Romanian Reports in Physics **72**, 409 (2020)

CLEANSING DATA FROM THE HIGH-POWER LASER SYSTEM IN ELI-NP: A HOLISTIC SYSTEM APPROACH

G. KOLLIOPOULOS, G. PRODAN, B. BOISDEFFRE, I. DANCUS

Extreme Light Infrastructure – Nuclear Physics & "Horia Hulubei" – National Institute for R&D in Physics and Nuclear Engineering, Laser System Department, P.O. Box MG-6, RO-077125 Bucharest–Magurele, Romania
Corresponding author, e-mail: *georgios.kolliopoulos@eli-np.ro*

Received December 13, 2019

Abstract. A method aiming to reduce the amount of data that should be stored at the ELI-NP data center is presented here. Based on a holistic system approach ([1] – L. Bertalanffy, General System Theory, George Braziller, Inc. 1968) on the High Power Laser System (HPLS), it consists of a binary classifier which marks the records either as "useful" or "useless".

Key words: binary classifier, High Power Laser System, holistic system approach.

1. INTRODUCTION

The European Strategy Forum on Research Infrastructures decided, in 2006, to build a pan-European distributed research facility, the Extreme Light Infrastructure (ELI) [2]. Three European countries, the Czech Republic, Hungary and Romania, would host different but complementary experiments in three distinct facilities: ELI-beamlines [3], ELI-ALPS [4] and ELI-NP [5, 6] respectively; all three will make use of High-Power Laser Systems (HPLS).

The HPLS in ELI-NP has been already implemented in Magurele – Romania by Thales (*https://www.thalesgroup.com*). It is composed of two arms (Arm A and Arm B), each of which can deliver ultra-short pulsed lasers of three distinct power values at three distinct repetition rates: 100 TW at 10Hz, 1 PW at 1 Hz and 10 PW at 1/60 Hz. The HPLS comprises hundreds of devices integrated in a Supervisory Control and Data Acquisition (SCADA) type system [7, 8]. Each one of these devices, according to its position in the System, operates with the proper repetition rate: 10 Hz, 1 Hz or 1/60 Hz. During operation hours, data are continuously acquired concerning both the current condition of each device (if it is activated, if it works properly, etc.) and the measurements which are being performed. The frequency of acquisition follows the repetition rate corresponding to each specific device. As a result, from one single device operating at a repetition rate of 10 Hz, half a million of records are acquired for just 14 hours of operation (*e.g.* between 8:00 am and 10:00 pm on a working day).

The total size of data acquired daily can exceed the 700 GB (350 GB from each one of the two Arms). More than 90% the total volume of these data corresponds to spectra and beam-profile pictures, acquired by spectrometers and cameras, respectively. It has been realized though that a big number of acquired images do not contain any laser beam-profile and an equally big number of acquired spectral records do not contain any spectrum of the laser pulse. Diagnostic devices which are left activated without being triggered externally, will keep recording data even without any presence of the laser beam. This results in a big number of records, especially from devices operating with a repetition rate of 10 Hz, without any useful information.

In order to eliminate pictures without valuable information, an algorithm has been already proposed [9] to detect the presence (or absence) of a laser beam profile in them. In this work we offer an alternative solution: a universal binary classifier (UBC) is introduced, which distinguishes the potentially useful data (in terms of images, spectra, energy values, etc.) from the ones that do not contain any signal at all. This UBC avoids the time-consuming direct processing of the large objects (pictures as large 2-dimensional matrices, spectra as long arrays) themselves and focuses instead on the necessary conditions which need to be fulfilled for the corresponding monitored laser unit to be in an activation state. The proposed Universal Binary Classifier has been already applied to a small part of the data recorded, during the implementation phase of the HPLS, at the exit of one of the laser units. Nevertheless, the idea generalizes to the data recorded at the exit of every laser unit of any similar High Power Laser System in the world.

This work is the beginning of a larger project that aims to use Big Data tools in order to analyze systematically the data generated during the HPLS operation in order to optimize the maintenance and the operation performances. In the future, laser system's data will be correlated with experimental data acquired by the ELI-NP control and monitoring system for experiments [10].

2. THE METHOD

For cameras to capture beam profiles and spectrometers to record laser pulse spectra, laser activity is a prerequisite. Instead of examining separately each one of the pictures acquired by the cameras or each one of the spectra acquired by the spectrometers, it is much more convenient and much faster, to determine the time periods for which there is indication that lasing activity takes place at the specific unit. A kind of holistic system approach is necessary here; meaning that information acquired from different devices in the HPLS should be combined. An appropriate selection criterion, tailored to the specific features of each HPLS laser unit (oscillators, amplifiers, etc.), should be defined. As it is shown in Fig. 1, for all together the

diagnostic tools (cameras, spectrometers, etc.) attached to the unit, according to this criterion the recorded data can be classified as "useful" or "useless".

In this work, 1st Amplifier (AMP1) [11] on Arm B of the HPLS will be used as a study case. The cleansing process has been applied on spectra and beam profiles data recorded at the exit of the unit. In principle though, this holistic system approach can be applied on a variety of units that Amplify Light by Stimulated Emission of Radiation; only the selection criterion needs to be specified accordingly.

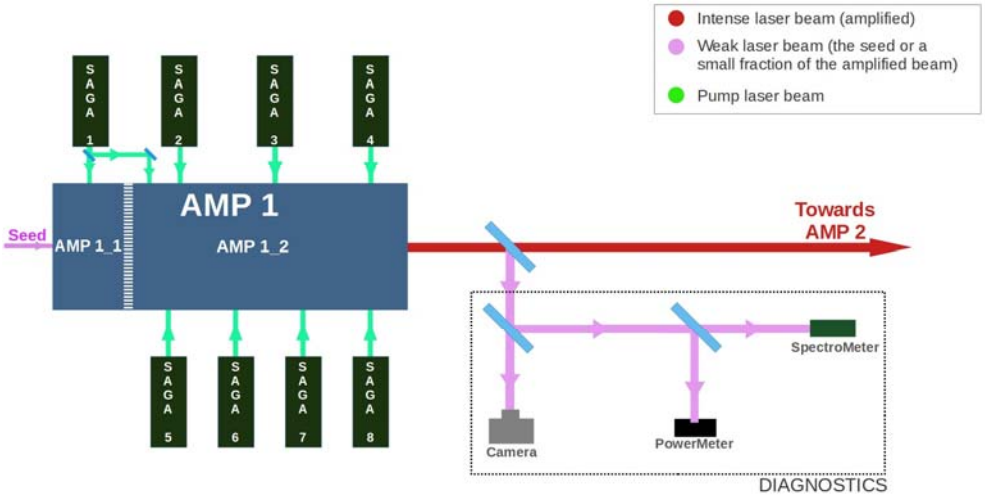

Fig. 1 – Schematics of 1st Amplifier (AMP1) of Arm B in the HPLS. Whenever there is not amplified laser beam coming out from the unit, all together the diagnostics do not record any signal.

Information in time about the condition of each one of the eight SAGA [12] pumps can be retrieved by the data records concerning these devices. A list of modulations in the pumping condition of AMP1 is saved in chronological order in a file of JavaScript Object Notation (JSON) [13] format, which offers a very fast parsing of the stored data. A graphical representation of this combined information, on November 29, 2018 in the afternoon, can be seen in Fig. 2.

Once the information on the activity of the pump-lasers is retrieved, the data cleansing can be seen as a three-step-process:

i. For every record, one takes into account the time of acquisition.
ii. By querying the JSON file, the pumping condition at the acquisition time is obtained.
iii. Application of the selection criterion. If AMP1, at the specific time, is pumped with at least 3 pumps and SAGA_1 is among them, the record is labeled as "useful"; if not, the record is labeled as "useless" and can be discarded.

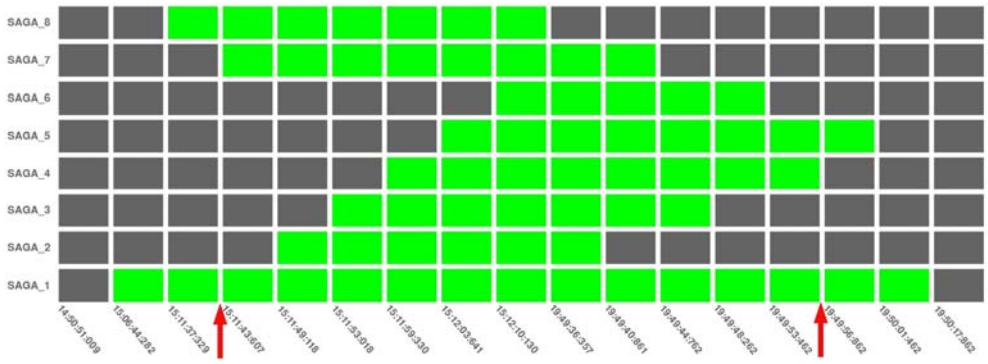

Fig. 2 – Pumping in AMP1 of ARM B on November 29, 2018 (afternoon). The rows correspond to each one of the pump-lasers. The two red arrows indicate the total time interval for which the selection criterion is fulfilled.

In Fig. 2, can be observed the subsequent activation of all the 8 pump-lasers (SAGAs) from 15:06:44:282 up to 15:12:10:130 (the time is given by the acquisition timestamp with the format: "hour:minute:second:millisecond"). In between 15:12:10:130 and 19:49:36:357, all the SAGA pumps are pumping the amplifier AMP1. After 19:49:36:357, it is observed the subsequent deactivation of the pumping-lasers until 19:50:17:862. According to the selection criterion, we should not be expecting any laser signal to be recorded by anyone of the diagnostic devices before 15:11:43:607 (activation of a 3rd pump-laser) and after 19:49:56:862 (less than 3 pump-lasers left active). This is indicated in Fig. 2 by the two red arrows.

AMP1 becomes fully operational only after the activation of all the 8 pump-lasers. Nevertheless, during the process of activating one by one the SAGAs, the system passes through a transition during a long part of which some signal can be detected. This signal is, as expected, smaller than the signal obtained at full operation; it can, however, offer useful information about the contribution of the individual pump-lasers in the amplification process.

In Fig. 3 one can see the transition phase depicted on the beam-profile and the spectrum acquired at the exit of the AMP1. The data were recorded on October 3rd, 2019. The careful reader will observe that some signal on the spectrometer starts to appear already after the activation of 4 pump-lasers. On the pictures recorded by the camera though, one can see a clear beam profile only after the activation of 6 pump-lasers. This is mainly due to the fact that the specific camera proves itself much more sensitive to noise in comparison to the spectrometer. In order to check this hypothesis a simple signal enhancing process is applied on the beam-profile pictures in the middle column of the Fig. 3. It consists on:

- selecting the pixels with a value above the mean value
- replacing their value by the maximum one

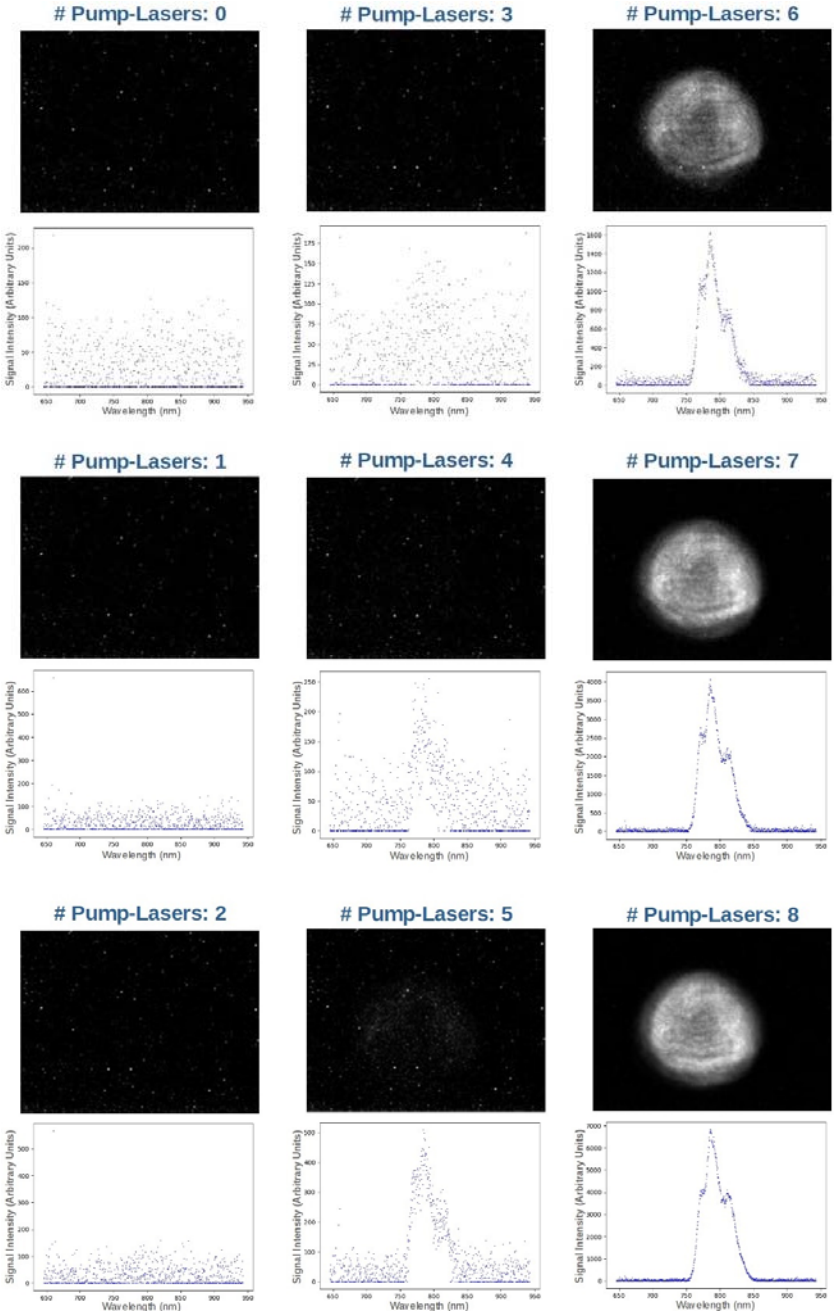

Fig. 3 – Data acquired by the camera and the spectrometer at the exit of AMP1 for different numbers of activated pump-lasers. The matrices corresponding to the pictures are normalized in respect to the maximum value. The gray-scale is linear in between 0 and 1. In the spectrograms, the numbers in the perpendicular axis are arbitrary. Nevertheless, in all the cases it has been used the same calibration.

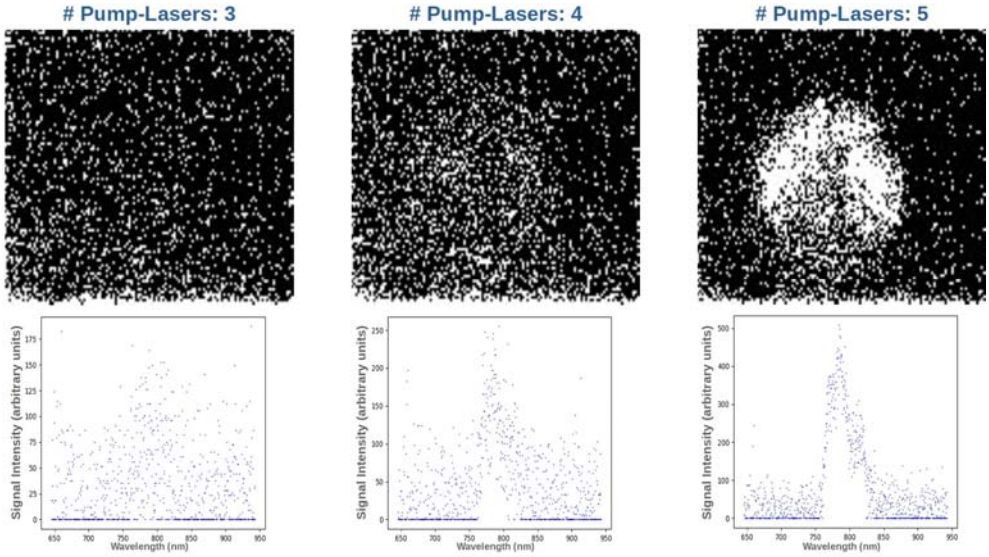

Fig. 4 – The three mid-column pictures in Fig. 3 after applying the signal enhancing process described in the text.

By applying this simple process on the beam-profile pictures with 3, 4 and 5 activated pump lasers, one obtains the results shown in Fig. 4. There, one can see a clear beam profile signal recorded with five activated pump-lasers. One can also observe some weak indication of signal even with four.

The question that is raised at this point is the following: since, after processing, some signal can be observed with only the activation of four pump-lasers, why was the selection criterion chosen to be looser than that? Why do we keep data acquired even with three activated pump-lasers? The answer is that the classifier is built to be biased on purpose; the selection is biased with respect to the rejection. It is thus preferred for the classifier, by mistake, to select data without any signal than to reject data with even some potentially useful weak signal.

3. EVALUATION AND DISCUSSION

The classification method presented in this work is about to be applied on hundreds of thousands of acquisitions per day by just one device. Consequently, it is not possible for a human to go through all these data and establish a reliable ground truth of significant size for the evaluation of the UBC. In a similar context, a model for the evaluation of a binary classifier without ground truth has been recently proposed [14]. Here, in order to circumvent the absence of ground truth, a second binary classifier (SBC), for data classification on spectra only, has been built and its classification results have been cross-checked with the results of the

one proposed in this paper. Furthermore, one by one the points of disagreement between the two classifiers have been examined on close inspection by human.

Data acquired during four days of important activity (meaning several hours of operation) in AMP1 (November 20, 2018; November 28, 2018; November 29, 2018; October 3, 2019), are examined by both the binary classifiers. The cases for which there is a disagreement between the two are taken into account and examined afterwards one by one. A false rejection or a false selection made by the UBC can be caught by the SBC and confirmed by the direct examination which follows. An overview of the evaluation can be seen in Table 1.

Table 1

Overall evaluation taking into account four days of activity

Total Number of acquisitions	1,511,822
Selected by the UBC	995,822
Rejected by the UBC	516,000
Rejection Ratio	**34.13%**
Falsely Selected by the UBC according to the SBC	1161
Falsely Selected after examination by human	436
Falsely Rejected by the UBC according to the SBC	8
Falsely Rejected after examination by human	2
Ratio of False Selection by the UBC	**0.044%**
Ratio of False Rejection by the UBC	**0.0004%**

The SBC is a pattern recognition algorithm simplified enough in order to be fast. It is written in Python [15] and, using the NumPy library [16], goes through and examines each spectral array record in order to decide if it belongs to the "useful" class (it contains signal) or to the "useless" one (it does not contain signal). The examination is based on the following process:

– Evaluation of a Threshold Value (TV) which is considered to be an upper barrier of the noise and that can be exceeded only in the presence of a laser beam. The TV is calculated by examining a small part of the array at a spectral area far away from where the laser pulse spectrum is expected to be found.
– Through an iterative process, the TV is compared with the values of the spectral array. If at least 10 values are found to be bigger than the TV, the spectral array is considered to contain signal.

The SBC offers a valuable cross-check on the results of the UBC, as it consists of a completely different classification approach. Wherever they fail, they fail for different reasons. As it is presented in Fig. 5, the SBC happens to mis-classify a few very-low-signal cases as no-signal ones; it is also sensitive to a few cases of exceptional noise. The UBC on the other hand, fails either in case of a major system failure or due to imperfect synchronization between the different devices in the system. This imperfect synchronization comes from the fact that the timestamp

given to each record is a software timestamp provided by the computer in which the specific record is saved. The software synchronization in between various computers in HPLS is based on a Network Time Protocol (NTP) [17].

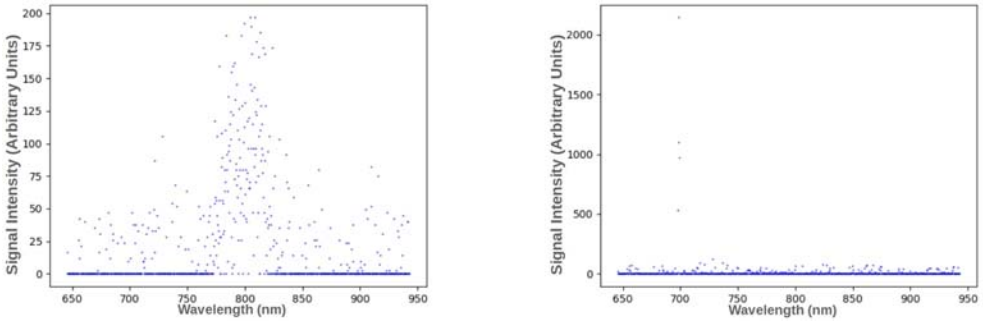

Fig. 5 – Two characteristic cases of mis-classification by the second binary classifier (SBC). In the picture on the left can be seen a very low signal spectrum which is labeled as "no signal". In the picture on the right, the SBC is confused by the four isolated points with extremely high values of noise and labels the record as "containing signal". The spectrum of the HPLS laser pulse can be found centered approximately at 800 nm, not at 700 nm [9].

As mentioned before, the selection criterion of the UBC is made on purpose to be loose. For this reason, one can observe a discrepancy of two orders of magnitude between the "False Selection" and the "False Rejection" ratios in Table 1 (two last rows). Both the spectra which appear on the 8th line to be "Falsely Rejected after examination by human" have been confirmed to belong to rare cases of an unusually poor synchronization between the spectrometer and the corresponding pump-laser.

An advantage of the UBC based on a holistic system approach is that it works for data acquired by any of the different diagnostic tools; cameras, spectrometers, energymeters, etc., all together. This is the reason why there is no need to evaluate it separately for data of another kind (*e.g.* beam profiles). If, at the exit of a specific unit, there is a laser beam for the spectrometer to record a spectral signal, it will be also there for the camera to record a beam-profile. If the beam-profile does not appear, it is a matter either of the sensitivity of the instrument or of the attenuation in front of it; a post-processing may reveal signal which is not able to be noticed directly (see Figs. 3 and 4).

In Table 1 can be seen that, for the four days which are examined, the "Rejection Ratio" is above 34%. This means that approximately the 1/3 of the acquisitions are not found to contain any useful information and, as a result, should not be kept. One has to keep in mind that 34% of economy in storage space is just indicative for four days of important activity in AMP1 on Arm B during the implementation phase of the HPLS (which ended in October 2019). It would not be correct, based on this number, to generalize for the whole period of the implementation phase which comprised days with very limited (or even not at all) laser activity in the specific unit but with the diagnostics let on. What can be said though is that the

use of the UBC will offer a cleansed database which, apart from a storage space economy, will allow any query to it to be faster and interesting data analysis projects to take place.

The classification process of the UBC is based on an algorithm which performs a small number of elementary operations. For each record to be classified, is taken into account only the acquisition time (which is just one value) and the chronologically sorted list of modulations in the activity condition of the correspondent laser unit. Performing a series of comparisons in a sorted list [18], the UBC identifies the proper activity condition time interval (see Fig. 2) and concludes if it should be expected for the record to contain signal or not. Avoiding to examine directly the content of the acquisition, the UBC becomes independent of the resolution used for the acquired data (images and spectra). Thus, the UBC offers the advantage that, for any big collection of data, the time complexity of the classification process depends only on the number of records and not on the size of each one of them.

4. SUMMARY

Based on a holistic approach to the system (combining information acquired by several devices and not just by one) a binary classifier has been built, able to distinguish in between "useful" and "useless" records at the exit of the 1st Amplifier (AMP1) on Arm B in the HPLS. This classifier is universal as it can be applied on any kind of data acquired by diagnostic instruments (pictures, spectra, energy values, etc.). For the evaluation of the proposed classifier it has been built a second classifier based on a completely different approach; it is a simplified spectrum pattern recognition algorithm. Both have been used for the spectra acquired during four different days of important activity in AMP1 and the results have been compared with each other. The efficiency of the proposed universal binary classifier has been demonstrated to be very high on the specific sample of days. The execution time of the universal binary classifier based on holistic system approach does not depend on the resolution of the acquired data.

Acknowledgments. This work was carried out under contract sponsored by the Ministry of Research and Innovation: PN 19 06 01 05. We thank our colleagues from the ELI-NP Laser System Department and Thales for their continuous support, in particular Andrei Gradinariu and Cristian Capiteanu.

REFERENCES

1. L. Von Bertalanffy, *General System Theory*, George Braziller, Inc., 1968.
2. G. Mourou et al., *ELI – Extreme Light Infrastructure Science and Technology with Ultra-Intense Lasers Whitebook*. [Online] Available from: https://eli-laser.eu/media/1019/eli-whitebook.pdf
3. B. Rus et al., *ELI-Beamlines laser systems: Status and design options*, Proc SPIE. 8780.

4. D. Charalambidis et al., *ELI-ALPS: implementation status and first commissioning experiments (Conference Presentation),* Proc. SPIE 11039, Research Using Extreme Light: *Entering New Frontiers with Petawatt-Class Lasers IV*, 110390M (14 May 2019).
5. F. Lureau et al., *10 PetaWatt Laser System for Extreme Light Physics*, Laser Congress 2019 (ASSL, LAC, LS&C), OSA Technical Digest (Optical Society of America, 2019), paper ATh1A.5.
6. S. Gales et al., *The Extreme Light Infrastructure – Nuclear Physics (ELI-NP) facility: new horizons in physics with 10 PW ultra-intense lasers and 20 MeV brilliant gamma beams*, Progress in Physics **81**, *9* (2018).
7. S. A. Boyer, *SCADA Supervisory Control and Data Acquisition*, USA: ISA – International Society of Automation, 2010.
8. A. Gotz et al., *TANGO V8 – Another Turbo Charged Major Release*, Proceedings of ICALEPCS 2013, San Francisco, CA, USA, 2013.
9. B. De Boisdeffre et al., *Images Processing Techniques and Data Analysis Applied to High-Power Laser Systems*, MMEDIA, 2019, p. 49.
10. M. Cernaianu et al., *Monitoring and Control Systems for Experiments at ELI-NP*, Romanian Reports in Physics, **68**, Supplement, S349–S443 (2016).
11. *The White Book of ELI Nuclear Physics Bucharest-Magurele, Romania*. [Online] Available from: https://eli-np.ro/whitebook.php
12. THALES, *SAGA HP Flashlamp-Pumped Nd:YAG Laser Series*. [Online] Available from: www.thales-laser.com
13. Ecma International, *Standard ECMA-404, The JSON Data Interchange Syntax*, 2nd Edition, December 2017. [Online] Available from: www.ecma-international.org
14. M. Fedorchuk et al., *Binary Classifier Evaluation Without Ground Truth*, Ninth International Conference on Advances in Pattern Recognition (ICAPR-2017), Dec 2017, Bangalore, India. hal-01680358
15. Brian Jones, David Beazley, *Python Cookbook*, 3rd Edition, May 2013.
16. S. Van der Walt et al., *The Numpy array: a structure for efficient numerical computation*, Computing in Science & Engineering, **13**, *2*, 22–30 (2011).
17. D. L. Mills, *Internet Time Synchronization: The Network Time Protocol*, IEEE Transactions on Communications, **39**, *10* (1991).
18. D. E. Knuth, *The Art of Computer Programming*, 3rd Edition, Volume 1, 1997.