|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| **Public Document** | |
| N63 | 2020/11/18 |
| Source | MPAI-2 |
| Title | MPAI Application Note #7 - MPAI-CUI – Compression and Understanding of Industrial Data |
| Target | MPAI Members |

### MPAI Application Note #7

### Compression and Understanding of Industrial Data (MPAI-CUI)

**Proponents**: Guido Perboli (POLITO), Valeria Lazzaroli (Arisk), Mariangela Rosano (POLITO)

**Description**: Most economic organizations, e.g., companies, etc., produce large quantities of data, often because these are required by regulation. Users of these data maybe the company itself or Fintech and Insurtech services who need to access the flow of company data to assess and mon­itor financial and organizational performance, as well as the impact of vertical risks (e.g., cyber, seismic, etc.). For example, nowadays companies heavily rely on the security and dependability of their Information System for all the categories of workers, including the management of Industrial Control Systems. Adding into the risk analysis process cybersecurity-related parameters will help a more precise estimation of the actual risk exposure. Cybersecurity data will help a reassessment of financial parameters based on risk analysis data.

The sheer amount of data that need to be exchanged is an issue. Analysing those data by humans is typically on­erous and may miss vitally important information. Artificial intelligence (AI) may help reduce the amount of data with a controlled loss of information and extract the most relevant information from the data. AI is considered the most promising means to achieve the goal.

Unfortunately, the syntax and semantics of the flow of data is high dependent on who has produced the data. The format of the date is typically a text file with a structure not designed for indexing, search and ex­traction. Therefore, in order to be able to apply AI technologies to meaningfully reduce the data flow, it is necessary to standardize the formats of the components of the data flow and make the data “AI friendly”.

**Comments**:

Recent regulations are imposing a constant monitoring (ideally monthly). Thus, there is the pos­sibility to have similar blocks of data in temporally consecutive sequences of data.

The company generating the data flow may need to perform compression and understanding for its own needs (e.g., to identify core and non-core data). Subsequent entities may perform further data compres­sion and transformation.

In general, compressed data should allow for easy data search and extraction.

In a first phase MPAI-CUI is addressing primarily risk identification.

**Examples**

MPAI-CUI may be used in a variety of contexts, such as:

1. To support the company’s board in deploying efficient strategies. A company can analyse its financial performance, identifying possible clues to the crisis or risk of bankruptcy years in advance. It may help the board of directors and decision-makers to make the proper decisions to avoid these situations, conduct what-if analysis, and devise efficient strategies.
2. To assess the financial health of companies that apply for funds/financial help. A financial institution that receives a request for financial help from a troubled company, can access its financial and organizational data and make an AI-based assessment of that company, as well as a prediction of future performance. This aids the financial institution to take the right decision in funding or not that company, having a broad vision of its situation.

To assess the risk in different fields considering non-core data (e.g., non-financial data). Accurate and targeted sharing of core and non-core data that ranges from the financial and organizational information to other types of risks that affect the business continuity (e.g., environmental, seismic, infrastructure, and cyber). As an example, the cybersecurity preparedness status of a company allow better estimation of the average production parameters, like the expected number of days of production lost, which are affected by cyberattacks (e.g., days the industrial plants are stopped, days personnel cannot perform their work due to unavailability of the information system). Several parameters need to be considered, which are obtained by direct acquisition of data from the target companies that perform a cybersecurity risk analysis.

1. To analyse the effects of disruptions on the national economy, e.g., performance evaluation by pre/post- pandemic analysis [1].

**Requirements:**

1. The standards of the data in the data flow should be AI friendly. In other words, the different data required to predict a crisis/bankruptcy of a company should be gathered, carefully selected and eventually completed to be suitable for automatic analysis and then treated by an AI-based algorithm.
2. The standard shall ensure efficiency of data structure, indexing and search, according to specific syntax and semantics.
3. The standard shall all the extraction of main parameters with indication of their semantics.
4. The standard shall support context-based compression (i.e., depending on the sequence of data).
5. The standard shall support lossless compression.
6. The standard shall support context-based filtering with different levels of details.

**Object of standard**:

Two main areas of standardization are identified:

1. **Inputs objects**: a first set of data input related to:
   1. ***Financial data input***
      1. Financial statements and fiscal yearly reports data (usually expressed in xls or xbrl formats). Their contents follow the accounting standards defined by the Organismo Italiano Contabilità (OIC) at the Italian level and the International Accounting Standards Board (IAS/IFRS) at the international level.
      2. Invoices. In Italy, the FatturaPA format is expressed in xml, but more in general invoices have to be compliant with the European standard EN 16931-1:2017.
      3. Semantics of governance elements.
      4. Other economic data as company size uniformly recognized according to the size of employees or to the economic activities (e.g., classification elaborated by Eurostat and OECD data); imports and exports, etc…
   2. ***Vertical risks*** data input: in a preliminary phase, we will consider seismic and cyber, as vertical risks of primary interest. In the future, the object of the standard could be extended to cover other risks (e.g., related to infrastructures, sustainability). Generally, at the international level, the ISO/IEC 31000 standard defines the principles and guidelines related to the input data to consider for the risk assessment and management.
      1. *Seismic risk.* AI algorithm may help to define a socio-economic and technological model that will support companies and institution in defining reconstruction plan properly. In this direction, data input to assess seismic risk according to ASTM standards, will integrate:
      2. Technical data related to the existing/needed infrastructures (i.e., geolocation coordinates), architectural and urban planning data, as well as output data from the Building Information Modelling;
      3. Socio-demographic data, i.e., statistical data collected by certified sources (e.g., ISTAT in Italy, World Bank, International Financial Statistics, Worl Economic Outlook Databases, International Monetary Fund Statistics Data) about population figures, and their characteristics and distribution.
      4. *Cyber risks*. Considering cybersecurity-related parameters in the risk assessment will help to understand and estimate the impact of the actual risk exposure on the company performance, its financial health and business continuity.

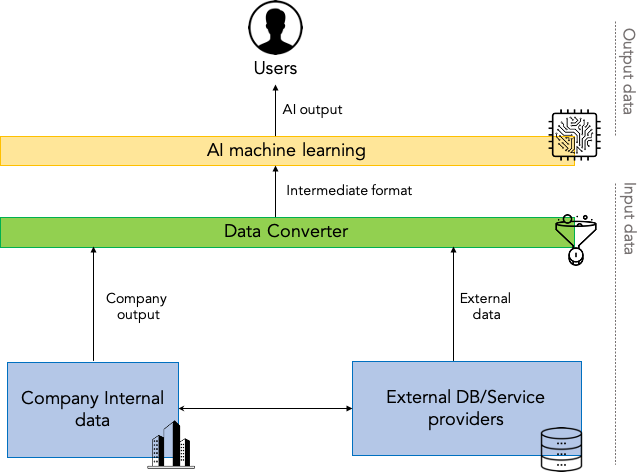
As an example, an effective system to back up the sensitive data and testing periodically its effectiveness can be of help. Moreover, having well defined incident responses and a team prepared to deal with them, can help minimizing the effects of attacks and the time to recover. Therefore, an initial set of internationally recognised (ISO/IEC 27000 Information security management) inputs to consider are:

1. Data related to assessment of organizational cyber management:
   1. *organizational-level incident management* (enumerate): no, simple plans available, detailed plan + IR Team, full integrated management (e.g., with a security operations center)
   2. *backup* management: no, user requested, automatic, automatic and tested
   3. *vulnerability management* (enumerate): no, assessement, management plans, with automatic tools
   4. *enterprise patch management* (enumerate): no, manual, automatic, testing
   5. *specific cybersecurity and testing personnel* (enumerate):no cybersecurity tasks, IT personnel with cybersecurity tasks, cybersecurity-trained roles
   6. *cybersecurity procedure and mitigation testing* (enumerate): no, occasional, planned, planned&frequent
   7. *risk analysis*: no, threats identified, assessment available, some mitigations implemented, full (mitigations implemented or justified, risks quantified)
2. Data related to prevention of cyber-attacks. Being able to detect anomalies into an information system can allow preventing some attacks or discovering them before attackers, which can be estimated considering:
3. *monitoring*: no, basic detection, integrated detection, organization-level, Security Information and Event Management (SIEM) software
4. *reaction*: no, planned pre-configured responses, organized (some autom­atic), integrated tool-based & human supervised
5. Data related to training. Studies report that the personnel that has followed spec­ific training on cybersecurity aspects that are relevant for their roles are more likely to avoid errors that may compromise the information systems of their companies (e.g., use better password, less likely to click on phishing emails). Training is even more important for personnel with cybersecurity-related responsibilities, hence:
6. *awareness/training cybersecurity personnel*: no, occasional, frequent training
7. *awareness/training other categories*: no, occasional, frequent, frequent and tailored per tasks.
8. Data related to legal issues. An additional field where measuring the preparedness is the risk of losses due to legal issues (e.g., GDPR fines for
9. *availability of cybesecurity certification* (enumerate): no, certifications relevant for the company business obtained
10. *compliance to regulations* (enumerate): no, minimum, adequate
11. Data for quantifying exposure. Whenever available, the following aggregated values can help to quantify the overall exposure to cybersecurity risks:
12. *overall risk exposure*: monetization value from the risk assessment phase (number, directly obtained by the target company)
13. *mitigated risk exposure*: monetization value after the risk mitigation phase (number, directly obtained by the target company)
14. *value of assets by security requirement*: (Confidentiality-asset value/ Integrity-asset value / Availability-asset value)
15. *percentage of the value of assets-by-security-requirement on the overall value of the company assets*: (Confidentiality-asset value/ Integrity-asset value / Availability-asset value).
16. **Output objects**: represented by the outcome of the AI-based assessment in a format known by the user (format: Json)? This format is expressed in terms of a set of indexes that reflect the health of a company, the appropriateness of the governance and the impact of risks on economic-financial parameters. Some of these indexes are in response to legal requirements on business bankruptcy and crisis (e.g., DSCR), others are computed by means of a proprietary machine learning algorithm.

More in detail, the parameters in output are:

* Risk index of the likelihood of company default in a time horizon of 36 months. It reflects the company performances based on the financial data.
* Index of business continuity reflects the impacts of other risks on the previous measure.
* Index of adequacy of the organizational model considers the impact of the governance on the performance of the company, highlighting possible issues in terms of conflict of interest or familiarity.
* Debt service coverage ratio is a measurement of a firm's available cash flow to pay current debt obligations

This is depicted in *Figure 1* where the object of the standard is identified to be the intermediate format and the AI machine output.

**

*Figure 1 – MPAI-CUI model*

In some cases, internationally agreed input data formats exist. In several other cases a variety of formats exist. In these cases, meta formats to which existing formats can be converted should be defined.

The current framework has been made as general as possible taking into consideration the wide range of issues related to risk management. We are expecting the architecture to be enriched and extended according to inclusion of other risks and eventually the synergies with other MPAI applications.

Data confidentiality, privacy issues, etc. are for further consideration.

**Benefits:** MPAI-CUI will bring benefits positively affecting:

1. Technology providers need not develop full applications to put to good use their technol­ogies. They can concentrate on gradually introducing AI-based technologies that will allow a transition from traditionally approaches based on statistical methods, overcoming their limitations. This will enhance the accuracy of prediction and improve user experience.
2. Service providers (e.g., Fintech and Insurtech companies, advisors, banks) can deliver accurate products and services supporting the decision-making process, with minimising time to market as the MPAI-CUI framework enables easy combination of internal and external-party compon­ents.
3. End users as companies and local government can obtain an AI-decision support system to assess the financial health and deploy efficient strategies and action plans.
4. Processing modules can be reused for different risk management applications.

**Bottlenecks:** The full potential of AI in MPAI-CUI would be limited by a market of AI-friendly data providers and by the adoption of a vast amount of information and data strictly dependent from the company and its context.

**Social aspects**:

A simplified access to the technologies under the MPAI-CUI standard will offer end users AI-based products promising for predictions and supporting the decision making in different contexts, reducing the effort of user in analyzing data and improving its experience which becomes more personal, but including a wide vision (e.g., thought benchmarking).

Moreover, the MPAI-CUI standard and the introduction of AI-based technologies will allow a transition from the present systems which are human readable, to machine readable technologies and services.

At the national level, governments can simulate the effects of public interventions and deploy proper strategies and plans in supporting the companies and economy.

**Success criteria**:

MPAI CUI becomes the bridge between traditional approaches in compliance with the actual regulation on prediction of business crisis and fully AI-based systems.

**References**

[1] Perboli G., Arabnezhad E., A Machine Learning-based DSS for Mid and Long-Term Company Crisis Prediction. CIRRELT-2020-29. July 2020.