
Culturomics: Reflections on the Potential of Big Data Discourse Analysis Methods for Identifying Research Trends

Vered Silber-Varod, The Open University of Israel, vereds@openu.ac.il

Yoram Eshet-Alkalai, The Open University of Israel, yorames@openu.ac.il

Nitza Geri, The Open University of Israel, nitzage@openu.ac.il

Abstract

This study examines the potential of big data discourse analysis (i.e., culturomics) to produce valuable knowledge, and suggests a mixed methods model for improving the effectiveness of culturomics. We argue that the importance and magnitude of using qualitative methods as complementing quantitative ones, depends on the scope of the analyzed data (i.e., the volume of data and the period it spans over). We demonstrate the merit of a mixed methods approach for culturomics analyses in the context of identifying research trends, by analyzing changes over a period of 15 years (2000-2014) in the terms used in the research literature related to learning technologies. The dataset was based on Google Scholar search query results. Three perspectives of analysis are presented: (1) Curves describing five main types of relative frequency trends (i.e., rising; stable; fall; rise and fall; rise and stable); (2) The top key-terms identified for each year; (3) A comparison of data from three datasets, which demonstrates the scope dimension of the mixed methods model for big data discourse analysis. This paper contributes to both theory and practice by providing a methodological approach that enables gaining insightful patterns and trends out of culturomics, by integrating quantitative and qualitative research methods.

Keywords: Culturomics, quantitative methods, discourse analysis, big data, textual analytics, learning technologies, mixed methods model for big data discourse analysis.

Introduction

Big data discourse analysis is a current trend in research and practice, which aims at extracting valuable knowledge from the ample digital textual data available in the information age. In the context of identifying cultural trends, this approach was initially termed in 2010 as culturomics (Bohannon, 2010; Michel et al., 2011). Culturomics focuses on the quantitative exploration of massive datasets of digital text. In this study, we examine the potential of culturomics and suggest a mixed methods model for big data discourse analysis, which argues that quantitative methods alone are not sufficient for effective big data discourse analysis and that in order to identify meaningful trends and patterns, qualitative methods should be applied alongside the quantitative ones.

The potential of culturomics, as well as the need for a mixed methods approach, is empirically demonstrated in this paper in the context of identifying research trends. We analyzed terminology related to the learning technologies research field over a period of fifteen years

(2000-2014), using a dataset based on search query results of Google Scholar, which is an open-source large-scale diachronic database.

Studies on the evolution of new research disciplines are usually conducted by experts in the studied field who have retrospective insights on it (inter alia, Belanger & Jordan, 2000; Laurillard, 2013). Another conventional method for studying the evolution of a discipline is by obtaining data via surveys (e.g., Hart, 2007-2014, an annual Learning Tools Survey that has been compiled from the votes of 1,038 learning professionals from 61 countries worldwide). Both methods require experts' participation and rely on human knowledge.

Rather than employing a quantitative approach, or a qualitative one, for investigating discourse trends, we propose a mixed methods model for big data discourse analysis, which provides a methodological approach for selecting the appropriate combination of quantitative and qualitative methods as a function of the data scope that encompasses both the amount of data and its temporal expansion.

Theoretical background

The massive data-driven approach emerged during the 1990ies from the corpus linguistics field of research (Sinclair 1991; 2004) as a standard methodology for learning about linguistic phenomena from textual (written or spoken) data (Webber, 2008). The main advantage of corpus linguistics is that it helps revealing patterns and features of a certain theme in big data sets, allowing a broader perspective on entire large corpora or texts rather than on specific phenomena within them (Rayson, 2008).

The term *culturomics* – the quantitative exploration of massive datasets of the digital text was coined by Michel et al. (2011). They demonstrated its utility by creating a dataset of trillions of words from 15 million books in Google Books, and making it accessible to the public so that researchers can search for patterns of linguistic change over time in the frequency distribution of words or phrases. Michel et al. (2011) claimed that utilizing unique text-analysis methods on corpora of massive data enables identifying and measuring cultural patterns of change, which are reflected by the language choices represented in the texts.

N-gram analysis is one of the main computational text mining techniques, which culturomics utilizes. An n-gram is a sequence of words of length n. An n-gram analysis is based on calculating the relative frequency that a certain n-gram appears in a dataset (for a description of n-gram analysis, see Soper & Turel, 2012). Michel and Lieberman Aiden (2011) demonstrated the value of the culturomics approach in a historical study of flu epidemics, where they showed how the search for the term *influenza* resulted in a graph whose peaks correlated with flu epidemics that killed massive populations throughout history.

Similarly, Bohannon (2011a) demonstrated the power of culturomics in his study of the Science Hall of Fame (SHoF) – a pantheon of the most famous scientists of the past two centuries. The common criteria for inclusion of scientists in SHoF are based on conventional rigorous metrics of scientific excellence, such as the amount of citations and journals' impact factor (Bar-Ilan, 2008). Bohannon (2011a) suggested culturomics as an alternative method for measuring a scientist's impact, by weighing the cultural footprint of scientists across societies and throughout history.

One of the advanced things one can do with frequency analysis of massive texts is to measure "how fast does the bubble burst?" (Michel & Lieberman Aiden, 2011), or in other words, what is the life span of scientific trends or a buzzword before they disappear from the research literature? In their study, Michel & Lieberman Aiden (2011) found that the "bubble" bursts faster and faster with each passing year, and claimed: "We are losing interest in the past more rapidly". Furthermore, an absence of certain terms or people during a period, or after a certain point of time, which a culturomic analysis reveals, might provide insights into cultural phenomena (Bohannon, 2010).

Following the rise of culturomics, it has been utilized in many other disciplines (Tahmasebi et al., 2015), such as history, literature, culture and social studies, (e.g., Hills & Adelman, 2015; Willems, 2013), accounting (Ahlawat & Ahlawat, 2013), and physics (Perc, 2013). Attempts have been made in various disciplines to apply a variety of quantitative computational methods on the available massive volumes of data, in order to make informed predictions of trends (e.g., Radinsky, Agichtein, Gibrilovich, & Markovitch, 2011; Radinsky, & Horvitz, 2013). Soper & Turel (2012) performed a culturomic analysis of the contents of the monthly magazine of the Association for Computing Machinery (ACM), *Communications of the ACM*, during the years 2000-2010, and demonstrated how it can be used to quantitatively explore changes over time in the identity and culture of an institution. Soper, Turel, & Geri (2014) used culturomics to explore the theoretical core of the information systems field. Other computational methods that were used to identify trends of change in research fields were citation analysis, co-citation analysis, bibliometric coupling, and co-word analysis (Van Den Besselaar & Heimeriks, 2006).

The field of learning technologies undergoes frequent changes as new technologies emerge, experienced, and adopted or abandoned in practice. These changes are reflected in the themes of publications in this research field (Geri, Blau, Caspi, Kalman, Silber-Varod, & Eshet-Alkalai, 2015). Raban and Gordon (2015) used bibliometric analysis methods in order to identify trends of change in the field of learning technologies over a period of five decades. While the quantitative methods that Raban and Gordon (2015) applied revealed some insights, some of their findings demonstrate the need for a complementary qualitative evaluation. Silber-Varod, Eshet-Alkalai, and Geri (2016) used a data-driven approach for analyzing the discourse of all the papers published in the proceedings volumes of the Chais Conference for the Study of Innovation and Learning Technologies, during the years 2006-2014. Chais conference is conducted annually in Israel. In terms of culturomic research, Silber-Varod et al.'s (2016) corpus, which included 730,000 words (in Hebrew and in English) of 553 articles, is considered small. Silber-Varod et al.'s (2016) study demonstrated the potential of such an analysis, by showing that Israeli researchers of learning technologies are concerned primarily with pedagogical aspects of technology adoption and use, as well as identifying prominent key terms. However, it also exemplified the limitations of quantitative methods in identifying trends during relatively short periods, and a profound qualitative interpretation of the results was required in order to gain valuable insights.

More recently, the use of visualization tools, such as Wordle, enabled a better illustration of these trends (Viegas, Wattenberg, & Feinberg, 2009). A well-known example of a visualization tool that represents a prediction of technological change over time is the Gartner's Hype Cycle of assimilating emerging technologies (Fenn, 1995), in which innovative technologies are plotted

on a graph according to their maturity and predicted the status of assimilation or disappearance (Linden, & Fenn, 2003).

Similar to Gartner's Hype Cycle (Fenn, 1999), and based on implications of our culturomic study of learning technologies discourse (Silber-Varod, et al., 2016), we hypothesize that terminology of the learning technology field would show five main patterns of change over time: rising; stable; fall; rise and fall; rise and stable. We argue that these patterns reflect their maturity status. The rising curve reflects terminology at the positive hype. The fall curve reflects terminology at the negative hype. The rise and fall curve reflects a terminology at the disillusionment point (Fenn, 1999); and the rise and stable curve reflect a plateau of productivity. Since our data spanned over the years 2000-2014, we would expect to find the rising curve for terms that represent technologies and research trends that emerged in the last years, and the fall curve for older terms. In the stable curve, we would predict to find the oldest core terms of the research discipline, and in the rise-fall curve, we would anticipate to find rather hype terms, which lost their relevance rapidly. Nevertheless, in order to fully interpret the findings, additional qualitative analysis would be required, as our mixed methods model for big data discourse analysis suggests.

A mixed methods model for big data discourse analysis

The ubiquitous presence of digital textual information, which is often available online, calls for appropriate methods and tools for content analysis, which would provide valuable knowledge. The information systems field has been facing similar challenges of extracting knowledge from data for several decades. The common four level hierarchy of knowledge extraction is, from bottom to top: Data, Information, Knowledge and Wisdom (DIKW), known as the DIKW hierarchy, the "Knowledge Pyramid" and the "Information Hierarchy" (Ackoff, 1989; Rowley, 2007; Zeleny, 1987). At the lower levels of this pyramid (data and information), mainly automatic quantitative tools are used to extract the relevant content, which is required for decision-making. However, at the higher levels of the pyramid (i.e., knowledge and wisdom), human intervention may be essential (Alavi & Leidner, 2001; Davenport & Prusak, 1998). Similar concerns were raised regarding the interpretation of quantitative culturomic analyses findings (Bohannon, 2011b), which emphasized the need for human discretion.

Akin to knowledge and wisdom extraction, which require certain degrees of human intervention, we suggest that qualitative methods should be systematically integrated in culturomic analyses. We propose a mixed methods model for big data discourse analysis in order to represent the interaction between the data scope, which is the amount of data and the period it spans over, and the type of appropriate analysis on a quantitative-qualitative scale. The larger the scope of a database, the more likely it is that quantitative methods may provide meaningful insights, and there would be less need for complementary qualitative analyses. Table 1 presents a schematic description of the mixed methods model for big data discourse analysis. The scope construct, with its two dimensions: Size of the dataset and period analyzed, provides a useful systematic approach for planning culturomic analyses. Moreover, the mixed methods model expands the possibility to perform valuable culturomic analyses of databases, which one or both dimensions of their scope limit the effectiveness of quantitative methods.

Table 1. A Mixed Methods Model for Big Data Discourse Analysis

Size of the dataset	Relatively small	Relatively big
Period analyzed		
Long period of time	Mixed methods	Mainly Quantitative methods
Short period of time	Mainly Qualitative methods	Mixed methods

Methodology

Search queries design - In order to identify trends and patterns of change over time in terminology related to learning technologies research, we used the open-source diachronic database of Google Scholar. Search engines are one of the most important tools that Internet users rely on. Popular search engines, such as google.com and yahoo.com, are powered by information retrieval, data mining, and machine learning algorithms. According to Chen, Kiciman, & Wang (2008), conventional search engines return deterministic results. Such objectivity provides dependable and reproducible results.

We compiled a list of 186 search terms. A similar amount of search queries, 200, was used by Segev, Aviv, & Baram-Tsabari (2014) for analyzing temporal patterns of scientific information seeking on Google and Wikipedia over a period of 264 weeks between December 2007 and December 2012. The 186 search terms list was based on two sources:

- Keywords used in the 553 papers presented at the annual Chais Conference for the Study of Innovation and Learning Technologies, over the years 2006-2014, as reported in Silber-Varod et al. (2016).
- The Directory of Learning & Performance Tools (Hart, 2007-2014), which lists over 2,000 terms related to learning and working in education and at the workplace. Terms were selected from the webpage "Top 100 Tools 2007-2014", which indicates the rating of each term since the first survey in 2007. From this list, we extracted the generic terms (not actual labels).

We cross validated the tools and the keywords from both lists, thus creating a new 186 terms list. Each of the 186 terms was searched twice according to the following setup:

1. Each query was performed using quotation marks for the term (e.g., "Facebook group"), in order to retrieve only the exact phrase. In addition, a set of sub-terms was used as anchors for all the queries in order to contextualize the results in the learning technologies field. The anchor sub-terms were "learning technologies" OR "learning technology", i.e., query results were of papers that included at least one of these terms.
2. The queries were performed per year, from 2000 to 2014. In addition to these 15 years, we searched for results until 1999. Thus, each term had results of 16 periods, with the set of anchor terms. In total, 2,976 search queries were carried out.
3. In order to validate the data, we duplicated the search queries after six months for a subset of 24 terms. Although the absolute numbers changed (higher results at the second run), the difference was minor, and more importantly, the relative frequencies were not

changed at all. This validation helps to overcome the effect of the limitations of Google Scholar as a research tool.

Limitations of Google Scholar search results - For the purpose of the present study, the advantages of using Google Scholar lie in its massive volume of data, each datum being identified by the year of publication, and in the open access that it provides for all. However, search engines have some limitations, and several methodological aspects should be clarified:

- Indeterminacy of date: Queries per year are not 100% accurate since there are documents (i.e., articles) that are not assigned to a specific year. In some cases, documents dated decades ago are assigned to recent years due to a new platform or website that publishes them. Thus, the results of the assignment of papers per year are not fully accurate. Nevertheless, our empirical study suggests that these deviations are negligible.
- Indeterminacy of context: Although the queries were carried out with anchor terms, the context of the search term is still vague to some extent, and not all papers in the results belong to the examined learning technologies field. It is possible that the search term will appear as a paper in the reference list only, or in an article that focuses on another research field.
- Indeterminacy of results: Search engines' purpose is to provide the user quick results of the input query. Therefore, even if the query resulted in tens of thousands results, the probability that an average user will look at each one of them is zero. However, for our purposes, it is important to be aware that even the tens of thousands of results are not to be considered the "whole population". Thus, we are limited in our knowledge about the full size of the query population.

Notwithstanding, although the results are inaccurate with regard to exact amounts of terms, in the relevant context, and during the examined period, the results provide satisfying trends, as demonstrated in the validity check we performed, which is described above (point 3).

Relative frequency and keyness analysis - The opening of the Internet for public use in the early 1990s resulted in a general exponential increase of all sorts of published content. It may be argued that since the amount of academic papers, which are published each year, grows, the most expected pattern of a frequency curve is a rising curve. Therefore, following prior culturomics analyses (Soper & Turel, 2012) we used relative frequencies of terms.

After recording the absolute frequency of each of the 2,976 query results, we calculated the relative frequency of each term, according to the following formula:

$$\# \text{ of search results per term per year } X / \text{ Total } \# \text{ of all search results per year } X.$$

This method allowed us to contextualize the amount of search results within our own database and create graphs that show trends of change over time in the research field of learning technologies, as represented by the distribution of major terms.

Furthermore, we tracked unique occurrences of wording and terminology throughout the years. For this purpose, we used an analysis based on the comparison of each year's frequency to a

reference corpus. The reference corpus consisted of previous years' frequencies, as this allows identifying unique occurrences of a term in a given year (Rayson, Berridge, & Francis 2004; Künstler, Maiwald, & Saage, 2008). The comparison was carried out by Log-likelihood calculations (Anthony, 2011; Rayson, 2008), which enable the comparison of frequencies between a given corpus and a reference corpus. Thus, a high Log-likelihood value represents high keyness (a method that uses the log-likelihood ratio statistic to compare frequencies and then rank them in terms of the significance of differences.) of a certain term in a certain year, while lower values represent low keyness.

The design of our research enables achieving systematic results via Google Scholar, which minimizes the aforementioned limitations of search engines. As our method is carried out within a given research field, and the analysis is based on relative frequencies calculated from a given database, we thus contend that the word frequency observations represent accurate trends.

Results

The results presented in this section demonstrate the potential of quantitative big data discourse analysis for identifying research trends. We present three general perspectives of analysis. The first perspective uses curves to describe five main types of relative frequency patterns found throughout the years. The second perspective provides the top key-terms per each year. These two analyses are based on the dataset of our Google Scholar query results. The third perspective demonstrates the scope dimension of the mixed methods model for big data discourse analysis, with data from three datasets.

Relative frequency curves

The rising curve - The rising curve is interesting for its starting point, i.e., whether the rise is evident from the first examined year (e.g., "Blended learning" in Figure 1), or it begins during the examined period (e.g., "Facebook group" in Figure 1). However, a rising curve may turn in the following years into a rise and fall curve, or a rise and stable curve, which are described below.

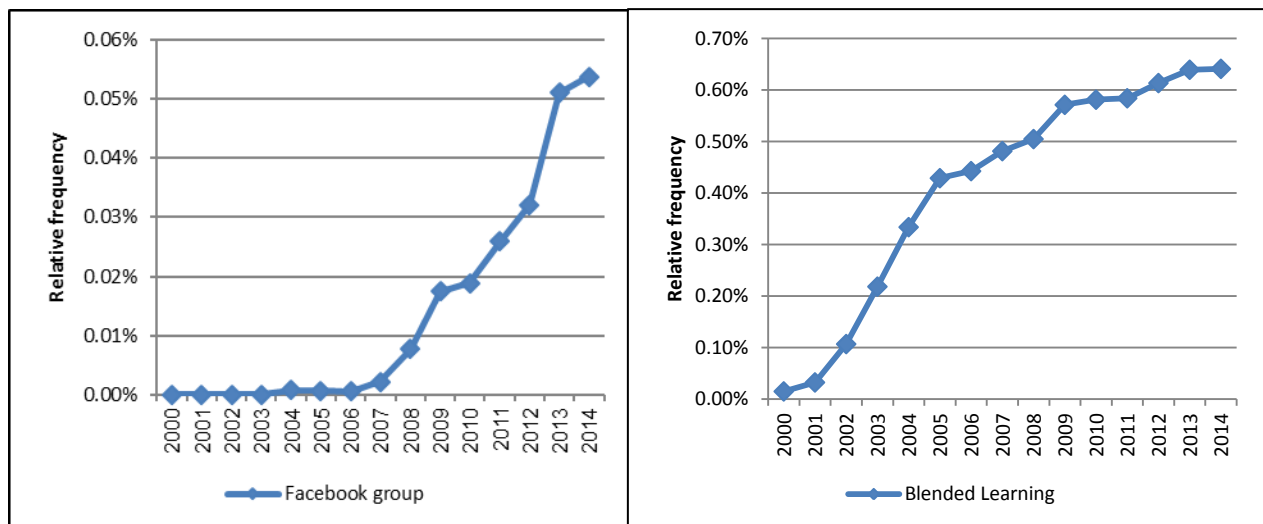


Figure 1. Examples of the rising curve of two terms in the field of learning technologies

The stable curve - A stable curve describes terms, which are expected to be included in many publications that belong to the examined field. The stable curves in our database included some core terms in the learning technologies field, such as Learning, Technology, Education, and Training (see Figure 2). We suggest that the slight decrease, which is shown in Figure 2, should be interpreted as a relative decrease, and not a decline of interest. As the field of research advances, there are more terms. Thus the relative weight of the core terms may decrease a bit.

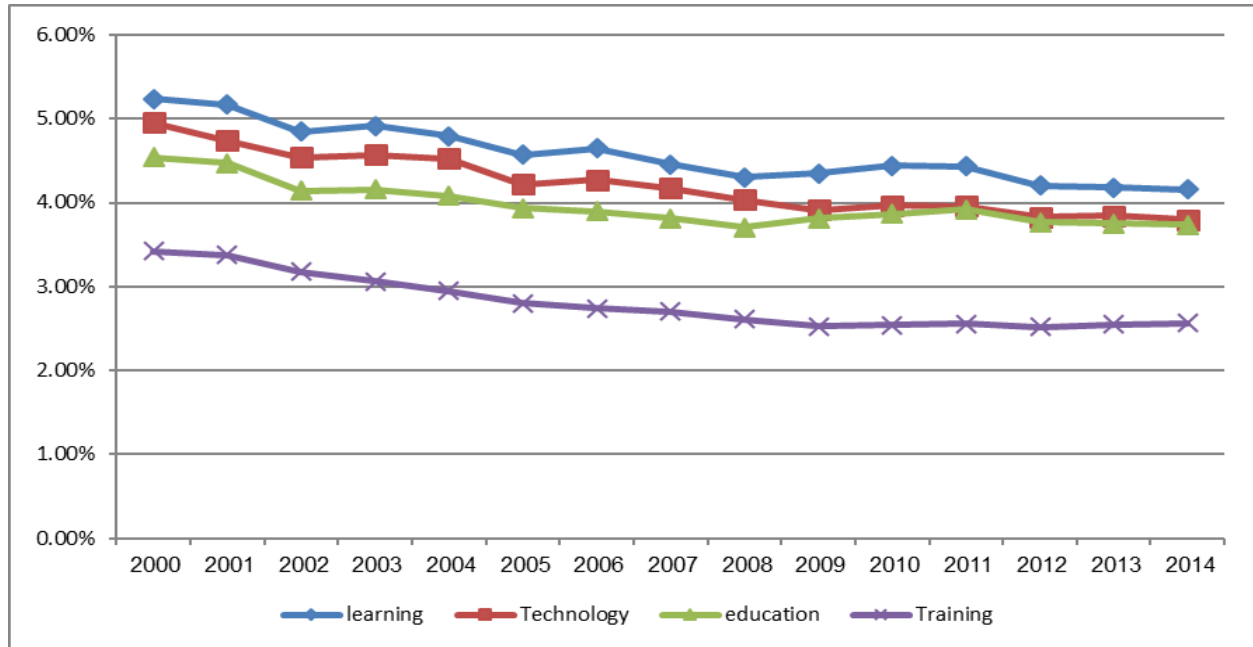


Figure 2. Examples of the stable curve of four core key terms in the field of learning technologies

The fall curve - There are terms that become relatively less prevalent over time, such as the term “technological change”, which is depicted in Figure 3. As with all other types of curves, the quantitative analysis just shows the decline in relative frequency, but it does not explain the reason for it.

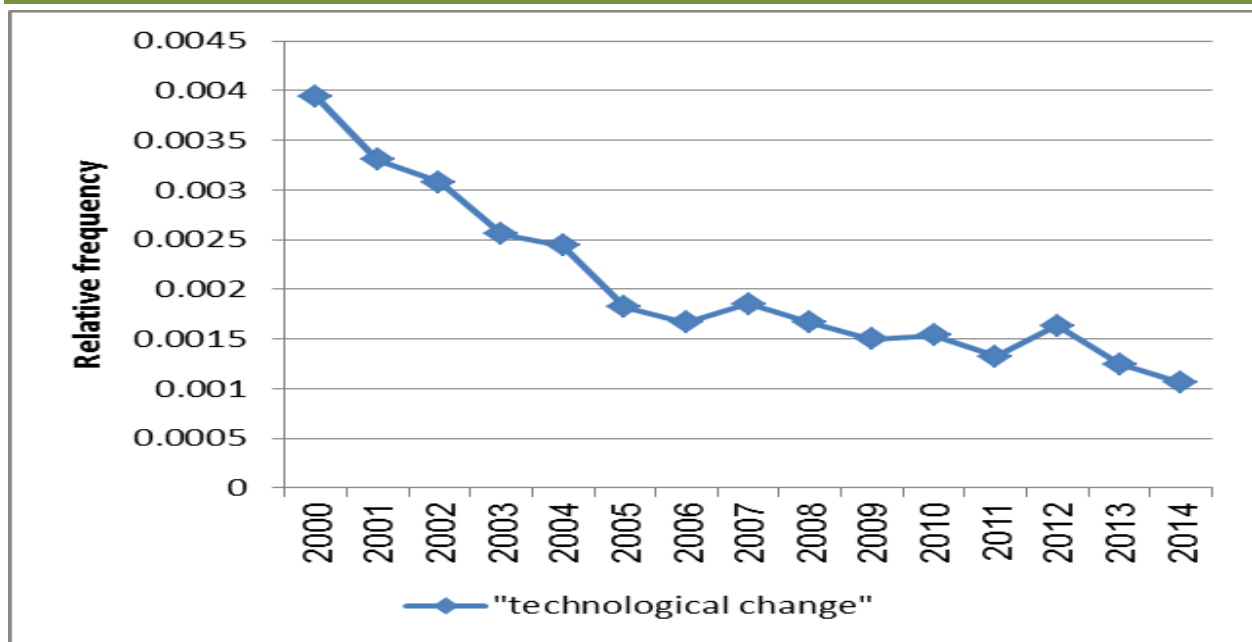


Figure 3. An example of a fall curve of a term in the field of learning technologies

The rise and fall curve - Another typical frequency pattern is the rise and fall curve (see "learning objects" in Figure 4). This curve shows the peak of interest in a term, which later becomes less prevalent. Sometimes, there are several peaks, such as the term "knowledge sharing" (see Figure 4), during the examined 15 years. In the case of one peak, the term emerged and later became less frequent for some reason. However, the case of several peaks, as demonstrated in figure 4, may be of a relatively important term in the examined context, i.e. "knowledge sharing" in this particular example, which every few years experiences a local peak of interest.

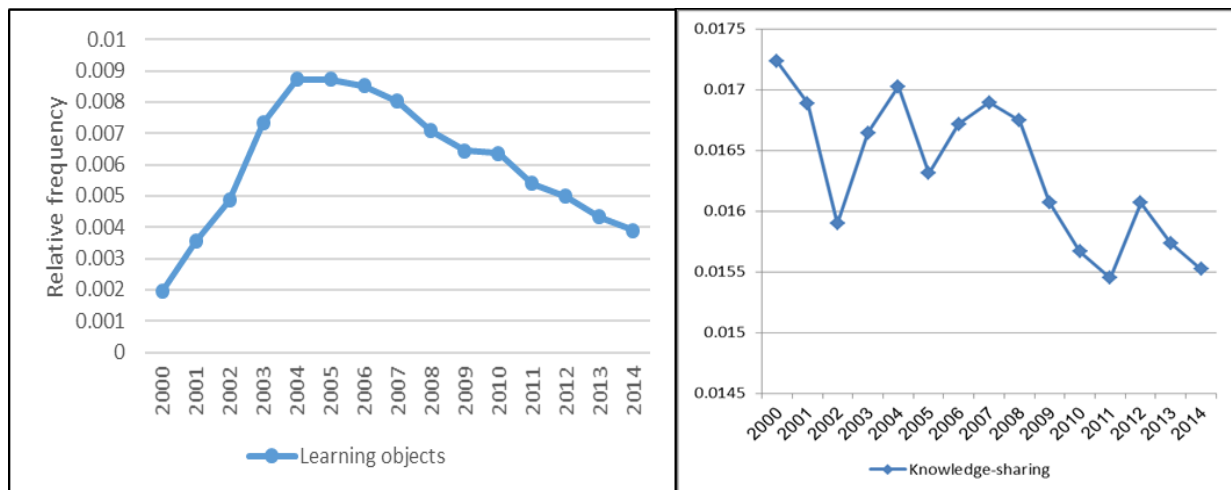


Figure 4. Example of two sorts of the rise and fall curve of terms in the field of learning technologies

The rise and stable curve - The rise and stable curve demonstrates a term that has reached its peak and still keeps this relative frequency of use ("e-learning" and "cyber bullying" in Figure 5).

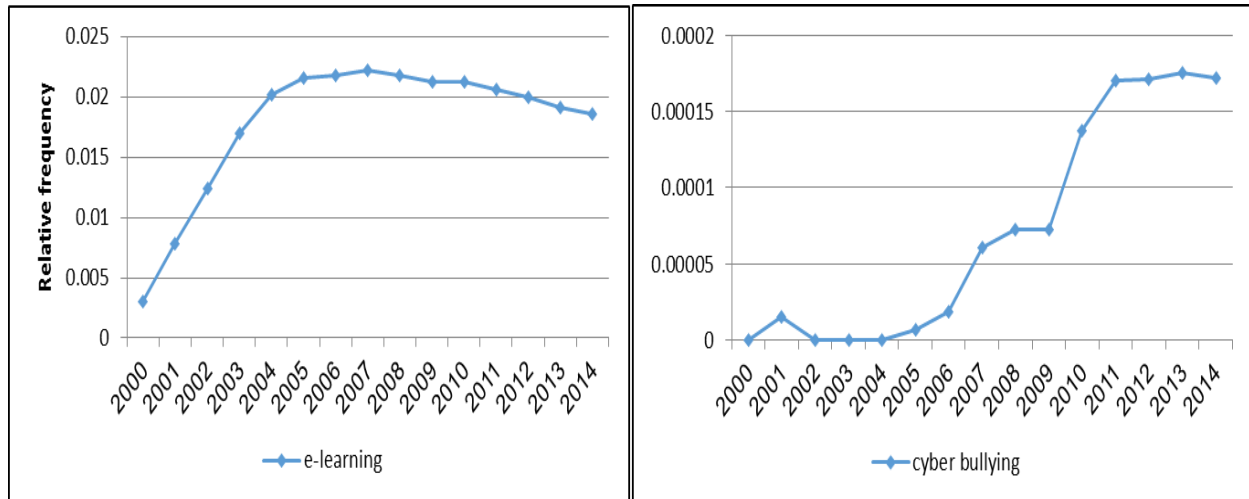


Figure 5. Example of two sorts of the rise and stable curve of terms in the field of learning technologies

The above five types of curves: rising; stable; fall; rise and fall; rise and stable, are the main types of relative frequency distribution patterns. However, they do not encompass the patterns of all the 186 terms we examined. Several terms had more than one rise or one fall, which represents a much less stable use of these terms.

Key Terms and Their Relation to the Emergence of Technologies

The keyness analysis is shown in Table 2, where we summarized the top three growing key terms (out of the 186 terms in the database) according to Log-Likelihood calculations. It can be seen in Table 2 that "Internet" was a relatively key term in 2000, and "e-learning" emerged from 2000 until 2006. Learning Management System (LMS) was one of the top key terms in 2003, and Moodle started its emergence in 2005. The term Wiki was among the key terms from 2006 until 2011. In 2011, "social networks" was one of the top key terms, and in 2013, research on Massive Open Online Courses (MOOCs) emerged prominently.

Following the results in Table 2, we performed an exploratory study, which aimed to identify the gap between the introduction of new technologies or systems and their evolvement in the research discourse. We looked for the time the technologies related to the terms in Table 2 emerged, and found the following: With regard to *Internet* (key term in 2000), by the end of 1994, the total number of websites was still relatively small, but many notable websites were already active. *Wikipedia* (key term in 2006) was launched in 2001. Moodle's (key term in 2005) first version was released in 2002, Facebook (key term in 2011) was launched in 2004, and the first MOOC (key term in 2013) was launched in 2008. When calculating the gap between the appearance of a certain technology and its research ascent, as represented by the keyness results,

we found that it took between three years to seven years for the technology to prominently emerge in the learning technologies research field.

Table 2. Top three key terms of the learning technologies research field

Year	Top Key Term	2nd top Key Term	3rd top Key Term
2000	Afterschool	online learners OR online learning	Internet
2001	e-learning	online learners OR online learning	Learning objects
2002	e-learning	ICT	online learners OR online learning
2003	e-learning	Learning objects	Learning Management System (LMS)
2004	e-learning	Learning objects	Blended learning
2005	e-learning	Blended learning	Moodle
2006	e-learning	Wikipedia / Wiki	Moodle
2007	Wikipedia / Wiki	web 2.0	Moodle
2008	web 2.0	Wikipedia / Wiki	Moodle
2009	web 2.0	Wikipedia / Wiki	Moodle
2010	web 2.0	Wikipedia / Wiki	Moodle
2011	social media	web 2.0	Facebook
2012	social media	Facebook	Twitter
2013	social media	MOOC	Facebook
2014	social media	MOOC	Gamification

Demonstration of the Mixed Methods Model for Big Data Discourse Analysis

An example of the scope parameter and its effect on the ability to interpret quantitative results is shown in Figure 6. The figure represents the relative frequencies of the term "collaboration" as were found in three databases: Google Books Ngram Viewer (1980-2008), Google Scholar (2000-2014), and Chais Conference Proceedings (2006-2014). The Google Ngram Viewer "collaboration" curve represents a rising curve with a steeper slope beginning in the 1990ies, and a stable use during 2005-2008. Google Scholar results reflect a stable use in the field of learning technologies during 2000-2014, and the conference proceedings' results show much more oscillations. However they are within 1% of the relative frequency scale.

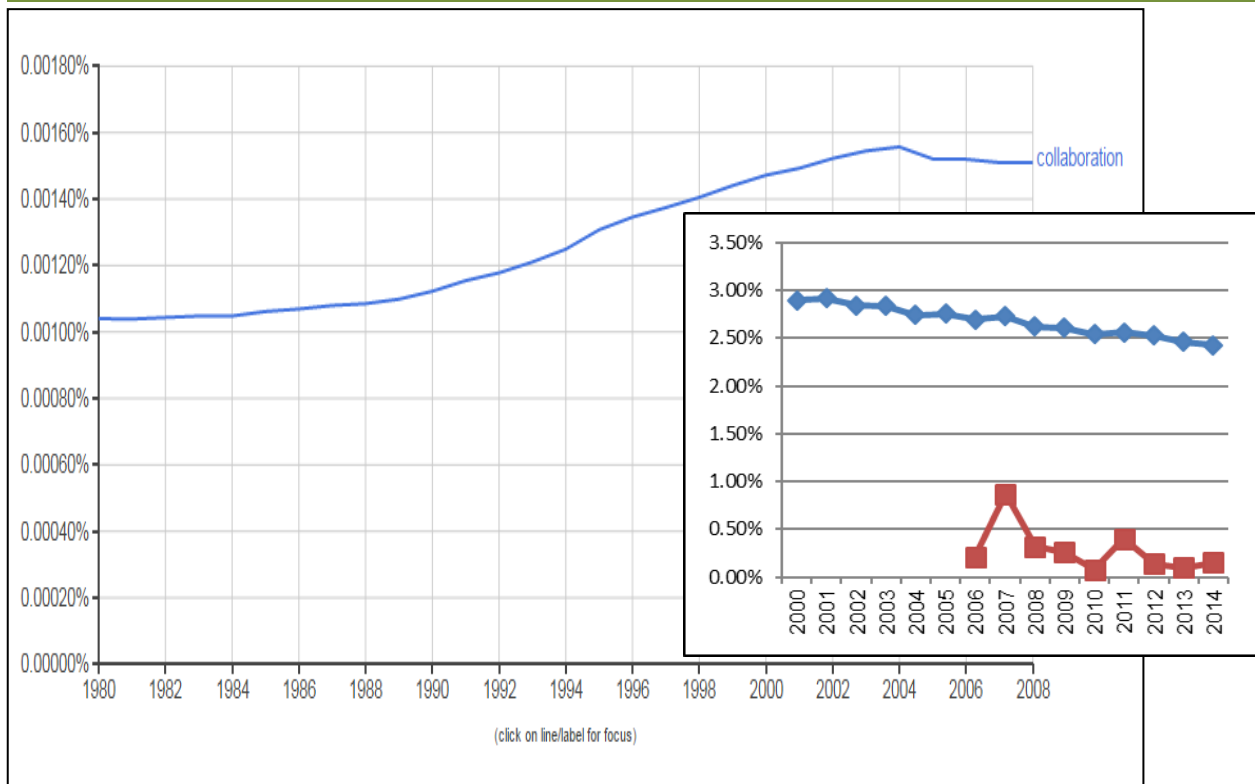


Figure 6. The relative frequency curves of the term "collaboration" as found in three databases: On top: Google Books Ngram Viewer (1980-2008); Middle: Google Scholar (2000-2014), and at the bottom Chais Conference Proceedings (2006-2014)

The mixed methods model for big data discourse analysis suggests that as the scope of a database increases, the presented perspective broadens; hence it is more likely that quantitative methods would provide meaningful results. As the scope decreases, researchers should be aware of the limited perspective and the need for complementing qualitative analyses for interpreting the results would increase.

Discussion

This paper introduced a novel mixed methods model for big data discourse analysis, and contributed to both theory and practice by providing a methodological approach that enables gaining knowledge out of big data discourse analysis, by integrating quantitative and qualitative research methods. Furthermore, the paper demonstrated the potential, as well as the limitations, of culturomics, by applying it for identifying trends of the learning technologies research field during the years 2000-2014. This paper focused on the methodological aspects of the suggested model. Therefore, further study, both quantitative and qualitative is required for elaborating the analysis, interpreting the results, and gaining a deeper understanding of research trends in the learning technologies field.

The implications of the mixed methods model for big data discourse analysis go beyond the demonstrated application of identifying research trends. The mixed methods model could be applied as a useful approach for other sorts of research that involves textual data, as well as for practical applications, such as analyzing commercial textual databases. The five types of term curves (rising; stable; fall; rise and fall; rise and stable) along with the keyness analysis may also serve as useful tools for identifying trends.

Conclusions

Culturomics (Bohannon, 2010; Michel et al., 2011) is an emerging research field, which relies on quantitative analysis methods. This paper suggests that systematically adding a qualitative aspect to a culturomic analysis may considerably improve the potential of gaining insightful findings out of big data discourse analysis, and provides an approach for selecting the appropriate mix of quantitative and qualitative methods. We have empirically shown the potential of culturomics for identifying research trends. Then, by comparing several corpora, we demonstrated that corpora with different scope require a different mix of quantitative and qualitative research methods. Furthermore, the mixed methods model for big data discourse analysis increases the possibilities of performing effective culturomic analyses of databases, which differ in their scope.

Acknowledgement

The authors gratefully acknowledge that this research was supported by the Open University of Israel's research fund (grant no. 504292). The authors would like to thank the research assistant, Daphna Amit.

References

- Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1), 3-9.
- Ahlawat, S., & Ahlawat, S. (2013). An innovative decade of enduring accounting ideas as seen through the lens of culturomics: 1900-1910. *American Journal of Business Research*, 6(1), 63-75.
- Anthony, L. (2011). Antconc version 3.2.4w. Center for English Language Education, Waseda University. Retrieved from: <http://www.antlab.sci.waseda.ac.jp/software.html>
- Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, 25(1), 107-136
- Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257-271.
- Belanger, F., & Jordan, D. (Eds.). (2000). *Evaluation and implementation of distance learning: Technologies, tools and techniques*. Hershey, PA: IGI Global.
- Bohannon, J. (2010). Google opens books to new cultural studies. *Science*, 330(6011), 1600-1600. Retrieved from: http://dericbownds.net/uploaded_images/Science-2010-Bohannon.pdf

-
- Bohannon, J. (2011a). The science hall of fame. *Science*, 331(6014), 143-143. Retrieved from: <http://www.sciencemag.org/content/331/6014/143.3.full>.
- Bohannon, J. (2011b). Google Books, Wikipedia, and the future of culturomics. *Science* 331: 135. 14 January 2011. Retrieved from: <http://www.terceracultura.net/tc/wp-content/uploads/2011/01/culturomics.pdf>
- Chen, S., Kiciman, E. M., & Wang, Y. M. (2008). U.S. Patent Application No. 12/137,508.
- Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business Press.
- Fenn, J. (1999). When to leap on the hype cycle: Research note. Stamford, CT: Gartner Group DF-08-6751. Retrieved from: http://www.cata.ca/_pvw522C275E/files/PDF/Resource_Centres/hightech/reports/indepstudies/Whentoleaponthehypecycle.pdf
- Geri, N, Blau, I., Caspi, A., Kalman, Y. M., Silber-Varod, V., & Eshet-Alkalai, Y. (2015). A decade of Chais conferences: Introduction to the IJELL special series of Chais conference 2015 best papers. *Interdisciplinary Journal of e-Skills and Life Long Learning*, 11, 149-157. Retrieved from <http://www.ijello.org/Volume11/IJELLv11p149-157ChaisPreface.pdf>
- Hart, J. (2007-2014). Top 100 Tools for Learning. *Directory of Learning & Performance Tools & Services*. Retrieved from: <http://c4lpt.co.uk/directory/top-100-tools/>
- Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, 143, 87-92
- Künstler, V., Maiwald, P., and Saage, S. (2008). A Guide to Corpus Linguistics. Retrieved from: <http://fss.plone.uni-giessen.de/fss/faculties/f05/engl/ling/help/materials/restricted/clguide.pdf/file/A%20Guide%20to%20Corpus%20Linguistics.pdf>
- Linden, A., & Fenn, J. (2003). Strategic Analysis Report Understanding Gartner's Hype Cycles. 30 May 2003. Retrieved from: <http://www.ask-force.org/web/Discourse/Linden-HypeCycle-2003.pdf>
- Laurillard, D. (2013). *Rethinking university teaching: A conversational framework for the effective use of learning technologies*. Routledge.
- Michel J.-B., & Lieberman Aiden, E. (2011). What we learned from 5 million books. TED presentation. Retrieved from: http://www.ted.com/talks/what_we_learned_from_5_million_books/transcript#t-407429
- Michel, J.-B., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331: 176.
- Perc, M. (2013). Self-organization of progress across the century of physics. *Scientific Reports*, 3. doi:10.1038/srep01720

-
- Raban, D. R., & Gordon, A. (2015). The effect of technology on learning research trends: a bibliometric analysis over five decades. *Scientometrics*, 105(1), 665-681.
- Radinsky, K., & Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 255-264). Retrieved from: <http://tx.technion.ac.il/~kirar/files/Radinsky-webtorealworld.pdf>
- Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World Wide Web* (pp. 337-346). Retrieved from: <http://www.gabrilovich.com/publications/papers/Radinsky2011WTS.pdf>
- Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *7th International Conference on Statistical analysis of textual data (JADT 2004)* (pp. 926-936).
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.
- Rowley, J. E. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of information science*, 33(2), 163-180.
- Segev, E., Aviv J. S., Baram-Tsabari, A. (2014). Temporal patterns of scientific information-seeking on Google and Wikipedia. In: Y. Eshet-Alkalai, A. Caspi, N. Geri, Y. Kalman, V. Silber-Varod, & Y. Yair (Eds.), *Learning in the Technological Era: Proceedings of the 9th Chais Conference for the Study of Innovation and Learning Technologies*. 193-200 [in Hebrew].
- Silber-Varod, V., Eshet-Alkalai, Y., Geri, N. (2016). Analyzing the Discourse of Chais Conferences for the Study of Innovation and Learning Technologies via a Data-Driven Approach. *Interdisciplinary Journal of e-Skills and Life Long Learning*, 12. (forthcoming).
- Sinclair, J. (1991). *Corpus, concordance, collocation* (Vol. 1). Oxford, UK: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London, UK: Routledge.
- Soper, D. S., & Turel, O. (2012). An n-gram analysis of Communications 2000--2010. *Communications of the ACM*, 55(5), 81-87.
- Soper, D. S., Turel, O., & Geri, N. (2014). The intellectual core of the IS field: a systematic exploration of theories in our top journals. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 4629-4638). IEEE.
- Tahmasebi, N., Borin, L., Capannini, G., Dubhashi, D., Exner, P., Forsberg, M., ... & Mogren, O. (2015). Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4), 169-187.
- Van Den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, 68(3), 377-393.

-
- Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with Wordle. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6), 1137-1144.
- Webber, B. (2008). Computational perspectives on discourse and dialogue. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.). *The handbook of discourse analysis* (pp. 798-816). Hoboken, NJ: John Wiley & Sons.
- Willems, K. (2013). 'Culturomics' and the representation of the language of the Third Reich in digitized German books. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, 18(1), 87-99.
- Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human systems management*, 7(1), 59-70.

Authors' Biographies

Vered Silber-Varod is a Research Fellow at The Research Center for the Study of Innovation in Learning Technologies, The Open University of Israel. She holds a B.A. in Political Science and French language and literature studies from the Hebrew University of Jerusalem, an M.A. with Magna Com Lauda in Hebrew language studies at the department of Hebrew and Semitic languages from Tel Aviv University, and a Ph.D. in Humanities from Tel Aviv University. Her research interests and publications focus on various aspects of linguistics, with expertise in speech prosody, acoustic phonetics, and quantitative analysis of written and spoken texts. Vered has over 10 years of experience as a Linguist at various institutes, including the ACLP – Afeka Center for Language Processing, Afeka College of Engineering. Personal site:

http://www.openu.ac.il/Personal_sites/vered-silber-varod/

Yoram Eshet-Alkalai is a Professor at the Open University of Israel, Department of Education & Psychology. He is Head of the M.A. program in educational technology, and the former Head of the Research Center for Innovation in Learning Technologies in the Open University. He holds a B.A. in Archeology (Hebrew University), M.Sc. in Geology (Hebrew University) and Ph.D. in Earth & Environmental Sciences (City University of NY). For a decade, he was the Head of the Instructional Design Program in the Tel Hai Academic College, and a senior researcher in the Geological Survey of Israel. He has 15 years of experience in developing technology-based instructional solutions for educational systems in Israel and the USA. In this capacity, he was involved in the design of hundreds of simulations, data-bases, tutorials and large-scale curriculum integration projects. His major research and publications interests focus on cognitive aspects of working with digital technologies, digital games and design principles of computer-based learning environments. Personal site:

http://www.openu.ac.il/Personal_sites/yoram-eshet.html

Nitza Geri is an Associate Professor at the Open University of Israel, Department of Management and Economics, and Head of the Research Center for Innovation in Learning Technologies. She holds a B.A. in Accounting and Economics, an M.Sc. in Management Sciences, and a Ph.D. in Technology and Information Systems Management from Tel-Aviv University. Nitza is a CPA (Israel) and prior to her academic career she had over 12 years of business experience. Her research interests and publications focus on various aspects of the value of information and knowledge, as well as adoption, implementation, and continued use of

information systems, including strategic information systems, e-business, information economics, attention economy, knowledge management, value creation and the Theory of Constraints, managerial aspects of e-learning systems. Personal site:

http://www.openu.ac.il/Personal_sites/nitza-geri.html