Available Online at http://www.journalajst.com

ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ISSN: 0976-3376

Asian Journal of Science and Technology
Vol. 11, Issue, 01, pp.10722-10733, January, 2020

RESEARCH ARTICLE

MLM-BASED LEARNING AND BOOSTING MODEL – PART 1: MULTI- SOURCES/RIGHTS OF DIGITAL RESOURCES TO BUILD UNIVERSAL KNOWLEDGE REPOSITORIES USING AN ENRICHED SEMANTIC MICRO METADATA HARVESTER ENGINE AND SEMANTIC SHARED KNOWLEDGE NOTICE

Ronald Brisebois[1], Apollinaire Nadembega[1,*], Toufic Hajj[1] and Cédric Charles[1]

In Media Technologies, Montréal, Canada

ARTICLE INFO

Article History:
Received 15th October, 2019
Received in revised form
09th November, 2019
Accepted 27th December, 2019
Published online 31st January, 2020

Key words:
Digital Resources, Entity Resolution, Machine Learning Models, Metadata, MicroMetadata, Scorm, Semantic .Shared Knowledge Notice.

ABSTRACT

The wide proliferation of various wireless communication systems and devices has led to the arrival of a massive amount of Digital Resources (DR) from multi-sources, various metadata and media. However, data integration has allowed the ability to provide to users a uniform interface for multiple heterogonous data sources, metadata and users. Hence, the problem of matching which contents or DR belong to a specific user interest that demands more attention. In this article, we proposed a different model named: Learning & Boosting Architecture Model (LBAM). LBAM has goals to identify evolving interests of a person and to potentially propose a personal agenda, channels and activities. The first process is based on the creation of a hub of multiple sources of Micro Metadata (MM) using a Semantic Enriched MM Harvestor, a Watch & Notify Engine and a Semantic Shared Knowledge Notice (SSKN). They are harvested through a process able to catalogue the rights, interests and novelties in a scorm notice. It uses Machine Learning Models to improve the auto cataloguing of the DRs. It includes a Semantic Learning Watch and Notify engine using SSKN that allows ways to find DR or Event novelties of DR according to the evolving user interests. Using simulation studies and prototypes, we demonstrate that LBAM slightly improves accuracy in harvesting treatment from Entity Resolution and Linked Data compared to existing models using SSKN. We also demonstrate the integration of MM rights in a notice compared to other existing architectures. This article is the first paper of multiple for the LBAM project.

Citation: ***Ronald Brisebois, Apollinaire Nadembega, Toufic Hajj and Cédric Charles*, 2020.** "Mlm-based learning & boosting model – part 1: multi-sources/rights of digital resources to build universal knowledge repositories using an enriched semantic micro metadata harvester engine and semantic shared knowledge notice", *Asian Journal of Science and Technology*, 11, (01), 10722-10733.

INTRODUCTION

The ties between Data Integration (DI) and Machine Learning Models (MLM) have always been apparent. However, the volume and variety of data consumed by modern analytical pipelines have greatly strengthened the connections between data integration and machine learning. DI systems are increasingly looking to use MLM to automate parts of different integration tasks, such as (i) data cataloguing and inferring the schema of raw data, (ii) data alignment, (iii) metadata enrichment and (iv) transformation recommendations for data normalization, while machine learning models are only as good as the data used for training, which means that one must utilize data from the greatest possible variety of sources (1). In a previous paper we proposed an enhanced model for classification of metadata and enriched metadata (1), see Fig. 1.

Corresponding author: **Ronald Brisebois,**
In Media Technologies, Montréal, Canada.

We have added the capability to manage for each metadata the sources to our existing model, the rights and the enrichments associated for any notice. We called metadata at the last level of the hierarchy Micro Metadata (MM). Entity Resolution (ER) is an unavoidable and arguably the most important problem in integrating data from multiple sources keeping track of the rights and sources. Whereas schema alignment is also important, it can often be solved manually because the size of a schema is typically small; in contrast, we often need to match at least thousands of entities from different sources, making manual solutions seldom an option. ER consists to three steps: blocking records that are likely to refer to the same real world entity, comparing pairs of records to decide if it is a match and clustering records according to pair wise matching results, such that each cluster corresponds to a real-world entity. The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 describes part 1 of MLM based Learning & Boosting Model and introduces its various algorithms while Section 4 presents the evaluation through a prototype and a number of simulations. Section 5 presents a summary and some future work.

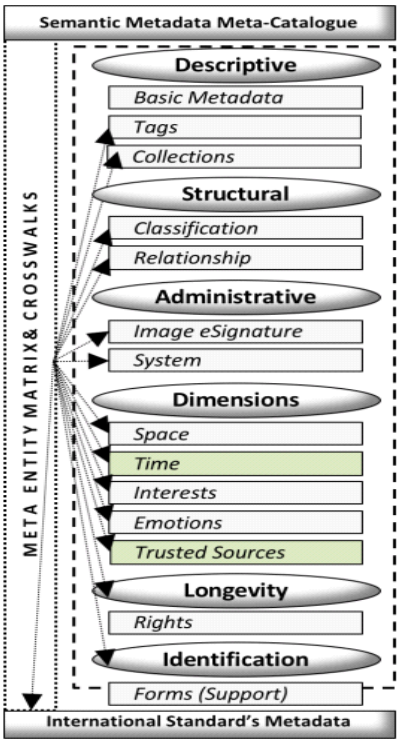

Fig. 1. Semantic metadata meta-catalogue enhanced classification

The other processes and LBAM architecture will be treated in future publications.

Related work: Building of a multi-source, multi-rights hub (2, 3) and multi-enrichments where the contents are linked as a structured LD, web harvesting process from multiple data sources, with their own unstructured data model remains a challenge. To achieve this there are four main processes: metadata harvesting (MH) (4-13), data integration (DI) (1)(14)(15) entity resolution (ER) (16)(17)(18)(19) (20)(21) (22) (23) (24)(25)(26)(27)(28)(29)(30)(31)(32) and content linkage (CL) (29)(33)(34) (35)(36)(37) (38)(39) (40) (41) (42) (43)(44)(45)(46)(47)(48)(49); several word are done for MH and DI; however, for ER and CL, there are still many issues to resolve. In this section we will focus on ER and CL, present an overview and the existing limits. Indeed, Big data integration consists to:

- ***Schema Mapping***: it refers to creating a mediated schema, and identifying the mappings between the mediated schema and the local schemas of the data sources to determine which attributes contain the same information; our previous studies proposed a model, call SMESE (2, 3).
- ***Content linkage***: refers to the task of identifying records linked to the same logical entity across different data sources. Record linkage (RL) is a process of finding records that correspond to the same entity from one or more data sources (50).
- ***ER or fusion***: it refers to resolving conflicts from different sources.

DI and CL are two axes of the Big Data research field. Big Data may be defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" or "data too big to be handled and analyzed by traditional database protocols such as SQL" (40). More authors assume that size is not the only feature of Big Data. They use the Five V's (Volume, Variety, Velocity, Value and Veracity) to characterize Big Data. Big data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner while cloud computing provides the underlying engine through the use of distributed data processing platforms. For example, the disambiguation pile process. One of the main step of disambiguation pile is the ERM (16) (17) (21) (18) (19) (20) (22) (51) (23). However, before ERM process, DI and CL are the first challenges of data management in the context of big data and cloud computing. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis (38). For example, the issue of merging Big Data catalogues in an already existing information system is discussed. In the context of this work, two issues of big data management are addressed: acquisition and organization. For acquisition, we have to acquire high speed data from a variety of sources and have to deal with diverse access protocols. Knoblock and Szekely (52) described how they exploited semantics to address the problem of big data variety. They proposed an approach to integrate data from multiple types of sources and in widely different formats, including both relational and hierarchical data. They implemented their approach to using semantics for big data integration in a system called Karma. Karma allows a user to import data from a wide variety of sources, clean and normalize the data, quickly build a model or semantic description of each source, and integrate the data across sources using this model. According to Authors, Karma performed an analysis of the data distribution in each column such as the frequency of different values, frequency of values whose type is different from that of the majority of values or frequency of null values.To illustrate the approach, they used a dataset from the cultural heritage domain in order to build a virtual museum that integrates the metadata about artwork several museums.

One main limitation of their approach is the fact that the data comes from an already structured database which is not typical most of the time. Karma is limited to find noisy, missing, or inconsistent data; unfortunately, we may conclude that Karma is not useful for entry resolution. Bellini *et al.*(50)

proposed a system for data integration and reconciliation of smart city related aspects as road graph, services available on the roads and traffic sensors. According to authors, their system allows managing a big data volume of data coming from a variety of sources considering both static and dynamic data which are mapped to a smart-city ontology called

KM4City. *Unfortunately, their KM4City (proposed knowledge model for Smart City) is limited to seven areas. In addition, they did not take into account the data generated by citizens. Finally, authors did not propose their own data integration model, they used an existing model called Pentaho Kettle formalism*.

Raul Castro *et al.* (53) proposed a data integration stack that provides low latency data access to support near real-time in addition to batch applications, called Liquid.

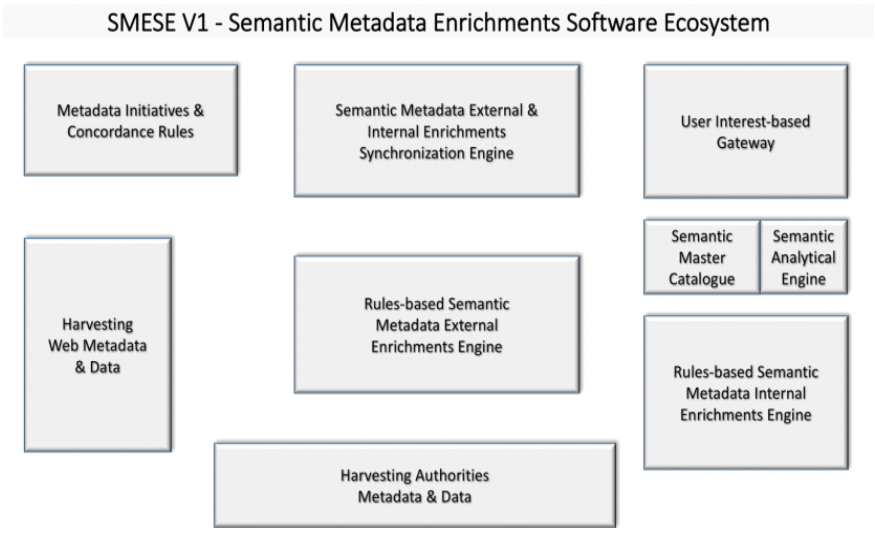

Fig. 2. Semantic Enriched Metadata Software Ecosystem (SMESE)

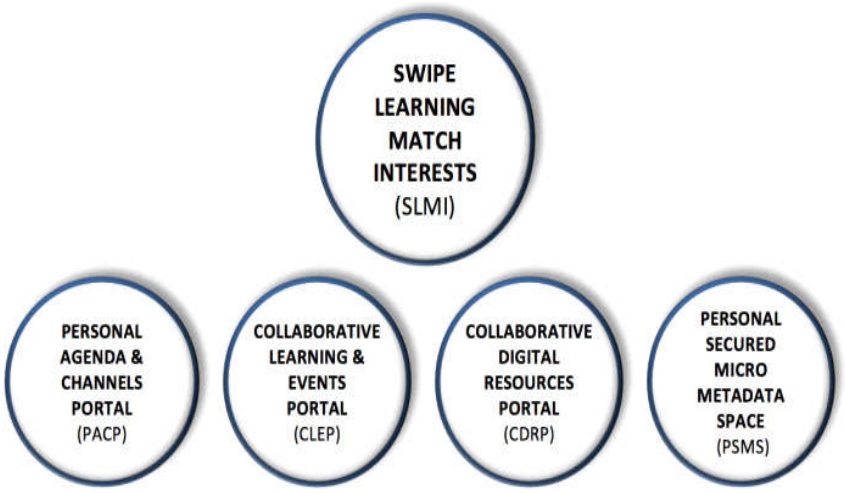

Fig. 3. LB project outputs

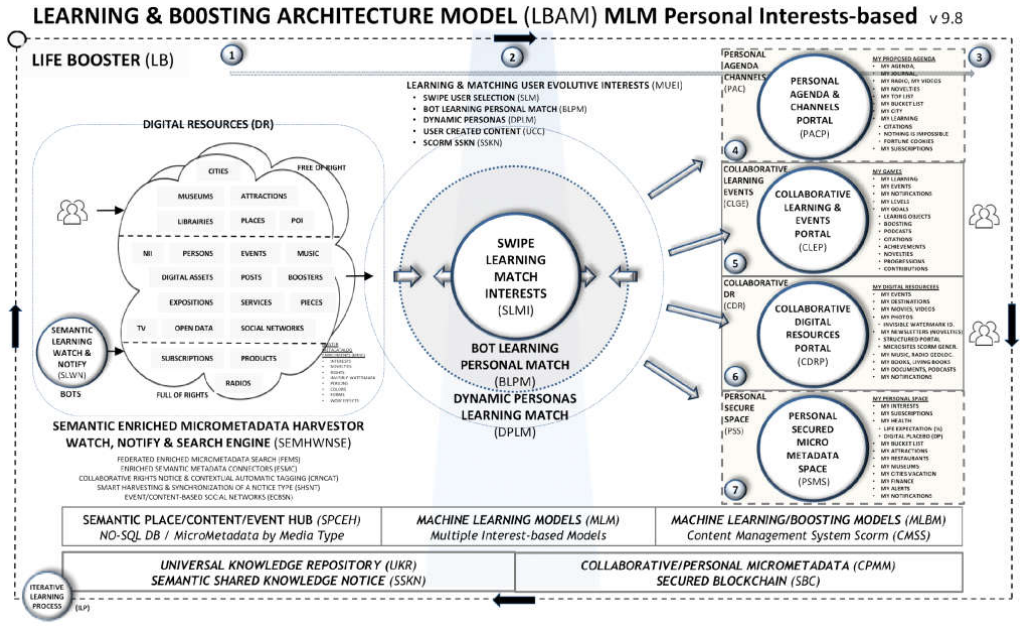

Fig. 4. LBAM Overview Model

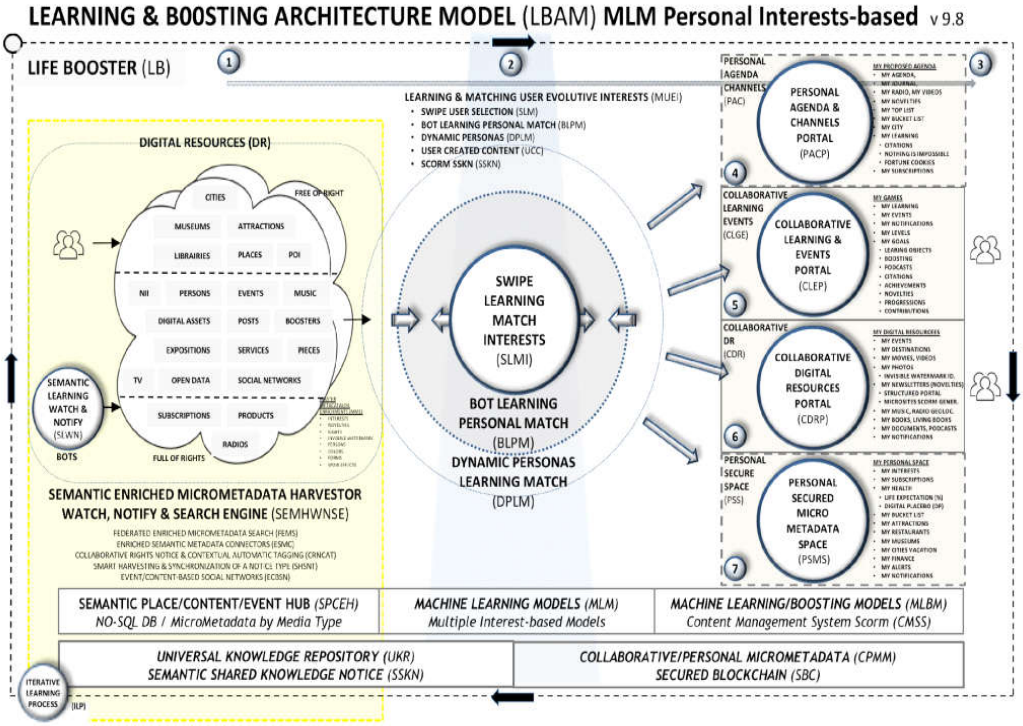

Fig. 5. LBAM Part 1 (in yellow)

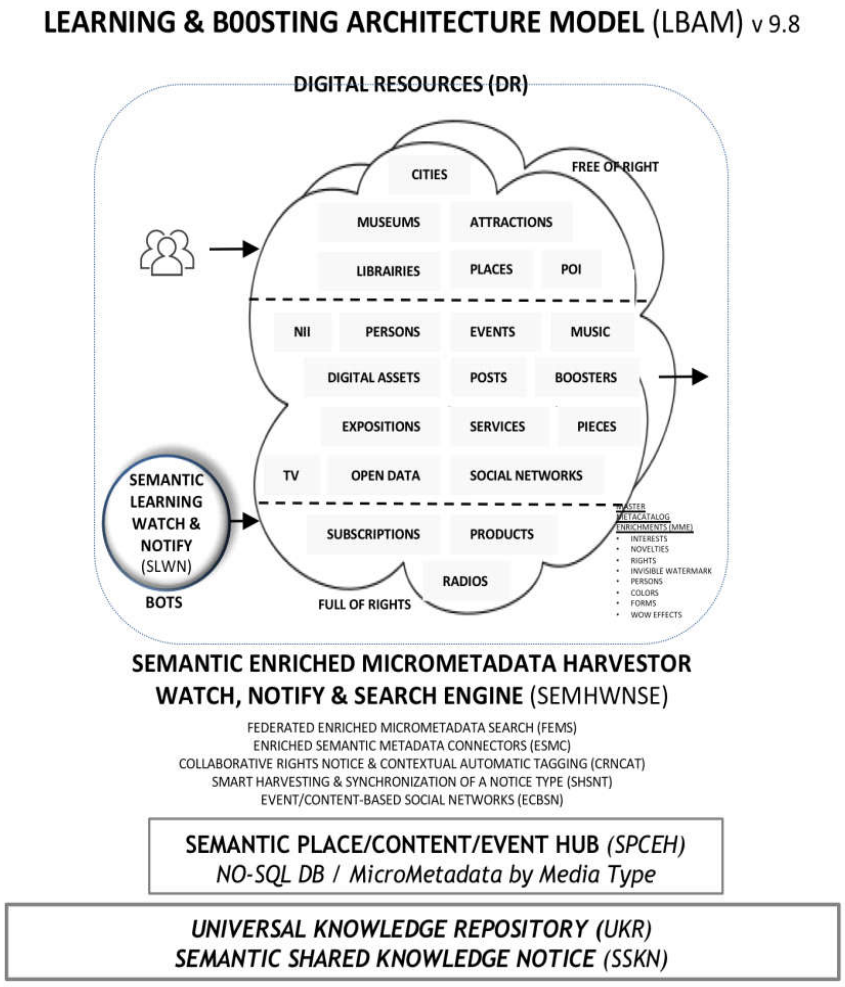

Fig. 6. Universal Knowledge Repository and Digital Resources

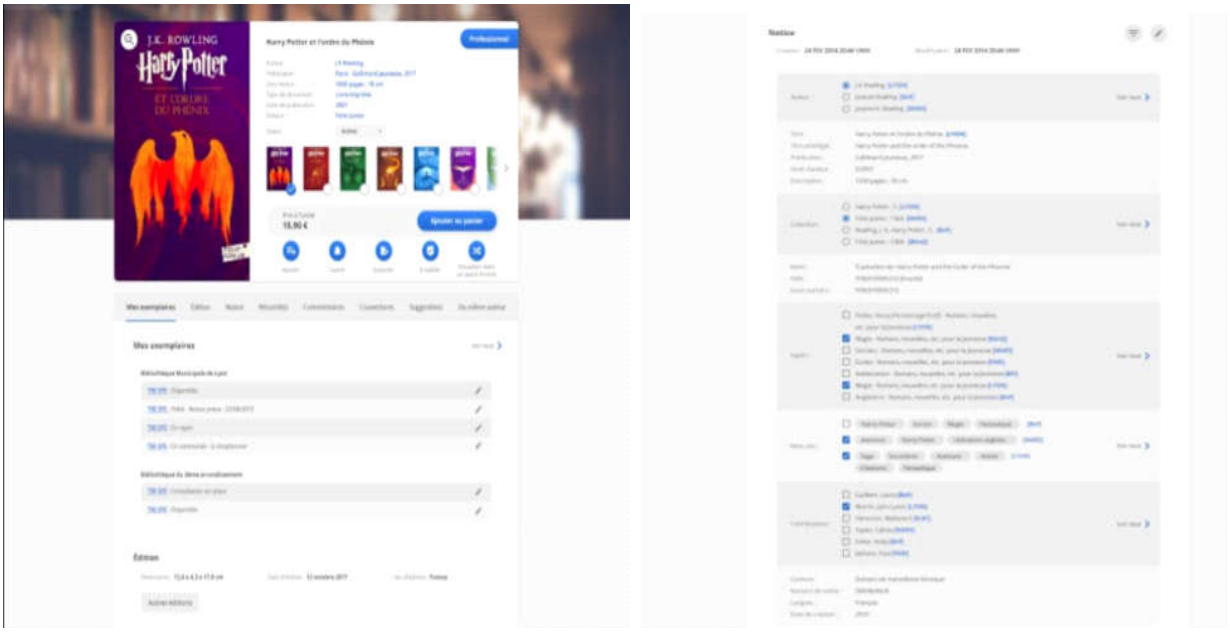

Fig. 7. A book SSKN notice ¨Harry Potter¨ – Collaborative cataloguing

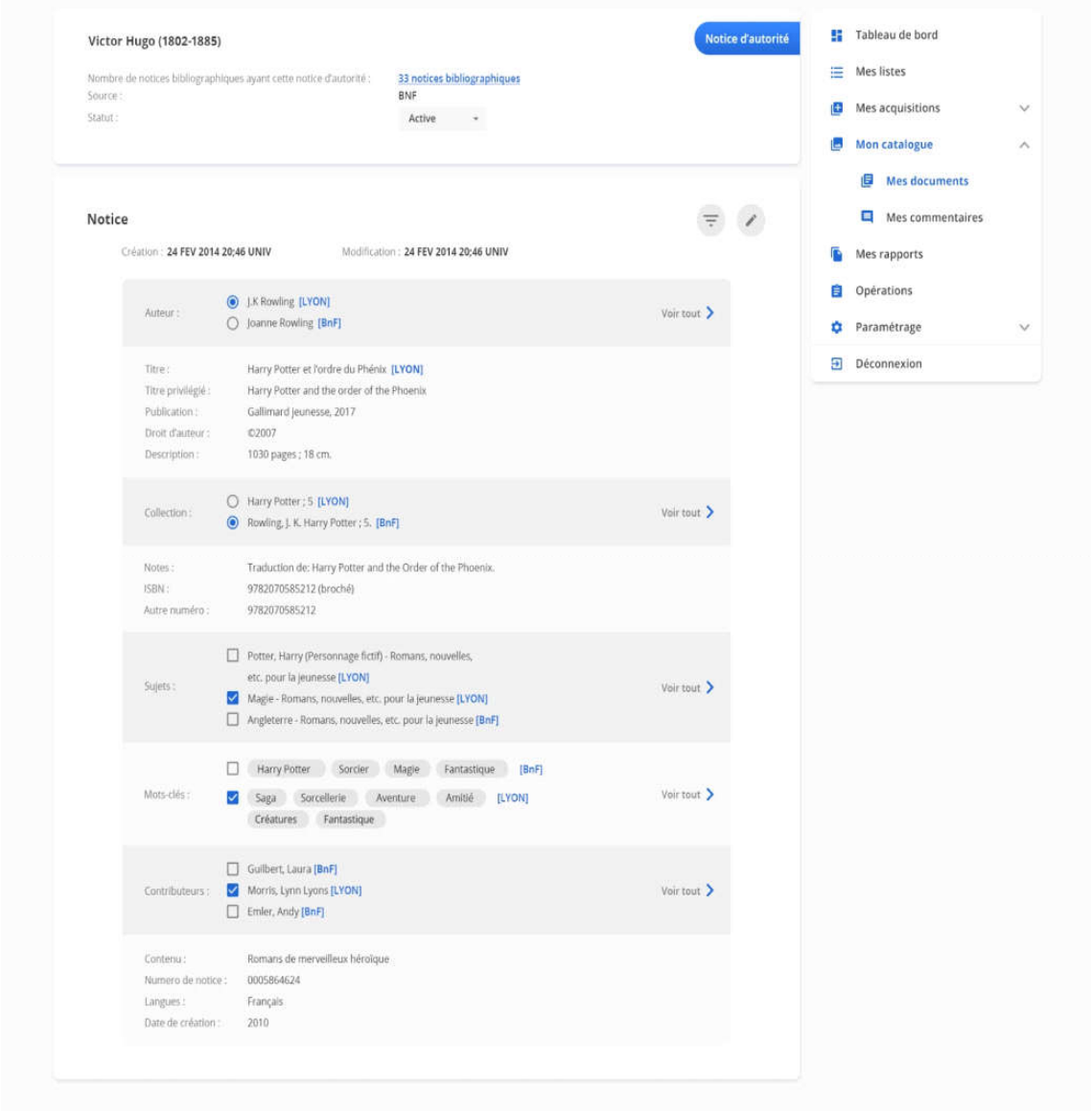

Fig. 8. A book SSKN notice ¨Victor Hugo¨ – Collaborative cataloguing

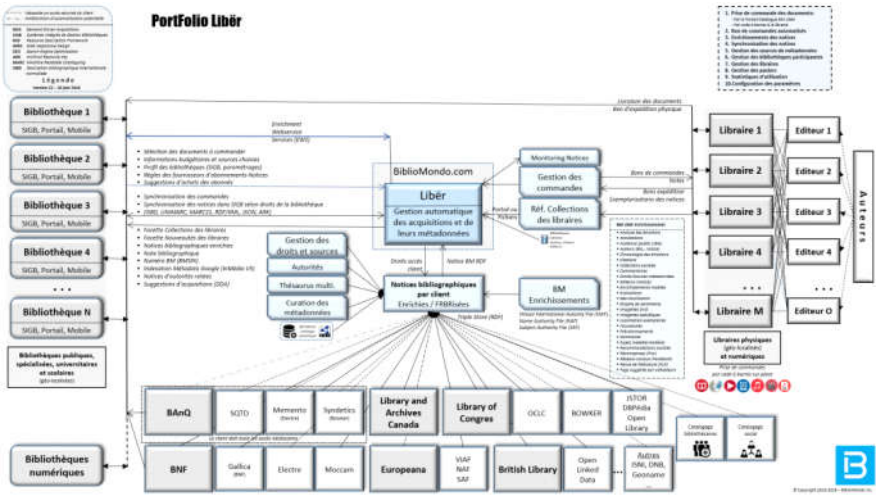

Fig. 9. Libër Project – Collaborative Cataloguing of Notices (CCN)

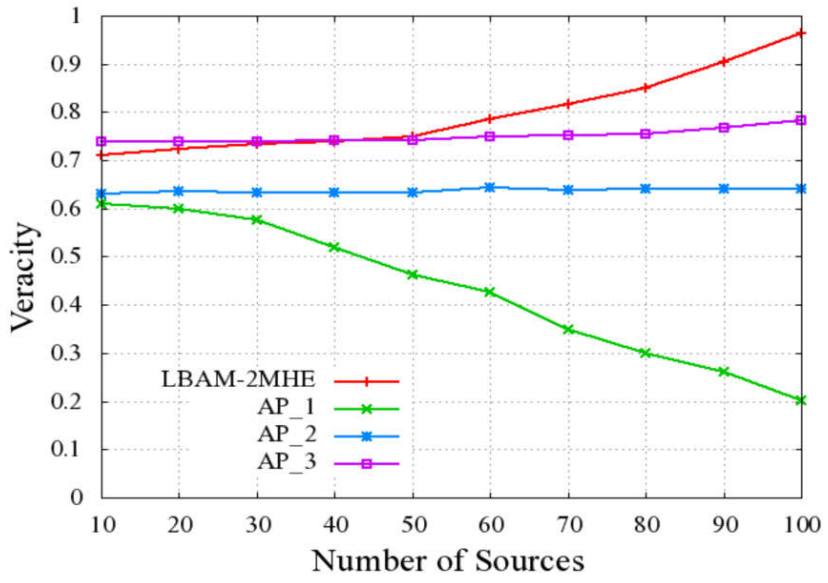

Fig. 10. Veracity VS Number of Sources

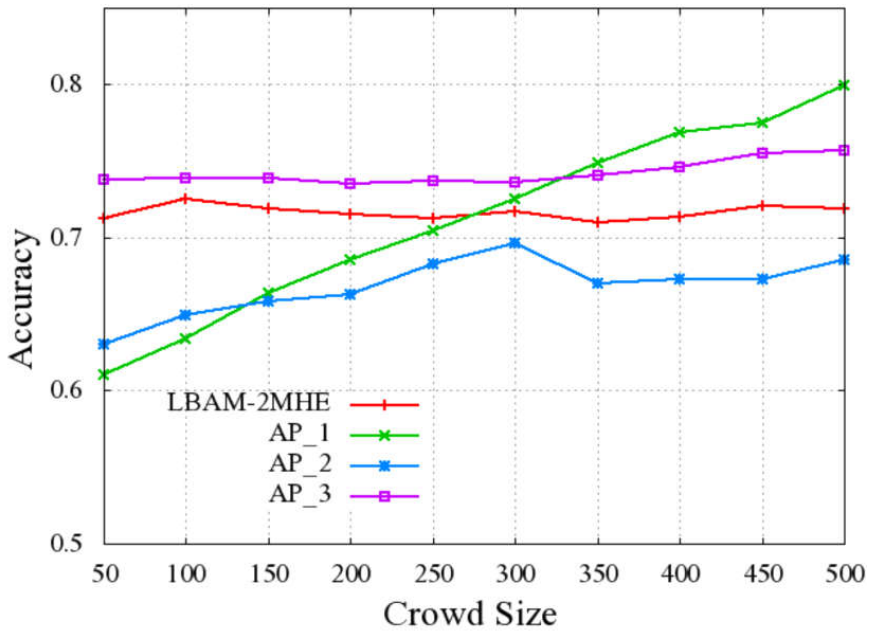

Fig. 11. Accuracy VS Crowd Size

Table 1. Prototype's evaluation dataset entity types

Entities Type	Number of entities
Books	1,183,437
Videos	1,435
Music	458
Museums	55,324
Documents	71,724

According to authors, Liquid consists of two cooperating layers: a messaging and processing layer. The messaging layer (based on uses Apache Kafka) provides data access based on metadata, which permits back-end systems to read data from specific points in time while the processing layer (based on Apache Samza) executes ETL-like jobs for back-end systems, guaranteeing low-latency data access. The two layers communicate by writing and reading data to and from two types of feeds stored in the messaging layer.

Unfortunately, Liquid is only for the messages instead of ER: Literature mentions that ERM is the more important task after data harvesting from multi-sources in the context of metadata integration in order to build a unified and trusted repository (UTR). According to (54), any important data management, such as ERM, cannot be completely addressed by existing algorithms and automated processes. These tasks can be enhanced through the use of human cognitive ability. The objective of ERM is to identify which records (entities) refer to the same real-world entity; this task is fundamental in data integration. Lots of approaches have been proposed to improve the quality of entity resolution such as combining different methods, an iterative approach, and the use of functional dependencies. Kardes *et al.* (55) proposed an entity resolution for the organization entity domain based on blocking and clustering strategies where all they have are the organization names and their relations with individuals. Authors assumed that if they show different representations of the same organization as separate institutions in a single person's profile, it will increase the performance of their ER approach in terms of accuracy. *The main limit of their approach is the fact that is based on person profile. How will their ER approach be implemented without person profiles?* Vesdapunt *et al.* (20) proposed a hybrid human-machine approach for solving the problem of Entity Resolution. In their approach, a machine learned model first assigns candidate pairs of records a probability of how likely they are to be duplicates, and then we ask humans questions about record pairs until we have completely resolved all records in our database. Authors considered the problem of devising optimal strategies for asking questions to the crowd, based on the pairwise matching probabilities that minimize the expected numbers of questions required.

This approach requests human contribution for certain ER and are not trusted. This task is different from user feedback to enrich a machine learning model. The accuracy of their approach is strongly linked to the quality of the crowd responses. Efthymiou *et al.* (56) focused on entity resolution in the Web of data performing blocking method. Blocking is used as a pre-processing step for ER to reduce the number of required comparisons. Specifically, the authors distinguished between data originating from sources in the center (i.e., heavily interlinked) and the periphery (i.e., sparsely interlinked) of the LOD cloud to capture the differences in the heterogeneity and overlap of entity descriptions. Thus, they studied the behavior of existing blocking algorithms for datasets exhibiting different semantic and structural characteristics. They presented the results of blocking in terms of owl: same As links and other kinds of links as a ground truth. *Unfortunately, authors' contribution are limited to the evaluation of a cluster of 15 machines using real data.* Whang *et al.* (51) explored a pay-as-you-go approach to entity resolution. They investigated how to maximize the progress of ER with a limited amount of work using "hints," which give information on records that are likely to refer to the same real-world entity. The author's approach addressed three important questions: how to construct the hints, how to use the hints, what cases does pay-as-you-go pay off? Unfortunately, the goal of this work is just to provide a unifying framework for hints and to evaluate the potential gains. *Their work is empirical by nature and the hints heuristic. Their work is proposed as representative cases*. Zhu *et al.* (57) addressed the problem of performing entity resolution on RDF graphs containing multiple types of nodes, using the links between instances of different types to improve accuracy. They modeled the observed RDF graph as a multi-type graph and formulate the collective entity resolution as a multi-type graph summarization problem; the goal is to transform the original k-type graph into another k-type summary graph composed of super nodes and super edges where each super node is a cluster of original vertices representing a latent entity, while super edges encode potentially valuable relations between those entities. The *author's approach is based on a metadata that have all the entities such as the manufacturer of product or authors of papers. In the context of Web Big Data, this case is very rare and cannot be applied to any domain. In addition, as (55), their approach is strongly linked to a specific metadata. An important question is what happens if this metadata is empty?* Jurek *et al.* (58) proposed a new approach to unsupervised content linkage based on a combination of ensemble learning and enhanced automatic self-learning.

Their approach incorporates an ensemble of learning and self-learning techniques into content linkage. They generated an ensemble of diverse self-learning models by applying different combinations of similarity measure. By using different combinations of similarity measures they generated different sets of similarity vectors that could be used to generate different self-learning models. To ensure high diversity among the self-learning models they applied the proposed seed Q-statistic diversity measure. They also used Contribution Ratios of BCs to eliminate those with very poor accuracy. *Authors combine existing approaches that improve the existing automatic self-learning technique for CL. In conclusion, several limitations are not solved (i.e., ER and CR) in order to build the semantic linked to Digital Assets, Persons, Events, Subscriptions and many related media.* Multi-Sources/Rights of Digital Resources to build Universal Knowledge Repositories (UKR) using an Enriched Semantic MM Engines and Semantic Shared Knowledge Notice (SSKN). In this section, we present the details of our proposed approach. First, we introduce MLM based Learning & Boosting Model and second, the details of LBAM algorithms and models (Part 1). From our previous research, we described that metadata in catalogues represent resource characteristics that can be indexed, queried and displayed by both humans and machine. The SMESE semantic ecosystem harvests and enrichs

metadata and MM. We can see in Fig. 2, the main components of the SMESE ecosystem.

Many aggregators harvest metadata and consequently data that, in the process, may become inaccurate because they did not look at the semantic context of the sources, the reputation of the source, neither to their timely accuracy, the usage of a meta-catalogue and the MM. The SMESE ecosystem defines crosswalks that. For further understanding about SMESE algorithms and processes to semantically enrich metadata using multiple metadata/data sources, refer to previous papers (59, 60). The Life Booster project proposed to use the SMESE platform for the creation of User Evolutive Interests, 3 portals (Personal Agenda & Channels, Collaborative Learning & Events, Collaborative Digital Resources) and 1 Personal User Space – see Fig. 3.

Overview of Life Booster project: The Learning & Boosting Architecture Model (LBAM) (Fig. 4), a Machine Learning Interest-based Model, has several goals: identify Matching Evolving User Interest (MEUI) of person and potentially to a Daily Agenda and Channels according to user interests who evolve periodically. This LBAM model is built from 3 main processes: a) Identification of the MM of Digital Resources (DR) including Events and their timeline (novelties) and ongoing enrichment, b) Matching Evolutive User Interest (MEUI) using a Chat Bot and a swipe action and c) The Daily Smart Booster Agenda created to suggest DR according to the evolutive user interests. This project is called Life Booster (LB) and is intended to keep track of the rights of the contents (Digital Assets) or Events, the Evolving User Interests and Machine Learning/Boosting Processes who are part of an Iterative Learning Process (IPL) shown in the next figure (see Fig. 5).

The first process (see Fig. 5) is based on the creation of a hub of secured multiple metadata using the Semantic Enriched MM Harvester, Watch, Notify & Search Engine linked to Users and Bots (SLWN) and includes multiple sources of rights and their aggregation into MM by Media Type using Multi Sources Semantic Knowledge (SSKN). These SSKN is used to create enriched MM. These Metadata are assembled through a Harvesting process able to catalogue the rights, interests and the novelties. This process includes Sub-processes named: Federated Enriched MM Search, Enriched Semantic Metadata Connectors, Collaborative Rights Notice & Contextual Automatic Tagging, Smart Harvesting & Synchronization of a Notice Type and Event/Content-based Social Network. This process also includes the ability for the User or the Merchant to create or update media and metadata. This harvesting process has to keep track of the Novelties as well. SLWN allows the ability to keep track of any event which may interest some watching and notifying process in the system. It includes the following main process: Federated Enriched MM Search, Enriched Semantic Metadata Connectors – (Enrichments are per examples: Interests, Novelties, Persons, etc.), Collaborative Rights Notice & Contextual Automatic Tagging, Smart Harvesting & Synchronization of a Notice Type and Event/Content-Based Social Networks. This process harvests Free of right and Full of Right DR and manages the MM multi-rights. The second process is mainly serves to identify the Matching Evolutive User Interest (MEUI) by an Algorithm of matching from four different levels of User Interests: The User Personal Interest using the real time Swipe Learning Match Interests (SLMI),the Interests of the Personas of the User using Dynamic Personas Learning Match (DPLM) – the Personas of the Users are categorized in 18 different personas in our model, the Bot swipe as a counterpart for Swipe Learning Match Interests (SLMI) using Bot Learning Match (BLM) – a simulator of automatic matching interests based on a set of user with the 95% of the same Personas and user Created Content (UCC) allowing to extract some behavior from the User. The Bot Learning Match (BLM) is an assisted process (Bot) allows User Interests for Digital Assets as Events to be matched, Photos, Persons, etc. This process uses Multiple Interest-based Models to learn the User Interests in different situations with the Swipe principle to like (right) or dislike (left), time of the day and contextual behavior. Using MLM, this process improve the MEUI identification over the learning process.

The third process focuses on the prediction of the daily evolving interests of each user and context regarding. The parts of the third process include: Personal Agenda & Channels Portal – it is a personal Journal, a personal Radio and a personal Channels Portal (PACP). Here we build a recommended agenda, journal, radio channel and video channel to a specific user according to all five processes of LBAM and their evolving personal interests. PACP proposes an Agenda for the day or the coming week to the User. Every day of this agenda is refined according to the usage and interests of the users. This process uses Machine Learning/Boosting Models to improve the cataloguing of the Digital Asset and Events, boost interest of User and improve the identification of the User Interests. This process also places an emphasis on Collaborative Learning & Events Portal (CLEP) and gives games or learning activities to do according to the User's Interest. The Collaborative Digital Resources Hub – Collaborative Digital Resources identifies potential Events and Media who could meet the Evolutive Interests of the User. The last process is Secured Personal MM Space (SPMS) - My Personal Space, where the user can manage their configuration, interests, digital resources, events and agenda and regroup all personal information and interests. This process is based on Machine Learning/Boosting models. The fourth process is the Personal Agenda & Channels Portal (PACP) process but with an emphasis on the personal channels process. It allows a method to propose a dedicated Personal Channel to a user according to their interests and available Digital Resources at a specific time. This Personal Agenda & Channels Portal is using MLBM evolving with time and all interactions with the user. The fifth process named Collaborative Learning & Events Portal (CLEP) includes the sharing of knowledge and gaming for the benefit of each user. The process includes the ability to create, reference, evaluate and organize content or knowledge in an evolutive learning process at different levels. It allows Digital Resources to be accessed and used by a multitude of users in multiple languages.

The sixth process is the Collaborative Digital Resources Portal (CDRP). The process includes My Newsletter who fulfill the CDRP to create content and digital resources per different interest categories and learning needs. This process also includes a CMS based Micro-Sites Generator using newsletter smart aggregation to create new content and knowledge as well as notifications and alerts according to the interests of the users, it is called Watch for me (What4me). The seventh

process is the Secured Personal MM Space (SPMS) but with an emphasis on Personal Metrics and Digital Placebo (DP).

The process includes in My Health, the Life expectation metric and the DP who intend to help User to reach a better level on MEUI. All these seven processes are embedded in a larger Machine Learning Mechanism allowing the ability to learn at different stages of the macro process and to improve all other learning processes. We call this critical process: Iterative Learning Process (ILP). We will explore more in details the first process of this model in this first article of LBAM. These other processes will be treated in future publications.

UKR Process & Algorithms: Many process are involved in the creation of the UKR and SSKN, see version 9.8 of the model Fig. 6. The semantic enriched MM Harvester is the cornerstone of the UKR and ultimately the Collaborative Digital Resources Portal (CDRP).

Semantic Share Knowledge Notice (SSKN): The Semantic Share Knowledge Notice is the base to catalogue multiple sources and multiple rights into MM, see Fig. 7 for an example – a book SSKN notice. This cataloguing includes the capability to support Scorm Notice for the Learning process. The Persona creates the space to identify and learning automatically from behavior of a group of people. It reflects the interests of a group of people. Notice View according to persona and cataloguing rules: Users of persona (A) will see the metadata which are catalogued/enriched by users of the same persona (A); for example, users of persona (A) will see the metadata "Description_Persona_A" while users of persona(B) will see the metadata "Description_Persona_B" of the same notice. The Fig. 8 shows a second example of a SSKN notice related to a book by Victor Hugo. We can see the MM and the collaborating cataloging notice with multiple rights, sources and enrichments.

Machine learning model (MLM): MLM algorithms are used at different levels in LBAM to identify the evolutive interests of users. It uses the same model as SMESE but enhances the process to identify MM sources in the structured environment and unstructured web.

Prototype Applications and Evaluation using simulations: In this section we present the experimental evaluation of our proposed approach. The objective of our experimental evaluation is to compare, according to the literature, more recent and performing algorithms on various types of entities.
In the Libër project, we can see the model of harvesting all MM to create de Libër Repository, see Fig. 9. This prototype has to primary goal of validating the accuracy and precision for:

- Entity resolution;
- Content linkage;
- MM rights and sources preservation;
- Universal Knowledge Repositories.

Simulation Setup and Datasets Characteristics: The Datasets we use were provided by five (5) data sources of various notice types. The overall datasets contain more than one million of entities and each entity contains MM. Datasets consist of five (5) types of real entities: books, videos, music, museum and documents. Table 1 shows each dataset entities types and their count.

Performance measurement criteria: The goal of the following section is to compare the performance of three different approaches with our proposal. Our goal is to investigate the benefits of operating at the finest level of granularity. To evaluate the behavior of comparison approaches, we employed two kinds of measures: the veracity of metadata in the central repositories when varying the number of sources to harvest and the accuracy when varying the crowd size that contribute to clean the harvested metadata. As comparison terms, we use the approaches described in (24), (57), and (50), which are referred to as AP_1, AP_2, and AP_3, respectively.

RESULTS AND DISCUSSION

For each run, we compute the performance metrics. To obtain the simulation results, we compute the average of the 10 runs.
In Fig. 10, we evaluate the average veracity of data varying with the number of sources to harvest while in Fig. 11 shows the average accuracy when varying the crowd size. In Fig. 10, we observe that for LBAM-2MHE and AP_3 (resp., AP_2) the average veracity increases (resp., remains constant) with the number of sources. We also observe that for the AP_1, the average veracity decreases with the number of sources. Fig. 10 also shows that LBAM-2MHE outperforms AP_1, AP_2, and AP_3.For example, LBAM-2MHE provides an average veracity of 0.797 for ten harvested sources, whereas AP_3 (more efficient than AP_1 and AP_2 in this scenario) provides an average of 0.751 for ten harvested sources. Overall, the average relative improvement of LBAM-2MHE compared with AP_3 is about 04% for ten harvested sources. We observe in Fig. 10 that, at fifty harvested sources, LBAM-2MHE increases faster than AP_3; that mains that LBAM-2MHE is more performing in the context of big data integration. This can be explained by the fact that LBAM-2MHE uses the watcher and notifier engine to validate the veracity of metadata.

LBAM-2MHE also includes a MM Model that increases the exactitude of metadata. Fig. 11 presents the average accuracy when varying the crowd size. The crowd size is the number of persons that contribute to improve the quality of data. In this scenario, only ten harvested sources of metadata are used. We observe that for LBAM-2MHE (resp., AP_1, AP_2 and AP_3) the average veracity remains constant (resp., increase) with the size of crowd. This means that LBAM-2MHE is not impact by the users contributions to improve the metadata quality. Despite this, LBAM-2MHE retains a higher performance than AP_1. For example, Fig. 11 shows that, LBAM-2MHE provides an average accuracy of 0.72 for a crowd size of fifty, whereas AP_1 provides an average of 0.70 for a crowd size of fifty. Overall, the average relative improvement of LBAM-2MHE compared with AP_1 is about 02% for fifty contributors (crowd size). We observe that AP_1 increases faster than AP_2 and AP_3 and signifies that AP_1 is more impacted by the crowd size. Remember that, as shown in the scenario of Fig. 10, LBAM-2MHE outperforms in the context of big data integration meaning that the amount of sources to harvest is large. We proposed a new model that aims is to create a hub of secured metadata in order to build the Semantic Place/Content/Event hub (SPCE) linked to Collaborative Digital Resources Portal. These metadata are assembled

through a harvesting process called Metadata Rights Interests Novelties Harvester (MRINH) that is able to catalogue the rights, the interests and the novelties.

The media could be with Open Rights or Free of Rights. Using simulation studies, we demonstrated that SPCE improves accuracy in estimates of treatment effects from ER and linked data compared to existent models. Finally, we put emphasis on the value of the metadata especially the MM. This article is the first part of our project, called LBAM.

Summary and future work: We have shown that it is possible and more accurate to harvest metadata using the MM classification. As an example, it is better to be able to harvest all the museum of the world to use a number of multi-sources, multi-rights of MM, using UKR and SSKN in a timely manner. Yet, there are many improvements that can be added to this model such as: improvements of the Harvesting Algorithms, improvements in the meta-catalogue structure and refinement of SSKN. Here are some of the future works that we looking to explore furthermore:

Second to Seventh Process: 2) Matching User Evolutive Interests (MUEI); 4) Personal Agenda & Channels Portal (PACP) Process; 5) Collaborative Learning & Events Portal (CLEP); 6) Collaborative Digital Resources Portal (CDRP); and 7) Personal Secured MM Space (PSMS). Process five, six and seven are the cornerstone to create the process 4 (PACP).

REFERENCES

(1) Dong, X. L. and Rekatsinas, T. 2018. "Data Integration and Machine Learning: A Natural Synergy," in Proceedings of the 2018 International Conference on Management of Data, Houston, TX, USA, pp. 1645–1650.

(2) Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "Efficient Scientific Research Literature Ranking Model based on Text and Data Mining Technique," *International Journal of Engineering Research And Management (IJERM),* vol. 04, no. 02, pp. 95-105, February 2017.

(3) Brisebois, R., Abran, A. and Nadembega, A. 2017. "A Semantic Metadata Enrichment Software Ecosystem (SMESE) based on a Multi-platform Metadata Model for Digital Libraries," *Journal of Software Engineering and Applications (JSEA),* vol. 10, pp. 370-405, April 30.

(4) Brisebois, R., Nadembega, A., N'techobo, P. and Djeuteu, H. L. 2017. "A semantic web metadata harvesting and enrichment model for digital library and social networks," *International Journal of Current Research (IJCR),* vol. 9, no. 10, pp. 59162-59171, October 2017.

(5) Vargiu, E. and Urru, M. 2013. "Exploiting web scraping in a collaborative filteringbased approach to web advertising," *Artificial Intelligence Research,* vol. 2, no. 1, pp. 44-54.

(6) S. Teli, "Metadata Harvesting From Selected Institutional Digital Repositories in India: A Model to Build a Central Repository," *International Journal of Innovative Research in Science,Engineering and Technology,* vol. 4, no. 4, pp. 1935-1942, 2015.

(7) S. Shi, C. Liu, Y. Shen, C. Yuan, and Y. Huang, "AutoRM: An effective approach for automatic Web data record mining," *Knowledge-Based Systems,* vol. 89, pp. 314-331, 2015/11/01/, 2015.

(8) N. R. Haddaway, "The Use of Web-scraping Software in Searching for Grey Literature," *The Grey Journal,* vol. 11, no. 3, 2015.

(9) V. B. Kadam, and G. K. Pakle, "A Survey on HTML Structure Aware and Tree Based Web Data Scraping Technique," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 2, pp. 1655-1658, 2014.

(10) D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," *Briefings in Bioinformatics,* vol. 15, no. 5, pp. 788-797, 2014.

(11) B. G. Dastidar, D. Banerjee, and S. Sengupta, "An Intelligent Survey of Personalized Information Retrieval using Web Scraper," *I.J. Education and Management Engineering,* vol. 5, pp. 24-31, 2016.

(12) A. Casali, C. Deco, and S. Beltramone, "An Assistant to Populate Repositories: Gathering Educational Digital Objects and Metadata Extraction," *IEEE RevistaIberoamericana de Tecnologias del Aprendizaje,* vol. 11, no. 2, pp. 87-94, 2016.

(13) G. Gupta, and I. Chhabra, "Optimized Template Detection and Extraction Algorithm for Web Scraping of Dynamic Web Pages," *Global Journal of Pure and Applied Mathematics,* vol. 13, no. 2, pp. 719-732, 2017.

(14) H.-N. Dai, R. C.-W. Wong, H. Wang, Z. Zheng, and A. V. Vasilakos, "Big Data Analytics for Large-scale Wireless Networks: Challenges and Opportunities," vol. 52, no. 5, pp. Article 99, 2019.

(15) J. Hui, L. Li, and Z. Zhang, "Integration of Big Data: A Survey," in 4th International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE 2018) Zhengzhou, China, 2018, pp. 101-121.

(16) R. C. Steorts, "Entity Resolution with Empirically Motivated Priors," *Bayesian Anal.,* vol. 10, no. 4, pp. 849-875, 2015/12, 2015.

(17) P. Christen, and R. W. Gayler, "Adaptive Temporal Entity Resolution on Dynamic Databases," in PAKDD 2013: Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg, 2013, pp. 558-569.

(18) A. Globerson, N. Lazic, S. Chakrabarti, A. Subramanya, M. Ringgaard, and F. Pereira, "Collective Entity Resolution with Multi-Focal Attention," in 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 2016, pp. 621–631.

(19) R. Nuray-Turan, D. V. Kalashnikov, and S. Mehrotra, "Adaptive Connection Strength Models for Relationship-Based Entity Resolution," vol. 4, no. 2, pp. Article 8, 2013.

(20) N. Vesdapunt, K. Bellare, and N. Dalvi, "Crowdsourcing algorithms for entity resolution," vol. 7, no. 12, pp. 1071–1082, 2014.

(21) B. Ramadan, and P. Christen, "Forest-Based Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China, 2014, pp. 1787–1790.

(22) D. Firmani, B. Saha, and D. Srivastava, "Online entity resolution using an Oracle," *Proceedings of the VLDB Endowment,* vol. 9, no. 5, pp. 384–395, 2016.

(23) G. Papadakis, G. Papastefanatos, T. Palpanas, and M. Koubarakis, *Scaling Entity Resolution to Large, Heterogeneous Data with Enhanced Meta-blocking*, 2016.

(24) C. Chai, G. Li, J. Li, D. Deng, and J. Feng, "Cost-Effective Crowdsourced Entity Resolution: A Partial-Order Approach," in Proceedings of the 2016 International Conference on Management of Data, San Francisco, California, USA, 2016, pp. 969–984.

(25) G. Simonini, S. Bergamaschi, and H. V. Jagadish, "BLAST: a loosely schema-aware meta-blocking approach for entity resolution," *Proceedings of the VLDB Endowment,* vol. 9, no. 12, pp. 1173–1184, 2016.

(26) A. Passos, V. Kumar, and A. McCallum, "Lexicon Infused Phrase Embeddings for Named Entity Resolution,"*arXiv e-prints*, https://ui.adsabs.harvard.edu/abs/2014arXiv1404.5367P, (April 01, 2014, 2014).

(27) T. Williams, and M. Scheutz, "POWER: A domain-independent algorithm for Probabilistic, Open-World Entity Resolution," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015, pp. 1230-1235.

(28) A. Elmagarmid, I. Ilyas, M. Ouzzani, J.-A. Quiané-Ruiz, N. Tang, and S. Yin, "NADEEF/ER: generic and interactive entity resolution," in ACM SIGMOD International Conference on Management of Data, Portland , OR , USA, 2014.

(29) M. Mountantonakis, and Y. Tzitzikas, "Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets," vol. 9, no. 3, pp. Article 15, 2018.

(30) G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl, "A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces," *IEEE Transactions on Knowledge and Data Engineering,* vol. 25, no. 12, pp. 2665-2682, 2013.

(31) B. Ramadan, P. Christen, H. Liang, and R. W. Gayler, "Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution," vol. 6, no. 4, pp. Article 15, 2015.

(32) J. Fisher, P. Christen, Q. Wang, and E. Rahm, "A Clustering-Based Framework to Control Block Sizes for Entity Resolution," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015, pp. 279–288.

(33) J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big data challenge: a data management perspective," *Frontiers of Computer Science,* vol. 7, no. 2, pp. 157-164, April 01, 2013.

(34) H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and its technical challenges," *COMMUNICATIONS OF THE ACM,* vol. 57, no. 07, pp. 86-94, 2014.

(35) X. L. Dong, and D. Srivastava, "Big data integration," in 2013 IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, QLD, Australia, 2013, pp. 1245-1248.

(36) C. L. Philip Chen, and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences,* vol. 275, pp. 314-347, 2014/08/10/, 2014.

(37) M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications,* vol. 19, no. 2, pp. 171-209, 2014/04/01, 2014.

(38) I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems,* vol. 47, pp. 98-115, 2015/01/01/, 2015.

(39) M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing,* vol. 79-80, pp. 3-15, 2015/05/01/, 2015.

(40) C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable Big Data: A survey," *Computer Science Review,* vol. 17, pp. 70-81, 2015/08/01/, 2015.

(41) L. Cai, and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal,* vol. 14, no. 02, pp. 1-10, 2015.

(42) M. Mountantonakis, and Y. Tzitzikas, "High Performance Methods for Linked Open Data Connectivity Analytics," *Information,* vol. 09, no. 06, 2018.

(43) J. H. Wortman, and J. P. Reiter, "Simultaneous record linkage and causal inference with propensity score subclassification," *Statistics in Medicine,* vol. 37, no. 24, pp. 3533-3546, 2018.

(44) J. Ferguson, A. Hannigan, and A. Stack, "A new computationally efficient algorithm for record linkage with field dependency and missing data imputation," *International Journal of Medical Informatics,* vol. 109, pp. 70-75, 2018/01/01/, 2018.

(45) A. Saeedi, M. Nentwig, E. Peukert, and E. Rahm, "Scalable Matching and Clustering of Entities with FAMER," *Complex Systems Informatics and Modeling Quarterly (CSIMQ),* vol. 95, no. 16, pp. 61-83, 2018.

(46) J. Zheng, Z. Lu, X. Zhao, and X. Li, "A Novel Holistic-based Entity Unifying Method for Heterogeneous Data," *Journal of Physics: Conference Series,* vol. 1302, pp. 042057, 2019/08, 2019.

(47) A. Saeedi, E. Peukert, and E. Rahm, "Using Link Features for Entity Clustering in Knowledge Graphs," in 15th International Conference, ESWC 2018 - The Semantic Web, Heraklion, Crete, Greece, 2018, pp. 576-592.

(48) S. Isaj, E. Zimányi, and T. B. Pedersen, "Multi-Source Spatial Entity Linkage," in Proceedings of the 16th International Symposium on Spatial and Temporal Databases, Vienna, Austria, 2019, pp. 1–10.

(49) D. Javdani, H. Rahmani, M. Allahgholi, and F. Karimkhani, "Deep Block: A Novel Blocking Approach for Entity Resolution using Deep Learning," in 5th IEEE International Conference on Web Research (ICWR2019), Tehran, Iran, 2019, pp. 41-44.

(50) P. Bellini, M. Benigni, R. Billero, P. Nesi, and N. Rauch, "Km4City ontology building vs data harvesting and cleaning for smart-city services," *Journal of Visual Languages & Computing,* vol. 25, no. 6, pp. 827-839, 2014/12/01/, 2014.

(51) S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-As-You-Go Entity Resolution," *IEEE Transactions on Knowledge and Data Engineering,* vol. 25, no. 5, pp. 1111-1124, 2013.

(52) K. Craig A., and S. Pedro, "Exploiting Semantics for Big Data Integration," *AI Magazine,* vol. 36, no. 1, pp. 25-38, 2015.
(53) R. C. Fernandez, P. Pietzuch, J. Kreps, N. Narkhede, J. Rao, J. Koshy, D. Lin, C. Riccomini, and G. Wang, "Liquid: Unifying Nearline and Offline Big Data Integration," in 7th Biennial Conference on Innovative Data Systems Research (CIDR '15), Asilomar, California, USA, 2015, pp. 1-8.
(54) G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced Data Management: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 9, pp. 2296-2319, 2016.
(55) H. Kardes, D. Konidena, S. Agrawal, M. Huff, and A. Sun, "Graph-based Approaches for Organization Entity Resolution in MapReduce," in TextGraphs-8 Graph-based Methods for Natural Language Processing, Seattle, Washington, USA, 2013, pp. 70–78.
(56) V. Efthymiou, K. Stefanidis, and V. Christophides, "Big data entity resolution: From highly to somehow similar entity descriptions in the Web," in 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 401-410.
(57) L. Zhu, M. Ghasemi-Gol, P. Szekely, A. Galstyan, and C. A. Knoblock, "Unsupervised Entity Resolution on Multi-type Graphs," in 15th International Semantic Web Conference - The Semantic Web, Kobe, Japan, 2016, pp. 649-667.
(58) A. Jurek, J. Hong, Y. Chi, and W. Liu, "A novel ensemble learning approach to unsupervised record linkage," *Information Systems,* vol. 71, pp. 40-54, 2017/11/01/, 2017.
(59) R. Brisebois, A. Nadembega, and T. Hajj, "Traceable and trusted smart harvesting algorithm from unstructured and structured web (smese-ttsha)," *International Journal of Current Research (IJCR),* vol. 11, no. 02, pp. 1050-1058, Feb. 2019, 2019.
(60) R. Brisebois, A. Nadembega, and T. Hajj, "Trusted smart harvesting algorithmbased on semantic relationship and social networks (smese-tsha)," *International Journal of Recent Scientific Research (IJRSR),* vol. 10, no. 01, pp. 30593-30604, January 2019, 2019.

****