ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 11, Issue, 02, pp.1050-1058, February, 2019

DOI: https://doi.org/10.24941/ijcr.34272.02.2019

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

TRACEABLE AND TRUSTED SMART HARVESTING ALGORITHM FROM UNSTRUCTURED AND STRUCTURED WEB (SMESE-TTSHA)

Ronald Brisebois, *Apollinaire Nadembega and Toufic Hajj

Im Media Technologies, Montréal, Canada

ARTICLE INFO

Article History:
Received 25th November, 2018
Received in revised form
18th December, 2018
Accepted 30th January, 2019
Published online 28th February, 2019

Key Words:

Block Chain, Entity Resolution, Metadata sources, Metadata Harvesting, Metadata Structure Detection, Machine Learning, Metadata Management, Entity Linking, Unstructured Web.

ABSTRACT

Entity Resolution and unstructured Web has recently attracted significant attentions and usage of Block Chain is increasing to try to solve different problems as traceability. Entity Resolution can be defined as the process of identifying, matching, verifying accuracy and try to get to the same metadata definition of the Entity from several sources including the unstructured web and structured databases. A new issue have been identified for Entity Resolution: What is information today could be wrong in an instant later, so we talk now of the value of the Entity Resolution in function of the time. In this paper, we address the issue of data and metadata timely integration from unstructured, structured and multi-sources. We propose a new semantic approach of data integration and Entity Resolution that aims to build a unified and trusted traceable repository (UTTR), called SMESE Traceable Trusted Smart Harvesting Algorithm from Unstructured and Structured Web (SMESE-TTSHA). SMESE-TTSHA is based on Traceable Smart Harvesting Strategies (TSHS) addresses the problem of performing Traceable Entity Resolution (MLM-TTSHA) using trusted, ranked sources and taking care of the value of the information at an instant *t*. We experimentally evaluate our SMESE-TTSHA approach on large real datasets and compare the performance results with those of existing approaches. Our experimental results show our proposed models perform well on the Traceable Entity Resolution compared to the existing approaches, while also satisfying constraint of the algorithm.

Corresponding author:

Citation: ***Ronald Brisebois, Apollinaire Nadembega and Toufic Hajj.*** **2019.** "Traceable and Trusted Smart Harvesting Algorithm from Unstructured and Structured Web (SMESE-TTSHA).", *International Journal of Current Research*, 11, (02), 1050-1058.

INTRODUCTION

Due to the open and decentralized nature of the Web, real world entities are usually described in multiple datasets using different URIs in a partial, overlapping and sometimes evolving way. Recognizing descriptions of the same real-world entities across, and sometimes within, data sources emerges as a central problem in the context of the Web of data. Addressing this problem, referred to as Entity Resolution (ER) (Steorts, 2015; Christen, 2013; Globerson *et al*., 2016; Nuray-Turan, 2013; Vesdapunt *et al*., 2014; Ramadan, 2014; Firmani *et al*., 2016; Whang, 2013; Papadakis *et al*., 2016) that is a prerequisite to various applications, namely, semantic search in terms of entities and their relations on top of the Web of text, interlinking entity descriptions in autonomous sources to strengthen the Web of data, and supporting deep reasoning using related ontologies to create the Web of knowledge.ER, resolving metadata and unstructured data is a long-standing challenge in database management, information retrieval, machine learning, natural language processing, and authority sources. ER is a common data cleaning task that involves determining which records from one or more data sets refer to the same real-world entities (Christen, 2013; Nuray-Turan, 2013; Fisher *et al*., 2015). ER is a well-known problem that has been extensively investigated in the past decades. Imagine that, given a very large collection of records from one or more data sets, how can we find records that actually refer to the same publication? To answer questions like this, we need to use entity resolution techniques. Additionally, some of the data sources might be noisier than the others and there might be different kind of typos that needs to be addressed. More, dynamic websites returns large number of entities from various sources when a user searches for a particular data in it. In the web of entities, entities are described by interlinked data and metadata rather than documents on the web. These web of entities keep undergoing dynamic changes and it becomes a very challenging task to visualize the relations between all these entities. Extraction of data from such sources becomes very tedious. It is needed to generate an optimized algorithm for the extraction of metadata and data from such sources and to make assessment on the quality and time value of the information. Data describing entities are made available in the Web under different formats (e.g., tabular, tree or graph) of varying structuredness. Traditional ER techniques, for instance, used for merging customer databases or library catalogues, are not suited for the Web of data, due to high heterogeneity (i.e., different properties are used to describe the

same kind of entity in different domains) and non-regularity in data structuring (i.e., even within the same domain, properties describing the same kind of entity significantly vary in terms of occurrences and types). Typically, an entity described in knowledge bases, such as Yago or Freebase, is declared to be instance of several semantic types, i.e., classes. One of the most popular approach for ER is crowd sourced. However existing techniques of Crowd sourced entity resolution either cannot achieve high quality or incur huge monetary costs. In addition, crowd sourced data management have three important problems:

- ***Quality Control:*** Workers may return noisy or incorrect results so effective techniques are required to achieve high quality;
- ***Cost Control:*** The crowd is not free, and cost control aims to reduce the monetary cost;
- ***Latency Control:*** The human workers can be slow, particularly compared to automated computing time scales, so latency control techniques are required (Li *et al*., 2016). The challenges in the management of the plethora of available linked datasets are due to the large quantities of linked data (volume), the dynamic aspects of data (velocity), the data originating from different domains and sources (variety), to assess and improve the accuracy of data (veracity), and an indication of the impact of data quality, both for decision making and monetary aspects (value) (Mountantonakis, 2018). The typical sources of big data are classified according to the way they are generated (Chen *et al*., 2013):

- ***User generated contents:*** From applications with massive users such as Tweeter, Facebook, Instagram, Blogs; messages, discussions, photos/videos posted are posted and shared by users. These data are directly contributed by users and are typically unstructured;
- ***Web data****:* They are crawled and processed to support applications such as Web search, mining and harvesting;
- ***Transactional data****:* They are generated by a large scale system due to massive operations/transactions processed by the system;
- ***Scientific data****:* They are collected from data-intensive experiments or applications;
- ***Graph data****:* they are formed by a very huge number of information nodes and the links between the nodes such as social networks and RDF knowledge bases.

To address the situation, we propose ahybrid human-machine approach for solving the problem of Traceable Entity Resolution (TER), called SMESE Timely Trusted Smart Harvesting Algorithm based on Semantic Relationship and Social Network (Smese-Ttsha). Smese-ttsha is a hybrid semantic approach of data integration and entity resolution that aims to build a unified, trusted and traceable repository (UTTR). The question is how the sources could be smart? SMESE-TTSHA is based on efficient traceable semantic harvesting strategies (TSHS) and machine learning model for repeatable entity resolution (MLM-TTSHA).TSHS addresses the problem of semantic harvesting based on authority file sources, sources classification model and the data graph model nodes exploration patterns while MLM-TTSHA addresses 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. SMESE-TTSHA characteristic are accurate of data/metadata, repeatable or traceabble, semantic enriched metadata and origin source based cataloging. SMESE-TTSHA allows to meet the challenges of Semantic cleaning process and Semantic watch process. SMESE-TTSHA is an extension of our previous works about SMESE (Brisebois *et al*., 2017; Brisebois, 2017), metadata enrichment (Brisebois *et al*., 2017; Brisebois *et al*., 2017; Brisebois *et al*., 2017), STELLAR (Brisebois *et al*., 2017; Brisebois *et al*., 2017; Brisebois, 2017; Brisebois *et al*., 2017) and Semantic Harvesting (Brisebois *et al*., 2017). The remainder of the paper is organized as follows. Section 2presentsthe related work. Section 3describes the algorithm (SMESE-TTSHA) and its various algorithms while Section 4 presents the evaluation through a number of simulations. Section 6 presents a summary and some suggestions for future work.

Related work: Data integration and record linkage (Mountantonakis, 2018; Chen, 2013; Jagadish, 2014; Dong, 2013; Philip, 2014; Chen *et al*., 2014; Hashem *et al*., 2015; Assunção, 2015; Kacfah Emani, 2015; Cai, 2015; Mountantonakis, 2018) and Entity Resolution (ER)(1-10, 12, 33-39) are two related research domains that aim to build a multi-catalog ecosystem (Brisebois *et al*., 2017; Brisebois *et al*., 2017) where the records are linked as a structured linked data ecosystem, web harvesting process (Brisebois *et al*., 2017; Vargiu, 2013; Teli *et al*., 2015; Shi *et al*., 2015; Haddaway, 2015; Kadam, 2014; Glez-Peña *et al*., 2014; Dastidar, 2016; Casali *et al*., 2016; Gupta, 2017) from different data sources, with their own unstructured data model, remains a challenge. Here, we focus only on Named Entity Resolution (NER); notice that our previous work (Brisebois, 2017) focus on metadata harvesting. Named Entity Resolution (NER) 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 traceable repository (UTTR). According to (Li *et al*., 2016), any important data management, such as NER, cannot be completely addressed by existing algorithms and automated processes; these tasks can be enhanced through the use of human cognitive ability.

Efthymiou *et al.* (2015) 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 is limited to the evaluation of a cluster of 15 machines using real data. Papadakis *et al.* (2016) proposed new meta-blocking methods that improve precision by up to an order of magnitude at a negligible cost to recall based on two combined techniques that reduce the overhead time of Meta-blocking by more than an order of magnitude. They introduced Block Filtering which intelligently removes entities from blocks (unnecessary entity) in order to shrink the blocking graph; the importance of a block for an individual entity is determined by the maximum number of blocks this entity participates in. Then, they accelerated the creation and the pruning of the blocking graph by minimizing the

computational cost for edge weighting using LeCoBI condition (Least Common Block Index).Authors do not clearly demonstrate how they crate the first blocking graph before applying the Block Filtering and Edge Weight Optimizing. Whang *et al.* (2013) 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. Authors approach addressed three important questions:

- How to construct the hints?
- How to use the hints? And
- In 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 are heuristics. Their work is proposed as representative cases and provided no formal guarantees.

Papadakis *et al*. (2013) proposed a novel framework that organizes existing blocking methods over highly heterogeneous information spaces (HHIS). Their framework comprises two layers where each targeting a different performance requirement. The effectiveness layer encompasses methods that create blocks in the context of HHIS, aiming at placing duplicate entities in at least one common block while the efficiency layer aims at processing blocks efficiently, discarding the repeated and unnecessary comparisons they contain. The goal of authors framework is to combine complementary blocking methods that can be easily tailored to the particular settings and requirements of each application. So authors did not proposed an overall approach for entity resolution. Jurek *et al.* (2017) proposed a new approach to unsupervised record linkage based on a combination of ensemble learning and enhanced automatic self-learning. Their approach incorporates ensemble learning and self-learning techniques into record 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 from the final ensemble. Authors just combine existing approaches that improve the existing automatic self-learning technique for RL.
As conclusion, we can claim the most of existing approaches are not based on time and are not traceable. We also understand that the best approach is one that uses at the least human contribution while achieving high accuracy. Finally, no approach shows how their TER is used to update the entity repository in a timely or traceable process.

SMESE Traceable Trusted Smart Harvesting Algorithm (SMESE-TTSHA) Models: In this section, we present the details of the proposed approach, called SMESE-TTSHA (SMESE Traceable Trusted Smart Harvesting Algorithm from unstructured and structured Web). First, we introduce the overview of SMESE-TTSHA and second, the details of SMESE-TTSHA algorithms and models. More specifically, we present (1) the SMESE-TTSHA architecture and relationship models between multi-sources entities that aims to show (i) the interoperability between SMESE-TTSHA components, (ii) the contribution of each component in the overall trusted architecture and (iii) the metadata harvesting strategies.

SMESE-TTSHA overview: Many aggregators harvest metadata and data that, in the process, may become inaccurate because they did not look at (1) the semantic context of the sources, (2) the reputation of the source, (3) neither to their timely accuracy and the usage of a meta-catalogue (master catalogue) (4)The timely accuracy is critical to ensure that we are talking of a trusted information. This timely accuracy specificity could be resolve by the usage of Block Chain Technics (BCT). The proposed SMESE ecosystem defines crosswalks that create metadata pathways to different sources of data and metadata; each pathway checks the structure of the metadata source and then performs data harvesting but keeping the timeline of the harvesting. For TTSHA, we enhance the classification of this model adding the traceability (time) of the harvesting.

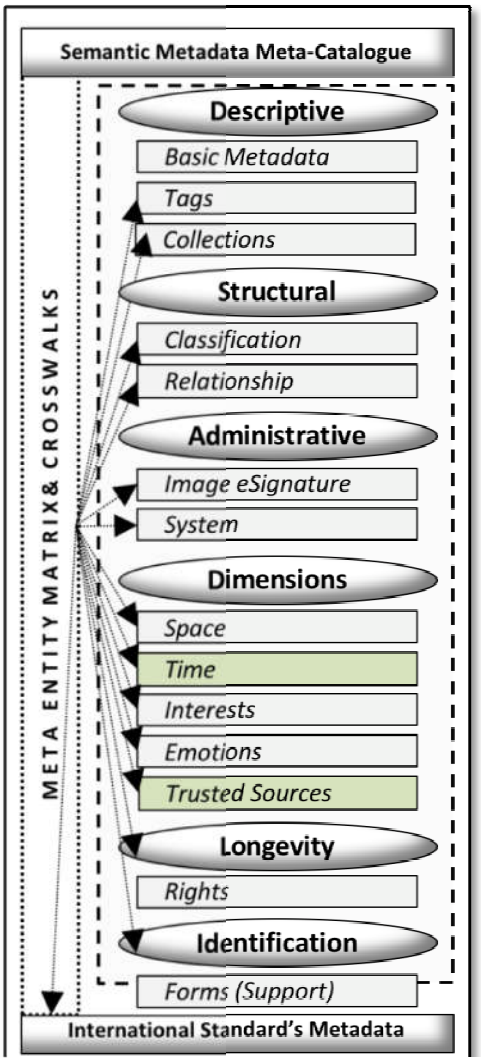

Fig. 1. Semantic metadata meta-catalogue enhanced classification

In the Fig. , we can see our model, named Timely Metadata Trusted Sources (TMTS) which is an extension of our previous model SMESE V3. TMTS represents the new model including the concept of timely knowledge experts and the concept of evolution of the metadata over time and relation with the evolution of the data as well. In the Fig. , we can see our previous model (top level) and new model (lower level) related to HAMD, this model from SMESE V3 evolves to a new

model. This new model simplified the involved authorities to make emphasis on the critical one in term of trusty and authority.

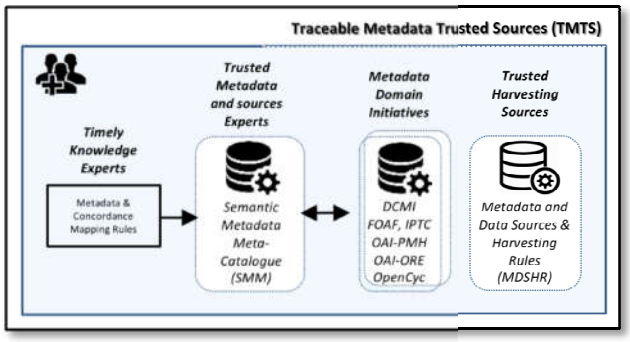

Fig. 2. Traceable Metadata Trusted Sources (TMTS)

This new model change mainly the order of the sources according to their trust of accuracy and ranking at a time ¨*t*¨.

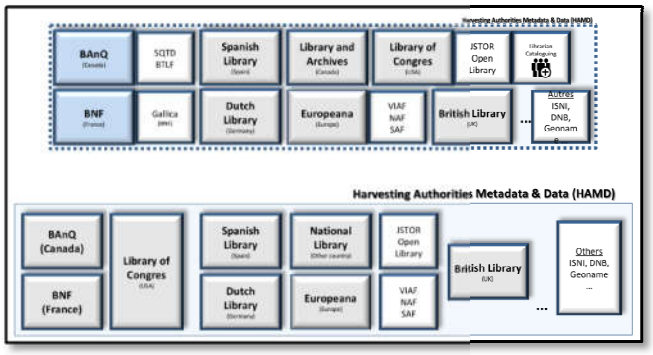

Fig. 3. Harvesting Authorities Metadata & Data (HAMD)

Criterion for identifying timely trusted sources (HAMD new version):

- Define by national or authority libraries;
- Define by international or authority associations;
- The authorities are ranked and timely.

SMESE-TTSHAAlgorithms

As mentioned above, SMESE-TTSHA consist of two main algorithms: TSHS (2) and MLM-TTSHA (2).

In this work, the semantic watching process will not be addressed; these axes of research will be addressed in the future works.

SHS is composed by:

- Experts-based sources analysis model;
- Semantic hierarchy strategy based harvesting algorithm.

MLM-TTSHA is composed by

- AI/MLM multi-sources metadata matching algorithm;
- MLM cleaning algorithm;
- AI/MLM enrichment algorithm.

In the following sections, we introduce, in details, TSHS (2) and then MLM-TTSHA (2).

Experts-based sources analysis model: Sources analysis process is a manual task that is performed by experts. The experts are selected according to the type of entities that SMESE-TTSHA needs to harvest; the experts are persons who work in the domain of expertise where the entities are classified and catalogued. For example, librarians are chosen as experts for the libraries and their associated entities such as books while the museologists are chosen as experts for the museums and their associated entities such as artworks. The main goal of experts is to identify relevant sources for specific entities and evaluate the metadata trust level of these sources; for example, for the names of museums in the world, VIAf (https://viaf.org/) is identified as the best while for the geolocation of the of museums in the world, Geonames (http://www.geonames.org/) is the best. For each source, we define a list of metadata to be completed or validated timely by experts:

Name; Website; Description; Source type; Contents type; Amount of contents; Harvesting technique; Harvesting strategy; Harvesting complexity level; Data format; Metadata structure; Source trust level per expert.

For example, *"source type"* may be: Museum, International Authority, National Library, National Research, knowledgeable source, International Association, Gallery, or City. And according to the source type metadata and their own experiences, experts may infer a source trust level that is used to compute sources *Accuracy Weight Factor* (*AWF*). As example (see Table), experts may define a source timely trust level (*STTL*) knowledge base according to the source type: For the metadata structure, experts need to identify the relationship entities of each sources. For example, Fig. shows the relationship between some authority file for three depth level metadata structure of sources while Fig. shows the relationship for four depth level. The depth level denotes the number of authorities file after the root authority file; in Fig., to reach "Freebase" after "VIAF", we have "Worldcat", "Wikidata" and then "Freebase" while in Fig. to reach "Freebase" after "ISNI", we have "VIAF", "Worldcat", "Wikidata" and then "Freebase". As mentioned, for each source, each expert needs to assign a STTL between 0 and 1. Then, based on the STTL, SMESE-TTSHA computes the AWF of each sources. To compute the AWF of sources, SMESE-TTSHA evaluates each expert weight. First, the universal weight is computed and the universal weight is used to compute geolocated-based weight. Universal weight of expert i is computed as follows:

$$W_i^U = \frac{\sum_{j=1}^{E-1} e_i^j}{E-1} \quad (1)$$

where denotes the evaluation of expert *i* by expert *j* while *E-1* denotes the total number of expert without expert *i*; indeed, the expert *i* does not evaluate himself and $1 \leq e_i^j \leq 100$. Based on the Universal weight, the geolocated-based weight of expert *i* is computed as follows:

$$W_i^L = \{^{W_i^G}_{100 \; if \; i \; resides \; in \; the \; locality \; L} \quad (2)$$

Then, based on the geolocated-based expert weight, SMESE-TTSHA computes the sources *Accuracy Weight Factor* (*AWF*) as follows:

$$AWF(s) = \frac{\sum_{i=1}^{E}(W_i^L \times STL(s,i))}{E} \quad (3)$$

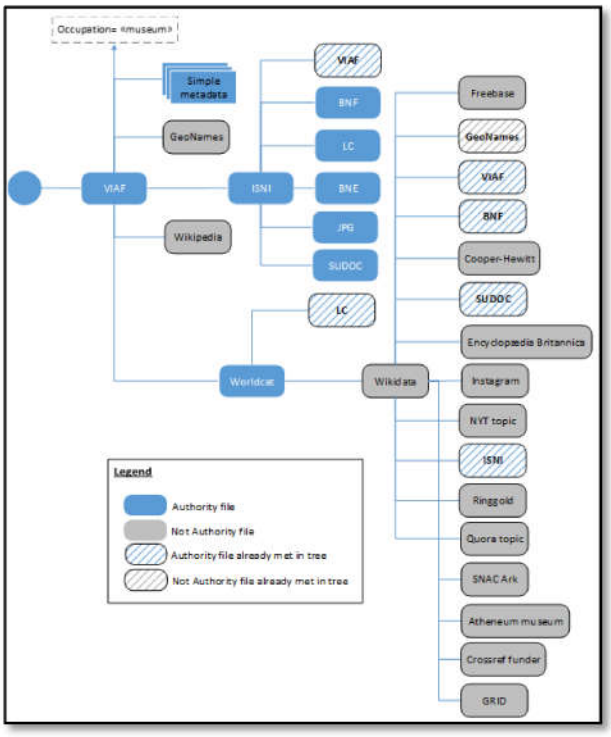

Fig. 4. Example of three depth level metadata structure of sources

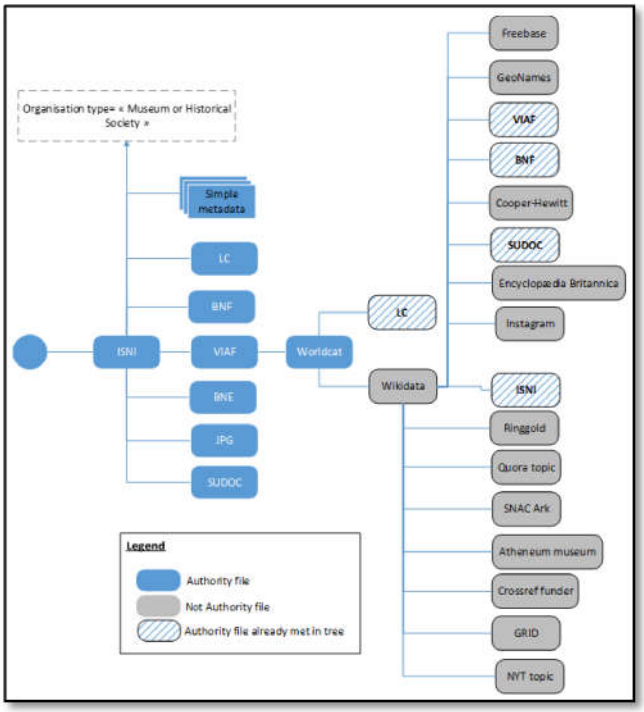

Fig. 5. Example of four depth level metadata structure of sources

Where *STL(s,i)* denotes the source trust level of source *s* assigned by expert *i*. In order to increase the accuracy of metadata, SMESE-TTSHA assigns the *mAWF* to each metadata of an entity based on the AWF of the source where the metadata is harvested. The following algorithm (see Table 2) presents the SMESE-TTSHA *mAWF* evaluation for a given metadata. The main contribution of *Mawf* evaluation algorithm is to update each metadata accuracy weight factor based on the sources having the biggest AWF. To also take into account the experts feedback, *mAWF(M)=100*when metadata *M* about entity *C* is validated by an expert that is affiliated to the entity C; for example, if a librarian of the Library of Congress validates the address of this library, mAWF (address) for the entity "Library of Congress" becomes 100. Notice that SMESE-TTSHA may suggest sourcesto the experts for analysis based on its knowledge database. Indeed, for a given source, SMESE-TTSHA compares the entities of this source with the entities of sources referred by experts. If the most important entities of sources referred by experts are found in the analysed source, SMESE-TTSHA infers that the analysed source must by analyse by experts. For example, if the k most famous entities of the world are catalogued on an analysed source and this source is not referred by experts, then this source is detected as source to suggest to experts.

Machine learning model for entity resolution (MLM-TTSHA): MLM-TTSHAgoal is to address the problem of entity resolution and entity enrichment. In details, MLM-TTSHA consists of automatic multi-sources metadata matching, cleaning and entity resolution and metadata enrichments using machine learning models and artificial intelligence algorithms. MLM for TTSHA algorithms try to predict trusted ranked sources of metadata. It uses the same model than SMESE but enhances the process to identify trusted ranked metadata sources in the structured environment and unstructured web and allow to keep all the history of the predicted and non-predicted trusted ranked sources of metadata. That's mean that a trusted source could become untrusted and the other way as well. The trust in regard to a source of metadata and data could change depending of many parameters Fig. 6 shows the MLM-TTSHA model. In the following sections, we present in details of SMESE-TTSHA.

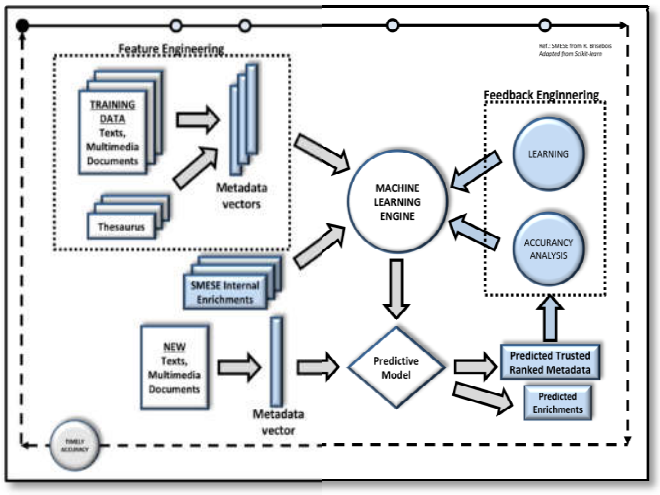

Fig. 6. MLM for Timely Trusted Smart Harvesting (TTSHA)

AI/MLM multi-sources metadata matching algorithm (*3MA):* Remember that the goal of SMESE-TTSHA is to harvest entities form various sources in order to build a unified, trusted and traceable repository (UTTR). In order to meet this goal, it is mandatory to identify each harvested metadata and match them with those which are listed in the same entity type on UTTR. Unfortunately, due to the large number of sources, this task becomes tedious. So, the role of AI/MLM multi-sources metadata matching algorithm is to perform this task automatically. Let $S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$ be the list of sources of same entities type *t* and *UTTR-t*= $\{m_1; m_2; m_3; \ldots; m_N\}$ be the list of metadata defined in the unified and trusted repository (UTTR) for entity *t*.Let $s_i = \{m_{i,1}; m_{i,2}; m_{1,3}; \ldots; m_{i,ki}\}$ be the list of metadata of entities type *t* in source s_i. Table 3 shows the AI/MLM multi-sources metadata matching algorithm.

Indeed, for each metadata of each source, 3MA performs a semantic similarity based on its function, called *Metadata Sem Sim()*. Table 4 shows the algorithm implemented by *Metadata Sem Sim* .Metadata Sem Sim is based on our function SEMD is proposed in our previous work (Brisebois *et al*., 2017). SEMD is computes the semantic distance between two terms in order to know if these terms are similar or not. The results of theAI/MLM multi-sources metadata matching algorithm is the matrix *M* of where for each metadata m_i of the UTTR-t, we identify the metadata vector $(m_{1,i}, \ldots, m_{j,i}, \ldots, m_{n,i})$. Equation (4) represents the matrix *M*.

$$M = \begin{pmatrix} m_1 \\ \ldots \\ m_i \\ \ldots \\ m_N \end{pmatrix} = \begin{pmatrix} m_{1,1} & \ldots & m_{n,1} \\ \ldots & \ldots & \ldots \\ m_{1,k1} & \ldots & m_{n,kn} \end{pmatrix} \quad (4)$$

Evaluation using simulations: In this section we present the experimental evaluation of our proposed approach, called SMESE-TTSHA. The objective of our experimental evaluation is to compare, according to the literature, more recent and performing algorithms on various types of entities. We use all entities for the testing and comparing set for all datasets.

Simulation Setup and Datasets Characteristics: To measure SMESE-TTSHA performance, a simulator program has been developed using Java code. The server characteristics for the simulations were: Dell Inc. Power Edge R630 with 96 Ghz (4 x Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz, 10 core and 20 threads per CPU) and 256 GB memory running VMWare ESXi 6.0. The Datasets we use was provided by forty-three (45) data sources of various types such as Government Departments (culture and tourism), National Library, International Authority, Notional associations, National Research, knowledgeable source, National Authority, Association Culturelle, Museum and Gallery, some of the data sources are Virtual International Authority File (VIAF), International Standard Name Identifier (ISNI), Bibliothèquenationale de France (BNF), Bibliothèque et Archives Nationales du Québec (BANQ), Système Universitaire de Documentation (SUDOC), Getty Union List of Artist Names (JPG), Wikidata and Wikipedia. The overall datasets contains millions of entities and each entity contains metadata including the title, country, city, artist, address, latitude, longitude, and type of entity. The datasets consist of four (4) types of real entities: Museum, Place, Artwork and Artist. Table 5 shows each dataset entities types and their count.

Performance measurement criteria: As the quality of the results, i.e. the performance of the algorithm can be determined in terms of the metrics used to evaluate the entity resolution. As in (Efthymiou *et al*., 2015; Hakan Kardes, 2013; Zhu *et al*., 2016), thesame performance metrics can be used for comparison: *Precision and Recall*.For example, accuracy is related to rate of true entity resolution (entities detected as duplicate and those detected as non-duplicate). *Recall*measures what fraction of the known matches are candidate matches while *Precision* measures what fraction of the candidate matches are known matches. True Positive (TP) denotes the case when a pair of entities is detected by a scheme as the same entity and whose the experts mention that it's the same entity. False Positive (FP) denotes the case when a pair of entities is detected by a scheme as the same entity and whose the experts mention that it is not the same entity. False Negative (FN) denotes the case when a pair of entities is detected by a scheme as not the same entity and whose the experts mention that it's the same entity. All remaining pairs of entities are considered to be True Negatives (TN). The metrics can be described in terms of this definitions, as seen in the following equations:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

We also evaluate the scalability of the proposed approach in terms of *running time* (Mountantonakis, 2018; Zhu *et al*., 2016) we report how running time varies with the size of data to evaluate the scalability.

RESULTS AND DISCUSSION

Simulation results are averaged over multiple runs; indeed, the simulation program is run more than 50 times; one run of the simulation program provides ten prediction units; a prediction unit contains a destination and the path toward this destination.

Table 1. Sources timely trust level knowledge base

Source Type	STTL
Recognized international/national authority	1
Recognized national association/researcher/derivative authority	0.95
Very credible sources	0.90
Credible sources	0.85
Knowledgeable sources	0.80
Others sources	0.65

Table 2. SMESE-TTSHAmAWF evaluation algorithm

Pseudo code: *SMESE-TTSHAmAWF evaluation algorithm*
For metadata *M* harvested from a source *s*
IF M is a Geo Location metadata and Source s is GeoNames
mAWF(M)=100
ELSE
IF mAWF(M) is null
mAWF(M) = AWF(s)
ELSE
IFmAWF(M) ≠ 100
IFAWF(s) = 100,
mAWF(M) = 100
ELSE
mAWF(M)= Biggest between {mAWF(M), AWF(s)}

Table 3. AI/MLM multi-sources metadata matching algorithm

Pseudo code: *Multi-sources Metadata Matching Algorithm (3MA)*
1. For each source s_i; i=1 to n
2. For j=1 to ki
3. For each metadata of UTTR-s; q=1 to N
4. Val($m_{i,j}$) = *MetadataSemSim(m_q; $m_{i,j}$)*
5. MATCH $m_{i,j}$ with m_q where Val($m_{i,j}$) is the biggest
6. RETURN*M*

For each run, we compute each performance criteria using their equation, respectively; thus, to obtain the simulation results shown in Figs. 7 and 9, we compute the average of the 50 runs. The overall datasets were divided into 10 subsets with IDs assigned to each of them. In Fig. 7 and 8, the average precision and average recall varying with the datasets ID while in Fig. 9 the average running time varying with the weeks rang. Fig. 7 shows the average precision when varying the Datasets ID.

We observe that SMESE-TTSHA outperforms ER1, ER2 and ER3; for example, SMESE-TTSHA provides an average

Table 4. MetadataSemSim function

Pseudo code: *Function MetadataSemSim (A,B)*
1. $\leftrightarrow_{AB}$= *SEMDis* between A and B
2. IF $\leftrightarrow_{AB}$ between 0.90 and 1.0
3. RETURN1
4. ELSE
5. IF $\leftrightarrow_{AB}$ between 0.8 and 0.9
6. IF value type of A is same than type of B
7. RETURN0.95
8. ELSE
9. RETURN $\leftrightarrow_{AB}$
10. ELSE
11. RETURN 0
12. SENDMatching request to experts

Table 5. Evaluation datasets entities types

Entities Type	Number of entities
Museums and Galleries	84,069
Artworks	12,034,926
Places (Countries, States and Cities)	3,133,006
Artists	1,519,354

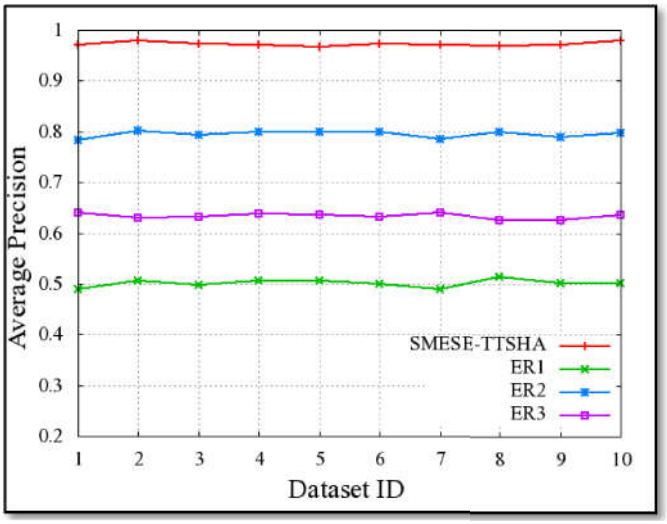

Fig.7. Precision VS Dataset ID

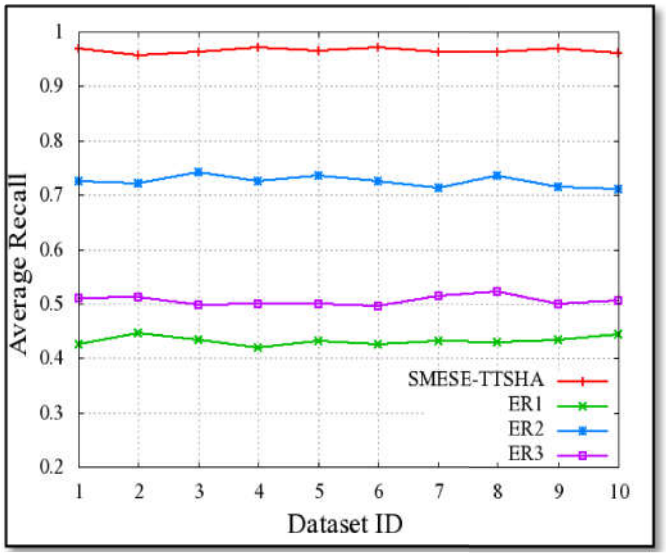

Fig. 8. Recall VS Dataset ID

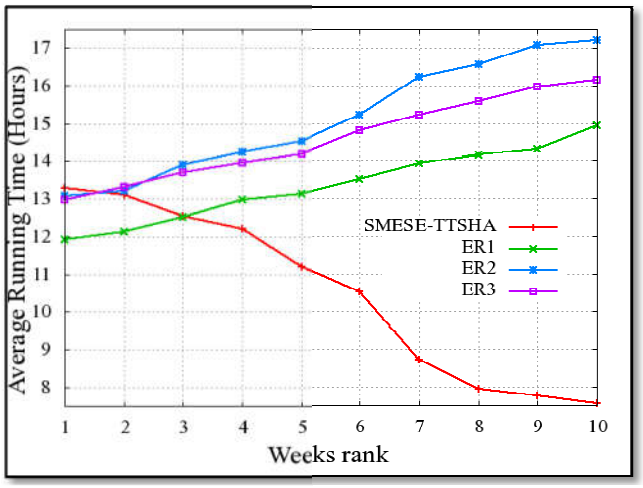

Fig. 9. Running time VS Weeks rank

precision of 0.97 per Dataset, whereas ER2 (more efficient than ER1 and ER3 in this scenario) provides an average of 0.79 per Dataset; overall, the average relative improvement (defined as (average precision of SMESE-TTSHA — average precision of ER2)) of SMESE-TTSHA compared with ER2 (resp. ER1 and ER3) is about 18% (resp. 48% and 34%) per Dataset. This can be explained by the fact that SMESE-TTSHA uses a hierarchy of trusted sources for entity harvesting process. Fig. 8shows that SMESE-TTSHA outperforms ER1, ER2 and ER3; SMESE-TTSHA provides an average precision of 0.96 per Dataset, whereas ER2 (more efficient than ER1 and ER3 in this scenario) provides an average of 0.72 per Dataset. The average relative improvement (defined as (average recall of SMESE-TTSHA — average recall of ER2)) of SMESE-TTSHA compared with ER2 (resp. ER1 and ER3) is about 24% (resp. 53% and 46%) per Dataset. This is mainly due to the fact that SMESE-TTSHA detects well the true negative candidates in contrast to ER1, ER2 and ER3. Fig. 9 shows the average running time when varying the weeks rang. In this scenario, the simulations are run ten times during ten weeks; indeed, at the beginning of each week, the simulation is run in order to take into account the new entities added by the sources. We observe that for SMESE-TTSHA, the average running time decreases with the weeks rang while for ER1, ER2 and ER3, the average running time increases with the weeks rang; this can be explained by the fact that SMESE-TTSHA applies a repeatable smart approach which only treats the news entities added by the sources in contrast to ER1, ER2 and ER3 which re-treat all the entities. For example, at the 10^{th} time of running, SMESE-TTSHA requires about 8.7 hours while ER1 requires about 15 hours. However, we observe that at the first time of running, ER1 outperforms SMESE-TTSHA, ER2 and ER3. The average relative improvement (defined as (average recall of ER1— average recall of SMESE-TTSHA)) of ER1compared with SMESE-TTSHA (resp. ER2 and ER3) is about 1.38 hours (resp. 1.18 hours and 1.07 hours) per Dataset. In summary, the analysis of the simulation results shows that schemes that use human annotation combine to machine learning model (MLM) outperform schemes that are limited to human annotation or machine learning model. We also observe that schemes that use machine learning model (MLM) outperform schemes that are limited to human annotation. However, schemes that use machine learning model (MLM) require more running time than those which use machine learning model (MLM).Finally, repeatable approach requires less running time with the number of times the process is run.

Summary and future work

We have shown that it is possible and more accurate to harvest and trace (traceability of the results) metadata and data using trusted sources instead of to harvest just sources of metadata and data. So, a source will have a ranking of accuracy (trust) and a period of time of validity of this ranked trust. As an example, it better to be able to harvest all the museum of the world to use timely a number of trusted and ranked sources of metadata and data than just to harvest the web or some databases without any guidance about their relevancy and accuracy. That means that a trusted ranked sources of metadata as a life time and may change rapidly over time. The meta-catalogue that we built in SMESE project as to include the list of trusted sources of metadata related to a type of object and trusted thesaurus and their traceability. Yet, there is room for improvement if we look to build application to structure the unstructured web. Here are some of the future work that we looking to explore:

- AI/MLM sources automatic analysis to identify trusted ranked and the traceability of sources of metadata and data;
- The verification/validation process of the trusted sources of metadata or how to validate automatically the ranking and traceability of a metadata source.

REFERENCES

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

Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "An Assisted Literature Review using Machine Learning Models to Identify and Build a Literature Corpus," *International Journal of Engineering and Science Invention (IJESI),* vol. 6, no. 7, pp. 72-84.

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.

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.

Brisebois, R., Abran, A. and Nadembega, A. 2017. "A Semantic Metadata Enrichment Software Ecosystem based on Metadata and Affinity Models," *International Journal of Information Technology and Computer Science (IJITCS),* vol. 9, no. 8, pp. 1-13, August 2017.

Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "A Semantic Metadata Enrichment Software Ecosystem based on Machine Learning to Analyse Topic, Sentiment and Emotions," *International Journal of Recent Scientific Research (IJRSR),* vol. 8, no. 4, pp. 16698-16714, April.

Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "A Semantic Metadata Enrichment Software Ecosystem based on Topic Metadata Enrichments," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 7, no. 3, pp. 1-23, May 2017.

Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "An Assisted Literature Review using Machine Learning Models to Recommend a Relevant Reference Papers List," *International Scientific Research Organization Journal,* vol. 02, no. 02, pp. 1-24, November 2017.

Brisebois, R., Abran, A., Nadembega, A. and N'techobo, P. 2017. "Text and Data Mining & Machine Learning Models to Build an Assisted Literature Review with Relevant Papers," *International Journal of Scientific Research in Information Systems and Engineering (IJSRISE),* vol. 03, no. 01, pp. 6-27, April 2017.

Brisebois, R., Abran, A., Nadembega, A., and N'techobo, P. 2017. "A Semantic Metadata Enrichment Software Ecosystem based on Sentiment and Emotion Metadata Enrichments," *International Journal of Scientific Research in Science Engineering and Technology (IJSRSET),* vol. 03, no. 02, pp. 625-641, March-April 2017.

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.

Cai, L. and Zhu, Y. 2015. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal,* vol. 14, no. 2.

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

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

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

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

Christen, P. and Gayler, R. W. "Adaptive Temporal Entity Resolution on Dynamic Databases," *Advances in Knowledge Discovery and Data Mining.* pp. 558-569.

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

Dong, X. L. and Srivastava, D. "Big data integration." pp. 1245-1248.

Efthymiou, V., Stefanidis, K. and Christophides, V. "Big data entity resolution: From highly to somehow similar entity descriptions in the Web." pp. 401-410.

Elmagarmid, A., Ilyas, I. F., Ouzzani, M., Quian, J.A. 2014. #233, -Ruiz, N. Tang, and S. Yin, "NADEEF/ER: generic and interactive entity resolution," in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, Utah, USA, 2014, pp. 1071-1074.

Firmani, D., Saha, B. and Srivastava, D. 2016. "Online entity resolution using an Oracle," *Proc. VLDB Endow.,* vol. 9, no. 5, pp. 384-395.

Fisher, J., Christen, P., Wang, Q. and Rahm, E.2015. "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, pp. 279-288.

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

Globerson, A., Lazic, N., Chakrabarti, S., Subramanya, A., Ringgaard, M. and Pereira, F. "Collective Entity Resolution with Multi-Focal Attention." pp. 621–631.

Gupta, G. and Chhabra, I. 2017. "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.

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

Hakan Kardes, Deepak Konidena, Siddharth Agrawal, Micah Huff, and A. Sun, "Graph-based Approaches for Organization Entity Resolution in MapReduce."

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

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R. and Shahabi, C. 2014. "Big data and its technical challenges," *Commun. ACM,* vol. 57, no. 7, pp. 86-94.

Jurek, A., Hong, J., Chi, Y. and Liu, W. 2017. "A novel ensemble learning approach to unsupervised record linkage," *Information Systems,* vol. 71, pp. 40-54, 2017/11/01/.

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

Kadam, V. B. and Pakle, G. K. 2014. "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.

Li, G., Wang, J., Zheng, Y. and Franklin, M. J. 2016. "Crowdsourced Data Management: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 9, pp. 2296-2319.

Mountantonakis, M. and Tzitzikas, Y. 2018. "High Performance Methods for Linked Open Data Connectivity Analytics," *Information,* vol. 9, no. 134, pp. 1-33, 3 June 2018.

Mountantonakis, M. and Tzitzikas, Y. 2018. "Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets," *J. Data and Information Quality,* vol. 9, no. 3, pp. 1-49.

Nuray-Turan, R., Kalashnikov, D. V. and Mehrotra, S. 2013. "Adaptive Connection Strength Models for Relationship-Based Entity Resolution," *J. Data and Information Quality,* vol. 4, no. 2, pp. 1-22.

Papadakis, G., Ioannou, E., Palpanas, T., Niederée, C. and Nejdl, W. 2013. "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.

Papadakis, G., Papastefanatos, G., Palpanas, T. and Koubarakis, M. "Scaling Entity Resolution to Large, Heterogeneous Data with Enhanced Meta-blocking." pp. 1-12.

Passos, A., Kumar, V. and McCallum, A. 2014."Lexicon Infused Phrase Embeddings for Named Entity Resolution," *corr,* vol. abs/1404.5367.

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

Ramadan, B., and Christen, P. 2014. "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, pp. 1787-1790.

Ramadan, B., Christen, P., Liang, H. and Gayler, R. W. 2015. "Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution," *J. Data and Information Quality,* vol. 6, no. 4, pp. 1-29.

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

Simonini, G., Bergamaschi, S. and Jagadish, H. V. 2016. "BLAST: a loosely schema-aware meta-blocking approach for entity resolution," *Proc. VLDB Endow.,* vol. 9, no. 12, pp. 1173-1184.

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

Teli, S. 2015. "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.

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.

Vesdapunt, N., Bellare, K. and Dalvi, N. 2014. "Crowdsourcing algorithms for entity resolution," *Proc. VLDB Endow.,* vol. 7, no. 12, pp. 1071-1082.

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

Williams, T. and Scheutz, M. "POWER: A domain-independent algorithm for Probabilistic, Open-World Entity Resolution." pp. 1230-1235.

Zhu, L., Ghasemi-Gol, M., Szekely, P., Galstyan, A. and Knoblock, C. A. 2016. "Unsupervised Entity Resolution on Multi-type Graphs," *The Semantic Web – ISWC.* pp. 649-667.

****