Zbornik pete nacionalne konferencije sa međunarodnim učešćem pod nazivom

Primena slobodnog softvera i otvorenog hardvera PSSOH 2022

U Beogradu, oktobra 2022. godine

Univerzitet u Beogradu – Elektrotehnički fakultet

Zbornik pete nacionalne konferencije sa međunarodnim učešćem pod nazivom

Primena slobodnog softvera i otvorenog hardvera PSSOH 2022

u Beogradu, februara 2023. godine

Izdavači / Publishers
Univerzitet u Beogradu - Elektrotehnički fakultet /
University of Belgrade – School of Electrical Engineering i / and
Akademska Misao / Academic Mind

Štampa / Printed by
Akademska Misao / Academic Mind

ISBN: 978-86-7466-973-0

Tiraž / Number of copies: 50

Mesto i godina izdanja / Place and year of publication
Beograd, 2023. / Belgrade, 2023

University of Belgrade – School of Electrical Engineering

Proceedings of the Fifth National Conference with International Participation titled

Application of Free Software and Open Hardware PSSOH 2022

in Belgrade, February 2023.

Predgovor petoj PSSOH konferenciji

U 2022. godini održana je peta jubilarna dvojezična nacionalna PSSOH konferencija sa međunarodnim učešćem. Iako su nam uslovi dopustili da konferenciju organizujemo uživo, omogućili smo snimanje predavanja i postavljanje na YouTube kanal, uz saglasnost predavača. Zbog nemogućnosti učešća predavača uživo, jedno predavanje je održano online. Ove godine uživo održavanje konferencije nije uspelo da privuče veliki broj učesnika, ali smo jako zadovljni diskusijama i umrežavanjem koje je usledilo u pauzama i nakon konferencije. Naša želja da negujemo nacionalne skupove u Republici Srbiji i da ujedno promovišemo Elektrotehnički fakultet, Univerziteta u Beogradu kao mesto na kome je moguće razmenjivati ideje i predstavljati rezultate rada iz oblasti koje pokriva konferencija je ispunjena.

Zbornik publikujemo nakon što je konferencija završena iz istih razloga kao ranijih godina, a to je mogućnost javne recenzije radova, ali i rasterećenje sopstvenih obaveza i rasporeda. Kako se na PSSOH konferenciji neguje dobra praksa otvorene nauke, nikome ništa nismo naplatili, a svi radovi su dostupni u otvorenom pristupu, uključujući i elektronsku verziju PSSOH zbornika (dijamantski pristup).

Plenarnu sesiju smo ove godine zadržali, ali uz samo jedno predavanje i to Itala Vinjolija koji je ovog puta došao lično u Beograd, na čemu smo mu veoma zahvalni. Pristizala su nam pozitivna iskustva iz publike nedeljama i mesecima nakon predavanja Itala Vinjolija – posebno je vladao veliki interes za strategiju monetizacije i komercijalizacije softvera otvorenog koda.

Pored uvodne sesije gde su se prisutnima dekan Elektrotehničkog fakulteta dr Dejan Gvozdić, redovni profesor ispred našeg Fakulteta i Nadica Miljković ispred Uredničkog odbora i plenarne sesije, izlaganje radova je bilo organizovano u još jednoj sesiji sa pet predavanja po pozivu. Miloš Đorđević je ispred CMS (eng. *Compact Moun Solenoid*) kolaboracije predstavio iskustva u deljenju otvorenih podataka u CERN-u (eng. *The European Organization for Nuclear Research*), Tamara Vučenović je održala predavanje koje je imalo za cilj da predstavi fenomen lažnih vesti, Ana Gavrovska je predstavila alate za analizu video signala i podelila svoja iskustva, Kannika Thaimai i Ivana Madžarević su predstavite UNLOCK aceleratora i njegov značaj za otvorene inovacije, dok su Andrijana Todosijević, Katarina Simonović i Anđela Arsović doprinele tekućim PSSOH temama sa svojim iskustvom u korišćenju alata otvorenog koda za upravljanje logovima i vizuelizaciju u AMRES-u (Akademska mreža Republike Srbije). Učesnici su imali prilike da čuju različita iskustva u radu sa otvorenim podacima i softverskim alatima, kao i da se upoznaju sa iskustvima u deljenju otvorenih podataka u CERNu- i da steknu nova znanja o tome kako da odgovorno koriste dostupne informacije. Konferenciju je ispred uredničkog odbora zatvorio Miloš Cvetanović. Postkonferencijski događaji su ove godine izostali, ali to ne znači da je interes za ranije post-konferencijske događaje izostao.

Na svim prethodnim, ali i budućim PSSOH konferencijama svi gosti i predavači su dobrodošli ako žele da pomognu u skladu sa svojim mogućnostima i sve tretiramo ravnopravno, što je tradicija pokreta slobodnog softvera i otvorenog hardvera, a sada i tradicija PSSOH konferencije. Kako smo najavili u 2018. i realizovali kasnije, i ove godine smo ponudili svim autorima koji su na Elektrotehničkom fakultetu Univerziteta u Beogradu objavili otvorene nastavne materijale da ih predstave u Zborniku, te Zbornik sadrži i ovu sekciju.

Organizacija PSSOH konferencije je podržana od strane velikog broja pojedinaca, ustanova, kompanija i udruženja i ovde ih je nemoguće sve pobrojati. Najzahvalniji smo našim donatorima iz Akademske Misli iz Beograda, ali i LotusFlare kompaniji koja je pokrila troškove štampanja majica sa PSSOH logom. Zahvalni smo svim predavačima koji su se odazvali našem pozivu i upotpunili sadržaj naše konferencije. Bez Nikole Todorovića i Miloša Bjelića koji predsedavaju Organizacionim odborom ne bi smo mogli da zamislimo PSSOH. Veoma smo zahvalni članovima Organizacionog odbora Dragici Nikolić, Dejanu Petkoviću, Jovanu Sandiću, Mihajlu Pavloviću, Živani Garašević, Danilu Đokiću, Sari Živković, Pavlu Radojkoviću i Leni Milovanović. Zahvalnost dugujemo i svim članovima naučnog i spoljnjeg organizacionog odbora. Kako nismo u mogućnosti da sve koji su nam pomogli nabrojimo, izvinjavamo se svima koje nismo spomenuli.

U duhu PSSOH tema i sa željom da promovišemo slobodan softver i uz veliku zahvalnost Italu Vinjoliju na nesebičnoj podršci, ovaj Zbornik smo pripremili u programskom paketu LibreOffice.

u Beogradu, 14. februara 2023. godine

Uredнички odbor PSSOH konferencije

Foreword to the Fifth PSSOH Conference

The fifth quinquennial bilingual national PSSOH conference with international participation is held in 2022. Although we organized an in-person event, we recorded lectures and with the lecturer's consent uploaded them to our YouTube channel. Due to the overlapping schedule, one lecture was held online. The live PSSOH 2022 conference did not attract a larger number of participants, but we are satisfied with the discussions and overall networking. We are proud for being able to foster national gatherings in the Republic of Serbia and to promote the School of Electrical Engineering, University of Belgrade (ETF) as a place with vivid atmosphere that nourishes free exchange of ideas and results in line with the conference topics.

We publish the Proceedings after the conference for the same reasons as previously, to enable public review of the papers and to attenuate our own obligations and schedules. As PSSOH conference complies to the open science principles, we did not charge article processing charges and fees – all papers are available in open access, including the electronic version of the PSSOH proceedings (Diamond Open Access).

We kept the plenary session this year, but with only one lecture held by Italo Vignoli, who came to Belgrade personally this time, for which we are very grateful. We received positive feedback from the audience for weeks and months after Italo Vignoli's lecture - in particular, there was great interest in the monetization strategy for the commercialization of open source software.

In addition to the Introductory session, where ETF Dean Prof. Dejan Gvozdić and Editor Nadica Miljković welcomed all participants, and Plenary session, the main PSSOH session with five invited lectures was held. Milos Dordevic, in front of the CMS (Compact Moun Solenoid) collaboration, presented experiences in sharing open data at CERN (The European Organization for Nuclear Research), Tamara Vučenović gave a lecture on the fake news phenomenon, Ana Gavrovska presented open tools for video signal analysis and shared her experiences, Kannika Thaimai and Ivana Madžarević presented the UNLOCK accelerator for open innovation, while Andrijana Todosijević, Katarina Simonović, and Anđela Arsović contributed to current PSSOH topics with their experiences in using open source tools for log management and visualization in AMRES (Academic Network of the Republic of Serbia). The participants had the opportunity to discover different experiences related to the open data and open software tools, as well as to learn about sharing open data practices at CERN – as well as to gain new skills how to responsibly use/judge available information. Miloš Cvetanović closed the conference in front of the Editorial board. Post-conference events were not organized in 2022, but we are satisfied with the current interest in our past post-conference events.

At all previous and future PSSOH conferences, all guests and lecturers are welcome if they want to help according to their capabilities and we treat everyone equally, which is tradition of the free software and open hardware movement, and of the PSSOH conference. As we announced in 2018, we offered all Authors who published open teaching materials at ETF to present them in the Proceedings.

The organization of the PSSOH conference is supported by many individuals, institutions, companies and associations, and it is impossible to list them all here. We are most grateful to our donors from Academic Mind from Belgrade, but also to the company LotusFlare, which covered the costs of printing PSSOH T-shirts. We are grateful to all lecturers for their kind contribution. Without Nikola Todorović and Miloš Bjelić, who serve as Chairs of the Organizing Committee, we would not be able to imagine PSSOH. We are very grateful to the members of the Organizing Committee Dragica Nikolić, Dejan Petković, Jovan Sandić, Mihajlo Pavlović, Živana Garašević, Danilo Đokić, Sara Živković, Pavlo Radojković and Lena Milovanović. We also owe thanks to all members of the Scientific and External Organizational boards. As we are not able to list everyone who helped us, we apologize in case we failed to mention anyone.

According to the PSSOH themes and with aim to promote application of free software and with kind gratitude to Italo Vignoli for his selfless support, this Proceedings is prepared in LibreOffice.

in Belgrade, February 14, 2022.
Editorial Board of the PSSOH Conference

Sadržaj / Table of Contents

FOSS Sustainability

Italo Vignoli
One of the founders and team members of The Document Foundation and main spokesperson
italo@documentfoundation.org

Announcement and Outline for Plenary Lecture

This is the third invited lecture held by Italo Vignoli from The Document foundation at the PSSOH conference. Mr. Vignoli's presentation covered topics related to the Free and Open-source Software (FOSS) projects sustainability. Our participants had a chance to learn about LibreOffice experiences with special interest (from the audience) on the monetization strategy for commercialization of open-source projects. The presentation held by Italo Vignoli at PSSOH 2022 is available on YouTube (https://youtu.be/Peir_qkEYeg, Accessed on February 27, 2023) and on Zenodo repository (DOI: https://doi.org/10.5281/zenodo.7244311).

Keywords: Free and Open-source Software; FOSS, Sustainability; Open Standards; LibreOffice; Open Document Format.

PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7189233

Open Data from CMS at CERN: Status and Plans

Milos Dordevic[1] on behalf of the CMS Collaboration

1: Vinca Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, 11351, Vinca, Belgrade, Serbia
e-mail: milos.dordevic@cern.ch

Abstract

The CMS Collaboration at the CERN LHC has released to the public more than two petabytes of open data. The large parts of these datasets that were used in the data analyses have led to the discovery of the Higgs boson in 2012. This open data, apart from its originality and scientific value, has already facilitated public results produced by non-collaboration members, within the high energy physics or even from different fields, thus improving the knowledge exchange and supplementing the original research with new ideas and insights. The CMS open data is already used for educational and outreach purposes, through providing the hand-on exercises for CERN Masterclass and other events. A brief introduction to the LHC and the CMS Experiment is outlined, followed by the description of real-time data selection by the Trigger system and an overview of raw data to physics results pathways. Information is given on how to access the CMS open data using virtual machines and containers, followed by presenting in more detail an example of a simplified CMS analysis at the Trigger level. The possibilities to address an application of Machine Learning in high energy physics using the CMS open data are shown as well, concluding with the user feedback, and an opinion from Nature Physics.

Key words: [cern, cms, open, data, status]

1 Introduction

At the CERN Open Data portal [1], an interested subject can find more than two petabytes of open data released by the CMS Collaboration at the CERN Large Hadron Collider (LHC) [2]. These are the original datasets from high energy proton-proton collisions recorded by the CMS Experiment in 2010, 2011 and 2012. A large portion of these datasets, released now to the public view and scrutiny, was used to discover the Higgs boson in 2012 by the CMS Collaboration. According to the CMS Open Data policy [3], the CMS Collaboration has committed to release 100% of its analysable data within ten years of collecting it. The other LHC experiments, ATLAS, LHCb and ALICE, have also released their data to the public, with the corresponding usage policies [4, 5, 6]. The CMS Open Data is nowadays being used in scientific research, both within the high energy physics community [7, 8] and by non-collaboration researchers. It is also used for educational and outreach purposes all around the world, enabling them to promote the field and attracting young people to cutting-edge research being done at CERN.

Section 2 gives a brief overview of the LHC and CMS Experiment, presenting the basics of real-time data selection using the Trigger system, as well as the analysis pathways needed to convert the raw data to physics results. In Section 3, the motivation to release the CMS open data, along with the details of how to access it and examples of its usage, are outlined. Section 4 reports on the results published

using CMS open data and Section 5 gives a summary with plans for further usage and improvements.

2 Overview of the LHC and the CMS Experiment at CERN

The Large Hadron Collider (LHC) at CERN is in operation for more than a decade already, steadily delivering high intensity particle beam collisions at record breaking energies. The products of these high-energy collisions are being recorded by the most powerful ”cameras”, being the particle detectors, wrapped around the beam interaction points. The CMS detector is a multilayered, general-purpose detector designed primarily for discovery and characterisation of the Higgs boson, precision Standard Model measurements and Beyond Standard Model searches (e.g. Supersymmetry or Extra Dimensions). The particle beams colliding at the LHC are organised in bunches, with around 100 billion protons in each bunch at the design luminosity value. The bunch collisions happen at a rate of 40 MHz, while the production of the Higgs boson and potential new particles are very rare events, occurring at the rate of the order of 1 Hz or lower, respectively. Recording all the collision data would generate more than 50 terabytes each second and storing such an amount of data would be impossible and, in fact, not needed for the physics goals of the CMS experiment. Selective read-out of the particle collisions is performed in real time by the Trigger system [9]. The CMS Trigger system is implemented in two tiers, the Level-1 trigger (L1) based on custom-made fast electronics that reduces the rate to about 100 kHz, and the High Level Trigger (HLT) using a software running on computing farm, further reducing the rate to about 1.3 kHz, saving full event content and performing prompt reconstruction. Figure 1 presents a graphic image of the LHC along with the map of the area. Figure 2 shows a 3D model of the CMS experiment, revealing its very complex, multilayered internal structure.

The raw data that comes out of the online system is followed by the event reconstruction which refines this data, also applying calibrations and performing the creation of higher-level physics objects. This procedure is usually referred to as skimming (also slimming, thinning, etc.), which reduces the data size, but also increases its usability, both by reducing the number of events and compressing the event format. Event information from each step in the simulation and reconstruction chain is logically grouped into what is called a data tier. Examples of data tiers include RAW and RECO [12]. RAW represents the detector data after online formatting, the L1 trigger result, the result of the HLT selections and potentially some of the higher level quantities which are calculated during HLT processing. It would be impossible to perform physics analysis using raw primary datasets, hence the RECO data tier is made, representing first level of the event reconstruction. Detector-level information is passed through the various reconstruction algorithms. Tracking, vertexing, and Particle Flow are performed in this step and then some basic physics object collections are created. This step is very CPU intensive. The further reduction to physics objects is performed when creating the AOD (Analysis Object Data) data tier, where only some hits and other detector-level info is kept, making the physics object collections as priority and retaining some supporting information from RECO. The AOD data tier can be further skimmed into more compact data tiers, which will be described in Section 3.3. An example of tracking, ECAL and HCAL-based objects corresponding to basic data tiers is shown in Figure 3.

3 The CMS open data: motivation, access and examples

3.1 The motivation for releasing the CMS open data to the public

There is a strong motivation and multiple reasons to release the CMS data for public use. First and foremost, science results should be inclusive and open to everyone. By making the CMS data open, more people become engaged with research. There is an important educational aspect to it, since it provides an ideal platform for teaching and exercises, thus attracting students to particle physics. It is also another way to directly return something to society. The CMS open data improves communication and facilitates the exchange of knowledge with the researchers first within the same, particle physics

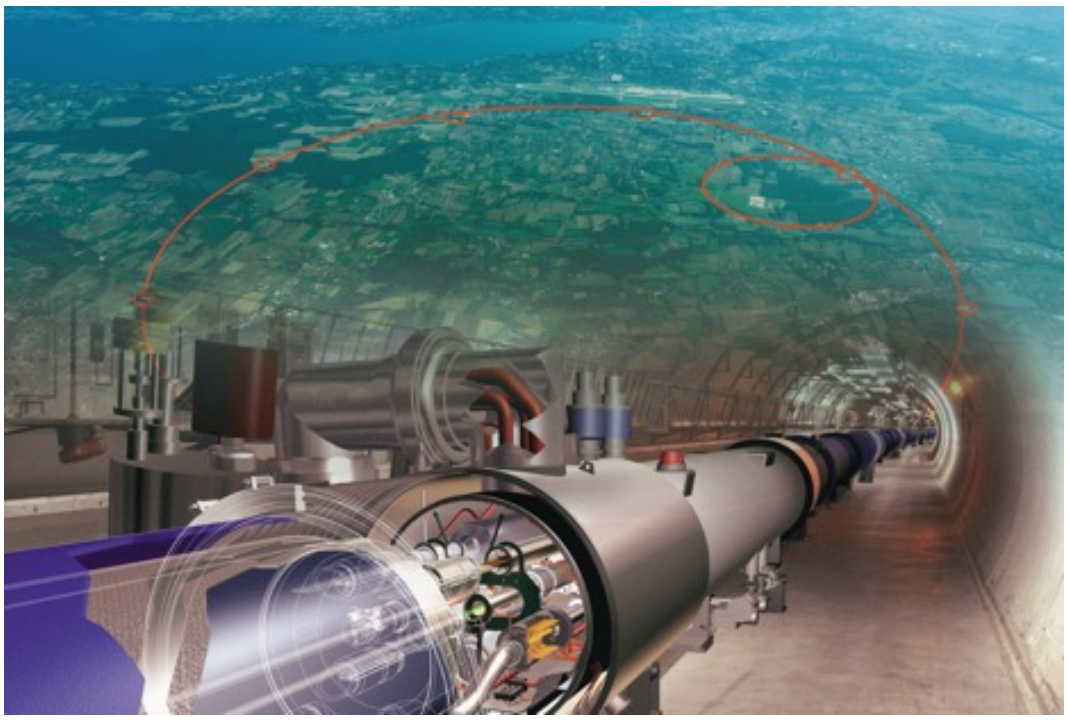


Figure 1: The Large Hadron Collider at CERN.

research field, but also from other scientific areas, such as machine learning and data science topics.

3.2 Data levels and access: virtual machines and containers

The starting point for the usage of the CMS open data is the webpage [1] where a user can browse more than two petabytes of particle physics data, allowing to inspect different datasets, environments, software and many examples, each well documented. The CMS open data has the following levels:

- Level 1 : Open access publication and additional numerical data
- Level 2 : Simplified data for Outreach and Education
- Level 3 : Reconstructed data and the software to analyse them
- Level 4 : Raw data, and the software to reconstruct and analyse it

as defined in the CMS data policy [3], starting either from raw experimental or simulated data, going through the reconstructed data and the datasets with the higher level of abstraction generated by analysis workflows, and ultimately all the way to data which are represented in scientific publications. Each of these levels enables different opportunities for long-term re-use, but also poses different challenges for data preservation. Level 1 corresponds to publications, with additional documentation provided, in order to put the results in context and understand the analysis procedures, some additional numerical data which did not or could not appear in the publications. Examples of these would be the cross sections of different physics processes, given as a function of multiple variables. The Level 2 includes simplified data formats, such as, for example, multi-dimensional distributions of analysis variables, or the four- vectors of particles or jets, energy clusters and tracks. These could be re-used immediately for theory interpretations, some limited physics analysis, but also for educational and outreach purposes. Level 3 represents the reconstructed data and simulations, released together with the corresponding

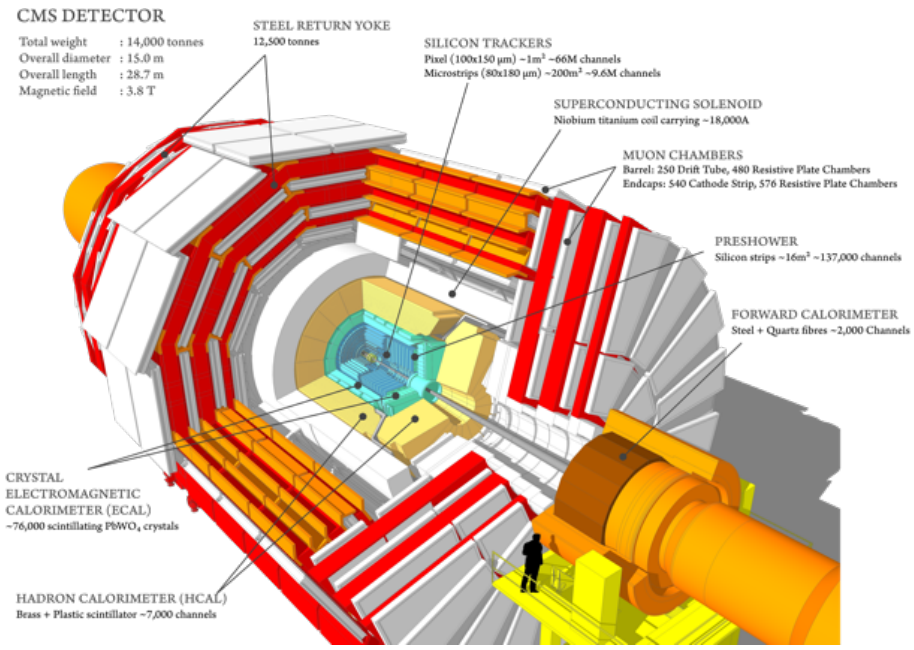


Figure 2: The CMS Experiment at CERN LHC.

software, analysis workflows and documentation that is needed to access this data, reproduce published analyses, or to perform new analyses without requiring re-reconstruction of the data or running the new simulations. Level 4 is the raw data and the software and documentation needed to access, reconstruct and analyse this data. This is the most complex level, which also requires the highest computational effort. An example of using Level 3 CMS open data will be presented in Section 3.4.

Access to the CMS open data is provided via the usage of the CERN Virtual Machine (VM) [10], or the Docker container [11], providing the usage of the CMSSW [12] environment and the ROOT [13] framework. The setup of the CERN VM is enabled through the usage of the Oracle VM VirtualBox software, a free, open source and multiplatform application to run virtual machines. It provides a base for the operating system which is compatible with the CMSSW environment needed to run the analysis on the open data. Following the download of the required software, the VM file size on the host machine is at the order of a few GBs. Docker is a free, commercial implementation of a software container, as an alternative to the VM images. There are different kinds of container images available at the Docker Hub and CERN GitLab, from the light-weight ones to the images containing full CMSSW installation that is at the order of several GB, allowing them to preserve complete CMS physics analyses. Both VM and Docker organisation and structures compared, are shown in Figure 4.

3.3 Data format of CMS open data

Most of the CMS open data is published in the Analysis Object Data (AOD) format which includes serialised C++ objects requiring the CMSSW environment and ROOT framework to be read. Each of the AOD events holds about 500 kB of information, resulting generally with large files. The CMS Collaboration has thus developed derived data formats called MiniAOD [14], and its successor NanoAOD [15]. While the MiniAOD is similar to AOD and also storing serialised C++ objects, NanoAOD is based on storing the basic types such as floats, integers and arrays thereof. Table 1 presents an example of the variable content of the muons collection, embedded in the NanoAOD data format.

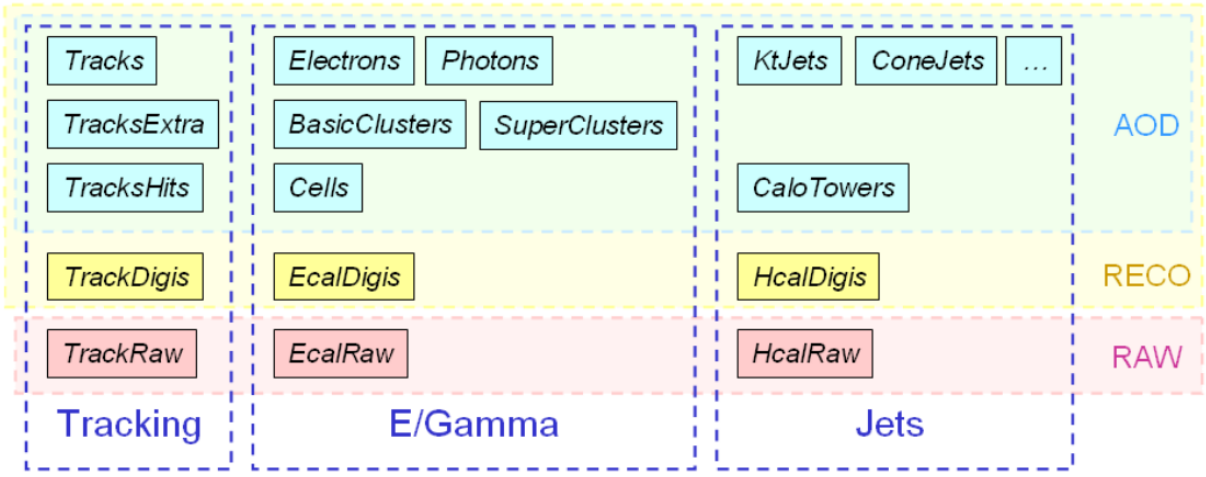


Figure 3: Tracking, ECAL, and HCAL-based objects corresponding to different CMS data tiers.

Table 1: Data format of a muon collection in the NanoAOD.

Variable	Type	Description
nMuon	unsigned int	Number of muons in this event
Muon_pt	float[nMuon]	Transverse momentum of the muons
Muon_eta	float[nMuon]	Pseudorapidity of the muons
Muon_phi	float[nMuon]	Azimuth of the muons
Muon_mass	float[nMuon]	Mass of the muons
Muon_charge	int[nMuon]	Charge of the muons (either 1 or -1)

3.4 Examples of usage: Higgs boson and CMS Trigger System

The CMS open data has provided an example [16] of a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis, published in [17]. This example enables different levels of complexity, from the one useful for educational purposes, to the more complex one requiring at least some minimal understanding of the content of the paper [17]. The example uses legacy versions of the original CMS datasets in the AOD data format, but slightly different than in the original publication. Many of the data selection cuts are the same, however this CMS open data analysis is still a much simplified reimplementation of the original CMS Higgs to four leptons analysis. The four lepton invariant mass spectrum in $H \rightarrow 4l$, as obtained using CMS open data, is outlined in Figure 5.

An example [18] needed to extract the Trigger information from the CMS Open/Legacy data is provided, with its implementation in C++ code and configuration in Python [19]. In this example one can find the following analysers, each performing some of basic Trigger analyses using CMS open data:

- GeneralInfoAnalyzer : several C++ snippets on how to access trigger information such as metadata, prescales, module information, etc.
- ModuleInTriggerAnalyzer : shows how to dump all the modules for a specific trigger and/or obtain the last active module (filter) of a trigger.
- TriggerMatchingAnalyzer : how to match reconstructed tracks to objects that fired a trigger (or possible set of triggers) that contain a specific module.
- TriggerSimplePrescalesAnalyzer : use wildcards to access different versions of the same trigger, check their L1 and HLT prescales, and whether the trigger accepted the event or not.

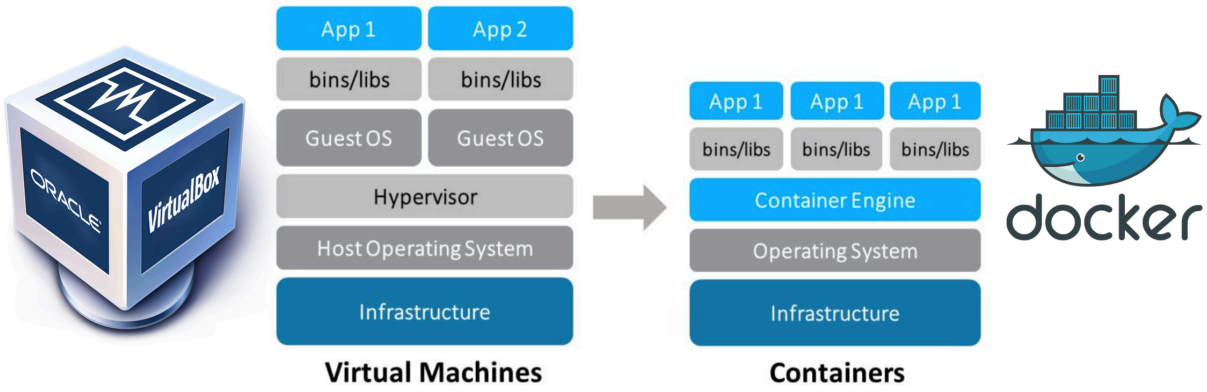


Figure 4: Oracle VM VirtualBox (left). Docker software container (right).

Here we will focus on the ModuleInTriggerAnalyzer and present an example of running it, together with the required commands and the obtained output, using the CERN VM. Example of a typical HLT path used in CMS Trigger is shown in Figure 6. After setting up the VM, in the terminal, the list of commands is required to be typed in to run the analyser. These commands include the creation of the CMSSW environment, downloading of the corresponding analyser from the github repository, navigation to the directory where the example is installed, linking required databases for replicating the analysis conditions and finally running the examples. The described workflow is the following:

```
cmsrel CMSSW_5_3_32
cd CMSSW_5_3_32/src/
cmsenv
git clone -b 2011 git://github.com/cms-legacydata-analyses/TriggerInfoTool.git
cd TriggerInfoTool
cd {packagename}
scram b
ln -s python/{configname} .
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA
    FT_53_LV5_AN1
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db
    FT_53_LV5_AN1_RUNA.db
ls -l
ls -l /cvmfs/
cmsRun {configname} > full.log 2>&1 &       (checking w/ "tail -f full.log")
```

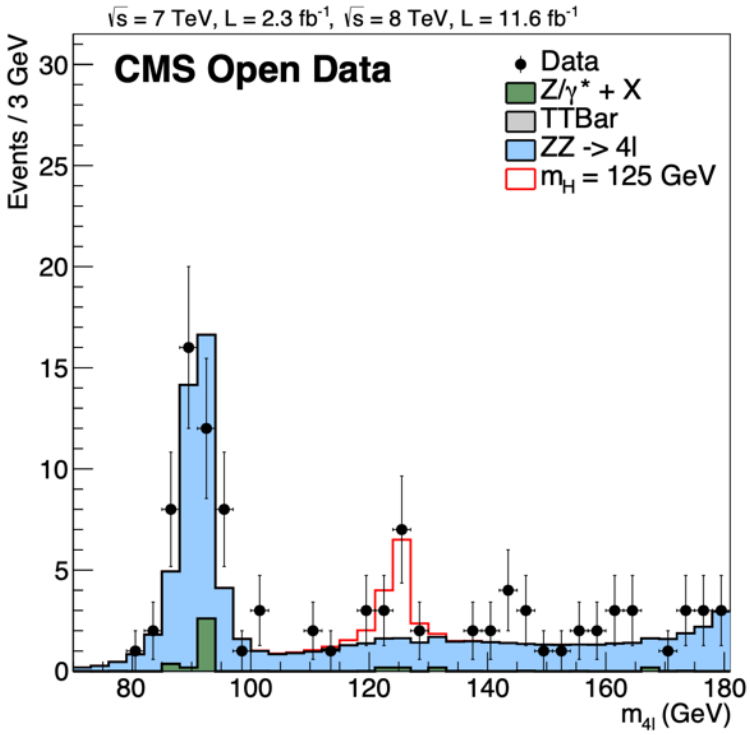


Figure 5: The four lepton invariant mass in the $H \rightarrow 4l$ analysis using CMS open data.

Following the execution of these commands, the corresponding output file is provided. The selection of this output, listing all the modules contained in the Trigger path HLT_Jet190_v6, as well as the event processing information with the selected event range are printed, as shown in Figure 7. The last module that has fired in this trigger path for each event is highlighted with the corresponding colour. This is a basic example of Trigger analysis, but such evaluations are very often made in CMS Trigger.

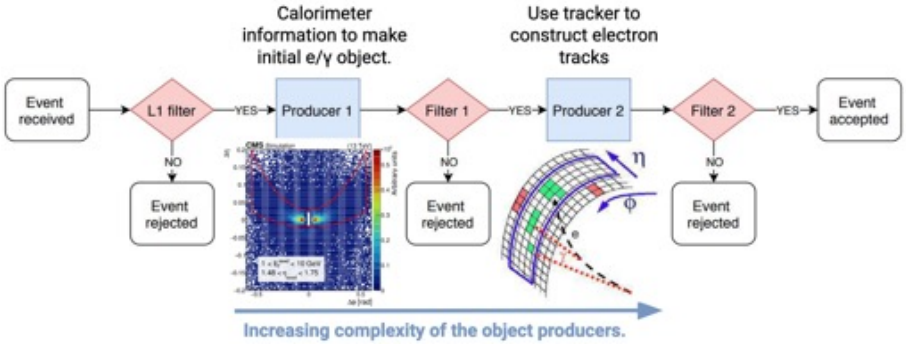


Figure 6: The example of the HLT path, showing increasing complexity of the object producers.

3.5 CMS Open data for machine learning in high energy physics

The CMS Open Data has also addressed the continuously growing application of machine learning (ML) to various challenges in high energy physics [20]. In the paper [21], it is clearly outlined that collaboration with data science and ML community is taken as a high priority in helping to advance the application of state-of-the-art algorithms to high energy physics. The ML datasets, derived from millions of CMS simulation events, focus on solving a number of problems in particle identification, tracking and distinguishing between multiple collisions in each bunch crossing (pileup). Reconstructed data and simulations released are from the CASTOR calorimeter, used by CMS in 2010,

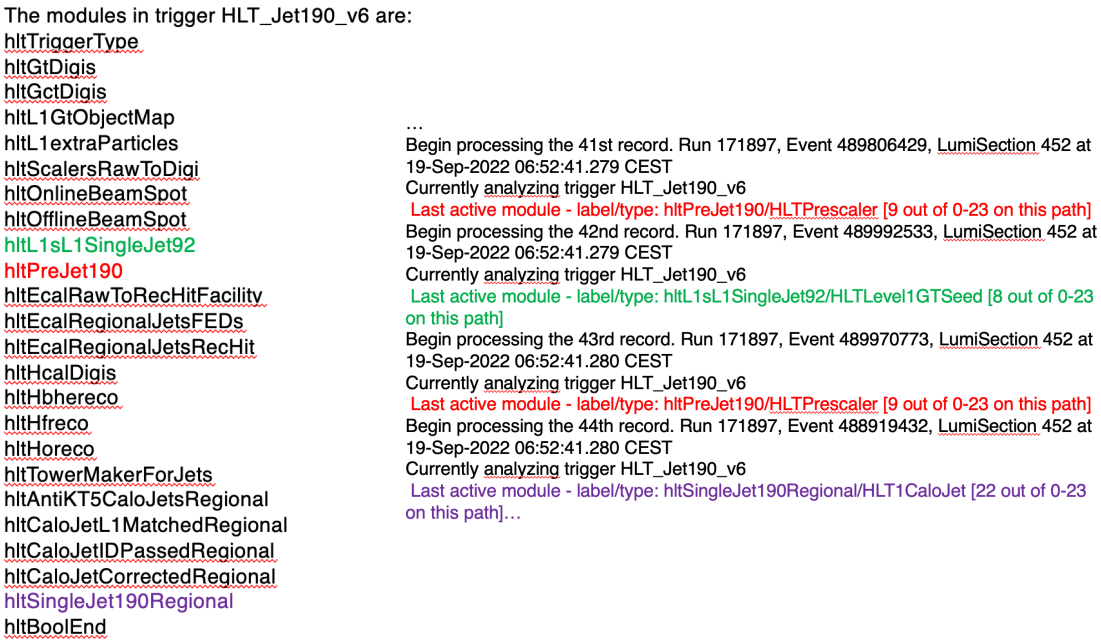


Figure 7: The output of running the ModuleInTriggerAnalyzer with CMS open data. The modules contained in trigger path HLT_Jet190_v6 (left). Output content showing event processing (right).

representing the first release of data from the very-forward region of CMS, with full instructions on access and usage.

4 Published results using CMS Open data and users feedback

There are already several scientific papers published with CMS open data. Some examples are the papers [22] and [23] which studied the QCD splitting functions and jet substructure, respectively, both being very common subjects of studies within the real LHC experiments. These are very clear examples of how the CMS open data can be used by theorists to extract (and publish) the valuable physics information. In the paper [23], the jet transverse momentum was extracted like in many original CMS analyses, and this is shown in Figure 8. We will not this time go into details of this paper. However, there are some very interesting lessons that the authors of this paper have kindly shared with the public, based on their own experience with using the CMS Open Data, quoting them precisely:

- "We then converted AOD files into a text-based MIT Open Data (MOD) format to facilitate the use of external analysis tools."

- “From the physics perspective, our experience with the CMS Open Data was fantastic”
- “From a technical perspective, though, we have encountered a number of challenges”

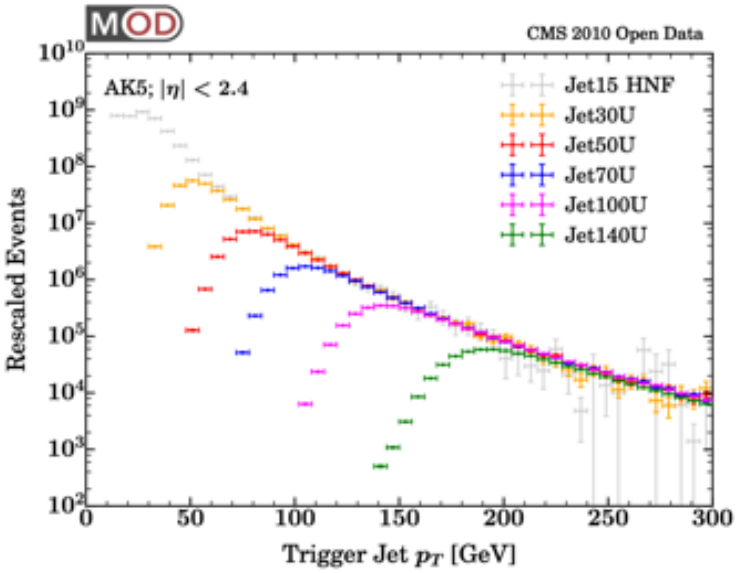


Figure 8: Leading jet transverse momentum spectrum, as obtained from CMS open data.

There is also an opinion article on the CERN open data published in Nature Physics [24]. Here are a few quotes from this article, that the author of this manuscript found to be the most interesting to cite:

- “only those who spent years building the experiment have earned quick access”
- “other scientists can analyse the data while LHC is still running, testing unconventional strategies”
- “public data can complement the overall research effort”

These quotes stressed some of the most important points related to releasing the CMS data to the public. Even though some of the users have had difficulties in parsing the available data format, faced complexity of the CMSSW software environment or sometimes found it challenging to browse through the documentation, all the analysts who used open data have highlighted the importance of the research-like quality of the data released by the CMS Collaboration and rooted for its continuation. User experience and feedback are ranked among the most important for the future of CMS open data.

5 Summary

The CMS Collaboration is leading the Open Data effort within the LHC experiments at CERN. The services to locate, browse and detailed descriptions on how to use the CMS Open Data are all provided and broadly used for education, outreach and scientific purposes. The means of using Virtual Machines and Docker containers have allowed access to the original CMS data, with the possibility, in the latter case, even to preserve a full CMS analysis. In this paper, examples of performing a simplified Higgs to four leptons CMS analysis, as well as how to access and run the basic CMS Trigger analysis, are presented, with the latter also specifying the user commands and the output. The use of CMS Open Data is already spreading throughout the scientific community, mainly in high energy physics theory, but also in other related disciplines such as machine learning and data science. According to the CMS open data policy, the public release of new CMS data is becoming imminent and also improvements to the documentation and use are continuously being made.

References

[1] CERN, The CERN Open Data portal, https://opendata.cern.ch.
[2] The CMS Collaboration, JINST 3 (2008) S08004.
[3] The CMS Collaboration, CMS preservation policy, 10.7483/OPENDATA.CMS.7347.JDWH.
[4] The ATLAS Collaboration, ATLAS preservation policy, 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ.
[5] The LHCb Collaboration, LHCb preservation policy, 10.7483/OPENDATA.LHCb.HKJW.TWSZ.
[6] The ALICE Collaboration, ALICE preservation policy, 10.7483/OPENDATA.ALICE.54NE.X2EA.
[7] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, J. Thaler, Jet substructure studies with CMS open data, Phys. Rev. D 96, 074003 (2017).
[8] C. Cesarotti, Y. Soreq, M.J. Strassler, J. Thaler, W. Xue, Searching in CMS open data for dimuon resonances with substantial transverse momentum, Phys. Rev. D 100, 015021 (2019).
[9] The CMS Collaboration, The CMS trigger system, JINST 12 (2017) P01020.
[10] CERN, The CERN Open Data portal, http://opendata.cern.ch/docs/cms-virtual-machine-2011.
[11] CERN, The CERN Open Data portal, http://opendata.cern.ch/docs/cms-guide-docker.
[12] The CMS collaboration, CMS Offline Software, https://github.com/cms-sw/cmssw.
[13] R. Brun, F. Rademakers, ROOT - An object oriented data analysis framework (1997).
[14] G. Petrucciani, A. Rizzi,C. Vuosalo, Mini-AOD: A New Analysis Data Format for CMS (2015), http://dx.doi.org 10.1088/1742-6596/664/7/072052.
[15] A. Rizzi, G. Petrucciani, M. Peruzzi (CMS Collaboration), EPJ Web Conf. 214, 06021, 6 p, (2019).
[16] CERN, The CERN Open Data portal, http://opendata.cern.ch/record/5500.
[17] The CMS Collaboration, S. Chatrchyan et al., Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys.Lett.B 716 (2012) 30-61.
[18] CERN, The CERN Open Data portal, http://opendata.cern.ch/record/5004.
[19] CERN, The CERN Open Data portal, https://github.com/cms-opendata-analyses/TriggerInfoTool/tree/2011.
[20] CERN, The CERN Open Data portal, https://home.cern/news/news/knowledge-sharing/cms-releases-open-data-machine-learning.
[21] K. Albertsson et al., Machine Learning in High Energy Physics Community White Paper, arXiv:1807.02876.
[22] A. Larkoski et al., Exposing the QCD splitting functions with CMS open data, Phys. Rev. Lett. 119, 132003 (2017).
[23] A. Tripathee et al., Jet Substructure Studies with CMS Open Data, Phys. Rev. D 96, 074003 (2017).
[24] M. Strassler, J. Thaler, Slow and steady, Nature Physics 15, 725 (2019).

PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7324850

KAKO PREPOZNATI LAŽNE VESTI U SAVREMENOM TEHNOLOŠKOM OKRUŽENJU

Primeri iz prakse

Tamara Vučenović

Univerzitet Metropolitan, Fakultet za menadžment, 11000, Beograd, Srbija,
elektronska pošta: *tamara.vucenovic@metropolitan.ac.rs*

REZIME

U radu se razmatra fenomen/pojam ’’lažnih vesti’’, ukratko se ukazuje na njegovo značenje, zašto je važno da naučimo da prepoznamo lažne vesti, kao i koji mehanizmi i konteksti utiču na nekritičko prihvatanje dezinformacija. Takođe, pojam onlajn građanskog novinarstva je kratko predstavljen, pre svega imajući u vidu njegov značaj za savremeni medijski kontekst, u kome se generišu i dele i lažne vesti. U završnom delu rada prezentovan je izbor organizacija i relevantnih adresa na internetu gde se, uz brojne primere, može naučiti kako da se prepoznaju/provere lažne vesti u savremenom tehnološkom okruženju, već prezasićenom ogromnim brojem informacija.

Ključne reči: lažne vesti, mediji, digitalne tehnologije, IT, dezinformacije, obrazovanje.

1 Uvod – šta su lažne vesti?

’’Prema Google Scholar-u, 2210 publikacija na engleskom jeziku sa „lažnim vestima“ u naslovu pojavilo se od januara 2017. do početka 2020.godine, dok je, u svim godinama koje su prethodile i uključujući 2016, bilo ukupno 73.” [1 p5-6]

Tehnološki razvoj, razvoj digitalnih medija, posebno veb 2.0/3.0 veb okruženja, radikalno menjaju način na koji se ljudi svakodnevno odnose i deluju sa medijima: upotreba je jednostavnija, digitalni mediji sve dostupniji, korisnici, koji su u prošlosti uglavnom bili konzumenti, mogu da postanu kreatori sadržaja koji svoju poruku mogu da pošalju milionima ljudi. Dakle, izmenjen je medijski prostor, ali i načini na koje planiramo, kreiramo, realizujemo i distribuiramo medijske sadržaje. “Hiperprodukcija informacija u digitalnoj formi otežava proces procenjivanja i vrednovanja informacija i ovu činjenicu je neophodno imati u vidu kada pokušavamo da razumemo relaciju korisnik-sadržaj.” [2 p44]. Čarli Gir (Charlie Gere) određuje život u razvijenom svetu kao „društvo prezasićeno digitalnom tehnologijom” [3 p192] U, slobodno možemo reći, eri društvenih medija, vesti stižu do korisnika sa različitih strana, a tradicionalni ’’čuvari kapija’’, odnosno urednici i novinari, izgleda da više nisu neophodni. Slavne ličnosti i političari imaju direktan pristup milionima pratilaca. Bez adekvatnog proveravanja informacija, lako nasedamo na obmane koje onda mogu da postanu viralne.

Praktična i teorijska (stručna i naučna), razmatranja pojma ’’lažne vesti’’ su poslednjih godina sve brojnija, prateći, u izvesnom smislu porast popularnosti ovog termina, koji je do pažnje javnosti dospeo posebno nakon predsedničkih izbora u SAD-u, 2016. godine. Kolinsov rečnik (Collins

Dictionary) je odabrao upravo termin "lažne vesti" ("fake news") kao reč godine u 2017, jer je upotreba ovog termina te godine povećana 365%. Prema definciji ovog rečnika pod ovim pojmom se podrazumeva "lažna, često senzacionalna, informacija koja se širi pod maskom novinskog izveštavanja". [4]

Dakle, iako mnogi novinari i profesori novinarstva smatraju da ne postoje lažne vesti, već su vesti – ili vesti ili to nisu, većina istraživača je saglasna da su to zapravo netačne inofrmacije, koje mogu da obmanu čitaoce i navedu ih da poveruju da je vest koju čitaju tačna i kredibilna. Dok se neke netačne informacije/lažne vesti šire sa namerom da se zaradi novac od advertajzinga, neke su i posledica neobaveštenosti/pogrešne obaveštenosti ljudi/korisnika interneta koji lažne vesti potom šire na društvenim mrežama. Takođe, u studiji koju je sproveo Univerzitet Stanford (2019), [5 p3] navodi se da su istraživanja još 2016. pokazala da mladim ljudima u USA nedostaju osnovne veštine digitalne evaluacije, dok je istraživanje čiji su rezultati objavljeni 2019., takođe bilo zabrinjavajuće. Naime 2/3 učenika (od ukupno 3,446) nije moglo da prepozna i/ili objasni razliku između vesti i reklama i/ili oglasa, te je zaključeno da je neophodno da unapređenje digitalnih kompetencija mladih bude zasnovano na istraživanjima.

2 Onlajn građansko novinarstvo

Građansko novinarstvo, posebno u digitalnom medijskom okruženju ima veliki značaj za kritičko razmatranje medija, a ova forma participacije u javnom diskursu je posebno postala značajna u eri u kojoj je veća dostupnost/korišćenje mobilnih tehnologija dovelo do toga da veliki broj korisnika danas može da kreira informacije i šalje ih sa bilo kog mesta na planeti. Građansko novinarstvo bi se najšire moglo odrediti kao proces sakupljanja, pripremanja, širenja i analize vesti, informacija i drugih oblika novinarskog izražavanja, u kome javno mnjenje ima aktivnu i ključnu ulogu. Naravno, ideja da bi građani trebalo da participiraju u kreiranju informacija i da se uključe u novinarstvo nije nastala u informacionom društvu. Šejn Boumani Kris Vilis (Shayne Bowman, Chris Willis): „Namera ovog učešća javnosti je da obezbede nezavisne, pouzdane, tačne, relevantne informacije širokog raspona, kakve demokratija i zahteva.” [7 p9] Loren Hertel (Lauren Hertel) takođe uviđa da živimo u informacijski prezasićenom medijsko-društvenom kontekstu i navodi da u ovom trenutku jednostavno postoji previše informacija, kao i da se njihov broj svakodnevno povećava:‘’Ja sam preživjela rak. Idem na Internet i tražim informacije o raku, i ima ih mnogo. Dio toga su zaista valjani medicinski savjeti i dobri izvori, dok je dio jednostavno zastrašujući. Tu su informacije koje plaše ljude, tu su ljudi koji pokušavaju da prodaju zastrašujuće stvari žrtvama raka... Kako da pronađete ono što vam je potrebno? Na kraju sam se oslonila u velikoj mjeri na nekoliko novinara koji su pažljivo ispitali puno tih informacija i prikupili sigurne i dobre izvore i tekstove o ovoj temi…” [8 p170]. Novinar-građanin (amater) je vođen pre svega unutrašnjim porivom da saopšti okruženju nešto što smatra vrednim saopštavanja, prema ličnim kriterijumima raspoloženja/interesa, i u skladu sa veštinama koje poseduje. Odgovornost, uređivanje, autonomnost, originalnost i profesionalnost predstavljaju neke od osnovnih postulata profesionalnog bavljenja novinarstvom.

‘’U čemu je razlika između 11. 09. u Njujorku? U Njujorku je armija građana dostavljala medijima. Izveštaje očevidaca fotografije, ali sebe nisu videli kao građane-reportere. Bili su samo obični građani u vanrednoj situaciji. To nije bio slučaj sa blogerima, moblogerima i video blogerima 07. jula u Londonu. Četiri godine nakon 11. septembra ljudi su svesni svoje uloge - ne samo kao autora vesti, veći kao reporter i producenata. Bilo je to ozbiljno upozorenje dotadašnjim medijskim monopolima - o moći i domašaju interneta.” [8 p182].

Ova situacija bila jeo zbiljno ‘’upozorenje’’ dotadašnjim medijskim monopolima - o moći 0 i domašaju interneta, iako su građani-novinari ostali ’’meta’’ napada, uglavnom zbog nedostatka kvaliteta i sadržaja, zbog toga što pišu uglavnom iz hobija, te u tom smislu ne mogu da zamene profesionalne novinare, posvećene svom poslu svakodnevno.

3 Zašto je važno da naučimo da ‘’razotkrijemo’’ lažne vesti?

Savremeni čovek je uglavnom u velikoj meri okružen medijima, stoga možemo da kažemo da živimo u/na/sa medijima, te da na izvestan način i sami postajemo – mediji. Ili kako veoma dobro ukazuje Dalibor Petrović, kritički tumačeći medijski posredovano društvo i tvrdeći da su onlajn platforme za društveno umrežavanje zapravo finalni proizvod medijskog terora nad ljudima: „Mediji su svuda oko nas. Oni su nas naučili da se ništa ne sme propustiti sećanju i neizbežnom zaboravu ... i samo ono što je prikazano postaje istinito, dok sve ostalo zauvek nestaje u medijskom mraku. U strahu od zaborava mi mahnito beležimo svaki delić naših života." [9 p168]

Osim što podrivaju temelje novinarstva i demokratije, što mogu ozbiljno da ugroze nečiji život (čemu posebno svedočimo tokom pandemije korona virusa kada su izazivale veliki strah i paniku kod mnogih ljudi širom sveta), činjenica da poslednjihgodinaupravodruštvenemrežepostaju mesto gde se većinaljudiinformiše - a nečitajući kredibilne novinarske i uredničke tekstove – čini proces razotkrivanja dezinformacija još težim i izazovnijim. Društveni mediji igraju ključnu ulogu u potrošnji vesti, prema podacima iz avgusta 2022. godine, polovina odraslih u SAD bar ponekad dobija vesti sa društvenih medija, Fejsbuk nadmašuje sve druge sajtove društvenih medija. Otprilike trećina odraslih u SAD (31%) kaže da redovno dobijaju vesti sa Fejsbuka. [10 p2] Ekspert za komunikacije Barbara Alvarez (Barbara Alvarez) upozorava da ukoliko nemamo znanje da pravilno identifikujemo/prepoznamo lažne vesti, onda društvene mreže i veb sajtovi koji ih prenose, mogu lako da prevare ‘’neobučeno’’ oko, da poveruje da je to verodostojan izvor, jer je njihov glavni cilj – obmana. Ako pogledamo predsedničke izbore u USA, 2016. godine, videćemo da su se upravo društveni mediji pokazali kao pogodno tlo za širenje dezinformacija i uticaj na birače i procese donošenja odluka. [11 p173]

Šošana Zubof (Shoshana Zuboff), doba u kome živimo naziva dobom nadzornog kapitalizma i uviđa da, iako se trudimo da nazremo obrise digitalne ere, mi i dalje zapravo ’’bauljamo’’ po mraku, a nadzorni kapitalizam ide u sasvim drugom smeru u odnosu na rani digitalni san" [12 p19]. Menja se i utiče na ponašanje ljudi, uglavnom bez njihovog znanja, brižljivim formulisanjem poruka koje će na njih imati najviše uticaja. Dakle, važno je pitanje kako savremeni medijski i tehnološki kontekst utiču na slobodu govora i na demokratske procese i sisteme vrednosti u društvu, kakav je zaista potencijal mreže za otvorenu diskusiju i dijalog. Nadzorni kapitalizam je u osnovi ekonomski sistem koji daje za pravo da se naše iskustvo, želje, osećanja…prepoznaju, prikupljaju, obrađuju i da nam se onda ponude proizvodi koji će biti bazirani na tome. Drugim rečima, neovlašćeno se trguje na tržištu našeg budućeg ponašanja.

I na kraju, etička pitanja imaju fundamentalan značaj za ljudski napredak. U tehnološki sve razvijenijem svetu, u kome i dalje brojni stanovnici planete Zemlje nemaju ni osnovne uslove za život, a kamoli internet, i u kom se vodi fundamentalna borba između profita i društvenog (i individualnog) dobra, ko o etičkim pitanjima treba da brine? Da li o tome šta je za ljude dobro treba da razmišlja inženjer koji kreira aplikacije koje prikupljaju i obrađuju naše podatke, neko iz marketinga, država, sami korisnici? Da li nas tehnologije menjaju ili mi, izmenjeni, menjamo njih?

4 Zašto mnogi veruju u lažne vesti?

Doktor S. Šijam Sundar (S. Shyam Sundar) istražuje društveno-psihološke efekte različitih digitalnih medija, od veb sajtova i društvenih mreža do mobilnih telefona, robotike i internet stvari. ukazuje na značaj jedne od ključnih karakteristika lažnih vesti/dezinformacija – na psihološki značaj izvora. Sundar analizira tehnološke i psihološke procese koji objašnjavaju kako su se pojavile lažne vesti, zašto su promenile novinarstvo i zašto im verujemo, zbog čega su onlajn mediji pogodni za njihovo ’’bujanje’’ i kako se od njih možemo zaštiti u budućnosti [13]. Drugim rečima, razmatra šta nam psihologija govori o dubljoj motivaciji koja utiče na našu pažnju i ponašanje i koja nas ponekad čini samo zupčanicima u viralnoj mašini dezinformacija. Publika je ključan element u distribuiranju dezinformacija, a digitalni mediji su omogućili publici da postane kreator i distributer različitog sadržaja – audio, video, vesti, linkova, informacija itd. U tolikom obimu da ne bismo pogrešili ako bismo tvrdili da mnogi danas sebe i vide kao novinare, koji im više nisu neophodni kao posrednici između izvora i čitalaca. Sa druge strane, upravo mogućnost da svako može da kreira i prenosi informacije, osim pozitivnih, u smislu demokratizacije određenih procesa, ima za posledicu i širenje različitih vrsta sadržaja putem interneta, što predstavlja ’’plodno tlo’’ i za širenje lažnih vesti. Sandarova istraživanja pokazala su da ljudi ne primećuju razliku između lažnih i pravih vesti, odnosno između profesionalnog izvora i izvora laika/amatera, kao i da im to nije mnogo važno, odnosno ne bave se time. Takođe, ljudi najviše veruju preporuci prijatelja i vestima/pričama koje su se svidele drugim korisnicima. Dakle, kritički odnos prema medijskim sadržajima, poznavanje načina na koji funkcionišu mediji i razumevanje suštinskog značaja izvora informacija, često – izostaju. Informacija koja dolazi do krajnjeg korisnika uglavnom potiče iz više izvora, koje je potrebno znati da bismo mogli da procenimo da li je vest koju smo dobili tačna ili ne. Ono što se često zaboravlja, a veoma je važno u profesionalnom novinarstvu, je značaj neposrednog izvora. Društveni medij može da bude legitiman izvor informacija ukoliko ima odgovornog urednika i profesionalne novinare. Na internetu imamo čitav lanac izvora i nije jednostavno razotkriti odakle je vest potekla, ko ju je ’’pustio’’. Čitaoci obraćaju pažnju na slojevitost izvora i zanimaju se za njih jedino ukoliko im je priča zaista važna, inače se retko i pitaju o kredibilnosti izvora vesti koje konzumiraju. Sve češće čujemo da ljudi kažu da su nešto pročitali na nekom agregatoru (Comcast, Verizon itd.) ili na Fejsbuku, a to nisu relevantni i pouzdani izvori za informisanje. Posebno, imajući u vidu činjenicu da aplikacije koje koristimo prilagođavamo sebi, da smo manje kritični prema sadržajima koje smo sami odabrali da pratimo i delimo (gde vidimo sebe kao izvor), kao i prema sadržajima/vestima koje dele naši prijatelji. Koji su efekti? Sandar zaključuje da upravo iz ovih razloga, između ostalog, ’’nasedamo’’ na lažne vesti – podcenjivanje profesionalnih novinarskih vesti, ignorisanje problema višeslojnih izvora, zato što prijatelji ili mi sami – postajemo izvor. Način na koji primamo informacije i i kvalitet informacija koje primamo značajno utiču na proces donošenja odluka.

Sandar navodi da svojevrstan proces selekcije informacija, bile one lažne ili ne, odigrava se zahvaljujući sledećim psihološkim konceptima:

1. Teorije zavere – određeni ljudi imaju dobro razrađene teorije šta je u pozadini nekih pojava i događaja, a koje nisu zasnovane na činjenicama, te su skloniji da poveruju u informacije koje idu u prilog teoriji u koju veruju.
2. Potvrda pristrasnosti – ljudi imaju sklonost da traže potvrdu za ono u šta već veruju, pa tako i u medijima tražimo one priče koje potvrđuju naša uverenja.
3. Selektivna izloženost – u mnoštvu medija koji su nam na raspolaganju pravimo selekciju onih koje ćemo pratiti, opet u skladu sa našim uverenjima i interesovanjima.

4. Selektivna percepcija – ljudi percipiraju samo one aspekte priče/sadržaja koji su u skladu sa njihovim uverenjima.
5. Filter mehurići – sve gore navadeno ima za posledicu da funkcionišemo u okviru svojih ’’filter mehurića’’, gde mi sami selektivno filtriramo informacije u skladu sa našim uverenjima i pogledom na svet.

Zašto ”bujaju” lažne vesti?

1. Prezasićenost informacijama
2. Multiplikovani izvori vesti – nemogućnost razlikovanja kredibilnih izvora
3. Druge korisnike vrednujemo više od profesionalnih izvora
4. Ja kao izvor – manji pregled informacija
5. Za informacionu pismenost je kritična ’’source literacy’’ – pažnja fokusirana na izvore, na motivaciju i pristrasnost izvora, uviđanje neophodnosti proveravanja više zvora itd.

I na kraju, samo da kratko pomenemo dipfejk (*deepfake*) tehnologiju (kojom mogu da se generišu lažni video snimci), koja se kontinuirano unapređuje i sadržaj generisan modelima mašinskog učenja postaje sve uverljiviji. Prema dostupnim informacijama, dipfejk sadržaji prvi put su se pojavili 2016. godine kada je jedan Reddit korisnik počeo da implementira lica poznatih ličnosti u pornografske sadržaje. Praksa se uskoro proširila i na druge sadržaje, a ovako proizvedeni video zapisi izgledaju sve bliži realnosti i sve je lakše poverovati da su ljudi rekli ili uradili stvari koje nisu.

5 Kako prepoznati/dekonstruisati lažne vesti – primeri iz prakse

Invazija lažnih vesti podriva medijsku pismenost, ali i same medije dovodeći nas do dileme - da li uopšte i kako možemo danas da verujemo medijima? U tzv. tradicionalnim medijima znalo se gde stoje tabloidi (na kom mestu u supermarketu ili na trafici) a gde relevantni izvori informacija. Sada Njujork tajms i tabloid, ili Vreme i Informer, koegzistiraju na internetu. Naizgled potpuno ravnopravno. Čini se da nam je sve više neophodna jedna nova vrsta medijske pismenosti, za onlajn svet, da bismo razlikovali šta je šta, odnosno informaciju od dezinformacije, kada nam sve stiže u istom fidu, na npr. fejsbuku ili tviteru. Potrebno je više istraživanja i edukacije – da učimo kako da čitamo vesti, odnosno kako da prepoznamo lažne vesti i dođemo do pravog smisla u ovom, u tom pogledu, sve zagađenijem i toksičnijem informacionom okruženju. Drugim rečima, potrebna su nam znanja koja su nekada bila potrebna samo novinarima, i koji su već na prvoj godini studija učili da treba da imaju najmanje dva pouzdana izvora za jednu informaciju, da upoređuju različite članke, prate kontekst i ’’trag novca’’.

Prema Vodiču za lažne vesti i dezinformacije, Novosadske novinarske škole, možemo da prepoznamo/dekonstruišemo kroz šest koraka verifikacije [14 p 12-13]:

1. Poreklo - da li je sadržaj originalan?
2. Izvor - ko je postavio sadržaj na internet?
3. Datum nastanka - kada je nastala fotografija (foto-forenzika)?
4. Gde je sadržaj zaista nastao (gugl mape kao koristan alat)?
5. Kako stupiti u kontakt sa autorom?
6. Da li postoji dozvola za objavljivanje?

Za prepoznavanje lažne vesti potrebno je primeniti alate za proveru informacije u skladu sa sledećim pravilima: [15]

1. Provera vebsajta na kojem smo pronašli informaciju. Kartice „pravni podaci“ ili „o nama“ generalno pomažu u određivanju vrste sajta koji konsultujemo (blog, humoristični sajt itd). Ako je u pitanju društvena mreža, proveriti nalog koji je podelio/objavio informaciju (parodijski, institucionalni i sl).
2. Pogledati datum objavljivanja informacije. U današnje vreme, informacija brzo stari ili je u međuvremenu već demantovana/proverena.
3. Proveriti identitet autora informacije (novinar? stručnjak za datu temu? građanin?). Zapitajte se koji je njegov/njen cilj - da nas informiše, podeli svoj stav ili da manipuliše nama?
4. Istražiti poreklo informacije - gde je prvo objavljena? Na internetu se informacije često dele, objavljuju, ali i deformišu, izvlače iz konteksta ili tumače. Zato je važno otkriti odakle dolazi informacija.
5. Postaviti prava pitanja, biti radoznao i sumnjati. Kritičko razmišljanje je najefikasnije sredstvo kojim raspolažemo u zaštiti od lažnih vesti i teorija zavere.

6 Internet adrese sa primerima iz prakse

U završnom delu rada prezentovan je izbor organizacija i relevantnih adresa na internetu gde se, uz brojne primere i postepeno, korak po korak, može naučiti kako da ‘’raskrinkamo’’ lažne vesti.

1. Raskrinkavanje.rs
Ovaj projekat portala KRIK, koji se već više godina hrabro i uspešno bavi razotkrivanjem korupcije i kriminala, pre svega ima za cilj borbu protiv medijskog dezinformisanja. Predstavljanje dezinformacija kao činjenice vide kao sve dominantniji trend. Raskrinkavanje beleži objavljivanje lažnih i neproverenih vesti, kao i razne oblike kršenja pravila novinarske profesije koja bi trebalo da osigura objektivno, istinito i nezavisno informisanje javnosti u Srbiji. Uz raskrinkavanje lažnih vesti, na portalu se može naći i edukativni materijal, kao i zanimljivi kvizovi koji mogu da unaprede naše znanje iz ovih oblasti. Zajedno sa kolegama iz regiona osnovali su regionalnu mrežu medija koji otkrivaju lažne vesti SEE Check. [16]

2. Istinomer.rs
Istinomer je osnovala organizacija CRTA, bavi se proverom činjenica kroz ocenjivanje izjava političara, donosioca odluka i javnih zvaničnika, prema kriterijumima – doslednosti, istinitosti i ispunjenja obećanja [17]. U stvaranju sajta učestvuju i čitaoci, koji mogu da predlože temu za istraživanje. Za temu rada je važno da dodamo i da je Istinomer u julu 2020. godine postao zvaničan partner Fejsbuku u suzbijanju dezinformacija na ovoj društvenoj mreži.

3. Fakenews.rs tragač
FakeNews tragač pokrenula je Novosadska novinarska škola 2017. godine, a posvećen je borbi protiv dezinformacija u medijima koji sadržaje objavljuju na srpskom jeziku. Među brojnim zanimljivim tekstovima na ovoj internet adresi možemo da izdvojimo tekst pod naslovom ‘’Evergrin fotografije praznih rafova u Evropi: prizori nestašice za svaku priliku’’ u kome se na duhovit, ali kritičan način piše o tekstovima domaćih medija o nestašicama hrane u Evropi i USA, koje su pratile fotografije praznih rafova koje su nisu autentične [18]. Rad FakeNews Tragača [19] funkcioniše prema sledećoj metodologiji:

1. Predmet analize
2. Prijava sadržaja
3. Analiza sadržaja
4. Redakcijski rad
5. Ispravke grešaka
6. Nepristrasnost u radu
7. Medijska i naučna pismenost

U kontekstu ovog rada, posebno je važno skrenuti pažnju na elektronski Vodič za borbu protiv lažnih vesti [14], koji su pripremili novinari Fejknjuz tragača, a koji sadrži brojne korisne savete i sugestije u vezi sa verifikacijom sadržaja sa interneta. Kako utvrditi originalnu verziju fotografije? Kako prepoznati da li je i na koji način snimak izvučen iz konteksta, kako prepoznati fotomontažu? itd.

4. SHARE Resurs centar
Centralizovani portal SHARE Fondacije, SHARE Resurscentar, na kome se nalaze tekstovi, publikacije, stručne analize, baze znanja, snimci predavanja, kao i drugi sadržaji i materijali SHARE Fondacije. Objedinjeni portal omogućava jednostavan i pregledan pristup materijalima o lažnim vestima i različitim vrstama sadržaja (video materijali, publikacije, itd). Osim analitičkih tekstova, istraživanja o onlajn medijima u Srbiji, na ovojadresi se mogu pronaći i sadržaji o temama koje nisu u dovoljnoj meri pristune u javnosti, a značajne su, kao što su na primer – odnos tehnologije i jezika, sloboda izražavanja na internetu, digitalna bezbednost i demokratija itd. [20]

5. Projekat NEZAVERENI u BG i L.A. / dekonstrukcija teorija zavere i lažnih vesti
“Rex Preprič” je pokušaj da se putem prepričavanja naučnih radova iz određene oblasti dostignuća nauke približe široj publici. Kulturni centar Rex je, uz pomoć konsultanata i saradnika iz SAD i Srbije, istraživao fenomene teorija zavere i lažnih vesti sa različitih teorijskih i praktičnih-aktivističkih tačaka gledanja i u različitim profesionalnim, medijskim i aktivističkim poljima. Projekat kroz bavljenje ovim fenomenima kritički istražuje i objašnjava rastuće nepoverenje u bazična dostignuća i temelje nauke, kulture, medija i tehnologije. [25]

6. Stranicu “Medijska pismenost” pokrenuo je tim koji uređuje i vodi platformu Raskrinkavanje.ba, posvećenu proveri tačnosti medijskih informacija (*fact-checking* medija), sa posebnim fokusom na online medije. Kakvi mediji treba da budu, kako se odbraniti od manipulacija, oblici manipulacija i kome se obratiti ako ih uočite, uz zaseban link za resurse, gde se može pronaći raznovrstan i koristan materijal iz ovih oblasti. [26]

7. AFP je pokrenuo servis za digitalnu proveru u Francuskoj 2017. godine i postao je vodeća *fact-checking* organizacija u svetu, sa specijalizovanim novinarima u zemljama od Sjedinjenih Američkih Država do Mijanmara. Od jula 2020. AFP je započeo *fact-checking* na srpskom. ‘’Pored tradicionalnih novinarskih veština, za proveru onlajn informacija koristimo određeni broj jednostavnih alatki, kao i zdrav razum i mnogo opreza.’’ [27]

8. Platforma TALMIL stavlja na raspolaganje pedagoške sadržaje na jezicima koji se govore u Albaniji, Bosni i Hercegovini, na Kosovu, u Severnoj Makedoniji, Crnoj Gori i Srbiji, za nastavu i učenje medijske i informacione pismenosti. Iako na portalu piše da su materijali namenjeni za

srednjoškolski uzrast, izvanredan materijal, na više jezika i namenjen nastavnicima, mladima i medijima, sa odličnim primerima i velikim brojem odličnih resursa i alata, zanimljivim edukativnim kvizovima, predstavlja izvanredan prostor za učenje/informisanje u vezi sa medijskom pismenošću, medijima i lažnim vestima. [28]

7 Zaključne napomene

Umeće pažljivog čitanja i promišljanja pročitanog, veština prepoznavanja dezinformacija i manipulacija koje, kako smo videli, u sve većem broju, dolaze sa različitih strana i sa različitim ciljevima, izvesna odgovornost prema pročitanom/deljenom/preporučenom sadržaju, kritički pristup medijima i tehnologijama i upućenost u njihov uticaj na čoveka – predstavljaju osnovu za dublje razumevanje digitalnog, medijskog okruženja. Besplatni, profesionalni sadržaji koji su nam dostupni onlajn predstavljaju dragocene izvore i podršku ovom procesu učenja i razotkrivanja. Jer, na kraju, odabrati pozudane izvore i prave informacije nije više jednostavno ni profesionalcima ni amaterima …

Zahvalnica

Zahvaljujem se Elektrotehničkom fakultetu Univerziteta u Beogradu i Organizacionom odboru na pozivu da učestvujem na konferenciji “Primena Slobodnog Softvera i Otvorenog Hardvera”, jedinstvenoj po svom konceptu i temama koje ima u fokusu.

Literatura i veb izvori

[1] Allen J. at all. Evaluating the fake news problem at the scale of the information ecosystem. Sci. Adv. 6, eaay3539, 2020.
[2] Vučenović T. Biblioteka kao nosilac participativnih praksi u kulturi u kontekstu informacionog društva, doktorska disertacija, Univerzitet u Beogradu, Filološki fakultet, 2016, COBISS.SR-ID – 48692751 http://phaidrabg.bg.ac.rs/o:14689
[3] Gir Č. Digitalna kultura.[A.L. Todorović, prev.]. Beograd: Clio, 2011. [Gere C. Digital culture. 2002].
[4] Collins online dictionary and reference resources, Collins COBUILD Advanced Learner’s Dictionary, Harper Collins Publishers, [Internet], citirano 2022, dostupno na https://www.collinsdictionary.com
[5] Breakstone J. at all. Students’ civic online reasoning: A national portrait. Stanford History Education Group & Gibson Consulting, [Internet], objavljeno2019, navedeno 2022, dostupno na https://purl.stanford.edu/gf151tb4868
[6] Valencia College library, [Internet], navedeno 2022, dostupno na https://libguides.valenciacollege.edu/c.php?g=612299&p=4251522
[7] Bowman S. and Willis C., We Media | How audiences are shaping the future of news and information, The Media Center at American Press Institute, 07. 2003: objavljeno 2022, dostupno na http://www.hypergene.net/wemedia/download/we_media.pdf
[8] Vučenović T. Građansko novinarstvo u digitalnom 21. veku. Kultura br. 132. Beograd: Zavod za proučavanje kulturnog razvitka, 2011.
[9] Petrović D. Društvenost u doba interneta, Novi Sad:Akademska knjiga, 2013.
[10] Social Media and News Fact Sheet, News consumption on social media, Survey of U.S. adults conducted July 18 – Aug. 21, 2022, Pew research center [Internet], dostupno na https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/
[11] Bradshaw S. at all. Sourcing and Automation of Political News and Information over Social Media in the United States, 2016-2018, Political Communication, 37:2, 173-193, 2020. DOI: 10.1080/10584609.2019.1663322
[12] Zubof Š. Doba nadzornog kapitalizma, [J. Petrović. prev.]. Clio: Beograd 2020. [Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. 2019].
[13] Shyam Sundar S. The Conversation, [Internet], objavljeno 08.12.2016., navedeno 2022, dostupno na https://theconversation.com/why-do-we-fall-for-fake-news-69829

[14] Vodič za borbu protiv lažnih vesti, Novosadska novinarska škola [Internet], objavljen 31.01. 2018., citirano 2022, dostupan na https://issuu.com/novinarska-skola/docs/fake_news_vodic
[15] Pedagoška platforma za nastavu i učenje medijske i informacione pismenosti na zapadnom Balkanu, Talmil.org, [Internet], navedeno 2022, dostupno na https://talmil.org/sr/5-kako-prepoznati-lazne-vesti/
[16] Portal Raskrikavanje, Mreža za istraživanje kriminala i korupcije KRIK, [Internet], navedeno 2022, dostupno na https://www.raskrikavanje.rs/all.php
[17] Istinomer, CRTA, [Internet], navedeno 2022, dostupno na https://www.istinomer.rs/
[18] Fakenews tragač, Novosadska novinarska škola [Internet], objavljeno 09.05.2022, navedeno 2022, dostupno na https://fakenews.rs/2022/05/09/evergrin-fotografije-praznih-rafova/
[19] Fakenews tragač, Novosadska novinarska škola [Internet], citirano 2022, dostupno na https://fakenews.rs/o-nama/
[20] SHARE resurs centar, SHARE fondacija, [Internet], navedeno 2022, dostupno na https://resursi.sharefoundation.info/sr/o-nama/
[21] Kulturni centar REX, Projekat ‘’Rex preprič’’, Fond B92, [Internet], navedeno 2022, dostupno na http://rex.fondb92.org/sr/rex-prepric-nezavereni-a-prepricavani-eid-2400.1.213.html
[22] Platforma Raskrinkavanje.ba, projekat Medijska pismnost,UG “Zašto ne”, [Internet], navedeno 2022, dostupno na https://medijskapismenost.raskrinkavanje.ba/
[23] Servis za digitalnu proveru, AFP-Agence France-Presse [Internet], citirano 2022, dostupno na https://cinjenice.afp.com/list

Dodatak – sajtovi posvećeni dezinformacijama, teorijama zavere i proveri vesti na stranom jeziku

1. BBC, https://www.bbc.com/blueroom, https://www.bbc.co.uk/beyondfakenews/, https://www.bbc.co.uk/academy/, navedeno 2022.
2. Biblioteka Gradskog Univerziteta u Njujorku, https://library.csi.cuny.edu/c.php?g=452334&p=7765919, navedeno 2022.
3. Biblioteka Valensia koledža na Floridi, https://libguides.valenciacollege.edu/c.php?g=612299&p=4251645, navedeno 2022.
4. Interdisciplinarna međunarodna mreža sa ciljem dubljeg i sveobuhvatnijeg razumevanja teorija zavere, https://conspiracytheories.eu/, navedeno 2022.
5. Vašington post, https://www.washingtonpost.com/politics/2022/04/06/ukraine-biolab-conspiracy-theory-quickly-went-viral-it-took-weeks-pinpoint-source/, navedeno 2022.
6. Spotting misleading or fake news, https://cdnapisec.kaltura.com/html5/html5lib/v1.8.9/mwEmbedFrame.php/p/811482/uiconf_id/21849131/entry_id/1_gxg0lb6a?wid=_811482&iframeembed=true&playerId=kaltura_player&entry_id=1_gxg0lb6a, navedeno 2022.

Napomena

Prevodi citata preuzetih iz izdanja na engleskom jeziku su autorkini.

PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7096251

Video analysis using open-source FFmpeg tool and selection of codecs

Ana Gavrovska

University of Belgrade - School of Electrical Engineering,
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
e-mail addresses: anaga777@gmail.com, anaga777@etf.rs

ABSTRACT

This paper presents a brief overview of a free and open-source multimedia framework called FFmpeg with emphasis on video analysis and selection of codecs (CoDec - Coding-Decoding/Compression-Decompression). Filtering, trimming, transcoding, streaming, multiplexing, and other possibilities are available using this tool which is commonly applied for multimedia handling, telecommunications and video traffic monitoring, forensics and, generally, practical projects. Commands can be used in the command line interface with input and output files, selected actions and parameters. There are several available FFmpeg libraries for developers, where some of them enable implementation of relatively novel encoding/decoding solutions. Quality conditions may be set differently according to specific control modes. This cross-platform and open solution is beneficial for understanding both basic and complex approaches for video and multimedia manipulation, and is relevant for educational purposes. Regardless of the framework, codecs are evolving towards the new generation and some of the general trends in this domain are presented here. Finally, attention is given to some of the licensing considerations.

Keywords: open-source, ffmpeg, video, codecs, multimedia compression, presets

1 Introduction

For content production and distribution, it is important to cope with a large amount of media [1-4]. During the last decades, video has become dominant in digital communication and a large part of global telecommunication traffic. Video production, visual advertising, sharing platforms, video conference, unified communication and collaboration products, immersive media, have led to the need of efficient video delivery models. Video demand on the internet has grown significantly to 80% of all traffic in 2021 compared to 67% in 2016 according to Cisco report [5]. Video web services have skyrocketed in the period of Covid-19 pandemic [6]. Thus, new media technologies have caused revising issues of content delivery

congestion.

Encoding and decoding video and audio streams, i.e. codecs, and other relevant tools for developers enable practical implementations [1-2, 7-8]. In addition to coding-decoding, usually compression-decompression tasks are performed to deal with bandwidth bottleneck and limitations of storage. Lossless compression approaches are developed having in mind retaining all information without losing media quality, while mostly used lossy ones make the trade-off, meaning transmission channel or storage space is less occupied with decent approximation of encoded data. Methods for editing, control, quality evaluation, encapsulation, and many other processes are related to coding and media handling. Different encoded data can be combined by multiplexing and data elements (streams) should coexist within container formats. This provides interoperability within ICT (Information and Communication Technology) systems that is essential. Benefits of IP (Internet Protocol) communications provided flexibility and scalability in media distribution [9, 10].

Free and open-source frameworks and projects are necessary for providing basics of understanding multimedia technologies that are becoming more complex and of wide-ranging interest among engineers, developers, and stakeholders. FFmpeg is one of the leading multimedia framework available for implementation of codecs, processing, streaming, multiplexing and similar media handling options and manipulations with high portability over different platforms, environments, and machine architectures [11, 12].

In this paper several main aspects related to modern codecs in general and FFmpeg usage are discussed. The paper is organized as follows. Brief description of popular formats and main FFmpeg framework characteristics can be found in Section II after the introduction. Basic tools for treating media are described, such as setting video quality according to predefined parameters and chosen codec. Some of the examples, which include preset options, are presented in Section III. Video streaming is part of our daily lives, and adoption of new standards and licensing efforts are also important issues to sum things up in Section IV. Finally, Section V is dedicated to conclusion.

2 Codec tools

A large number of specific techniques nowadays make a solution satisfying for media delivery. Committees, corporations, communities, and alliances design variety of standards and formats. For example, MPEG stands for Motion Picture Experts Group, and is dedicated to media coding and compression, transmission and formats. In video and multimedia technology field MPEGx and H26x standards have taken the central place over the years [1, 13].

In order to analyze some of the codec tools, FFmpeg (Fast Forward MPEG) is applied here. It represents one of the most powerful multimedia environments with options suitable for many users and developers working in different fields that include video and multimedia processing. It supports different formats from old-fashioned ones to novel solutions, where most video programs include this solution as a part of their processing pipeline [10, 11]. Cross-platform libraries for processing like OpenCV can also use the FFmpeg results as backend for recording, media converting and audio and video streaming, and this backend usage is not the rare case [14]. This practical project is free to use and one of the most deserving for this are developers Fabris Belard and Michael Niedermaier, and many other

project participants.

2.1 FFmpeg tools and libraries

FFmpeg includes three basic tools, i.e. commands that can be used in working with media content: *ffmpeg*, *ffprobe* and *ffplay*. By using *ffmpeg* command, some of the popular codec can be applied or format conversion can be performed. On the other hand, the *ffplay* command serves to reproduce files in order to visualize the content. The professionals parse video content and monitor traffic quantities via *ffprobe* command. FFmpeg comes with libraries: *libavutil* (intended for simplifying programmable approaches and routines), *libavcodec* (containing encoders and decoders), *libavformat* (dedicated to multiplexing and demultiplexing), *libavdevice* (containing input and output devices for grabbing and rendering), *libavfilter* (holding filters for multimedia needs), *libswscale* (for scaling and conversion between color and pixel based formats), and *libswresample* (for resampling, rematrixing and sample format conversions). FFmpeg software may be useful in overcoming the need of special hardware in order to use main codecs and formats especially the ones for distribution over the Internet. The project has technical documentation, automated testing environment (fate), bug tracker, and wiki page. The main licence is GPL (GNU General Public License) or LGPL (GNU Lesser GPU) [11].

The commands are easily applied in command line (CLI - *Command Line Interface/Interpreter*) with chosen input and output files, selected actions and parameters. In Fig. 1 typical video processing that includes demultiplexing (demux) of input file, decoding, encoding, and multiplexing (mux) obtaining output file. The media content is kept in containers. Input and output files can be: mpeg, avi (audio video interleaved), m4v that is similar to mp4 (MPEG-4 Part 14), etc. Some of the codec examples are also given in Fig. 1 [11, 12]. Extending the framework is possible as well [15].

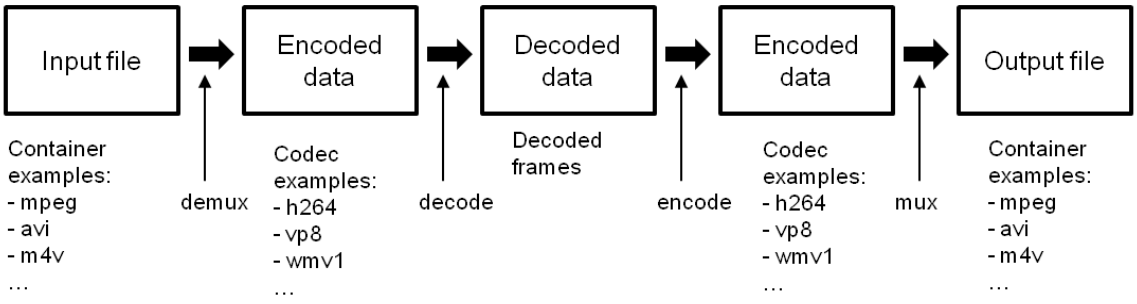


Figure 1. Typical video processing.

2.2 Video coding steps and standards

In video coding there are several main steps including: partitioning, prediction, data transformation and quantization, filtering improvements and entropy coding, as can be seen in Fig. 2. Each new standard brings novel advancements, like more partitions and larger blocks (superblocks), more intra and inter predictions, functions, filters and introducing AI (*Artificial Intelligence*) to coding and compression [6, 13, 15].

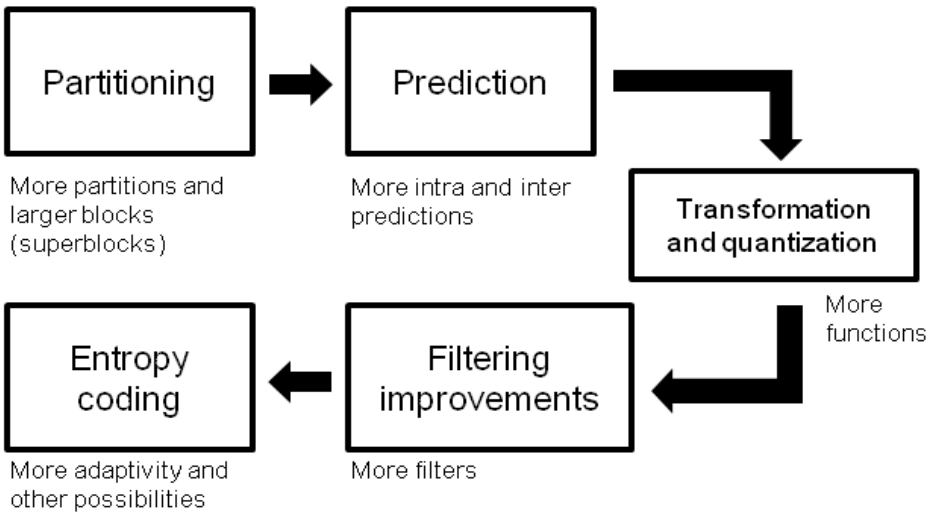


Figure 2. Illustration of main video coding steps and advancement.

The fundamentals in streaming are made by MPEG-2 standard that was popular in practical implementations like broadcasting. This practical usefulness has been continued by MPEG-4 by introducing additional enhancements. H.264 or Advance Video Coding (AVC) standard (MPEG-4 Part 10) is a well known standard from 2003 found by ITU-T and ISO/IEC. AVC is still in force for generic audiovisual services [16]. One of the most popular formats is still mp4 (MPEG-4 Part 14) representing digital multimedia container format for data encapsulation. A new standard is developed after H.264/AVC called H.265 or HEVC (High Efficiency Video Coding) or MPEG-H Part 2. HEVC is initially published in 2013 as the new generation standard. It has offered bitrate reduction of thirty to fifty percent to achieve comparable quality to H.264 [17]. Unfortunately, it has not proved to be the right solution as its predecessor despite its technical characteristics.

In 2013 Google released VP9 video coding format. Its development is based on previous VP8 standard (initally On2 technologies) and is intended for internet media delivery [18-20]. It has similar quality ratings with HEVC but with different implementation and traffic effects. Google's VP9 has been supported by popular web browsers and video platforms. In 2018 AOMedia (Alliance for Open Media) developed AV1 standard (AOMedia Video 1), and united top tech leaders for the next generation of media delivery over internet. The developed AV1 is considered a general-purpose open and royalty-free video coding format, and a successor of VP9 [18].

When efficient delivery over Internet is the target, the most common web codec formats and corresponding containers listed by Mozilla are presented in Table 1 [21]. Depending on source video format and set configuration parameters different coding results can be obtained. Supported codecs and formats by FFmpeg can be listed by using adequate commands [11]. This is shown in Fig. 3, where explanation of abbreviations is also given.

Table 1. Popular codecs and containers [21].

Codec	Full codec name	Container support
AV1	AOMedia Video 1	MP4, WebM
AVC (H.264)	Advanced Video Coding	3GP, MP4
H.263	H.263 Video	3GP,
HEVC (H.265)	High Efficiency Video Coding	MP4
MP4V-ES	MPEG-4 Video Elemental Stream	3GP, MP4
MPEG-1	MPEG-1 Part 2 Visual	MPEG, Quick Time
MPEG-2	MPEG-2 Part 2 Visual	MP4, MPEG, Quick Time
Theora	Theora	Ogg
VP8	Video Processor 8	3GP, Ogg, WebM
VP9	Video Processor 9	MP4, Ogg, WebM

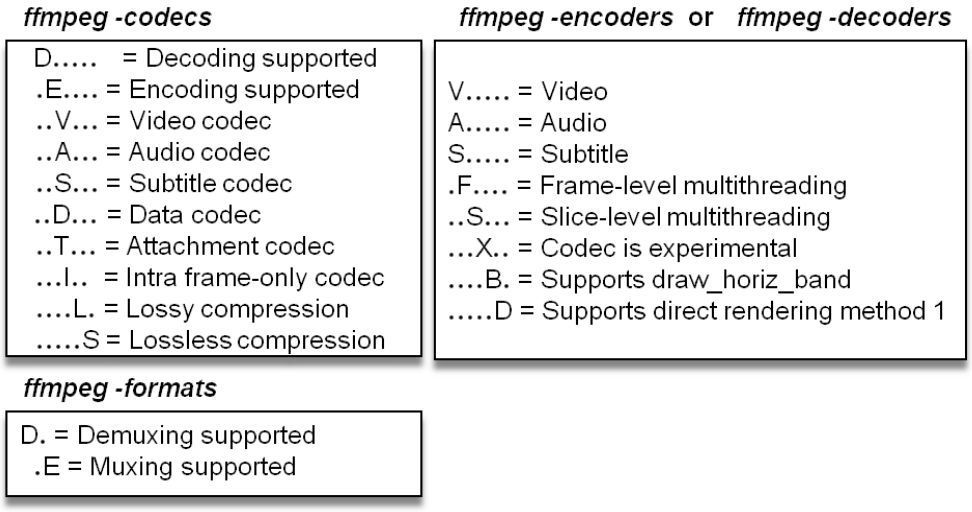


Figure 3. Commands for listing supported codecs and formats.

3 Use case for media transcoding and preset options

Video (and audio) can be manipulated by (trans-)coding and controlling its quality. This can be done in many ways like setting constant factor quality, buffer size, or using constant, constrained or variable bit rate [11]. In Fig. 4 commands applied here are presented. Firstly, audio is removed, and then constant rate factor (crf) is selected as a model for controlling output. The option is available for popular codecs and keeps the output quality level by rate control method. For VP9 and AV1, the crf usually takes values from 23 to 63, and in the case popular codecs x264 and x265 it matches quality ranges from 19 to 41 [22]. Lower crf values correspond to higher quality. The parameter crf chosen here is 35, and *libsvtav1* which represents the AV1 codec or to be more precisely SVT-AV1 (Scalable Video Technology for AV1) codec [23, 20]. Video of about ten minutes length (19036 frames) and 30fps [24] is tested according different preset options (codec Lavc59.39.100 libsvtav1). A preset is a collection of options that can provide certain encoding speed that slower preset may provide

higher quality per filesize with more time needed to encode [12]. Similar preset options are available for AVC and HEVC. The preset options are focused on speed and codec complexity. In the case of SVT-AV1 in Fig. 4 preset option (value 10) is shown with no additional tuning applied here. Supported preset options ranges from 0 to 13 with 13 for debugging and higher speed for higher preset value.

```
ffmpeg -i av_input -c copy -an  v_file
ffmpeg -i v_file  -c:v libsvtav1 -crf 35 v_output
ffmpeg -i v_file  -c:v libsvtav1 -preset 10 -crf 35 v_output
```

Figure 4. Examples of ffmpeg commands.

The obtained results for SVT-AV1 and five preset options are given in Table 2, without direct measuring the video quality, and there seems that there are no significant difference after reproduction to a standard viewer under common circumstances (ten volunteers using LED 23" monitor). On the other hand, time or speed needed for transcoding is quite different. Here, when no preset option is applied, as in Fig. 4, the same results as with preset 10 are obtained. The lowest filesize is for preset option 8, as shown in Table 2.

Table 2. SVT-AV1 results for different coding speed.

No.	Preset option	fps	Lsize	bitrate	speed
1	12	69	151584kB	1957.1kbits/s	2.29x
2	10	40	141200kB	1823.0kbits/s	1.32x
3	8	17	136125kB	1757.5kbits/s	0.583x
4	6	6.3	145011kB	1872.2kbits/s	0.212x
5	4	1.5	147742kB	1907.5kbits/s	0.0485x

Table 3. H.264 results for different coding speed.

No.	Preset option	fps	Lsize	bitrate	speed
1	ultrafast	199	645603kB	8335.3kbits/s	6.63x
2	superfast	116	249509kB	3221.7kbits/s	3.87x
3	veryfast	96	143810kB	1856.9kbits/s	3.21x
4	faster	58	169341kB	2186.6kbits/s	1.92x
5	fast	50	179782kB	2321.4kbits/s	1.68x
6	medium	43	178491kB	2304.7kbits/s	1.44x

Number of frame per second (fps) for output is 30fps, but for processing the video it varies. When applying the preset option for the libx264 and crf 25, without tuning, from medium to ultrafast, higher speeds are obtained, as in Table 3. Medium is default preset option and it may range from veryslow to ultrafast. For the settings here, the lowest file size and bitrate are obtained for preset option veryfast. The quality difference exists, meaning often higher quality for slower coding [25]. This trend is illustrated in Fig. 5 based on VMAF (Video Multimethod Assessment Fusion) scores. One may still consider that there is no significant difference to a standard viewer in this 1080p case. So, preset default option may respond to higher speed at the price of slightly decrease of quality. Some of the details of the presets can be found in [26].

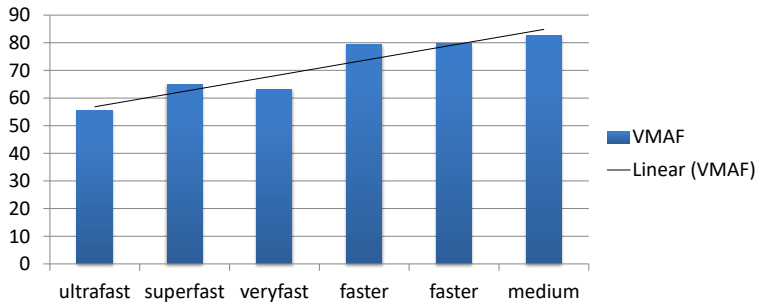


Figure 5. VMAF scores using six presets.

4 Towards new trends in video codecs

Joint Video Exploration Team (JVET) has been working on the Versatile Video Codec (VVC) or H.266 as the successor of HEVC. It is expected that new VVC could deliver the same quality with up to 50% improvement compared to HEVC. MPEG released this new standard in 2020, where UHD 4k/8k, high dynamic range (HDR), VR (Virtual Reality) and omnidirectional 360 degrees video are taken into account [13, 27]. MPEG also proposed Coding-independent code points (CICP) and Versatile Supplemental Enhancement Information (VSEI) for improved interpretation of bitstreams and its parameters [13]. EVC (Essential Video Coding) or MPEG-5 Part 1 addresses cases where coding standards have not be adopted despite technical characteristics. MPEG-5 part 2 or LCEVC (Low Complexity Enhancement Video Coding) is developed, where enhancement layer is specified for streaming when combined with a base encoded video [13, 28]. Wireless and mobile multimedia delivery solutions are especially valuable, and there is a need for focusing on real-time transmission [28-29]. According to Bitmovin's Industry Report H.264 is dominant in 2020 [30]. In [31] it is obtained that besides x265 and VP9, which can qualify for real-time performance, AV1, VVC, and HM (HEVC test model) incur long compression times, and their compression efficiency is the advantage rather than real-time applications.

Patents do have affects on standards and applications [32]. Different patent holders and patent pools have been related to standards implementation, like [33, 34]. AI is also an important trend, and MPAI community is interested in AI-based coding with licensing formats [35]. Information of legal status of some of the solutions is difficult to answer whether one is a lawyer or not. Industrial aspects and, generally, profit should not affect the speed of general acceptance of the new standards. The novel standards are mostly results of united work, as well as the selection of core technologies, but the cooperation often makes conflicts [36]. FRAND means fair, reasonable and non-discriminatory, and licenses may be considered FRAND or non-FRAND. Having in mind impartiality challenge existed in some standards, the licensing for new solutions are of great interest. Fair, adaptive, but balanced approaches are expected for future media codec market, industry, and research needs.

5 Conclusion

Open-source and free video and multimedia technology frameworks, tools and projects play important role in education and research, especially when it comes to video and multimedia engineering fields. The necessity of such projects should be recognized at a larger scale due

to necessity of future video and multimedia experts. Moreover, the impression is that such tools can be extremely useful to the professionals for further research and making extensions. New codec solutions should have in mind different presets and optimizations in order to deal with transmission challenges, as well as practical licensing approaches.

References

[1] Chellappa R, Theodoridis S. Academic Press Library in Signal Processing, Volume 7: Array, Radar and Communications Engineering. Academic Press; 2017 Dec 1.

[2] Zhang T, Mao S. An overview of emerging video coding standards. GetMobile: Mobile Computing and Communications. 2019 May 2;22(4):13-20.

[3] Ma N. Distributed video coding scheme of multimedia data compression algorithm for wireless sensor networks. EURASIP Journal on Wireless Communications and Networking. 2019 Dec;2019(1):1-9.

[4] Hossain K, Roy S. A data compression and storage optimization framework for iot sensor data in cloud storage. In2018 21st International Conference of Computer and Information Technology (ICCIT) 2018 Dec 21 (pp. 1-6). IEEE.

[5] Cisco VNI Complete Forecast Highlights. Global - 2021 Forecast Highlights.https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf (last accessed 01.07.2022.)

[6] Pelurson S, Cozanet J, Guionnet T, Abdoli M, Biatek T. AI-Based Saliency-Aware Video Coding. SMPTE Motion Imaging Journal. 2022 May 10;131(4):21-9.

[7] Gavrovska A. Uvod u savremene video tehnologije i sisteme. Akademska misao, 2021.

[8] Ozer J. What is Codec? https://www.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=74487 (last accessed 01.07.2022.)

[9] Ulas D. Digital transformation process and SMEs. Procedia Computer Science. 2019 Jan 1;158:662-71.

[10] Kale V. Digital transformation of enterprise architecture. CRC Press; 2019 Jul 8.

[11] FFmpeg. https://ffmpeg.org/, https://trac.ffmpeg.org/ (last accessed 01.07.2022.)

[12] Ferrando N. "FFmpeg - From Zero to Hero". https://ffmpegfromzerotohero.com/ (last accessed 01.07.2022.)

[13] MPEG. https://www.mpeg.org/ (last accessed 15.07.2022.)

[14] OpenCV - Open source Coputer Vision, Video I/O with OpenCV Overview, https://docs.opencv.org/4.x/d0/da7/videoio_overview.html (last accessed 15.07.2022.)

[15] Wu X, Qu P, Wang S, Xie L, Dong J. Extend the FFmpeg Framework to Analyze Media Content. arXiv preprint arXiv:2103.03539. 2021 Mar 5.

[16] ITU-T H.264 : Advanced video coding for generic audiovisual services. https://www.itu.int/rec/T-REC-H.264-202108-I/en (last accessed 15.07.2022.)

[17] Bitmovin to Bitmovin's Video Developer Report. https://go.bitmovin.com/video-developer-report-2020 (last accessed 15.07.2022.)

[18] Alliance for Open Media. https://aomedia.org/ (last accessed 15.07.2022.)

[19] Chen Y, Murherjee D, Han J, Grange A, Xu Y, Liu Z, Parker S, Chen C, Su H, Joshi U, Chiang CH. An overview of core coding tools in the AV1 video codec. In2018 Picture Coding Symposium (PCS) 2018 Jun 24 (pp. 41-45). IEEE.

[20] Gavrovska AM, Milivojevic MS, Zajic G. Analysis of SVT-AV1 format for 4k video delivery. In 2020 28th Telecommunications Forum (TELFOR) 2020 Nov 24 (pp. 1-4). IEEE.

[21] Web video codec guide, https://developer.mozilla.org/en-US/docs/Web/Media/Formats/Video_codecs (last accessed 15.07.2022.)

[22] Wu PH, Katsavounidis I, Lei Z, Ronca D, Tmar H, Abdelkafi O, Cheung C, Amara FB, Kossentini F. Towards much better SVT-AV1 quality-cycles tradeoffs for VOD applications. InApplications of Digital Image Processing XLIV 2021 Aug 1 (Vol. 11842, pp. 236-256). SPIE.

[23] SVT-AV1. https://gitlab.com/AOMediaCodec/SVT-AV1 (last accessed 15.07.2022.)

[24] Big Buck Bunny, h HYPERLINK "https://peach.blender.org/"ttps://peach.blender.org (last accessed 01.07.2022.)

[25] Ozer J. Introduction to ABR Production & Delivery, https://www.streamingmedia.com Streaming Media West 2019 (last accessed 01.07.2022.)

[26] OBS, Streaming with x264, https://obsproject.com/blog/streaming-with-x264#presets 2017 (last accessed 01.07.2022.)
[27] Bross B, Wang YK, Ye Y, Liu S, Chen J, Sullivan GJ, Ohm JR. Overview of the versatile video coding (VVC) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology. 2021 Aug 2;31(10):3736-64.
[28] Minopoulos G, Memos VA, Psannis KE, Ishibashi Y. Comparison of video codecs performance for real-time transmission. In2020 2nd International Conference on Computer Communication and the Internet (ICCCI) 2020 Jun 26 (pp. 110-114). IEEE.
[29]] Fujihashi, T., Koike-Akino, T. and Watanabe, T., 2021. Soft Delivery: Survey on A New Paradigm for Wireless and Mobile Multimedia Streaming. *arXiv preprint arXiv:2111.08189*.
[30] Bitmovin to Bitmovin's Video Developer Report. https://go.bitmovin.com/video-developer-report-2020 (last accessed 01.07.2022.)
[31] Panayides, A. S., Pattichis, M. S., Pantziaris, M., Constantinides, A. G., & Pattichis, C. S. (2020). The battle of the video codecs in the healthcare domain-a comparative performance evaluation study leveraging VVC and AV1. *IEEE Access*, *8*, 11469-11481.
[32] Pfeiffer, S., 2009. Patents and their effect on Standards: Open video codecs for HTML5. *International Free and Open Source Software Law Review*, *1*(2), pp.131-138
[33] MPEG-LA. https://www.mpegla.com/ (last accessed 01.07.2022.)
[34] Access Advance. https://accessadvance.com/ last accessed 01.07.2022.)
[35] MPAI community. https://mpai.community/standards/mpai-evc/about-mpai-evc/ last accessed 01.07.2022.)
[36] Jones SL, Leiponen A, Vasudeva G. The evolution of cooperation in the face of conflict: Evidence from the innovation ecosystem for mobile telecom standards development. Strategic Management Journal. 2021 Apr;42(4):710-40.

PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7139336

Driving Innovation in Free Knowledge with UNLOCK Accelerator

Kannika Thaimai[1], Ivana Madžarević[2]

1: Wikimedia Deutschland, 10963, Berlin, Germany
2: Wikimedia Serbia, 11000, Belgrade, Serbia
e-mail addresses: kannika.thaimai@wikimedia.de, ivana.madzarevic@vikimedija.org

ABSTRACT

This paper deals with the topic of innovation in free knowledge and how these innovations can lead to overcoming existing challenges in achieving knowledge equity. Wikimedia Accelerator UNLOCK, which in its third edition is focused on regional collaboration between Western Balkans and German-speaking areas, was presented as one of the programs which promotes the power of open innovation and impact-driven ideas that make knowledge accessible to everyone. The paper first refers to the existing obstacles within the Wikimedia projects that this program aims to overcome, which relate to equal access to knowledge, targeted diversity in terms of content on Wikipedia and other Wikimedia projects, as well as diversity in terms of those contributing to free knowledge. In addition, UNLOCK is a tailor-made accelerator that supports ideas outside the Wikimedia movement, thus inviting participants to gain skills by having peer-to-peer exchange, exchange of successful models and includes a mentoring program. The paper presents the methodology that was used, as well as the mapping of the innovative capacity of the movement. Finally, the projects that are being implemented at the time of writing the paper and the outcomes that have been achieved are listed.

Keywords: free knowledge, open software, innovation, open data, regional collaboration, knowledge equity.

1 Introduction

Wikimedia is one of the biggest free knowledge movements which promotes projects with open access and free content such as Wikipedia, Wikimedia Commons, Wikidata and others. It is a global movement whose mission is to bring free educational content to the world. [1] The current Wikimedia technologies, platforms, projects, policies, knowledge formats, editing rules, social structures and governance systems have somewhat organically and unsystematically grown. In order for us to implement knowledge equity, to enable emerging and marginalized communities to join and participate, we will not only need to remove barriers or adjust those technologies, platforms, projects, etc. but we will also need to create an environment where these new communities can devise their own technologies, systems, social structures, policies and governance. Innovation here is much more than new gadgets. Explicitly it includes policies, processes, formats and social innovations as well. We hope

that these innovations will open the doors to new people and new content, helping to grow a diverse and vibrant movement.

1.1 How can we drive innovation for free knowledge?

UNLOCK Accelerator is an innovation-driving program that aims to promote new free knowledge projects. UNLOCK supports participating teams over a set period of time in validating, testing and further developing their project ideas in a structured manner with the help of coaching, exchange and collaboration, a network of experts and, if required, a scholarship. The program is open to Wikimedia volunteers as well as aims to tap into new communities in order to attract innovators who have not yet been part of the Wikimedia movement. UNLOCK was initially launched by Wikimedia Deutschland in the German-speaking region in 2020 and scaled to the European level in 2021. In these editions, the program was able to generate over 80 applications and supported 10 projects. This year Wikimedia Deutschland, Wikimedia Serbia and Impact Hub Belgrade have joined forces to co-design and co-host the program to support projects teams from the German-speaking region (Austria, Germany, Switzerland) as well as the Western Balkans (Albania, Bosnia and Herzegovina, Kosovo, Montenegro, North Macedonia, Serbia).

2 Background

The Wikimedia Foundation operates eleven content projects that follow the free content model, with their main goal being the dissemination of knowledge. [2] Existing Wikimedia projects use Creative Commons licenses and are free in terms of using the content, as well as making contributions by increasing the free content on these platforms. The current Wikimedia technologies, platforms, projects, policies, knowledge formats, editing rules, social structures and governance systems have changed over time. New functionalities have been included in the systems themselves, the user experience has been improved, and the experience from the editor's point of view, tools have been created to measure outcomes in order to increase the satisfaction and retention of editors, as well as the quality of the edited content. Over time, the focus of the movement and projects have expanded. Although the essence was the creation and promotion of free knowledge, since the beginning of the creation of Wikipedia and other projects, new goals have been developed, where, in addition to increasing the free knowledge, we are now working to a large extent on increasing accessibility and creating equal opportunities for various marginalized and sensitive communities.

2.1 Challenges of current Wikimedia-projects

However, projects still have certain challenges in terms of access to knowledge and creating the environment where we are closer to achieving knowledge equity every day. In terms of who is reading Wikimedia projects, here are some statistics from survey Wikimedia Foundation has conducted:

- Across regions, men tend to read Wikipedia more often than women. Though awareness and usage of Wikipedia are high for both men and women in many regions of the world, based on reader surveys one-third (33 percent) of Wikipedia readers over the age of 18 on any given day are women. [3]

- The same survey showed that men on average also read more articles when they visit Wikipedia than women. As such, many of the top-read articles on Wikipedia draw almost exclusively readers who are men. [4]

In addition to the above mentioned, there is also the question of meeting the needs of editors - do we have tools that can ensure uninterrupted editing of members of marginalized and / or underrepresented communities, do we have tools that can provide enrichment of content about notable people from socially sensitive groups. When we look at the statics who is contributing to Wikimedia projects, here is the current state of it:

- Wikimedia contributors are 87% male. Almost half live in Europe and one-fifth in Northern America, as compared to 9.7% and 4.8% of the global population.
- Fewer than 1% of Wikipedia's editor base in the U.S. identify as Black or African American.
- Although women were still markedly underrepresented among contributors, there was a modest increase in women contributors between 2019 (11.5%) and 2020 (15.0%).
- Only 1.5% of Wikipedia editors are based in Africa, although people in Africa comprise 17% of the world's population. [5]
- More men than women have tried to edit Wikipedia at least once. Across Wikipedia users in all six of the regions surveyed (the United States, Mexico, Egypt, Nigeria, Germany, and India), 27 percent of male respondents had edited Wikipedia at least once, while only 21 percent of female respondents had. [6][7]

Members of marginalized communities can often be discriminated against, and in order to avoid these situations, the Wikimedia Movement created the Universal Code of Conduct. This Universal Code of Conduct (UCoC) defines a minimum set of guidelines of expected and unacceptable behavior. It applies to everyone who interacts and contributes to online and offline Wikimedia projects and spaces.[1] Although the code has been adopted, through innovation we can contribute to its implementation.

In order for us to enable emerging and marginalized communities to join and be involved in the projects, we have to face the challenges and create an environment where these new communities can devise their own technologies, systems, social structures, policies and governance. This can be done through the introduction of innovations and new systems that have not been implemented so far or have been implemented to a certain extent, but not enough to create a final solution. Introducing innovations is a process that implies constant listening to the needs of the community, repeating several iterations until the final solution is reached, and this is more demanding when talking about innovation in free knowledge.

2.2 Achieving knowledge equity

Wikimedia Foundation works on ensuring that everyone, everywhere has equitable access to create and consume information. Looking back in history, knowledge was not accessible to all and has been concentrated in the hands of the few. Marginalized groups' histories and perspectives have been excluded by structures of power and privilege. Creating Wikipedia was a starting point to revolutionize this model, as this was the first and world's largest, free,

1 https://meta.wikimedia.org/wiki/Universal_Code_of_Conduct

collaboratively-sourced encyclopedia. [8] Over the years, a lot of volunteers joined the free movement, uploading and editing various content that was visible and easily accessible to all the people in the world. Yet, Wikipedia and other Wikimedia projects do not currently reflect the world's diversity. Specifically, our projects are largely missing the histories, stories, and contexts of: women and nonbinary people; those within the LGBTQI+ community; people with disabilities; and those within the global majority, including Black people, Indigenous peoples, and people of color. [9]

If we look at the current state,the following barriers exist and are experienced on Wikimedia projects:

- Content on the projects is still not fully adapted to blind and visually impaired people.
- Biographies about women who meet Wikipedia's criteria for inclusion are more frequently considered non-notable and nominated for deletion compared to men's biographies.
- Women and non-binary contributors have identified systemic bias in policies; lack of awareness and implicit bias within community; and poor community health as the biggest obstacles to achieving gender equity in the Wikimedia movement.
- The guidelines on reliable sources in the English, French and Spanish language editions of Wikipedia impact contributors and the inclusion of content for marginalized communities.
- Many readers have challenges accessing Wikipedia (and other internet sites) because of issues with infrastructure, cost of data, and more. [10]

In order to make these projects sustainable, which are of crucial importance for humanity, we must be guided by the strategic recommendations of the Wikimedia movement, which, among other things, cover the issue of relevance. This implies that the question of whether the projects are relevant enough, whether they are still important and necessary, is being asked again and again. In order to stay relevant and the environment where we can fight above mentioned challenges, innovation has to be a part of our path.

In this case innovations include policies, processes, formats and social innovations as well. We hope that these innovations will open the doors to new people and new content, helping to grow a diverse and vibrant movement. UNLOCK was created to drive innovation – thereby investing in people and communities who are working on new free knowledge projects that address knowledge equity. Besides, the collaboration with Wikimedia affiliates as well as with a player from the innovation ecosystem will also provide learning opportunities for us as a movement, on how to collaborate with people and institutions that are not from the immediate open knowledge movement. Creating new alliances in the innovation field will strengthen our movement both regionally and as a whole. UNLOCK is not a "one size fits all" solution; rather it should evolve into something that ultimately becomes freely adaptable, reusable and changeable depending on the geographic, cultural and economic context. This process of adaptation, or better, contextualization – a central movement strategy principle – is an area where we as a movement still have much to test and to learn. We see great learning opportunities through the contextualisation of the program with the Western Balkan partners – Wikimedia affiliates and those from the innovation ecosystem alike.

3 Methodology

With a clear and bold vision[2], with concrete recommendations[3] and guiding principles[4] in ‘our bag’, the Wikimedia Movement Strategy[5] is undergoing implementation. In doing so, social and technical innovations are going to be crucial to master the challenges of today and tomorrow in order to equitably create knowledge. Strengthening the innovative capacity of the movement will be key to address gaps in knowledge equity, and therefore, to stay relevant and attractive as a movement in the future. But how to ‘innovate in free knowledge’?

3.1 Innovation framework

In 2022 we published a first analysis on where the Wikimedia movement is at when it comes to innovation [11]. Our methodology towards innovation is based on a very common innovation framework [12] – defined, validated and continuously developed – by Nesta (the National Endowment for Science, Technology and the Arts) and the Young Foundation. The framework clearly shows that innovation is neither simple nor predictable, but rather a complex story of loops and leaps within a process, structured into different phases (see figure 1).

Figure 1: By Kannika Thaimai (WMDE) - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=114682763.jpg

Not every innovation moves through the seven stages sequentially – stages can overlap or skip, while some innovations jump directly into practice or even scaling. We apply this framework as it provides a common language for thinking about how to support innovation more systematically and thereby taking into consideration the different formats, techniques, tools, processes of support that innovators and innovations might require at each stage in order to grow.

2 https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017
3 https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Recommendations
4 https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Recommendations/Movement_Strategy_Principles
5 https://meta.wikimedia.org/wiki/Movement_Strategy

3.2 Mapping Wikimedia innovative capacity

In the past two decades of the Wikimedia Movement, structures and processes as well as communities of volunteers and free knowledge enthusiasts have been created. These have driven the incremental development within existing Wikimedia projects.

Some examples: The technical community can get together at the annual Wikimedia Hackathon and get involved in the development of the MediaWiki[6] and many other areas in Wikimedia's technical ecosystem [13]. And structures that are already in place make it easier to integrate additional ideas and prototypes relating to already existent Wikimedia projects into the established ecosystem. This is apparent in the development of Wikidata[7], the Wikibase Ecosystem[8], Wikimedia Commons[9] and other Wikimedia projects. Our mapping has also clearly shown some significant shortcomings when it comes to experimenting with or promoting unfamiliar and novel ideas and projects. There may be many great ideas out there but currently only few structures and resources of support to truly foster them, which is equally necessary for the Wikimedia movement to evolve, grow, and become sustainable in the long run. Further readings and deep dives into the mapping can be found in the analysis 'Building an ecosystem to innovate in free knowledge' [14].

3.3 UNLOCK Accelerator: Making room for new free knowledge ideas and projects

Among the few support structures already in place is the Wikimedia Accelerator UNLOCK. At UNLOCK, we do not only support volunteers who are already active in the movement or involved in existing Wikimedia projects, but also people who have not been a part of our movement so far. Because this is also particularly about allowing new impulses to evolve (this includes ideas, solutions and people) that we as a movement have not even seen and taken into consideration. This may include new formats to make knowledge accessible as well as more diverse solutions for consuming, sharing and creating knowledge.

UNLOCK is a structured cohort-based program and supports participating teams in developing their idea into a functional prototype. UNLOCK provides:

- Coaching & mentoring: All participating teams will work together with experienced coaches. They will be at the team's side from the early stages of envisioning their project to the development of a plan for its realization and finally the implementation itself. Participants will be challenged to refine and further develop their idea and impact, and will learn how to sustain their projects and initiatives.

- Skill development: Project teams will not only develop a feasible and viable concept based on their idea, but will learn new methods, mindsets and tools that can help them with many other team projects and ventures. UNLOCK workshops and working sessions include agile and co-creative methods such as design sprint, design thinking just to name but a few.

- Impactful communities: Wikimedia Deutschland and Wikimedia Serbia are passionate about Free Knowledge and boast a wealth of experience in the field. Participants will benefit from both our expertise and connections.

6 https://www.mediawiki.org/wiki/MediaWiki
7 https://www.wikidata.org/wiki/Wikidata:Main_Page
8 https://wikiba.se/
9 https://commons.wikimedia.org/wiki/Main_Page

- Financial support: If needed, we can offer financial support for your participation in the program. The scholarships are intended to contribute to the cost of living so that participants can participate in the UNLOCK Accelerator full-time.

In addition to this structured format, UNLOCK 2022 focused on a cross-regional approach with its collaboration among Wikimedia Deutschland, Wikimedia Serbia and Impact Hub Belgrade. What are our core motivations for this collaboration?

- We aim to strengthen and build our capacities through joint growth, frequent exchange and by challenging one another – allowing us to further advance the program.

- By coming together we can also expand our international networks. This not only helps to increase the awareness for the program, and in return attract and support even more innovators from the regions; but it also allows us to pull from a larger pool of experts who can lend their knowledge and skills set to our program participants.

- Working across sectors, we hope to bring the Open Knowledge community and innovation driven communities closer together. We hope that they will find fruitful and long-term ways of collaborating as well, strengthening an innovation ecosystem for Free Knowledge. This is what we aim for.

4 Results

UNLOCK is committed to encourage and challenge by providing an experimental space for projects where, in particular, risks can be taken and exchanging learnings regarding best practices and failures is encouraged. We believe that multiple rounds of trial and failure are integral to the process of developing innovative solutions. Experimental spaces are particularly relevant in the German-speaking and Western Balkan context, as a culture of failure is still not strongly developed, in contrast to the USA [15]. Even though many people and institutions support civil society projects: free knowledge and open source technology and innovation certainly doesn't have it easy in these regions, and there are hardly any low-threshold funding programs for emerging, non-profit free knowledge and open source projects [16].

4.1 Output

For this year’s open call we received 34 applications from 12 countries and a total of 104 applicants. Out of these 104, 53% are from the Western Balkans region, 24% from the German-speaking area and 23% “others” (incl. people from European and non-European countries; plus those without information).

Seven innovative, bold, engaging and diverse projects – with a total of 24 participants from Albania, Germany, Montenegro and Serbia – will be supported within the UNLOCK program. These projects show and have convinced us – the jury and the program team – throughout the selection process [17] with new perspectives for free knowledge – and addressing knowledge equity – in different regional as well as thematic contexts:

- activist.org[10] – an open source platform that breaks down barriers to becoming politically active and thereby connects people and organizations from different social and activist movements.

10 https://www.wikimedia.de/unlock/unlock-projects/activist-org/

- Game of political participation[11] – encouraging young people in the Western Balkans to familiarize themselves with political decision making and political systems through elements of gamification.
- f[ai]r[12] – establishing an ethics certification for digital applications through a holistic examination of the AI system in the social context, addressing aspects of bias, discrimination, diversity and inclusion.
- Inclusio[13] – providing user-generated audio descriptions of visual content to the blind and visually impaired. Ideally the solution could be connected and tested on structured data in Wikimedia Commons.
- macht.sprache.[14] – fostering politically sensitive translation through an open source platform that allows for crowdsourcing and discussing politically sensitive terms and their translations, and through a tool to help translate with sensitivity.
- MOCI SPACE[15] – a digital space to connect activists, grassroots initiatives and civil society actors in the Western Balkans and that allows for co-creating, publishing and sharing knowledge by making use of the Matrix protocol for federated communication.
- P2P Wiki for indigenous wisdom and biodiversity[16] – an open source tool to collect and safeguard indigenous knowledge, and to raise awareness about biodiversity with a P2P offline-first methodology.

The funded projects show varying degrees of proximity or distance to existing Wikimedia projects. For example, activist.org uses interfaces to Wikidata or Inclusio is exploring options to connect and test structured data in Wikimedia Commons. Other projects operate independently of existing Wikimedia projects. There is potential to strengthen the connection to existing Wikimedia projects and thus develop them further. However, this would depend on various factors including project vision and commitment as well as (technical) feasibility.

By the end of the program, each project will showcase their results in a Demo Day. UNLOCK demo days, the communication around the event, and our general outreach activities bring people into contact with innovative free knowledge and open source ideas and concepts and their implications for various communities in the German-speaking as well as Western Balkan areas. In this way, interest in and understanding of how making knowledge free, more inclusive and equitable can affect individuals and how individuals can als get involved into these free knowledge and open source projects.

5 Discussion

Furthermore, UNLOCK creates further results that might not be measured in the classical sense. Running three UNLOCK editions clearly unfolded the potential of the program geared towards innovation: the Wikimedia Accelerator provided the participants[17] with the necessary knowledge and skills-enhancing methods to drive the implementation of their projects forward in the shortest possible time. With UNLOCK, we create awareness in new

11 https://www.wikimedia.de/unlock/unlock-projects/game-of-political-participation/
12 https://www.wikimedia.de/unlock/unlock-projects/fair/
13 https://www.wikimedia.de/unlock/unlock-projects/inclusio/
14 https://www.wikimedia.de/unlock/unlock-projects/macht-sprache/
15 https://www.wikimedia.de/unlock/unlock-projects/moci-space/
16 https://www.wikimedia.de/unlock/unlock-projects/p2p/
17 UNLOCK Accelerator supported in total 17 project teams with the total number of participants of 56 in the years 2020-2022.

communities that have not been addressed by Wikimedia Deutschland so far – communities from the innovation context and the field of social entrepreneurship. We have also been able to activate teams from these communities to participate[18]. We have created access to Wikimedia for these people and projects, and in some cases also to existing Wikimedia projects. UNLOCK has established a space for exchange and knowledge transfer between projects inside and outside the movement.

With such an innovation-driving format like UNLOCK we could accelerate ideas to be turned into prototypes that aim to address knowledge equity. As of now, 17 prototypes could be developed. And not all of them will sustain – from the first two editions 6 out 10 projects are still active. We strive for diverse and inclusive projects that have to be applied bearing in mind the quality. In order to qualitatively assess the impact of the projects – meaning those that truly achieve knowledge equity and sustain – more time is needed.

It has also become apparent (and confirmed by former participants) that UNLOCK provides participating teams with opportunities for further personal as well as professional development. The projects value the program's agile and people-centered approach to responding to the needs of the teams. This results in participants being able to develop their skills, not overexert themselves in volunteering, and maintain an intrinsic motivation for their project.

With respect to the cross-regional collaboration, we have learned how this could be beneficial to the participants and program as a whole. Wikimedia Deutschland and Wikimedia Serbia are part of the large Wikimedia free movement and brought together their expertise and resources in terms of free knowledge, open software, open resources, peer to peer exchange and tailored training for the participants. Impact Hub Belgrade provided support in terms of building entrepreneurial spirit and providing mentors for the teams. With these joint resources and forces, partners bring the quality of the program to the new level.

One – if not the greatest – challenge in driving innovation is the sustainability of the supported projects. UNLOCK is not designed for long-term support. And as described above, innovation is a process with different stages. In the long run, we hope to be able to create an innovation ecosystem within the Wikimedia movement. This means that we – as a movement – must take a more systematic approach in order to not only set selective and short-term impulses for innovations, but to create a long-term, sustainable impact with different actors and stakeholders as well as our actions.

References

[1] Wikimedia Foudation, official website: https://www.wikimedia.org/, last accessed 21.09.2022.

[2] Meta-Wiki, the global community site for the Wikimedia Foundation's projects: https://meta.wikimedia.org/wiki/Wikimedia_projects, last accessed 15.09.2022.

[3] Global Reach team of Wikimedia Foudation, Research:Characterizing Wikipedia Reader Behaviour: https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Reader_Behaviour/Demographics_and_Wikipedia_use_cases#Reader_Surveys, last accessed 21.09.2022.

[4] Global gender differences in Wikipedia readership, Isaac Johnson, Florian Lemmerich, Diego Saez-Trumper, Robert Wes, Markus Strohmaier and Leila Zia: https://arxiv.org/pdf/2007.10403.pdf

18 Based on our internal application evaluation for all three rounds: more than 80% of applications have been submitted by people from outside the movement.

[5] Wikimedia Community Insights/Community Insights 2020 Report: https://meta.wikimedia.org/wiki/Community_Insights/Community_Insights_2020_Report, last accessed 21.09.2022.
[6] Communications/YouGov survey on women and Wikipedia: https://meta.wikimedia.org/wiki/Communications/YouGov_survey_on_women_and_Wikipedia, last accessed 21.09.2022.
[7] Wikimedia Foundation, *Open the knowledge: Change the stats*: https://wikimediafoundation.org/our-work/open-the-knowledge/#a3-change-the-stats, last accessed 19.09.2022.
[8] Wikimedia Foundation, *Education: Promoting Knowledge Equity*: https://wikimediafoundation.org/our-work/education/promoting-knowledge-equity/, last accessed 18.09.2022
[9] Wikimedia Foundation, *Open the knowledge: Help us achieve knowledge equity* https://wikimediafoundation.org/our-work/open-the-knowledge/#a1-help-us-achieve-knowledge-equity, last accessed 18.09.2022
[10] Wikimedia Foundation, *Open the knowledge: Change the stats*: https://wikimediafoundation.org/our-work/open-the-knowledge/#a3-change-the-stats, last accessed 19.09.2022.
[11] Wikimedia Deutschland/Innovation Engine/Building an ecosystem to innovate in free knowledge: https://meta.wikimedia.org/wiki/Wikimedia_Deutschland/Innovation_Engine/Building_an_ecosystem_to_innovate_in_free_knowledge, last accessed 21.09.2022.
[12]https://media.nesta.org.uk/documents/the%20open%20book%20of%20social%20innovation.pdf
[13] Wikimedia Foundation, *Wikimedia technical areas*: https://commons.wikimedia.org/wiki/File:Wm_technical_areas.svg, last accessed 21.09.2022.
[14] Wikimedia Deutschland/Innovation Engine/Building an ecosystem to innovate in free knowledge: https://meta.wikimedia.org/wiki/Wikimedia_Deutschland/Innovation_Engine/Building_an_ecosystem_to_innovate_in_free_knowledge, last accessed 21.09.2022.
[15] Ezell, S. / Marxgut, P. (2015): Comparing American and European Innovation Culture. https://www2.itif.org/2015-comparing-american-european-innovation-cultures.pdf, last accessed 21.09.2022.
[16] Leu, P (2022): Funding for Future Wirkungen eines leichtgewichtigen und niederschwelligen Förderinstruments für Open-Source-Software. https://craft.stiftung-mercator.ch/files/Dokumente/Publikationen/PrototypeFund_Handbuch.pdf, last accessed 21.09.2022.
[17] Wikimedia Accelerator UNLOCK: UNLOCK 2022 selection process: https://www.wikimedia.de/unlock/unlock-blog/unlock-2022-selection-process, last accessed 21.09.2022.

PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7139492

Upravljanje logovima i vizualizacija statistika korišćenja AMRES servisa upotrebom alata otvorenog koda

Todosijević Andrijana[1], Simonović Katarina[1], Arsović Anđela[1]

1: Akademska mreža Republike Srbije - AMRES, 11000, Beograd, Srbija
elektronska pošta: andrijana.todosijevic@amres.ac.rs, katarina.simonovic@amres.ac.rs, andjela.arsovic@amres.ac.rs

REZIME

Log poruke predstavljaju automatski dokumentovane događaje u formi hronoloških zapisa koji sadrže različite informacije o IT sistemu i mreži. Upravljanje log porukama je od velikog značaja za svaku organizaciju, pa i Akademsku mrežu Republike Srbije (AMRES) i omogućava efikasnu i kvalitetnu analizu rada i upotrebu kako servisa, tako i mreže u celini. Od izuzetne važnosti je i mogućnost brze i jednostavne pretrage velikog broja generisanih log poruka, rešavanje problema i izdvajanje bitnih podataka za kasniju upotrebu. Elastic Stack softver je sveobuhvatan alat otvorenog koda koji omogućava prikupljanje i pretragu velike količine log poruka različitog tipa, kreiranje dinamičkih izveštaja i grafičkog prikaza željenih rezultata. U radu su razmatrani i detaljno objašnjeni procesi prikupljanja i analize log poruka AMRES eduroam servisa i dati primeri upotrebe Grafana alata otvorenog koda u prikazu statistika korišćenja servisa od strane AMRES krajnjih korisnika.

Ključne reči: logovi, upravljanje logovima, Elastic Stack, Grafana, eduroam.

1 Upravljanje logovima

Upravljanje log porukama je složen proces koji za cilj ima generisanje, prenos, skladištenje, zatim i analizu velike količine podataka u okviru informacionog sistema [1]. Log poruke se sastoje od hronoloških zapisa koji sadrže različite informacije i predstavljaju automatski dokumentovane događaje u samom sistemu i mreži. Prvobitno, logovi su korišćeni za identifikaciju sigurnosnih incidenata i rešavanje problema, ali danas imaju mnogo dodatnih i podjednako značajnih funkcija. Koriste se za optimizovanje performansi servisa i mreže, praćenje ponašanja korisnika i generisanje podataka korisnih za istraživanje i analizu njihovih aktivnosti. Rast broja, obima i raznovrsnosti logova praćen je povećanjem potrebe za upravljanjem log porukama. Upravljanje logovima je ključan segment zaštite i održavanja funkcionisanja servisa i mreže. Sposobnost prikupljanja različitih log poruka sa više izvora na jednom mestu, kao i njihova automatska pretraga i analiza, od velikog su značaja za svako IT okruženje. Veliki broj alata i softvera omogućavaju brzu i uspešnu analizu problema, kao i trenutno delovanje i akcije bez potrebe za manuelnim prikupljanjem, organizovanjem i pretragom velike količine podataka. Koristeći ove mogućnosti i funkcionalnosti, organizacija može na veoma efikasan način da održava mrežu i servise. Svaka organizacija ima višestruku korist od procesa sakupljanja i upravljanja logovima. Na ovaj način je omogućeno da se svi detalji čuvaju u obliku zapisa za određeni vremenski period. Rutinski pregledi i analiza logova su ključni za identifikaciju incidenata, problema u funkcionisanju mreže i servisa, kao i rešavanje istih. Takođe, mogu imati veliku ulogu u

analizi ponašanja krajnjih korisnika, biti deo internih istraživanja, uspostavljanja osnova i identifikacije operacionih trendova i dugoročnih problema [1]. U radu je opisano upravljanje logovima kojima se monitorišu servisi i aplikacije u AMRES mreži, kao i alati koji se u tu svrhu koriste, navedeni u Tabeli 1.

Tabela 1: Alati koji se koriste u procesu upravljanja logovima.

Naziv softvera	Tip softvera	Funkcija softvera
freeradius	alat otvorenog koda	RADIUS server
syslog-ng	alat otvorenog koda	generisanje i prikupljanje log poruka
Logstash	alat otvorenog koda	prikupljanje i obrada log poruka
Elasticsearch	alat otvorenog koda	indeksiranje, skladištenje, pretraga i analiza log poruka
Kibana	alat otvorenog koda	vizualizacija, pretraga i analiza log poruka
Grafana	alat otvorenog koda	vizualizacija metrika i vremenskih serija log poruka

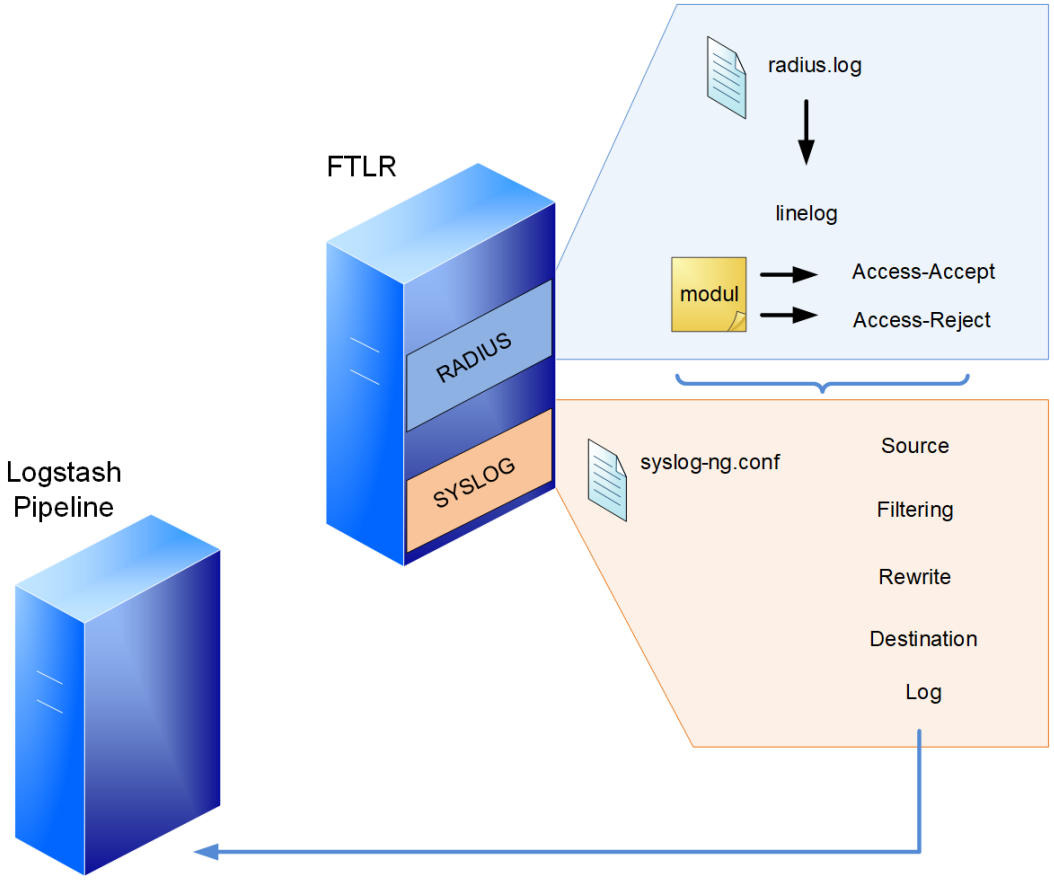
AMRES prati rad i korišćenje više servisa, a implementacija i funkcionisanje Elastic Stack [2] i Grafana [3] softvera razmatrana je sa aspekta primene u eduroam servisu. U Poglavlju 2 objašnjen je način generisanja logova i dat prikaz osnovne konfiguracije syslog-ng protokola koji se za tu priliku koristi. Implementacija Elastic Stack okruženja razmatrana je u Poglavlju 3. Opisana je konfiguracija Logstash pipeline alata za obradu i filtriranje log poruka, kao i Elasticsearch klastera za prikupljanje i skladištenje logova. Razmatrani su i način pretraživanja, prikaza i analize izdvojenih rezultata korišćenjem Grafana alata.

2 Generisanje i prikupljanje log poruka

Infrastruktura upravljanja log porukama obuhvata hardver, softver, mrežu i uređaje koji se koriste za njihovo generisanje, prikupljanje, skladištenje, analizu i upotrebu [1]. Ova infrastruktura takođe obuhvata i nekoliko funkcionalnosti koje predstavljaju dopunu prethodno navedenim procesima. Neke od njih su i parsiranje logova (izdvajanje podataka iz loga tako da se parsirane vrednosti mogu koristiti kao inputi za druge log procese), filtriranje i agregaciju događaja. Proces prikupljanja i skladištenja podrazumeva redukciju, kompresiju, konverziju, arhiviranje logova, kao i rotaciju log fajlova prema određenom rasporedu i proveru njihovog integriteta. Poslednji korak je analiza logova koja se sastoji od mapiranja zapisa iz jednog ili više izvora (na osnovu IP adrese, DNS imena, tipa događaja itd.), prikaza logova i kreiranja izveštaja. Postoje različiti načini generisanja i prikupljanja logova. Iako se u okviru Eastic Stack softvera u ovu svrhu koriste Beats alati, za potrebe praćenja rada određenih AMRES servisa od značaja, koristi se syslog-ng softver koji je dostupan za Linux platforme.

2.1 Osnovna konfiguracija syslog-ng softvera za potrebe eduroam servisa

Kao jedan od najvažnijih servisa, eduroam zahteva slanje najdetaljnijih log poruka i u tu svrhu se koristi syslog-ng softver [4]. Log poruke se prosleđuju Logstash pipeline serveru, koji zatim vrši njihovu dalju obradu. Poruke koje se šalju su dobijene kao rezultat pokušaja autentifikacije krajnjih korisnika, a generiše ih RADIUS deamon. Detaljno objašenjenje svih poruka i pojedinačnih parametara iz tih poruka je dato u Poglavlju 2.2. Postupak generisanja i slanja logova razlikuje su u zavisnosti od servisa koji se monitoriše. Kako se ovaj rad fokusira na analizu log poruka eduroam servisa, na Slici 1 prikazan je način komunikacije između sistemskih komponenti koje se nalaze na istom serveru (RADIUS server, syslog-ng deamon) i udaljenog Logstash pipeline servera.



Slika 1: Postupak generisaja i prikupljanja log poruka eduroam servisa.

Osnovni nivo konfiguracije zasniva se na tri komponente:

- izvor,
- odredište i
- Log sekcija.

Konfiguracioni fajl *syslog-ng.conf* mora imati bar tri osnovna parametra, *source*, *log* i *destination*, koja su data u nastavku:

```
source s_local {
       system();
       internal();
};
destination d_logstash {
       udp("147.91.x.x" port(514));
};
log {
       source(s_local);
       destination(d_logstash);
};
```

U ovom okruženju, puna putanja do konfiguracionog fajla je */usr/local/etc/*. U konfiguraciji je označen deo koji predstavlja default UDP (User Datagram Protocol) port za syslog protokol, preko koga se omogućava slanje logova na udaljenu lokaciju [5].

2.2 RADIUS log poruke

FTLR (*Federation Top-Level RADIUS*) server je RADIUS server kroz koji prolaze autentifikacioni zahtevi prilikom pokušaja povezivanja na eduroam. Ovi autentifikacioni zahtevi mogu poticati od korisnika iz AMRES mreže, kao i od inostranih korisnika. AMRES FTLR serveri su realizovani primenom FreeRADIUS softvera [6]. Ovaj softver prema početnoj konfiguraciji beleži sve autentifikacione zahteve u *radius.log* fajl. Rezultat autentifikacije može biti poruka „Login OK“ ili „Login incorrect“. Uz primenu FreeRADIUS linelog modula, ove poruke se prepisuju tako da budu usklađene sa RADIUS RFC 2865 preporukom [7], pa za neuspešnu autentifikaciju imaju vrednost „Access-Reject“ , dok za uspešnu autentifikaciju imaju vrednost „Access-Accept“. U nastavku je dat primer konfiguracije linelog modula:

```
linelog logstash {
       filename = syslog
       format = ""
       reference = "%{%{reply:Packet-Type}:-format}"
       Access-Accept ="Access-Accept: IdP=%{tolower:%{Realm}} MAC=%{Calling-Station-
           Id} AP=%{Called-Station-Id} RP=%{Operator-Name}"
       Access-Reject ="Access-Reject: IdP=%{tolower:%{Realm}} MAC=%{Calling-Station-
           Id} AP=%{Called-Station-Id} RP=%{Operator-Name}"
```

Ovi podaci se zatim šalju syslog-ng softveru, koji filtrira, prepisuje i usmerava log poruke ka udaljenom Logstash pipeline serveru. Atribut „Called-Station-Id“ je u formatu Base Radio MAC:SSID (npr. 00--00-00-00-00-00:eduroam), koji je nedovoljno razumljiv [5]. Kako bi se dobila prepoznatljiva vrednost za AP (Access Point) atribut „Called-Station-Id“ u AP delu poruke, koriste se dva načina mapiranja:

- *rewrite* sekcija syslog-ng softvera, čime se ovaj atribut prepisuje u format koji se odnosi na lokaciju na kojoj je AP postavljen (npr. cisco1142-rcub-studenjak5), preporučuje se kada postoji manji broj AP uređaja. Nakon što log poruka prođe postupak generisanja i obrade, njen konačan format je:

  ```
  Jan 28 15:37:21 ftlr1 radiusd[31369]: Access-Accept: IdP=etf.bg.ac.rs MAC
     =48-50-73-x-x-x AP=cisco1142-rcub-studenjak5 RP=1rcub.bg.ac.rs
  ```

- *Lookup* fajl sa zapisima koje koristi Logstash pipeline, koji će biti objašnjen u Poglavlju 3.1.

RADIUS log poruka se sastoji od sledećih podataka:

- Access-Accept/Access-Reject – rezultat autentifikacije,
- IdP – domen institucije,
- MAC – MAC adresa korisničkog uređaja,
- AP – niz karaktera koji predstavlja naziv AP uređaja, na osnovu koga se određuje lokacija AP-a,
- RP – RADIUS atribut *Operator-Name* na osnovu koga se određuje Davalac Resursa, tj. kojoj instituciji pripadaju AP uređaji.

3 Elastic Stack softver otvorenog koda

Za prikupljanje i skladištenje log poruka AMRES koristi deo Elastic Stack softvera koji predstavlja skup alata otvorenog koda i formira veoma moćnu platformu za upravljanje logovima, prikupljanje i obradu podataka iz više izvora, centralizovano skladištenje na skalabilan način, uključujući i skup alata za njihovu analizu i pravljenje izveštaja. Elastic Stack softver obuhvata sledeće komponente:

- Beats
- Elasticsearch
- Logstash
- Kibana

Beats komponenta predstavlja skup alata koji služe sa formatiranje log poruka i polja u okviru log zapisa i slanje istih ka Logstash pipeline serveru. Elasticseach je NoSQL baza podataka [8] za indeksiranje, skladištenje, pretragu i analizu velike količine podataka, obezbeđujući RESTful API [9] interfejse i JSON [10] format za rad sa podacima. Ovaj alat nudi maksimalnu pouzdanost, lako upravljanje, jednostavnu implementaciju, ali i mogućnost naprednih upita. Logstash pipeline je agregator podataka koji se koristi za prikupljanje log poruka sa više izvora, omogućuje transformacije nestruktuiranih podataka, filtriranje i slanje podataka u Elasticsearch bazu. Kibana je alat za vizualizaciju podataka, implementiran tako da dopunjuje Elasticsearch i omogućuje pretragu, pregled i interakciju sa indeksiranim podacima u realnom vremenu. Koristi se primarno za analizu log poruka i omogućuje tekstualne upite i pretragu podataka [2]. Grafana je još jedan alat otvorenog koda koji se koristi za vizualizaciju metrika i vremenskih serija za log poruke koje se dobijaju iz različitih izvora podataka, uključujući Elasticsearch, kao i kreiranje dinamičkih izveštaja i grafičkog prikaza željenih rezultata.

3.1 Konfiguracija Logstash pipeline softvera

U prethodnom poglavlju opisan je syslog protokol, pomoću koga se šalju podaci sa udaljene lokacije, tj. RADIUS servera na Logstash pipeline server. Na samom Logstash pipeline serveru takođe je potrebno konfigurisati syslog-ng servis, koji upisuje log poruke u fajlove, koje zatim čita Logstash pipeline softver. Konfiguracioni fajl *syslog-ng.conf* mora imati bar tri osnovna parametra, *source*, *log* i *destination*, prikazana u nastavku:

```
source s_udp { udp(); };
log {
  source(s_udp);
  destination (d_logstash);
};
destination d_logstash {
       file("/opt/logstash/$SOURCEIP/$FACILITY-$YEAR-$MONTH-$DAY"
       owner("logstash") group("logstash") perm(0600)
       create_dirs(yes) dir_perm(0770));
};
```

Uloga Logstash pipeline softvera je da učita, formatira i filtrira log poruke, a primer konfiguracionog fajla je dat u nastavku:

```
input {
    file {
        path => "/opt/logstash/147.91.x.x/*"
        start_position => "beginning"
        sincedb_path => "/dev/null"
    }
}

filter {

    grok {
        patterns_dir => ["./patterns"]
        match => { "message" => "%{TIMESTAMP_ISO8601:time} %{SYSLOGHOST:
            syslog_hostname} %{DATA:syslog_program}(?:\[%{POSINT:syslog_pid}\])?: %{
            ACCESS:access}: IdP=%{IDP:IdP} MAC=%{MAC:MAC} AP=%{AP:AP} RP=%{RP:RP}"}

    }

        translate {
        source => "AP"
        target => "[APalias]"
        dictionary_path => "/usr/share/logstash/eduroam_lookup.json"
        fallback => "Unknown"
        override => true
      }
if [APalias] == "Unknown" {
        mutate {
        rename => {"[APalias]" => "[APnew][AP_name]"}
        add_field => {
             "[APalias][Grad]" => "Unknown"
             "[APalias][Lokacija]" => "Unknown"
             "[APalias][Latitude]" => "Unknown"
             "[APalias][Longitude]" => "Unknown"
             }
        }
    }

       mutate {
       remove_field => [ "@version", "syslog_program", "log", "@timestamp", "
           syslog_pid", "event", "host" ]
       }

}
output {

    elasticsearch {
         ssl => true
         ssl_certificate_verification => true
         cacert => "/etc/elasticsearch/certs/http_ca.crt"
         hosts => "https://147.91.x.x:9200"
         index => "monitoring"
         user => "elastic"
         password => "xxx"
    }
```

```
    stdout { codec => rubydebug }
}
```

Logstash pipeline softver čita log poruke koje syslog zapisuje u fajlove na serveru, za šta je zaslužan *input* segment, na početku konfiguracionog fajla. U okviru *filter* segmenta, log poruka se formatira na osnovu polja koja Logstash pipeline softver prepoznaje kao obrasce, a koja su definisana u *patterns-dir* parametru, a zatim mapirana u *match* parametru. *Patterns* fajl je dat u nastavku:

```
ACCESS .*
IDP .*
MAC .*
AP .*
RP .*
Longitude .*
Lokacija .*
AP_name .*
Latitude .*
Grad .*
```

Na osnovu podataka koji su poslati Logstash pipeline serveru od strane RADIUS servera ne mogu se dobiti svi željeni podaci o uspešnim i neuspešnim autentifikacijama. Zato je potrebno uvesti dodatne podatke kroz *lookup* fajl, koji se zatim povezuju sa podatkom o AP MAC adresi u *translate* segmentu. Ukoliko se taj podatak ne nalazi u fajlu, parametrima loga se dodaju “Unknown” vrednosti. Deo *lookup* fajla koji AMRES koristi dat je u Tabeli 2.

Tabela 2: Prikaz dela eduroam lookup fajla koji se koristi za formatiranje log poruke u Logstash pipeline softveru.

Lokacija	Grad	APmac	APname	Latitude	Longitude
ETF	Beograd	00-3a-7d-xx-xx-xx:eduroam	cisco2702-amres-bg.etf1	44.80556	20.47623
ETF	Beograd	00-3a-7d-xx-xx-xx:eduroam	cisco2702-amres-bg.etf10	44.80556	20.47623

Na osnovu uvedenih dodatnih podataka kreiran je sadržajniji log zapis i obogaćena je mogućnost pretrage i analize log poruke, a pridodat je grafički prikaz željenih izveštaja. U *output* sekciji pored konfiguracije destinacije na koju Logstash pipeline prosleđuje formatirane i filtrirane logove, najvažnija opcija je specificiranje indeksa za obrađene log poruke koji se šalju Elasticsearch softveru. Primer novokreirane log poruke prikazan je u nastavku:

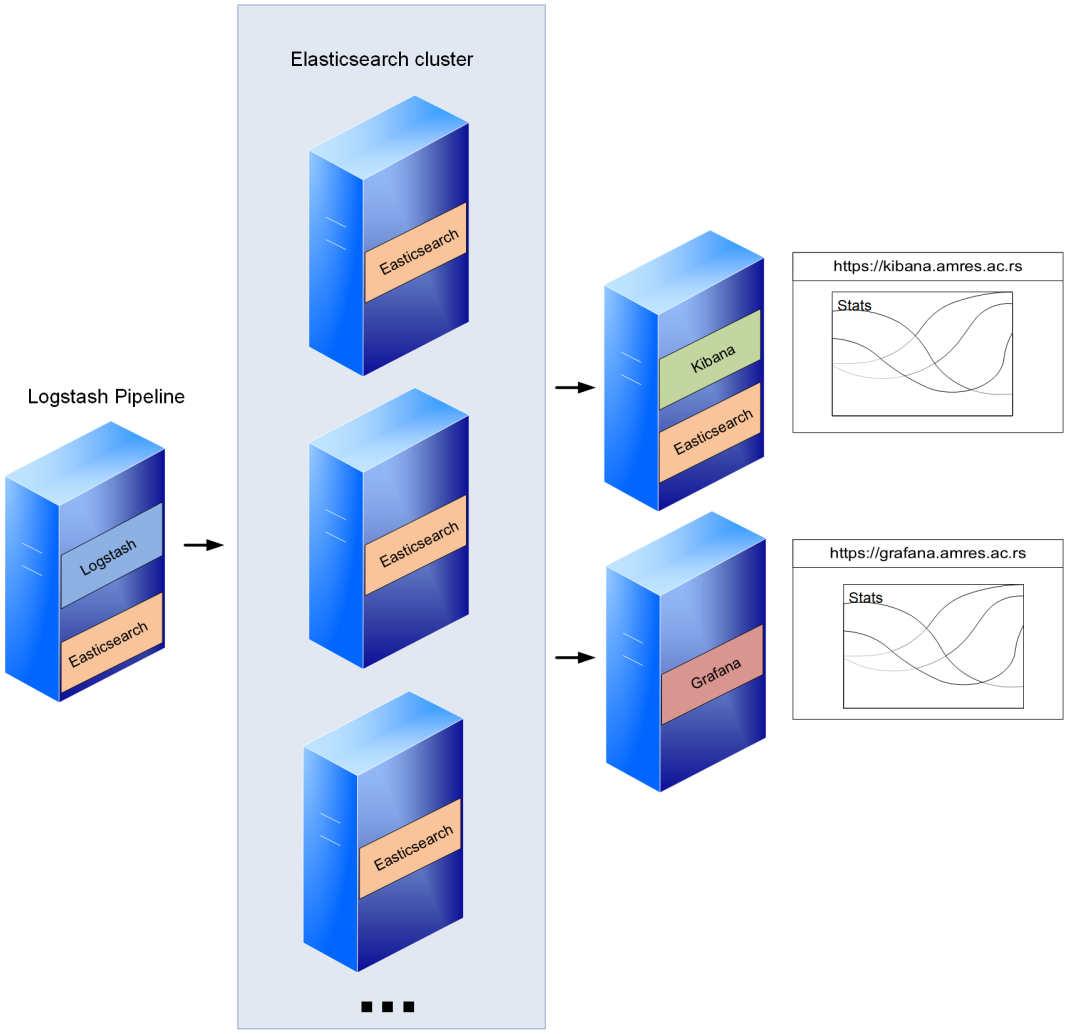
```
{
             "time" => "2022-06-01T15:30:01+02:00",
               "AP" => "00-3a-7d-xx-xx-xx:eduroam",
               "RP" => "1amres.ac.rs",
   "syslog_hostname" => "147.91.x.x",
              "MAC" => "b2-f8-f8-xx-xx-xx",
          "message" => "2022-06-01T15:30:01+02:00 147.91.x.x radiusd[15246]: Access-
              Accept: IdP=edu.arh.bg.ac.rs MAC=b2-f8-f8-xx-xx-xx AP=00-3a-7d-xx-xx-
              xx:eduroam RP=1amres.ac.rs",
           "access" => "Access-Accept",
          "APalias" => {
       "Lokacija" => "Elektrotehnicki fakultet Univerziteta u Beogradu",
```

```
        "Latitude" => "44.805563",
           "Grad" => "Beograd",
       "Longitude" => "20.47623",
         "AP_name" => "cisco2702-amres-bg.etf30"
   },
              "IdP" => "edu.arh.bg.ac.rs"
}
```

3.2 Konfiguracija Elasticsearch softvera

Filtrirane i formatirane log poruke se prosleđuju Elasticsearch softveru. Elasticsearch softver skladišti podatke pri čemu je najoptimalnije rešenje konfigurisati klaster koji se sastoji od jednog master noda i više običnih nodova, nad kojima se vrše upiti od strane Kibana komponente softvera, a koji omogućuje veću pouzdanost i skalabilnost celog sistema.

Slika topologije za prikupljanje i skladištenje logova, uključujući Elasticsearch klaster, prikazana je na Slici 2.



Slika 2: Postupak prikupljanja i skladištenja log poruka AMRES servisa.

Sve informacije o konfiguraciji klastera dostupne su u izlazu sledece komande koju omogućuje Elastic-

search API:

```
# curl --cacert /etc/elasticsearch/certs/http_ca.crt -u elastic https://147.91.x.x
   :9200/_cat/nodes?v
Enter host password for user 'elastic':
ip  heap.percent ram.percent cpu load_1m load_5m load_15m node.role master name
147.91.x.x        77         97  4    0.54   0.40    0.28 cdfhilmrstw -   node-3
147.91.x.x        50         97  21   0.96   0.75    0.45 cdfhilmrstw -   node-4
147.91.x.x        55         98  1    0.07   0.10    0.07 -         -     node-2
147.91.x.x        66         96  3    0.01   0.06    0.05 cdfhilmrstw *   node-1
```

U izlazu ove komande se može uočiti da je node-1 master nod, a njegova funkcionalnost se može promeniti. Funkcionalnost noda zavisi od uloge koju vrši u klasteru. Da bi Logstash upešno prosleđivao logove i upisivao u bazu, u okviru Elasticsearch softvera definisan je index na master nodu, što se postiže sledećom komandom:

```
# curl -X PUT --cacert /etc/elasticsearch/certs/http_ca.crt -u elastic https
   ://147.91.x.x:9200/monitoring?pretty
Enter host password for user 'elastic':
{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "monitoring"
}
```

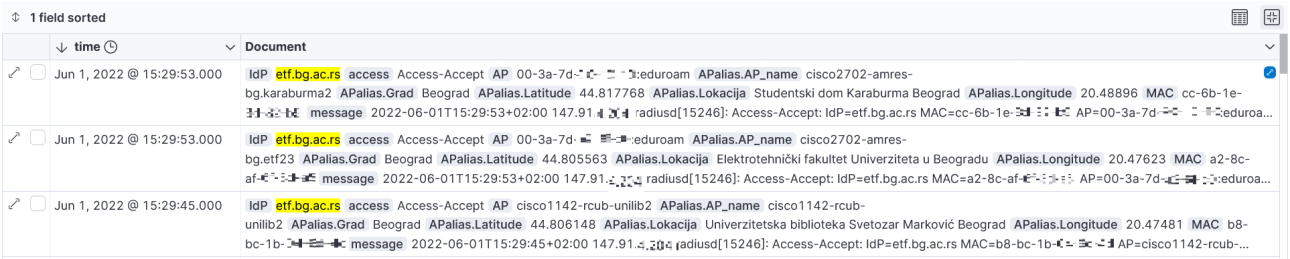
Nakon uspešne konfiguracije, korišćenjem Elasticsearch API komande mogu se pretražiti prikupljeni logovi, a u nastavku je dat primer strukture jednog takvog loga u JSON formatu, na osnovu upita i pretrage po domenu Davaoca Identiteta, tj. domena institucije:

```
# curl -X GET --cacert /etc/elasticsearch/certs/http_ca.crt -u elastic https
   ://147.91.x.x:9200/monitoring/_search?pretty -H 'Content-Type: application/json'
    -d'
> {
>   "query": {
>     "match": {
>       "IdP":"etf.bg.ac.rs"
>     }
>   }
> }
> '
Enter host password for user 'elastic':
{
  "took" : 23,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 319,
      "relation" : "eq"
    },
```

```
"max_score" : 2.3934784,
"hits" : [
  {

    "_index" : "monitoring",
    "_id" : "TTWV2YIBCUCBlsdkFJDV",
    "_score" : 2.3934784,
    "_source" : {
      "access" : "Access-Accept",
      "RP" : "1amres.ac.rs",
      "AP" : "00-3a-7d-x-x-x:eduroam",
      "IdP" : "etf.bg.ac.rs",
      "syslog_hostname" : "147.91.x.x",
      "time" : "2022-06-01T15:14:24+02:00",
      "APalias" : {
        "Longitude" : "20.461087",
        "Grad" : "Beograd",
        "AP_name" : "cisco2702-amres-bg.med22",
        "Latitude" : "44.797416",
        "Lokacija" : "Medicinski fakultet Univerziteta u Beogradu"
      },
      "message" : "2022-06-01T15:14:24+02:00 147.91.x.x radiusd[15246]: Access-
         Accept: IdP=etf.bg.ac.rs MAC=98-0d-51-x-x-x AP=00-3a-7d-x-x-x:eduroam RP
         =1amres.ac.rs",
      "MAC" : "98-0d-51-x-x-x"
    }
  },
```


Za razliku od Elasticsearch softvera koji prikazuje podatke u JSON formatu, Kibana omogućuje grafički prikaz log poruka i njihovu dalju obradu u smislu kreiranja statistika i vizualizaciju podataka. Primer log poruke koju prikazuje Kibana dat je na Slici 3.



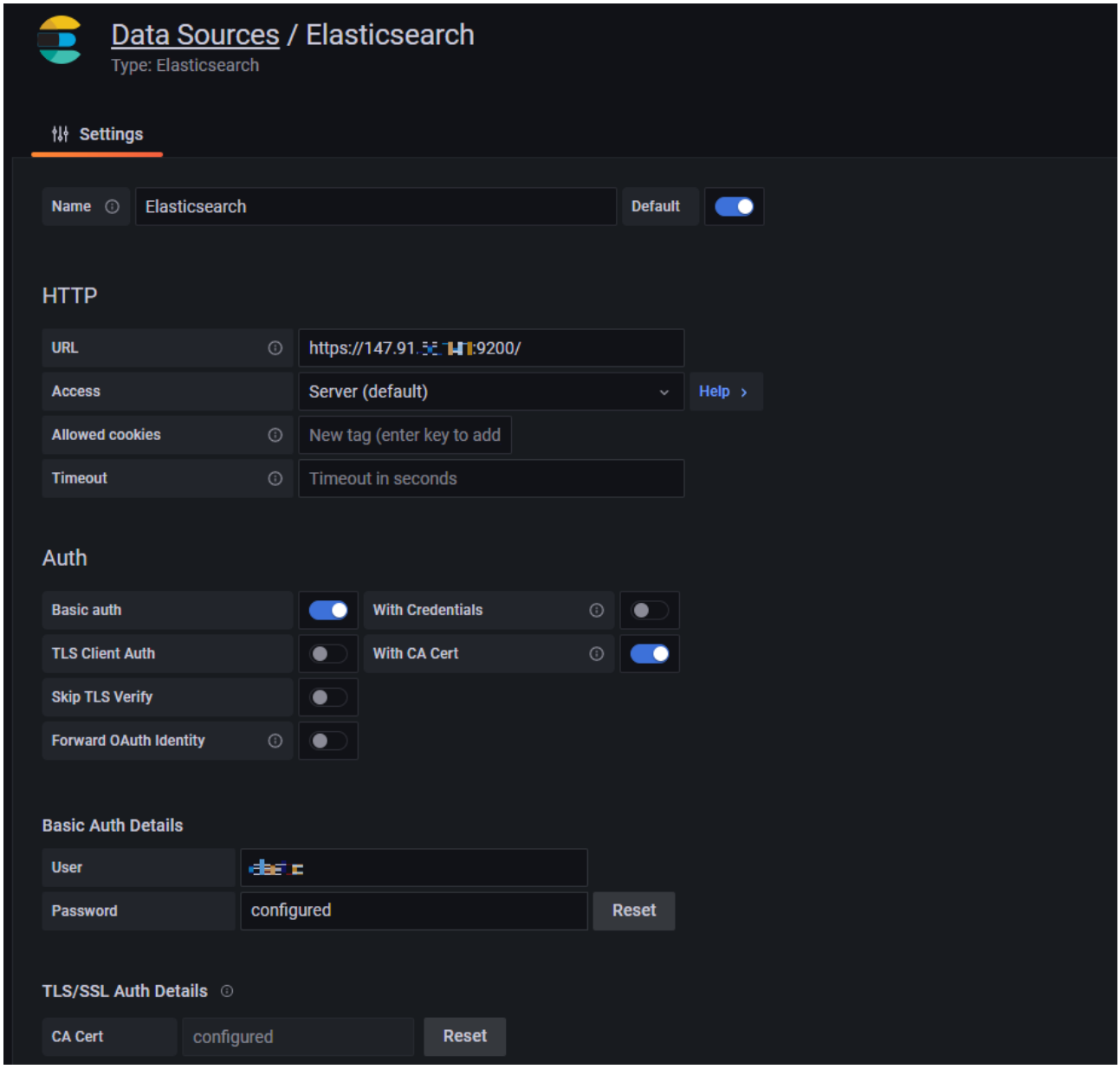
Slika 3: Primer formatirane log poruke koju prikazuje Kibana.

Sa slike se mogu videti uspešne autentifikacije AMRES krajnjih korisnika koji imaju otvoren digitalni identitet kod matične institucije koja je u ovom slučaju Elektrotehnički fakultet, a prijavili su se na eduroam servis sa tri različite lokacije:

- Studentski dom Karaburma Beograd,
- Univerzitetska Biblioteka Svetozar Marković Beograd,
- Elektrotehnički fakultet Univerziteta u Beogradu.

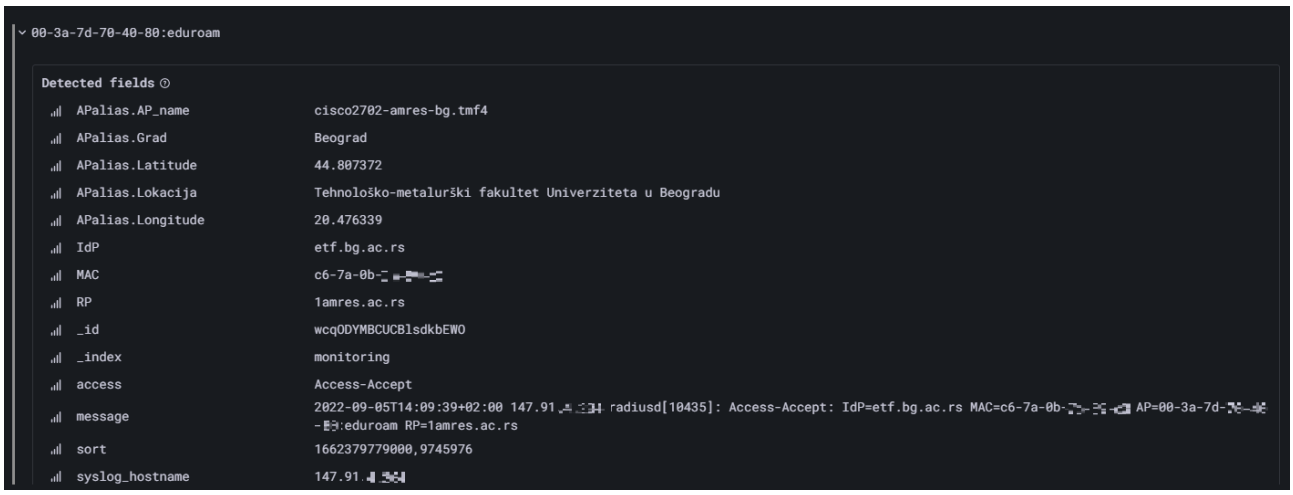
3.3 Konfiguracija Grafana softvera

Elasticsearch softver pruža mogućnost pristupa podacima koje skladišti u JSON formatu, koristeći Elasticsearch API. Dodatno, veliki broj aplikacija sada obuhvata i dodatke za integraciju sa Elasticsearch izvorom podataka. U ovu grupu se ubraja i Grafana softver, koji predstavlja veoma moćan alat za analizu i grafički prikaz podataka. Podešavanje izvora log poruka koje Grafana čita i koristi je veoma jednostavno i prikazano je na Slici 4.



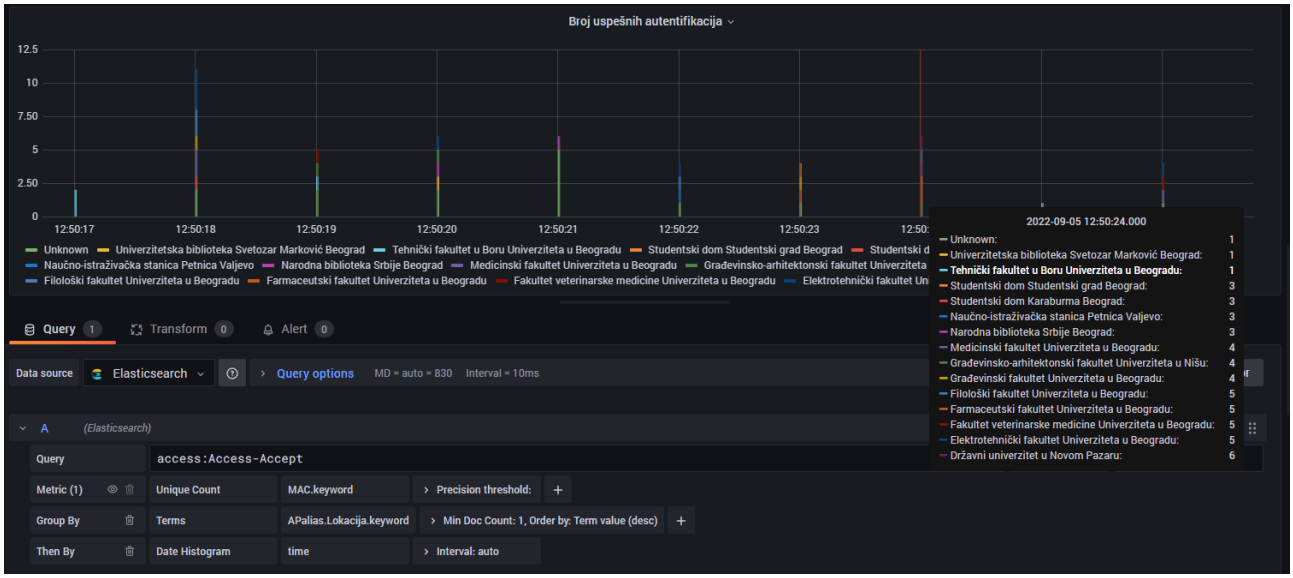
Slika 4: Konfiguracija Elasticsearch izvora podataka u okviru Grafana softvera.

Da bi Grafana softver mogao da učita log poruke sa udaljenog servera, potrebno je podesiti adresu Elasticsearch softvera i autentifikacione parametre, pošto je Elasticsearch zaštićen sertifikatom i za pristup podacima je neophodna autentifikacija i CA sertifikat. Nakon uspešne autentifikacije, log poruke su dostupne za prikaz i dalje korišćenje u okviru upita za pretragu, filtriranje, transformacije, itd. Na Slici 5 dat je prikaz logova u Grafana softveru.



Slika 5: Primer log poruka koju prikazuje Grafana softver.

Slika 6 ilustruje samo jedan od mnogobrojnih upita koji se mogu upotrebiti za grafički prikaz statistika dobijenih na osnovu prikupljenih podataka.



Slika 6: Primer Grafana upita kojim se vizualizuju statistike korišćenja eduroam servisa

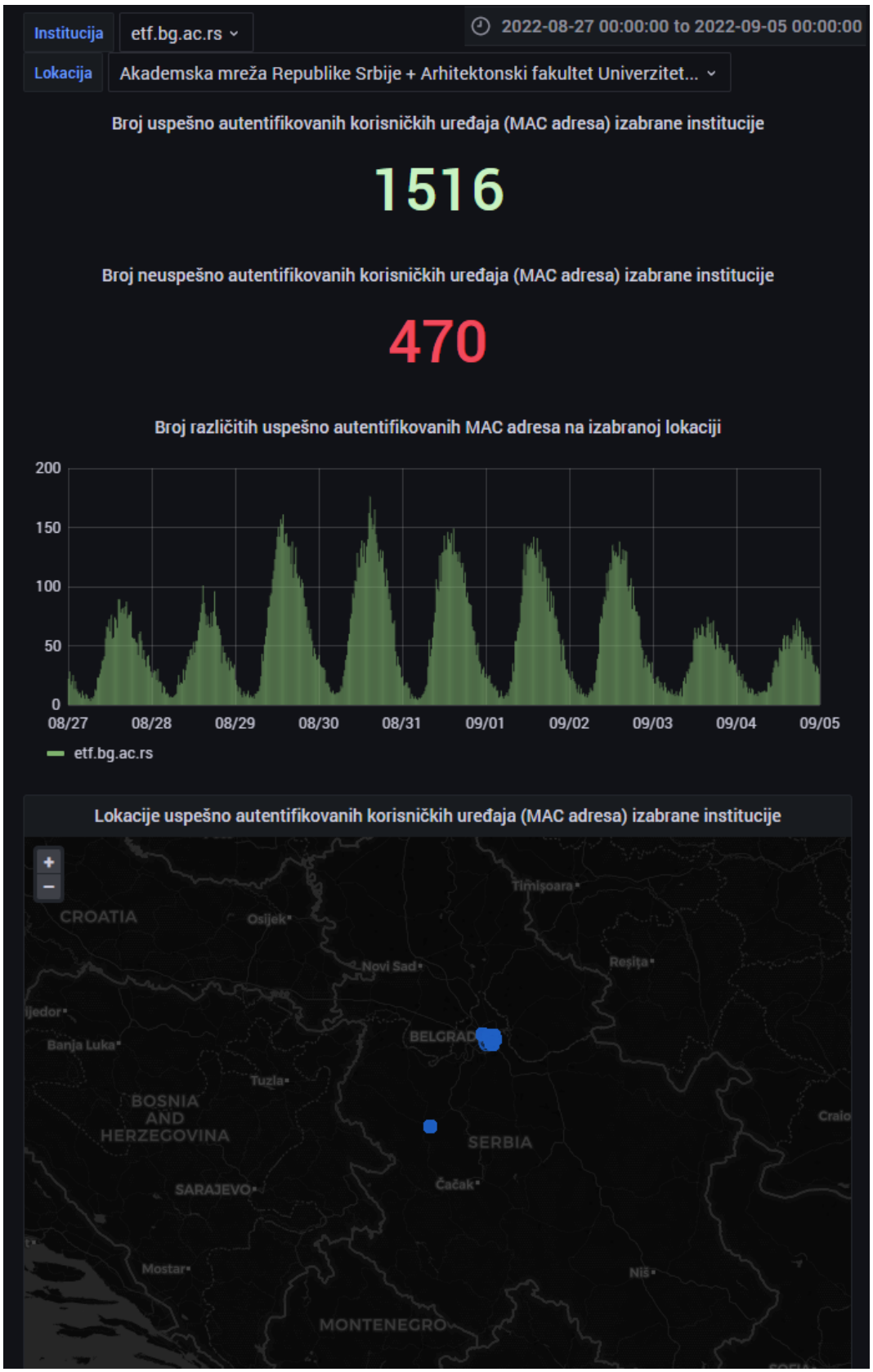
Pretraga u okviru Grafana softvera omogućava detaljnu analizu log poruka i olakšava praćenje performansi i korišćenja servisa. Svaki realizovan upit može se sačuvati kao panel, za kasnije korišćenje. Sačuvani paneli se zatim mogu spojiti u *Dashboard*. U okviru Grafana softvera *dashboard* predstavlja zajednički prikaz kreiranih izveštaja, koji mogu biti u različitim formatima.

AMRES prati i izdvaja veliki broj podataka iz log poruka i pravi statistike o različitim aspektima korišćenja eduroam servisa. AMRES prati veliki broj parametara bitnih za eduroam servis, koji se mogu podeliti u pet grupa:

- Svi korisnici,
- AMRES korisnici,

- Korišćenje po instituciji,
- Korišćenje po lokaciji,
- Strani korisnici.

Za ove grupe informacija od značaja kreiran je Dashboard, čiji je primer prikazan na Slici 7.



Slika 7: Primer grafičkog prikaza podataka dobijenih primenom različitih upita u okviru Grafana softvera.

4 Zaključak

Elastic Stack softver pruža velike mogućnosti za prikupljanje log poruka i na taj način efikasno i kvalitetno prati i analizira rad i korišćenje servisa u mreži. Softver koji je razmatran je više nego dovoljan alat koji administratorima omogućava da na brz i efikasan način prikupljaju i pretražuju logove iz različitih izvora, vrše monitoring servisa i prate ponašanje korisnika. Implementirani sistem predstavlja sveobuhvatno i skalabilno okruženje, u smislu količine i obima log poruka koje je potrebno skladištiti i obraditi. Takođe, pruža veliku pouzdanost u radu i mogućnost gubitaka podataka je svedena na minimum. Iako je uloga Grafana softvera u ovom radu vizualizacija i analiza logova, on se primarno koristi i kao alat za monitoring rada servisa i prikaz i analizu različitih metrika.

Literatura

[1] Murugiah Souppaya and Karen Scarfone. Nist special publication 800-92, Guide to computer security log management, 09 2006.

[2] Elastic stack. https://www.elastic.co/. Accessed: 2022-09-01.

[3] Grafana. https://grafana.com/. Accessed: 2022-09-01.

[4] syslog-ng website. https://www.syslog-ng.com/. Accessed: 2022-09-01.

[5] Splunk log management. https://archive.geant.org/projects/gn3/geant/services/cbp/Documents/cbp-48_splunk_log_management_amres.pdf. Accessed: 2016-03.

[6] freeradius website. https://freeradius.org/. Accessed: 2022-09-01.

[7] RFC2865. https://www.rfc-editor.org/rfc/rfc2865. Accessed: 2022-09-01.

[8] Elasticsearch as a NoSQL database. https://www.elastic.co/blog/found-elasticsearch-as-nosql. Accessed: 2022-09-01.

[9] Elasticsearch RESTful API. https://www.redhat.com/en/topics/api/what-is-a-rest-api. Accessed: 2022-09-01.

[10] JSON. https://www.json.org/json-en.html. Accessed: 2022-09-01.

Spisak autora(ki) i predavača / List of Authors and Speakers


Универзитет у Београду – Електротехнички факултет
Микроталасна мерења
-уџбеник у електронском облику-
Побудна антена
AUT
10 dB
≤ −0,25 dB !
Мерење
Симулација
PNA
OCXO
YIG
LO
DUT
Миодраг С. Тасић
Београд, 2022.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

004(082)

НАЦИОНАЛНА конференције са међународним учешћем "Примена слободног софтвера и отвореног хардвера" (5 ; 2022 ; Београд)
Zbornik pete nacionalne konferencije sa međunarodnim učešćem pod nazivom Primena slobodnog softvera i otvorenog hardvera PSSOH 2022, [Beograd] / [urednički i organizacioni odbor, editorial and organizational board Nadica Miljković, Predrag Pejović, Miloš Cvetanović]. - Beograd : Univerzitet, Elektrotehnički fakultet : Akademska Misao = Belgrade : University, School of Electrical Engineering : Academic Mind, 2023 (Beograd : Akademska Misao = Academic Mind). - 59 str. : ilustr. ; 30 cm

Na spor. nasl. str.: Proceedings of the Fifth National Conference with International Participation titled Application of free software and open hardware PSSOH 2022. - Radovi na srp. i engl. jeziku. - Tiraž 50. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-7466-973-0 (AM)

а) Рачунарство -- Зборници

COBISS.SR-ID 117321737

Konferenciju podržali

LOTUS
FLARE