Challenges Due to Excessive Amount of Online Data and (Mis)Information

Katarina Rojko, Dejan Jelovac

Fakulteta za informacijske študije v Novem mestu

Ljubljanska cesta 31a, 8000 Novo mesto, Slovenia

{katarina.rojko, dejan.jelovac}@fis.unm.si

Abstract *The overload of data and information (including misinformation) that fill the World Wide Web and which are generally in circulation, became our contemporary problem, as information reliability is very often doubtful. Insofar as there is more information in circulation, their usage is reduced proportionally and, consequently, their meaning is reduced proportionally. Digital skills of EU citizens in general are underdeveloped, while amount of digital data grows exponentially. For this reason, constant development of digital competences and ICT literacy as a result of systematic education, social control, criteria for verifying reality, recognition of fake data, etc, is required.*

Keywords digital data, data quality, digital skills, ICT literacy, (mis)information

1 Introduction

Internet enables efficient, cost-effective data collection and facilitates access to large amount of data. However, since its use requires certain skills, among others, Hewson, Vogel and Laurent (2015) wrote a guide how to use internet as a tool for conducting research. Such guides are required since many researchers lack of skills for conducting research, using options offered by immense amount of data available online (Rojko, 2016). Moreover, the entire society has access to online data conditionally, and people are looking for information on World Wide Web (WWW) all the time.

The excessive amount of data is another problem, and every second it is more serious, as data is growing at a rapid pace in contemporary information society. Micro focus (2017) stated that 44 billion GB of data was created per day in 2016, while the growth to 462 billion GB of data created per day is predicted in 2025. Micro focus also revealed that 90% of data created is unstructured, so big data analytics and archiving tools are critical in being able to manage all this data.

Every day, we create 2.5 quintillion bytes of data. To put that into perspective, 90% of the data in the world today has been created in the last two years alone – and with new emerging technologies, devices, and sensors, the data growth rate will likely accelerate even more (IBM Marketing Cloud, 2016). For this reason, we are overwhelmed by the excessive amount of information. Namely, regardless of industry or profession, people need the right information at the right time to make truly confident and well-judged, productive decisions. But due to rapidly growing amount of information, it is harder and harder to separate the "signals from the noise", and to "discern the insights from the hindsight" (IBM Marketing Cloud, 2016).

As the internet has become flooded with untrustworthy information, some of which is intentionally misleading, it is necessary to know, how to recognize misinformation. However, the main question is, how to find and recognize "correct" information? For illustration, in 2016 there was 215 billion of emails sent every day, and 269 billion in 2017, while more than half were spam (Micro Focus, 2017). How can "typical user" deal with this problem? Good digital skills are important, but there is much more to know, and for this reason already since the turn of the century handbooks as "Web of Deception: Misinformation on the Internet" from 2002, by Anne P. Mintz, ed., are published.

There is another dilemma – a reflection on what is "correct" data. Seifert (2017) argues that what must be added to our understanding of misinformation in the "post-truth era" is our experience of misinformation, as the processing of information changed in the "post-truth" world.

In June 2016, the U.K. held a referendum on its membership in the European Union, and in November

2016, the U.S. held its national elections. In the run-up to both of these important decisional events, the internet with its burgeoning collection of information dissemination appli-cations, influenced the decisions of voters (Cerf, 2017). The disturbing aspect of these (and many other decisional events) is the quantity of poor-quality content, the production of deliberately false information, and the reinforcement of bad information through the social media. Besides, people began to read superficially, since because of the excessive amount of data, it is even more difficult to deepen into a single subject to read, and the quantity often wins.

In summer 2017, Pew Research Center and Elon University's Imagining the Internet Center conducted a large canvassing of technologists, scholars, practitioners, strategic thinkers and others, asking them to choose one of the two answer options:

a) The information environment will improve – in the next 10 years, on balance, the information environment will be improved by changes that reduce the spread of lies and other misinformation online.
b) The information environment will NOT improve – in the next 10 years, on balance, the information environment will NOT be improved by changes designed to reduce the spread of lies and other misinformation online.

Out of 1,116 respondents, 51% chose the option b), and 49% chose the answer a). Participants were next asked to explain their answers. In continuation we reveal some of these follow-up responses (Pew Research Center, 2017):

The quality of information will not improve in the coming years, because technology can't improve human nature all that much. Christian H. Huitema

In the arms race between those who want to falsify information and those who want to produce accurate information, the former will always have an advantage. David Conrad

We live in an era where most people get their 'news' via social media and it is very easy to spread fake news. … Given that there is freedom of speech, I wonder how the situation can ever improve. Anonymous project leader for a science institute

In order to reduce the spread of fake news, we must deincentivize it financially. Amber Case

When the television became popular, people also believed everything on TV was true. It's how people choose to react and access to information and news that's important, not the mechanisms that distribute them. Irene Wu

We can't machine-learn our way out of this disaster, which is actually a perfect storm of poor civics knowledge and poor information literacy. Mike DeVito

These responses clearly indicate the problem we are facing today. Some respondents are optimists, while others believe that information environment will not improve, as it is becoming harder and harder to find the right information, and predict that a larger digital divide[1] will form.

The internet's continuous growth and accelerating innovation also allows more people and artificial intelligence to create and instantly spread manipulative narratives. Furthermore, human tendencies and infoglut drive people apart and make it harder for them to agree on "common knowledge" (Pew Research Center, 2017).

Beside structured data, there is also much wider pool of unstructured data and different advanced tools are getting available for analysis of. But unstructured online data analysis and textual analytics require business intelligence skills and use of appropriate software tools. This means that not only digital skills are sufficient; there is also a condition of access to required sources, most often determined by the income level or inclusion in certain communities (Rojko, 2016).

Due to above mentioned concerns we decided to take a focus on needed skills for finding the information online and digital skills are one of the most important for the ability to find reliable information in required time.

2 Research methodology

Within the initial phase of the research, we did a systematic analysis of sources and a review of literature, in combination with experiences obtained during past two decades of active internet usage. This provided us with the theoretical and practical basis for formulation of our hypotheses.

For the purpose of empirical analysis, we used the data from Eurostat, which allows conclusions on reliable official dataset of data and thus increases certainty of our findings. The basis for the analysis

[1] Classical sociological theories of inequality, as well as empirical evidence (Ragnedda and Muschert, 2013), define digital divide as the unequal access and utility of internet communications technologies and explore, how it has the potential to replicate existing social inequalities, as well as create new forms of stratification. They examine how various demographic and socio-economic factors including income, education, age and gender, as well as infrastructure, products and services affect the internet use and access.

were digital[2] skills of individuals, variables that show level of digital competences.

To determine different impacts on digital skills, we also studied certain social exclusion indicators, and compared digital skills' levels to information[3] skills' and software[4] skills' levels.

We used only indicators' values for the last available year, to present the most current situation. We nonetheless encountered certain limitations, e.g. data for the age group under 16 was not available. Nevertheless, qualitative data analysis in combination with the literature review and obtained practical experiences, enabled us to conduct critical assessment of sources and theories, and credible verification of set hypotheses.

3 Research goal

Besides presented impact factors form Eurostat database, also others have significant impact on the ability for successful separation of the right from wrong, misleading information. Those are e.g. the kind of information we are looking for, level of ICT availability, usage and data quality, personal ability to develop a critical distance to obtained information, etc.

The goal of our research was to present the most up-to-date situation to convince readers that this topic requires much greater attention and coordinated actions of all actors involved to decrease levels of poor civics knowledge and poor information literacy. This can be improved only by better awareness, constant education, continuous research and data accessibility. Namely, other studies do not focus on current situation from the angle of challenges and conditions for society's prosperity based on information access and use, although it is generally accepted that "The oil of 21st century is data".

We decided to focus on measurable data, including data on digital skills in frame of European Union (EU) as a whole (28, 15), and we exposed situation in Slovenia and Croatia. We also researched other Eurostat's "Digital Society" data, to find explanations for observed situation, which enabled us to make conclusions that are more credible. Besides, we added some figures on internet size and number of internet users.

Based on initial data and literature review, we have set the following theses:

- Digital skills of EU citizens do not reach sufficient level to enable them to find reliable information on world wide web in required time.

- Development of digital skills is crucial in the contemporary period of excessive amount of online data and (mis)information.

4 Data Analysis

To support our theses, we firstly checked the level of digital skills in EU 28 and EU 15, while we also focused on situation in Slovenia and Croatia. For the purpose of our research, we decided also to consider certain social exclusion parameters that supposed to have impact on the level of digital skills. Moreover, we compared digital to information and software skills, to provide explanations for different rates.

Nonetheless, we are aware that the skills' rates do not answer the question, who is able to find reliable information faster; they only provide measurable indicators for exploring certain obstacles and conditions for it.

Data below thus show the most recent levels of digital skills in EU and other measurable indicators that can serve as the explanation of different skills' levels and provide understanding why certain groups of people have problems to find reliable information on the internet, and why there is a greater potential to mislead them by partial or wrong information.

[2] Persons that have been using internet during last 3 months are attributed a score on four digital competence domains: information, communication, content-creation and problem-solving, depending the activities they have been able to do. The scores are basic, above basic and below basic. Individuals not using internet are classified without digital skills. The four digital competence domains are aggregated in four logical groups (Eurostat, 2018).

[3] Information processing skills refers to the ability to identify, locate, retrieve, store, organize and analyze digital information, judging its relevance and purpose. The indicator is based on five activities internet users have been able to do online during previous 3 months. The scores are basic, above basic and none. Individuals not using internet are classified without digital skills (Eurostat, 2016).

[4] Software skills for content manipulation refer to the ability to create and edit new content (from word processing to images and video); integrate and re-elaborate previous knowledge and content; produce creative expressions, media outputs and programming; deal with and apply intellectual property rights and licenses. The indicator is based on six activities internet users have been able to do during previous 3 months. The scores are basic, above basic and none. Individuals not using internet are classified without digital skills (Eurostat, 2016).

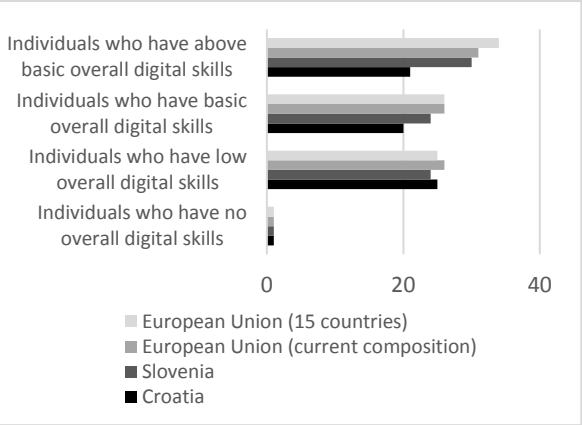

Figure 1. Digital skills of individuals (%) in 2017 (Eurostat, 2018)

Figure 1 shows the rates of digital skills in EU in 2017 and includes comparison with situation in Slovenia and Croatia. In EU 28, 52% of individuals have only low or basic overall digital skills, while 31% have above the basic overall digital skills, and 1% has no overall digital skills. In EU15 digital skills are, as expected, developed better, as 34% of individuals have above the basic overall skills.

Comparison of Slovenia and Croatia to the EU average reveals that both Slovenia and Croatia lag behind with 30% and 21% of individuals with above the basic overall digital skills respectively, while this percentage in EU28 varies from 10% in Romania to 58% in Iceland. Slovenia occupies nineteen spot in this term, while Croatia twenty-sixth, which indicates, that Slovenia and Croatia have to invest stronger in the development of digital skills, to avoid staying the laggards in this term.

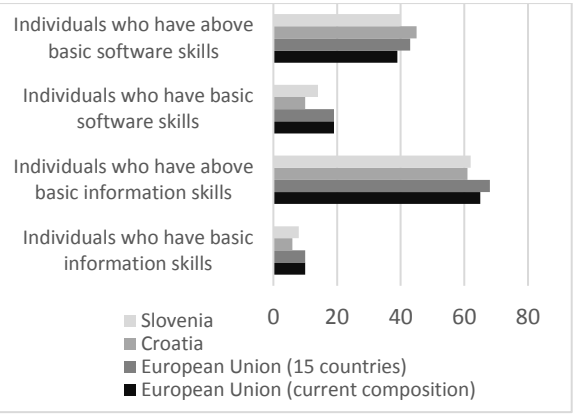

Figure 2. Information and software skills of individuals (%) in 2015 (Eurostat, 2016)

From the Eurostat's (2016) data on information and software skills' rates, we can find out that in 2015 as much as 65% of individuals had above the basic information skills in EU 28, which is more than double in comparison with above the average digital skills (28%) in the same year.

This clearly indicates that digital skills are significantly underdeveloped and that with such insufficient level, a considerable number of individuals cannot exploit their potential to search for digital information successfully.

Even worse situation is observed in terms of software skills' comparison, as Eurostat (2016) data revealed that 42% of individuals in EU28 had no software skills at all in 2015. Furthermore, the data on both information and software skills' rates in total, again reveal a smaller rate of skills in Slovenia and Croatia as in the EU 28.

In sought of the answer, why there is such a big gap between individuals in terms of digital, informational and software skills' rates, we decided to look at three social exclusion indicators (55 to 74 years old; low education; unemployed or inactive or retired) which showed us the correlation (Eurostat, 2016).

We could conclude, that in all three different kinds of skills, there is a big gap between all individuals and individuals marked with at least one or two social exclusion indicators, however still the biggest difference was in terms of information skills - the most "traditional" skill (Eurostat, 2016).

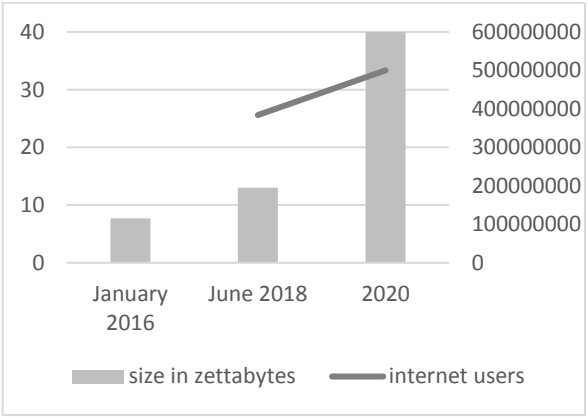

Figure 3. The size of the internet and number of internet users (live-counter.com, 2018)

The size of the internet doubles about every 2 years (live-counter.com, 2018). In July 2018, the counter revealed that the size of internet was 13,4 zettabytes, while, year and a half earlier (in January 2016) it was 7,7. In another year and a half (by 2020), the amount of data is expected to enlarge to 40 zettabytes. It is estimated that by then 50 billion devices will be connected to the internet. Those devices are used currently (July 2018) used by 3,8 billion of internet users, while by 2020, the total number of internet users is predicted to rise to over 5 billion (live-counter.com, 2018).

Data presented in Figure 3, clearly show huge data growth in terms of quantity in past two years, while the projections for the next years do not exhibit any

slowing. This clearly indicates that in the next "Big data years" digital skills will be even more required.

5 Discussion

Based on our research we were able to confirm our first thesis that digital skills of EU citizens do not reach sufficient level to enable them to find reliable information on world wide web in required time. Besides digital skills, also information and software skills are important basis.

Data we used show the most recent levels of digital skills in EU and other measurable indicators that can serve as the explanation of different skills' levels and provide understanding why certain groups of people have problems to find reliable information on the internet and why there is greater potential to mislead them by partial or wrong information.

There are certain conditions that must be met, to be able to look over the WWW for the right information. There is a requirement of having access to and be able to use internet connected device in order to access online sources. Certain groups of people are in disadvantage in this requirement, and for this reason we also considered social exclusion indicators.

Besides, our research revealed a big gap between digital, information and software skills. Comparing only above the basic skills of individuals with high, medium and low education, the smallest difference in skills' rates is among individuals with high formal education, while the biggest difference in skills' rates is among individuals with no or low education.

Moreover, we found out that digital skills are significantly underdeveloped, meaning that with the low rate of digital skills, notable amount of individuals cannot exploit the potential they have with own information skills. Thus much greater attention should be devoted to obtaining digital skills, especially among individuals with no or low formal education, while also among mid and high educated individuals.

Nonetheless, the target value in percentage with respect to the digital skills level in general, should be the same or close to informational skills' rates, we argue. Namely, if these levels would be close or the same, much greater exploitation of internet as an immense source of information would be possible.

Furthermore, there are also other requirements that must be met, as is the knowledge how to check data reliability, having accesses to different databases, and ability to develop critical distance to (mis)information. Here, dissemination of fresh knowledge and solutions by universities and ICT suppliers can help to solve the problem of society on how to find and recognize the "correct" information in Big- data years.

However, since the ability to find the reliable information on WWW is not only a technical issue but also a social issue, as internet changed the ways of information obtaining, constant education from early childhood, technology possession, information and digital skills' development, etc. are required. In addition, the knowledge where to find appropriate data with the ability to access this data is conditional. Here notable difference among different age groups and education level are observed, while also other factors certainly have impact, as income, location, origin, etc.

Based on presented data we also found out, that Slovenia and Croatia lag behind EU in terms of digital skills level, which leads to slower transition to knowledge society. Slower transition to knowledge society might further lead to consequent slower economic growth, as the quick access to information is becoming even more important in globalized e-society.

We could also confirm our second thesis that development of digital skills is crucial in the contemporary period of excessive amount of online data and (mis)information. Namely, the amount of data is growing at a very rapid pace, as presented by the live-counter.com (2018) data, while vast majority data of is available online and generated only in recent years. For this reason, people with underdeveloped digital skills are in big disadvantage, which is also recognized by governments, who regularly provide free options for the inclusion of certain groups of people (usually marked with social exclusion indicators) into the programs which help them to raise the level of ICT literacy and digital skills.

6 Conclusion

In the digital era the ability to find correct information is conditional, as individuals without good skills for online research experience big disadvantage. In addition, the future of internet as a source of reliable information might become questionable, as it has become flooded with untrustworthy information, some of which is intentionally misleading or corrupted.

Only a small share of society will find, use and perhaps pay premium for information from reliable sources. Those will separate from those who are not selective enough or who do not or are not able to invest either time or money in doing so. For this reason, concerns about how vast majority of society will be able to find and use accurate information, are justified.

Moreover people usually do not use or have access to big data software tools, thus the flood of digital data and sources available is making it even more difficult now for them to avoid (mis)information gathering, as was in the period when there was no or less digital data available.

However, the amount of data is growing exponentially, and finding the correct information becomes restricted to WWW users who are able to follow the changing situation in globalized "post-truth" society.

It is necessary to know, how to recognize misinformation. Good digital skills are important, but we also have to be able to recognize poor-quality content, the deliberately false information, and the bad information, many of these spread through the social media.

The rise of "fake news" and the proliferation of narratives that spread online are challenging also publishers and platforms. Some are trying to stop the spread of false information, but their task is Sisyphus' work.

7 References

Cerf, V. G. (2017). Information and Misinformation on the Internet. *Communications of the ACM*, 60(1), 9-9. ACM, NY, USA. Retrieved from https://dl.acm.org/citation.cfm?id=3018809

Eurostat. (2016). *Digital society*. Retrieved from: http://ec.europa.eu/eurostat/data/database

Eurostat. (2018). *Digital economy and society*. Retrieved from: http://ec.europa.eu/eurostat/data/database

Finnegan, R. (2005). *Participating in the Knowledge Society: Researchers Beyond the University Walls*. Palgrave Macmillan UK.

Halfpenny, P., & Procter, R. (Eds.). (2015). *Innovations in digital research methods.* Los Angeles [etc.]: Sage.

Hewson, C., Vogel, & C., Laurent, D. 2016. *Internet research methods.* Second Edi. Los Angeles [etc.]: Sage.

IBM Marketing Cloud. (2016). *10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations.* Retrieved from https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl12345usen-20170719.pdf

Live-counter.com. (2018). *How big is the internet.* Retrieved from http://www.live-counter.com/how-big-is-the-internet/

Micro focus. (2017). *Growth of internet data in 2017*. Retrieved from https://www.slideshare.net/Micro-Focus/growth-of-internet-data-2017.

Ó Dochartaigh, N. (2012). *Internet Research Skills*. Sage.

Pew Research Center. (2017). *The Future of Truth and Misinformation Online.* Report, October 19, 2017. Retrieved from http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/

Ragnedda, M., & Muschert, W. G. (2013). *The Digital Divide: The Internet and Social Inequality in International Perspective*. Routledge.

Rojko, K. (2016). Requirements and obstacles of e-Research, *Research in Social Change*, 8(2), 53-74.

Salmons, J. (2016). *Doing qualitative research online*. Los Angeles [etc.]: Sage.

Seifert, M. C. (2017). The Distributed Influence of Misinformation. *Journal of Applied Research in Memory and Cognition*, 6(4), 397-400. Elsevier Inc.

Van Deursen, A., & van Dijk, J. (2014). *Digital Skills: Unlocking the Information Society*. Springer.

Wishart, J., & Thomas, M. (2016). *E-Research in Educational Contexts: The Roles of Technologies, Ethics and Social Media.* Taylor & Francis.