HLIF2024: a Competition for High-Level Information Fusion

Claire Laudy
Thales
Palaiseau, France
claire.laudy@thalesgroup.com

Nicolas Museux
Thales
Palaiseau, France
nicolas.museux@thalesgroup.com

Simon Fossier
Thales
Palaiseau, France
simon.fossier@thalesgroup.com

Céline Reverdy
Thales
Palaiseau, France
celine.reverdy@thalesgroup.com

Tom Fougère
Thales
Toulouse, France
tom.fougere@thalesgroup.com

Amandine Audouy
Thales
Toulouse, France
amandine.audouy@fr.thalesgroup.com

Clara Lopez
Thales
Toulouse, France
clara.lopez@thalesgroup.com

Florent Chenevier
Thales
Toulouse, France
florent.chenevier@fr.thalesgroup.com

***Abstract*—This paper presents the HLIF2024 competition. HLIF2024 is the first competition for High-Level Information Fusion solutions, co-organized with the FUSION2024 conference. The paper presents the challenge use case associated with the competition: information extraction and fusion from Notices to Air Missions data. The challenge is to automate the processing of this data, in order to support pilots with their decision making tasks during flights. To simulate this, the challenge involves a fictitious flight during which information is requested by the pilot. To provide pertinent information to the pilots, the participants are asked to populate a given domain ontology that can then be queried to answer the challenge questions. The participants solutions are evaluated with regard to the correctness, completeness and preciseness of their answers to the questions. The motivation behind the use case is presented, together with the practice dataset that was built and provided to the competition participants. The evaluation methodology and metrics used to rank the participants solutions are detailed. The results of the competition will be presented at the FUSION2024 conference.**

***Index Terms*—High-level information Fusion, Evaluation, Competition, Challenge, Benchmark**

I. Introduction

Algorithm evaluation is essential, yet it is a difficult task. In most cases and for domains such as signal and data processing, there exist reference evaluation databases, solution baselines and competitions that enable evaluating one's algorithm or system with regard to the state-of-the-art solutions. Evaluation benchmarks and competitions exist for images [1], text [2] [3], video, audio and numerical data processing systems [4]. Regarding the classes of algorithms, there exist benchmarks and competitions for the evaluation of machine learning algorithms, automatic clustering, automatic classification or natural language processing systems. For instance, [5] and [6] provides numerous datasets for testing and evaluating machine learning algorithms. The Wikipedia encyclopedia also provides a list of datasets available for machine learning research [7].

Regarding problem resolution algorithms, such as optimization or satisfaction methods, benchmarks exist and are used for the organization of competitions or challenges associated with the scientific conferences of these domains (e.g. The Genetic and Evolutionary Computation Conference (GECCO) [8] or the biennial French Operational Research and Decision Support Society (ROADEF) challenge [9]). In these benchmarks, the optimal solution is known, or at least a set of best-known solutions. This makes it possible to evaluate any algorithm and compare it to state-of-the-art or well-known algorithms.

Within the visual analytics community, the Visual Analytics Science and Technology (VAST) Challenge [10] proposes to solve a new problem every year. This challenge is of special interest to us, as the question raised aims at getting awareness of an (ongoing) situation and answer the Who-What-Where-When-How question. High level information fusion shares this objective of understanding a situation. However, the role played by the human analyst within a visual analytics system is important. The system does not provide the answer to the challenge by itself but rather supports the human analyst.

A first synthetic dataset for evaluation of hard and soft data fusion system was proposed in [11], inspired by a Counter Insurgency scenario. It integrates data coming from both physical and human sensors. In [12], a benchmark was specifically developed for high level information fusion purposes. It contains information already extracted from different information media and thus focuses on the fusion task, rather than the information extraction from soft data task.

However, no existing and regularly organised challenge or

competition exist for the evaluation and comparison of these classes of algorithms. Through the organisation of HLIF2024, we propose to pave the way to such regular events. For HLIF2024, we propose to the participants to challenge their high level information extraction and fusion solutions on a real problem, proposed by one of the Thales Global Business Units. The objective is to provide solutions in order to support airline pilots in the analysis of the data, in particular the so-called NOTAMs (Notices to Air Missions) [13], that they receive during the flight briefing phase.

This paper presents the HLIF2024 competition that we organised, in collaboration with FUSION2024 conference. We describe the practice benchmark, the evaluation metrics and the evaluation process we developed for HLIF2024. Section II presents the context and challenge associated to the chosen use case. We detail the tasks and questions to which participants have to answer as well as the practice dataset, made of the Notice to Air Mission dataset and the domain ontology, that was developed and provided to the participants in the early phase of the competition. Section III describes the methodology and metrics that we use to rank the solutions. We describe the answer validation process, the metrics used for evaluation of the participants' solutions and the evaluation pipeline that we developed. We conclude in section IV, presenting our vision for future HLIF competitions.

II. Benchmark of NOTAM data

In the section, we present the dataset that was built and made available to participants for practice purposes. The aim was to give the participants the opportunity to test their solutions on a dataset similar to the one that was to be used for the evaluation, without revealing this last dataset before the evaluation phase of the competition. The evaluation dataset will be made available after the competition, through a Thales GitHub project (https://github.com/ThalesGroup).

A. The difficulty of interpreting NOTAMs

Notice to Air Missions[1], abbreviated as NOTAM [13], are short messages published by governmental agencies in order to inform about evolution of the flight-related infrastructures. NOTAMs may be issued by national or international civil aviation agencies, air traffic controllers, the military, weather agencies, airports for instance. These messages are generally available to pilots, air traffic controllers, flight planners and other aviation stakeholders. In particular, pilots use them in order be informed of evolution of the flight-related infrastructures, so that they can take appropriate action.

A NOTAM can contain information on various subjects such as temporary airport closures, airspace restrictions, recently erected obstacles, faulty navigation equipment, changes in flight procedures, weather warnings etc. Figure 1 depicts an example of a NOTAM message

The NOTAM information system is a worldwide system created in 1947, and is used by all stakeholders of the aviation domain. NOTAMs are part of an old, worldwide system that has little evolved, causing numerous problems of interpretation and understanding by users, and leading to accidents and delays.

The system is assessed to be highly unsatisfactory. For instance, among 2100 pilots surveyed in 2019 by fixingnotam.org, *72% often have difficulty understanding a NOTAM* and *74% regularly miss critical information* [14].

Pilots are required to read and understand the huge amount of NOTAM related to their flight. This raises one difficulty: the cognitive overload linked to this task can be dangerous and cause accidents or near-accidents, such as the one involving the flight MH17 over Ukraine. In that accident, the restrictions issued by the Federal Aviation Administration were not noted by the automated systems of Malaysian Airlines. The initial route of the flight was not changed, crossing a restricted area due to the armed conflict in Ukraine [15].

B. The challenge story and tasks

In that context, we propose a challenge for high level information extraction and fusion systems, which aims at processing NOTAM data and making it available through an ontology. Our aim is to compare and evaluate the different existing approaches for high-level information extraction and fusion and tp assess their suitability to the problem.

Ontologies are a commonly used tool for representing high level information. Many researchers in high level information fusion have been integrating ontologies as part of their work, in order to represent and store high level information extracted from both hard and soft data. Furthermore, using a unique and federated representation of a business domain (such as an ontology) enables to easily compare the contents, and thus the information gathered on a situation. For these reasons, our approach to evaluate and compare high level information fusion solutions is based on comparing the contents of a reference ontology (contained in the part called A-box in the ontology domain) with the contents of an ontology populated using a high level information fusion solution.

The story of the challenge is the following. *The company AirNotes, specialised in developing state-of-the-art Artificial Intelligence (AI)-based solutions for aviation, is designing NoteOnTo, a new smart assistant to pilots. NoteOnTo is in charge of collecting and analysing NOTAM messages and providing a comprehensive overview of their contents to pilots during their flights.*

With NoteOnTo, AirNotes provides a smart assistant to pilots. NoteOnTo reads and analyses the NOTAMs and stores all current, non-redundant and up-to-date information in a knowledge base. This knowledge base can then be queried easily, either by the pilots themselves through pre-defined queries, or using other tools of the AirNotes smart software suite Air3S.

The task proposed within the HLIF2024 competition is to develop the algorithm that acquires textual raw NOTAM data and analyses it with regard to a domain ontology. The

[1]alternatively: Notice to Air Man, Notice to Air Men, or more recently Notice to Air Operations

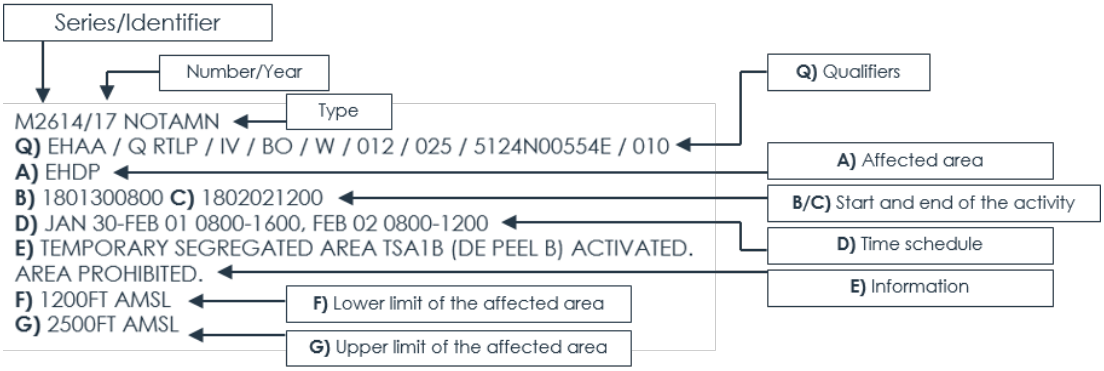

Figure 1. An example of NOTAM message

Table I
FLIGHT PLAN INFORMATION FOR THE PRACTICE DATASET

Airport name	OACI code	Function
Amsterdam Schiphol AMS	EHAM	Departure
Toulouse-Blagnac TLS	LFBO	Arrival
Brussels BRU	EBBR	Alternate airport 1
Paris Charles de Gaulle CDG	LFPG	Alternate airport 2
Limoges-Bellegarde LIG	LFBL	Alternate airport 3
Bordeaux-Mérignac BOD	LFBD	Alternate arrival airport 1
Carcassonne CCF	LFMK	Alternate arrival airport 2

information embedded in the NOTAM is stored in a populated instance of the ontology.

To populate the ontology, first, data items extracted from different NOTAMs must be associated. Indeed, although all the aviation related infrastructures (airports, runways, taxiways etc.) have unique identifiers in the aviation databases, the way they are referenced to in the NOTAMs may differ. These discrepancies between names may either come from the habits the humans that wrote the NOTAMs have, or even due to unintentional misspelling.

Secondly, the information about one specific infrastructure or event may be spread across several NOTAMs. This forces pilots and automated analysis systems to have a fusion process that detects that several NOTAMs concern the same entity and fuse the information spread across these NOTAMs.

C. Questions to be answered by competitors

As a provider of information to pilots about their flight, the contestants had to provide information regarding the different airports and highlight potential impacts of some of the NOTAMs on the flight plan.

The considered flight plans were provided to the contestants, with information on the departure and arrival airports, the different alternate airports and the estimated dates and times of departure and arrival.

As an example, the information contained in Table I was given with regard to the practice dataset, together with the following flight dates and times:

Estimated time of departure: Sunday 31/12/2023 19:40
Estimated time of arrival: Sunday 31/12/2023 21:30

The competitors then had to provide information about the following three instructions.

Q1: What is all the information contained in the NOTAMs that is applicable to the departure and arrival airports for the period of time going from 40 minutes before the estimated time of departure to an hour and a half after the estimated time of arrival?

Q2: During the flight, the pilot gets aware of a modification on the the arrival airport infrastructure that makes it impossible for him/her to land there. What are the IRI[2] of the individual NOTAM conveying this information in the ontology?

Q3: The pilot decides to land on one of the alternate arrival airports and needs support in choosing the most suitable one. What is the list of IRIs of the alternative airports in the ontology, populated with the information linked to landing and arrival procedures in these airports for the period of time going from 40 minutes before the estimated time of departure to an hour and a half after the estimated time of arrival?

The information fusion problem consists in feeding properly the ontology with the raw NOTAMs to answer completely and correctly to these questions.

D. The NOTAM ontology

As part of the benchmark, we provide an ontology that must be populated with information extracted from the NOTAM dataset. This NOTAM ontology was built upon concepts from the BEST [16] and ATMONTO [17] ontologies that describe the Air Traffic Management domain.

ATMONTO, the NASA ATM (Air Traffic Management) Ontology describes classes, properties, and relationships relevant to the domain of Air Traffic Management, and represents information pertinent to a broad and diverse set of interacting components in the US and the global airspace, including flights, aircraft, manufacturers, airports, airlines, air routes, facilities, air traffic advisories, weather phenomena, and many others. Three different variants of the ATM Ontology are

[2]Internationalized Resource Identifier, extension of URI (Universal Resource Identifier) with Universal Character Set.

provided: atmontoCore, atmonto, and atmontoPlus. We used concepts from atmontoCore.

The BEST project investigated on how semantic technologies can be used for ATM purposes. The project developed and provided an ontology for the ATM domain ontology that takes into account the relevant ATM information models and concepts.

From the core concepts imported from BEST and ATMONTO and relying on our domain knowledge, we built a T-Box (Terminological part of the ontology. i.e. the part containing the domain model) containing the description of the information that can handled through NOTAMs to which we also added the concept of NOTAM. These concepts are organised in a typology of classes that represent the different types of NOTAMs, according to their categories. We also added concepts linked to the impacts that information potentially conveyed by NOTAMs may have on a flight.

E. The practice dataset

A practice dataset was made available to the registered participants of the challenge. It is composed of 206 fictional NOTAM messages. Our scenario flight departs from Amsterdam-Schiphol and arrives at Toulouse-Blagnac. Among the 206 NOTAM messages for our scenario, 3 were entirely designed and written by ourselves. They are the focal point of our scenario, and will be central to the results provided by the evaluated algorithms. These NOTAMs are focused on the following infrastructures and events:

- *Fireworks at Disneyland Paris* that are planned to take place for 60 minutes in a circular area, centered on a given coordinates, on the 17^{th}, 24^{th}, 25^{th} and 31^{st} of December and 1^{st} and 5^{th} of January;
- *Fireworks at Atomium Brussels* that are planned to take place for 60 minutes in a circular area, centered on a given coordinates, on the 24^{th}, 25^{th} and 31^{st} of December and 1^{st} of January;
- *Works on a runway in Toulouse-Blagnac airport* that results in a shorter runway.

Other NOTAMs were added to the practice dataset, in order to "hide" these three NOTAMs. To avoid having systems that learned to recognize the real NOTAMs from the fake ones using open and available NOTAM databases, these NOTAMs also had to be fictitious. In order to build a coherent and plausible scenario, we used real NOTAMs as examples, that we altered. To get data for the scenario, we thus gathered real NOTAMs from the FAA Defense Internet NOTAM Service [18], concerning the following airports :

- Amsterdam-Schiphol
- Bruxelles-National
- Paris-Charles de Gaulle
- Bordeaux
- Toulouse-Blagnac
- Carcassonne

The NOTAM messages were imported on the 24^{th} of November 2023. We then had to alter the data, so that it reflects the date of our scenario. To do so, we used the *Faker* python library [19]. *Faker* is a Python package that enables generating fake data.

Each fake NOTAM is built following the subsequent process:

1) Get an example NOTAM from the real database;
2) Generate a new date by adding a fixed number of days to B and C fields of the NOTAM;
3) Replace any date in the E field (free text) by a new date by adding the same number of days.

III. Evaluation of the solutions

A. The validation process

Answering the challenge consists in providing two files:

1) the given ontology enriched with the information contained in the NOTAMs,
2) a JSON file pointing the individuals in the ontology that answer the asked questions.

In other words, the ontology is the knowledge base and the JSON file indicates the entry-points of the knowledge base that constitute the answers to the challenge.

The ontology follows the OWL2 syntax. It is not expected from the participants to add new concepts or relations in such ontology. Even if it is not forbidden, it is unproductive to do so as these added concepts or relations cannot be part of the expected solutions. Therefore using them will degrade the satisfactory degree of the metrics and so, the final score. However, the validation process does not check the consistency of this populated ontology regarding its expected content. It validates only its overall syntax and checks that the individuals indicated as an answer to a question exists in this ontology and is an instance of a concept.

The JSON object needs to follow a defined format, and we will check its compliance in terms of format and content (i.e. the pointed individuals are indeed in the ontology).

Here is an example of such JSON object following the expected syntax :

```
{ "Q1": {"hasValue": ["
    ↪ https://myontology.owl#$p1", "
    ↪ https://myontology.owl#$p2"]},
  "Q2": {"hasValue": ["
      ↪ https://myontology.owl#$p3"]},
  "Q3": {"hasValue": ["
      ↪ https://myontology.owl#$p4"]}  }
```

In order to limit issues due to wrong syntax or bad contents, we have defined some constraints to frame the submissions:

- The ontology exists: the ontology file is not empty, well written and the given namespace in the URIs is the same as the one of the initially given ontology.
- The JSON is properly formatted and respects the defined format.
- The `Q1` ... `Qn` keys correspond to the question numbers, and `hasValue` is the associated answer.
- There is at least one question being answered.
- An answer cannot be empty.

- Every question answered must have at least one answer, and nothing else than answers.
- The value of an answer is an array of URIs.
- Duplicates within an answer to a question are not allowed.
- Each question answered has its own answer (but individuals can belong to several answers).
- The individuals URIs are present in the populated ontology provided along with the JSON.

Most of these constraints are properties described in the following JSON schema provided to the participants. The schema is compliant with the specification draft-2019-09 [20].

```
{   "title": "Answer",
    "description": "Answers-schema to the questions
        ↪ of the 2024 HLIF challenge",
    "type": "object",
    "additionalProperties": false,
    "minProperties": 1,
    "maxProperties": 3,
    "patternProperties": {
        "^(Q|q)([0-9]+)": {
            "type": "object",
            "additionalProperties": false,
            "required": ["hasValue"],
            "properties": {
                "hasValue": {
                    "type": "array",
                    "items": {
                        "type": "string",
                        "format": "uri"},
                    "minItems": 1,
                    "uniqueItems": true}}}} }
```

It is then easy for the participants to check locally if their answers are well formatted.

In addition, a RESTful GET service, available during the practice phase only, gives the opportunity to the participants to test the compliance of their solutions to the constraints listed above. At the time this paper is written, the participants can get details about the REST service at `https://hlif-challenge.s3g-labs.fr/rest`.

The result is a Boolean value indicating if the provided solution is valid or not, and a descriptive message in case it is not valid. If there is more than one issue, the service indicates only the first it has met. So the process must be iterated until there are no more errors.

B. The evaluation metrics

The evaluation of a high level fusion algorithm must be performed over two points of view. The first one deals with functional capability, and the second one deals with operational capabilities.

The functional capabilities cover the quality of the provided result. It is evaluated, dependently to an expected result, regarding **its correctness** (to which degree it contains only the correct answer and no incorrect answer), **its completeness** (to which degree it contains the full expected result), **its preciseness** (how much the correct answer is at the expected level of accuracy).

These criteria are relevant on one execution of the algorithm, nevertheless it is not enough to perform a rigorous evaluation. The stability and the robustness are two other functional capabilities that must be taken into account, statistically over a series of executions. The stability indicates how the algorithm behaves if it is executed more than once on the same set of pieces of information, whereas the robustness indicates how the algorithm behaves accordingly to changes in its inputs.

The metrics used in order to compare the results of the solutions of the different teams are the ones defined and mathematically described in [12]. Here we quickly remind their respective definitions, where r_q is the solution to the question q provided by an algorithmic solution designed by a participant, and r_q^* is the corresponding expected solution.

Definition 1 : Correctness metric
The correctness metric, $\mu_{\text{correct}}(q) \in [0,1]$, measures the amount of information of the result r_q that is also contained in the expected result r_q^*, compared to the overall amount of information contained in r. In essence, the percentage of r_q that is in r_q^*.

Definition 2 : Completeness metric
The completeness metric, $\mu_{\text{complete}}(q) \in [0,1]$, measures the amount of information of the result r_q that is also contained in the expected result r_q^*, compared to the overall amount of information contained in r_q^*. In essence, the percentage of r_q^* that is in r_q.

Then, providing a solution that is correct and complete means that it contains the full expected result, and only the expected result.

Definition 3 : Preciseness metric
The preciseness metric, $\mu_{\text{precise}}(q) \in [0,1]$, measures the mean square discrepancy between the level of precision of expected result r_q^* with the one provided by the result r_q. It says how much the result r_q is over- or under-informative compared to the expected result r_q^*.

The evaluation of the answer to a question q is then the vector $\mu(q)$, of which the coordinates are the evaluations of the result r_q on each of the above criteria:

$$\mu(q) = (\mu_{\text{complete}}(q), \mu_{\text{correct}}(q), \mu_{\text{precise}}(q))$$

These values are aggregated to obtain the score of the given solution. Our aggregator takes into account the relative importance and interaction of the evaluation-criteria (modeling the complementarity or the redundancy of such criteria). This involves the 2-additive Choquet sum commonly used in multi-criteria decision aid [21]–[24].

$$\begin{aligned}\text{score}(\mu(q)) \quad = \quad & \sum_i \text{Importance}(\mu_i(q)) \\ & -\frac{1}{2}\sum_{\{i,j\}} \text{Interaction}(\mu_i(q), \mu_j(q))\end{aligned}$$

The global evaluation score of the answer to the set of N questions is the average of the aggregated scores of each question q:

$$\text{eval} = \frac{1}{N}\sum_q \text{score}(\mu(q))$$

This is justified by the fact that all answers have the same degree of relative importance, i.e. the same contribution to the global score, and are mutually independent. It is not necessary to answer well the first question to answer well the second or the third. In a context where it would not be the case, because the questions would have different difficulty levels for instance, a weighted average would be enough while the independence is maintained. Otherwise, an aggregator as the one we have used in the evaluation of one answer is more appropriate.

To support the participants in the design of their information fusion solution, we provide also a RESTful GET service allowing them to have feedback about the quality of their solutions. It returns the scores for each answer, one per metric, the aggregated result and the global score. Again, these webservices are available only during the practice phase.

C. The HLIF2024 evaluation pipeline

For operational use, an algorithm dedicated to such a task needs to perform in constrained conditions. In particular, it has at its disposal a certain amount of computing power (in terms of memory, processor load, distributed or centralized processing), it can have dependencies to external resources (prior models, external datasets for instance). Machine learning techniques or advanced natural language processing (NLP) treatments are typical models for processing textual data. Learning models such as SVM [25] or XGBoost [26], fine-tuning pre-trained transformers such as popular models BERT [27] or T5 [28], or prompt-based learning of generative AI based on the GPT model [29] are very relevant options in this context. In the case where these models are necessary for the processing of the dataset, they have to be embedded in the solution provided by the participants.

As the proposed problem is not too hard computationally, we assume that the computation should not take too much time. Therefore, we assume that taking too much time to provide a result is significant of a computational issue. Beyond an arbitrarily chosen amount of time (the same for everyone), processing will be interrupted and results available at that moment will be evaluated in the same way as the others.

In the context of HLIF2024, rather than assessing the operational capabilities of the solutions as criteria for evaluation, we specified requirements that solutions should meet for the competition. As the challenge aims at evaluating solutions that could typically be embedded in aircraft equipment, we allow limited resources. For instance, there is no Internet access at run-time during evaluation of the solutions. Nevertheless, the solutions may be prepared and built using any available resources.

The hardware settings used to run the participants solutions are the following ones:

- 16 Intel® Xeon® Silver 4114 CPU cores
- 40 GiB of RAM
- only CPU use (no GPU)
- no Internet access

To maximize the portability of solutions, contestants are required to encapsulate their algorithms inside a Docker [30] image. This ensures the complete consistency of the solution's behaviour at run-time, by isolating the execution environment of the algorithm from the host machine's environment configuration. In order to standardize the way that contestants' solutions behave in terms of input and output data exchange, as well as ensuring proper access rights management to run on the execution environment provided by the organizers, the contestants are required to follow some guidelines to build their image in the expected format. These requirements are detailed at `https://hlif-challenge.s3g-labs.fr/docker`. A Dockerfile template is also provided to ensure that these requirements are respected during the Docker image building process.

The challenge has been designed in two main phases:

1) **An initial practice phase**, where the contestants get acquainted with the problem, the data format and the expected results. They can configure and improve their algorithms with a practice dataset, and access a REST service to get feedback regarding the correctness of the structure/format of their outputs and their evaluation on each criterion. The contestants can also validate the docker building by submitting it on the competition's web site. If everything is good, the evaluation scores are then visible.
2) **A final evaluation phase**, where a similar problem is presented. This new problem involves a different location, retains the same data format and the same type of expected results, however it involves processing a new and a larger dataset. Each contestant provides an encapsulated algorithm that has been tuned for the challenge; this algorithm will be run by the organizers directly, from which the value of each criterion will be evaluated.

The processing pipeline for solution evaluation can be described as follows:

1) The contestants provide Docker [30] images containing their algorithms, ready to run, to a webservice hosted by the organizers.
2) These images are loaded in an isolated execution environment managed by the organizers. This environment verifies that the Docker images have been well-configured by the participants, without installation problems, misconfiguration, or any other issues.
3) The evaluation dataset, unknown by the contestants, is provided as input to the contestants algorithms, which are then run.
4) The output of the algorithms is analyzed in terms of format: do results match the expected structure? The responsibility of result format compatibility is left to the contestants, and no ad-hoc modification will be done by the organizers.
5) Finally, the evaluation of each criterion is performed for this solution, and these criteria are aggregated to obtain the final score.

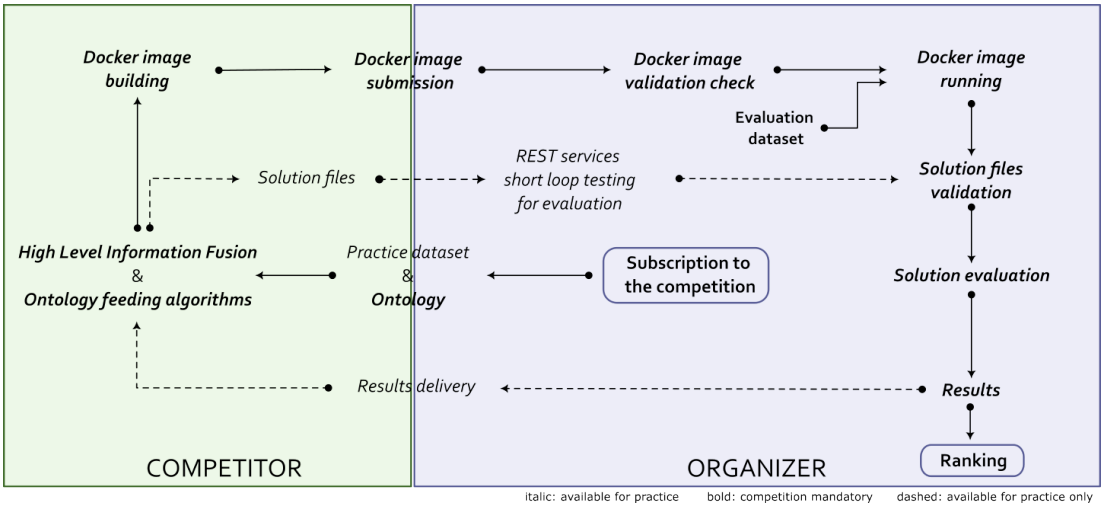

Figure 2. The complete processing pipeline from subscription to solutions ranking

The complete processing pipeline from the subscription to the competition to the ranking of all provided solutions run on the evaluation dataset is shown in Figure 2.

IV. Conclusion

In this paper, we present the first high-level information fusion competition – HLIF2024 – that we organised with the support of the FUSION2024 conference. The paper describes the practice dataset, the competition question and the evaluation process that we deployed for the competition.

HLIF2024 is a first attempt to organize such an event in the high level information fusion scientific community. It proposed not only evaluation and reference datasets, specifically built for High level information fusion purpose, such as the SYNCOIN [11] and Ravixe [31] datasets were, but also a fully automated and reproducible evaluation service aiming at comparing different solutions, on a single problem, with a single evaluation process. We hope that this event will pave the way to regular competitions that would benefit the whole community.

In future occurrences of the competition, we believe that more complex scenarios, such as the ones proposed in the VAST challenge, but adapted to the specific task of information fusion, should be developed. For instance, these scenarios should include:

- several events involving several actors and/or objects, locations, etc.
- more classes of objects interacting with each other,
- several information sources.

Furthermore, we would like to assess the management of uncertainty achieved by the solutions. To do so, we will introduce the uncertainty evaluation metrics presented in [12], and add other evaluation criteria.

References

[1] "Computer vision test images," 2005, [Online; accessed 2024-03-12]. [Online]. Available: https://www.cs.cmu.edu/~cil/v-images.html

[2] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," 2020.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," 2016.

[4] Kaggle, "Kaggle competitions," [Online; accessed 2024-03-12]. [Online]. Available: https://www.kaggle.com/competitions

[5] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[6] R. S. Olson, W. L. Cava, P. Orzechowski, R. J. Urbanowicz, and J. H. Moore, "Pmlb: A large benchmark suite for machine learning evaluation and comparison," 2017.

[7] Wikipedia contributors, "List of datasets for machine-learning research — Wikipedia, the free encyclopedia," 2004, [Online; accessed 2024-03-12]. [Online]. Available: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

[8] GECCO2024 organizing comitee, "Gecco2024 competitions," 2024, [Online; accessed 2024-03-12]. [Online]. Available: https://gecco-2024.sigevo.org/Competitions

[9] ROADEF organizing comitee, "Roadef main page," 2024, [Online; accessed 2024-03-12]. [Online]. Available: https://www.roadef.org/societe-francaise-recherche-operationnelle-aide-decision

[10] K. Cook, G. Grinstein, and M. Whiting, "The vast challenge: history, scope, and outcomes: An introduction to the special issue," *Information Visualization*, vol. 13, no. 4, pp. 301–312, 2014. [Online]. Available: https://doi.org/10.1177/1473871613490678

[11] J. L. Graham, J. Rimland, and D. L. Hall, "A coin-inspired synthetic dataset for qualitative evaluation of hard and soft fusion systems," in *Fusion 2011 - 14th International Conference on Information Fusion*, ser. Fusion 2011 - 14th International Conference on Information Fusion, 2011, 14th International Conference on Information Fusion, Fusion 2011 ; Conference date: 05-07-2011 Through 08-07-2011.

[12] C. Laudy and N. Museux, "How to evaluate high level fusion algorithms?" in *22^{nd} International Conference on Information Fusion (FUSION'19)*, Ottawa, Canada, 07 2019.

[13] Wikipedia contributors, "Notam — Wikipedia, the free encyclopedia," 2004, [Online; accessed 05-February-2024]. [Online]. Available: https://en.wikipedia.org/wiki/NOTAM

[14] OPSGROUP, "International notam survey – final report 2020," 2020, [Online; accessed 23-February-2024]. [Online]. Available: https://fixingnotams.org/wp-content/uploads/2020/12/International-NOTAM-Survey-Report-2020-OPSGROUP.pdf

[15] Wikipedia contributors, "Malaysia airlines flight 17 — Wikipedia, the free encyclopedia," 2004, [Online; accessed 07-February-2024]. [Online]. Available: https://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_17

[16] I. Kovacic, D. Steiner, C. Schuetz, B. Neumayr, F. Burgstaller, M. Schrefl, and S. Wilson, "Ontology-based data description and dis-

covery in a swim environment," in *2017 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, 2017, pp. 5A4–1–5A4–13.

[17] R. M. Keller, "The nasa air traffic management ontology: Technical documentation," 2017, [Online; accessed 07-February-2024]. [Online]. Available: https://ntrs.nasa.gov/citations/20170006095

[18] "Faa defense internet notam service," https://www.notams.faa.gov/dinsQueryWeb/, [Online; accessed 26-February-2024].

[19] "Faker's documentation," 2024, [Online; accessed 07-February-2024]. [Online]. Available: https://faker.readthedocs.io/en/master/#

[20] JSON Schema Community, "Json schema specification draft 2029-09," https://json-schema.org/specification-links#draft-2019-09-(formerly-known-as-draft-8), 2019, accessed: 2023-11.

[21] M. Grabisch, "k-order additive discrete fuzzy measures and their representation," *Fuzzy Sets and Systems*, vol. 92, no. 2, pp. 167 – 189, 1997, fuzzy Measures and Integrals. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011497001681

[22] ——, "The möbius transform on symmetric ordered structures and its application to capacities on finite sets," *Discrete Mathematics*, vol. 287, no. 1, pp. 17 – 34, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012365X04002936

[23] M. Grabisch and C. Labreuche, *Fuzzy Measures and Integrals in MCDA*. New York, NY: Springer New York, 2005, pp. 563–604. [Online]. Available: https://doi.org/10.1007/0-387-23081-5_14

[24] ——, "A decade of application of the choquet and sugeno integrals in multi-criteria decision aid," *4OR*, vol. 6, no. 1, pp. 1–44, Mar 2008. [Online]. Available: https://doi.org/10.1007/s10288-007-0064-2

[25] Wikipedia contributors, "Support vector machine — Wikipedia, the free encyclopedia," 1992, [Online; accessed 26-February-2024]. [Online]. Available: https://en.wikipedia.org/wiki/Support_vector_machine

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system (arxiv.org)," 2016, [Online; accessed 26-February-2024]. [Online]. Available: https://arxiv.org/abs/1603.02754

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding (arxiv.org)," 2018, [Online; accessed 26-February-2024]. [Online]. Available: https://arxiv.org/abs/1810.04805

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer (arxiv.org)," 2019, [Online; accessed 26-February-2024]. [Online]. Available: https://arxiv.org/abs/1810.04805

[29] G. Yenduri, R. M, C. S. G, S. Y, G. Srivastava, P. K. R. Maddikunta, D. R. G, R. H. Jhaveri, P. B, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions (arxiv.org)," 2023, [Online; accessed 26-February-2024]. [Online]. Available: https://arxiv.org/abs/2305.10435

[30] Docker Inc, "Docker: Software development and containerization technologies," https://www.docker.com/, 2013, accessed: 2024-02.

[31] N. Museux, C. Laudy, and M. Florea, "Houses bombing in ravixe: A bench for high level fusion evaluation," in *22nd International Conference on Information Fusion (FUSION'19)*, Ottawa, Canada, 07 2019.