (Variety). On the other hand, if the analysis of the viewers' preferences is done using real time (Velocity) twitter traffic (Volume) generated by the Hashtag (#) of a new drama then it is accepted as a big data application
- Identify how public broadcasters are using the application: We read the document abstract searching for text describing any of the five stages of the broadcasters' value chain presented in this paper. For example: if the analysis in real-time of the twitter traffic previously described was used solely to improve a text processing technique then it was removed. However if the analysis was done to provide feedback to the drama producers such as viewers' demographics (Production), scenes that generated the most traffic (Experience) or favorite actors (Production). Then the paper was accepted.

## VI RESULTS

### *A. Dataset acquisition:*

**The big-data dataset:** The big-data set was composed of 44.801 articles and 642 knowledge converging keywords. The number of documents per year, for the selected big-data set of documents, is shown in Fig 2.

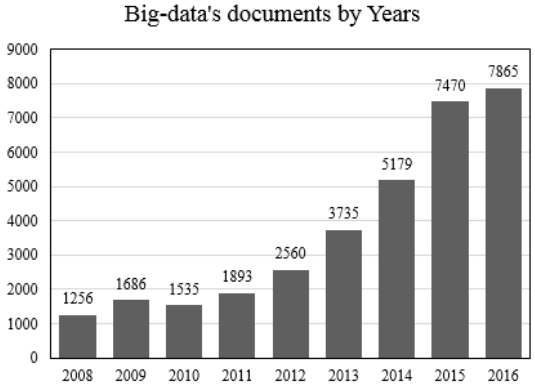


Fig. 2. Number of documents with populated author keyowrd field by year, for the big-data document set.

BBC's dataset is composed of 557 papers, the documents which did not have the author keyword field populated were removed, and 204 documents remained.

[1] The details of this disambiguation process can be found in (Navas and Miyazaki 2018a)

NHK dataset is composed of 3.262 papers, after removal of papers without author keywords, 1.065 remained..

*B. Creation of keyword co-occurrence network for big-data and priorization of keywords*

We used the set of knowledge converging keywords from our previous work comprised of 642 keywords. In Table 1, we show ten of them which were representative of topics converging into big-data.

TABLE I. TOP TEN KEYWORDS PER YEAR BIG DATASET [2000-2016]

| **Keywords representative of the knowledge converging into Big-data** |
|---|
| Natural language processing |
| Artificial Intelligence (AI) |
| Learning |
| Machine Learning |
| Pattern Recognition |
| Data science |
| General-purpose computing on graphics processing units |
| Social media |
| Speech recognition |
| Big-data |
| Ontology (information science) |

*C. Identify the shared keywords between the big-data dataset and public broadcaster's dataset, over the years (2008-2016)*

We obtained the list of shared keywords for the BBC, NHK and the big-data knowledge convergent keywords and prepared them for further processing.

C. ANALYZE AND IDENTIFY TECHNOLOGICAL TRAJECTORIES

We present one example of detecting the technological trajectories behind the shared keywords for BBC and NHK. BBC's accumulated number of papers per shared keyword for the period of 2007-2016 is shown in Fig 3.

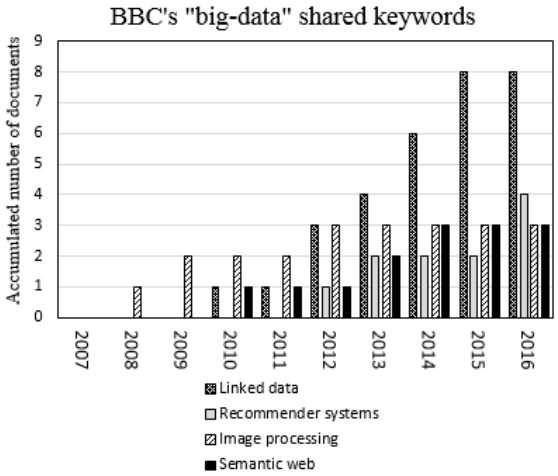


Fig 3 BBC's big-data shared keywords' accumulated documents for 2000-2016.

The shared keywords which accumulated more papers over time were "Linked data," followed by "Recommender systems," "Image processing" and "Semantic web" for BBC.

The number of documents connected to the shared keyword "linked data" has been increasing over time. They are related to the automatization of content publishing and program indexing, in the first years. Later in 2014, the papers explain applications related to automatic quiz generation.

Following the information of the documents in the BBC document set and from the big-data document set, we identified a project related to information retrieval which was the "The world radio service." The objective of this project was to index more than 70.000 radio programs, about three years' worth of continuous audio. To index these programs, BBC used to employ human indexers and the programs before this practice started were not indexed. The big-data characteristics of Volume and Variety were present in considerable amount of radio programs. (BBC&RD 2014).

In Table 2, the applications identified for the shared keywords for the BBC are presented.

TABLE 2 BBC'S SHARED KEYWORDS APPLICATIONS' DESCRIPTION.

| Trajectories | Description |
|---|---|
| Linked data | Automatic radio program indexing<br>Automatic content publishing on the web<br>Quiz generation |
| Recommender systems | Understanding user experience in recommender systems.<br>Recommender system |

| | |
|---|---|
| | Optimization of large-scale on-demand TV based on user viewing profiles. |
| Semantic web | Web community analysis based on user interactions.<br>Automatic and crowdsourced tagging of the BBC's radio archive |
| Image processing | Image-based camera tracking |

The implementation of big-data technologies by BBC started after 2008, reflecting the effects of the big-data dataset which spans from 2008-2016. The applications related to "linked data" explain the usage of data to automatize two processes inside the BBC, to publish web content and index the radio archives.

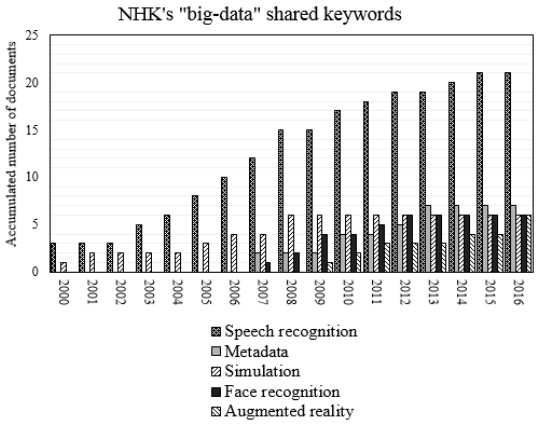


Fig 4 NHK's shared keywords' accumulated documents, for the years 2000-2016.

At NHK, from Fig 4, it can be seen-that the two shared keywords started before 2008. We checked these two keywords to learn the legacy applications that they were related were enhanced using big-data technologies. The shared keywords from which technological trajectories are identified are "speech recognition," "Metadata," "Simulation," "Face recognition" and "Augmented reality" (AR).

From Table 3, we identified two trajectories, Improvement of AR service provision on mobile devices and visualization of data using AR.

From the analysis of the documents linked to AR in the NHK dataset, we identified that in early years NHK reported the development of camera tracking technology to add AR elements to their content, e.g., sports or variety programs. Subsequently, the documents describe the analysis of the implementation of AR-enabled TV and considerations about the usage of AR on mobile devices.

Following the information from the documents of the big-data and public broadcaster's sets the analysis revealed that NHK is implementing big-data technologies to improve its AR TV services.

TABLE 3 NHK'S SHARED KEYWORDS APPLICATIONS' DESCRIPTION

| Trajectories | Description |
|---|---|
| Speech recognition | Automatic transcription and captioning of news programs.<br>speech recognition process improvement .<br>Real-time language model and acoustic models update for real-time program captioning.<br>Language model training improvement |
| Metadata | Broadcast programs metadata used for access control system.<br>Automatic creation of sport program scenes metadata.<br>Creating summaries of news programs using camera shot roles.<br>Automatic scene detection. |
| Augmented reality | Camera tracking using 3d models<br>Synchronization of 3D images when a program is seen on multiple screens<br>Age estimation for multiple screen AR secure viewing |
| Face recognition | Soccer players tracking and recognition<br>Video indexing by automatic face image registration |

From the analysis of documents linked to NHK's shared keywords, we identified two trajectories, first described as the synergy of face and speech recognition which are used to generate metadata to index NHK's programs. Second, speech recognition as it helps NHK's process of automatic speech transcription and supports other services; and third, augmented reality which is supported by face recognition technologies to track camera position and provide AR TV

A summary of the potential applications in the broadcasters' value chain of the previously discussed applications is introduced in Table 4.

TABLE 4 IDENTFIED TRAJECTORIES FOR BBC AND NHK AND THE STAGE OF THE BROADCASTERS' VALUE CHAIN

| Public broadcaster | Trajectory | Value chain |
|---|---|---|
| NHK | Speech recognition | Production |
| | Generate metadata to index NHK's programs | Discover |
| | Augmented Reality | Production and experience |
| BBC | Linked data | Discover |
| | Recommender systems | Discover |
| | Semantic Web | Production |

| | | |
|---|---|---|
| | Image processing | Deliver |

## Ⅶ Conclusions

We used keyword analysis and the concepts of the Three Vs of big data and the public broadcasters value chain to identify applications emerging from the knowledge convergence of BBC, NHK and big data.

Furthermore, we propose that that the identified applications reflected the big data strategy of each broadcaster. In the case of NHK, they were adopting big data to strengthen their 'Production' and 'Experience' part of the value chain, whereas BBC was trying to strengthen their 'Discover' part of the value chain through big data.

A limitation of this study is the considerable amount of human intervention required to identify the applications emerging from the BBC, NHK and big-data knowledge convergence.

## References


Ávila-Robinson, A., & Miyazaki, K. (2014). Assessing nanotechnology potentials : interplay between the paths of knowledge evolution and the patterns of competence building. International Journal of Technology Intelligence and Planning, 10(1), 1–28. doi:10.1504/IJTIP.2014.066709

BBC&RD. (2014). The World Service Radio Archive - BBC R&D. BBC Research and Development. http://www.bbc.co.uk/rd/projects/worldservice-archive-proto.

BBC. (2013). Public and private broadcasters across the world - the race to the top.

Callon, M., Laredo, P., Rabeharisoa, V., Gonard, T., & Leray, T. (1992). The management and evaluation of technological programs and the dynamics of techno-economic networks: The case of the AFME. Research Policy, 21(3), 215–236. doi:10.1016/0048-7333(92)90017-X

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In AIP Conference proceedings 1644 (Vol. 97, pp. 97–104). AIP Publishing. doi:10.1063/1.4907823

Huang, Y., Schuehle, J., Porter, A. L., & Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'Big Data.' Scientometrics, 105(3), 2005–2022. doi:10.1007/s11192-015-1638-y

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. Research Policy, 18(4), 209–223. doi:10.1016/0048-7333(89)90016-4

Metcalfe, J. S., & Miles, I. (2000). Innovation systems in the service economy. (C. Antonelli & B. Carlsson, Eds.) (1st ed.). New York, New York, USA: Springer Science+Business Media, LLC. doi:10.1007/978-1-4615-4425-8

Miyazaki, K. (1995). Building competencies in the firm: Lessons from Japanese and European optoelectronics (1st ed.). New york: ST. MARTIN'S PRESS, INC. doi:10.1007/978-1-349-2387

Nakagawa, T. (2014). Overview of Big Data Uses and Applications. Broadcast Technology, (55), 16–17.

Ruiz-Navas, S. & Miyazaki, K. (2018b). Developing a framework to track knowledge convergence in "big data." International Journal of Technology Intelligence and Planning, 12(2), 121. doi:10.1504/IJTIP.2018.096101

OECD. (2014). Data-driven Innovation for Growth and Well-being: Interim Synthesis Report, (Xx), 86.

Oltra, V., & Saint Jean, M. (2009). Variety of technological trajectories in low emission vehicles (LEVs): A patent data analysis. Journal of Cleaner Production, 17(2), 201–213. doi:10.1016/j.jclepro.2008.04.023

Ruiz-navas, S., & Miyazaki, K. (2014). Unveiling the Knowledge Convergence on " Big data " : Analysis of Scientific Keywords . Asialics 2014, 1–23.

Ruiz-Navas, S., & Miyazaki, K. (2015). Big data enabled service innovation in broadcasting. In ASIALICS 2015 "Innovation Driven Natural Resource Based Industry." Yogiakarta.

Ruiz-Navas, S., & Miyazaki, K. (2018a). A complement to lexical query's search-term selection for emerging technologies: the case of "big data." Scientometrics, 117(1), 141–162. doi:10.1007/s11192-018-2857-9

Yuan, F., & Miyazaki, K. (2013). Understanding the Dynamic Nature of Technological Change Using Trajectory Identification Based on Patent Citation Network in the Electric Vehicles Industry. In PICMET 14 conference Infrastructure and Service Integration (pp. 2780–2790).

Teece, D., & Pisano, G. (1994). The dynamic capabilities of firms: An introduction. Industrial and Corporate Change, 3(3), 537–556. doi:10.1093/icc/3.3.537-a