Harvesting Knowledge from Cultural Heritage Artifacts in Museums of India

Abhilasha Sancheti[1], Paridhi Maheshwari[2], Rajat Chaturvedi[3],
Anish V. Monsy[4], Tanya Goyal[5], and Balaji Vasan Srinivasan[1]

[1] Big data Experience Lab, Adobe Research, Bangalore, India
[2] Department of Electrical Engineering, Indian Institute of Technology, Kanpur
[3] Department of Computer Science, Indian Institute of Technology, Bombay, India
[4] Department of Computer Science, Indian Institute of Technology, Guwahati, India
[5] Department of Computer Science, University of Texas at Austin, Austin, TX, USA
sancheti@adobe.com, 1997.paridhi@gmail.com, chaturvedirajat96@gmail.com,
anishvmonsy2@gmail.com, tanyagoyal.93@gmail.com, balsrini@adobe.com

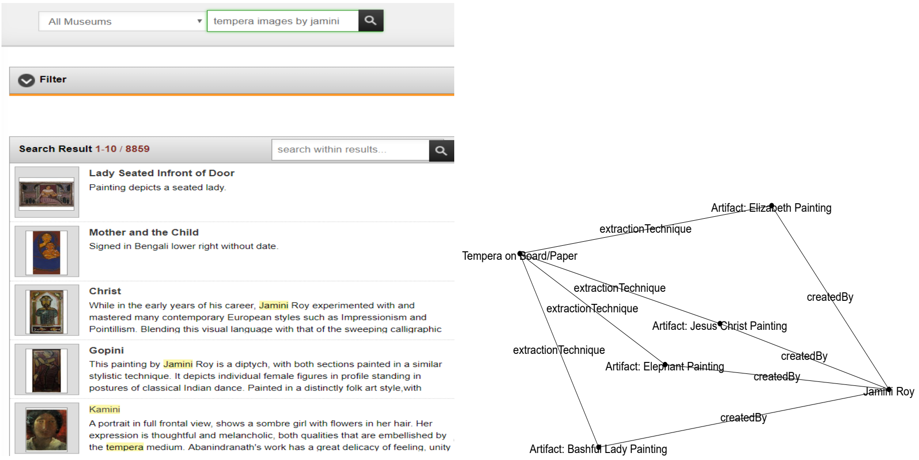
Abstract. Recent efforts towards digitization of cultural heritage artifacts have resulted in a surge of information around these artifacts. However, the organization of these artifacts falls short with respect to accessing the facts across these entities. In this paper, we present a method to harvest the knowledge and form a knowledge graph from the digitized artifacts in the Museums of India repository via distant supervision to enable better accessibility of the facts and ability to extract new insights around the artifacts. Triples extracted from an open information extractor are first canonicalized to a standard taxonomy based on a metric-based scoring. Since a standard taxonomy is insufficient to capture all the relationships, we propose a sequential clustering based approach to add artifact specific relationships to the taxonomy (and to the knowledge graph). The graph is enriched by inferring missing facts based on a probabilistic soft logic approach seeded from a frequent item set framework. Human evaluation of the final knowledge graph showed an accuracy of 75% on par with knowledge bases like DBpedia.

1 Introduction

Cultural heritage represents a legacy of traditions and customs inherited from the past, maintained in the present and preserved for the benefit of future generations. As an attempt to reach wider audiences, various museums across the globe have digitized their artifact collections [6, 9, 14] and have made them available on web portals to facilitate better availability of the artifacts data to the public. However, in the absence of a proper organization, the large amount of digital content in these portals can be overwhelming and infeasible to interpret the information associated with the artifacts.

For a cultural enthusiast, a simple keyword search might not always fetch what (s)he is looking for since some of the information can span details from multiple artifacts. Standard information retrieval system cannot satisfy such needs since they serve information from a single source only. For example, Fig.

1(a) shows a sample query “tempera images by Jamini” to Museums of India (MOI) [14], an online portal about cultural artifacts in India. This illustrates that the current organization of the artifacts does not capture the specific style of paintings by an author. There could be several such aspects that could be useful for gaining insights and discovering relationships between the artifacts. This calls for a systematic approach to harvest the knowledge from cultural artifacts in order to enhance the understanding and facilitate the organization around them to enable better accessibility of the facts around these artifacts.



tempera images by jamini

Search Result 1-10 / 8859

Lady Seated Infront of Door
Painting depicts a seated lady.

Mother and the Child
Signed in Bengali lower right without date.

Christ
While in the early years of his career, Jamini Roy experimented with and mastered many contemporary European styles such as Impressionism and Pointillism. Blending this visual language with that of the sweeping calligraphic

Gopini
This painting by Jamini Roy is a diptych, with both sections painted in a similar stylistic technique. It depicts individual female figures in profile standing in postures of classical Indian dance. Painted in a distinctly folk art style,with

Kamini
A portrait in full frontal view, shows a sombre girl with flowers in her hair. Her expression is thoughtful and melancholic, both qualities that are embellished by the tempera medium. Abanindranath's work has a great delicacy of feeling, unity

(a) Query Result from Museums of India Portal

(b) Query Result from CultKB

Fig. 1. Search result for query paintings by ’Jamini Roy’ using ’Tempera’ from the Museums of India web portal and the constructed CultKB

There is a growing body of work that focuses on harvesting knowledge from structured and unstructured data sources [23] towards building a knowledge base. Such knowledge bases/graphs serve as an excellent organization for insightful explorations as well as cross-artifact fact extraction/retrieval. Popular knowledge bases like DBpedia [2], NELL [7], YAGO [24] contain “facts” of the form “subject-predicate-object” and are generally extracted from a generic corpus like Wikipedia and canonicalized based on a standard taxonomy.

While knowledge graphs offer a good solution towards exploration, standard taxonomies are insufficient to capture the facts in cultural artifacts and render the taxonomies from standard knowledge bases inapplicable to the needs of cultural artifacts. However, building such a taxonomy from scratch specific to the cultural artifacts is also infeasible since this requires significant input from domain-experts. To address these challenges, we start with a standard taxonomy and harvest the facts from artifacts canonicalized to this taxonomy. Simultaneously, we also enrich the taxonomy to cover the needs of the cultural heritage artifacts by adding new artifact-specific relationships to the taxonomy (and the corresponding facts to the knowledge graph). The proposed approach addresses 3 major challenges:

1. The meta-data in the digitized cultural artifacts do not always have well-formed text and hence can result in noisy facts. Therefore, processing these noisy facts to extract meaningful facts via appropriate canonicalization is the first major challenge that we address in our approach.
2. Since standard taxonomies are insufficient to canonicalize all the facts that exist in the data from a specific domain, building a systematic approach to enrich the taxonomy with domain-specific relationships is the second challenge that we address. Identifying the new predicates for the taxonomy would require de-duplicating their multiple representations and reducing the overall noise in the extracted facts.
3. Finally, the uniqueness of the data about cultural artifacts provides an opportunity to look for patterns in the extracted facts to infer new/missing facts and enrich the knowledge graph with additional relationships that are not already present in the data.

Fig. 1(b) shows a list of paintings made by the artist using tempera technique in response to the same query in Fig. 1(a) extracted from our proposed knowledge graph built on the MOI [14] data.

2 Related Work

Knowledge harvesting deals with extracting meaningful relationships and constructing knowledge graphs from text and other unstructured as well as structured sources [15]. Knowledge graph extraction involves the problem of inferring entities (nodes) and their relations/predicates (edges) from uncertain data while simultaneously incorporating constraints imposed by ontological inferences [23]. Entities in uncertain data might appear in different forms due to mis-spelled usages, use of synonyms or any other factors. Therefore, entities and relations extracted from the uncertain data are canonicalized, which is the process of standardizing the extracted facts to a taxonomy to achieve a consistent knowledge graph. Ontologies in the taxonomy aid in adding constraints to the facts for maintaining consistency and meaningfulness of the extracted facts [7].

There exists a number of large-scale **publicly available knowledge bases** like YAGO [24] , DBpedia [2] , Freebase [5], etc. While DBpedia [2] builds upon the structured info-boxes of Wikipedia, YAGO [24] automatically derives its facts from Wikipedia and Wordnet using a combination of rule-based and heuristic approaches. But these works deal with knowledge covering a broad range of real-world concepts and are not restricted to any particular domain. There exists very limited work on building knowledge bases for a specific domain. Kobren et al. [16] build a knowledge base of scientists and their affiliation via crowdsourcing. Similarly, Zhao et al. [26] use crowdsourcing to build a software-engineering knowledge base from StackOverflow. However, given the limited expertise available for cultural artifacts, such crowd-sourced approaches are not feasible in our case.

Developments in **digitizing cultural artifacts** have led to a few efforts to understand and organize such cultural artifacts. Agirre et al. [1] developed a system, PATH to aid people in navigating through the Europeana [9] artifacts. PATH measures artifact similarity to Wikipedia articles/entities by comparing the topics generated from each artifacts metadata using Latent Dirichlet Allocation (LDA) with the Wikipedia topics. The matched Wikipedia articles/entities are used to generate hierarchies which help in browsing and exploring the artifacts. Fernando et al. [11] explored techniques to automatically add Wikipedia links to resources in order to provide relevant background information. However, these approaches are not suitable to organize our data owing to the limited information available about Indian cultural heritage on open knowledge sources like Wikipedia. This restrains the use of external data sources.

With these limitations in mind, we propose a novel algorithm to harvest knowledge from a cultural artifact corpus [14], canonicalize it to a standard taxonomy and simultaneously enrich the taxonomy, and finally refine and infer any missing information in the extracted facts. The proposed approach is designed for distant supervision and hence can scale without annotations from a human expert.

3 Harvesting data from MOI [14] into CultKB

Table 1(a) shows the list of museums and their artifacts from the portal. Table. 1(b) shows the distribution of different categories of artifacts in the data. The portal currently hosts information of over $90k$ artifacts including paintings, manuscripts, coins, and sculptures.

Table 1. Statistics around various artifacts in the Museums of India web portal

(a) Artifacts in different museusm

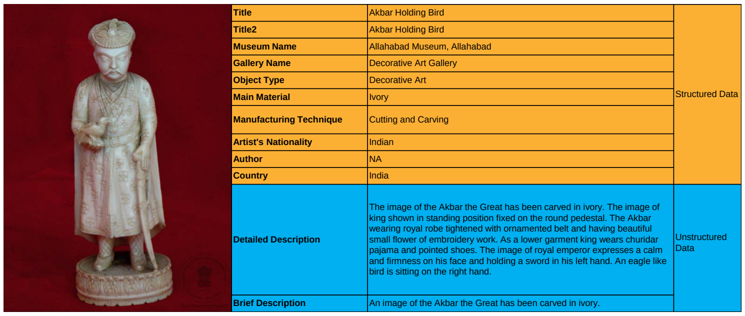
Museums	Artifacts
Salar Jung Museum, Hyderabad	23, 981
National Museum, New Delhi	21, 384
The Allahabad Museum	13, 277
Indian Museum, Kolkata	12, 228
Nagarjunakonda Museum	8, 400
Victoria Memorial Hall, Kolkata	2, 900
National Gallery of Modern Art, New Delhi	5, 423
National Gallery of Modern Art, Mumbai	1, 400
National Gallery of Modern Art, Bengaluru	500
Goa Museum	700

(b) Artifacts in top 10 categories

Category	Artifacts
Painting	10207
Decorative Art	7227
Manuscript	7152
Coin	6714
Terracotta	4010
Miniature Painting	3889
Soldier	3843
Porcelain	3799
Anthropology	3710
Central Asian Antiquities	2930

Fig. 2 shows a sample artifact with the associated meta-data. The artifact meta-data is in the form of field-value pairs which includes the structured data such as the title, creator and year of work along with the unstructured data such as brief and detailed description about the artifact. We build **CultKB** our knowledge base of cultural artifacts from MOI by harvesting this meta-data.

We begin with the canonicalization of the structured data to the YAGO taxonomy [24]. We canonicalize the unstructured data to the YAGO taxonomy using a voting based mechanism across different scoring functions. To account



Title	Akbar Holding Bird	Structured Data
Title2	Akbar Holding Bird	
Museum Name	Allahabad Museum, Allahabad	
Gallery Name	Decorative Art Gallery	
Object Type	Decorative Art	
Main Material	Ivory	
Manufacturing Technique	Cutting and Carving	
Artist's Nationality	Indian	
Author	NA	
Country	India	
Detailed Description	The image of the Akbar the Great has been carved in ivory. The image of king shown in standing position fixed on the round pedestal. The Akbar wearing royal robe tightened with ornamented belt and having beautiful small flower of embroidery work. As a lower garment king wears churidar pajama and pointed shoes. The image of royal emperor expresses a calm and firmness on his face and holding a sword in his left hand. An eagle like bird is sitting on the right hand.	Unstructured Data
Brief Description	An image of the Akbar the Great has been carved in ivory.	

Fig. 2. A sample artifact along with the associated structured and unstructured information from Museums of India

for predicates not in the taxonomy, we use a density-based spatial clustering approach to identify valid predicates and de-noise their multiple manifestations. We finally use a probabilistic soft logic based approach to infer missing facts in the constructed knowledge graph.

Canonicalization of Structured Data: The structured data, in the form of field-value pairs, naturally occurs in the <subject, predicate, object> format with each distinct field representing an edge between the artifact and the corresponding field value. Since the number of predicates in the structured data was small, we identified the predicates in the structured data as a part of preprocessing and manually mapped them to the appropriate predicates in the YAGO taxonomy [24] after an initial set of candidate predicates being extracted via string matching. The triples thus extracted are directly added to our knowledge graph which has the subject/object as its nodes and the predicates as edges.

Canonicalization of Unstructured Data: For canonicalizing the unstructured text, the artifact description is preprocessed to resolve all co-referencing pronouns using the Stanford Co-reference Parser [17]. All possible triples are extracted from the processed text based on an open information extraction (OpenIE) architecture [4, 10]. OpenIE architecture identifies relation phrases in sentences based on syntactic and lexical constraints and assigns a pair of noun arguments for each extracted relation. For each triple, the entity type of subject and object are recognized using the Stanford Named Entity Recognizer [12]. The OpenIE triple extraction is based on the sentence structure analysis and therefore tends to be noisy.

To reduce the noisy triples and resolve redundant and ambiguous facts, the entities (subject and object) and the predicates are mapped to the YAGO taxonomy [24]. For the entities, an edit distance is computed from the matching entities in YAGO and the map beyond a threshold (σ_{entity}) is used as the canonicalized entity.

The canonicalization of predicates is constrained on the nature of entities associated in the artifact triples and that of YAGO triples by incorporating the ontological knowledge of the relationships between entity types to remove noisy triples. For example, the domain and range constraints `DOMAIN`(isWrittenBy, book) and `RANGE`(isWrittenBy, person) specify that the relation 'isWrittenBy' is a mapping from entities with type book to entities with type person. The appropriate YAGO predicate for a given triple is then identified based on an ensemble of three approaches:

1. *Semantic Mapping:* The first approach captures the semantic similarity of words in the phrase and the YAGO relations using a vector space model. It involves computing the cosine similarity between the Word2Vec embeddings [18, 19] of the relationships from artifact triples and those from YAGO. Word2Vec captures the semantic space of the words and therefore such a measure maps the relationships based on their semantic closeness to the relationships in the YAGO taxonomy.
2. *Syntactic Mapping:* In this approach, the resemblance of two predicates is determined by the resemblance of the main verbs. A dependency parser is used to extract the dependency tree from the unstructured source text and a network of "cognitive synonyms" [20] of the root verb of the dependency tree is identified. This network of synonyms is compared with the root verbs of the YAGO relations to establish a correspondence between relations in the syntactic sense.
3. *Pattern based Mapping:* Two verbal phrases are likely to be similar if they share some common pattern of words, with a possible difference of some words like helper verbs and adjectives. With this intuition, the last approach is extended from [21] which obtains textual patterns in binary relations, transforms them into syntactic-ontologic-lexical patterns using frequent item set mining [13] and constructs a taxonomy for these patterns. We match the closest YAGO relationship corresponding to a current pattern taxonomy triple (including the respective POS tags) and assign it as the predicate.

An empirical threshold is used for every approach to find suitable predicate in the taxonomy. A ranked order of target predicates (beyond the threshold) from all the 3 methods is combined based on a voting mechanism to determine the best canonicalized relationship for the current triple.

Enriching the taxonomy with cultural heritage specific predicates: Canonicalization based on a standard taxonomy does not standardize all the extracted triples due to the uniqueness of the relationships in the cultural artifacts. Since the OpenIE predicates are extracted from the sentences, there are multiple manifestations of the same relationships in the database. This calls for a novel approach to enrich the initial taxonomy with predicates specific to the cultural heritage.

Starting with the mapped and unmapped relations, cosine similarity between the Word2vec [18, 19] embeddings of the relationships is used to perform a density-based spatial clustering (DBSCAN) [8]. DBScan is capable of identifying the number of clusters simultaneously. This resulted in $20,000$ different relation types being grouped into 7000 clusters. Incorporating a constraint of maintaining the same NER tags of the subject and object throughout the cluster resulted in partitioning into 9,000 clusters.

For the rest of the clusters, if a predicate from YAGO taxonomy is a part of the cluster, the cluster is tagged with the corresponding YAGO predicate and all the facts are updated with this predicate. In the absence of such a predicate, a representative predicate was chosen based on its frequency of occurrence in the corpus. The NER tags of the subject and object of the associated predicate are used to define the domain and range of the new relationship. Clusters with a significant relation (based on the threshold) are added to the taxonomy and the rest are ignored.

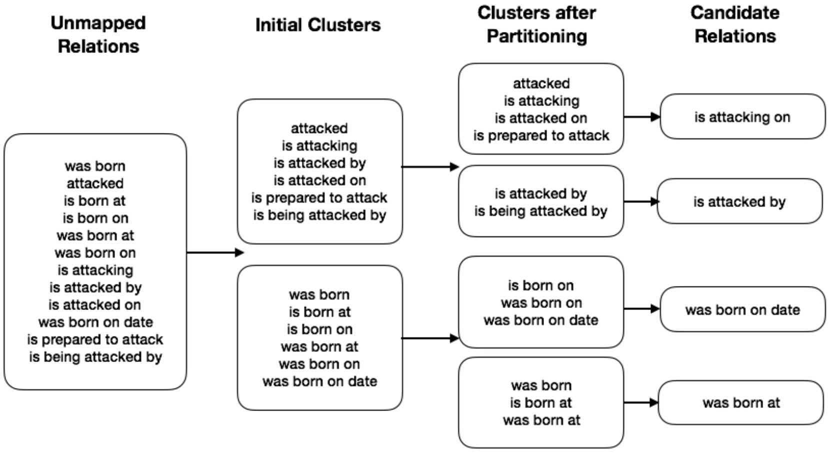


Fig. 3. Illustration of Clustering Algorithm

Figure 3 shows an illustration of various steps in the clustering process. The unmapped relations are first clustered depending upon the verb ‘attack’ or ‘born’. ‘was born at’ and ‘was born on’ are originally in the same cluster but the corresponding NER tags are ‘location’ and ‘date’ respectively. Hence, they are partitioned into different clusters.

Inferring Missing Information: Since the facts are extracted from manually curated content, the knowledge graph is subject to the Open World Assumption, which states that *any missing triple is not necessarily false, just unknown.* The knowledge graph is enriched with new triples based on a probabilistic approach that simultaneously identifies the missing information, strengthens the confidence value of correct facts and resolves conflicts in the data.

Logical rules are extracted via association rule mining [13] taking into account the Partial Completeness Assumption (PCA) which states that if there is at least one object associated with a subject through a relation then the relationship is considered complete. This implies that the PCA makes predictions only for those entities that have an object for the relationship, and remains silent otherwise.

Logical rules, of the form,

$$< E_1, R_1, E_2 > \wedge < E_2, R_2, E_3 > \wedge \ldots \wedge < E_n, R_n, E_{n+1} > \Rightarrow < E_1, R_{n+1}, R_n >,$$

encode frequent correlations in the data. The left-hand side of the implication is called the body and the right-hand side is the head. The rules are assigned a normalized confidence score based on their support in the extracted knowledge base. A support for a rule is defined as the number of distinct subject and object pairs in the head of all the instantiations that appear in the knowledge base. The confidence score is calculated as the ratio of the support of the rule to the number of all the known true facts together with the assumed false facts in the extracted knowledge graph.

A Probabilistic Soft Logic (PSL) model [22] is defined based on the rules from the frequent items that includes the input set of rules along with the predicted triples. PSL minimizes a Markov Hinge-Loss function [3] that uses the input triples and their confidence scores to infer new facts along with their probabilistic confidence. PSL forms a probability distribution over all the interpretations/facts possible out of the derived and extracted facts and then infers the “most likely” facts. The task of “most likely explanation” inference corresponds to finding the confidence of each fact in the knowledge graph that maximizes the probability distribution over the derived facts. Confidence scores of facts endorsed by multiple rules are amplified, thus reinforcing the correct triples in the knowledge graph.

4 Evaluation of CultKB

We extracted $847,547$ facts from the structured data input of $90,193$ artifacts. The canonicalization of triples from the unstructured data to the YAGO taxonomy yielded 3615 more facts. The unmapped relations from the above step went through the clustering phase and gave us further $147,176$ facts and added $5,502$ new relationships to the taxonomy. Finally, the enrichment phase added $408,752$ more facts to the knowledge base summing up to an overall of $1,407,090$ facts.

In the absence of a gold standard dataset, we test the correctness of the facts in the knowledge base via human annotations from Amazon Mechanical Turk (AMT). Each AMT worker was presented with a text snippet from the Museums of India dataset to evaluate the correctness of the facts extracted from them. Each worker annotates 3 facts extracted from the presented passage, with one of the subject, predicate or object missing. The worker is tasked with selecting the appropriate option for the missing part from a list of options while ensuring the correctness of the completed fact. Occasionally, none of the options may correspond to the correct fact. We, therefore, allowed the AMT worker to opt for “None of the above” for such cases. Note that we are not evaluating the correctness of facts itself, but only the “correctness of the facts” as present in the text. Such an evaluation technique allows us to evaluate the correctness of the facts extracted by the algorithm, as well as establish ground truth for future experiments.

Each fact or triple is evaluated by 3 annotators. We used the Cohen's Kappa score to check for inter-annotator agreement which measures the agreement between categorical options, while simultaneously accounting for agreement by chance. Hence it is more robust than simple percentage calculation. We simulated two annotators by randomly selecting 2 turkers for each fact and calculating the agreement between them. This is repeated for $1,000$ iterations and we report on the median of these iterations. We obtain a Cohen's Kappa score of 0.763 with 95% confidence interval of 0.0455 indicating high inter-annotator agreement. More details about the evaluation is provided in the supplementary material.

Table 2. Accuracy of facts in CultKB. We also report on the Wilson interval for $\alpha = 5\%$ to ensure that the accuracy values are significant.

Stage	Accuracy-Interval
YAGO Canonicalized	$63.03\% \pm 18.15\%$
Sequential Clustering	$82.16\% \pm 6.18\%$
Overall after Enrichment	$75.50\% \pm 6.67\%$

Table 2 shows the accuracy of CultKB facts extracted. The accuracy of the facts canonicalized to YAGO (where both the predicates and the entities are canonicalized) are lower than the rest but is reasonable at 63.03% indicating that the canonicalization to a taxonomy is fruitful when the entire triple can be canonicalized.

The accuracy increases when the predicates are enriched using the clustering technique and this further establishes the need for building a base taxonomy to the needs of the cultural artifacts. The higher accuracy also justifies the ability of the proposed approach to introduce the culture specific predicates thus addressing the inadequacies of the standard taxonomy.

An overall accuracy of $75.50\% \pm 6.67\%$ of the facts is comparable to that of DBpedia [2] (81% [25]) built from a cleaner and more structured source establishing the integrity of the constructed knowledge base.

Exploring CultKB: Table. 3 shows the frequency of facts with a given relationship for the top 20 relationships in CultKB; note that there exists a long tail of relations with lower frequency counts. This count varies from as high as $1,578$ for relation "painted" to as low as 69 for the relation "created". The relations 'painted', 'is Fragment of', 'painted from', 'is decorated with', 'has depicted a portrait of' reflect the facts around the intrinsic details of an artifact itself, while relations like 'studied art at', 'visited', 'created', 'belongs to' reflect the information about the artist involved.

The constructed knowledge graph facilitates a **navigation** through the various artifacts of the Museums of India and allows to hop between different artifacts sharing the same facets. Figure 4 shows such a sub graph of CultKB.

The artifacts are labelled with their corresponding title. The labels on edges are the relationships between nodes. We can visualize information such as 'placedIn',

Table 3. Distribution of different relationships in CultKB

Predicate	Count	Predicate	Count
painted	1578	belongs to	600
is Fragment of	544	is written in	497
placedIn	483	depicts	466
studied art at	376	visited	201
is Head of	194	is A handle of	158
consists of	131	was born in	131
has studied	129	is written by	126
painted from	123	is decorated with	106
is Drawing of	100	is seated in	88
is seated on	86	has depicted a portrait of	74

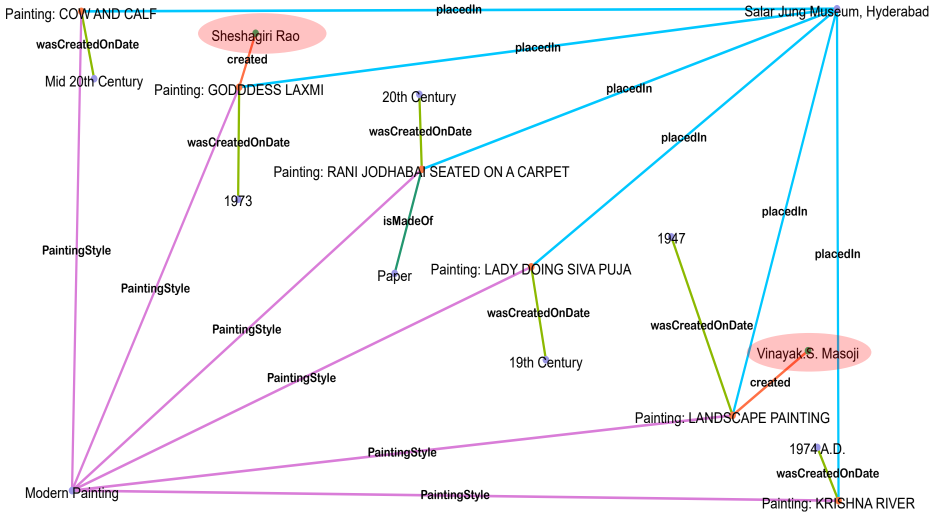


Fig. 4. Random subgraph from CultKB

'created', 'wasCreatedOnDate' for an artifact as well as can also see how different artifacts are related to each other. It is easy to see that the artifacts are placed in the 'Salar Jung Museum, Hyderabad'. Note that such a navigation is much richer than the one proposed in PATHS [1] since PATHS connects related artifacts without providing any reason for connections. But our knowledge graph representation naturally allows for a deeper artifact navigation experience.

A combination of the navigation experience and the retrievability of the organized data in CultKB allows for interesting knowledge discovery from the data. For example, the path between two artists 'Sheshagiri Rao' and 'Vinayak.S. Masoji' in Fig. 4, whose paintings are housed in Salar Jung Museum, reveals that the two painters are part of the school of "Modern Paintings". Such connections are impossible without an organized representation like CultKB.

The knowledge graph also aids in easy accessibility of facts in the original data. Recall the example in Fig. 1, where a query on "tempera images by jamini" on the Museums of India portal yields irrelevant results (Fig. 1(a)). The structured knowledge representation in CultKB facilitated the results via a "path query" that connects the entities 'Jamini Roy' and 'Tempera Images' in the graph yielding the result in Fig. 1(b) which shows that there are four paintings by 'Jamini Roy' in the tempera style. Note that algorithms to understand and serve such queries are beyond the scope of this paper, but CultKB can aid in serving such queries.

5 Conclusion

We studied the problems with the accessibility of cultural heritage artifacts and proposed a novel approach to construct a knowledge base for the artifacts in Museums of India. The need for such a domain-specific knowledge base is justified due to the lack of facts supporting Indian cultural artifacts present in global knowledge bases like YAGO [24], DBpedia [2]. Evaluation of the constructed knowledge base with human annotators showed acceptable accuracy along the scales of existing knowledge bases. The structured knowledge graph thus obtained facilitates both knowledge discovery and enhanced retrieval of the cultural artifacts. Although, we had applied the proposed approach to the domain of cultural artifacts, the approach is generic and can be easily extended to other domains as well.

References

1. Agirre, E., Aletras, N., Clough, P.D., Fernando, S., Goodale, P., Hall, M.M., Soroa, A., Stevenson, M.: Paths: A system for accessing cultural heritage collections. In: ACL (Conference System Demonstrations). pp. 151–156 (2013)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The Semantic Web pp. 722–735 (2007)
3. Bach, S.H., Broecheler, M., Huang, B., Getoor, L.: Hinge-loss markov random fields and probabilistic soft logic. arXiv preprint arXiv:1505.04406 (2015)
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI. vol. 7, pp. 2670–2676 (2007)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM (2008)
6. British Museum, W.: `http://www.britishmuseum.org/`
7. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI. vol. 5, p. 3 (2010)
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD (1996)
9. Europeana Museums, W.: `https://www.europeana.eu/portal/en`
10. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. Association for Computational Linguistics (2011)

11. Fernando, S., Stevenson, M.: Adapting wikification to cultural heritage. In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 101–106. Association for Computational Linguistics (2012)
12. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005)
13. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: Amie: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd international conference on World Wide Web. pp. 413–422. ACM (2013)
14. Museums of India, W.: `http://museumsofindia.gov.in`
15. Jay Pujara, Sameer Singh, B.D.: Knowledge graph construction from text. In: AAAI Tutorial (2017)
16. Kobren, A., Logan, T., Sampangi, S., McCallum, A.: Domain specific knowledge base construction via crowdsourcing. In: Neural Information Processing Systems Workshop on Automated Knowledge Base Construction AKBC, Montreal, Canada (2014)
17. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In: Proceedings of the fifteenth conference on computational natural language learning: Shared task. pp. 28–34. Association for Computational Linguistics (2011)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
20. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
21. Nakashole, N., Weikum, G., Suchanek, F.: Patty: A taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1135–1145. Association for Computational Linguistics (2012)
22. Pujara, J., Miao, H., Getoor, L., Cohen, W.: Knowledge graph identification. In: International Semantic Web Conference (ISWC) (2013), winner of Best Student Paper award
23. Pujara, J., Miao, H., Getoor, L., Cohen, W.W.: Using semantics and statistics to turn data into knowledge. AI Magazine 36(1), 65–74 (2015)
24. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
25. Zaveri, A., Kontokostas, D., Sherif, M.A., Bühmann, L., Morsey, M., Auer, S., Lehmann, J.: User-driven quality evaluation of dbpedia. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 97–104. ACM (2013)
26. Zhao, X., Xing, Z., Kabir, M.A., Sawada, N., Li, J., Lin, S.W.: Hdskg: Harvesting domain specific knowledge graph from content of webpages. In: Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on. pp. 56–67. IEEE (2017)