Semantic data integration in upgrading hydro power plants cyber security

Z. Tabak*, H. Keko ** and S. Sučić**
* Elektroprivreda HZ HB, Mostar, Bosnia and Herzegovina
** KONČAR - Digital Ltd., Zagreb, Croatia
zoran.tabak@ephzhb.ba; hrvoje.keko@koncar.hr; stjepan.sucic@koncar.hr

***Abstract* - In the recent years, we have witnessed quite notable cyber-attacks targeting industrial automation control systems. Upgrading their cyber security is a challenge, not only due to long equipment lifetimes and legacy protocols originally designed to run in air-gapped networks. Even where multiple data sources are available and collection established, data interpretation usable across the different data sources remains a challenge. A modern hydro power plant contains the data sources that range from the classical distributed control systems to newer IoT-based data sources, embedded directly within the plant equipment and deeply integrated in the process. Even abundant collected data does not solve the security problems by itself. The interpretation of data semantics is limited as the data is effectively siloed. In this paper, the relevance of semantic integration of diverse data sources is presented in the context of a hydro power plant. The proposed semantic integration would increase the data interoperability, unlocking the data siloes and thus allowing ingestion of complementary data sources. The principal target of the data interoperability is to support the data-enhanced cyber security in an operational hydro power plant context. Furthermore, the opening of the data siloes would enable additional usage of the existing data sources in a structured semantically enriched form.**

Keywords – cyber security, industrial automation, hydro power plants, communication and information technologies, semantic alignment

I. INTRODUCTION

Hydroelectric power has been a crucial part of electric power systems for centuries, and the key developments in hydro power date back to mid-19th century. Hydro power will remain important well into the future. With years of increased installation of solar and wind power, a large share of world's renewable energy is produced in the hydroelectric power plants. Year-to-year growth in hydropower installation remains at more than 3% [1]. Hydropower not only provides an energy source with no emissions of greenhouse gases: its water management infrastructure is tied to potable water management, irrigation and flood management and control. In short, the importance of hydropower for a society that is ever more reliant on electric power and facing climate extremes can't be overstated.

Hydroelectric power plants have been early adopters of digital technologies. Starting with the early turbine-generator illuminated boards and panels populated with indicators and individual control stations dedicated to control elements, the first direct digital control (DDC) systems established in the 1970s have been mimicking the interfaces of the previous illuminated boards. These systems have largely been ineffective as these have been "zoomed into" a comparatively small subset of the process at any time. Though the user could switch the displayed process subset, only a reduced dataset could be visible at any time. This was known as "keyhole" effect. The user has effectively been peeking into the process seeing only a very reduced number of process variables.

The next generation of distributed control systems (DCS) [2] emerged with more performant and more capable computer displays available in the 1980s. These systems, for the first time, focused on the operator tasks. Historical information and trends could be monitored, and thus the process dynamics could be followed by the operator. Gradual improvements to the computer systems have allowed higher screen resolutions and multiple window support in the graphical user interfaces, which in turn allowed seeing more variables at the same time.

Further developments in the 1990s have delivered the first instances of expert support systems with help and guidance messages to the operator included in the functionalities of most control systems. In a typical setting, the control system functionalities have been expanded to the one of a knowledge base and guide. These support systems provide additional information to the operator, but no decisions are taken automatically by these systems – instead, the operator can rely on the support from the additional information integral to the control system in front of his eyes. These were the early precursors to today's artificial intelligence and machine learning techniques, started emerging already in mid to late 1980s. While automatic expert systems and neural networks have been utilized to model and eventually optimize functions lateral to the main hydro power plant process but have not been a part of the main line of work and have not been used as support in the operator decision making.

In the recent decades the shift from the highly customized proprietary technologies and designs to practically the same commodity infrastructure: the same hardware architecture and the same operating systems, as in conventional IT business. The industrial automation control systems have migrated and embraced commercial computing platforms based on x86 processor architecture and Windows or Linux platforms. In other words: the recent versions of automation platforms do not use exotic operating systems and databases anymore, but instead run on the very same Windows and Linux platforms that the

common IT utilizes. Though domain-specific highly performant historian databases such as Osisoft PI exist, the commonly used relational databases such as PostgreSQL, Oracle or MS SQL Server are also very commonly used. The application scope of the recent generation of supervisory control and data acquisition systems (SCADAs) has widened, too. Nowadays these systems serve not only for direct control as a means for the operator to directly control the process, but also as a source of timely information gathered directly from the process and provided to other parties concerned with plant information. This goes beyond basic operations and includes scheduled maintenance, ownership supervision, market participation, and other business functions.

In other words, the normal power plant operation process cannot be considered fully detached from the rest of the world. There is information flow in the form of relevant measurements flowing to and from general purpose office networks. In turn, the classic pretext of using air gaps to consider the process networks separated from the "general purpose" office ones is not true anymore.

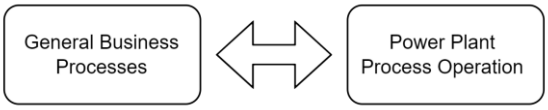

Figure 1. Information exchange in modern hydro power plant operation

This enables streamlined and optimized operation based on live data from the process and improved business performance, as the process decision can be helped from market-driven inputs. This, however, also opens an indirect path for cyber-attacks on the industrial automation control systems. The viable attack surface is more pronounced, and this calls for early detection techniques and adequate risk management.

Even in a mature technology such as hydropower, innovative digital technologies have been emerging recently. Examples include advanced monitoring technologies for rotating machines [3] and power electronics based technologies to increase the flexibility of the existing power plants, to cope with the increased flexibility demands in highly renewable power systems [4]. These new digital technologies represent additional sources of data about the power plant process.

This paper proposes to approach the issue directly and take advantage of diverse data sources to enable the extraction of the necessary knowledge and increase the cyber security and resilience. The paper is structured as follows: first, the typical context of cyber security in industrial automation and particularly in a hydro power plant is outlined in Chapter 2. The security operations center context is described in Chapter 3. Chapter 4 outlines the current state of semantic data integration in industrial automation and presents the requirements for semantically integrated data model targeting cyber security, outlining the requirements of a scalable and modular data model, capable of handling the diverse scales and contexts.

II. CYBER SECURITY IN INDUSTRIAL AUTOMATION CONTROL SYSTEMS

As already indicated, the recent IACS systems resemble general purpose IT systems [5], as a modern SCADA or a modern DCS use the same hardware and software platform as a general-purpose conventional IT system. The only practical differences are the controllers and actuators, connected via standard interfaces of a common computer. In fact, it is quite common for the IACS vendors to offload the hardware purchase to the clients, where the general IT departments are then tasked with supporting the IACS with servers, workstations, networking equipment, operating systems etc.

The "general purpose" IT and industrial automation context differ very notably in the approach to safety and security. While a failure of a general IT system or even a breach does not cause immediate effects on health and safety, in in the IACSs and in operational technology in a wider sense, an unrecoverable error in operation can directly cause serious damages in the real (physical) world. The devices connected to apparently the same computer systems can directly actuate the physical systems and cause expensive failures and even deaths. There is a particularly egregious example from recent history from the hydro power plant context in the Russian Sayano-Shushenskaya [6] hydro power station incident. The turbine at that power station had a relatively narrow band of operation where it operated with lower vibrations. After repairs and upgrades to electro-hydraulic regulators, one of the turbines previously also exhibiting signs of vibrational problems has worked as a slack output regulator, thus changing constantly its power output. On August 17th, 2009, the 920-ton rotor shaft of the turbine shot up from its seat. The turbine room has been flooded, 75 people died in the accident, aluminum smelting plants nearby had to stop operation, transformer oil spill spread 80 km downstream. It took more than 5 years to complete the required repairs.

The general IT systems tend to have a short upgrade cycle and are typically replaced or upgraded in the cycles not exceeding three to five years. Any upgrade to power plant automation systems requires a restart of the acceptance testing procedures so the automation systems are expected to remain in operation for a decade or more. For this reason, it often incorporates old and obsolete hardware and software. While Ethernet and TCP/IP is ubiquitous today, in the power plants there are numerous legacy communication protocols – often encapsulated or converted so they can work over Ethernet and TCP/IP. In most power plants there will be other types of "field bus" local networking systems as well.

In the most difficult cases, the original vendors are not even existent anymore due to mergers and acquisitions. The vendors supporting their previous lines of products might end as well. An example from Croatia is a line of protection relays from a very prominent manufacturer, purchased in the late 1990s during the war reparation years. That line of equipment is no longer supported by the original manufacturer - the only official path towards

support is the replacement of the equipment with a newer generation. Nota bene, there are solutions enabling compatibility with the newest standards and protocols, albeit offered by an alternative local manufacturer.

In the general IT context, patching and updating is a daily occurrence and is done even if no staging and testing is in place. On the contrary, power plant systems often never get patched or updated due to the principle of not touching operational systems unnecessarily. For most industrial power plants, the cost of shutting them down is the principal cost component in an upgrade procedure – thus the systems are expected to operate for decades with no shutdowns. The critical IACS systems often employ full redundancy allowing hot (live) repairs and the topology of the system is typically not configured with cost as the primary driver, but with the ability to limit the impact of a failure. Hierarchical design is also a norm, allowing local overrides.

The reality is that numerous critical industrial infrastructures in the world are controlled and protected by computer-based systems that present a dangerously problematic breeding grounds for cyber security. The breach effects range from the theft of trade secrets or business-affecting production schedules, loss of integrity and reliability to equipment damage, death, violation of environmental conditions, risking public health and even national security. Hydro power plants with large reservoirs fit all the above characteristics as uncontrolled reservoir spills can damage a large area downstream. The Croatian public is acutely aware because of the sabotage of the Peruća dam in southern Croatia during the war in 1993. After the departure of UNPROFOR protection forces from the dam, the Serbian attackers reached the dam and tried to deploy and detonate more than 20 tons of explosives. If the attack had been successful, more than 240 square kilometers downstream would be affected. The Croatian forces have in a swift action removed the Serbian forces, and the dam has luckily been quickly repaired afterwards.

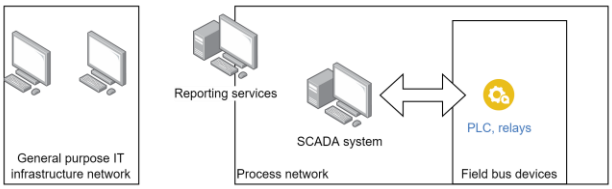

Figure 2. A typical high-level view of an architecture in IACS of a power plant

Today's plant networks in both office and process "sides" are primarily based on Ethernet and TCP/IP, with prominent other standards for the field bus local networking. As indicated before, typically there are effective bridges from the office side to process side include reporting services taking the data from automation systems and analyzing them and providing for normal business practices. The separation between the IT and process networks are often implemented via VLANs in routing equipment so the physical layer is shared, and the plant networks can be multiplexed and multipurposed, so the same physical channel carries process control and message data traffic as well as other purposes such as VoIP. Recent uses of tablet PCs and augmented reality for assisted plant maintenance and monitoring pose additional requirements to connectivity within the plant.

Finally, the IoT and miniaturization of sensors as a trend has reached all the way to the deepest nooks of the hydro power plant process: the newest generation of transformer and generator monitoring tools have their sensors embedded within the generators [3] and transformers [7] and measure vibrations and oil characteristics. Given the importance and price of the principal plant equipment their existence is more than welcome – however these systems often operate in a separate fashion from the rest of the systems and often use their own vertical network infrastructure to exfiltrate the data from the actual sensors.

There are different pivot points and points of possible attack to the hydro power plant network. The obvious ones go from wireless connectivity nowadays offered through integrated wireless and Bluetooth adapters in laptops and desktops that can be used for rogue access, to cellular modems embedded in the devices within the system so the vendor can manage the systems in a streamlined fashion. Even if air gap is not ever breached, the infection can spread as is well known from the very well documented Stuxnet [8] attack via infected media. Beyond numerous obvious and less obvious attack vectors to a typical hydro power plant, there are also various pockets and pillars of data collection and extraction caused by decades of gradual implementations and upgrades. The IoT trend mentioned above is a particular culprit here – either because of non-existence of a scaffolding that would provide infrastructural support for the IoT sensors

Supply chain threats are also not to be overlooked as hidden functions can be included in rogue firmware (as also seen from the Stuxnet example), providing backdoor access – as the Maersk cryptolocking [9] example shows quite clearly.

Obviously – there does not exist a single answer that fits all the above purposes. Not every cyber security control and mitigation strategy is applicable to industrial automation systems, and even less to hydro power plants with varying ages of the equipment spanning decades. However – not all issues are technical. In many cases, technical issues seem to be solved for decades – but the procedural implementation issues remain a challenge. For this reason, there is a whole series of cyber security in industrial automation oriented standards, closely related to IEC 27001 [10] series of standards, called IEC 62443 [11]. These standards focus on business policies and procedural issues, as opposed to technical standards similar to IEC 62351 [12] that closely define the technical details of encryption implementation. This is a particularly difficult challenge in industrial automation that requires operating procedures to be extremely efficient, especially in the emergency situations. In other words – as with everything in IACS design, no amount of cyber security can justify inability to act in an emergency. The ability to operate in a critical situation is paramount.

III. CYBER SECURITY RISKS HANDLING AND SECURITY OPERATIONS CENTRES

There are two typical levels or types of risk considered in cyber security: static and dynamic risk. Static risk pertains to risk assessment under hypothesis of slowly changing circumstances, and the dynamic risk assessment considers the underlying conditions change more frequent or that there are short notice changes in the system. Typically, it is considered that cyber security should be assessed from both points of view. In most cases the risk modeling is challenging due to difficulties in handling the data from the processes within the power plant.

A typical approach to handling dynamic risks is via a centralized Security Operations Centre (SOC) that detects real time anomalies because of risks that are identified and processed in the framework of static risk analysis. This SOC approach that analyses the data flows without further interpretation of the data flow context have become widely available in the recent years.

Sifting the legitimate situations from the illegitimate ones is challenging due to complex procedures of data acquisition so the SOCs commonly employ machine learning and data mining techniques. In fact, the bulk of security operations center operation is constantly performing a classification task. Where both the data collected from endpoints and generated (synthetic) data enter the classifier that determines whether the current situation is an incident or not.

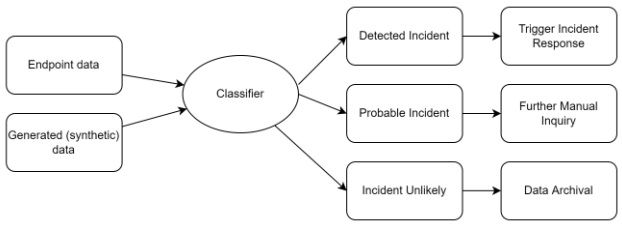

Figure 3. Typical operations of a security operations center

A typical classifier setup is to classify the situation into three typical classes: incident detected, incident probable and incident unlikely. When an incident is detected an incident response is triggered so the incident can be contained [13]. In most cases there are several additional layers of decision making: whether the collected data is sufficient and whether the mitigation succeeded. An undecided class of probable incidents triggers notifications and manual actions, while normal operation where incident is unlikely is typically followed by logged data archival. In most cases, the current SOC approaches include a limited insight into the semantic interpretation of collected logs, especially when industrial automation protocols are considered.

IV. SEMANTIC INTEGRATION IN INDUSTRIAL AUTOMATION

The IEC 61850 [14] is de facto a lingua franca in current industrial automation control systems. Originally it has been designed as an interoperable standard for electrical substations. In practice it is the first widely adopted telecontrol standard that goes beyond the notion of a signal or a command and includes the data semantics within the protocol: effectively the IEC 61850 introduces semantic interpretation of the message payload within the protocol itself. Within hydro power plants, in the last decade there have been many installations of IEC 61850 systems as well.

In a wider context, the IEC 61968/61970 [15] CIM Common Information Model is a series of standards under development that aim towards standardizing the information exchange between electrical distribution systems. This is a series of standards representing all major objects in an electric utility enterprise. It provides a standard way of representing power system resources as object classes and attributes, along with their relationships. It does facilitate the integration of applications developed independently by different vendors, however it might not be best suited approach for large-scale data gathering as it is, effectively, an ontology of everything related to electricity distribution enterprise. There have been several attempts of creating an overarching ontological model of all data. However, when creating a semantic model, its effectiveness is inversely proportional to its complexity. Numerous layers of abstraction rapidly create a barrier for implementation.

An orthogonal approach to creating a strict ontology is to utilize a set of tags as markers of collected data. In the context of building automation, the Project Haystack [16] is an open industry initiative, focused on providing a common metadata methodology for building automation so the value from the vast quantity of data being generated by the smart devices in homes, buildings, factories, and cities. Project Haystack approach is adopted in automation, control, energy, HVAC, lighting, and other environmental systems. In Haystack approach, the tags are designed as semantic carriers – akin to markup language describing the data. At a first glance this approach may appear deceptively simple, however the actual Haystack definition extends towards a standard specification on defining and describing device descriptive data, a standard taxonomy (a default set of tags), and a set of software tools and API specification to ease the model adoption. The principal promise of Project Haystack is to make the data with the accompanying tags "self-describing". The parallels with the security context and diverse types of sensors in a hydro power plant are obvious. However, the loose tagging approach is not without its drawbacks: redundant tags can appear, and while tag taxonomy is comparatively simpler than a strict ontology, its maintenance may prove to be challenging.

The SYNERGY Horizon 2020 project (GA #872734) has proposed a more elaborated SYNERGY Common Information Model, taking inspiration from both ontological models and tagging based models and paying particular attention to the data model lifecycle management. The goal of the SYNERGY CIM is to model the structure of the data exchanged between any stakeholders of the electricity data value chain, and it needs to provide a proper representation of the knowledge of the electricity data value chain. At the same time, it must allow enough flexibility to capture different combinations and relations of the data, while still keeping

the semantic information of the data. More can be found in [17].

The above favorable characteristics and good performance when coupled with large datasets indicate that the similar approach could be employed within the context of cyber security in industrial automation of hydro power plants. A derivative of the approach in the Project Haystack and SYNERGY CIM models could serve as a starting point towards deeper analysis of the data within the security-related operations of a hydro power plant.

This paper makes a proposition that gathering the data from diverse sources within a hydro power plant should be coupled with a data model that keeps the original semantic information together with the data, allowing the structured approach to the data interpretation without hindrances to system analysis performance. The original information must be kept so the actions to manage cyber security can be properly targeted.

It is not only the post festum analyses that could benefit here - deeper analysis involving information from multiple sources and going beyond the transactive logs and communication protocol logs can also be useful in dynamic real time analyses, as is common in security operations centers. The tagging semantic information can be used as effective filter for the classifier in the automated SOC, and at the same time, the same tagging-based information provides a backbone for avoiding and filtering out erroneous conclusions in the SOC classifier approach.

V. CONCLUSION

In a typical hydro power plant, the operational data sources are ever increasing putting additional burden on power plant operators. As the infrastructure for data collection becomes more affordable, all the parties concerned with plant operation request increased volumes of data to be used in optimal modeling of operations, maintenance, and other business aspects. However, these data are typically siloed by operational context, and sometimes even by a vendor providing the service in a separate data collection vertical. The approach where separate system aspects create their own data verticals to exfiltrate the required data is not sustainable both from economic standpoint and from the cyber security risk management standpoint. Not only is this capital expensive and not streamlined, it is also comparatively insecure. It is also not reasonable to expect that many different vendors with different focuses and contexts would align with a common data repository as an intermediary. Such repository, if it neglects the importance of maintaining the original data semantics, would be difficult to maintain and would pose a friction point for the operation.

To exploit the hidden semantic potential of all the data being constantly gathered in the power plant, this paper proposes that a data repository that uses semantic integration as a key building block is installed. In this repository, for all data items, the original semantics (such as measurement units) and the original data sources are always clearly known and kept. This way, the additional value from the existing data sources thus can be extracted. The principal benefits to cyber security stem from increased observability of the process – without an additional pivoting point towards the process network. The proposed concept might seem viable only in post-festum analyses – the security operations center approach can also profit from having deeper semantic insight into data being collected.

ACKNOWLEDGMENT

Research leading to these findings has been, in part, supported by the SYNERGY Horizon 2020 project. The SYNERGY project has received funding from the European Union's Horizon 2020 Research and Innovation program under No. 872734.

VI. REFERENCES

[1] Å. Killingtveit, "15 - Hydroelectric Power," in Future Energy (Third Edition), T. M. Letcher, Ed. Elsevier, 2020, pp. 315–330. doi: 10.1016/B978-0-08-102886-5.00015-3.
[2] B. R. Mehta and Y. J. Reddy, "Chapter 3 - Distributed control system," in Industrial Process Automation Systems, B. R. Mehta and Y. J. Reddy, Eds. Oxford: Butterworth-Heinemann, 2015, pp. 75–133. doi: 10.1016/B978-0-12-800939-0.00006-1.
[3] "VESKI ltd." https://veski.hr/index.php?page=condition-m (accessed Feb. 22, 2022).
[4] "XFLEX HYDRO - EU Horizon 2020 project," XFLEX HYDRO - EU Horizon 2020 project. https://xflexhydro.net (accessed Feb. 22, 2022).
[5] W. Shaw, "Cyber Security and Automation Systems (NRC presentation)." [Online]. Available: https://www.nrc.gov/docs/ML1319/ML13198A409.pdf
[6] "Sayano Shushenskaya accident – presenting a possible direct cause - International Water Power." https://www.waterpowermagazine.com/features/featuresayano-shushenskaya-accident-presenting-a-possible-direct-cause (accessed Feb. 22, 2022).
[7] "Transformer monitoring," KONČAR. https://www.koncar-institut.hr/en/?solution_group=transformer-monitoring (accessed Feb. 22, 2022).
[8] K. Zetter, Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon, Reprint edition. New York: Crown, 2015.
[9] A. Greenberg, Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers. New York: Doubleday, 2019.
[10] 14:00-17:00, "ISO/IEC 27000:2018," ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/39/73906.html (accessed Feb. 22, 2022).
[11] "IEC 62443-4-1:2018 | IEC Webstore." https://webstore.iec.ch/publication/33615 (accessed Feb. 22, 2022).
[12] "IEC 62351:2022 SER | IEC Webstore | cyber security, smart city." https://webstore.iec.ch/publication/6912 (accessed Feb. 22, 2022).
[13] "Cybersecurity Framework," NIST, Nov. 12, 2013. https://www.nist.gov/cyberframework (accessed Feb. 22, 2022).
[14] "IEC 61850:2022 SER | IEC Webstore | LVDC." https://webstore.iec.ch/publication/6028 (accessed Feb. 22, 2022).
[15] "IEC 61970:2022 SER | IEC Webstore | automation, cyber security, smart city, smart energy, smart grid, CGMES." https://webstore.iec.ch/publication/61167 (accessed Feb. 22, 2022).
[16] "Home – Project Haystack." https://project-haystack.org/ (accessed Feb. 22, 2022).
[17] The SYNERGY Consortium, "SYNERGY D3.1 Common Information Model." [Online]. Available: https://www.synergyh2020.eu/wp-content/uploads/2021/11/20210608_suite5_D3.1_SYNERGY-Common-Information-Model_pu_v1.0.pdf