Journal of Information Technology and Applications

(BANJA LUKA)

JITA

Exchange of Information and Knowledge in Research

APEIRON

THE AIM AND SCOPE

The aim and scope of the Journal of Information Technology and Applications (JITA) is:
• to provide international dissemination of contributions in field of Information Technology,
• to promote exchange of information and knowledge in research work and
• to explore the new developments and inventions related to the use of Information Technology towards the structuring of an Information Society.
JITA provides a medium for exchanging research results and achievements accomplished by the scientific community from academia and industry.

ebscohost.com		road.issn.org	
citefactor.org/contact		citefactor.org	
scholar.google.com		cosmosimpactfactor.com	
doisrpska.nub.rs			
crossref.org			

Printed on acid-free paper

JITA

Contents

EDITORIAL

Dear authors,
Dear readers,

Welcome to the twenty-second edition of the Journal of Information Technology and Applications (JITA), published by Pan-European University APEIRON Banja Luka.

The Journal of Information Technology and Application (JITA) publishes quality, original papers that contribute to the methodology of IT research as well as good examples of practical applications.

JITA journal is a scientific and professional information magazine that aims to promise a forum for engineering, research and academics, the university and industry to showcase their scientific careers and whether it will make them learn and explore.

JITA offers the perfect convenience to share scientific resources, share your own research highlights, and innovate science with the entire scientific world.

Our mission is to promote and establish cooperation and dialogue between academic institutions and research institutes belonging to the field of information technology, which are as follows:

- Promoting the best research output in the academic field.
- Support in the professional training of IT staff.
- Introducing computer scientists with advanced modern technological achievements.
- Announcement of newspapers in the most advanced information technologies.

Journal of Information Technology and Applications is a scholarly open access, peer-reviewed, fully refereed journal, which provides a platform for scientists and academicians all over the world to discuss new issues and developments in different areas of informatics.

The topics covered by this journal include, but are not limited to:

- Advances in Networks
- Artificial Intelligence
- Intelligent Systems
- Computer Modelling
- Operations Management and Information Systems
- Statistics and Informatics
- Mobile Computing
- Machine Learning Research
- Mathematics and Computer Science
- Bioinformatics
- Medical Imaging
- Ophthalmology & Visual Science
- Sustainability Management and Information Technologies
- Discrete Probability Distributions
- Science Development
- Innovation

We can proudly point out that JITA started with the selection of quality articles and has continued as such to this day. Members of the review team come from all continents, as well as the paper authors. With two independent "blind" reviews, we manage to select the highest quality articles and publish them afterwards.

Gratitude
On behalf of the Editorial Board, we would like to thank the authors for their high quality contributions, and also the reviewers for the effort and time invested into the preparation of the Journal of Information Technology and Applications.

Acknowledgments
We would like to thank many who have read and/or commented on earlier versions of this journal including Leonid A. Baranov, Yuri M. Inkov, Hristo Hristov, Sanja Bauk, Dimitris Kanellopoulos and Goran Djukanović. However, any errors or shortcomings remain my full responsibility.

Conflicts of Interest
The author declares no conflict of interest.

Editor-in-Chief, Zoran Avramović
zoran.z.avramovic@apeiron-edu.eu
Editors, Gordana Radić, Dušan Starčević

Demystification of RNAseq Quality Control

Dragana Dudić[1], Bojana Banović Đeri[2], Vesna Pajić[3] and Gordana Pavlović-Lažetić[4]

[1]Faculty of Informatics and Computer Science, University Union Nikola Tesla, Belgrade, Serbia,
ddudic@unionnikolatesla.edu.rs
[2]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia,
bojanabanovic@imgge.bg.ac.rs
[3]Seven Bridges Genomics, Belgrade, Serbia, vesna.pajic@gmail.com
[4]Faculty of Mathematics, University of Belgrade, Serbia, gordana@matf.bg.ac.rs

Abstract: Next Generation Sequencing (NGS) analysis has become a widely used method for studying the structure of DNA and RNA, but complexity of the procedure leads to obtaining error-prone datasets which need to be cleansed in order to avoid misinterpretation of data. We address the usage and proper interpretations of characteristic metrics for RNA sequencing (RNAseq) quality control, implemented in and reported by FastQC, and provide a comprehensive guidance for their assessment in the context of total RNAseq quality control of Illumina raw reads. Additionally, we give recommendations how to adequately perform the quality control preprocessing step of raw total RNAseq Illumina reads according to the obtained results of the quality control evaluation step; the aim is to provide the best dataset to downstream analysis, rather than to get better FastQC results. We also tested effects of different preprocessing approaches to the downstream analysis and recommended the most suitable approach.

Keywords: data preprocessing, Illumina sequencing, NGS analysis, quality control, sequence analysis, total RNAseq.

Introduction

High throughput sequencing technologies provide a way of studying the structure of genetic material, both DNA and RNA. Complexity of the sequencing procedures results in error-prone data sets, which need to be properly treated in order to obtain relevant results from downstream analyses. This process is more complex in the case of RNA sequencing (RNAseq), as it includes additional step of reverse transcription of RNA molecules to complementary DNA (cDNA), after which the common steps (as for DNA sequencing) are performed: amplification, fragmentation, purification, adaptor ligation, and sequencing. On the other hand, RNAseq today is intensively used for a number of analyses, such as characterization of transcriptional activity, quantification of gene expression, differential gene expression, analysis of alternative splicing, functional anal-

ysis, gene fusion detection, etc. Because of demands for the greatest possible reliability of data that will be used in such analysis, it is of the highest importance to estimate the quality of obtained reads and how the quality of reads will or could affect final results of the analysis [1]. In many cases, it is better to omit low quality reads from further processing than to cause the misinterpretation of data.

Quality control consists of two steps: evaluation and preprocessing. The evaluation step consists of a number of metrics which indicate quality of assessed raw reads, while the preprocessing step includes sequence filtering according to results of the evaluation step. Since there is a plethora of available tools, it is challenging for new users to appropriately choose between these tools and to adapt to new ones. Some of them are general tools, meant for both DNA and RNAseq raw reads, including both steps of

quality control like NGS QC Toolkit, FASTX-Toolkit, PRINSEQ, QC-Chain, FaQCs, HTQC and others are designed just for one of the tasks. While there are plenty of tools for preprocessing available, i.e. Trimmomatic [2], SolexaQA, Trim Galore!, Cutadapt, DeconSeq, ConDeTri, Sicle, Scythe, Seqtk, SortMeRNA, BBDuk and BBSplit, VecScreen, Kraken, none of them performs all the preprocessing tasks. On the other hand, there are just several tools designed for the purpose of evaluation of raw reads (FastQC [3], seqTools, fastqp) that provide standard metrics for estimation of raw data quality. Discovery of contaminant sequences (sequences that originated from other organisms than the one that was sequenced) demands specific evaluation tools and they include tools like FastQ Screen and VecScreen. Because it is necessary for quality preprocessing and further steps of downstream RNAseq analysis to estimate error probability in raw reads, we focus on raw reads quality evaluation. The vast majority of researchers today use FastQC [3] in combination with some of the tools specially designed for filtering NGS data. However, FastQC operates by creating flags on datasets based on several metrics and their expected values in the case of DNA experiments which is often misleading for RNAseq experiments. Additionally, special attention should be paid in the preprocessing step in order to enable the assessment of reliable results for further analysis of RNAseq data because the stringent approach, which is widely used, might not be the best choice for transcriptome data.

In this paper, we address the usage and proper interpretation of metrics for RNAseq quality control, implemented in and reported by FastQC, such as Per base sequence content, Per sequence GC content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, and Kmer Content and provide a comprehensive guidance for their assessment in the context of total RNAseq quality control of Illumina raw reads. Also, we give recommendations on how to adequately perform the preprocessing step of raw total RNAseq Illumina reads according to the obtained results of the evaluation step, with an aim not to get better FastQC results, but to provide the best dataset to downstream analysis.

QC evaluation

Evaluation of raw reads is the first step in sequencing analysis and serves to determine the validity of sequenced data. One of the mostly used tools for quality control is FastQC, developed by the Babraham Institute in Cambridge. It is included in many bioinformatics software (Galaxy, Illumina BaseSpace, GenePattern, Chipster, Yabi, Taverna, KNIME, Tavaxy, BioDT), and in that way it became a sort of a standard tool for evaluation of NGS data.

This tool takes fastq, SAM or BAM file as input, quickly performs quality analysis of provided data and outputs results in html format. Quality analysis consists of 12 metrics (Basic Statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, Adapter Content, Kmer Content). Most of the metrics are represented graphically, only Basic Statistics and Overrepresented sequences are presented in tabular format. The exception is the Kmer content module, which is shown in both ways (graphically and in tabular format). For each metric a section in the resulting report is made, flagged as 'PASSED', 'WARNING' or 'FAILED'. The flags are given based on the expected values for sequencing DNA data. At FastQC download page examples of good and bad data for different sequencing platforms are available. Illumina platform is widely used for conducting RNAseq experiments, and for Illumina data the FastQC author suggests that a good report should be mostly flagged as PASSED - only one checkpoint (Kmer content) is flagged as WARNING. However, since the FastQC tool expects diversity and randomness in data even a slight deviation will issue a warning, while severe one will result in failure. Because the correct interpretation of the FastQC report is crucial for other steps of the analysis, flags provided by the program itself cannot be taken for granted. Although the checkpoints like Basic Statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base N content and Adapter Content have universal interpretation and high quality data should pass all of them, other checkpoints are specific for different types of sequencing experiments (DNAseq or RNAseq). All of the specific checkpoints
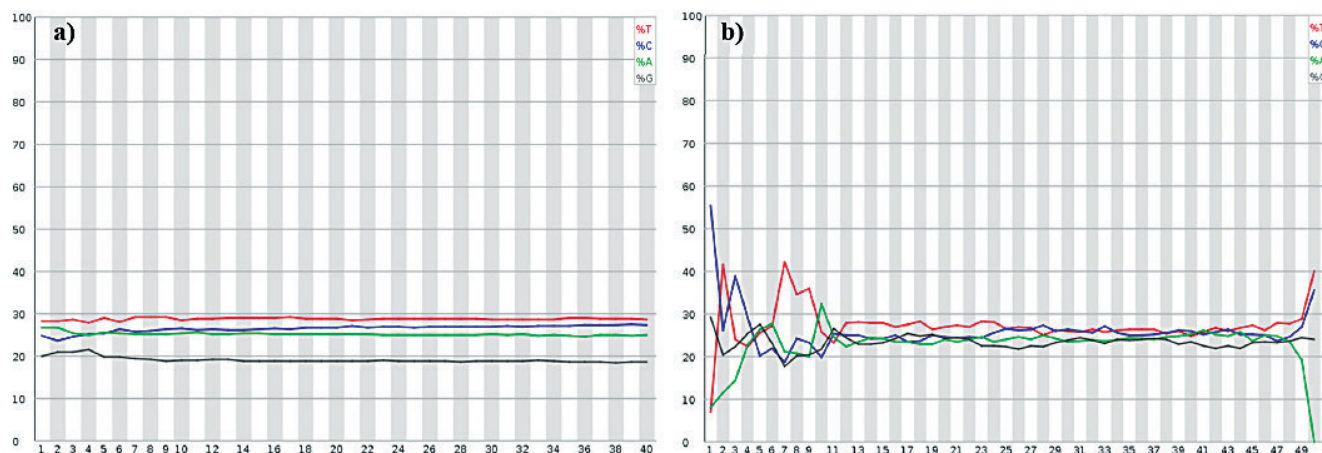
Figure 1. *Per base sequence content graphs showing the read position on X-axis and percentage of each base on Y-axis, a) DNA-seq Illumina data - passed b) total RNAseq Illumina data - failed*

will be discussed briefly in the context of Illumina RNAseq experiments evaluated with FastQC v0.11.5.

Per base sequence content

This checkpoint shows the percentage of each base in each position and it is represented with a graph. According to FastQC interpretation little to no difference is expected between representations of bases, presented as four smooth lines around 25% (Fig. 1a). Warning is issued if the difference between base pairs is greater than 10% in any position, while failure is shown if this difference is greater than 20% (Fig. 1b). Analysis of the most of Illumina RNAseq data issues a failure in this checkpoint, usually not because of any of the common reasons. As stated in help files, but not pointed out in results, the main reason of such failure is the way in which RNAseq libraries are produced (by priming using random hexamers, which is why the bias at the start positions is expected). This is particularly the characteristic of the first 10-13 positions at the 5′-end Illumina RNAseq data and read count reweighting scheme is proposed in order to reduce the impact of this bias [5]. Alternative approach would be to use oligo(dt) priming, but the same study showed that, in this case, data would be highly biased toward 3′-end, and that bias cannot be easily mitigated. Presence of polyA/T tails is another cause of bias at the 3′-end, unrelated with the way of priming. Also, we noted that the abscissa of the representation plot is not equally divided; first nine positions are given separately for each data, and the longer the read the more distant other points become, and sometimes

even given as ranges. For long reads this gives a false impression that the data at the start positions are more deviated.

All Illumina RNAseq data will issue a warning or failure in the Per base sequence content module. The choice of further steps depends on the main goal of sequencing and the nature of available referent resources for the organism under investigation. If de novo assembling is in plan, it is advised to remove first 10-13 bases, while in other cases these bases should be retained.

Per sequence GC content

This plot shows the GC content of each sequence for each position compared to the modeled normal distribution, because it is expected that the random library has a nearly normal distribution. It is misleading to interpret the modeled distribution as a curve that shows information from the reference genome or transcriptome of the sequenced organism. This is just a Gaussian distribution parameterized according to the mean and variance of the GC content of the provided reads. Reads will pass this checkpoint if the GC content curve does not deviate too much from the modeled distribution (Fig. 2a). If sequences outside of the normal distribution comprise more than 15% of the total, FastQC will raise a warning, while failure is given if these reads comprise more than 30% of the total (Fig. 2b).
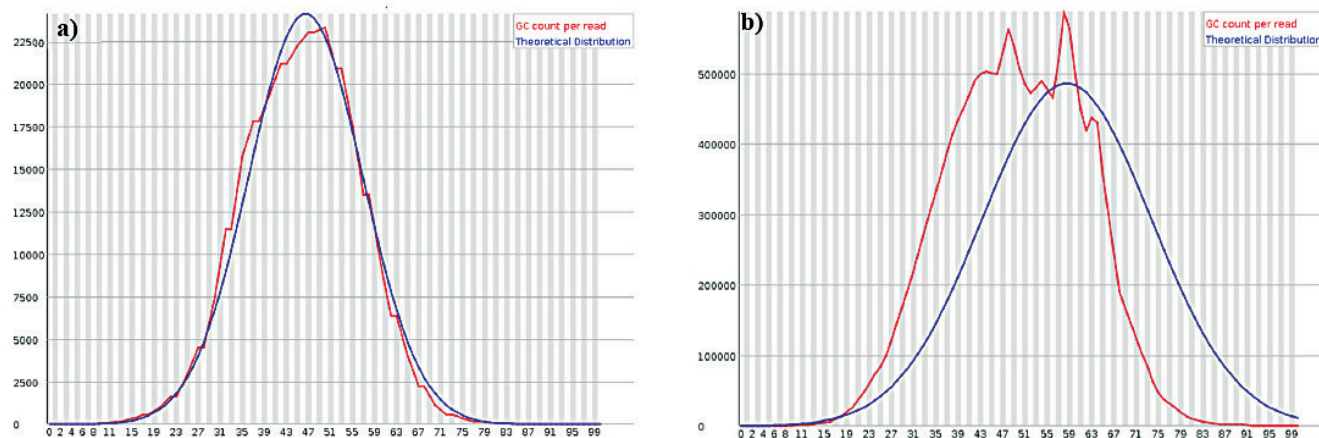
Figure 2. Per sequence GC content shows the mean GC content on X-axis, while the number of reads is shown on Y-axis; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-failed

Different types of deviation have different manifestations. If the plot is shifted in comparison with the modeled distribution, the reason is systematic bias and FastQC will not flag this as an error. Another reason for the shifted distribution could be the presence of polyA/T tails (when shifted to the left) and rRNA richness (when shifted to the right) in data [6]. On the other side, the same reasons could manifest like shoulders or peaks in the distribution plot, only in that case it will be reported as error, because any distribution that has peaks and/or is not unimodal is by default addressed as an error. In the distribution plot, different kinds of contamination (adapter dimers and sequences from other organisms) are usually represented with peaks. The shape of the distribution curve is also affected by the short reads, characteristic for total RNAseq reads, and duplicate-rich reads, which are characteristic for all RNAseq experiments. Additionally, in RNAseq data for organisms with highly repetitive genomes GC content varies due to the presence of some classes of transposons which are GC rich [7][8]. As a consequence of aforementioned, GC content plot could take some form of bimodal distribution. These flaws affect many metrics used in QC and a common way of dealing with these flaws of RNAseq data is to perform, after removing contamination, de novo assembling of reads into larger contigs which should mitigate or even completely eliminate the second peak and smooth the main peak.

Sequence Length Distribution

For sequence length distribution it is expected to be uniform (Fig. 3a). If the read length is variable, warning is issued (Fig. 3b), and failure is caused due to the presence of sequences of zero length.
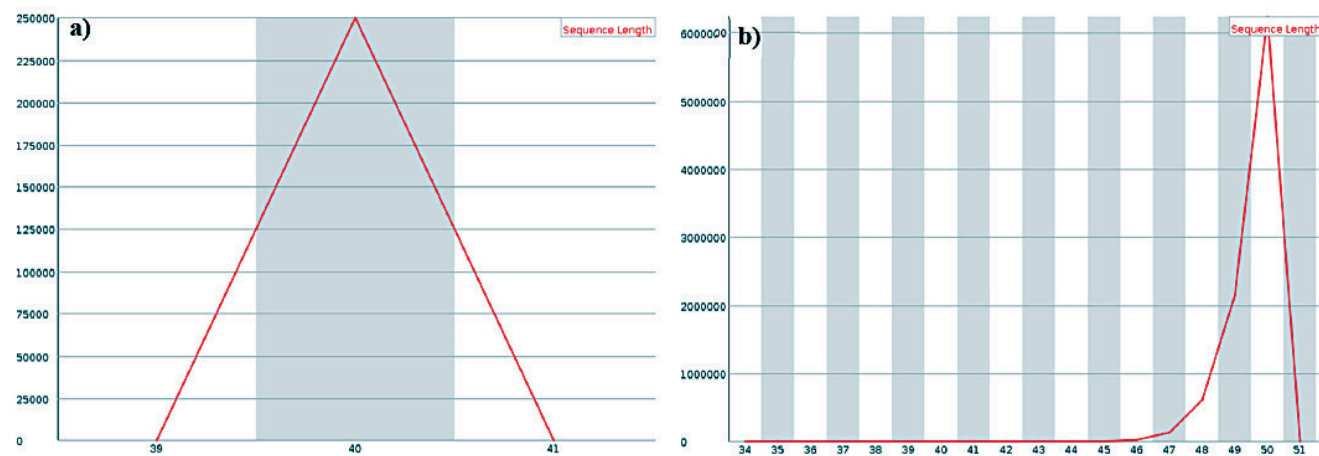

Figure 3. Sequence length distribution shows different sequence lengths on X-axis and Y-axis presents number of reads; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-warning
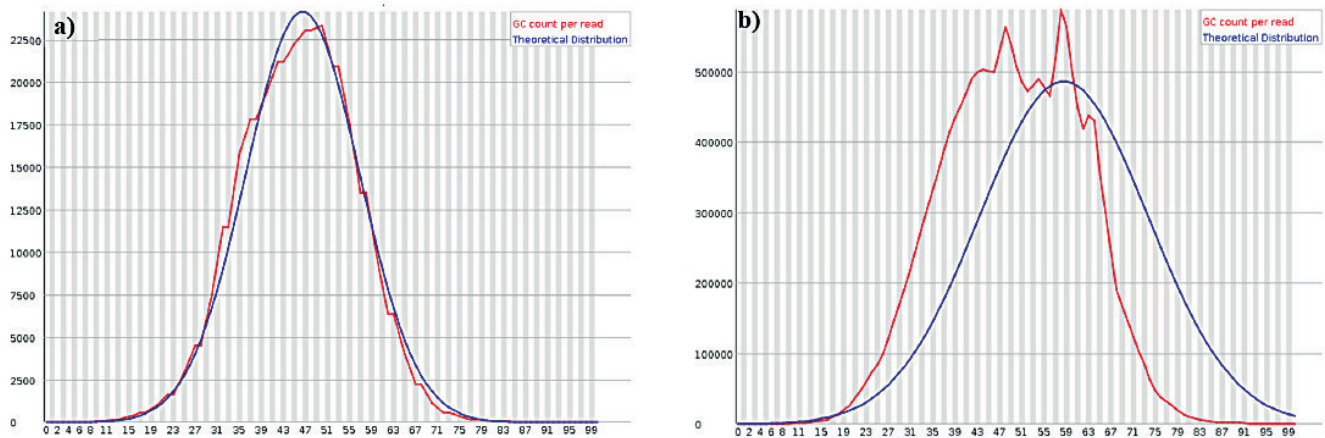
Figure 4. Sequence duplication level shows different degrees of duplication on X-axis and percentage of the duplication is shown on Y-axis; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-failed

The Help files specify that for some high throughput sequencers it is usual to have different read lengths although this will give a warning. For Illumina it is expected to have the equal length for all reads, but in the case of total RNAseq, some RNA molecules are shorter (like small and microRNA), which is why warnings in this module can be ignored.

Sequence Duplication Levels

This checkpoint estimates the degree of duplication using an exact sequence match in a specific way. Common duplicates include PCR and optical duplicates. PCR duplication can arise in transcriptome data because of the specific way of sample preparation and optical duplicates which are the consequence of reading the same cluster twice or more times on sequencer.

Due to reducing memory usage, only sequences that firstly occur in the first 100,000 sequences are taken into account. Additionally, sequences longer than 75 bp are truncated to 50bp. The aforementioned optimizations should provide a good estimate of the data, but this is not the case for RNAseq reads. RNAseq reads tend to be several times longer. Because the transcriptome data include products of alternative splicing, it is possible that different RNAseq reads have some of the first bases the same and the rest of the sequence different. Also, in RNAseq data we expect a plethora of highly expressed sequences and for highly expressed sequences it is expected to have multiple occurrences in the dataset. Regardless, if duplicate sequences make up more than 20% of the sequences taken into account,

FastQC will issue a warning, and if there is more than 50% of the duplicate sequences, this checkpoint will raise an error (Fig. 4b).

In section "Common reasons for warnings" of FastQC Help files it is indicated that warnings for RNAseq data could arise because of the nature of the experiment. We state, based on the aforementioned, that if PCR duplicates and contaminants problems are excluded, any kind of error in this module does not affect the overall quality of the RNAseq data even if duplication level is higher than 50%.

Overrepresented sequences

All sequences that represent more than 0.1% of total sequences are labeled as overrepresented and presented in a table with count, percentage and possible source of the sequence (Fig. 5). Warning will be issued if at least one overrepresented sequence is found. A failure will arise if there is at least one sequence that represents more than 1% of the total data.

To conserve memory, this checkpoint looks for candidates for overrepresented sequences in the first 100,000 sequences, but candidates are tracked through the whole dataset, therefore more overrepresented sequences could be found in the dataset. Databases of known contaminants are queried by overrepresented sequences that were found and best hits are being reported. Since Illumina RNAseq library preparation includes adapters (that are usually removed, but sometimes some of them remain in the sequences), this module will label them and they should be removed from the data.

Sequence	Count	Percentage	Possible Source
CTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGGT	57399	0.6247347688043816	No Hit
CCCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATAT	37086	0.40364664255264543	No Hit
CCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATATTGATG	36448	0.3967026055050105	No Hit
CCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	31942	0.34765898334726314	No Hit
CTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACCTT	30483	0.33177912433080653	No Hit
CTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATATTG	29650	0.3227126935146939	No Hit
CTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	28330	0.30834572031269064	No Hit
CCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATATT	28306	0.308084502618108877	No Hit
CCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATATTG	24959	0.2716555182945445	No Hit
CAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	23543	0.25624367431421374	No Hit
GGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCAGC	16550	0.18013136855541934	No Hit
CTGCTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGT	14718	0.16019175120233609	No Hit
GTCTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACC	14691	0.15989788129593144	No Hit
GCTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGG	14422	0.15697006630249294	No Hit
CTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCCATATTGATGC	13736	0.14950359386569428	No Hit
CTCGCTATGTTGCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATC	13654	0.14861110007587286	No Hit
GCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGAT	11675	0.12707152434347557	No Hit

Figure 5 Overrepresented sequences in totalRNAseq data - warning

Kmer Content

In total RNA sequencing a lot of different types of molecules are sequenced. Because of the presence of small RNA, mitochondrial and chloroplastic (in case of plants) RNA, residual rRNA (mostly removed during a library preparation, but some rRNA may still remain), overrepresentation of the sequences is expected to some extent. Also, alternative splicing and differential gene expression affect the number of overrepresented sequences. Taking into account the aforementioned, we conclude that warnings and failures can be ignored as long as adapters are trimmed.

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
AAGAACG	2050	0.0	26.984587	43
TTCGGCG	620	0.0	24.305752	44
AGACCGC	2055	0.0	24.061537	4
GACCGCG	2085	0.0	23.81967	5
GGACCTT	8550	0.0	23.700302	44
ACCGCGT	2110	0.0	23.532833	6
CGCGTTC	2130	0.0	23.516998	8
CCGCGTT	2120	0.0	23.422468	7
CAGACCG	2275	0.0	22.404818	3
TACGGAG	2230	0.0	22.28056	34
ATACGGA	2250	0.0	22.083471	33
AATACGG	2270	0.0	21.984428	32

Figure 6. List of k-mers

The K-mer module is similar to the overrepresentation module. This checkpoint seeks for portions of the sequences, 7-mers that show positional bias according to the binomial test. In that way it could identify parts of sequences that can cause problems in further analysis. It is represented dually. For graphical representation it is characteristic that values on the X-axis are unequally represented. In tabular representation count, p-value, frequency and position in the read are shown, enabling to reconstruct the longer kmers (Fig. 6).

Because this type of analysis is slow, only 2% of the data is processed. The sequenced data is considered good if the presence of k-mers is balanced. If a binomial p-value is less than 0.01 for any k-mer, warning will be given (Fig. 7), and if p-value is less than 0.00001 a failure will occur.

This is another module affected by the Illumina bias at the 5'-end of reads and adapters and polyA tails at the 3'-end of the RNAseq reads, but these errors can be ignored as long as adapters are trimmed.

QC preprocessing

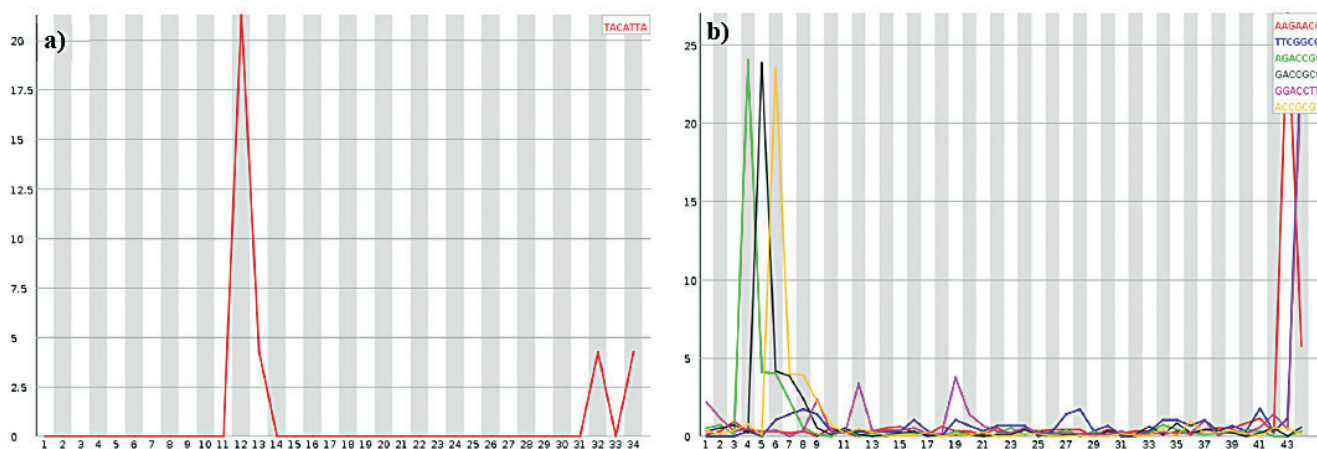Because regions of low quality carry less information of interest, the main step in preprocessing

Figure 7. K-mer content shows position in the reads on X-axis and relative enrichment of a k-mer on Y-axis; a) DNA-seq Illumina data-warning; b) total RNAseq Illumina data-failed

is to deal with these kinds of sequences. Other steps in the preprocessing of NGS data depend on results obtained by the QC evaluation software. As stated in previous section, Illumina's characteristic 5' nucleotide bias, PCR and optical duplicates, polyA/T tails, rRNA and different kinds of contaminants, like adapters and genomic sequences that belong to different organisms than the sequenced one, are recognized by the QC evaluation tools as errors in data that should be removed in the QC preprocessing step. Many QC preprocessing tools have similar functionalities but they differ in speed, available resources (like adapter files) and effectiveness.

Processing of low quality data

The presence of low quality data is mainly the result of systematic errors and in the case of Illumina platform, common error is substitution. There are three general approaches in the preprocessing of low quality NGS data: error correction, masking and trimming. For error correction, there are three groups of methods applicable to Illumina data: k-spectrum-based methods (Rcorrector [9], Quake, Reptile, Hammer, Musket, Bless, Bloocoo, Lighter, Trowel), multiple sequence alignment (SEECER Karect, Coral, ECHO) and suffix array/tree-based methods (Fiona, SHREC, HSHREC, HiTEC, RACER) with only SEECER and Rcorrector being specially designed for RNAseq data. Brief description and evaluation of all methods can be found in [10], while the evaluation of different k-spectrum-based tools, as the dominant method for Illumina data, is available in [11]. Because k-spectrum-based error correction

method uses only local information from sequenced reads, it is not suitable for analyses that include the read mapping task since mapping is done globally. On the other hand, this method is very useful for analyses with de novo assembling in plan because it reduces complexity of created graphs and speeds up the whole process of assembling [12]. Moreover, this error correction method is an integral part of some assembling tools like SOAPdenovo, ALLPATHS_LG, SGA and SPAdes.

When downstream analyses include read mapping, common approaches for dealing with low quality bases are masking and trimming. Masking is a non-invasive process of dealing with low quality data which implies substitution of low quality bases with N's. This is mostly performed with in house scripts, but there is also a tool, FASTQ Masker, included in FASTX-toolkit.

Although masking is a lenient method and it is shown that masking is more effective than trimming in SNP detection [13], trimming is the commonly used technique for this task. Trimming implies removing low-quality parts of sequences and because this is the main task of QC preprocessing, most of the general QC preprocessing tools have this functionality. Window-based (ConDeTri, PRINSEQ, Sickle, SolexaQA, Trimmomatic) and running sum (Cutadapt, SolexaQA with -bwa option, Seqtk) are two types of algorithm families used in trimming software, where the first approach gives better results for common RNAseq tasks like SNP calling, gene expression analysis and de novo assembly [14]. Trimming has different levels of stringency ranging from

mild, with removing bases with quality less than 5 (recommended by [15]), to more aggressive approach with quality threshold between 20 and 30 (recommended by [14]). Because the main goal of RNAseq is to get as much as possible information from the data, we recommend a milder approach for trimming.

Clipping 5' end

To overcome Illumina 5' end sequence content bias caused by random hexamers (presented on Per base sequence content plot), two approaches are suggested. One of them is a correction method specially designed for solving this kind of problem, described in   [16] and the other one is clipping. Although clipping is always a loss of information (because it includes removing the correct bases too), this is a widely adopted QC technique for eliminating 5'end Illumina bias. Alternative to overcome this problem is not to eliminate the bias. Rationale behind this is that we can try to conduct downstream analyses without clipping the leading bases, and if we get poor results, we can always go one step back and use one of the aforementioned methods to overcome it.

Adapter removal

In total RNAseq it is common that sequenced fragments are shorter than the read length, which may result in adapter sequences left over. The presence of adapters is checked in QC evaluation, and if adapters are highly abundant, they will be listed in overrepresented sequences. Also, the presence of adapter sequences affects GC content and K-mer content plot. According to the aforementioned it is clear that all adapter sequences should be removed. This can be done with any general QC preprocessing tool.

Removing contaminants

Contamination of RNA sequences may arise from different reasons and it can affect several QC evaluation checkpoints, like Per sequence GC content, Sequence duplication level and Overrepresented sequences. If the goal is just to identify contaminants the appropriate tool is FastQ Screen or VecScreen, while tools like DeconSeq and BBSplit are used for actual decontamination. Another method

for removing contaminants includes read mapping to possible contaminants genomes, but the same method can be used in order to check the percentage of contamination. For any kind of contamination detection or removal it is necessary to provide sequences/genomes of possible contaminants. Common contaminants in Illumina total RNAseq is PhiX control viral DNA, Illumina TruSeq primers, vectors like plasmid, phage, cosmid, BAC, PAC, YAC and transposable elements from the cloning host which is usually Escherichia coli or yeast. Other possible contaminants are the result of impurities in the RNA sample and include microbes and other organisms that are being sequenced in the lab. The list of possible contaminants is not finite, and we cannot know in advance which contaminants are expected to be present in the dataset.

Complete list of contaminants can be obtained using BLAST, but this is a time consuming task because of the huge number of sequences in the RNAseq dataset. Similar, but less time consuming approach would be to randomly select some reasonable number of sequences (eg. 500) from a transcriptome dataset and to blast them. Another approach would be to execute initial mapping and to BLAST unmapped sequences. Anyway, if contaminants do not comprise a significant number of sequences they will not affect the read mapping step, but for assembling they should be removed.

Removing PCR and optical duplicates

As stated in the previous section, although duplicates affect sequence duplication levels plot and GC content plot, removing them in RNAseq data can be more harmful than useful because read counts play a significant role in the downstream analysis of RNAseq data. If there is a good reason for removing the duplicates, one should be aware that, because it is much easier to locate duplicates after then before read mapping, available tools require a BAM file as input. In order to identify duplicates, Picard EstimateLibraryComplexity can be used. Widely used tools are Samtools rmdup (for removing PCR duplicates) and Picard MarkDuplicates (for marking or removing optical duplicates).

Removing polyA/T tails and rRNA

During the process of cleavage and polyadenyl-

ation, mRNA is enriched with polyA/T tails near the 3'end. The presence of polyA/T tails in the RNAseq can be seen in several modules of FastQC report: Per base sequence content, Per sequence GC content and Kmer content. PolyA/T tails can be removed by tools like PRINSEQ (options –trim_tail_right and –trim_tail_left) or with in house scripts. But, one should have a good reason for removing polyA/T tails because although they affect QC evaluation, they have no influence on read mapping (most mappers are not affected by presence of polyA/T tails) and, actually, they are beneficial for de novo assembling (because the polyA/T tail marks the end of transcript).

As part of the transcriptome, rRNA is expected to be present in the dataset to some extent. High abundance of rRNA can affect some checkpoints in QC evaluation, like GC content plot and overrepresented sequences, and they are often flagged as contaminants. Usually, rRNA is not of interest for downstream analysis and, as all other contaminants, it can be removed from the dataset with specific tools (SortMeRNA, BBDuk) or with mapping to the set of rRNA sequences available in sources like SILVA and Rfam database. But, the removal is not necessary because the high level of rRNA presence in RNAseq data does not represent actual contamination of the dataset. It just means that the dataset contains less mRNA than it is expected and because the presence of rRNA doesn't affect the next step of analysis (read mapping or de novo assembling), it is not necessary to remove rRNA from the dataset. Nevertheless, the exact level of rRNA presence in the dataset should be determined in order to resolve the amount of RNA of interest, using tools for identifying the level of contamination like FastQ Screen.

Sometimes rRNA has influence on specific RNA-seq analysis and, in that case it can be masked. For example, in order to increase FPKM values (which have influence on differential gene expression) it is desirable to exclude rRNA, tRNA, mitochondrial and chloroplastic RNA and similar elements from analysis. This can be done with Cufflinks tool using –M option.

Use case – evaluation of QC preprocessing methods

According to the aforementioned, we formed two groups of proposed QC preprocessing steps:

lenient and stringent, both intended to be used afterwards for read mapping and de novo assembling. Lenient group for mapping included: (1) trimming a read when average quality over a 4bp sliding window drops below 5 – Q5, (2) combination of Q5 with removal of adapters – AQ5, (3) combination of contamination removal with AQ5 – CAQ5, while stringent one comprised: (1) trimming a read when average quality over a 4bp sliding window drops below 25 – Q25, (2) combining adapter removal with Q25 – AQ25, and (3) removal of contamination combined with AQ25 – CAQ25. If the ending dataset contained reads shorter than 25bp, such reads were discarded. For the assembly, we used the dataset with removed adapters and contamination (CA) as a starting point. Lenient group was composed of the CAQ25 dataset and error corrected CAQ5 dataset (EC). Stringent group was formed by removing the first 11 bases from EC and CAQ25 datasets forming ECC and CAQ25C datasets, respectively. With lenient QC preprocessing steps, the resulting dataset will retain more information, and with stringent one the preprocessed dataset will consist of higher quality sequences. Both groups of approaches were compared to the naïve approach which refers to raw data analyses (R).

The dataset

To test the impact of two groups of QC preprocessing steps, we used a publicly available dataset of human breast cancer transcriptome with NCBI SRA accession number SRR2753165. This dataset contains raw Illumina Hiseq 2500 35-50bp single end transcriptome data, with cDNA Library being constructed using TruSeq stranded total RNA with Ribo-Zero Gold (for rRNA removal). Such useful information can guide us in determination of used adapters and primers as well as in the understanding of some QC evaluation results. In this example, the shape of the curve in the sequence length distribution plot is expected to be as in Fig. 3b because sequences are of variable length. Also, because the dataset is composed of transcriptome sequences, the peaks in the right part of the sequence duplication level graph are expected (as in Fig .4b).

Other FastQC results of interest are presented in Figures 1b, 2b, 5, and 7b. Deviations in 3'end in distribution of bases in sequences (Fig. 1b), in k-mer

content plot (Fig. 7b) and shift to the right in the GC content plot (Fig. 2b) imply the high abundance of polyA/T tails. We used NCBI BLAST [17] to determine origins of overrepresented sequences (Fig. 5). In this case, they were all small nuclear RNA (snRNA): 16 of them were 7SL RNA (type of SRP RNA and part of Alu transposable element), 3 were 7SK RNA, 3 were uncharacterized SRP RNA and 1 was uncharacterized snRNA. Both 7SL RNA and 7SK RNA are highly abundant GC-rich sequences. SRP RNA defines perinucleoar compartment [18] which is known to be present in breast cancer genomic data [19]. This explains the peculiar shape of GC content plot (Fig. 2b) to some extent. Other reasons include higher number of duplication sequences and abundance of rRNA sequences in the dataset.

To get deeper insight in the dataset content, we checked the level of contamination using Fastq Screen (Fig. 8). In order to determine the list of possible contaminants from other species, we sampled 500 reads from unmapped reads and blasted them against the human reference genome. Sampled sequences that remained uncharacterized were blasted against BLAST nt database. We detected contamination with Enterobacteria phage phiX174, so we used a standard list of contaminants. With less than 1% of PhiX and vectors from the UniVec [20] database, we may state that the dataset contamination is low. Further steps depend on the choice of the following step in downstream analyses. For the

mapping step there is no need to clean the dataset because the contaminants will not map to the reference sequence. On the other hand, the assembly step requires as much clean data as possible and removal of any amount of contaminants is required. Also, it should be noticed that nearly 1% of the dataset consists of rRNA, which is not severe, but it surely influences the shape of GC content graph.

Presence of polyA/T tails can easily be checked by using simple grep command. In this dataset, around 7% of sequences have some form of polyadenilation. When the reads are short, the tails can be sequenced through and that causes the occurrence of other nucleotides after polyA/T tail. But, this is not a problem, because many read mappers and de novo assemblers perform well with polyadenilation, so it is not necessary to remove it from the dataset in any case.

As it is shown in Fig. 5 and Fig. 8, there is no significant abundance of adapters. Nevertheless, all adapters should be removed from the dataset, independently of the choice of the following steps.

For quality trimming, to remove adapters and to discard reads shorter than 25bp (which can arise after quality trimming and adapter removal), we used Trimmomatic v0.36 because it is easy-to-use, fast, lightweight window-based QC preprocessing tool, specially designed for Illumina data. To remove PhiX and vector sequence contamination we used read mapping with TopHat v2.1.1 [21].
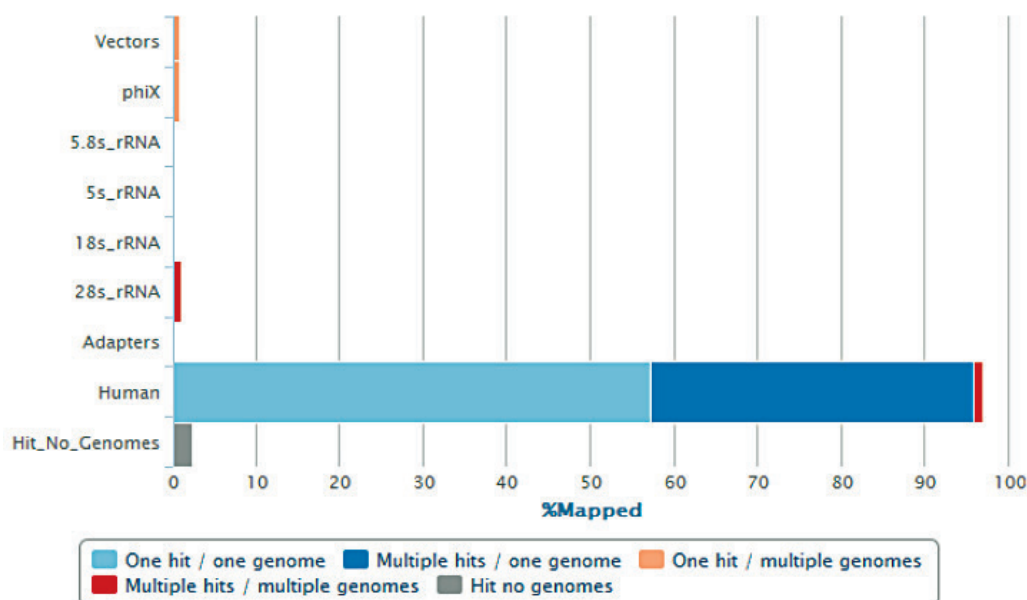

Figure 8. Level of contamination using Fastq Screen

Mapping

In order to compare different approaches and their benefit to the results of mapping step, we chose widely used read mapper Tophat2 and compared its performance when using default parameters, as is done in many studies, with its performance when using chosen set of parameters that contribute to the higher accuracy and confidence of results. As a reference genome we used GRCh38.p11 with appropriate GFF files. For comparison we used statistics from Tophat2 (Table 1a) taking into account not just read mapping rates but also the number of useful sequences. Additionally, we used Qualimap v.2.2.1 to count the number of reads mapping that belong to exonic (90.53%), intronic (7.61%) and intergenic (1.87%) regions as well as to determine duplication rates (29.09% for raw data and lenient approaches and 28.71% for stringent approaches) and uniformity of coverage through 5' and 3' bias (5' bias is 0.28, 3' bias is 0.56 and ratio between 5' and 3' bias is 0.84).

Similar studies, taking into account only the quality trimming, showed that mapping rate increases with the stringency of quality trimming [14], while the absolute number of aligned reads decreases [22]. We can state the similar - QC preprocessing approaches that include lenient quality trimming are beneficial in absolute number of aligned reads and in that way they contribute to overall amount of usable information in final dataset, but only one approach, CAQ5, provide the increase in mapping rates as well. The reason for this is the presence of 73583 contaminant sequences in AQ5 dataset, which increases the number of sequences in the dataset, but not the number of mapped sequences. On the other hand, the absolute number of multi aligned reads in a lenient QC preprocessing approach is also increased when compared to a stringent method, but the overall percentage is lower in lenient methods.

When we conducted a similar analysis with Tophat2, only this time we used setting of additional parameters as more appropriate for the nature of the set and the sequenced data themselves, we obtained the higher mapping rate for all approaches, but the conclusion was the same. Since we expected a variability in the dataset, we increased the number of mismatches (-N 4) and number of mutations expressed through edit distance metric (--read-edit-dist 5). Also, we demanded realignment of reads with edit distance higher than 2 (--read-realign-edit-dist 2) and no multihits (-g 1). Results are given in Table 1b. Again, we used Qualimap to determine quality of mapping (duplication rate: 27.35%-27.73%; mapping to specific regions: exonic - 91.16%, intronic - 6.96%, intergenic - 1.88%; 5' bias: 0.16, 3' bias: 0.61, 5'-3' bias: 0.3) and it also revealed the overall better quality of mapping when appropriate parameters in Tophat 2 were used.
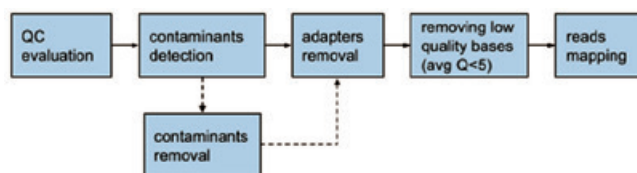

Figure 9. Proposed workflow for reads mapping task

Appropriate parameters contributed to the improvement of results in all approaches, with CAQ5 giving the best result. The proposed workflow is given in Fig. 9.

Table 1. Comparison of Tophat2 mapping results. R- raw data, C- removed contamination, A – removed adapters, Q5 – trimmed reads using sliding window with quality threshold set to 5, Q25 – trimmed reads using sliding window with quality threshold set to 25. Reads shorter than 25bp were discarded.

a) mapping performed with default parameters							
Data	**R**	**Q5**	**Q25**	**AQ5**	**AQ25**	**CAQ5**	**CAQ25**
No. of mapped seq	9049533	9051992	8913954	9051962	8913946	9051626	8913640
Mapping rate	98.5%	98.5%	98.7%	98.5%	98.7%	99.3%	99.5%
% of multi aligned	19.8%	19.8%	19.9%	19.8%	19.9%	19.8%	19.9%
b) mapping performed with parameters –g 1 –N 4 --read-edit-dist 5 --read-realign-edit-dist 2							
No. of mapped seq	9072621	9075046	8929607	9075035	8929538	9074742	8929260
Mapping rate	98.7%	98.8%	98.9%	98.8%	98.9%	99.6%	99.6%

De novo assembly

An accurate assembly requires a high quality clean data, which can be provided in one of two ways: by using stringent QC preprocessing approach on cleaned data (CAQ25) or by using lenient approach on clean data (CAQ5) combined with error correction (EC) performed using Rcorrector (a tool specially designed for this task in Illumina RNAseq data), with maximum number of correction set to 5. Additional steps in QC preprocessing, like elimination of Illumina specific bias, can contribute to the quality of assembly. According to Fig. 1b we determined that the first 11 bases are affected with Illumina bias so we used Trimmomatic to remove them from EC and CAQ25 datasets. In that way we formed two additional datasets, ECC and CAQ25C, in order to test the impact of clipping of biased bases on de novo assembly. All five datasets (R, CAQ25, EC, CAQ25c and ECc) were assembled using Trinity v2.4.0 [23] with minimal contig length set to 30 (because initial reads are short, ranging from 25 to 50bp). We didn't change the values of other default parameters because in Trinity they are set to give the best results for standard RNAseq data. For the evaluation of the assembly we mutually compared values of several standard metrics: N50, average contig length (both measured using in house perl script), reads mapping rate to transcriptome, transcriptome mapping rate to reference transcriptome and the number of proteins that matched the assembled transcriptome more than 80% (Table 2). All read mappings to transcriptome were conducted by using bowtie2 [24] with parameter --local, while the number of proteins was obtained by using BLASTX [25] and by the utility perl script provided within Trinity installation.

A desirable resulting assembly has the greater assembly length and the smaller number of contigs. That is why we observed the ratio between these two values in different approaches, with an aim to obtain the larger number represented as average contig length. Because contigs do not have uniform length, for mapping of reads to assembly and for mapping of assembly to reference transcriptome we observed only mapping rates and not the number of mapped contigs. Datasets used in stringent approaches included the clipping of the 10 bases from the 5' end, which made the sequences shorter and led to significant loss of information. This was reflected in the results by larger number of contigs and smaller assembly length, which is why the assembly had low average contig length. Another affected metric is the number of proteins which matched transcriptome in high percentage, used to reflect the number of almost full transcripts, showing that, although the mapping rate to reference transcriptome was high, obtained transcripts were not complete. On the other hand, it was beneficial for the mapping task, which was expected because it is easier to fit in the shorter than longer sequence in the reference sequence, but the longer ones are more reliable. Aforementioned is even more noticeable in the used dataset because the initial dataset consisted of very short sequences (35-50 bp) and clipping made the great amount of extremely short sequences which were not suited for further analysis. Methods in the lenient group gave better results, with similar results in most of the metrics, but with EC approach giving the highest number of >80% complete transcripts. According to the aforementioned, we recommend using EC approach.

Table 2. *Comparison of Trinity results for differently preprocessed data with parameter --min_contig_length set to 30. R – raw data, C – removed contamination, A – removed adapters, Q5 – trimmed reads using sliding window with quality threshold set to 5, Q25 – trimmed data using sliding window with quality threshold set to 25, EC – CAQ5 with mostly 5 error corrected bases, c – clipped 10 bases from 5' end. Reads shorter than 25bp were discarded.*

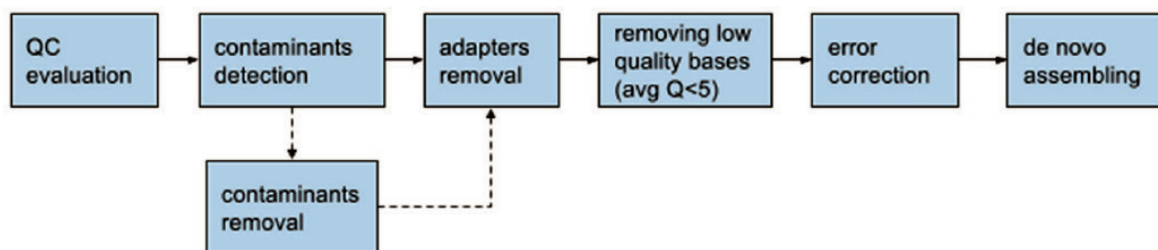| Data	R	CAQ25	EC	CAQ25c	ECc
N50	69183	69101	68014	120637	120571
average contig length	100.19	99.56	100.81	63.83	64.35
% of reads mapped to transcriptome	92.39	92.23	92.38	90.24	90.50
% of transcriptome mapped to reference transcriptome	86.49	86.71	86.31	90.21	89.95
No. of proteins matched the transcriptome >80%	2784	2793	2939	922	979

Figure 10. Proposed workflow for de novo assembly task

Conclusion

Specifics of RNA sequencing and RNAseq datasets have to be taken into account when choosing the best approach in the evaluation, preprocessing and analysis steps. Widely used software for quality control FastQC is not adjusted to RNAseq data which is why an adequate interpretation of obtained results is necessary, but the ultimate choice of appropriate QC preprocessing steps depends on the following step of downstream analysis - mapping or de novo assembling. Generally, all levels of QC preprocessing can be grouped into three approaches: naive approach which presumes that raw data is a high quality data, lenient approach which preserves more information, and stringent approach which provides cleanest dataset. Evaluation of approaches demands using significant and reliable metrics like mapping rate, number of mapped sequences and multi-mapped reads ratio for mapping task and N50, average contig length, percentage of assembly mapped to reference transcriptome, ratio of unique reads mapped to transcriptome and number of proteins matched the transcriptome in more than 80% for assembly task. For the comparison of QC results between different approaches and as a use case for studying the influence of QC on further analyses steps, we used publicly available RNAseq raw data from well researched and reference resources rich genome to provide reliable results. Our results showed that when downstream analysis included mapping, lenient methods with adapter removal included gave the best results. Moreover, results revealed that an influence of parameter selection on the mapping task exists, which is why parameters should be selected according to the nature and purpose of the NGS experiment. For the assembling task, the best results were obtained by using a lenient approach with error correction.

References

[1] Sheng, Q., Vickers, K., Zhao, S. *et al.* Multi-perspective quality control of Illumina RNA sequencing data analysis. Briefings in Functional Genomics 2017; 16(4): 194–204.

[2] Bolger, A. M., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 2014; 30(15): 2114–20.

[3] FastQC tool. Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [cited 24 August 2021]

[4] Okonechnikov, K., Conesa, A., García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data, Bioinformatics 2016; 32(2): 292–94.

[5] Hansen, K. D., Brenner, S. E., Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 2010; 38(12).

[6] Delhomme, N., Mähler, N., Schiffthaler, B., *et al.* Guidelines for RNA-Seq data analysis. Epigenesys; 17 November 2014.

[7] Meyers, B. C., Tingey, S. V., Morgante, M. Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. Genome Research 2001; 11(10):1660–76.

[8] Gu, Z., Wang, H., Nekrutenko, A. Li, W.H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 2000; 259(1-2), pp. 81-8.

[9] Song, L. Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience 2015; 4(1),

[10] Yang, X., Chockalingam, S. Aluru, S. A survey of error-correction methods for next-generation sequencing. Briefings in Bioinformatics 2013; 14(1), 56–66.

[11] Akogwu, I., Wang, N., Zhang, C., Gong, P. A comparative study of k-spectrum-based error correction methods for next-generation sequencing data analysis. Human Genomics 2016;10(20).

[12] Gabaldón, T. Alioto, T. Whole-Genome Sequencing Recommendations. In A.M. Aransay, J.L. Lavín Trueba Eds. Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing, Springer International Publishing 2016; pp.13-41.

[13] Yun, S., Yun, S. Masking as an effective quality control method for next-generation sequencing data analysis. BMC Bioinformatics 2014; 15(1).

[14] Del Fabbro, C., Scalabrin, S., Morgante, M. Giorgi, F.M. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis, PLoS ONE 2013; 8(12).

[15] MacManes, M. D. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics

2014; 5(13).

[16] Hansen, K. D., Brenner, S. E., Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 2010; 38 (12).

[17] Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T.L. NCBI BLAST: a better web interface. Nucleic Acids Research 2008; 36:W5-9.

[18] Wang, C., Politz, J. C., Pederson, T., Huang, S. RNA Polymerase III Transcripts and the PTB Protein Are Essential for the Integrity of the Perinucleolar Compartment. Molecular Biology of the Cell 2003,14(6):2425–35.

[19] Kamath, R., Thor, A., Wang, C. *et al.* Perinucleolar Compartment Prevalence Has an Independent Prognostic Value for Breast Cancer. Cancer Research 2005; 85(1):246-53.

[20] UniVec database. Available at: ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/ [cited 24 August 2021]

[21] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R, Salzberg, S.L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 2013, 14(4

[22] Williams, C. R., Baccarella, A., Parrish, J. Z., Kim, C. C. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 2016;17.

[23] Grabherr, M. G., Haas, B. J., Yassour, M., et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 2011; 29(7):644–52.

[24] Langmead, B., Salzberg, S. L. Fast gapped-read alignment with Bowtie 2, Nature Methods 2012; 9(4):357–59.

[25] Gish, W. States, D.J. Identification of protein coding regions by database similarity search. Nature Genetics 1993, 3:266-72.

ABOUT THE AUTHORS

Dragana Dudić is an Assistant at the Faculty of Informatics and Computer Science at the University Union Nikola Tesla for several Computer Science subjects. She obtained BSc and MSc in Computer Science at the Faculty of Mathematics, University of Belgrade where she is finishing her PhD in the field of Bioinformatics. Dragana is interested in research of repetitive sequences, especially human and plant transcriptomics.

Bojana Banović Đeri is a Research Associate in the Laboratory for Plant Molecular Biology in the Institute of Molecular Genetics and Genetic Engineering, University of Belgrade. In 2013. obtained PhD in Molecular biology at the Faculty of Biology, University of Belgrade in the field of plants' molecular biology, genetics and proteomics (characterization of buckwheat heteromorphic self-incompatibility).

Vesna Pajić is a Bioinformatician at Seven Briges Genomics. She has a strong education professional with a PhD. focused in Computer Science. Vesna is experienced in both academia and the biotechnology industry. She is interested in applications of mathematics and CS in life sciences and skilled in computer applications, bioinformatics, and management.

Gordana Pavlović-Lazetić is a Professor of Computer Science at Faculty of Mathematics, University of Belgrade. Gordana has forty years of teaching experience at the University of Belgrade in different programming languages and database courses. Her recent research interests are in bioinformatics and natural language processing. She spent two years as a researcher at the University of California, Berkley, and the Relational Technology Inc, as a consultant for extending the Relational Database System Ingres to manage text.

FOR CITATION

<A Day at School>: A Serious Game for Social Skills Training

Filimonas Papadiou[1], Fotis Lazarinis[2], Dimitris Kanellopoulos[3],

[1] Hellenic Open University, GR 26335, Patras, Greece. filimonas@gmail.com
[2] Hellenic Open University, GR 26335, Patras, Greece. fotis.lazarinis@ac.eap.gr
[3] Department of Mathematics, University of Patras, GR 26500, Greece. d_kan2006@yahoo.gr

Abstract: Soft skills are the personal characteristics of an individual that enhance his/her interactions, career prospects, and job performance. Soft skills include social skills which incorporate characteristics like empathy, self-control, socialization, and friendliness. The development of soft skills at an early age is vital. Currently, there are few serious games for social skills training aimed at primary school pupils. A serious game does not only provide fun but a player can discover knowledge about himself. This paper presents a serious game named "**A Day at School**" that helps primary school pupils to develop social skills through an educational scenario. In this scenario, the hero of the game faces various situations during a usual day at school. The scenario deals with the situations of bullying, racism, and social awareness of children. By using the educational application, pupils discover appropriate behavior and get the first stimulus for acquiring their social skills. The serious game helps them to socialize and gain the basis to cultivate empathy, friendliness, and self-control. Primary school pupils and teachers evaluated the serious game. The results showed that teachers found that the game is suitable for teaching purposes and its graphical user interface (GUI) is appealing.

Keywords: Serious games; Soft skills; Social skills; Educational games.

INTRODUCTION

In the last decades, the use of Information and Communication Technologies (ICTs) for teaching and e-learning purposes has increased significantly [4][8]. Many governments have made large investments in equipping schools with smart computing devices and software tools (games, text, animations, video, and audio clips). In addition, the adoption of e-learning services in teaching practices has spread rapidly. ICTs provide access to a wide range of e-learning resources, and thus pupils can improve their educational outcomes [21]. Also, the use of ICTs can reduce educational costs in the long run [6]. It can promote personalized teaching and better monitoring of pupils' progress [8]. ICTs can increase pupils' flexibility and autonomy while improving their learning attitudes and experiences [6]. Furthermore, ICTs can be used to improve teaching materials and make e-lessons more comprehensive, engaging, and interactive [4] [13].

In the near past, technical skills (also known as hard skills) were the only skills required for professional employment. Nowadays, technical skills are not enough in the current workplace [3]. Soft skills are required as well. Soft skills include social skills which incorporate characteristics like empathy, self-control, socialization, and friendliness. ICTs can develop people's soft skills which employers want from them [15]. There has been a lot of research ([2] [3][15][18]) that focuses on the importance of soft skills in the workplace. In many cases, successful job placement depends on people's soft skills, while technical knowledge plays a secondary role. As the workplace progressively seeks employees who are mature and socially well-adjusted, soft skills are valued as the number one qualification for entering the

workplace successfully. Many studies such as [1][3] [7][15][18] suggest that starting the development of soft skills at an early age is vital. For instance, at an early age, primary school pupils can play serious games aimed at their social skills training. Notably, primary school teachers are often unaware of the importance of soft skills and they cannot teach social skills lessons. Moreover, they are not aware that their pupils have not acquired proper soft skills. Discovery learning through serious games can solve this problem. Moreover, serious games, based on discovery learning, can help pupils to develop their social skills through experimentation and practice.

This paper presents a new serious game (named "A Day at School") that helps primary school pupils to develop their social skills. Also, it presents its evaluation by teachers and pupils.

SERIOUS GAMES FOR SOCIAL SKILLS TRAINING

Social skills include verbal and non-verbal communication such as speech, gesture, facial expression, and body language. In particular, they include the traits of humor, empathy, self-control, socialization, fierce personality, subtlety, friendliness, and body language [19]. They are interpersonal, human behavioral, or management skills necessary to apply technical skills and knowledge in the workplace [14] [22]. A person has acquired social skills if s(he) knows how to behave in social situations and understands rules when communicating with others [7]. A lack of soft skills can ruin the promising career of a person who has the technical ability and work experience but not interpersonal skills [1][5][10].

In a school environment, empathy, self-control, socialization, and friendliness are important for creating and maintaining friendships. Interactions of children are not always smooth and a child must be able to use appropriate tactics to resolve conflicts. Also, children need to be 'empathetic'. Empathy is the ability to understand (or feel) what another person is experiencing inside. Self-control is the ability to control ourselves, our emotions, and our behavior by anticipating situations. This is how we get better results in our social relationships. With self-control, pupils can effectively deal with conflicts that may arise in a school environment. Socialization is a sociological term that means the process of transmitting to the individual the morals, customs, rules, and

ideologies necessary for participation in society. Friendliness and friendship are important factors in the socialization of children and the development of healthy relationships.

Teachers must be aware when their pupils have not acquired the necessary soft skills. In this case, pupils should be encouraged to improve their soft skills by reading books, attending classes, participating in activities to broaden their horizons, such as presentations and discussions, or playing games. In addition, soft skills courses could be integrated into curricula. However, current curricula are already overloaded with technical skills lessons.

Serious games using discovery learning

A serious game includes elements of entertainment combined with a practical dimension [20]. The main purpose of a serious game is not entertainment, enjoyment, or amusement. Currently, the use of serious games for social skills development in the classroom is accelerating at a remarkable rate [23]. Pupils learn by playing serious games which can teach them social skills and concepts. Ref [23] is a systematic review that summarizes the current existing literature on social skill serious games for young people ages 5 to 19. A serious game can enhance the user's experience through interaction in education, training, health, or interpersonal communication [12] [23].

The most serious games are based on the discovery learning approach. Discovery learning is an inquiry-based teaching method (related to constructivism) that focuses on understanding the structures and scientific principles of a subject matter [17]. Discovery learning is directly linked to our experiences, it arises from and is influenced by the context in which it takes place. Also, it results from experimentation and practice. Consequently, pupils can develop social skills through experimentation whereby they can build their knowledge in the domain of social skills and derive rules and conclusions from the results of their experiences [23].

THE GAME: "A DAY AT SCHOOL"

We developed a serious game, named "A Day at School". In this game, the player is asked to move the hero of the game and find the best way to help friends and classmates in various situations such as

bullying, racism, etc. This educational application helps primary school pupils to develop their social skills through a simple educational scenario. By playing the serious game, pupils discover appropriate behavior and get the first stimulus as a trigger for acquiring social skills.

Design and Implementation

We developed the serious game by following the software design and development processes. Based on the discussion with teachers, we created the script and determined the functional and non-functional requirements of the game. We designed the use case diagram with the main functions of the application with which the user interacts. We wrote the source code of the serious game by using the *Pygame* and *Pygame_GUI* packages [9]. We created the graphics of the game by using the open-source application *Inkscape* [11]. Figure 1 depicts the application structure and system architecture.
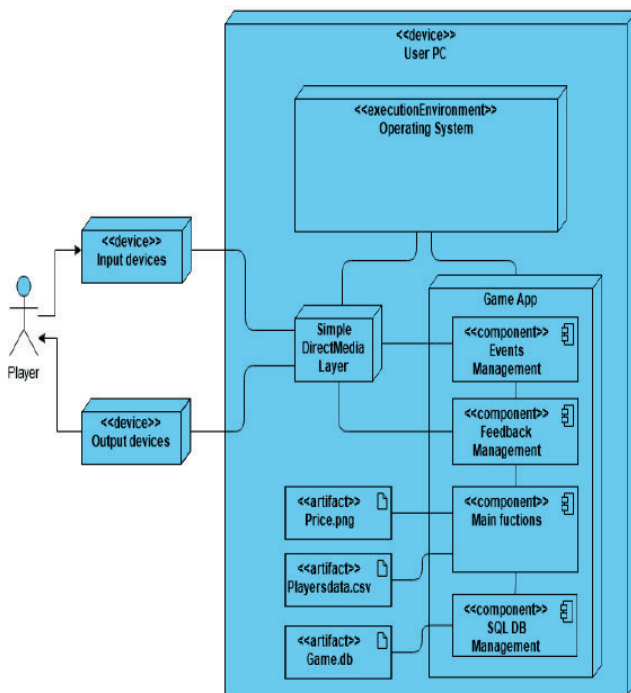
Figure 1. Deployment diagram

The application connects to the input devices via the *Simple Direct Media Layer*, through which the input events are received for management by the application. The output devices give feedback to the player. In addition to the main functions, the application has less important functions for exporting and managing data from the *SQL Database* as well as *Award extraction*. For the development of the application, we wrote the code by using the Python programming language, while we designed the prototype graphics by using an open-source editing application.

The Educational Scenario of the Game

The educational scenario takes place in a virtual primary school environment. In this scenario, the game hero faces various situations during a usual day at school. The scenario deals with the situations of bullying, racism, and social awareness of children. Bullying is the use of violence, verbal or physical, between children to cause fear, pain, or distress. Racism is the perception that people are not all equal to each other, but are divided into superior and inferior, distinguished either by skin color, ethnicity, religion, gender, or sexual orientation. Empathy is defined as identifying with another person's situation and understanding his behavior, motivations, and needs.

Based on these situations (bullying, racism, etc.), we created an educational scenario. Through these situations, pupils are trained in empathy, self-control, socialization, and friendliness. The goal of the educational scenario is for the pupil to discover knowledge on his/her own in an explorative way. In this scenario, the pupil can think and explore which is the appropriate choice/option so that s(he) can earn the corresponding points. There are positive, neutral, and negative situations for the pupil to understand the correctness of the actions of the game hero by winning enough, a few or no points. At each step of the game, if the player has not won any points, an opportunity is given to try again. In the design of the educational scenario, there is positive feedback so that the player is motivated to interact positively and perform the required actions. There are no negative scores. All players may receive a prize in the form of a commendation.

The educational scenario consists of six levels of training.

Analysis of Training Levels

Each level concerns training in a different subject. At each level, possible options are proposed to the user, and s(he) is asked to decide which option/solution is the correct one and earns the corresponding points. At each level, the user is asked

to decide whether the hero will do the right thing or not. Below, we analyze each level.

Level 1: Socialization on the way to school (Socialization - Empathy)

At the traffic light (located near to school) our game hero sees a classmate on the street. The classmate is sitting in a wheelchair and cannot activate the pedestrian crossing to cross the street. What the hero should do? There are three options:

1. He approaches the traffic light to activate it. Therefore, he can cross the street and continue on his way, not caring about his classmate.
2. He approaches the traffic light to turn it on. He helps his classmate to safely cross the street and takes her to school.
3. He carelessly passes the traffic light without helping and continues to school.
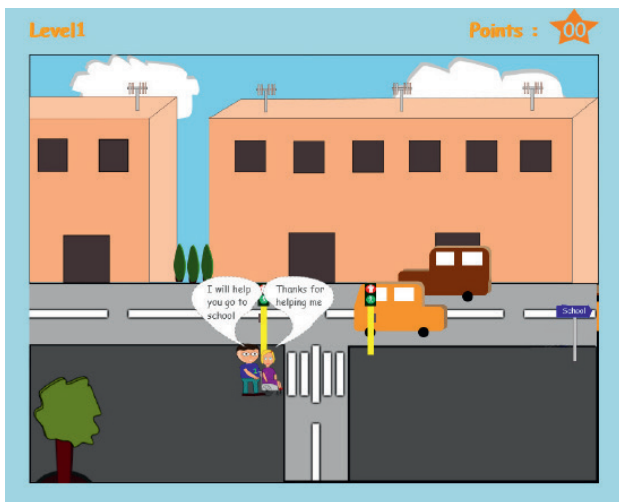
Figure 2. The hero is ready to help (Level 1)
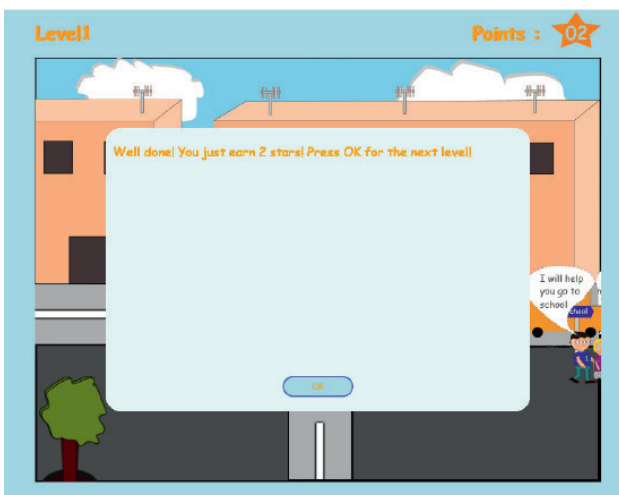
Figure 3. Successful completion of Level 1

Level 2: School violence - bullying in the classroom (Friendliness - Self-control)

As classmates leave the classroom, some of them knock down a classmate. What the hero should do? There are three options:

1. He goes to his classmate and remarks to the others that violence is not the solution.
2. He walks out of the classroom, wondering if he should talk to the teacher.
3. He goes over to his classmate and mocks him.
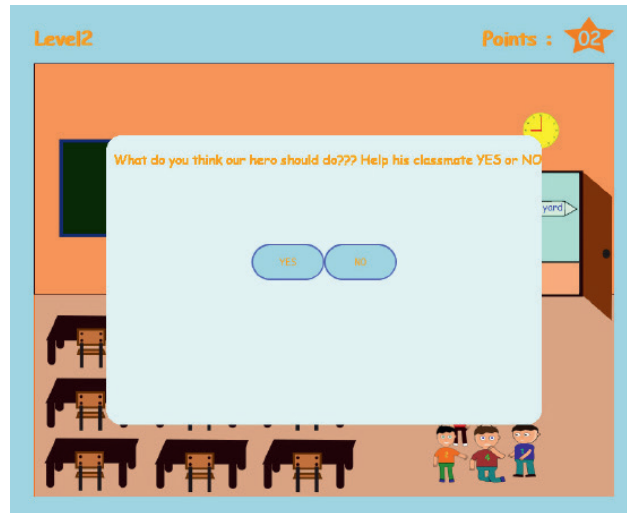
Figure 4. The pupil is asked to decide what the hero will do (Level 2)

Level 3: Distinctions between pupils at recess (Empathy - Friendship)

While the hero is with his classmates (next to them), he notices an isolated child being made fun of. What the hero should do? There are three options:

1. He approaches the child and tells him to join the group.
2. He approaches the child and tells him that he should just sit there since they do not want him in their group.
3. He goes to the dining room for his lunch. He hopes to find his teacher.
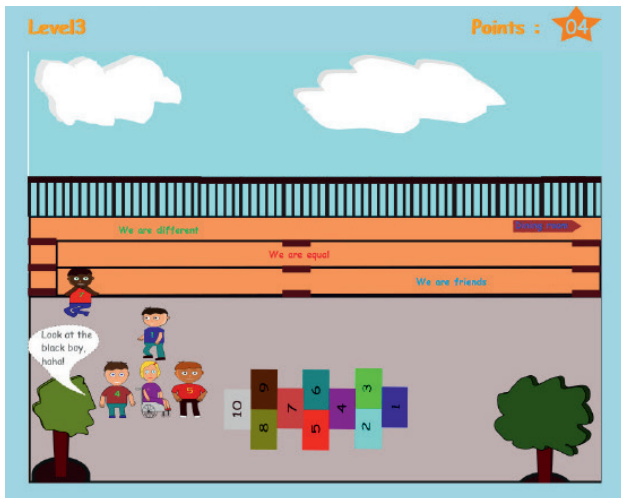
Figure 5. The hero is ready for action (Level 3)

Level 4: At meal-time (Empathy - Friendship)

The hero observes that a child is sitting without eating. What should the hero do? There are three options:

1. He approaches the child and discreetly gives some of his food. He tells to the child to talk about its problem to the teacher.
2. He approaches the child and shouts "Why are you sitting here if you have nothing to eat?"
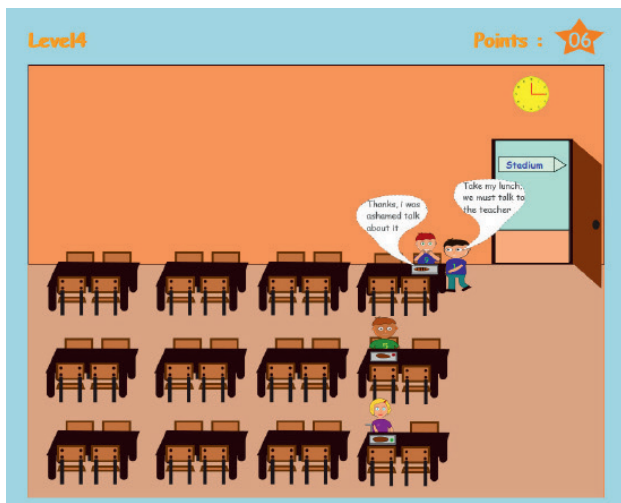3. He leaves the dining room to the stadium to talk to the teacher.

Figure 6. The hero helps his classmate (Level 4)

Level 5: Distinctions between pupils in gymnastics (Empathy - Friendship)

A classmate cannot cope in sports and sits in the stands. What should the hero do? There are three options:

1. He approaches his classmate and tells him that it's okay if he can't play but if he wants he can be a referee.
2. He approaches his classmate and tells him that since he cannot play, he shouldn't take part in gymnastics.
3. He leaves to go to his classroom and get ready for leave from the school.
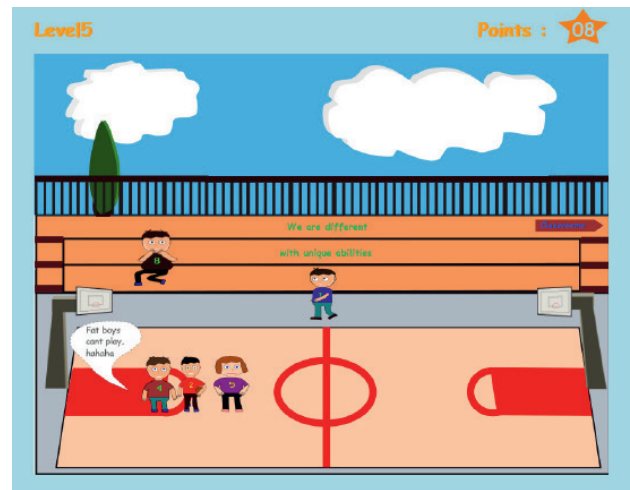
Figure 7. The hero is ready to help his classmate in the stands (Level 5)

Level 6: School violence-bullying (Friendliness - Self-control)

In this training level, the hero is the victim of verbal bullying by a group of kids. What the hero should do? There are three options:

1. He thinks he should not pay attention and leaves for his house heading for the park.
2. He returns to school to inform his teacher and the Principal.
3. He approaches the group of children and replies by mocking them. Then, he leaves towards the park.
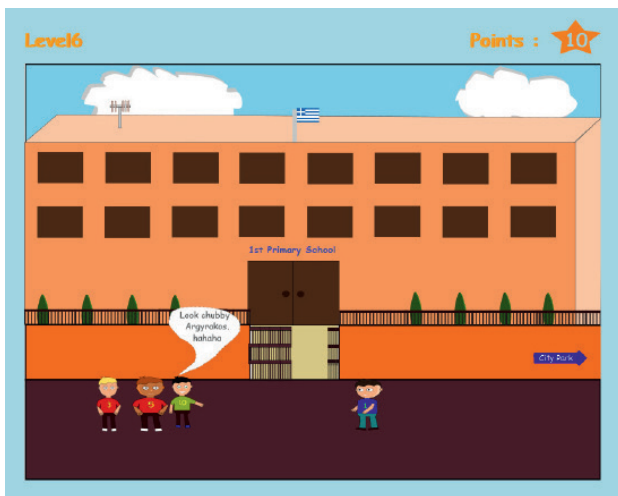
Figure 8. The hero is verbally bullied (Level 6)

EVALUATION OF THE GAME

Methodology

Teachers and pupils evaluated the game. Twenty-six (26) teachers of primary school education evaluated the game. Also, 28 primary school pupils participated in the evaluation process (57.1% were boys and 42.9% were girls). Among them, 7.1% were pupils in Primary Class A, 7.2% in Primary Class B, 14.3% in Primary Class C, 14.3% in Primary Class D, 35.7% in Primary Class E, and 21.4% in Primary Class F. The evaluation took place online from March 14, 2021, to March 20, 2021. For the evaluation of the game, we used questionnaires with metrics and Likert scales [16]. One questionnaire for teachers (see **Table 1** in **Appendix**) and two questionnaires for pupils (see **Tables 2 and 3** in Appendix) were distributed by email to investigate the subjective satisfaction of teachers and pupils. Teachers and pupils completed their questionnaires online through a free form creation service.

Table 1 shows the Questionnaire and the teachers' responses (%) per question. Table 1 includes questions about the teachers' experience with serious games, the design, and the educational scenario of the game, whether the intentions of the game are clear. Finally, this questionnaire asks the teachers' opinions about the effectiveness of the serious game and its possible integration into the teaching process.

Pupils were asked to answer questions before and after the playing of the game. The evaluation investigates:

- The experience of teachers and pupils in serious games, and how much they have been involved into the educational process.
- The effectiveness of the application on pupils, (i.e., whether they learned through the game).
- The suitability of the application for use in the school environment.
- The usability in terms of ease of use.
- Accuracy in the design of the application concerning the scenario and the topic it deals with.
- User satisfaction with the design of the application environment (i.e., graphics, sounds, etc.).

The first Questionnaire **(Table 2)** asks pupils questions about whether they play computer games and whether they have played serious games. Pupils were asked to answer how they would behave in certain situations. In the second Questionnaire **(Table 3)**, some questions are asked again to find out if a pupil has learned through the game, and additional questions are asked about the usability and design of the game. The Likert scale was not used in the pupil questionnaires because this would confuse them due to their age.

Results Analysis

Teachers: A large percentage (84.6%) of teachers had no previous experience with serious games. However, they understood the purpose of the game. They found that the design of the graphical user interface (GUI) is appealing to pupils, and considered it suitable for education. Regarding the design of the game, 92.3% of the teachers answered that they would not change anything, with the remaining 7.7% concerning the graphics. Regarding the effectiveness of the app, teachers were completely confident or almost confident that the app can raise pupils' awareness and develop social skills **(Figure 10).** A small percentage of teachers could not answer with certainty about the suitability of the application for use in the educational process. Almost 70% agreed with its suitability **(Figure 10).** Regarding the suitability of the serious game for the development of social skills, 69.3% of teachers considered it suitable (or rather suitable) with the remaining percentage expressing doubt **(Figure 10).** In terms of the accuracy of the game and the
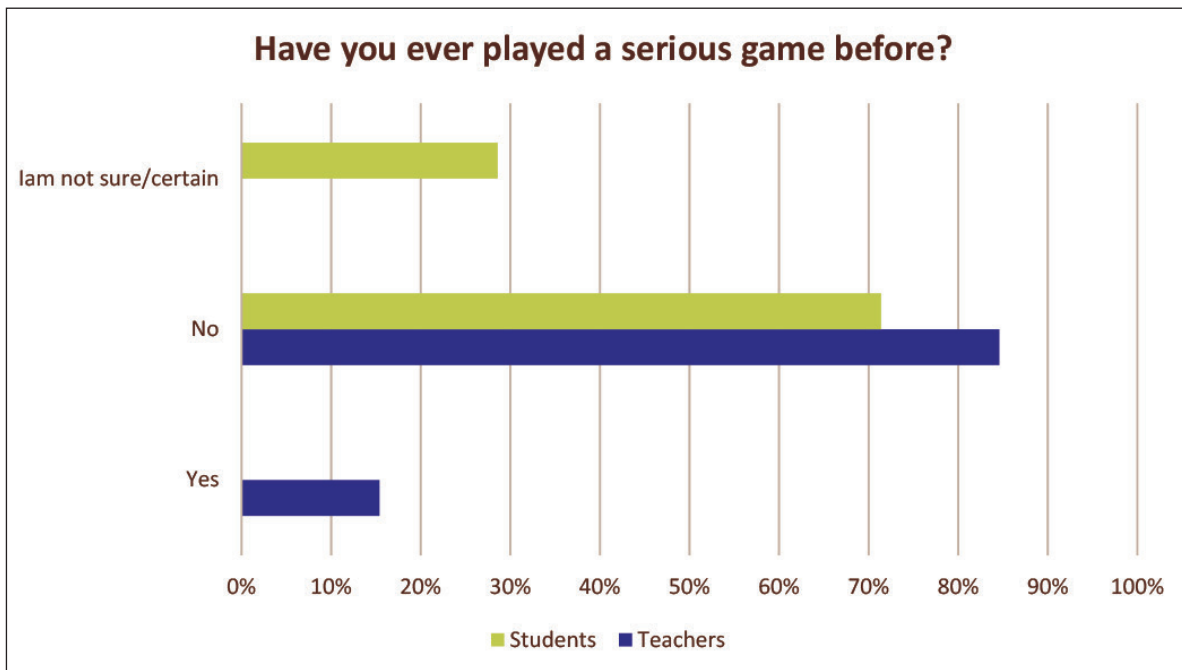
Figure 9. Use of serious games by teachers and pupils

scenario, this was considered to be understood by all teachers with almost half of them stating this with absolute confidence. With the same results, the approach to the topics dealt with in the game was judged to be apt. Still, 92.3% felt that the design matched the scenario, with only a small percentage of doubt.

In terms of user satisfaction, teachers found the design of the game to be appealing to the children's interest and the flow would not tire the user. Moreover, they did not suggest any change in the design of the game.

Pupils: Most of the pupils did not know the term "serious games" but they understood what the
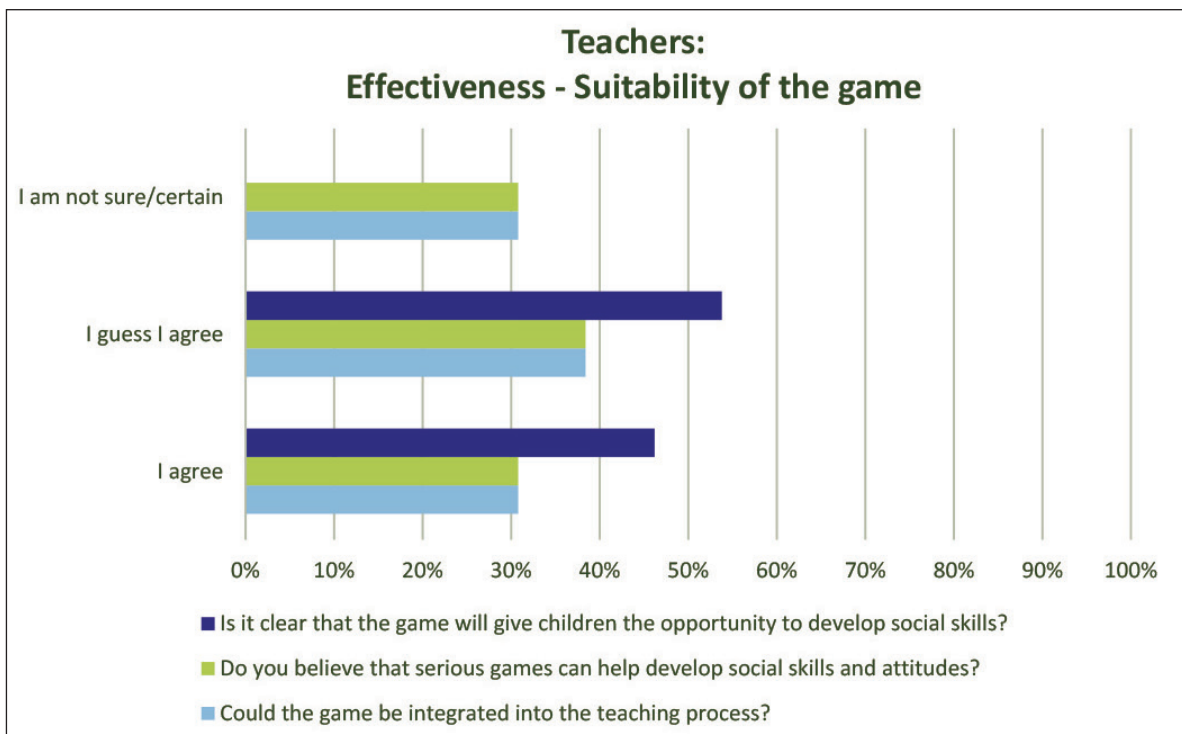
Figure 10. Teachers Effectiveness - Suitability of the application
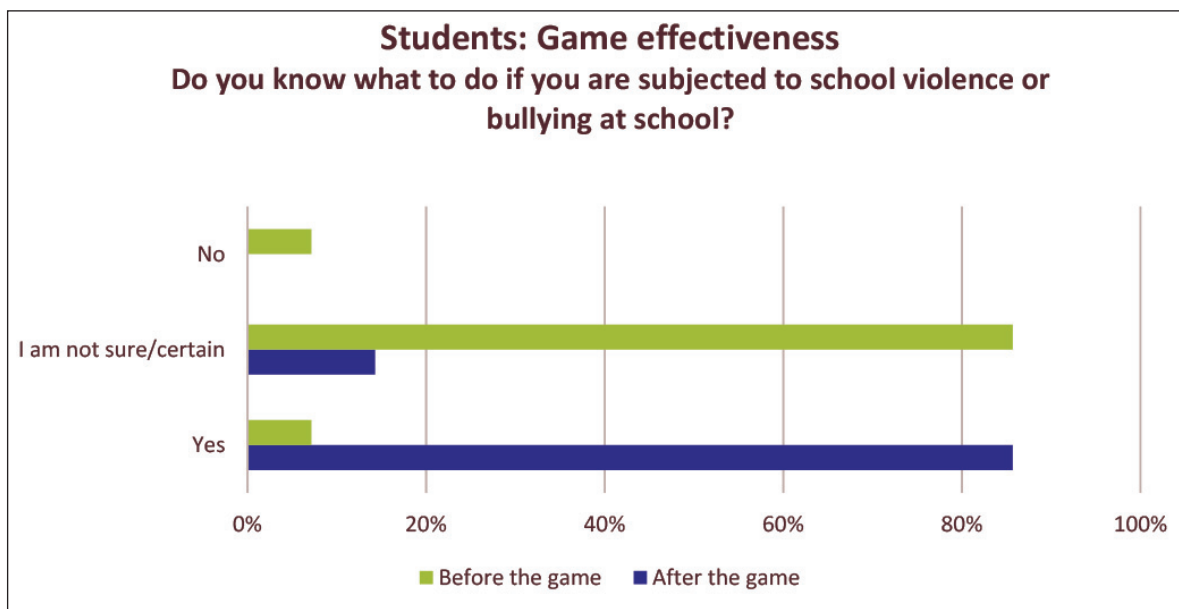
Figure 11. Effectiveness on pupils - Before and after playing the game

teaching purpose of the game was. The pupils had not used serious games at all or were not sure about it **(Figure 9).** Among the pupils, after the game, 92.9% said that they now know what bullying is, that they have understood what discrimination between people is with 71.4%. 85.7% responded that they know what to do if they are bullied and with the same percentage that the game helped them to understand how to deal with different situations **(Figure 11).**

All pupils said that: (1) they should help those in need; (2) they would help a classmate at school; and

(3) they would inform their teacher (**Figure 12**). Before pupils play the game, only 21.4% said they knew what school violence was. 78.6% were unable to say for sure. 85.7% were not sure about discrimination between people, while 14.3% did not know. A percentage of 71.4% and 78.6% said they were not sure if they would help a classmate at school with the remaining percentage being sure they would. Concerning bullying, 85.7% said they were not sure what to do if they were bullied with the remaining percentage split between those who either knew or did not know. After playing the game, most pupils
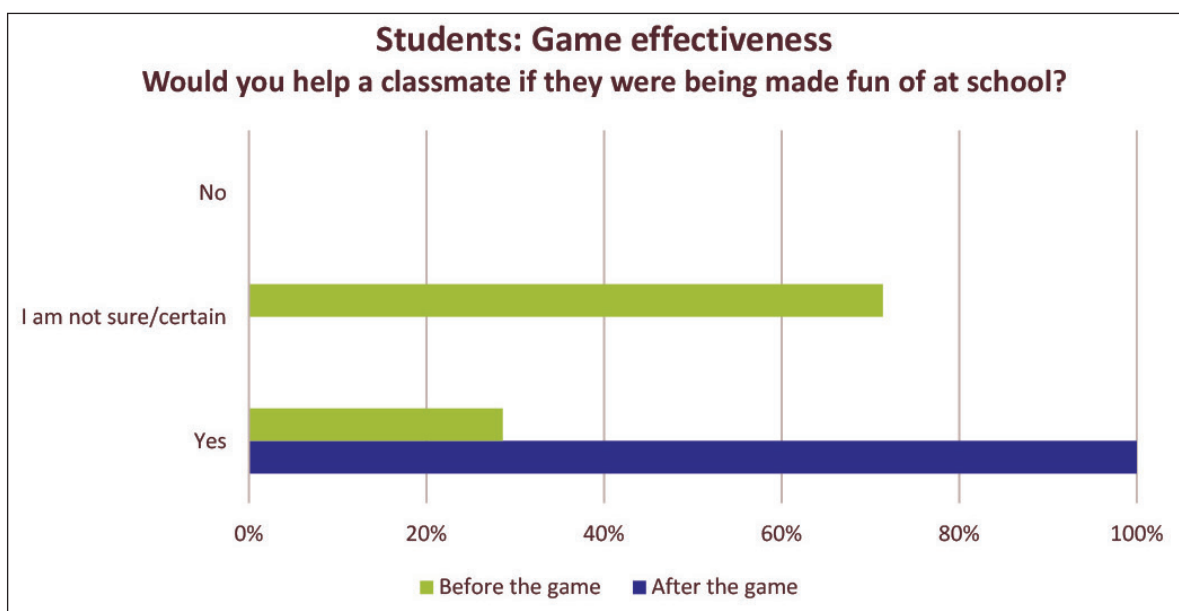
Figure 12. Effectiveness on pupils - Before and after playing the game

answered positively to all the above questions. All of them found that the game is easy to handle, with useful help provided.

By analyzing the above results, we conclude:

- Serious games have not been integrated into the educational process or widely accepted. The highest percentage (84.6%) of teachers had not used a serious game so far. The highest percentage of pupils was not sure if they were aware of serious games even though all pupils have played computer games before. Even if they have played a serious game, they were not sure about the term called "serious game".
- The game is effective as it convinced the teachers that it can raise awareness of social issues among pupils. By playing the game, pupils clarified the concepts of bullying and racism. It convinced them that they should talk to their teacher about all the issues addressed in the game. Moreover, although before the game, most of the pupils were not sure how they would help a classmate. After playing it, all of them answered positively. Therefore, the developed game helped them to socialize and gain the basis to cultivate empathy, friendliness, and self-control.
- The game is suitable for integration into the educational process (with almost 70% certainty) because the majority of teachers evaluated it with absolute (or relative) confidence. All the teachers were convinced that serious games are suitable for the development of social skills.
- The game is easy to play and attractive to users. The provided help system is satisfactory for pupils.
- The GUI of the game is satisfactory.
- The educational scenario is complete for addressing the issues of social skills training. It is appropriate and easy to understand by the teachers, which was reflected in the pupil's results after using the game.
- The purpose of the game has been achieved to a large extent. Some of the pupils may not have understood the social concepts dealt with in the game. However, after playing the game, all of them answered that they would help a classmate and would talk to their teacher.

CONCLUSION

This paper presented a game called "A Day at School" that is effective for primary school pupils for social skills training. Teachers found the educational scenario (script) appropriate. Moreover, they would like to integrate this game into the e-learning procedure. Also, pupils liked the game, understood the educational scenario of the game, and responded positively to what the game aimed to teach them. By using this game, pupils were triggered to develop and strengthen the social skills they discovered by using it.

Soon, we will expand our game by developing chapters with sub-levels where the user would select a category to be trained in other social skills subjects. We will develop an integrated platform to train primary school pupils in all social skills. Finally, we aim to use more effective graphic design capabilities that will enable the game even more attractive.

REFERENCES

[1] Bates, C. C., & Morgan, D. N. (2018). Literacy leadership: The importance of soft skills. *The Reading Teacher*, *72*(3), 412-415.

[2] Caputo, F., Cillo, V., Candelo, E., & Liu, Y. (2019). Innovating through digital revolution: The role of soft skills and Big Data in increasing firm performance. *Management Decision*, *57*(8), 2032-2051

[3] Cimatti, B. (2016). Definition, development, assessment of soft skills and their role for the quality of organizations and enterprises. *International Journal for quality research*, *10*(1).

[4] Comi, S. L., Argentin, G., Gui, M., Origo, F., & Pagani, L. (2017). Is it the way they use it? Teachers, ICT and student achievement. *Economics of Education Review, 56*, 24-39.

[5] de Vasconcelos, J. B., Kimble, C., & Rocha, Á. (2016). A special issue on knowledge and competence management: Developing Enterprise solutions. *Information Systems Frontiers*, *18*(6), 1035-1039.

[6] De Witte, K., & Rogge, N. (2014). Does ICT matter for effectiveness and efficiency in mathematics education?. *Computers & Education*, *75*, 173-184.

[7] Dowd, T. P., & Tierney, J. (2005). *Teaching social skills to youth: A step-by-step guide to 182 basic to complex skills plus helpful teaching techniques*. Boys Town Press.

[8] Falck, O., Mang, C., & Woessmann, L. (2018). Virtually no effect? Different uses of classroom computers and their effect on student achievement. *Oxford Bulletin of Economics and Statistics*, *80*(1), 1-38.

[9] Kelly, S. (2016). Basic introduction to pygame. In *Python, PyGame and Raspberry Pi Game Development* (pp. 59-65). Apress, Berkeley, CA.

[10] Kimble, C., de Vasconcelos, J. B., & Rocha, Á. (2016). Competence management in knowledge intensive organizations using consensual knowledge and ontologies. *Information Systems Frontiers*, *18*(6), 1119-1130.

[11] Kirsanov, D. (2021). *The Book of Inkscape: The Definitive Guide to the Graphics Editor*. No Starch Press.

[12] Laamarti, F., Eid, M., & El Saddik, A. (2014). An overview of serious games. *International Journal of Computer Games Technology*, 2014.

[13] Livingstone, S. (2015). Critical reflections on the benefits of ICT in education. In *Digital Technologies in the Lives of Young People* (pp. 19-34). Routledge.

[14] Matteson, M. L., Anderson, L., & Boyden, C. (2016). "Soft skills": A phrase in search of meaning. *portal: Libraries and the Academy*, *16*(1), 71-88.

[15] Mitchell, G. W., Skinner, L. B., & White, B. J. (2010). Essential soft skills for success in the twenty-first century workforce as perceived by business educators. *Delta Pi Epsilon Journal, 52*(1).

[16] Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann.

[17] Nurdin, K., Muh, H. S., & Muhammad, M. H. (2019). The implementation of inquiry-discovery learning. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 7(1).

[18] Remedios, R. (2012). The role of soft skills in employability. *International Journal of Management Research and Reviews*, *2*(7), 1285.

[19] Robles, M. M. (2012). Executive perceptions of the top 10 soft skills needed in today's workplace. *Business Communication Quarterly*, 75(4), 453-465.

[20] Shute, V. J., Ventura, M., & Bauer, M. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In *Serious games* (pp. 317-343). Routledge.

[21] Spiezia, V. (2011). Does computer use increase educational achievements? Student-level evidence from PISA. *OECD Journal: Economic Studies*, 2010(1), 1-22.

[22] Weber, M. R., Finley, D. A., Crawford, A., & Rivera Jr, D. (2009). An exploratory study identifying soft skill competencies in entry-level managers. *Tourism and Hospitality Research*, *9*(4), 353-361.

[23] Zheng, L. R., Oberle, C. M., Hawkes-Robinson, W. A., & Daniau, S. (2021). Serious Games as a Complementary Tool for Social Skill Development in Young People: A Systematic Review of the Literature. *Simulation & Gaming*, 10468781211031283.

APPENDIX: Questionnaires for teachers and pupils

Table 1. Teachers: Results of their Questionnaire

Reply Question	I agree (%)	I guess I agree (%)	I am not sure / certain (%)	I guess I disagree (%)	I disagree (%)
1. Have you ever played a serious game before?	15,4	-	-	-	84,6
2. Do you think is easy to understand the game scenario?	53,8	46,2	-	-	-
3. The design of the game is attractive to attract the interest of children?	100	-	-	-	-
4. Do the game flow and design tire the user?	76,9	23,1	-	-	-
5. Does the design of the game match the script?	61,5	30,8	7,7	-	-
6. Is it clear that the game wants to raise social awareness among children?	38,5	53,8	7,7	-	-
7. Is it clear that the game wants to make children aware of discrimination between people?	53,8	38,5	7,7	-	-
8. Is it clear that the game wants to make children aware of bullying?	53,8	38,5	7,7	-	-
9. Is it clear that the game will give children the opportunity to develop social skills?	46,2	53,8	-	-	-
10. Is the game's approach to the issues it addresses apt?	84,6	15,4	-	-	-
11. Could the game be integrated into the teaching process?	30,8	38,4	30,8	-	-
12. Do you believe that serious games can help develop social skills and attitudes?	30,8	38,4	30,8	-	-

Table 2. Pupils: Results of the first Questionnaire

Reply Question	Y e s (%)	I am not sure/ certain (%)	No (%)
1. Have you ever played a serious game before?	-	28,6	71,4
2. Have you ever played video games before?	100	-	-
3. Do you know what school violence/bullying is?	21,4	78,6	-
4. Do you know what discrimination between people is? (racism)	14,3	85,7	-
5. Do you think we should help those in need?	92,9	7,1	-
6. Would you help a classmate if s(he) was being made fun of at school?	28,6	71,4	-
7. Would you help a classmate if s(he) was being beaten at school?	21,4	78,6	-
8. Do you know what to do if you are subjected to school violence or bullying at school?	7,15	85,7	7,15

Table 3. Pupils: Results of the second Questionnaire

Reply Question	Y e s (%)	I am not sure/ certain (%)	No (%)
1. Do you know what school violence/bullying is?	92,9	7,1	-
2. Do you know what discrimination between people is? (racism)	71,4	28,6	-
3. Do you think we should help those in need?	100	-	-
4. Would you help a classmate if they were being made fun of at school?	100	-	-
5. Would you help a classmate if they were being beaten at school?	100	-	-
6. Do you know what to do if you are subjected to school violence or bullying at school?	85,7	14,3	-
7. Would you inform your teacher if a classmate was being beaten at school?	100	-	-
8. Would you inform your teacher if there was discrimination between children in your school?	100	-	-
9. Would you inform your teacher if you saw that a child from school was in serious trouble?	100	-	-
10. Do you think the game helped you to understand how to deal with different situations in your school and to be more socially aware/sensitive?	85,7	14,3	-
11. Did you like the images and colors of the game?	71,4	28,6	-
12. Was the game easy to use?	100	-	-
13. Was the help in the game understandable?	100	-	-

ABOUT THE AUTHORS
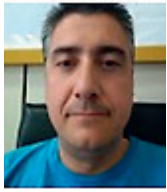
Filimonas Papadiou is an engineer in electronics. He received a Master's degree from Hellenic Open University in Information Technology Systems. Among others, his interests include software applications development. He works at Hellenic Air Navigation Service Provider (HANSP) as a Flight Information Service Officer (FISO).

Fotis Lazarinis holds a Ph.D. in educational technology from Teesside University, UK. He works as an adjunct lecturer at the Hellenic Open University. He participates in various research and development projects with a focus on educational technology. He has co-authored more than 50 papers in international journals and conferences.
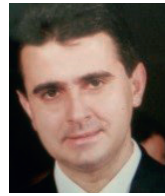
Dimitris Kanellopoulos is a member of the Educational Software Development Laboratory in the Department of Mathematics at the University of Patras, Greece. He received a Diploma in electrical engineering and a Ph.D. in electrical and computer engineering from the University of Patras. He is a member of the IEEE Technical Committee on multimedia communications. His current research interests are multimedia networks and wireless ad-hoc networks. He has published numerous articles in refereed journals and conference proceedings, and he has edited two books on "Multimedia Networking".

FOR CITATION

On the Possibility of Embedding the Mechanism of Linguistic Anticipation into Speech Recognition Systems

Daniel Kurushin[1], Natalia Nesterova[2], Olga Soboleva[3]

[1]Perm National Research Polytechnic University, Perm, Russia, kurushin.daniel@ya.ru
[2]Perm National Research Polytechnic University, Perm, Russia, nest-nat@ya.ru
[3]Perm National Research Polytechnic University, Perm, Russia, olga.v.soboleva@gmail.com

Abstract: The paper deals with the problems of modeling speech recognition systems. The authors proposed to use the mechanism of linguistic anticipation in the speech recognition systems. It is known that anticipation is a kind of phenomenon of anticipatory reflection, which can provide an opportunity for the subject to "look into the future." Anticipation is believed to be an effective method of improving reading technique in children as it enables to increase the speed of reading [1]. The similarity of the learning processes of the human brain and artificial neural-like algorithms allows to suggest that the inclusion of anticipation mechanisms into the operation of the speech recognition algorithm can improve the quality of the system. The paper presents the experiment carried out with the purpose to study the probability of increasing the quality of modern speech recognition systems provided that linguistic anticipation is embedded into such a system. The obtained results are discussed and possible directions for further work on this topic are considered.

Keywords: natural language processing, speech recognition systems, language models, anticipation.

Introduction

The term "anticipation" (lat. *anticipatio*) is known to be introduced into scientific discourse by V. Wundt [2] in 1880. There are other decignations for this mental mechanism, among which the following are the most well-known in Russian science: "ustanovka" (readiness to foresee and to expect future events) (D.N. Uznadze), "probabilistic forseeing" (I.M. Feigenber) and "anticipatory reflection" (P.K.Anokhin). However, today it is the anticipation that can be considered an umbrella term for "forward vision" (the literal translation of the word "anticipation" from Latin).

B.F. Lomov and E.N. Surkov (1988) define anticipation as "the ability to act and make certain decisions with a certain time-spatial anticipation in relation to expected future events" [3]. We agree that this definition of the psychological mechanism of foreseeing is rather exact, but it describes anticipation on the whole while we are interested in the so-called linguistic (language) anticipation, which is a special manifestation of the universal mechanism i.e. the effect of the subsequent language form on the form preceding it in the sequence. Linguistic anticipation allows you to predict the content of the text by the name, surname of the author, epigraph, etc., as well as to restore the missing elements of the text and guess the author's train of thought. Anticipation is considered an effective way of teaching reading techniques, because with systematic training children learn to guess a word by initial letters, a phrase by initial words, the content of the text by initial phrases, which significantly accelerates the pace of reading.

In our study, we made an attempt to consider the possibility of embedding the mechanism of linguis-

tic anticipation into speech recognition systems. It is known that the creation of a reliable speech recognition system resistant to noise with a low error rate is one of the urgent tasks of modern robotics [4]. For the solution of this task it is necessary to involve data from applied linguistics, neurolinguistics and psycholinguistics.

Over the past decades a large number of speech recognition systems have been designed. On the one hand, there are commercial projects, such as Microsoft Speech-To-Text [5], Google Speech API [6] and Yandex SpeechKit [7]. On the other hand, there are also open source systems that allow you to control the recognition parameters yourself and integrate yourself into a third-party software. Among open-source recognizers there are such systems as Mozilla DeepSpeech [8], Kaldi [9], CMU Sphinx [10]. The main difference between open-source systems and commercial ones is their autonomy (independence from Internet connection), anonymity (achieved through source code control), as well as flexibility of settings.

However the systems mentioned above have certain problems with speech recognition often inferior to commercial ones in quality. This is quite understandable since large human and financial resources are spent on the development and maintenance of commercial systems. They have large client bases which allow to make algorithm configuration more effective. Companies such as Yandex can afford to distribute micro-tasks for training their systems to the user community through services such as "Yandex.Toloka" [11].

PROBLEM FORMULATION

One of the most popular ways to improve and configure speech recognition systems with source code is to modernize the dictionary [12] and the language model of the system, because these "out of the box" entities are actually not suitable for use and they in its turn play a very important role in the recognition process.

In the course of early studies [13] it was revealed that the CMU Sphinx system achieved the best performance indicators when using medium-volume language models. But due to the fact that the medium-volume model is inherently built for a specific subject area, it can be concluded that for the correct operation of the system it is necessary to use several language models that are connected as needed.

Based on the definition of linguistic anticipation given above we understand the *anticipation* as a predictive process of loading language models into the speech recognition system relative to the intended stage of the conversation. We call a *language model* a set of coefficients formed during the training of a classifier, forming a probability distribution on a set of dictionary sequences [14].

In the course of the study three algorithms for simulation of anticipation were developed:
1) based on the classification of texts,
2) based on the transition probability matrix,
3) based on the dialog scheme.

The algorithm based on the classification of the texts is based on the idea that in the phrase uttered by a person in most cases there will be hints (perhaps difficult to notice) on the subsequent topic of the dialogue. The idea was put forward to use a classifier [13] which is given the last recognized phrase at the input, and at the output the classifier shows which of the language models the current phrase belongs to. It also shows the probabilities of correlation between the phrase and the subsequent language model.

The basis of the algorithm of the transition probability matrix was Markov chains [15]. The mathematical model of a Markov chain is based on the probability matrices of transitions of the system into possible states at each of the steps. In our case we take the existing language models as states, i.e. we have six states. It is supposed to build a transition matrix, where each cell of the matrix means the probability of transition to the next step from model to model.

The algorithm based on the dialog scheme [16] is the least labor-intensive from the point of view of implementation, but also the most unpredictable from the point of view of the result. To ensure the effectiveness of this method, it is necessary to create a reference scheme of the dialog, which is based on the language models loaded into the recognition system one at a time.

EXPERIMENT AND ITS RESULTS

The aim of the study was an experimental investigation meant to measure the performance of vari-

ous speech recognition systems. Standard and predictive algorithms for using language models based on industrial volumes of data were chosen for the experiment. The initial data for building model were taken from the project "Open speech to text", published in May 2019. The collected data include 10 thousand hours of annotated oral speech collected from various Internet sources, such as books, calls, YouTube videos, etc. (a total size of over 600 gigabytes).

The test data for the experiment was a thousand pre-dictated audio files. Metrics such as WER and RTF were used to measure the performance of systems [17]. The Word Error Rate (BER), or error measure, was implemented to evaluate the accuracy of the recognition system and is described by Equation 1.

$$WER= \frac{I+D+S}{N}, \quad (1)$$

where I is the number of inserted characters, D is the number of deleted characters, S is the number of replaced characters, N is the number of characters in the recognized word. Since in some cases the numerator value may be greater than the denominator value, the *WER* value can be greater than 1.

The *RTF* metric in turn is used to estimate the time spent on speech fragment recognition. *RTF* is described by equation 2 and is the ratio of the time spent by the system on recognition. The *TPP* length of the audio file is given in seconds *LEN*.

$$RTF= \frac{TPP}{LEN} \quad (2)$$

In the experiment the performance indicators of three autonomous speech recognition systems were compared. They are:

1) CMU Sphinx,
2) Kaldi,
3) Mozilla Deep speech.

The record of the experiment is presented in Table 1.

During the experiment it was revealed that the methods of constructing and using language models based on anticipation show a positive result only in the case of the CMU Sphinx system. The other systems worsen the indicators. Therefore one can make the following conclusion:

1) The CMU Sphinx system is based on fairly old speech recognition algorithms and therefore reducing the training sample for a given context has a positive effect on its operation.

2) More modern systems, such as Kaldi and Mozilla, are able to work with large training samples. A decrease in the sample size generally has a negative effect on classifying systems [18-20].

3) When teaching on large samples, which happens in modern systems, some patterns of dialogue development are taken into account "automatically".

Table 1 shows that the usage of Kaldi system demonstrates the best results. The percentage of recognition errors for it was only 19%, and the indicator of the time spent on recognition is considered practically a reference for open source systems.

Conclusion

During the experiments a rather contradictory result was obtained. On the one hand, the use of anticipation for some systems gives a positive effect as expected. On the other hand, for more modern systems the effect is absent or may even be regarded as negative. In addition, the probability of erroneous recognition at 19% is still high enough to talk about the applicability of the system in real conditions. In addition, it can be noted that, for example. CMU Sphinx gives a better result using the transition probability matrix than Mozilla DeepSpeech without anticipation. This suggests that it is possible to

Table 1 – The results of the experiment

System	CMU Sphinx		Kaldi		Mozilla DeepSpeech	
Metric	WER	RTF	WER	RTF	WER	RTF
The original algorithm	0.79	3.2	0.19	1.1	0.48	1.6
Based on text classification	0.32	1.7	0.30	1.0	0.49	1.6
Based on the transition probability matrix	0.29	1.7	0.28	1.2	0.48	1.5
Based on the dialog scheme	0.47	1.6	0.69	1.1	0.59	1.6

WER: less is better.

RTF: Less is better.

create more advanced anticipation mechanisms and more advanced mechanisms for embedding the anticipation model into the speech recognition mechanism.

One of the solutions to this problem, in our opinion, can be a formalized representation of the subject model, presented in the form of the graph which can explicate the denotatum structure of a unit of speech. The algorithm for generating such a graph was developed by the famous Russian linguist Anatoly Ivanovich Novikov [21]. In the future, this algorithm was adapted for machine use [22], and at the moment there is reason to believe that the modeling of future speech recognition systems will be based on the mechanisms of anticipation, which, in turn, will be based, among other things, on a model of a specific subject area.

References

[1]   Anticipaciya: chto eto, kak ona proyavlyaetsya i korrektiruetsya. – URL: https://glpni.ru/lekapctva/antitsipatsiya-chto-eto-kak-ona-proyavlyaetsya-i-korrektiruetsya.html (data obrashcheniya: 26.11.2021)

[2]   Vundt V. Osnovaniya fiziologicheskoj psihologii. Ob elementah dushevnoj zhizni. Intensivnost' oshchushcheniya. Per. s nem. M.: URSS, 2010.

[3]   Lomov B.F., Surkov E.N. Anticipaciya v strukture deyatel'nosti. M., 1980.

[4]   Sal'nikov I.S. Aktual'naya problematika fundamental'nyh i prikladnyh issledovanij v oblasti komp'yuternyh tekhnologij, mobil'noj svyazi, robototekhniki i iskusstvennogo intellekta // Problemy iskusstvennogo intellekta. 2018. №3 (10).

[5]   Microsoft Speech-To-Text // Microsoft Speech API. – URL: https://docs.microsoft.com/en-us/azure/cognitive-services/speech/home (data obrashcheniya: 26.11.2021).

[6]   Google Speech API // Cloud STT. – URL: https://cloud.google.com/speech-to-text/ (data obrashcheniya: 26.11.2021).

[7]   Yandex SpeechKit // Cloud – Raspoznavanie rechi – Tekhnologii YAndeksa. – URL: https://tech.yandex.ru/speechkit/cloud/doc/guide/concepts/asr-overview-technology-docpage/ (data obrashcheniya: 26.11.2021).

[8]   A TensorFlow implementation of Baidu's DeepSpeech architecture // github.com. – URL: https://github.com/mozilla/DeepSpeech (data obrashcheniya: 26.11.2021).

[9]   Kaldi // Kaldi Speech Recognition Toolkit. – URL: https://github.com/kaldi-asr/kaldi (data obrashcheniya: 26.11.2021).

[10]   Alimuradov A.K., Tychkov A.YU., Zareckij A.P., Kuleshov A.P., CHurakov P.P., Kvitka YU.S. Metod povysheniya effektivnosti golosovogo upravleniya na osnove komplementarnoj mnozhestvennoj dekompozicii na empiricheskie mody // Trudy MFTI. 2017. №2 (34).

[11]   YAndeks.Toloka. – URL: https://toloka.yandex.ru/ (data obrashcheniya: 26.11.2021).

[12]   Raspoznavanie rechi s pomoshch'yu CMU Sphinx / Habr // habr.com. – URL: https://habr.com/ru/post/267539/ (data obrashcheniya: 26.11.2021).

[13]   Obzor metodov klassifikacii v mashinnom obuchenii s pomoshch'yu Scikit-Learn // tproger.ru. – URL: https://tproger.ru/translations/scikit-learn-in-python/ (data obrashcheniya: 30.11.2021).

[14]   Grinin Igor' Leonidovich Rabota modeli generacii teksta s pomoshch'yu nejronnyh setej kak sostavnoj sistemy: modul'nyj analiz modul' pervyj. YAzykovaya model': rabota s tekstovymi vhozhdeniyami // Innovacii i investicii. 2020. №7.

[15]   CHto takoe skrytye modeli Markova / Habr // habr.com.. – URL: https://habr.com/ru/post/135281/ (data obrashcheniya: 30.10.2021).

[16]   GitHub - daniel-kurushin/three_word_english // github.com URL: https://github.com/daniel-kurushin/three_word_english/blob/master/10.jpg (data obrashcheniya: 30.10.2021).

[17]   Karpov A. A., Kipyatkova I. S. Metodologiya ocenivaniya raboty sistem avtomaticheskogo raspoznavaniya rechi // Priborostroenie. 2012. №11.

[18]   Mingalev A. V., Belov A. V., Gabdullin I. M., Agafonova R. R, SHusharin S. N. Raspoznavanie test-ob"ektov na teplovizionnyh izobrazheniyah // KO. 2019. №3.

[19]   Novikova N. M., Noaman S. A. Komp'yuternaya model' statisticheskogo raspoznavaniya izobrazhenij // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Seriya: Ekonomika. Informatika. 2012. №13-1 (132).

[20]   Mokeev V. V., Tomilov S. V. O reshenii zadachi raspoznavaniya izobrazhenij metodom glavnyh komponent i linejnym diskriminantnym analizom // KO. 2014. №4.

[21]   Novikov A.I. Semantika teksta i ee formalizaciya. – M.: Nauka, 1983.

[22]   Kurushin D.S., Nesterova N.M., Chovich L.I. // XIII međunarodni naučno-stručni skup Informacione tehnologije za e-obrazovanje: Zbornik radova, Banja Luka, Izdavač: Panevropski univerzitet "APEIRON", 2021. p. 81-86.

ABOUT THE AUTHORS

Daniel S. Kurushin, born in Perm, Russia, holds PhD in technical sciences. He graduated from Perm National Research Polytechnic University where he works at present as an Associate Professor at the Department of Information Technology and Automated Systems. For five years he had been working at Vienna University of Technology as a researcher. His main research interests are in artificial intelligence, robotics and systems of technical vision.

Natalia M. Nesterova, Doctor in Philology, works as a professor at the Department of Foreign Languages, Linguistics and Translation of Perm National Research Polytechnic University. She received her PhD at the Institute of Linguistics of Russian Academy of Sciences (Moscow), later she received a doctorate degree at Perm State University. Her research interests are in applied linguistics, psycholingustics and speech perception mechanisms.

Olga V. Soboleva, born in Perm, Russia, holds PhD in philology. She graduated from Perm State University and at present works as an Associate Professor at the Department of Foreign Languages, Linguistics and Translation of Perm National Research Polytechnic University. She is specializing in teaching Russian as a foreign language, information technologies in linguistics and automatic natural language processing systems.

E-Learning Platform Directions and Future Expansion With Case Study

Nedeljko Šikanjić[1], Zoran Ž. Avramović[2]

[1]*PhD student at Pan-European University Apeiron, Banjaluka*

[2]*Pan-European University Apeiron, Banjaluka*

(nedeljko.s.sikanjic, zoran.z.avramovic)@apeiron-edu.eu

Abstract: When we look at the current situation in the world we can see that world shifts into digital era. This means, it will also influence the learning and educational section. In this science paper we will analyze e-learning platform architecture, propose architecture based on the teaching process and perform comparative analysis of leading e-learning provides.

Keywords: E-learning, Education, Databases.

Introduction

When we look at the current situation in the world we can see that world shifts into digital era. This means, it will also influence the learning and educational section.

We can narrow down few of the advantages and disadvantages researched with students and their professors.

Benefits are that lectures are being recorded and posted for the easy access for students. Also, another benefit is that students could reach teacher and professors at any time. Response from professor and teaching stuff was in time and with consultation provided. Another benefit is that location is totally irrelevant, so the approach was flexible for most of the students and professors.

Disadvantages are lack of social interaction in term of everything was done online. There is certain physical need to be established between people in general so the same goes for the teachers and students.

Techniques being used in e learning were the same as in face to face class rooms. This means, lectures, exams and essays were being organized online. When it comes to lectures, benefit was that it was being streamed and recorded at the same time, which helped students to pick up some stuff they missed or some issues were overlooked and this way they could rewind or go back to part of the lecture they needed to get explanation.

From the part of submitting essays, this was similar process as it was before. Students were already used to send written essays to the professors, so in e-learning this process was not changed at all.

However, from the part of exam or taking exam process, this required some adjustments. For example, if exams was structured in real face to face situation, now it needed to be stricter as online, professor has no influence on the student environment. Of course, this was also being possible to track by having included online camera, and with camera with sound on, professor or teaching stuff, had possibility to overview the exam process. Another benefit of e-learning is screen sharing, as with this method interaction with students on exams or labs, was almost the same as in real face to face situations.

With this being said, facts that points to e-learning, has proven that new digitalization era is already in place, and that e-learning provides similar or same effects as regular face to face learning process.

Factor that is important for e-learning to be successful is the modernization and fact that students

in their early age have learned how internet is being searched and how to use different kind of applications. They are also familiar with capability of retrieving desired information over internet and how to work with various forms of documents.

Depending of area where e-learning can be implemented, providing knowledge without limits from seafarers and providing centralized certifications for example with using Moodle [1], to concepts of providing e-learning in space program [2].

Here we will analyze e-learning tools architecture, their structure and how they are helping improving learning process.

Architecture of E-Learning Platform

When human society begin their innovation era, they basically imitated forms in nature and around us. This has helped from the early stage such as Da Vinci innovations like flying machine, to help humans advance and prosper.

Same thing applies to the e-learning platform. To have successful e-learning platform that fulfills its goal, we need to imitate the learning process in educational systems.

Structure is to have entities or objects such as classrooms, subjects and some form of chat interaction.

Entities are designed based on the domain level they present, with properties as holders or keepers of the values in the right form. Within concept of domain, processes must be executed in right way where the relationship between entities and domains should provide the real life examples.

Relationships between entities should follow loose coupled principle, where based on the poten-

Fig. 1. *Architecture of e-learning platform*

tial growth of e-platform, should be enough flexible to implement new concept needed to follow learning concepts.

With chat interaction, we can solve everything. We just add the entities such as students, location or classroom and subject for what chat relates to. Also, email is another tool that also is being added to e-learning platform. Scheduling or appointments for calendars makes automation to perfection, as the students and professors can organize the learning activities.

Regarding storing data we can use document oriented databases or we can use relational database systems. For example, one of the e-learning providers Udemy is using document oriented database called MongoDB, and other e-learning provider such is Codeacademy is using SQLite as one of the databases.

We can use advantage of new database concepts for storing large data volumes [3], however this all depends on how we want to scale learning content including tests and certification data.

Types of E-Learning Platforms

We can divide e-learning platforms based on different criteria. For example, one important criteria is financial criteria. In this way we can say there:
- Free (open source)
- Paid (commercial)

If we look by the type of access we can divide into:
- Computer/Device based (offline)
- Online (internet/cloud) based

If we look by the type of usage criteria we can say:
- Company based (for employees)
- Massive Open Online Courses (MOOCS)

When we look at financial aspect, it is pretty simple to recognize what are the benefits. However, just because it is free, it doesn't mean it has less quality, but it has tendency of having non-assigned administrators or teams that are supervising the quality of the learning content. On the other hand, paid courses have this advantage, as they can finance these resources.

Type of access of the learning platform we can say that everything that is offline and can be access by the computer or device (mobile), it has the benefit of using it where there is no access to the in-

ternet. This now is not that much of importance as most of the online courses can be downloaded as offline content, so it can be viewed later in the offline environment.
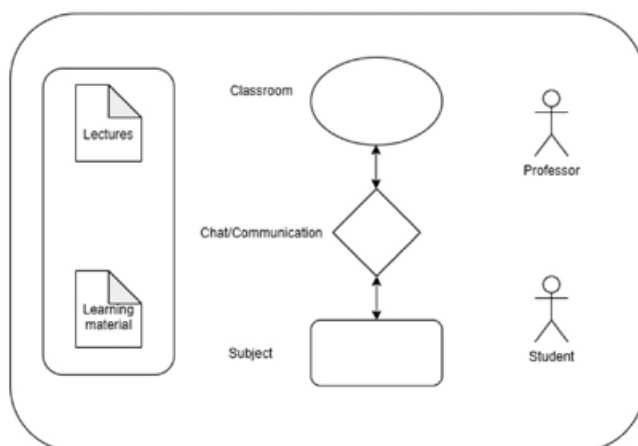
Types of usage in the term of who are teachers and students, we can say that companies where we can have also other instructions such as schools or universities, have dedicated teachers and learning content is restricted to the audience such as employees or students. One very important e-learning type of platform, which is gaining significant market share in the e-learning community is MOOCS. They have been constructed in the way that teacher can be student and student can be teacher. There are different types of content validation if the learning content is good or not. One of the main criteria is the reviewing system by the students, which have been proven as one of key value to recognize if course provides value to the students.

UDEMY CASE STUDY

Here we will address Udemy as one of the leaders in the e-learning platforms. Udemy is not that old (started by getting funding in 2010) company which has started their platform with only one course and it was a video they published on the YouTube platform. In Udemy, they wanted to solve one problem and it is called "chicken-egg" problem. By this, people from Udemy, didn't know how to get teacher to create and publish their courses on the platform where are not students, and they didn't know how to get students on the platform that has no teachers with courses.

Interesting was how they manage to use outsourcing platform called ODesk (previous Upwork), where they managed to reach to each freelance which had different skills, in order to get them to start a course from their branch on Udemy platform. This later proved to be huge success in getting the right teacher and by having the right learning content, they managed to get students to enroll in these courses.

Udemy had a opportunity to start with business to client B2C model but later switched to the busi-
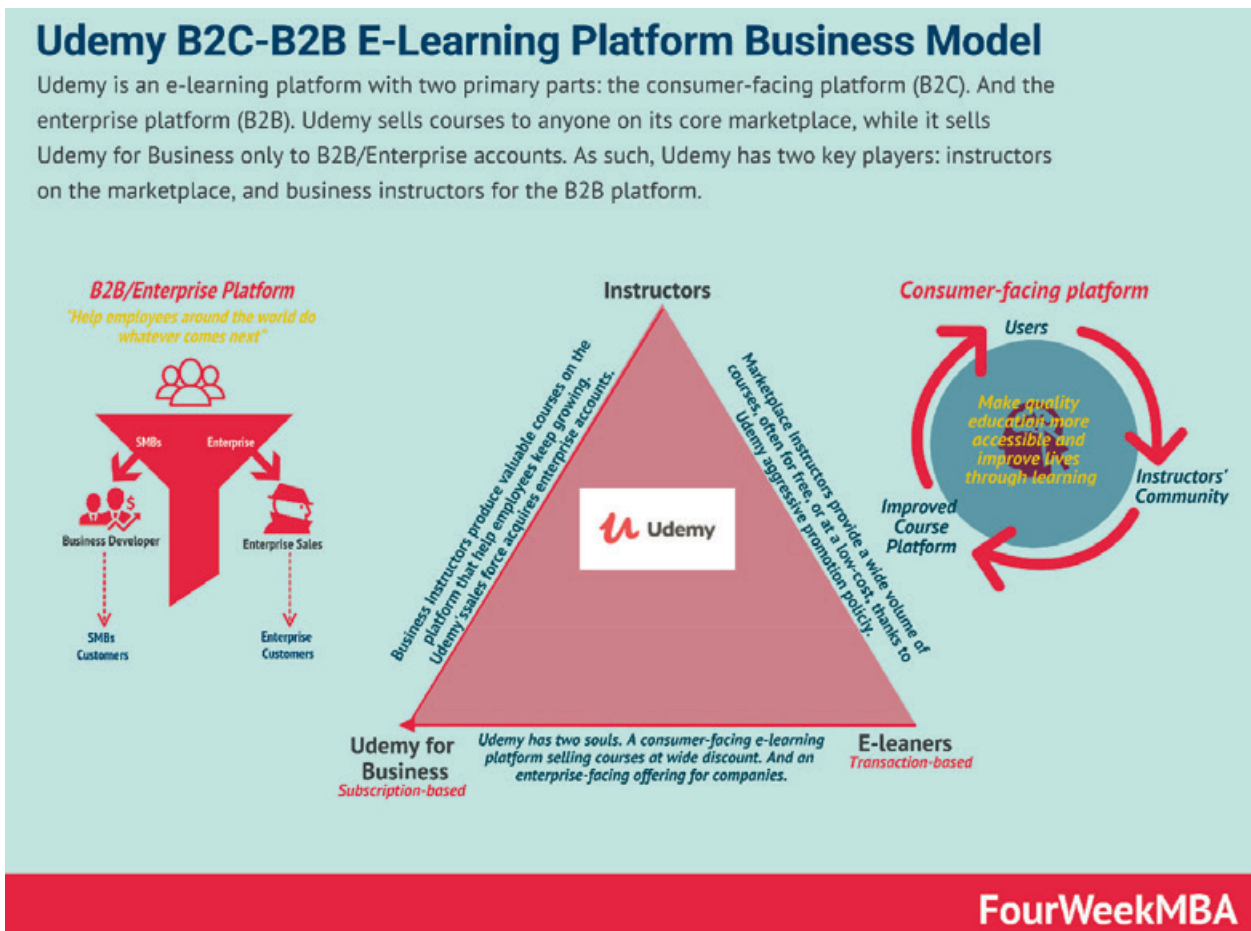
Fig. 2. Number of students on e-platforms in millions

ness to business B2C model [4]. Also Udemy has some courses for free, and this bring more students to the platform itself. With this approach, and by previously mentioned types of e-learning platform, we can see that Udemy has mixed different types into hybrid model, and by doing this approach has made a huge success and became a one of the leaders in e-learning business market.

For comparison, Udemy in 2015 had 8 million students, 32 000 courses and 18 000 teachers [5].

Today, Udemy in 2021 has 44 million students, 143 000 courses and 65 000 teachers.

Below is the chart of comparison with other popular e-learning platforms in term of quantity of courses provided on each of them.
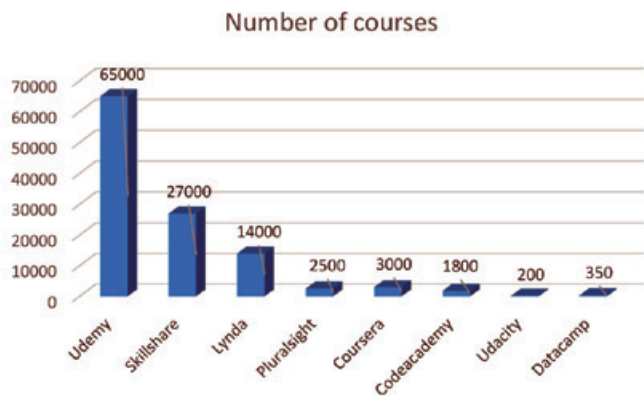
Fig. 3. *Number of courses on e-platforms*

As we can see on the graph above, Udemy is leader in providing the absolute biggest number of courses on e-learning platform. Following is Skilshare e-

learning platform with 27 000 courses, Lynda with 14 000 courses, then Pluralsight and Coursera with aproximalty 3000 courses, Codeacademy with less than 2000 courses, Datacamp and Udacity respectively with 350 and 200 courses.

To analyze more deeply, here are the description in the table of pricing models and coverage of topics.

As we can see in the table above, pricing models are pretty similar except for some e-learning providers like Udacity where courses have period of couple of months to be attended. However, some providers also provide cooperation with Universities such is Coursera, where you can also enroll in degree program and earn an official university degree. When it comes to topics of courses, they also tend to be almost the same except for some e-learning providers like Pluralsight, Datacamp and Codeacademy, which are specialized in certain topics like coding and database development.

Another comparative analysis needed to be done in order to get more information are number of students attending each of e-learning platforms.
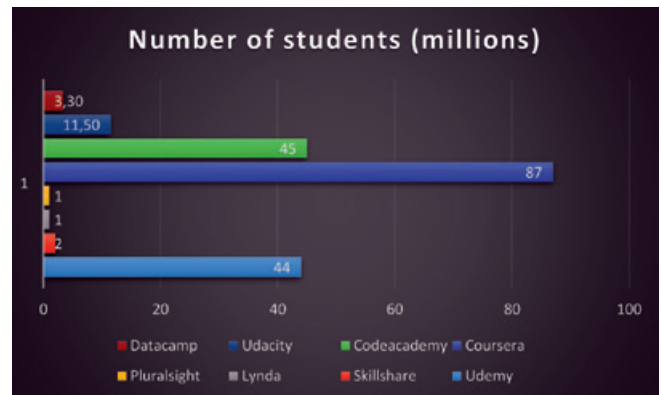
Fig. 4. *Number of students on e-platforms in millions*

Table 1. *E-learning pricing models and coverage of topics*

Name	Pricing model	Coverage of topics
Udemy	0-199$ (mostly they are at 12$)	Best Selling, Development, Design, Business, Personal Development
Skillshare	32$ monthly	Animation/Creative, Business/Marketing, Lifestyle/Productivity
Lynda	$25 monthly	Business, Creative, Technology
Pluralsight	$35 monthly	Technology/Development
Coursera	30-80 monthly	Data Science, Business, Computer Science, Personal Development, Information Technology, Language Learning, Health, Social Sciences, Math and Logic, Physical Science and Engineering, Arts and Humanities
Codeacademy	40$ monthly	IT/Development
Datacamp	29$ monthly	Database/Data Development
Udacity	718$- 2154$	Programming and development, Artificial Intelligence, Cloud Computing, Data Science, Business, Autonomous Systems, Career/Personal Development

On the graph above is presented the number of students in millions on e-platforms, where we can see disproportional numbers to the previously presented number of course categories and prices. For example, Coursera [6] has the highest number of students, 87 million enrolled in their course but the number of courses are 3000 compared to Udemy which has 65000 courses and 44 million students enrolled. However, what is symptomatic that with Coursera students can pursue the university degree and with Udemy it is not possible.

Another example which can prove paradigm that number of courses are not valid factor for gaining students is Codeacademy [7], which has 45 million students with nearly 2000 courses. This proves that niche e-learning platforms are also viable in providing value, knowledge to students and contributing to the e-learning market share.

Another conclusion what can be reached from these comparative analysis is that Udacity [8] with high pricing courses has 11.5 million students. We need to mention that this e-learning platform has accomplished that success with 200 courses.

Conclusion

E-learning has proven to be mandatory part of human society. Pandemic that unfortunately occurred in 2020/2021 has just pushed the limit of introducing knowledge to students. With understanding fast developing technology, we should embrace the new methods of grasping knowledge and sharing the knowledge with others.

In this science paper, was researched the future possible development of e-learning platforms, also was analyzed the current e-learning leaders with their various approaches, including number of courses, financial models and knowledge areas that are being covered.

We can conclude that e-learning has bright future, with making students grasp knowledge in easy and understandable way. Also, teachers are assimilating to the new technologies, using it to provide quality content for students and taking advantage of the e-learning platforms to make it possible to validate students' knowledge and track progress of each student.

References

[1] Sanja Bauk, Michael Kopp, Avramović Zoran: A Case Study On Introducing E-learning Into Seafarers' Education , JITA – Journal of Information Technology and Applications, PanEuropien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina JITA 3(2013) 1:34-42, (37.018.43:004.738.5), (DOI: 10.7251/JIT1301034B), Volume 3, Number 1, ISSN 2232-9625 (print), ISSN 2233-0194 (online).

[2] The "eLearning in Space" project, Arizona State University, United States of America, https://interplanetary.asu.edu/eLearningin Space (accessed on Sept. 15 2021)

[3] Nedeljko Šikanjić, Zoran Ž. Avramović, NEW APPROACH OF STORING AND RETRIEVING LARGE DATA VOLUMES, JITA – Journal of Information Technology and Applications, PanEuropien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina JITA 9(2019) 2:89-98, (UDC: 004.6:658.4]:519.8), (DOI: 10.7251/JIT1902089S), Volume 9, Number 2, ISSN 2232-9625 (print), ISSN 2233-0194 (online)

[4] https://fourweekmba.com/udemy-business-model/ (accessed on Sept. 20 2021)

[5] https://about.udemy.com/press-releases/140-year-over-year-growth-in-number-of-online-courses-taken-in-the-uk/ (accessed on Sept. 22 2021)

[6] https://about.coursera.org/press#:~:text=Coursera%20was%20launched%20in%202012,as%20of%20June%2030%2C%202021. (accessed on Sept. 24 2021)

[7] https://en.wikipedia.org/wiki/Codecademy (accessed on Sept. 24 2021)

[8] https://e-student.org/udacity-eview/#:~:text=By%20now%2C%20the%20number%20of,on%20artificial%20intelligence%2C%20for%20example (accessed on Sept. 24 2021)

ABOUT THE AUTHORS

Nedeljko Šikanjić holds a Magister degree in Informatics and Computer Science and has worked for more than 15 years as a Software and Database Architect/Engineer. His main fields of studies are in the area of advanced Databases and Software Architectures. He has been a holder of an active Microsoft Certified Trainer Certificate since 2012 and has been teaching courses on various topics in Information Technologies. Doctoral studies of the third degree enrolled in the academic 2017/2018.

Zoran Ž. Avramović was born in Serbia, on September 10, 1953. He finished elementary school and high school in Loznica with great success. He was awarded several diplomas by Nikola Tesla and Mihailo Petrović Alas. He graduated on time at the University of Belgrade - Faculty of Electrical Engineering, with an average grade of 9.72 in five-year studies. He received his master's degree at that faculty (all excellent grades, exams and master's degrees), and then obtained a doctorate in technical sciences (in 1988). As an excellent student of the University, he had the right and at the same time studied mathematics at the Faculty of Mathematics in Belgrade. He was the champion of Serbia in mathematics ("first prize") and Yugoslavia in electrical engineering ("gold medal").

Software Platforms Based on the Principles of Graphic Design, Automatic Command Generation and Visual Programming

Dražen Marinković, Zoran Ž. Avramović

Pan-European University APEIRON Banja Luka, Republika Srpska, BiH

(drazen.m.marinkovic, zoran.z.avramovic)@apeiron-edu.eu

Abstract: This paper presents a new approach to software application development using a graphical interface. The approach is based on a combination of drag and drop elements and logic based on the model's own concept. Low code platforms and principles have been developed and are still being developed precisely to enable the rapid creation and use of applications that meet all the special needs and requirements of various organizations. No code platforms allow professionals and laymen to create applications via graphical user interfaces without any prior knowledge or qualifications in programming. However, code platforms are closely related to low code platforms because they are both created with a similar goal, based on a very similar way of working and almost the same principles of operation. Many vendors point out that the future of software development is based on configuration, not program. We believe that eliminating code is one way to bring development to higher standards in application development. One of the biggest advantages of the LC/NC platform is that they allow us to take advantage of innate problem solving and human skills by removing at least a significant number, if not all barriers to implementing software solutions in today's software world.

Keywords: low code, no code, visually integrated development environment, low-skilled people, professional developers.

Introduction

Low code/no code is an approach to software application development using a graphical interface. The approach is based on a combination of drag and drop elements and logic based on the model's own concept. This approach aims to increase the number of those developers who can participate in the processes of creating and using software to achieve a business goal or facilitate some part of the business process.

Development low code platforms are called visually integrated development environments (IDEs).

The approach to software application development according to the low code/no code concept generally follows the following steps: request determination, API selection, workflow creation, application development, data model, user interface using visual IDE, API connection, SQL queries custom software solution, software user acceptance testing, application publishing and upgrading as needed.

Development using the low code/no code platform helps low-skilled people to create applications or some of its components, without the need for the participation of professional developers. Low code development is also beneficial to professional developers, as it shortens the time needed to develop projects and opens up the possibility for them to distribute parts of their projects to low-skilled staff.

Low code

Low code development platform offers a working environment for creating applications using a graphical user interface, instead of traditional programming. The low code approach to application development reduces the amount of programming, enabling accelerated application development and delivery. The benefit of this approach is that a much larger group of people can participate in the development of the application.

Low code platforms come from fourth-generation programming languages and tools for rapid application development. Similar to their predecessors, low code platforms are based on the principles of model-based design, automatic command generation, and visual programming. Low code platforms came into use in 2011. The term "low code" was first used on 9 June 2014 by Forrester Research.

Increasing the need for automation and constant new applications for all types of business processes places a requirement on developers to create specific applications that are tailored to the specific needs of each organization individually. Low code platforms and principles have been developed and are still being developed precisely to enable the rapid creation and use of applications that meet all the special needs and requirements of various organizations.

Some of the well-known platforms for low code development are:
- Quixy
- Visual Lansa
- Creatio
- GeneXus
- Appian
- Salesforce Lightning
- OutSystems

No code

No code platforms allow professionals and laymen to create applications via graphical user interfaces without any prior knowledge or qualifications in programming. However, code platforms are closely related to low code platforms because they are both created with a similar goal, based on a very similar way of working and almost the same principles of operation. The popularity of the no code principle of application development is constantly growing, primarily because it offers a huge number of possibilities without requiring any prior knowledge.

No code applications are used to meet the needs of companies that want to digitize processes through applications. Nevertheless, code tools are often created with the goal of adapting to business users, as opposed to traditional IT tools and technologies. This approach aims to accelerate the development cycle by bypassing time constraints, small budgets

Figure 1: LowCode / https://www.outsystems.com /

or insufficient resources to develop a complete software. The transition from traditional programming to abbreviated development methods is also making changes in the roles of traditional IT teams.

Some of the well-known platforms for no code development are:

- Zoho Creator
- Caspian
- Tigersheet
- Mendix
- Quickbase
- Retool
- Microsoft Power Apps
- OutSystems
- AirTabel

The difference between low code and no code

Low code application development is a convenient way for developers to develop applications quickly and with a relatively small amount of programming.

However, application code is created by users who do not need to know any programming language to create applications.

While no code platforms are more suitable for business users and non-professionals because the use of the platform does not require any programming knowledge, low code platforms are more suitable for more advanced users and more skilled developers.

LOW CODE/NO CODE IN IMPLEMENTATION

Practical implementation

There are, primarily, three types of applications that can be built on LC/NC platforms:

- **Portals and Web applications** - a good solution for those who want to create a personal or business website or blog.
- **Back Office applications** - provide ease of administration, automation of some routine or frequent tasks, creating reports
- **Mobile applications** - the most common and widespread type of application today, LC/NC provides companies with the opportunity to get closer to their customers in this way.

Advantages of low code/no code applications

- **Cost reduction**

With LC/NC platforms, development costs are greatly reduced, both due to the reduction of development complexity and the need to hire professional programmers.

- **Speed**

Development with LC/NC platforms allows companies and individuals to develop and modify their applications faster. The skills required to create applications using LC/NC platforms are much lower compared to traditional development.

Furthermore, all configuration processes, including the development of the program itself, its compilation and evaluation, and debugging, take place much faster in LC/NC development than in traditional.

- **Facilitated user experience**

LC/NC platforms automate many more operations critical to the user experience. The simplicity of these applications is reflected in their handling as much as in their creation.

- **Easier path from idea to realization**

By using the LC/NC platform for the development of applications specific to certain types of work, it is

Figure 2: Low Code No Code / https://quixy.com/blog/no-code-versus-low-code/

possible to create an application for the person who should perform the work. In this way, the possibility of misunderstanding and disagreement between the one who creates the application and the one who should use it is eliminated. In this way, one of the frequent phases in the design and implementation of software systems is eliminated, and that is working with the client. With LC/NC applications, there is usually no need to collect user requests or user impressions.

- **Compatibility**

One of the main advantages of low code/no code applications is that they can run on almost all operating systems, and most types of devices.

Disadvantages of low code/no code applications

- **Third party dependence**

With a low code/no code platform, the operation of the application depends on the owner of the platform almost as much as on the one who creates it. The creator of the application is not able to work with any background functions or to release updates. For all such jobs they depend on the owner of the platform and their legal engagement in case of any problem.

- **Unable to customize**

Low code/no code platforms generally offer very few options for developing custom software or custom applications.

LC/NC platforms have very few options for integration with existing software solutions, which can be a significant problem for companies or individuals who already have a large amount of data contained in a previously used application.

- **Lack of developers**

The number of developers who have experienced low code/no code platforms is very small, and the number of other potential developers is even smaller. Companies and individuals who want to develop applications through these platforms are usually doomed to their abilities and the very small number of available tutorials and help forums.

- **Impossibility of quality creation of larger software systems**

The low code/no code approach to application development is a good solution for small or medium applications, but complex, large applications with many special functions and procedures still remain strictly in the domain of traditional programming and large, familiar programming languages.

COMPARISON OF LC/NC APPROACH TO TRADITIONAL PROGRAMMING

While low code/no code application development platforms allow developers, users, and businesses to quickly develop applications using a variety of visual interfaces, traditional application development methods use conventional methods that allow developers to create large and powerful applications. Some of the basic differences in these approaches to development are:

- **Prerequisites required**

The traditional approach to application development generally involves the need to know the various tools and functions that developers use in creating complex applications. To be used properly, these applications require highly technical and specialized skills in this area.

Low code/no code platforms usually have a small set of tools that help develop different types of applications without much trouble. Application development with this approach becomes much easier for non-professional users, but the type and complexity of applications that can be developed still remains on the side of traditional programming languages.

- **Application quality**

Due to complex technologies and a very elaborate nature, traditional application development platforms usually produce errors and are themselves often quite complex, making it difficult for users to use applications.

In contrast, low code/no code platforms generally do not produce application design errors. This makes it easier to handle the application on the user side, but only applies to some smaller, more general applications.

- **Price**

Traditional software is quite expensive to build and can range from a few hundred KM to several hundred thousand KM for application design and development, depending on the size and complexity of the system. This is generally a better option for large companies that need extremely complex and specific software.

LC /NC platforms are much cheaper than traditional software for most companies. This is primarily because only access to the platform is paid for, not the development of the application itself.

- **Speed of development**

Traditional development platforms generally have a complicated setup system, which makes them less agile. Due to the complex codes, it takes a lot of time to learn and use them correctly.

By comparison, code platforms are much easier to use and operate, due to their drag and drop functions. Unlike traditional development, the code does not require any program development, but everything is done through inferfaces and already defined elements.

- **Maintenance**

In traditional programming, all kinds of modifications and upgrades go through the developer or team that developed that application. This process usually requires the adaptation of both the technologies used and the program itself. The software of the LC/NC application is usually maintained by the one whose platform it is, which makes the matter much easier for the company or the developer.

On the other hand, traditional software in most cases comes with professional support in charge of all kinds of corrections, troubleshooting and upgrades. In LC/NC applications, part of the job would be the responsibility of the company or individual who developed the application, which, depending on the problem, can be quite a time consuming process.

Conclusion

Without program development, organizations can separate the benefits of software from programming pitfalls. This paradigm is more important than ever because traditional methods have proven incapable of keeping up with the expectations of today's modern affiliate.

Many vendors point out that the future of software development is based on configuration, not program. We believe that eliminating code is one way to bring development to higher standards in application development.

One of the biggest advantages of LC/NC platforms is that they allow us to take advantage of innate problem solving and human skills by removing at least a significant portion, if not all the barriers to implementing software solutions in today's software world. Simply put, LC/NC increases the potential for innovation. This should resonate with organizations that have realized the need to become software-driven if they are to survive, or even thrive, in the years to come. Finally, the benefits of the LC/NC platform are crystallized in their ability to empower the entrepreneurial people in the organization to unleash their visions, ideas and creativity with minimal support through the construction of smart software solutions.

References

[1]   "10 best low-code development platformas in 2021," 27.09.2021 [Online]. Available at: https://www.software-testinghelp.com/low-code-development-platforms/. [Last accessed 01.10.2021].

[2]   T. Shukla, "8 best no-code platforms for building your app," 10.08.2021 [Online]. Available at: https://tiger-sheet.com/blog/the-8-best-no-code-platforms-for-building-your-app/. [Last accessed 30.09.2021].

[3]   B. Kumar, "Low code vs traditional development," 02.06.2020 [Online]. Available at: https://www.zoho.com/creator/decode/low-code-vs-traditional-development. [Last accessed 30.09.2021].

[4]   "No Code/Low Code Vs. Traditional Development - Which Team Should You Pick?," [Online]. Available at: https://marutitech.com/no-code-low-code-vs-traditional-development/#Low-CodeNo-Code_Application_Development. [Last accessed 01.10.2021].

[5]   "Low-code development platform," 23.08.2021 [Online]. Available at: https://en.wikipedia.org/wiki/Low-code_development_platform. [Last accessed 29. 09.2021].

[6]   "No-code development platform," 23.07.2021 [Online]. Available at: https://en.wikipedia.org/wiki/No-code_development_platform. [Last accessed 29.09.2021].

[7]   "What is low-code?," 2021 [Online]. Available at: https://www.creatio.com/page/low-code. [Last accessed 29.09.2021].

[8]   "No-code/low-code development," 27.07.2021 [Online]. Available at: https://golden.com/wiki/No-code_%2F_low-code_development-NMGMEA6/activity. [Last accessed 30.09.2021].

[9]   M. Azoff, "Why low-code/no-code is on the rise," 05.08.2021 [Online]. Available at: https://sdtimes.com/lowcode/why-low-code-no-code-is-on-the-rise/. [Last accessed 01.10.2021].

[10] B. McHugh, "What are the benefits of low-code IT development," [Online]. Available at: https://www.advsyscon.com/blog/benefits-of-low-code-it-development/. [Last accessed 30.09.2021].

ABOUT THE AUTHORS

Assistant professor **Dražen Marinković** was born on 1978 in Banjaluka. He finished elementary school and high school in electrical engineering. The Faculty of Business Informatics at the Pan-European University APEIRON Banjaluka enrolled in the academic year 2009/2010. years and graduated with an average grade of 9.81, defending a master's thesis entitled: "NoSQL databases in theory and practice" with a grade of 10. Academic 2015/2016 enrolls in doctoral academic studies of the third cycle at the Pan-European University APEIRON in Banjaluka, at the Faculty of Information Technology, and in the doctoral program "Information Systems in Communications and Logistics" where he passed all exams with excellent grades - excellent, 10. During his doctoral academic studies, as an author and co-author, he published 15 scientific and professional papers, two of which were published in international peer-reviewed journals.

Professor emeritus **Zoran Avramović** was elected a full member of the Russian Academy of Transport (RAT, St. Petersburg, Russia, since 1995), the Russian Academy of Natural Sciences (RAEN, Moscow, Russia, since 2001), the Serbian Academy of Engineering (IAS, Belgrade, since 2004). (formerly: Yugoslav Academy of Engineering - JINA) and the Academy of Electrical Engineering of the Russian Federation (RAEN, Moscow, Russia, since 2007). He is the scientific secretary of the Electrical Engineering Department of the Engineering Academy of Serbia. So far Avramović has published: 356 scientific and professional papers.

Accelerated Process of Digital Transformation - The Impact and Consequences of Covid-19

Mihajlo Travar[1], Igor Dugonjić[2], Saša Ristić[3]

[1]*Energy Regulatory Commission Republic of Srpska, BiH, mihajlotravar@gmail.com*
[2]*Pan-European University "Apeiron", BiH, igor.r.dugonjic@apeiron-edu.eu*
[3]*Deloitte d.o.o., BiH, sristic1992@gmail.com*

Abstract: Due to the current pandemic caused by the COVID-19 virus, the world is changing rapidly along with digital technologies that transform every aspect of life, society and the economy. To prevent a complete collapse and suspension of all business processes, companies were forced to organize remote work, i.e. workers perform their daily work activities from their homes. The situation in which the world is currently in clearly indicates that digital transformation is something that should be a priority. Digital transformation is changing the way of doing and developing the business, new opportunities for economic progress in the public and private sectors. It allows companies to survive and focus on innovation, increasing their competitiveness. We can say with certainty that digital transformation means much more than complete integration of digital technologies. It also means digitalization and business processes and models automation, marketing, sales, digital purchase, Big Data, and related processes, and is based on five different areas, which include customers, competition, value, innovation and data.

Keywords: Digital transformation, Information technology, Business process, Impact, Pandemic Covid-19.

Introduction

When the pandemic Covid-19 hit the Republic of Srpska and Bosnia and Herzegovina, the authorities faced conflicting obligations to protect public health, on the one hand, and respect for individual human rights, on the other hand. In order to slow down the spread of the virus, measures and restrictions have been introduced, which in some cases violate human rights. The reason could be that decision-makers did not have enough understanding to respond to the pandemic, but also the urgency of the measures themselves and their implementation, as in many cases they were adopted practically overnight. This paper discusses the impact of the response on the economic activity of citizens and the way everyday business takes place. During the research, the authors analysed already available data and content on respect for human rights, the impacts of measures and information about them, with a qualitative assessment of the impact of the measures on certain activities.

Digital transformation is a term that differs significantly from the term "digitalization", because it includes much more than that, and first of all significant changes in business models, in the way of living, affects our everyday habits and expectations regarding the quality of living. This, of course, greatly increases our dependence on the Internet and ICT (Information and Communication Technologies) solutions. The issue of cyber security and privacy, the degree of trust in technologies, the simplicity of using it by a wide population without technical prior knowledge, are some of the basic preconditions for providing mass acceptance. The situation in the world related to the Covid-19 pandemic gives us a significant argument for that, and it is especially important to maintain economic activity and reduce the damage that occurred.

Digital technology allows the world to continue to function during a pandemic, that certain processes can run smoothly, but that technology could today, as in the future, be a means of combating a

pandemic. First and foremost, in countries that are technological leaders in the world, artificial intelligence is used to find a cure and vaccine, which repeatedly shortens the time required to obtain results. In South Korea, artificial intelligence has been successfully used to find tests for the coronavirus in two weeks, while without the application of artificial intelligence, that work would last much longer. In addition, transforming conferences into digital online platforms, working from home, sending and receiving requests and responses for public administration, helps us keep jobs and processes in a state of emergency, especially to reduce the need for direct contact and thus stop further spread of the virus. It is clear to us that much of what has been applied so far will undoubtedly survive and remain even after the pandemic emergency passes. This situation has certainly affected many areas permanently, and the digital transformation of society will be accelerated in all directions.

Inevitably, as in other countries in the world, the pandemic Covid-19 negatively affected the economy of the Republic of Srpska and Bosnia and Herzegovina as a whole, which resulted in a decline in economic activity due to the economy closure. Consequently, pandemic Covid-19 significantly has imapcted businesses, economy and society, and accelerated companies digital transformation process. Therefore, it is important to explore opportunities and seek a way for companies to turn to mitigate the economic consequences of a pandemic, and at the same time adapt to the circumstances to be able to deal with it. Recognizing Covid-19 as a major driver of digital transformation of companies in Bosnia and Herzegovina, and as a process necessary to keep companies competitive in the market, in this paper authors discuss significance of digital transformation, as an essential move to economic recovery.

DIGITAL TRANSFORMATION - BASIC INFORMATION

The digital transformation began in the 1960s, when the development of programs and application solutions for automated data processing begins, which has had a significant impact on productivity growth and global economic development. Information technologies are no longer simple tools that serve to support the business but allow for a significant change in existing processes and the value chain. The application of digital technologies, transforms the way we work, changing the consciousness of employees, but nowadays it also plays a significant role in the process of changing the whole society. A whole range of different term definitions of Digital Transformation can be found in the literature, but almost each includes a couple of key components: digitalization, business transformation, collaboration or cooperation, modern digital technologies, value chain, and sometimes even the digital economy. Perhaps it is best to describe Digital Transformation as a set of all these elements. We can also say that digital transformation represents the networking of companies and customers, through all segments of the value-added chain, with the application of new technologies. As such, it requires skills that include extracting and exchanging data, as well as analysing them, and converting data into available information. This information should be used for decision making to initiate activities that increase the company's performance and market participation. Digital transformation and the transformation process involves companies, business models, processes, relationships, products, etc. For the success of digital transformation, it is necessary to change the awareness of employees and continuously implement their education and provide a large degree of IT literacy to be ready for the challenges posed. Companies need to invest and adopt completely new IT strategies that are managed by digital technology, as well as a competitive offering based on digital processes. The three basic steps define digital transformation within modern companies:

Digital work environment - application of technologies such as mobile devices, tablets and laptops; collaborative tools and social networks; application of technologies that are initially intended for private users.

Digital user experience - optimal user experience is a key factor in the digital transformation process. The negative experience of the user has a direct impact on the perception and sale of the product, which can be reflected in the company's business. Companies face the development of comprehensive, individualized systems that provide the user with optimization of experience in all digital and traditional business spheres. Although the way of displaying and designing such a system (front-end) plays the

most important role for the user (due to the interaction with the user), for optimal user experience it is necessary to perform the Digital Transformation of all processes in the company (logistics, accounting, storage and development). Therefore, in addition to the transformation of marketing and sales, Digital Transformation must include customer-oriented digitalization and integration of the entire processes in the company (front-end and back-end processes).

Digital business models and ecosystems - standalone companies will no longer be able to withstand the pressures of competition in the market (especially small and medium-sized enterprises), which will lead to the formation of digital ecosystems and service networks in which companies will connect and develop collaborative business models.

Digital Transformation and Its Significance for the Republic of Srpska and Bosnia and Herzegovina

In the Republic of Srpska and Bosnia and Herzegovina, the prevailing opinion is that digitalization is significant for business, but development and implementation of digital strategies are delayed. There is awareness in private sector that digitalization helps to achieve better future performance, however, companies confront obstacle which interfere with the process of transformation, such as lack of professional staff and technical knowledge to improve the digitalization process and funding. One of the main problems around the world, including Bosnia and Herzegovina, is that it is difficult for companies to translate digital transformation into business opportunities. The reasons are numerous for companies to access digital transformation, and given the dimension of the pandemic Covid-19, digital transformation is not an option for the companies of the Republic of Srpska and Bosnia and Herzegovina, but a matter of their survival. In previous years, Bosnia and Herzegovina has benefited from programs and funds of the EU and a number of international organizations, that support digital transformation. It is known that a faster transition to economic digitalization is a necessity and a step forward to economic and social integration, in preparation for BiH's membership in the EU. Digital Agenda for the Western Balkans is some of the main policy initiatives created at the digitalization policy level of the economy in BiH. In July 2017. the Trieste Summit was held, and since then the leaders of the Western Balkans officially recognized digital integration importance, as a key element of the plan to create the Regional Economic Area (REA).

In June 2018. the European Commission launched the Digital Agenda, one of the six main initiatives of the EU strategy for Western Balkans, to support the transformation of Western Balkan countries into a digital economy. It would bring faster economic growth, creating more jobs, improvement of the services, investments in digital infrastructure, digitalization of the industries, building capacity in digital security, strengthening the digital economy, development of the skills, e-administration development, e-purchase, roaming costs reduction (this was realized in June 2021. when roaming in the countries of the Western Balkans was abolished). The goal of this initiative is to support BiH and all Western Balkan countries towards successful digital transformation, and to be better prepared for the future in the EU. In October 2020, the European Commission adopted an Economic and Investment Plan for the Western Balkans, aiming to encourage a economic recovery, with an emphasis on the green and digital transition and development. Also, this plan encourages regional integration and rapprochement of the BiH and the Western Balkan countries to the EU. On the basis of the Digital agenda for the Western Balkans, this plan and investment offers an possibility to accelerate digitalization of the government administration, public services and businesses, in the line with EU values and framework. Throughout the pandemic customers turned dramatically to online channels and companies responded in the same way. BiH companies have done several changes in doing the business, and have been in an situation which demands reaction and change. Companies are looking for ways to improve productivity, offer the best possible services to their customers, facilitate way of communication, and approach to consumers with new sales channels. Among other things, companies should work to optimize work processes, and just simply to be innovative and provide competitive advantages, as the only option to remain vital in the business.

Many companies developed the capacity for
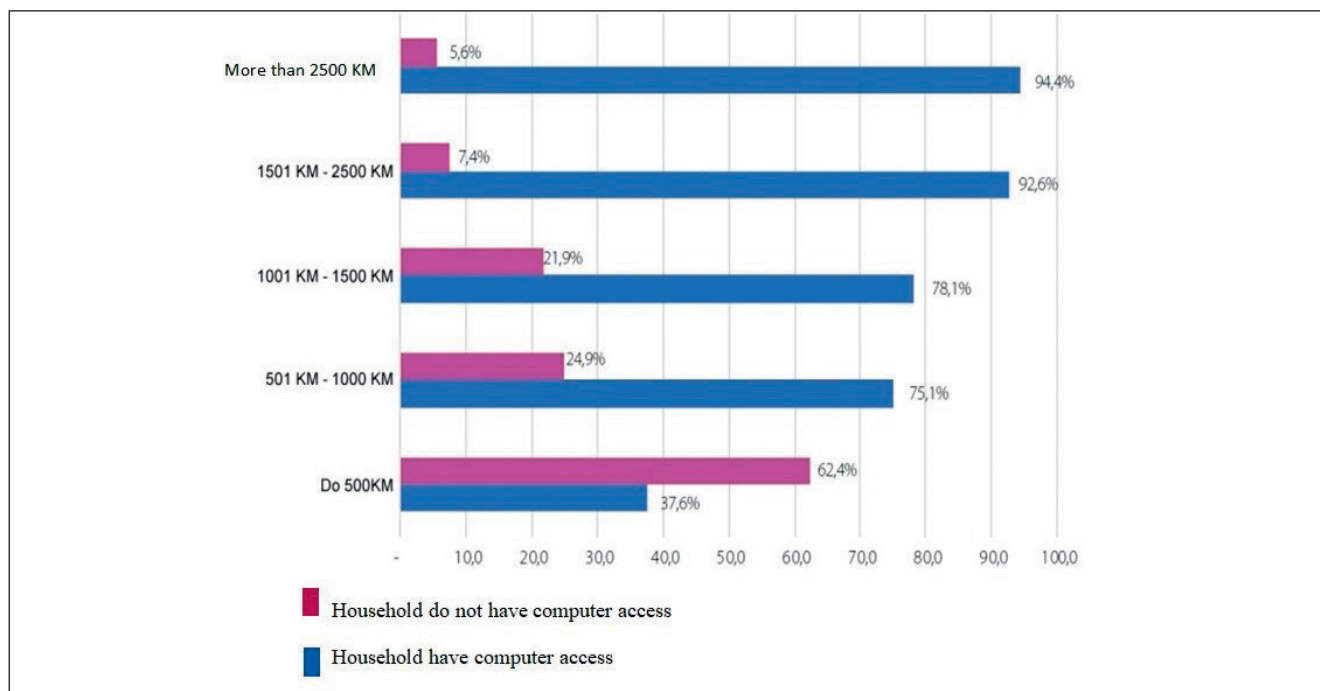
Figure 1. Percentage of households that have computer access, according to the monthly net, household income – BiH (Source: Agency for statistics of Bosnia and Herzegovina)

remote work; others created new sales channels through e-commerce platforms, while others changed business models, in an attempt to build value and competitive advantage. However, the Republic of Srpska/BiH were left behind in many of the digital transformation measures, such as using the broadband technology and other digital technologies, as well as the digital readiness of public and government administration. Altogether, including companies, government administration institutions, donors and other participants have to assist in overcoming this challenge through economic recovery of companies to create the best environment for business and increase the competitiveness of BiH companies. In order to support the economic recovery through digital transformation and mitigate the economic damage caused by the pandemic, the analysis carried out by the authors of this paper places emphasis on the following:

- Focus on administration digitalization and cutting red tape;
- Financing companies, in terms of digitalization of business processes and increasing business through e-commerce;
- Interinstitutional coordination and preparation of a concrete work plan related to the

Digital Agenda for the Western Balkans and the Economic and Investment Plan for the Western Balkans;
- Adjustment of all key participants, for the best possible usage digital transformation support schemes;
- Boosting current digital expertise in favour of digital transformation of the private sector and increasing expertise in digitalisation;
- Laws adoption to support the digitalization of public institutions and the private sector;
- Campaigning to promote digitalization and the potential it offers.

Insight into the report (USNews&WR, 2020) Bosnia and Herzegovina is not on the list of the most developed countries in the world, while the countries of the region: Croatia (44), Slovenia (56), and Serbia (72) are positioned on that list. Nevertheless, if an analysis of the report of the Agency for Statistics of Bosnia and Herzegovina on the use of Information and Communication Technologies in Bosnia and Herzegovina for 2020.is performed, it can be noticed that Bosnia and Herzegovina has a very high rate of application of digital technologies (possession and access to computers, i.e. information com-
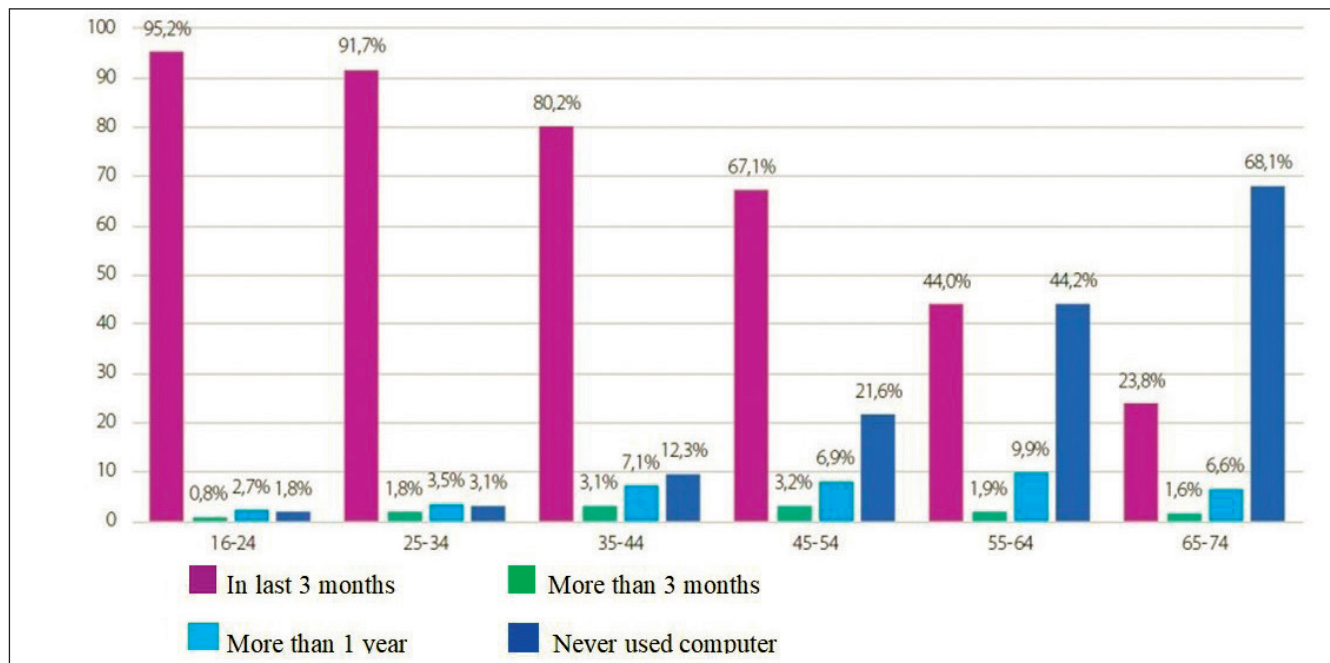
Figure 2. Percentage of computer users, by age – BiH (Source: Agency for statistics of Bosnia and Herzegovina)

munication technologies), although it is directly related to standard and monthly revenues of citizens (Figure 1.).

For the statistical image to be complete, it is necessary to look at the percentage of computer users (information communication technologies) by age. It can be observed that the application of digital technologies is present in a high percentage among the younger population and in a slightly smaller percentage in the middle age population, which is directly related to the entry into the digital age and the new industrial revolution (change of consciousness and culture of the social ecosystem).

The above analysis indisputably indicates a certain degree of digital literacy, but also the need for an agile change in the education system, and the introduction of new competencies in all areas of education that are in the new value chain, primarily engineering and economics. The recognized need to change and introduce digital competencies into the education system must be guided by good practices and recommendations by world organizations. UNESCO has defined the Competency Framework for Teachers (UNESCO 2018) in response to the needs of 2030. Sustainable Development Agenda, which recognizes the significant potential of information and communication technologies in the process of accelerating progress and increasing global

digital literacy. As potential challenges in this process, concepts have been identified: freely available educational resources, social networks, mobile technologies, internet stuff, artificial intelligence, virtual and augmented reality, processing large data sets, programming, ethics and privacy protection.

Joint Research Centre of the European Commission, defines the European Framework for the Development and Understanding of Digital Competence - DIGCOMP, which represents 21 competencies classified into five groups (Table 3.):

Information management - identification, locating, obtaining, storing, organizing and information analysis in digital form, where it is necessary to possess knowledge and skills to assess their relevance and purpose;

Communication skills in a digital environment - sharing resources through online tools; connecting with others and collaborating by digital tools; communicating and participating through communities and networks; building intercultural consciousness.

Creating digital content - creating and editing new content (word-processing, images and videos); integrating and refining previously acquired knowledge and content; making creative expressions, media elements and software programs; applying intellectual property rights and licenses.

Security in the digital environment – private and

data protection; digital identity protection; security measures implementation; safe and sustainable usage.

Problem solving - identifying digital demands and resources; decisions making about to choose and apply the most suitable digital tools according to needs or purpose; conceptual problems solving by digital resources; creative usage of technologies; solving technical problems; developing one's competencies and competencies of others.

Table 3. *DIGCOMP – competencies (Source: https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework)*

Dimension 1	Dimension 2	Dimension 3
5. Area	21 Competencies	Competency level
AREA	COMPETENCIES	
Information	1.1 Review, search and filter information	
	1.2 Information evaluation	
	1.3 Information and digital content management	
Communication	2.1 Digital technologies interaction	
	2.2 Digital technologies content sharing	
	2.3 Digital citizenship inclusion	
	2.4 Digital technologies collaboration	
	2.5 Correct and acceptable mode of communication	
	2.6 Digital identity management	
Content creating	3.1 Digital content development	
	3.2 Digital content integration and processing	
	3.3 Licensing and copyright	
	3.4 Programing	
Security	4.1 Device protection	
	4.2 Private data protection	
	4.3 Health care and wellness protection	
	4.4 Environmental protection	
Problem solving	5.1 Technical problems solving	
	5.2 Determining the needs and responsibilities of tech.	
	5.3 Digital technologies creative application	
	5.4 Determining digital companies deficiencies	

CONCLUSION

The pandemic has harmed both, consumers and businesses, changing the way people work, consume and spend free time. The pandemic Covid-19 has brought business and society to the necessity to adapt the new technologies. Regardless of the companies desire, digitalization has became necessary to facilitate companies, depending on measures and circumstances, to make changes, as: remote work, create channels to shop online, apply technology in everyday work, adjuste to consumer demand and expectations, identify suitable channels for communication and distribution, and other changes that have resulted from the need to adapt to the emerging situation.

Generally, digital technologies and digitalization are yet at the centre of the debate, due to their use during the outbreak of the pandemic Covid-19. In normal circumstances focus of the company is to increase revenues and market shares. In this pandemic situation, BiH companies are struggling in controlling the costs, maintaining their liquidity and survival. Pandemic Covid-19 continues to bring a crisis to public health, and make a severe economic consequence. After prevailing the pandemic Covid-19, companies have to evaluate the impact of these changes, and further to adjust the way to design, communicate and build value for the consumer.

The Digital transformation has helped many BiH companies to survive the effects of the the pandemic. This is related to the companies that have already started their digitalization before pandemic have started. Companies, that have not been prepared from the digitalization point of view, and did not have any action plan or strategy before the pandemic, responded without thinking to the process of digital transformation. It is clear now that it is necessary to adapt existing business models and processes, as well as regulations, to enable measurable and efficient work from home, i.e. remote work, which takes place on the digital platforms and which have all functionality that the company management needs.

By introducing digital technologies into daily work, the way of thinking and problem-solving approach of employees in production chains is changing, as well as the way users access new services. The question of the set of competencies that are necessary for the full integration of people and the working environment into the digital sphere arises, which leads to a radical change in the education system and gaining the new skills and knowledge needed to fully enjoy the services offered by digital technologies. Although the possibilities of working from home have been an available option since before, just the current pandemic has enabled such a system of work and has forced a large number of people to increasingly rely on digital technologies.

Literature:

[1] Abolhassan, F. (ed.), "The Drivers of Digital Transformation: Why There's No Way Around the Cloud", Springer International Publishing Switzerland, 2017.

[2] D., Korajčević, Š., "Use of information communication technologies in Bosnia and Herzegovina", Agency for Statistics of Bosnia and Herzegovina, 2020.

[3] Gigova, T., Valeva, K., Nikolova-Alexieva, V., "Digital Transformation – Opportunity for Industrial Growth", International Conference on Creative Business for Smart and Sustainable Growth (CREBUS), Sandanski, Bulgaria, pp. 1-4, 2019.

[4] Nicoletti, B., "Agile Procurement - Volume II: Designing and Implementing a Digital Transformation", Palgrave Macmillan, 2017.

[5] Schallmo, D. R. A., Williams, C. A., "Digital Transformation Now! - Guiding the Successful Digitalization of Your Business Model", Springer, 2018.

[6] Singh, S. K., Cha, J., Kim, T. W. and Park, J. H., "Machine Learning Based Distributed Big Data Analysis Framework for Next Generation Web in IOT." Computer Science and Information Systems 18(2):597–618, 2021.

[7] Tobji, M. A. B., Jallouli, R., Koubaa, Y. Nijholt. A. (ed.)., "Digital Economy: Emerging Technologies and Business Innovation", Springer Nature Switzerland, 2018.

[8] Travar, M., Dugonjić, I. and Ristić, S., "Analysis of Using Cloud Business in Bosnia and Herzegovina and the Region". JITA - Journal of Information Technology and Applications 9(2):118- 125, 2019.

[9] Travar, M., Šušić, I. and Ristić, S., "Importance of information technologies and ERP solutions for the business of the service sector of Republic of Srpska / BiH in the conditions of the pandemic Covid-19", XX International symposium INFOTEH – JAHORINA 20(8):23-27, 2021.

[10] Ustundag, A., Cevikcan, E. (ed.), "Industry 4.0: Managing The Digital Transformation", Springer International Publishing Switzerland, 2018.

[11] Agency for Statistics of Bosnia and Herzegovina, "Education" Last modified 31.10.2020.http://bhas.gov.ba/Calendar/Category/15

[12] United Nations, "World Population Prospects 2019", Department of Economic and Social Affairs Population Division, United Nations New York, 2019.

[13] https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework

About the authors

Mihajlo Travar earned his PhD at the Faculty of Mechanical Engineering, University of Belgrade. He is a member of Regulatory Commission for Energy of Republic of Srpska. Mr. Travar is Associate Professor at the 'University of Business Studies' in Banja Luka, where he gives lectures on the following subjects: Databases, Software Engineering, CASE Tools, Design Engineering and ERP Systems. He has written more than forty scientific papers in ICT, mechanical engineering and business organization.

Igor Dugonjić earned his Master's degree in computer science at the Faculty of Electrical Engineering, University of Banja Luka. He is doing his PhD at Pan-European University 'APEIRON' in Banja Luka. Mr. Dugonjić works as a medical equipment programming and maintenance engineer at the University Clinical Centre of The Republic of Srpska as well as a senior teaching assistant at the Pan-European University "APEIRON". He has written several scientific papers on medical ICT research.

Saša Ristić graduated on Faculty of Economics in Banja Luka. Currently he is on master studies on University of Banja Luka, department Accounting and audition. For several years he has been working at the largest IT Company in region on software solutions in SaaS distribution model. Besides that, he lectures on the seminars of Association of Accountants and Auditors of RS on the subject Information technology. He has written a few scientific works about ERP systems, electrical administration and IT

Architecture of GIS Solutions for Detection and Development of Wildfire Database

Saša Ljubojević[1], Zoran Ž. Avramović[2]

[1]PFE ˝Forests of the Republic of Srpska" JSC Sokolac, Banja Luka, sasa.ljubojevic@sumers.org
[2]Pan-European University Apeiron, Banja Luka, zoran.z.avramovic@apeiron-edu.eu

Abstract: This research paper presents organization of the business environment for work with geographic information systems (GIS) which are based on open source. The solution is completely open source: operating system, working environment and supporting apps. The architecture consists of: server, workstations, mobile devices and sensors. Software packages for each architecture segment will be displayed. The goal is to achieve a complete business environment for work with open source GIS, thus minimizing the costs of system development and maintenance. The illustrated example shows the possibility of applying GIS within a forestry company, in the field of wildfire monitoring and data collection and registering the possibility of wildfire occurrence using IoT.

Keywords: GIS, open source, IoT, wildfires, wildfire detection.

Introduction

Open source software is being increasingly used in business environments, primarily due to its cost as well as cost of licenses. As the software itself is not license limited, it can be therefore installed and used on an unlimited number of workstations. Open source software is available free of charge, unlike commercial software, which are sold under strictly defined licensing terms. Also, open source software is distributed with a license, and the most commonly used license is the General Public License (GPL). The GPL license defines that we can use the program for any purpose [1]. Nowadays, open source software does not lag behind commercial software. In some cases, it is even more developed than commercial software, and has wider field of application. The advantages of commercial software are reflected in its user friendliness and its adaptation to the needs of end users. Due to this reason, we come to the question of whether it is smarter to invest in software or into education of users.

This work presents the GIS architecture, established as open source software, which is needed for the complete functioning of a business organization. The architecture consists of server and client system elements. Establishment of such system provides management of spatial data within various business spheres, both for those which use spatial data as primary inputs, but also for those which use different inputs, and all can use open source software without spending financial resources for software licensing. Concrete example of the application of GIS is from the field of forestry, which refers to the establishment and maintenance of wildfire records at the territory of Republic of Srpska.

According to the Law on Forests of the Republic of Srpska [2], the beneficiary of forests and forest land owned by the Republic of Srpska is the Public Forest Enterprise "Forests of the Republic of Srpska" JSC Sokolac. The primary activity of the Public Company is the cultivation and protection of forests. Forest protection also includes protection against wildfires, and the obligation of the Public Company, in addition to the activities regarding fire protection, is to record and archive data on wildfires. The aim of this part of the research paper is to show the possibility of improving the current system of record-

ing wildfires using GIS, on the example of this Public Forest Enterprise. In addition to the records of recent wildfires, the procedure of collecting and entering historical data on wildfires is also important.

GIS DEVELOPMENT METHODOLOGY

The goal of this research paper is to present the development of GIS which is used in a business environment, and which is based on open source. This specific system was created by applying phase development, i.e. by applying incremental build model development. The structure of the GIS architecture is described in detail, which consists of servers, workstations, mobile devices and sensors for data collection including supporting software. Using the analysis methods, each individual element of the architecture will be analyzed and described. Using the synthesis methods, all elements of the architecture will be combined into one functional unit that enables work in a business environment.

SYSTEM ARCHITECTURE

This system consists of four primary elements: *server* which is the central unit for all data contained in the GIS and which enables a large number of users to mutually work and exchange spatial data; *client element* which enables the client workstation to access the central geo-base through appropriate software; *mobile application* for field data collection which can be used for data editing and data storing at client stations and also for data saving at central storage location, as well as the use of IoT for auto-

matic data collection using sensors which are delivered to a central storage location for further data processing.

Picture 1 shows architecture of above mentioned system including separate graphic elements with supporting links. Further, each software package shall be separately described of which the architecture is composed.

Table 1: Software packages and its components

Server	Workstations	Mobile device	IoT
Ubuntu Server	Ubuntu Desktop	Android	Arduino
PostgresSQL/PostGIS	QGIS	QField	Sensor
GeoServer			

Server

The operating system used for the server architecture is the **Ubuntu server** developed by Canonical, which was founded in 2004, and the current version of the server is 20.04 LTS, which has guaranteed updates and upgrades until 2032 [3]. It is a Linux operating system, based on the Debian distribution. A Server is a device that provides services to other software or devices called clients. This type of server architecture is called the client-server model, and this concept is based on distributed data processing whereas the programs and data are located in the most efficient places, so such systems are usually located in the local network [4].

Geo-base data used in this system is **PostgreSQL**. It is free and open source and is published under the PostgreSQL license [5]. It is visible on its official website that there are no plans to change the licensing conditions of the mentioned software, so it is to be expected that it will remain free forever. PostgreSQL, also known as Postgres, is a relational database system that uses the SQL language. The aforementioned software comes with many integrated features whose primary goal is to enable the creation of applications, protection and data integrity that help users manage their data regardless of their quantity. PostgreSQL is dated September 14th 2021 and is harmonized with at least 170 of the 179 mandatory functions with SQL: 2016 Core standard. It is important to be emphasized that no relational database is fully compliant with this standard [6].

The **PostGIS** software package is also open source, which provides support for spatial data management
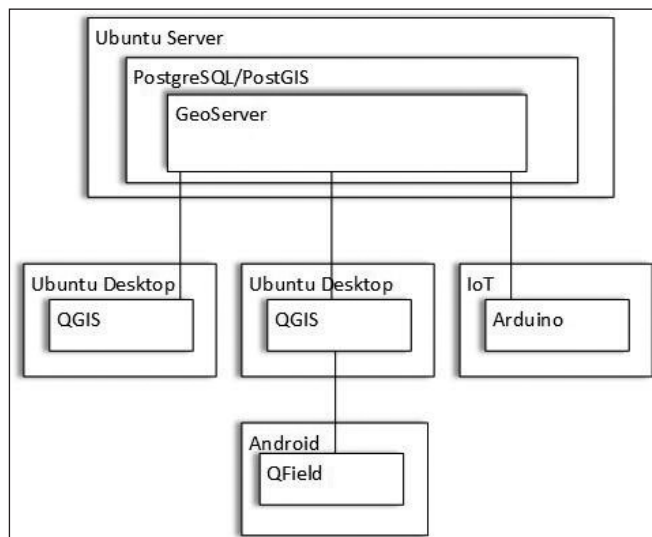
Picture 1: GIS architecture

in the PostgrSQL database. This extension allows you to run SQL queries over spatial data [7].

GeoServer is an open source server based on the Java programming language, which enables users to view and edit spatial data. This software package uses the Open Geospatial Consortium (OGC) standard, which allows great flexibility in creating maps and exchanging data with other systems. GeoServer has the ability to work according to the standards of Web Feature Service (WFS) and Web Coverage Services (WCS), which enables data sharing and data processing [8]. In addition to the above mentioned, it also includes certified high performance Web Map Service (WMS). With the help of these services, it is possible to embed data into other systems and websites. GeoServer is a fundamental component of the business GIS environment.

Workstations

After installation of server, which provides storage and distribution of data for further processing and editing by users, next step are client workstations. These workstations also use open source software and consist of an operating system, GIS and supporting software.

The workstation operating system is **Ubuntu Desktop** which comes from the same manufacturer as Ubuntu Server, which has been at the very top of popularity on DistroWatch [9] in recent years. Ubuntu desktop is based on Debian, which is created as a Linux desktop which is user-friendly [3].

After initial preparation of the operating system, it is necessary to install GIS on the workstation, and in this case it is **QGIS**. QGIS is one of the most popu-

lar GIS software solutions for open source desktops. With QGIS it is possible to create, modify, analyze and visualize spatial data. This software package supports numerous vector file types, raster file formats, as well as various databases [10].

In addition to a large number of built-in features, QGIS can be upgraded by a large number of add-ons that are available in the integrated repository. In addition to the operating system and GIS software, other supporting and needed software such as SAGA, GRASS and LibreOffice can be installed on workstations and thus have all applications needed for working in a business environment.

Mobile device

In order for GIS in a business environment to be completely functional, it is needed to enable data collection via mobile devices. The most important issue for all systems dealing with spatial data processing is field data collection. There are numerous open source software to perform this task. In our area, the most popular operating system for mobile devices is Linux based **Android**, which is open source, so the specified operating system for mobile devices will be the default and the system architecture will be built on it.

One of the most popular apps for fieldwork and open source data collection is **QField**. QField enables efficient work with outdoor spatial data, which is based on QGIS. Field work is not similar to the office work, because the work is executed on a smaller screen and the tasks are different. The primary goal of the QField application is to enable the user to perform field tasks within minimal working conditions. This means that only those functions that are essential for performing tasks in the field are available to the user on home screen [11].

IoT Sensor

IoTs represent the next generation of IT devices. IoT represents the use of different sensors and devices with unique identifiers that can send information about objects and people through the Internet, without requiring human-human or human-machine interaction, or is a communication network of physical objects with a unique IP address [12]. The purpose of the Internet of Things is to continuously monitor, detect and eliminate any possible irregu-
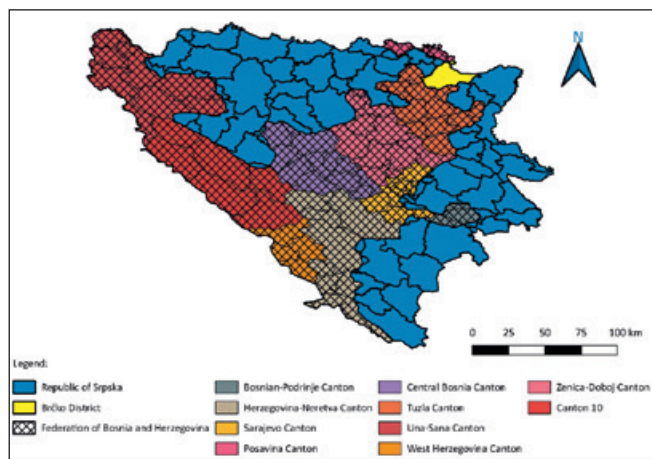
Picture 2. *Map created by QGIS*

larity and enable an alternative mode of operation of any process [4].

In this example, an Arduino microcontroller was used, which is a suitable device for prototyping general-purpose IoT devices. Arduino is an open source microcontroller platform for designing projects from the field of electronics. Its open source basis and its user friendliness have made the Arduino a functional choice for those looking to develop a variety of electronic projects [13].

Application of GIS System

Geographic information system can be applied in various areas of business, and in this example a review will be given regarding wildfire data collection. The need for such system exists in forestry companies. The application of GIS in the field of forestry is obvious, especially if we take into account that most of the business in forestry is related to the specific location, i.e. forest, regardless of activity type (forest inventory, forest restoration, forest protection, forest exploitation or any other related activity).

After implementation the GIS system, it is necessary to obtain a basic set of geo-spatial data. The data that represent the set of basic needed graphic images are: topographic map; digital terrain model (DEM); map of forests and forest land; map of forest management areas; road infrastructure map; watercourse identification map; administrative boundaries; map of mined and mine suspected areas; meteorological weather stations data (temperature and precipitation data) and similar. There is no limit of graphic images and survey data imported to GIS and all graphic images and survey data that can be obtained or designed are useful by creating new possibilities for future analyses.

After the completion of the preparatory phase related to the obtainment of graphic images and survey data, it is necessary to start developing GIS. Special attention in the process of developing GIS project should be paid to the selection of the reference coordinate system. For the purposes of this project, the Gauss-Krüger coordinate system (projection with 6 degrees wide zones) was used. In addition to the above, it is very important to define set of certain rules related to the organization of the GIS project, primarily for its easier handling afterwards, and especially regarding the possibility of future up-

grades or integration of this GIS system into different system.

Each GIS project, including this one, is organized through sets of different data that are displayed through layers, in order to separate different categories of data through different layers. All data contained in a GIS include spatial component and associated attribute data. The spatial component is visually displayed on one of the layers, while the attribute data are located in the background data geobase. Spatial data can be a point feature, a line data, and a spatial polygon data.

Regarding the database and its content, it is necessary to clearly define the organizational structure of the Public Forest Company with the corresponding codebook. The organizational structure should be presented through tables: forestry management area, forestry authorities, business unit, department, branch office, operational unit and forestry administration.

In addition to the above tables, another table should be created regarding wildfire records, containing spatial data related to wildfires, as well as other attribute data.

Previous records on wildfires were kept in prescribed forms, which is prescribed by the Ministry of Agriculture, Forestry and Water Management of the Republic of Srpska. In most cases, the lowest level of record keeping regarding the occurrence of wildfires is kept at the departmental level. This means that there is no exact geographical location of the wildfire determined, but only the indication of the micro-area where the fire had occurred. However, this is not the case when the records are not adequately kept or when a wildfire has affected a larger area than single forestry management area. In such case, only forestry department or numerous departments are entered in the records, but even then there is no information of the exact geographical wildfire location. Besides to the listed data on wildfire location, the document/form also contains other groups of data: type of fire and area affected by the fire; type of plants affected and type of burned forest stand; wildfire duration; damage caused per plant type and amount of damage, damage value; causes of fires and key factors in extinguishing fires.

All mentioned groups of attribute data contain fields that more clearly label and define wildfires.

For the purposes of this project, it is needed to collect historical data on wildfires, which are recorded in the prescribed forms by each forestry department in written form. Each form refers to single forestry department and single calendar year. These data should then be entered into the database and according to data on the wildfire location (forestry management department, business unit, department and branch office) should then be linked with associated layers, i.e. graphic images and survey data in the GIS.

Public Forest Company "Forests of the Republic of Srpska" JSC Sokolac has 32 organizational units.
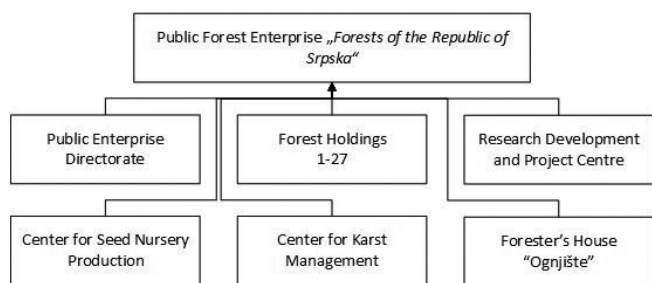
Picture 3. Organization of Public Forest Company "Forests of the Republic of Srpska" JSC Sokolac

Out of 32 organizational units, 28 organizational units keep records on wildfires in the prescribed forms (all Forestry Management Departments and the Karst Landscape Management Center). For the purpose of this research paper, historical data were collected, i.e. records on wildfires at the area governed by Forestry Managament Department "Gradiška" with its headquarters in city of Gradiška - beneficiary of forestry-economic department called
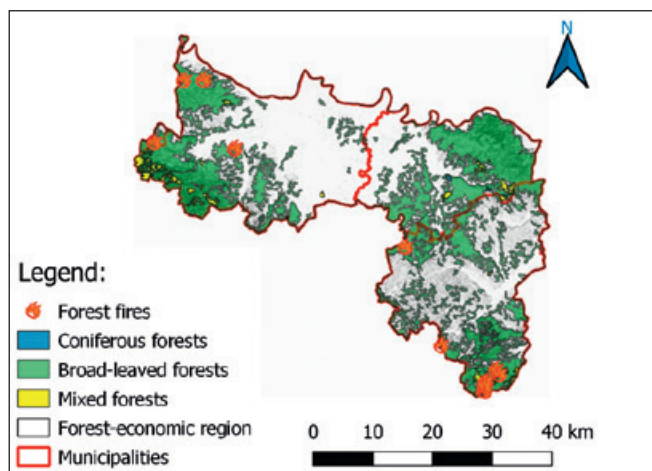
Picture 4. Overview of wildfire locations at Forestry Economic Department "Posavsko" for the period 2013-2020

"Posavsko". Historical data for the current subject period (ten years), starting from 2013, were collected. All collected data were in written form, and in accordance to data on economic units, departments and branch offices, each data was entered into map governed by forestry - economic department "Posavsko" in the form of points (centroids) as of the end of 2020.

Picture 4 shows the spatial structure of forests and forest land located within Forestry Economic Department "Posavsko" and occurrence of wildfires is presented within graphical image in the form of points. In addition to the map display, in accordance to the data entered into the spatial data geo-base, various analyses related to wildfires can be performed. In this example, digitized data for one organizational section of the Public Enterprise are presented, and data digitization for other organizational units is in progress. With the completion of the activities of digitization of historical data for the entire area of Republic of Srpska, beneficiary PFE "Forests of the Republic of Srpska" JSC Sokolac shall be in position to create future electronic archival records on wildfires. Some of the advantages of the use of GIS regarding wildfire records are digitization and centralization of databases, as well as numerous analyses that can be performed using imputed data.

Internet of Things and GIS

In addition to the records of wildfires, this research paper shall also show how to use sensors to collect data that may warn and indicate the potential occurrence of wildfires. For the purposes of this work, a prototype based on Arduino was developed, which contains a communication module and a temperature and humidity sensor DHT11.

Above mentioned DHT11 sensor is a simple temperature and humidity sensor, which is connected to the Arduino with just one pin bringing back two values: temperature and humidity. This sensor does not require calibration or additional time to be started. By connecting the sensor to the Arduino and by loading its software library, the sensor starts working. The sensor is simple, small in size and as such meets the minimum hardware needs for use within remote detection system. Once a communication channel is established with Arduino, which collects

temperature and humidity data, it can be used as an IoT data collection device.

The use of this type of sensor faces several challenges related to the power supply of the device. Since these devices are to be installed in a forest where the availability of electricity is limited, it is necessary to use some of the alternative power sources. In this particular case, an external battery was used, and it is also possible to use a solar power panel at appropriate clear locations, because in a forest it is difficult to find abundant sunlight location.
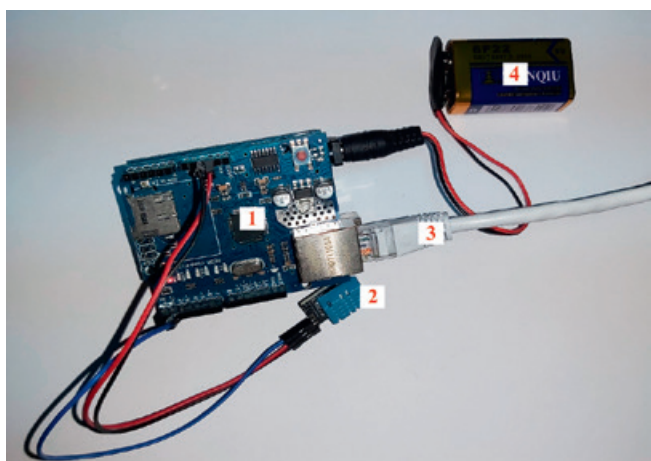
Picture 5. Prototype of IoT sensor (1-Arduino; 2-sensor; 3-UTP cable; 4-battery)

Also, one of the challenges is sending data to a central location, as the position of the sensor is primarily in remote forest areas, which, in most cases, are not covered by cellular network. Optimal method of sending data is via 3G or 4G cellular network and in the initial stages of system establishment it is important to select areas where communication channel can be established with high quality available signal to collect data and system to be tested. Since the tested area was not far away from urban environment, the prototype was connected with an RJ45 network cable. It should be noted that this type of connection should not be ruled out, because there are many forest field facilities that are suitable locations for the installation of sensors. Also, the sensor has the capability to store data on memory card, but in such case it loses its primary function of remote detection. This option only makes sense in collecting data and needed parameters for historical data analysis.

After loading and initial running of the program in Arduino, the system starts to deliver data to the geo-spatial database. The device should be placed at the intended location, and as the data are being recorded, the exact date and time of its reading is generated for each record.

ID bigint	Name text	Humidity double precision	Temperature double precision	Time timestamp without time zone
71	S1	45	28.9	2021-08-22 13:08:12
72	S1	45	28.9	2021-08-22 13:08:17
73	S1	46	29	2021-08-22 13:08:22
74	S1	46	29	2021-08-22 13:08:27
75	S1	46	29	2021-08-22 13:08:32
76	S1	46	29	2021-08-22 13:08:37
77	S1	46	29.1	2021-08-22 13:08:42
78	S1	46	29.1	2021-08-22 13:08:47
79	S1	46	29.1	2021-08-22 13:08:52
80	S1	46	29.1	2021-08-22 13:08:57
81	S1	46	29.2	2021-08-22 13:09:02
82	S1	46	29.2	2021-08-22 13:09:07
83	S1	46	29.2	2021-08-22 13:09:12
84	S1	46	29.2	2021-08-22 13:09:17

Picture 6. Image showing database recording (PostgreSQL)

Accordingly, a prototype device was developed for remote collection of data regarding temperature and humidity in a certain area. By installing a large number of IoT devices for remote monitoring of temperature and humidity, a clear overview of climatic conditions can be obtained. Detection of a reading that clearly shows a change of parameters in a certain area, in terms of a sudden increase in temperature including decrease in humidity in specific region, is a warning indication of the wildfire occurrence.

It is obvious that, by collecting a large amount of field data, the system can be perfected to the level to ignore natural phenomena and anomalies, and to be able to clearly separate and detect the changes that more precisely indicate the occurrence of wildfire. At the very beginning following installation of such system, a larger number of false alarms would occur due to various weather conditions and before clarification of received values regarding changes in temperature and humidity that indicate the occurrence of wildfire. Besides, field data regarding climate changes would be collected making it possible to mobilize the necessary manpower and machinery to a potentially specific endangered and critical forest area. It is a well-known fact that climatic characteristics and wildfire occurrence are associated, which is reflected by a larger number of wildfires during extreme hot and drought years. Therefore,

the temperature parameter was used as the most important data that characterizes the possibility of wildfire occurrence. For the purpose of this prototype, temperature of 45°C was defined as threshold value. This value was selected in accordance to the maximum recorded temperature at the territory of Republic of Srpska, which is 41°C [14].

This example shows case of one used sensor, which monitored changes during time interval of 5 seconds. This is probably an excessively demanding short interval, due to the needed time for weather condition to be changed and the amount of data that needs to be transferred to the central database. Table with collected data from the IoT sensor shows designed trigger which is activated once the temperature exceeded a defined value, in this case the defined 45°C.

Organization of the database should be established by setting an activation trigger for each new input data which analyzes the value of the new data in relation to the previous or several previous values checking the difference if any. This difference would be more precise after the transition period, following collection of large amount of data needed for defining and elimination of false alarms. Accordingly, the system becomes fully functional and applicable for practical use. In accordance to the data received from Public Forest Enterprise, we can select the period with the largest number of wildfires including wildfire incident map and in accordance to these data we can test the system in real life conditions, which may ultimately indicate all possible shortcomings and possibilities to improve the system.

Finally, it should be emphasized that this system is a prototype, and potential problems related to power supply and communication with the central location may arise due to its robustness, and prior to putting into operation of IoT sensor it is necessary to provide these resources.

CONCLUSION

This research paper presents the work method and type of software packages needed to create a business environment based on open source software. The architecture of the system is described, which consists of servers, workstations, mobile devices and sensors with supporting software whose central part is a group of open source software

packages: GeoServer, PostgreSQL/PostGIS, QGIS and QField. Using the listed software packages, it is possible to perform all the tasks that are expected from particular business system. A huge advantage of such system is that it is being distributed under an open source license which is free having no significant restrictions. In addition to the benefits reflected in reduced costs, there are many other benefits of using open source software, such as less time and less bureaucracy associated with purchasing licensed software. Although this licensed software is free, as in any other implementation there are costs related to training, implementation and development time for solution customization [15].

The second part of the research paper describes the use of open source GIS in the business environment of Public Forest Enterprise "Forests of the Republic of Srpska" JSC Sokolac, regarding wildfire database. Wildfire records were kept through manually filled prescribed forms at the level of single Forestry Management Department during single calendar year. It is not possible to perform any serious analysis in accordance to such database records. Provided solution enables digitization and centralization of the mentioned records and the overall result is the possibility of performance of various analyses using other data contained within the GIS. Some of the possible analyses are: periods of wildfire occurrence, wildfire altitude, distance from road infrastructure, average temperature and humidity during wildfires and many other analyses.

The sets of data are listed, i.e. layers required for such system, which can be used in the future for various activities of the Public Forest Enterprise. Historical wildfire occurrence data were collected and entered into the system for the Forestry Economic Department "Posavsko" for the period from 2013 -2020. The procedure of collecting historical data is presented, and activities on recording data for other Forestry Economic Departments are in progress. In addition to historical data, applicable method of real-time field data collecting through IoT is presented which therefore enables timely remote sensing of wildfire by processing these data. This research paper describes a prototype design based on an Arduino microcontroller, which meets the minimum operation requirements. It should be emphasized that the above prototype design can

perform required tasks, but may show difficulties in terms of power supply and communication with the central unit, and it is necessary in the future period to find a better solution for these shortcomings.

This research paper showed the possibility of using geographic information systems in an open source business environment, which can meet the needs of a large company such as the Public Forest Company "Forests of the Republic of Srpska" JSC Sokolac. Special attention is paid to wildfire record keeping and the presented solution design is capable to respond to a much larger number of requests not lagging behind commercial software of this type.

References

[1] Valade, Spring into Linux, Addison-Wesley/Pearson Education, 2005.
[2] Official Gazette of the Republic of Srpska, Forest Law. Banja Luka: Official Gazette of the Republic of Srpska, 75/08, page 17, 2008.
[3] Canonical Ltd. [Online]. Available at: https://ubuntu.com, [visited 10.20.2021].
[4] A. Langer, Analysis and Design of Next-Generation Software Architectures, Springer International Publishing, 2020.
[5] The PostgreSQL Global Development Group. [Online]. Available at: https://www.postgresql.org/about/licence, [visited 10.20.2021].
[6] The PostgreSQL Global Development Group. [Online]. Available at: https://www.postgresql.org/about, [visited 10.20.2021].
[7] Planet PostGIS. [Online]. Available at: https://postgis.net, [visited 10.20.2021].
[8] GeoServer. [Online]. Available at: http://geoserver.org/about, [visited 10.21.2021].
[9] DistroWatch. [Online]. Available at: https://distrowatch.com, [visited 10.20.2021].
[10] The Open Source Geospatial Foundation [Online]. Available at: https://www.osgeo.org/projects/qgis, [visited 22.10.2021].
[11] Opengis.ch. [Online]. Available at: https://qfield.org/docs/index.html, [visited 10.22.2021].
[12] Mr. S. Stojković, Collection and integration of spatial data, Belgrade: University of Belgrade – Faculty of Geography, 2020.
[13] S. Monk, Arduino: Programming introduction, Belgrade: Mikro Knjiga, 2017.
[14] Republic of Srpska Institute of Statistics. (2020). Statistical almanac of the Republic of Srpska (second, edited volume). [Online]. Available at: https://www2.rzs.rs.ba/static/uploads/bilteni/godisnjak/2020/StatistickiGodisnjak_2020_WEB_II.pdf, [visited 10.24.2021].
[15] (2020, Jun) PSD Citywide. [Online]. available: https://psdcitywide.com/articles/guide-open-source-gis-software-for-the-public-sector, [visited 10.24.2021].

About the authors

Saša Ljubojević, is a head of IT department in PFE Forestry of the Republic of Srpska JSC Sokolac. Master of Computer Science in the field of information systems (2019) at Pan-European University "Apeiron" Banja Luka. Currently attending doctoral studies at the Pan-European University "Apeiron" in the field of geographic information systems. In addition, I am Project Manager in the integration information system implementation in the PFE Forestry of the Republic of Srpska JSC Sokolac. Experienced System Administrator with a demonstrated history of working in the computer software industry.

Zoran Avramović born in Serbia. He graduated from the School of Electrical Engineering of the University of Belgrade. At this School he received a degree of a master, and then a PhD of technical sciences. PhD of Technical Sciences 1988. He is the professor of the Faculty of Transport and Traffic Engineering on the University of Belgrade (since 1989), professor of the School of Electrical Engineering on the State University of Belgrade (since 1992), professor on the Pan-European University "APEIRON", Banja Luka, Republic Srpska, Bosnia and Herzegovina (since 2007) and professor, Faculty of Transport and Communications, Berane, Montenegro (since 2009).

For citation

APPLICATION OF INFORMATION TECHNOLOGIES IN NEW FORMS OF TEACHING PROCESSES

Slavojka Lazić, Tijana Talić

Pan-European University APEIRON, Banja Luka, Republika Srpska, BiH

(slavojka.n.lazic, tijana.z.talic)@apeiron-edu.eu

Abstract: Educating young people is one of the most beautiful and humane vocations. The transfer of knowledge to young people and their introduction into the world of science requires a well-prepared and organized educator. The application of information technologies in education has become an everyday tool, so that its role in the educational process has come to the fore during the last two years. The Covid 19 pandemic brought new challenges to the education system in Republika Srpska. The best solutions for the teaching process were sought. At the beginning, the classes were conducted at a distance, last year in classrooms with classes shortened to 20 minutes, and this year the classes again last 45 minutes, with respect to protection measures. The paper will show how the students coped with all these changes and how much their knowledge of information technology helped them in all this. The research includes an analysis of data collected by a survey of high school students and refers to their attitudes towards the performance of the teaching process in the past few years.

Keywords: education, teaching process, IT.

INTRODUCTION

In just the last few years, the educational process in Republika Srpska has undergone changes that have not been recorded for decades. During the last and the year before last, education has undergone unexpected and sudden changes [1]. Education faced new challenges, which needed to be addressed as soon as possible.

The education of young people did not subside, it continued to take place, only in a different environment [2].

The regular teaching process has always involved meeting teachers and students in one place, at school. But almost half of the school year before last, students from all primary and secondary schools in Republika Srpska learned at a distance. The Covid 19 pandemic seems to have come to change both human mobility, habits and consciousness, and the foundation of society - education. The decisions and measures taken after that, before the start of the pandemic, were unthinkable. Last school year, school classes were more than halved and lasted only 20 minutes. In the new situation, both students and teachers managed. Who did better?

This school year began with a duration of 45 minutes, with respect for measures to protect everyone in the educational process [3]. According to the research, which was conducted among the students of the third grade of the secondary technical school, the students believe that the class lasts too long. It does not hold their attention, their concentration decreases, they think that they do not need so much time to acquire new knowledge - these are just some of the answers of students. Is it something else after all?

Are our students losing their work habits? Have the previous two school years, realized in difficult conditions, with shortened classes or in the form of online teaching, already introduced disorders in education that we will become aware of only later. Perhaps completely new norms of education are arriving: the duration of classes, the form of work, the method of teaching. We are perhaps in the greatest transition of the educational process ever.

Adapting Education to Sudden Changes

The education system of a country shows its current state and society's attitude towards the young generation. Organizing and performing the educational process is a very complex and demanding job [4]. It is the foundation of creating a new society, leadershiping and guiding young people towards new changes and challenges. The process of education is constantly changing and improving [5]. Teachers know that the ways in which students learn new content affect the quality and durability of their knowledge [6]. The application of information technologies in new forms of teaching processes is inevitable, no matter how they are performed today.

The Covid 19 pandemic, in early March last year, was the cause of sudden, forced changes in the education system. In a period of just over a year and a half, during the pandemic, the teaching process went through three school years. The year before last school year, at the beginning of the second semester, due to the Covid 19 pandemic, classes were immediately relocated from classrooms and realized at a distance. Last school year, students were returned to the classrooms, but classes were held with shortened classes. Schools were obliged, in order to protect students, to organize classes in smaller groups of students. In most schools, students in the class were divided into two groups, so classes were shortened from 45 to 20 minutes. This school year, classes are held again in classrooms, with classes of 45 minutes [7]. Due to the still present pandemic, the protection measures that the Institute for Public Health of the Republic of Srpska sent to all schools are respected.

Regular changes in the education system include analysis of the situation, detailed preparations and implementation of appropriate measures to eliminate possible shortcomings.

Changes in the way the teaching process has been carried out in the last few years have not had time for preparations, by no means for those that are systematic and comprehensive. They were coming quickly and they also needed to be resolved quickly. The most important thing was not to interrupt the teaching process and to realize the school year.

Teachers and students seemed to come together to overcome obstacles, caused by sudden changes.

The year before last, they tried to improve their computer literacy at the same time. Not in the classroom, but in their homes, students waited for class, learning about the importance of applying information technology, not only for fun, but also to improve their own education. On the other hand, teachers prepared and sent material to students every day and for several hours, examined them, received assignments, control, seminar and graduation papers....

When they somehow managed to master it all, the classes were returned to the classrooms. Classes were shortened, there was twice less time left to work in the classroom than before. New challenges awaited students and teachers again. And no matter how impossible it seemed, the content of the lesson planned for 45 minutes had to be realized in 20 minutes. And that is where the students and teachers menaged themselves again. Of course, there had to be a supplement to the appropriate elements of knowledge, even after class. Information technologies have become an instrument of complement and link of the new age between teachers and students.

Research on the Significance of IT Application in Teaching

The paper includes research on students' attitudes towards different ways of teaching: regular, with a duration of 45 minutes; teaching with classes shortened to 20 minutes and classes conducted at a distance. In just two school years, students had the opportunity to participate in all three forms of teaching.

The research included third grade students of secondary vocational school. Most of them will reach the age of majority this year, so they can be considered responsible when filling out the questionnaires. The students answered the six questions anonymously. The questions related to the teaching process, which they had been going through since the beginning of the Covid 19 pandemic.

The research included questions related to the collection of data on the quality and representation of the teaching process conducted at a distance and with reduced hours, their comparison with traditional teaching; comparing the success that students have achieved in different forms of teaching,

the shortcomings of certain forms of teaching and the problems they encountered during any form of teaching process.

The data collected from the students and their expressed opinions point to new facts, attitudes of young people and changes that may follow in the very near future.

ANALYSIS OF COLLECTED DATA

The analysis of the collected data can show the attitude of young people towards education. Did the irregularity of attending classes and shortened classes affect that, but the fact is that the collected data show that young people generally do not want teaching to take place with classes of 45 minutes. The most common reason was that they lose concentration in the second part of the class, they cannot be focused on the teaching content and their attention decreases.

When asked if they were satisfied with the ways of conducting online and abbreviated classes, the majority of students answered positively (95%). Only 5% of students stated that they were not satisfied, without giving reasons.

When asked, with which form of teaching they achieved the best success in school, only three students answered that they achieved the same success, regardless of the way of teaching. Other students gave different answers.

The results of the research on student achievement can be presented using diagrams (Figure 1).
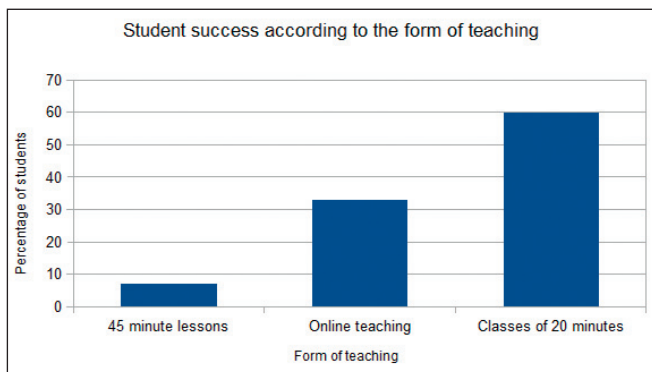
Figure 1. Students' opinions on the achieved success according to the form of teaching

The diagram shows that the largest number of students stated that they achieved the best results last year, with classes of 20 minutes (60%). One

third of students (33%) said that they achieved the best success during online classes and 7% of students said that this success was achieved during traditional classes of 45 minutes.

The students gave similar answers to the question, which form of teaching would they choose for this and the next school year, if they had a choice. More than four-fifths of students voted to come to school, but with shortened classes (81%), only 5% of students would choose teaching with 45-minute classes and 14% of students declared for online classes.

Comparing the previous answers, it can be noticed that one third of the students stated that they achieved the best success during online classes, but a significantly smaller number of them stated that they would like to continue this form of teaching. This shows that students still want to come to school. And in the oral conversation with the students, most of them state that they prefer attending classes at school and spending time with their classmates every day. However, either orally or through a survey, the majority of students say that classes should be shortened. When asked how realistic one school lesson should last, in order to be able to adopt all the new elements of knowledge, the students gave, perhaps unexpected answers (Figure 2).
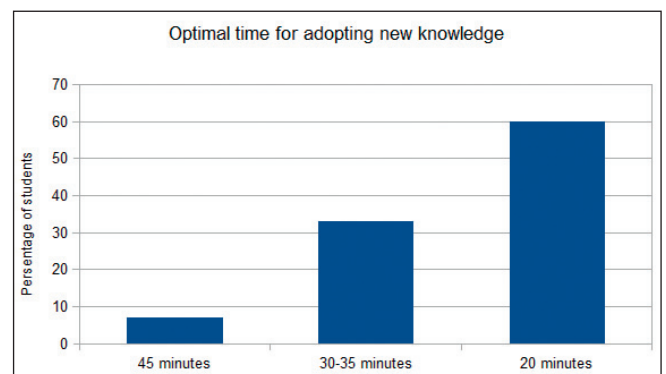
Figure 2. Students' opinion on the most favorable time for school hours

The diagram shows that most students believe that the optimal class time is between 30 and 35 minutes. When asked to state the reasons for shortening classes, they said that they thought that was enough to acquire new knowledge. Efficiency and effectiveness can be achieved during this time, with the appropriate organization of the class, the most

common are the comments of students. In the end, the students stated about the problems during the online classes or teaching with abbreviated classes. Students generally had no problems with any form of teaching (88%). Students who faced a problem during the implementation of the teaching process (12%), as the reason stated the inability to send materials, poor connection, weak signal on the network and the like.

THE ROLE OF SCHOOL IN YOUTH EDUCATION

According to the results of the research, students prefer to come to school, rather than learning at a distance. Most believe that the best results in learning are achieved by learning in the classroom. It is easier for them to follow the teaching contents, ask for explanations in case of misunderstandings or doubts, and finally, socialize with each other [8]. One student expressed himself very nicely, yet we are so young and eager to socialize.

The role of the school in the education of the younger generations is irreplaceable. It has always been a place of education but also ofupbringing young people. Education is somehow self-evident, but special attention should be paid to upbringing, especially today.

Students sometimes need only warm conversation and expression of understanding in order to achieve better success and increased responsibility.

And while the Covid 19 pandemic lasts, during which teachers and students are exposed to constant challenges, it is good to meet at school. They learned to use modern information technologies in the realization and improvement of the teaching process. They met with a completely new way of organizing and conducting classes. Now, just over a year after the declaration of the pandemic that is still going on, they have been returned to the original form of the teaching process, for 45 minutes. And while educators look forward to 45-minute classes, students generally want a shortened form of classes.

CONCLUSION

Students and teachers have faced the challenges that have been imposed on education in the last two years. Since the beginning of the Covid 19 pandem-

ic, the teaching process has gone through several forms. Distance teaching, classes shortened to 20 minutes, sending materials using modern IT tools, all these were new requirements.

Students and teachers went through it together. Maybe not quite the best, as it was wanted, but the teaching process did not suffer and managed to be realized.

Adapting the educational process to the difficult conditions of teaching implies its rationalization. Different forms of learning were used: e-learning at a distance, completely new platforms, the application of a virtual classroom and multimedia. The application of IT enables the realization and modernization of teaching content, and sometimes it is the only link between students and teachers.

After the changes, by returning the lessons to 45 minutes, it is noticed that the irregularity of the teaching process has left its mark. Research shows that, unlike teachers, most students are reluctant to accept a 45-minute class and prefer to shorten school hours.

LITERATURE
[1] https://www.unicef.org/bih/priče/covid-19-i-ponovno-otvaranje-škola
[2] Selimović, J. "Education during the pandemic", Faculty of Economics, Sarajevo, 2020.
[3] Ministry of Education and Culture of the Republic of Srpska, Beginning of the new school year, Banja Luka, 2021.
[4] Lazić, S., Avramović Ž. Z .: Digitization of traffic education in the Republic of Srpska, International scientific-practical conference, Russian University of Transport, Moscow, 2021.
[5] Lazić, S., Talić, T., Marinković, D .: The importance of modern information technologies in the difficult conditions of regular classes, IteO 2020, Pan-European University Apeiron, Banja Luka, 2020.
[6] Mandić, D. "Information Technology in Education", Srpsko Sarajevo, 2001.
[7] Republic Pedagogical Institute of Republika Srpska, www.rpz-rs.org.
[8] Lazić, S., Avramović Z .: The importance of IT in the online teaching process, 20th International Symposium INFO-TECH-JAHORINA, March 2021.

ABOUT THE AUTHORS

Slavojka Lazić, M.Sc. (1966), received her secondary education at the Gymnasium, Lukavac. She graduated from the Faculty of Transport and Traffic Engineering, University of Sarajevo, in 1991. Since 1994, she has been working as a professor of professional and theoretical teaching at the Technical School, Banja Luka. In the same educational institution, she performed the duties of assistant director for two terms (2010-2019).

She received her master's degree from the Faculty of Transport and Traffic Engineering, University of East Sarajevo, in 2010. At international and domestic conferences and scientific-professional gatherings, she has published a number of papers in the field of traffic and education.

Tijana Talić was born on January 1, 1983 in Banja Luka, Bosnia and Herzegovina. She graduated in 2008 from the Faculty of Natural Sciences and Mathematics in Banja Luka, Department of Mathematics and Informatics. She received her PhD in the field of computer science and informatics (2019) at the Faculty of Information Technology of the Pan-European University Apeiron. In 2019, at this faculty she was elected assistant professor in the fields of science: computer science and information science and bioinformatics.

INSTRUCTIONS FOR AUTHORS

The *Journal of Information Technology and Application (JITA)* publishes quality, original papers that contribute to the methodology of IT research as well as good examples of practical applications.

Authors are advised that adherence to the Instructions to Authors will help speed up the refereeing and production stages for most papers.

- Language and presentation
- Length of submissions
- Submission
- Contact details/biographies
- Title of the paper
- Abstract and keywords
- Figures and tables
- Sections
- Footnotes
- Special characters
- Spelling
- References
- Proofs
- PDF offprint
- Copyright and permissions
- Final material
- Correspondence
- Publication ethics

LANGUAGE AND PRESENTATION

Manuscripts should be written in English. All authors should obtain assistance in the editing of their papers for correct spelling and use of English grammar. Manuscripts should have double spacing, with ample margins and pages should be numbered consecutively. The Editors reserve the right to make changes that may clarify or condense papers where this is considered desirable.

LENGTH OF SUBMISSIONS

Papers should not normally exceed 12 Journal pages (about 8000 words). However, in certain circumstances (e.g., review papers) longer papers will be published.

SUBMISSION

Manuscripts must be submitted through the JITA online submission system.

Please read the instructions carefully before submitting your manuscript and ensure the main article files do not contain any author identifiable information.

Although PDF is acceptable for initial submission original source (i.e. MS Word) files will be required for typesetting etc.

CONTACT DETAILS/BIOGRAPHIES

A separate file containing the names and addresses of the authors, and the name and full contact details (full postal address, telephone and e-mail) of the author to whom correspondence is to be directed should be uploaded at the time of submission (you should select Contact details/Biographies as the file type). This file is not shown to reviewers. This file should also contain short biographies for each author (75 words maximum each) which will appear at the end of their paper.

The authors' names and addresses must not appear in the body of the manuscript, to preserve anonymity. Manuscripts containing author details of any kind will be returned for correction.

TITLE OF THE PAPER

The title of the paper should not be longer than 16 words.

ABSTRACT AND KEYWORDS

The first page of the manuscript should contain a summary of not more than 200 words. This should be self-contained and understandable by the general reader outside the context of the full paper. You should also add 3 to 6 keywords.

FIGURES AND TABLES

Figures which contain only textual rather than diagrammatic information should be designated Tables. Figures and tables should be numbered consecutively as they appear in the text. All figures and tables should have a caption.

SECTIONS

Sections and subsections should be clearly differentiated but should not be numbered.

FOOTNOTES

Papers must be written without the use of footnotes.

SPECIAL CHARACTERS

Mathematical expressions and Greek or other symbols should be written clearly with ample spacing. Any unusual characters should be indicated on a separate sheet.