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

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 10, Issue, 01(F), pp. 30593-30604, January, 2019

ISSN: 0976-3031

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

TRUSTED SMART HARVESTING ALGORITHMBASED ON SEMANTIC RELATIONSHIP AND SOCIAL NETWORKS (SMESE-TSHA)

Ronald Brisebois, Apollinaire Nadembega* and Toufic Hajj

InMedia Technologies, Montréal, Canada

DOI: http://dx.doi.org/10.24327/ijrsr.2019.1001.3088

ARTICLE INFO

Article History:

Received 12th October, 2018
Received in revised form 23rd November, 2018
Accepted 7th December, 2018
Published online 28th January, 2019

Key Words:

Entity resolution, metadatasources, metadata harvesting, social network, machine learning, meta-catalogue, metadata management, ontology alignment.

ABSTRACT

Crowd sourced and Entity Resolution has recently attracted significant attentions because it can harness the wisdom of crowd to improve the quality of Entity Resolution. Entity Resolution can be defined as the process of identifying, matching, verifying accuracy and merging metadata that correspond to the same entities from several databases. Two main issues have been identified for crowd sourced Entity Resolution: data, relation harvesting and integration, and named Entity Resolution. In this paper, we address the issue of data and metadata integration from multi-sources. We propose a new semantic approach of data integration, called SMESE Trusted Smart Harvesting Algorithm based on Semantic Relationship and Social Network (SMESE-TSHA). SMESE-TSHA is based on efficient Semantic Harvesting Strategies (SHS)addresses the problem of performing Entity Resolution (MLM-TSHA) using trusted and ranked sources.SHS addresses the problem of semantic harvesting based on authority file sources, sources classification model and the data graph model nodes exploration patterns while MLM-TSHA addresses the problem of performing Entity Resolution on RDF graphs containing multiple types of nodes. We experimentally evaluate our SMES-TSHA approach on large real datasets and compare the performance results with existing approaches. Our experimental results show our proposed models perform well on the Entity Resolution compared to the existing approaches, while also satisfying the running time restrictions.

INTRODUCTION

Massive amounts of data are nowadays being collected by most business and government organizations. Given that many of these organizations rely on information in their day-to-day operations, the quality of the collected data has a direct impact on the quality of the produced outcomes. Various data cleaning practices are employed to improve the collected data. One important practice in data cleaning is the task of identifying all records that refer to the same real-world entity. This process is called Entity Resolution (ER)[1-9] and can be applied to a single or to multiple data sources (within a single data source the process is called de-duplication); 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[2, 4, 10];ER is a well-known problem that has been extensively investigatedin 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. 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.

The management of the plethora of available linked datasets poses various challenges. There is a need for methods that can deal with large quantities of linked data (volume), to accommodate the dynamic aspects of data (velocity), to be able to uniformly deal with data originating from different domains and sources (variety), to assess and improve the accuracy of data (veracity), and to provide an indication of the impact of data quality, both for decision making and monetary aspects (value)[11]; these characteristics refer to the four dimensions related to big data[12-20]:

- Volume refers to the problem of how to deal with very large data sets, which typically requires execution in a distributed cloud-based infrastructure; data sizes will range from terabytes to zettabytes (that is, 10^{21} bytes).
- Variety refers to dealing with different types of sources, different formats of the data, and large numbers of sources. Much of the work on big data has focused on

Corresponding author:* **Apollinaire Nadembega
InMedia Technologies, Montréal, Canada

volume and velocity, but the problems of variety are equally important in solving many realworld problems[21];data comes in many different formats from structured data, organized according to some structures like the data record, to unstructured data, like image, sounds, and videos which are much more difficult to search and analyze.

- Velocity refers to dealing with real-time streaming data, such as video feeds, where it may be impossible to store all data for later processing;in many novel applications, data continuously arrives at possible very high frequencies, resulting in continuous high-speed data streams.
- Huge number of data sources – the real value of data sets is when these data sets are integrated and cross-correlated. Integration and cross-correlation among data sets from different sources allow one to uncover information and trends that often cannot be uncovered by looking at a data set in isolation.

Based on the description of ER and big data, the focus of this work is defined as the big data integration in order to build unified and qualified entities repository [14]. Big data integration consists to: (1) ***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 works proposed a model, call SMESE [22, 23]. (2) ***Record linkage***: it refers to the task of identifying records that refer to the same logical entity across different data sources, especially when they may or may not share a common identifier across the data sources.Record linkage (RL), also referred to as data matching or entity resolution, is a process of finding records that correspond to the same entity from one or more data source[24]. (3) ***Data fusion***: it refers to resolving conflicts from different sources.

Indeed, due to the open and decentralized nature of the Web, realworld 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 ER 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. Data describing entities are made available in the Web under different formats (e.g., tabular, tree or graph) of varyingstructuredness. Typically, an entity described in knowledge bases, such as Yago or Freebase, is declared to be instance of several semantic types, i.e., classes. The description of such an entity may employ properties from different vocabularies, resulting in quite different structural types even for descriptions of same type, e.g., personor place. 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:(1) Quality Control: Workers may return noisy or incorrect results so effective techniques are required to achieve high quality; (2) Cost Control: The crowd is not free, and cost control aims to reduce the monetary cost; (3) Latency Control: The human workers can be slow, particularly compared to automated computing time scales, so latencycontrol techniques are required[25].

To address the aforementioned challenges,we propose a hybrid human-machine approach for solving the problem of Entity Resolution (ER), called SMESE Trusted Smart Harvesting Algorithm based on Semantic Relationship and Social Network (SMESE-TSHA). SMESE-TSHA is a hybrid semantic approach of data integration and entity resolution that aims to build a unified, qualified and trusted repository (UTR). The question is how the sources could be smart? SMESE-TSHA is based on efficient semantic harvesting strategies (SHS) and machine learning model for entity resolution (MLM-TSHA).SHS addresses the problem of semantic harvesting based on authority file sources, sources classification model and the data graph model nodes exploration patterns while MLM-TSHA 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-TSHA characteristic are accurate of data/metadata, semantic enriched metadata and origin source based cataloging. SMESE-TSHA allows to meet the challenges of Semantic cleaning process (de-duplicate/merge/purge/ linking) and Semantic watch process (audit trail). SMESE-TSHA is an extension of our previous works about SMESE[22, 23], metadata enrichment [26-28], STELLAR [29-32] and Semantic Harvesting [33].The remainder of the paper is organized as follows. Section 2presentsthe related work. Section 3describes the algorithm(SMESE-TSHA) and its various algorithms while Section 4 presents the evaluation. Section 6 presents a summary and future work.

Related work

In order to build multi-catalog ecosystem[22, 23] where the records are linked asa structured linked data ecosystem,web harvesting process[33-42]from different data sources, with their own unstructured data model, remains a challenge. To present the related works, we focus on two research axes about semantic metadata harvesting from several sources:Data integration and record linkage [11-20, 43] and Name Entity Resolution (NER)[1-11, 44-50].

Data Integration (DI) and Record Linkage (RL)

The 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”[19]. 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.

Volume refers to the amount of all types of data generated from different sources and continue to expand while ***variety*** refers to the different types of data (e.g., video, image, text, audio, and data logs, in either structured or unstructured format) collected via sensors, smartphones, or social networks. ***Velocity*** refers to the speed of data transfer while ***value***, that is the he most important aspect of big data according to [17], refers to the process of discovering huge hidden values from large datasets with various types and rapid generation. ***Veracity*** refers to what is conform with truth or fact; in other words, Accuracy, Certainty, Precision; uncertainty can be caused by inconsistencies, model approximations, ambiguities, fraud, duplication, incompleteness, spam and latency. In this work, veracity is the key issues addressing.

Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis [17]; 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 (web, DBMS/OLTP, NoSQL, HDFS) and has to deal with diverse access protocols; For organization, we have to deal with various data formats (texts formats, compressed files, variously delimited, etc.) and must be able to parse them and extract the actual information like named entities, relation between them. Also, we have to be clean, put in a computable mode, structured or semi-structured, integrated and stored in the right location (existing data warehouse, data marts, Operational Data Store, Complex Event Processing engine, NoSQL database). Successful cleaning in Big Data architecture is not entirely guaranteed; in fact "the volume, velocity, variety, and variability of Big Data may preclude us from taking the time to cleanse it all thoroughly".

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 Named Entity Resolution (NER) [1-9]. However, before NER process, the data Integration (DI) and Record Linkage (RL) are the first challenges of data management in the context of big data and clouds computing.

Bellini*et al.*[24] proposed a system for data ingestion and reconciliation of smart cities 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 (Knowledge Model for City). Unfortunately, their KM4City (proposed knowledge model for Smart City) are limited to seven areas. In addition, they did not take into account the data generated by citizens. Finally, authors did not proposed their own data integration model, they used an existing model, call Pentaho Kettle formalism.

Knoblock and Szekely[21] 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 (that is, XML or JSON). They implemented their approach to using semantics for big data integration in a system called Karma. Karma allows a user to (1) import data from a wide variety of sources, (2) clean and normalize the data, (3) quickly build a model or semantic description of each source, and (4) 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 of the main limitations of their approach is the fact that the data comes from the already structured database; which is not the case most of the time. Karma is limited to find noisy, missing, or inconsistent data; unfortunately, we may conclude that Karma does not be useful for entry resolution.

Raul Castro *et al.*[51] proposed a data integration stack that provides low latency data access to support near real-time in addition to batch applications, called Liquid. According to authors, Liquid consists of two cooperating layers: 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 entities resolution.

Tong *et al.*[52] proposed a novel data cleaning platform for cleaning multi-version data on the Web, called Crowd Cleaner, via crowd sourcing approaches. Crowd Cleanerutilizes crowd sourcing-based approaches for detecting and repairing errors that usually cannot be solved by traditional data integration and cleaning techniques. According to authors, Crowd Cleaner does not only detect and repair false or delay versions of updates but also automatically determines which version of data should be accepted. Unfortunately, authors do not demonstrate the performance of their Crowd Cleaner.

Name Entity Resolution (NER)

Recently big data becomes the major challenge for data integration and cleaning. The research interests in the field of data integration and cleaning are[12]:

- ***User feedback and crowd sourcing***: Traditionally, the errors in schema mappings are expected to be fixed by domain experts. They propose a method to determine the order to confirm user feedbacks by evaluating the utilities of candidate matches.
- ***Uncertainty and provenance***: For this issue, the probabilistic model would be constructed to represent the data uncertainty and to make imprecise decisions.
- ***Pay-as-you-go***: This approach allows constructing an imperfect system which could provide necessary service, and to incrementally improve this system when there are more resources available like time and money.
- ***Entity matching and resolution(NER)***: The objective 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, iterative approach, and using functional dependencies.

Indeed, 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 repository (UTR). According to [25], 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. According to the literature review, crowd sourcing is an effective way to address such tasks by utilizing hundreds of thousands of workers (i.e., the crowd). Crowd sourcing allows solving computer-hard tasks and is benefit data management, such as data cleaning, data integration and knowledge construction. Thus, crowdsourced data management has become an area of increasing interest in research and experimentation. Unfortunately, quality control remains one of the main problems.

Vesdapunt *et al.* [5] 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 about 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 entities resolution and are not trusted; this task is different to user feedback to enrich a machine learning model. The accuracy of their approach is strongly linked to the quality of the crowd responses.

Kardes *et al.*[53]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 their ER approach will be implemented without person profiles?

Zhu*et al.*[54] 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 modelled 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.Authors approach is based on a metadata that have all the entities such as the manufacturer of product or authors of papers. In real life and in the context of Web Big Data, this case is very rare and cannot be applied to any domain. In addition, as [53], their approach is strongly linked to a specific metadata; What happens if this metadata is empty?

Whang *et al.*[55]studied the problem of resolving records with crowd sourcing where they asked questions to humans in order to guide ER into producing accurate results. Authors proposed algorithms that determine what pairs of records should be compared by the crowd. As Vesdapunt *et al.* [5], they applied transitive closure to reduce the number of questions to ask crowd workers. However, in contrast to [5], Whang *et al.*[55] used the humans for image similarity detection in their ER algorithm. According to authors, they used humans during the ER process itself; in contrast to those which used humans in an earlier trainingor in a later verification phase.

Mountantonakis and Tzitzikas[11]introduced methods for assessing the connectivity of large numbers of linked datasets, even if they come from different domains and sources, for assessing their connectivity and for providing value-added services such as global lookup services. They performed measurements related to the connectivity of more than two datasets, including provenance information. In other word, authors proposed a same As catalog for computing the symmetric and transitive closure of the owl: same As relationships encountered in the datasets. To construct their Same As catalog, authors constructed incrementally chains of owl: same As URIs where each URI becomes a member of a chain if and only if there is anowl: same As relationship with a URI that is already member of this chain.Their approach is only based on the transitive closure of "owl: same As" relationship; this method is the best for ER, but it is only apply when the entities have explicit owl: same As URIs.

Chai *et al.*[44] proposed a cost-effective crowd sourced entity resolution framework, called Partial-Order based cro Wdsourced Entity Resolution (POWER).The basic idea is that they defined a partial order on the record pairs and pruned many pairs that do not need to be asked based on the partial order. Specifically, they selected a pair as a question and ask the crowd to check whether the records in the pair refer to the same entity. After getting the answer of this pair, they inferred the answers of other pairs based on the partial order. Authors approach has the same limitation of the crowd sourced entity resolution technique.

As conclusion, we can claim the most of existing approaches are based on the crowd sourced whose the key components are the workers who answer about the similarity between pair entities. We also understand that the best approach is one that uses at the least human contribution while achieving high accuracy.

SMESE Trusted Smart Harvesting Algorithm (SMESE-TSHA)

In this section, we present the details of the proposed approach, called SMESE-TSHA who is based on Semantic Relationship and Social Network. First, we introduce SMESE-TSHA and second, the details of SMESE-TSHA algorithms and models. More specifically, we present (1) the SMESE-TSHA architecture and relationship models between multi-sources entities that aims to show (i) the interoperability between SMESE-TSHA components, (ii) the contribution of each component in the overall trusted architecture and (iii) the metadata harvesting strategies, (2) the efficient semantic harvesting strategies (SHS) that aims to perform a semantic harvesting based on authority trusted sources, sources classification model and the data graph model nodes exploration patterns, and (3) the machine learning model for ER (MLM-TSHA) that aims to performentity resolution (ER) on RDF graphs containing multiple types of nodes, using the links between instances of different types to improve accuracy and repeatability.

Fig 1 shows SMESE-TSHA prototype applied to unstructured Web and Museums.

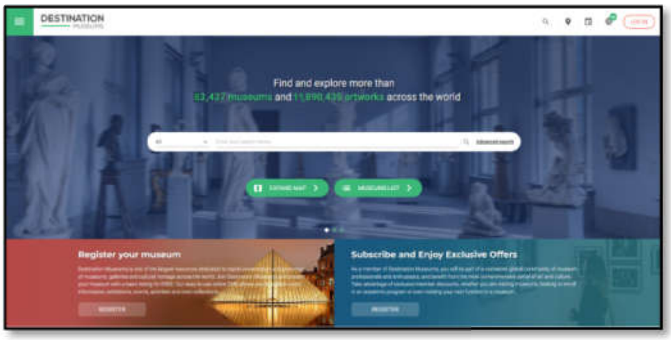
DESTINATION MUSEUMS
Find and explore more than
63,437 museums and 11,890,439 artworks across the world
Register your museum
Subscribe and Enjoy Exclusive Offers

Fig 1 SMESE-TSHA Prototype for Museums

SMESE-TSHAoverview

From our previous researches, we described that metadata in catalogues represent resource characteristics that can be indexed, queried and displayed by both humans and machine. This SMESE semantic ecosystem harvest and enrich metadata externally and internally. We can see in

Fig 2, the main components of the SMESE ecosystem.

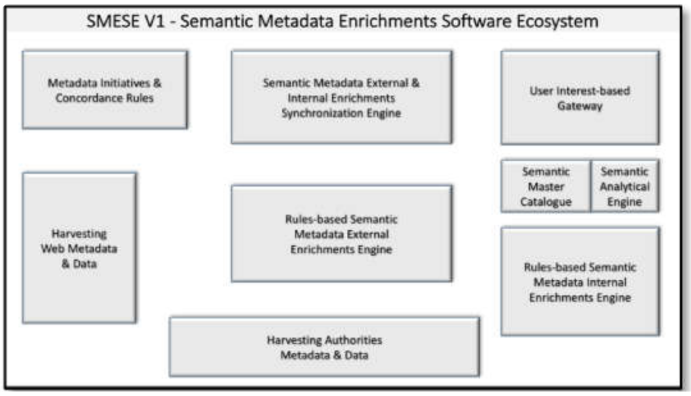
SMESE V1 - Semantic Metadata Enrichments Software Ecosystem
Metadata Initiatives & Concordance Rules
Semantic Metadata External & Internal Enrichments Synchronization Engine
User Interest-based Gateway
Semantic Master Catalogue
Semantic Analytical Engine
Harvesting Web Metadata & Data
Rules-based Semantic Metadata External Enrichments Engine
Rules-based Semantic Metadata Internal Enrichments Engine
Harvesting Authorities Metadata & Data

Fig 2 Semantic Enriched Metadata Software Ecosystem (SMESE V1)

Many metadata schemas exist to describe various types of content: structured and unstructured. It is another issue: How to make sure that the harvesting is accurate?(1) at a time *t* and (2) is repeatable over time to harvest new metadata and data at a time *t* + *p*, *t* denotes the time of the last harvesting and *p* time period elapsed after *t*.

Many aggregators harvest metadata and consequently 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). SMESE ecosystem defines crosswalks that create metadata pathways to different sources of data and metadata. Fig 3shows the semantic metadata meta-catalogu eclassification designed and implemented in the SMESE V1. For TSHA, we will enhance the classification of this model adding the trusted sources from where the metadata has been harvested and the ranking of the sources.

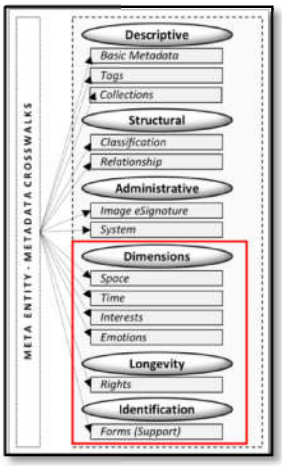

Fig 3 Semantic metadata meta-catalogue classification in the SMESE V1

The semantic annotation process of the new version of SMESE creates relationships between semantic models, such as ontologies, persons and qualified trusted sources. It may be characterized as the semantic enrichment of unstructured and structured content with new knowledge and linking these to relevant domain ontologies and knowledge bases. It typically requires annotating a potentially ambiguous entity mention with the canonical identifier of the correct unique entity.

As last year, an amount of 5 millions content have been harvested over a target amount of close to 500 millions, see the Table 1 (see next page)for an overview of the detail about harvested metadata and data. The text is analyzed by means of extensions of text mining algorithms such as latent Dirichlet allocation (LDA), latent semantic analysis (LSA), support vector machine (SVM) and k-Means.

Note that metadata modeling and an universal metadata model is the main focus of SMESE. Using simulation, the performance of SMESE was evaluated in terms of accuracy of topic detection and sentiment and emotion discovery. Existing approaches to enriching metadata were used for comparison. In

Fig 4, the SMESEplatformfrom previous research work is presented.

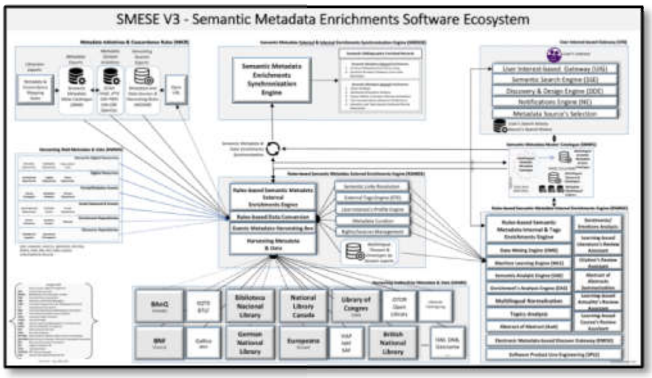

Fig 4 SMESE V3 – Semantic Metadata Enrichment Software Ecosystem

For understanding about SMESE V3 algorithms and processes to semantically enrich metadata using multiple metadata/data sources, refer to previous papers[56].

The unified meta-model from SMESE allows to build a Matrix of entity-metadata. In the next table 1, we have the detail about the meta-catalogue (master catalogue) and his evolution between 2017 and 2019:

Table 1 Master model items description

Number of Items	2017	2019	% Increase
Entity	214	338	*37%*
Metadata	1,548	2,096	*36%*
CrossWalk (Ontology)	26	69	*63%*
International DL Standards	3	46	*93%*
Semantic relationship (Meta)	362	525	*31%*

We can see the evolution of the master model items description related to SMESE V3 algorithms and processes using multiple metadata from different data sources. The number of entities increased from 214 in 2017 to 338 in 2019. An increase of 37%. This means that the unstructured world become a little bit more structured using many individual projects to build this new Master model of structured representation of the world. The number of metadata went up for 36% compare to two years ago. And the semantic relationship increase from 362 to 525, 31%. It seems that Entity, Metadata and Semantic relationship are growing at the same speed from the last two years. In the Fig 5, we can see our previous model about MICR, this model from SMESE V3 evolves to a new model named Traceable Metadata Trusted Sources (TMTS).

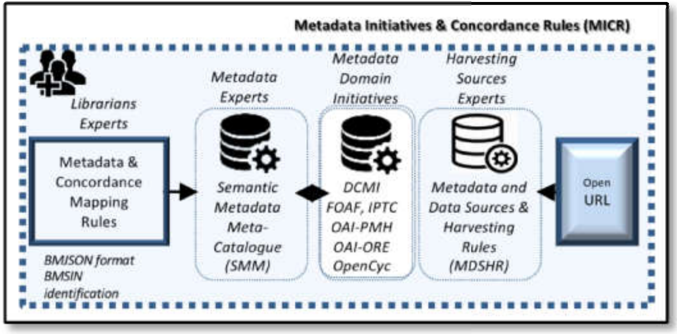
Metadata Initiatives & Concordance Rules (MICR)
Librarians Experts
Metadata & Concordance Mapping Rules
BMJSON format
BMSIN identification
Metadata Experts
Semantic Metadata Meta-Catalogue (SMM)
Metadata Domain Initiatives
DCMI
FOAF, IPTC
OAI-PMH
OAI-ORE
OpenCyc
Harvesting Sources Experts
Metadata and Data Sources & Harvesting Rules (MDSHR)
Open URL

Fig 5 Metadata Initiatives & Concordance Rules (MICR)

SMESE-TSHAAlgorithms

As mentioned above, SMESE-TSHA consist of two main algorithms: SHS and MLM-TSHA. SMESE-TSHA is composed to many processes. The running sequence of is important to optimize SMESE-TSHA; for example, multimedia contents downloading process must be performed after cleaning, verification and validation processes in order to avoid deleting downloaded contents which are double; assume that SMESE-TSHA downloads 3,000,000 images where 2,000,000 are duplicated.

The verification/validation process, contents downloading process and the semantic watching process will not be addressed; these axes of research will be addressed in future works. In following sections, we introduce, in details, SHS and then MLM-TSHA.

Semantic harvesting strategies (SHS)

The goal of SHS is to perform metadata harvesting in the strategic way in order to make the ER approach more simple and efficient. To achieve this goal, SHS begins by sources analysis by experts. The role of these experts is to analyse several metadata sources and classify them semantically. SHS define three harvesting strategies: (1) Strategy #1: Harvesting by hierarchy of trusted sources; (2) Strategy #2: Harvesting by source; ans (3) Strategy #3: Harvesting by interest.

In this work, we present the harvesting strategy by hierarchy.In the case of two other strategies, the users explicitly select the sources whose the entities metadata will be harvested. Specifically, "Harvesting by source" strategy allows users to select themselves the sources to harvest while "Harvesting by interest" strategy allows users to define selection parameters such as source type, source description, contents types, data format, metadata structure and source AWF.

Semantic hierarchy strategy based harvesting algorithm

Here, we present ourhierarchy strategy based harvesting algorithm. This approach is an extension of our previous harvesting algorithm proposed in [33]. Indeed, in [33], we proposed a semantic web metadata harvesting and enrichment model, called Semantic Universal Knowledge Model (SUKM). The goal of SUKM was to allow multi-sources and multi-entities types harvesting and enrichment in order to provide a semantic master entities repository as a rich semantic encyclopedia of knowledgeable entities. Unfortunately, SUKM harvesting algorithm did not apply hierarchy strategy that allows reducing the entity resolution process complexity in terms of time and space. SUKM harvesting algorithm applied source strategy where the sources to harvest are predefines. In this case, after the harvesting, SUKM applied deduplication and merge process to clean harvested entities.

The main contribution of hierarchy strategy based harvesting algorithm is the reduction of the metadata cleaning process, called Entity Resolution. For users, the benefit is the fact that entities are more quickly available for consultation. Hierarchy strategy harvesting consists in ordering the sources according to their trust level. The most trusted sources are harvested and the harvested metadata allows to enrich the previously harvested entities. Fig 6 shows an illustration of hierarchy strategy harvesting model. In our approach, the first trust level sources are the authority files; authority files are individual databases which allow to record detailed information about people, places, subjects, events and many more categories of information relevant about real world entities. The experts analysis process follows some constraints:

1. Where to find a specifictypes of entities
   - Types of entities, Description of entities, Time of the Harvesting.
2. Sources trust order
   - 1st: Authority filessources, 2nd: National libraries sources, 3rd: Derivative authority files sources, 4th: National associations sources and 5th: Researcher associations sources.
3. Analysis and definition the entities sources metadata
   - Name of source, Address of sources, Site web of sources, Location of source, Right of use of source, Trust level of source;
   - For which specific entities: (1) Location name; (2) City name; (3) Country name; (4) Corporate name; and (5) Person name.
4. Assignation of the sourcetrust level

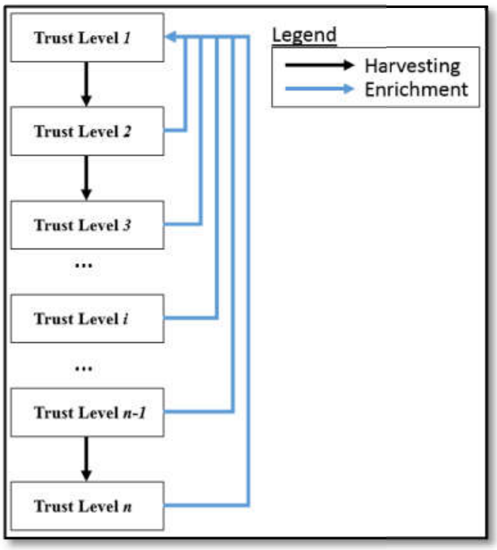

Fig 6 Illustration of hierarchy strategy based harvesting model

As shown the Fig 6, hierarchy strategy harvesting model favours more the sources with more trust level and uses the sources with lesstrust level as enrichment metadata; Table 2 shows the algorithm of hierarchy strategy harvesting.

Table 2 Hierarchy strategy based harvesting algorithm

```
Pseudo code: Hierarchy strategy based harvesting algorithm
Let S = {$s_1$, $s_2$, …, $s_n$}be the sources of entities
Let  eAWF(e) be the Accuracy Weight Factor of entity e
Let E be a list of entities of same type
Let TypeSimilarity (t, t’)be a function to identify the entities of same type
1.     $L_1$ =Downloadentities files of$s_1$
2.     Select entities eof same typet from $L_1$
3.     For each selected entity $e_1$
4.     eAWF($e_1$) = 100
5.     Addeinto E
6.     Select entities $e'_1$of type t’of from $L_1$ where TypeSimilarity (t, t’)
7.     For each selected entity $e'_1$
8.     eAWF($e'_1$) = 90
9.     Add$e'_1$into E
10.    For each entity e of E
11.    For each metadata mof e
12.    CALLSMESE-TSHA mAWF evaluation algorithm
13.    Indicate $s_1$as the source of the metadatam
14.    $L_2$ =Downloadentities files of$s_2$
15.    Select entities $e_2$of same typet from $L_2$
16.    Add$e_2$into E’
17.    eAWF($e_2$) = 100
18.    Select entities $e'_2$ of type t’ of from $L_1$ where TypeSimilarity (t, t’)
19.    Add$e'_2$into E’
20.    eAWF($e'_2$) = 90
21.    For each selected entity e’of E’
22.    IFe’already exists in E according to its ID
23.    e and  e’ are same entity in real world
24.    eAWF(e)=Biggest between {eAWF(e), eAWF(e’)}
25.    For each metadata mof e’
26.    CALLSMESE-TSHA mAWF evaluation algorithm
27.    IFmalready exists in e,
28.    Indicate $s_2$ as the source of the metadata m
29.    ELSE
30.    Addmin e
31.    Indicate $s_2$ as the source of the metadata m
32.    ELSE  //e’does not exist in E according to its ID
33.    Adde’ into E
34.    CALLSMESE-TSHA mAWF evaluation algorithm
35.    Indicate $s_2$as the source of the metadatam
36.    REPEATfrom 14 to 35 with$S - \{s1, s2\}$
```

From line 1 to 13, we harvest the entities of same type from the sources with more trust level which has the biggest count of entities. Then, from line 14 to 35, we harvest the second source which meets the both constraints: more trust level with biggest count of entities. In order word, the set S is sorted according to the two conditions (more trust level with biggest count of entities); after the harvesting of the first item of the set S, the line 14 to 35 are repeated for the rest of the set S; this represents the enrichment shows in the Fig 6 (see the previous page). This approach allows to harvest a lot of entities from the sources with more trust level. The function ***TypeSimilarity (t, t’)*** allows to identify the entity which is not evident to confirm its type; for example, VIAF assigns the value “MUSEUM” to the metadata “OCCUPATION” which allows to identify explicitly the entities type MUSEUM. Unfortunately, for certain entities, the metadata “OCCUPATION” are empty; in this case, it is not explicitly to identify the type of these entities; thus, based on the other metadata such as “NAME”, “CLASSIFICATION”, ***Type Similarity (t, t’)*** allows to infer the type of this entity. In order to keep the reference of the entities sources per metadata, SMESE-TSHA represents each metadata with the sources which have the same metadata value; for example, Table 3shows the sources referencing of metadata “Entity Name” for a museum.

Table 3 Referencing of metadata sources

Entity: *National Museum of American History*
Entity Name:

- National Museum of American History [s_1][s_9][s_{12}]
- National museum of American history Washington, D.C[s_2][s_3][s_8]
- National Museum of American History (Washington)[s_4][s_{11}][s_{13}]
- EstadosUnidos., Museum of History and Technology [s_5]
- Museum of History and Technology (EstadosUnidos)[s_{10}][s_{15}]
- National museum of American history (Washington, D.C)[s_6][s_7]
- National Museum of American History (U.S.)[s_{14}]

“National Museum of American History [s_1][s_9][s_{12}]” means that the name “National Museum of American History” is the same for the sources [s_1], [s_9] and [s_{12}].

Remember that SMESE-TSHA consist of two main algorithms that are: SHS and MLM-TSHA. After the introduction of SHS in the previous section, in next section, we introduce the second algorithm of SMESE-TSHA, called MLM-TSHA.

Machine learning model for entity resolution (MLM-TSHA)

MLM-TSHAgoal is to address the problem of entity resolution and entity enrichment. In details, MLM-TSHA consists of automatic multi-sources metadata matching, cleaning and entity resolutionand metadata enrichments using machine learning models and artificial intelligence algorithms.

MLM for TSHAalgorithmstry 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.

MLM cleaning algorithm

In this section, we present the algorithm uses by SMESE-TSHA to clean the harvested entities, called MLM based Entity Resolution (MLMER) from several sources with various data structure for unified and trusted repository (UTR). Our ER approach is a combination of learning and self-learning techniques into ER. Self-learning with automatic seed selection addresses the problem of lack of labelled datasets. The MLMERalgorithm is performed in six steps. As with a typical ER approach, blocking is the first step and it can be thought of as a pre-processing step. The second step of the MLMER algorithm is the selection of similarity measure schemes. In this step we search the whole space of all possible similarity measure schemes in order to select the most diverse subset of it. In the third step (seed selection with metadata weighting) each of the selected similarity measure schemes is first used to generate a set of similarity vectors. Then the automatic seed selection process is performed on each set of similarity vectors. As the output of this step different sets of seeds are selected. In the fourth step (Selecting the most diverse sets of seeds), the diversity between sets of seeds is measured using the proposed technique referred to as Seed Q Statistics. Only those most diverse sets of seeds are selected. In the fifth step the self-learning algorithm is applied with each of the selected sets of seeds. In the last step the proposed contribution ratios of base classifiers (BCs) are used to eliminate the weakest BCs from the final ensemble. Finally, for each pair of records the mode of the predictions of the selected self-learning models is provided as the final prediction.

For a given set of M similarity measures we could construct M^N possible similarity measure schemes of size N, where N is the number of metadata. We need to select those similarity measure schemes that produce the most diverse sets of similarity vectors from the given dataset. First, we select a set of similarity measure for each of the metadata ***f***. A pool of similarity measure schemes can then be generated, as a cross product of the sets of similarity measure selected for each of the metadata ***f***. Given a similarity measure m and a pair of entities e_1 and e_2 , each with N metadata, f_1 , . . . , f_N , the metadata similarity between e_1 and e_2 on metadata f_i , for i = 1 , 2 , . . . , N, is defined as: $m(e_1; e_2; f_i)$.For a given set of entities E and two similarity measures m_1 and m_2, let$\overrightarrow{m_1(E,f)}$ and $\overrightarrow{m_2(E,f)}$ be two vectors with each corresponding pair of elements in $\overrightarrow{m_1(E,f)}$ and $\overrightarrow{m_2(E,f)}$representing the metadata similarity between each possible pair of entities in E on metadata *f*. The similarity between m_i and m_j on metadata*f* is defined as:

$$\mathrm{Sim}_f(m_i,m_j)=\mathrm{CosSim}(\overrightarrow{m_1(E,f)}, \overrightarrow{m_2(E,f)}) \qquad (1)$$

CosSimmeasures the similarity between two vectors of an entity space by the cosine of the angle between them. We aim to select a set of similarity measures for each of the fields, in which no CosSim between two similarity measures is greater than a threshold.

Each of the selected similarity measure schemes is used to generate a set of similarity vectors for all the pairs of entities produced by the blocking process. For each vector set, a small group of similarity vectors are automatically labelled as match and non- match, which will be used as seeds in the self-learning process. A SVM is used in the self-learning process. To improve the efficiency of the learning process we apply the Stochastic Gradient Descent (SGD) algorithm for estimating the parameters of SVMs, which is very effective for large-scale online learning problems. With the proposed method the class probability distribution of an instance produced by the SVM-SGD algorithm as out- put is used to determine the level of its confidence on the classification of the instance. For example, if the class probability distribution of an instance is 0 on match and 1 on non-match then the SVM-SGD classifies the instance as non-match and the level of its confidence on the classification is 1.Following the self-learning process, a collection of classification models is generated. Since the proposed method is fully unsupervised we are not able to evaluate how good each of classification models is. Therefore, there is a risk of including classifiers with very poor accuracy (i.e., below 0.5) which are not valid in general, into the ensemble. In order to address this issue we propose a statistic which takes into account the contribution ratio of each individual base classifier to the final output of the ensemble. Each base classifier makes a prediction on each record pair as match or non-match. Following this, the mode of all the predictions by all the base classifiers is taken as the prediction of the ensemble.

Evaluation using simulations

In this section we present the experimental evaluation of our proposed approach, called SMESE-TSHA. The objective of our experimental evaluation is to compare, according to the literature, more recent and performing algorithms on various types of entities.

As comparison terms, we use the ER described in [5], [8] and [56], which are referred to as ER1, ER2, and ER3, respectively. ER1, ER2, and ER3were selected because, to the best of our knowledge, they represent the most recent work related to Entity Resolution that outperform existing approaches. Table 4shows the characteristics of DAMP, ER1, ER2, and ER3.

Table 4 Entity Resolution (ER) schemes forcomparison

Schemes	Human annotation	Machine learning
ER1[5]	YES	NO
ER2[8]	YES	YES
ER3[56]	NO	YES
SMESE-TSHA	YES	YES

According to the Table 4, ER2 and SMESE-TSHA combine Human annotation and machine learning model (MLM).

Simulation Setup and Datasets Characteristics

To measure SMESE-TSHA. ER1, ER2 and ER3 performance, a simulator program has been developed using Java. The server characteristics for the simulations were: Dell Inc. PowerEdge 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 VMW are ESXi 6.0.

The Datasets we use was provided by forty-three (43) data sources of various types. 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.

Table 5 Evaluation datasets entities types

Entities Type	Number of entities
Museums and Galleries	83,437
Artworks	11,890,435

Places (Countries, States and Cities)	2,562,458
Artists	1,071,724

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 [53, 54, 57], the same performance metrics can be used for comparison: ***Precision and Recall***. For example, accuracy is related to rate of true entity resolution. ***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:

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

We evaluate the scalability of the proposed approach in terms of *running time*[11, 54]; 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. For each run, we compute each criteria using their equation, respectively; thus, to obtain the simulation results shown in Figs. 8and 10, 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. 8 to 10, the average precision, average recall and average running time varying with the datasets ID.

Fig 7 shows the average precision when varying the Datasets ID.

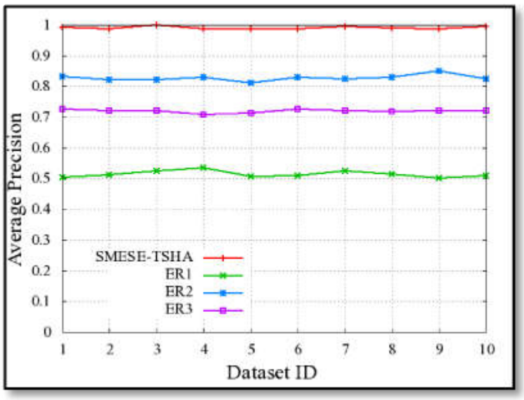

Fig 7 Precision VS Dataset ID

We observe that SMESE-TSHAoutperforms ER1, ER2 and ER3; for example, SMESE-TSHA provides an average precision of 0.98 per Dataset, whereas ER2 (more efficient than ER1 and ER3 in this scenario) provides an average of 0.83 per Dataset; overall, the average relative improvement (defined as [average precision of SMESE-TSHA— average precisionof ER2]) of SMESE-TSHA compared with ER2 (resp. ER1 and ER3) is about 15% (resp. 47% and 26%) per Dataset. This can be explained by the fact that SMESE-TSHAuses a hierarchy of trusted sources for harvesting process.

Fig 8 presents the average recall when varying the Datasets ID.

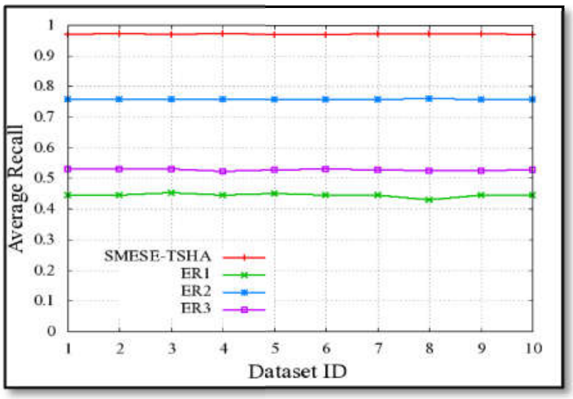

Fig 8 Recall VS Dataset ID

Fig 8 shows that SMESE-TSHAoutperforms ER1, ER2 and ER3; SMESE-TSHA provides an average precision of 0.97 per Dataset, whereas ER2 (more efficient than ER1 and ER3 in this scenario) provides an average of 0.78 per Dataset. The average relative improvement (defined as [average recall of SMESE-TSHA— average recallof ER2]) of SMESE-TSHA compared with ER2 (resp. ER1 and ER3) is about 19% (resp. 53% and 45%) per Dataset. This is mainly due to the fact that SMESE-TSHA detectswell the true negative candidates in contrast to ER1, ER2 and ER3.

Fig 9shows the average running time when varying the Datasets ID. We observe that ER1 outperformsSMESE-TSHA, ER2 and ER3. The average relative improvement (defined as [average recall of ER1— average recallofSMESE-TSHA]) of ER1compared with SMESE-TSHA (resp. ER2 and ER3) is about 1.38 hours (resp. 1.18 hours and 1.07 hours) per Dataset.

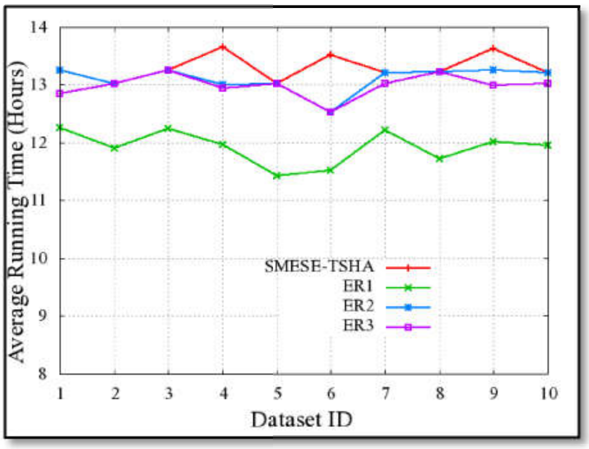

Fig 9 Running time VS Dataset ID

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.

Summary and future work

We have shown that it is possible and more accurate to harvest metadata and data using trusted sources instead of to harvest just sources of metadata and data. As an example, it better to harvest all the museum of the world to start with a number of trusted and ranked sourcesof metadata and data than just to harvest the web or some databases without any guidance about their relevancy and accuracy. The meta-catalogue that we build in SMESE project as to include the list of trusted sources of metadata related to a type of object, like a notice of Museum. So we need to add to our meta-catalogue to list of trusted sources of metadata and trusted thesaurus.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:

- the repeatable process of harvesting and the timely concept of trusted and ranked sources of metadata;
- the verification/validation process of the trusted sources of metadata or how to validate automatically the ranking of a metadata source.

References

1. R. C. Steorts, “Entity Resolution with Empirically Motivated Priors,” *Bayesian Anal.,* vol. 10, no. 4, pp. 849-875, 2015/12, 2015.
2. P. Christen, and R. W. Gayler, "Adaptive Temporal Entity Resolution on Dynamic Databases," *Advances in Knowledge Discovery and Data Mining.* pp. 558-569.
3. A. Globerson, N. Lazic, S. Chakrabarti, A. Subramanya, M. Ringgaard, and F. Pereira, "Collective Entity Resolution with Multi-Focal Attention." pp. 621–631.
4. R. Nuray-Turan, D. V. Kalashnikov, and S. Mehrotra, “Adaptive Connection Strength Models for Relationship-Based Entity Resolution,” *J. Data and Information Quality,* vol. 4, no. 2, pp. 1-22, 2013.
5. N. Vesdapunt, K. Bellare, and N. Dalvi, “Crowdsourcing algorithms for entity resolution,” *Proc. VLDB Endow.,* vol. 7, no. 12, pp. 1071-1082, 2014.
6. 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.
7. D. Firmani, B. Saha, and D. Srivastava, “Online entity resolution using an Oracle,” *Proc. VLDB Endow.,* vol. 9, no. 5, pp. 384-395, 2016.
8. 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.
9. G. Papadakis, G. Papastefanatos, T. Palpanas, and M. Koubarakis, "Scaling Entity Resolution to Large, Heterogeneous Data with Enhanced Meta-blocking." pp. 1-12.
10. 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.
11. M. Mountantonakis, and Y. Tzitzikas, “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, 2018.
12. 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.
13. H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, “Big data and its technical challenges,” *Commun. ACM,* vol. 57, no. 7, pp. 86-94, 2014.
14. X. L. Dong, and D. Srivastava, "Big data integration." pp. 1245-1248.
15. 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.
16. M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mobile Networks and Applications,* vol. 19, no. 2, pp. 171-209, April 01, 2014.
17. 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.
18. 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.
19. 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.
20. 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. 2, 2015.
21. K. Craig A., and S. Pedro, “Exploiting Semantics for Big Data Integration,” *AI Magazine,* vol. 36, no. 1, pp. 25-38, 2015.
22. R. Brisebois, A. Abran, and A. Nadembega, “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, 2017.
23. R. Brisebois, A. Abran, and A. Nadembega, “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, 2017.
24. 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.

25. 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.
26. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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,, 2017.
27. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
28. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
29. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
30. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
31. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
32. R. Brisebois, A. Abran, A. Nadembega, and P. N’techobo, “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, 2017.
33. R. Brisebois, A. Nadembega, P. N’techobo, and H. L. Djeuteu, “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, 2017.
34. E. Vargiu, and M. Urru, “Exploiting web scraping in a collaborative filteringbased approach to web advertising,” *Artificial Intelligence Research,* vol. 2, no. 1, pp. 44-54, 2013.
35. [35] 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.
36. 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.
37. N. R. Haddaway, “The Use of Web-scraping Software in Searching for Grey Literature,” *The Grey Journal,* vol. 11, no. 3, 2015.
38. 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.
39. 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.
40. 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.
41. A. Casali, C. Deco, and S. Beltramone, “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, 2016.
42. 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.
43. M. Mountantonakis, and Y. Tzitzikas, “High Performance Methods for Linked Open Data Connectivity Analytics,” *Information,* vol. 9, no. 134, pp. 1-33, 3 June 2018, 2018.
44. 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.
45. G. Simonini, S. Bergamaschi, and H. V. Jagadish, “BLAST: a loosely schema-aware meta-blocking approach for entity resolution,” *Proc. VLDB Endow.,* vol. 9, no. 12, pp. 1173-1184, 2016.
46. A. Passos, V. Kumar, and A. McCallum, “Lexicon Infused Phrase Embeddings for Named Entity Resolution,” *CoRR,* vol. abs/1404.5367, 2014.
47. T. Williams, and M. Scheutz, "POWER: A domain-independent algorithm for Probabilistic, Open-World Entity Resolution." pp. 1230-1235.
48. A. Elmagarmid, I. F. Ilyas, M. Ouzzani, J.-A. Quian, #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.
49. 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.

50. B. Ramadan, P. Christen, H. Liang, and R. W. Gayler, “Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution,” *J. Data and Information Quality,* vol. 6, no. 4, pp. 1-29, 2015.
51. F. Raul Castro, P. Peter, K. Jay, N. Neha, and R. Jun, "Liquid: Unifying Nearline and Offline Big Data Integration."
52. Y. Tong, C. C. Cao, C. J. Zhang, Y. Li, and L. Chen, "CrowdCleaner: Data cleaning for multi-version data on the web via crowdsourcing." pp. 1182-1185.
53. Hakan Kardes, Deepak Konidena, Siddharth Agrawal, Micah Huff, and A. Sun, "Graph-based Approaches for Organization Entity Resolution in MapReduce."
54. L. Zhu, M. Ghasemi-Gol, P. Szekely, A. Galstyan, and C. A. Knoblock, "Unsupervised Entity Resolution on Multi-type Graphs," *The Semantic Web – ISWC 2016.* pp. 649-667.
55. S. E. Whang, P. Lofgren, and H. Garcia-Molina, “Question selection for crowd entity resolution,” *Proc. VLDB Endow.,* vol. 6, no. 6, pp. 349-360, 2013.
56. 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.
57. V. Efthymiou, K. Stefanidis, and V. Christophides, "Big data entity resolution: From highly to somehow similar entity descriptions in the Web." pp. 401-410.

How to cite this article:

Ronald Brisebois, Apollinaire Nadembega and Toufic Hajj.2019, Trusted Smart Harvesting Algorithmbased on Semantic Relationship And Social Networks (Smese-Tsha). *Int J Recent Sci Res.* 10(01), pp. 30593-30604.
DOI: http://dx.doi.org/10.24327/ijrsr.2019.1001.3088

****