Harvesting Knowledge from Cultural Heritage Artifacts in the Museums of India

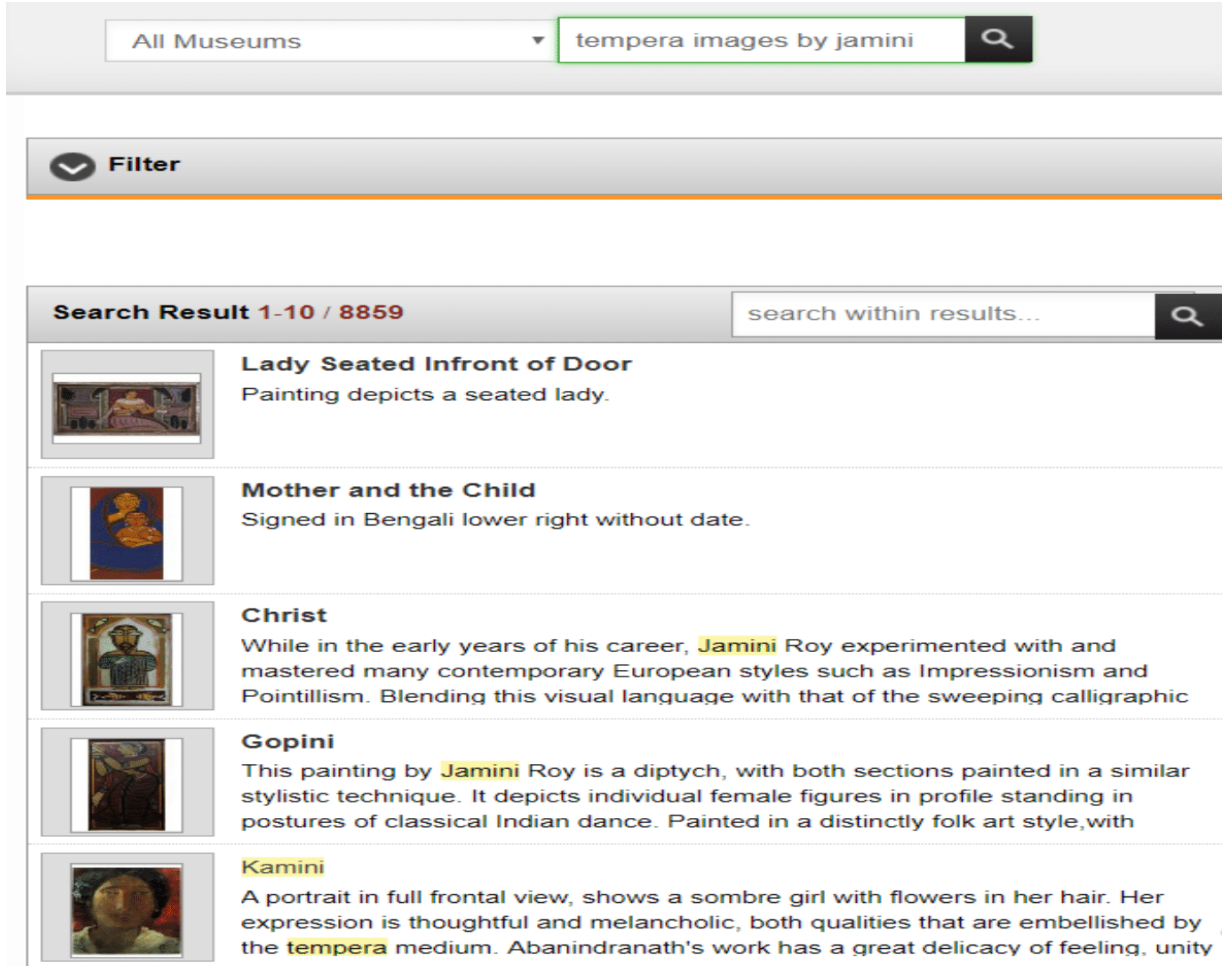
Abhilasha Sancheti, Paridhi Maheshwari, Rajat Chaturvedi, Anish V. Monsy, Tanya Goyal, Balaji Vasan Srinivasan

Takehiro Tobinaga

India's rich and diverse cultural heritage

- Recently several cultural artefacts have been digitized
  - E.g. Museums of India
- But …
  - Overwhelming digital content
  - Infeasible to interpret

When we look for information in "Museums Of India"



What can be done?

- Current organizations does not capture the specific style of painting.
- Standard information retrieval systems serve information from single source only

- Systematic approach to harvest knowledge is required
  - **Enhance the understanding**
  - **Facilitate organization**
  - **Better accessibility of facts**

Knowledge Graph Extraction

- Harvesting "knowledge" from structured and unstructured data sources
  - Eg. Dbpedia[1], NELL[2], YAGO[3]
- Knowledge base – triples of facts
  - Eg: < Nandalal_Bose > < alsoKnownAs > < Master_Moshai >
- Proposed solution to improve accessibility:
  - Build a knowledge graph for cultural artifacts
- Standard Taxonomies are insufficient to capture the facts in cultural artifacts

[1] 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)
[2] 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)
[3] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference onWorld WideWeb. pp. 697{706. ACM (2007)

Challenges

- Meta-data is not always in well formed text – can result in noisy facts
- Standard taxonomies are insufficient in canonicalizing facts from specific domain
- Inferring new and missing facts

Dataset

Structured and unstructured information about artifacts from "Museums Of India"

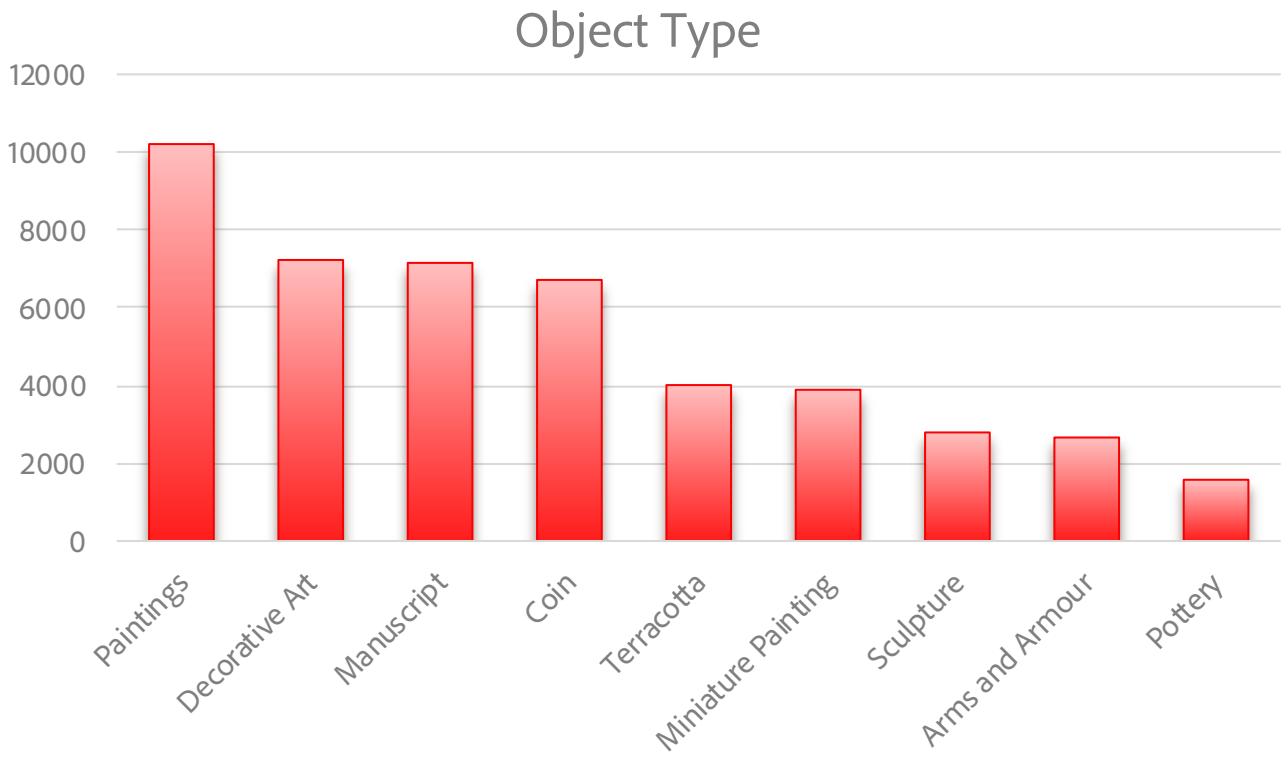
Title :	Akbar Holding Bird
Title2 :	Akbar Holding Bird
Museum Name :	Allahabad Museum, Allahabad
Gallery Name :	Decorative Art Gallery
Object Type :	Decorative Art
Main Material :	Ivory
Manufacturing Technique :	Cutting and Carving
Artist's Nationality :	Indian
Author :	NA
Country :	India

Structured Data

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.
Brief Description :	An image of the Akbar the Great has been carved in ivory.

Unstructured Data

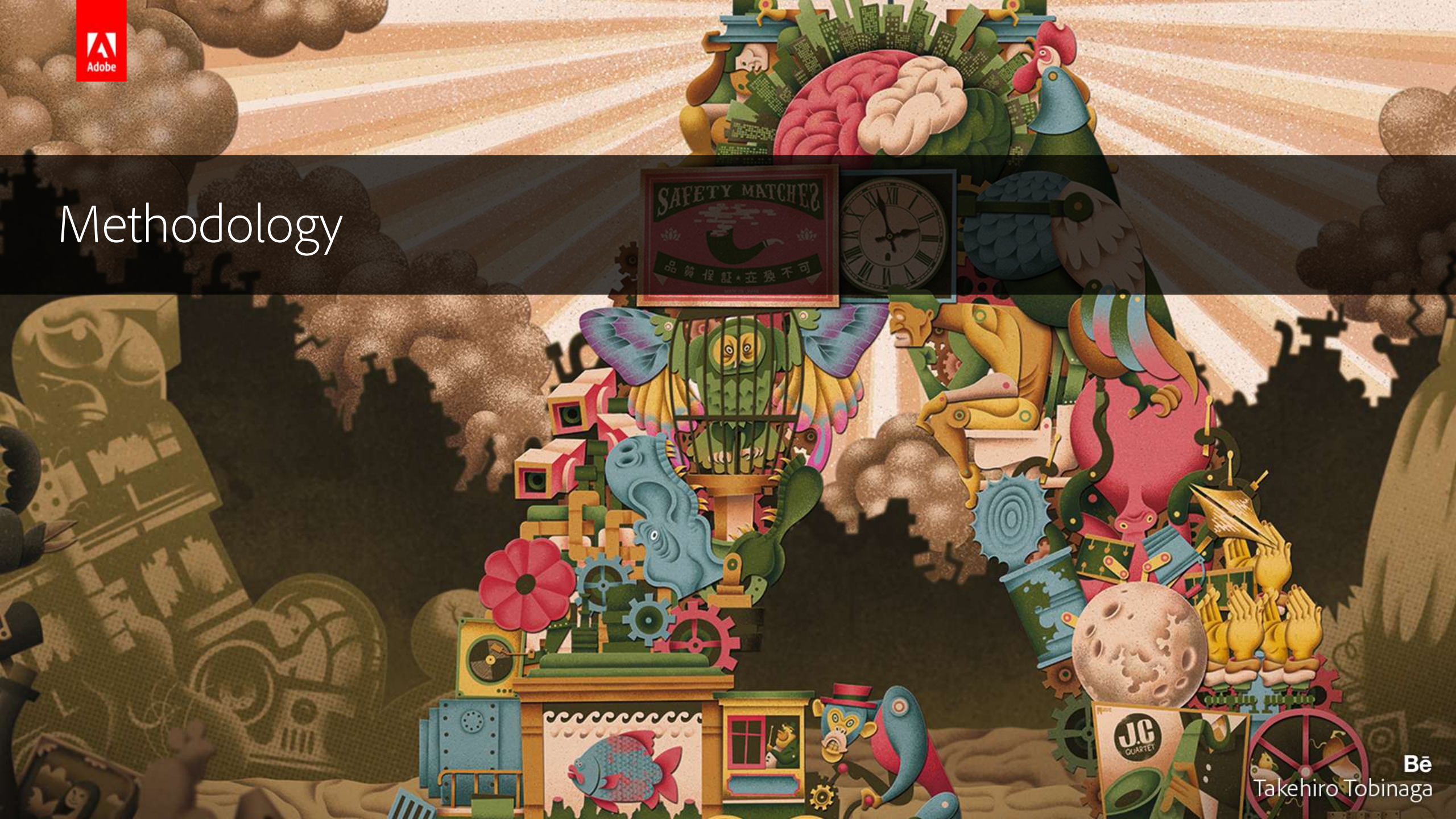
Dataset Statistics



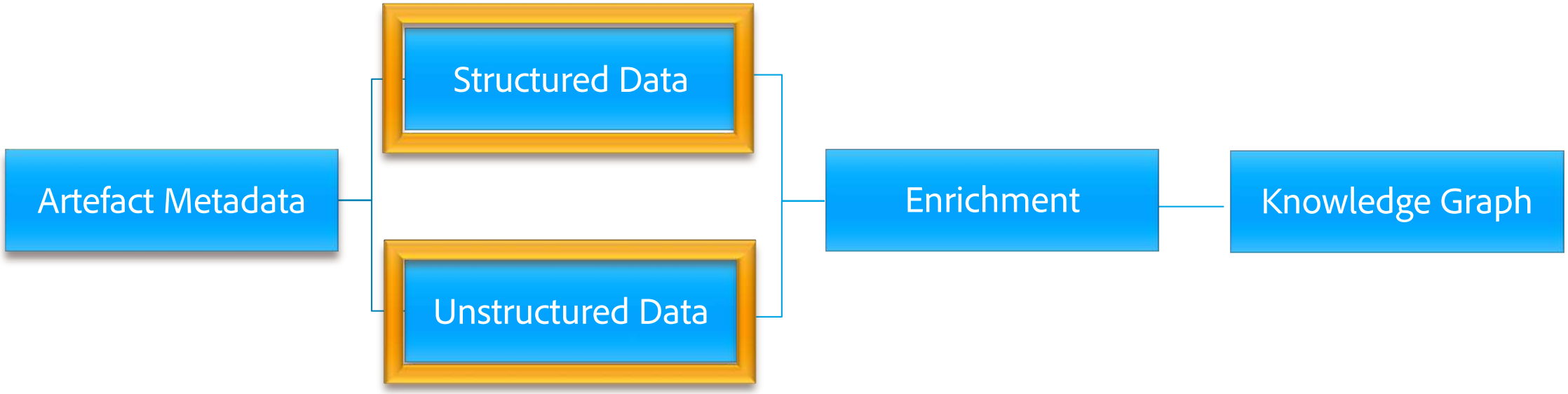
Counts

of museums: 10
# of artefacts: 90,193
# of object types: 258
# of distinct fields in structured data: 40
# of distinct fields in unstructured data: 3

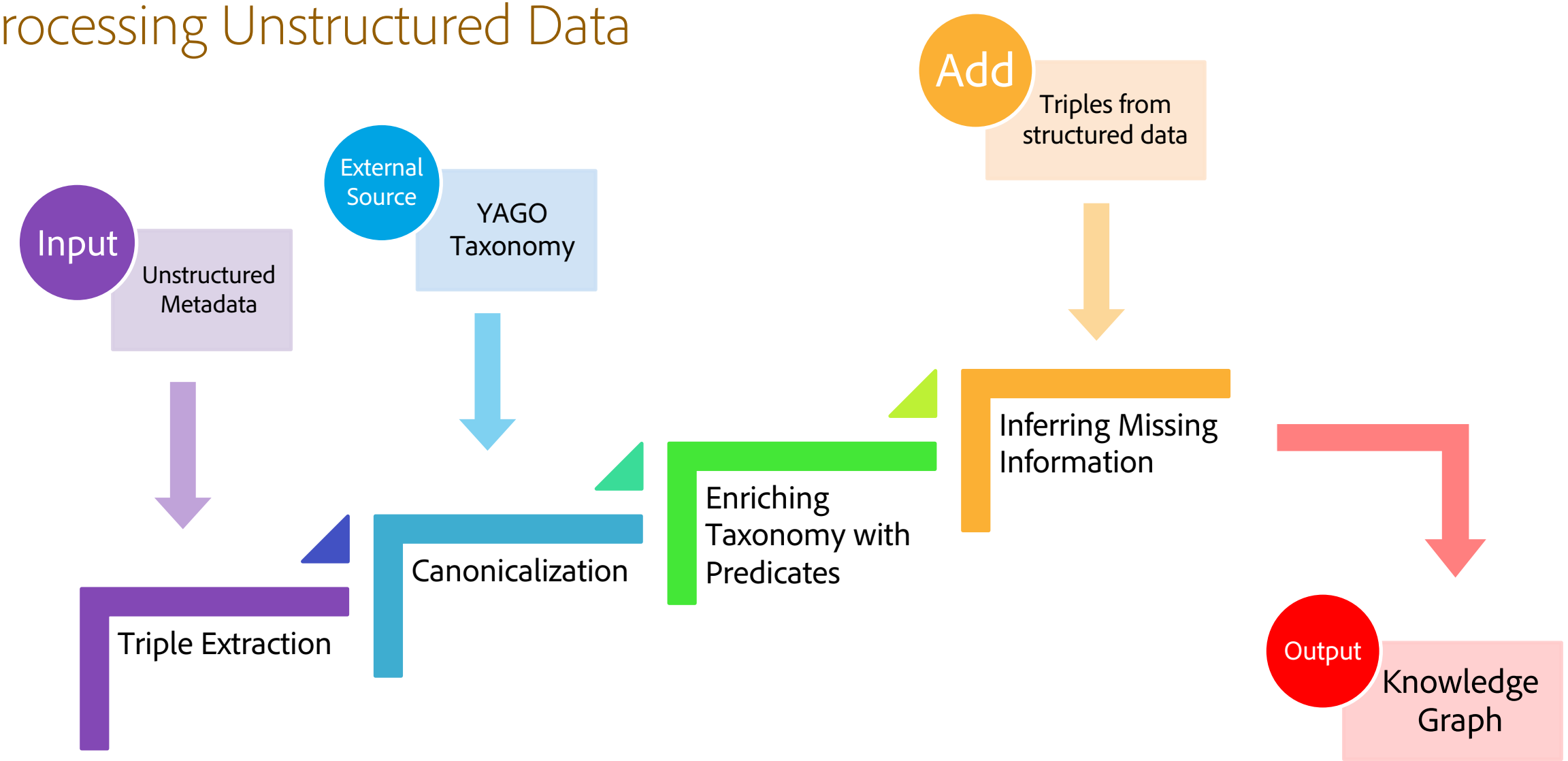
Methodology



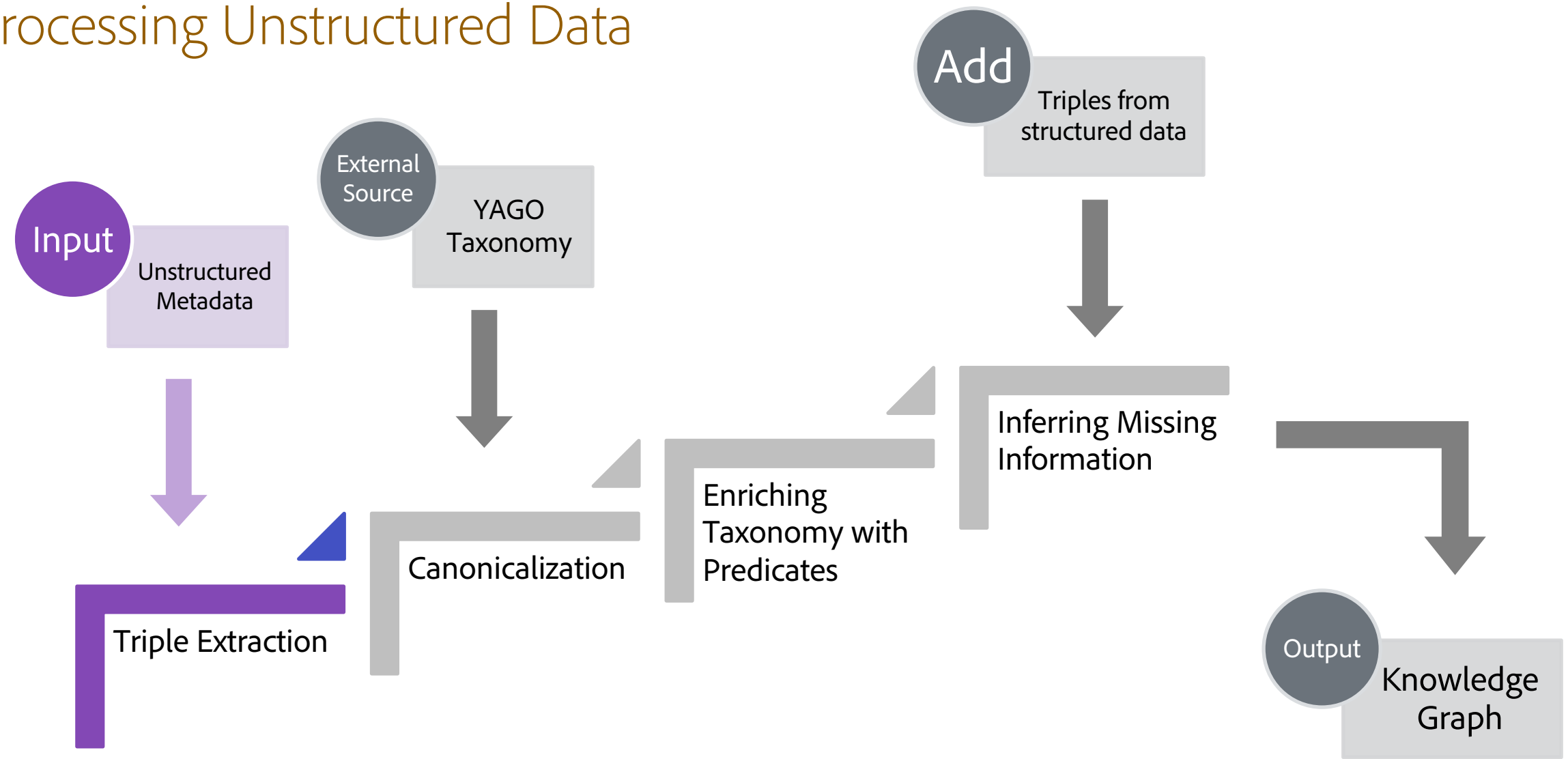
Approach: In a nutshell



Processing Unstructured Data



Processing Unstructured Data



Triple Extraction

Nandalal Bose had a close relationship with Gandhi and shared many of his ideals. He was the only artist patronized by the leader, who often insisted he had no time for art.

[1]

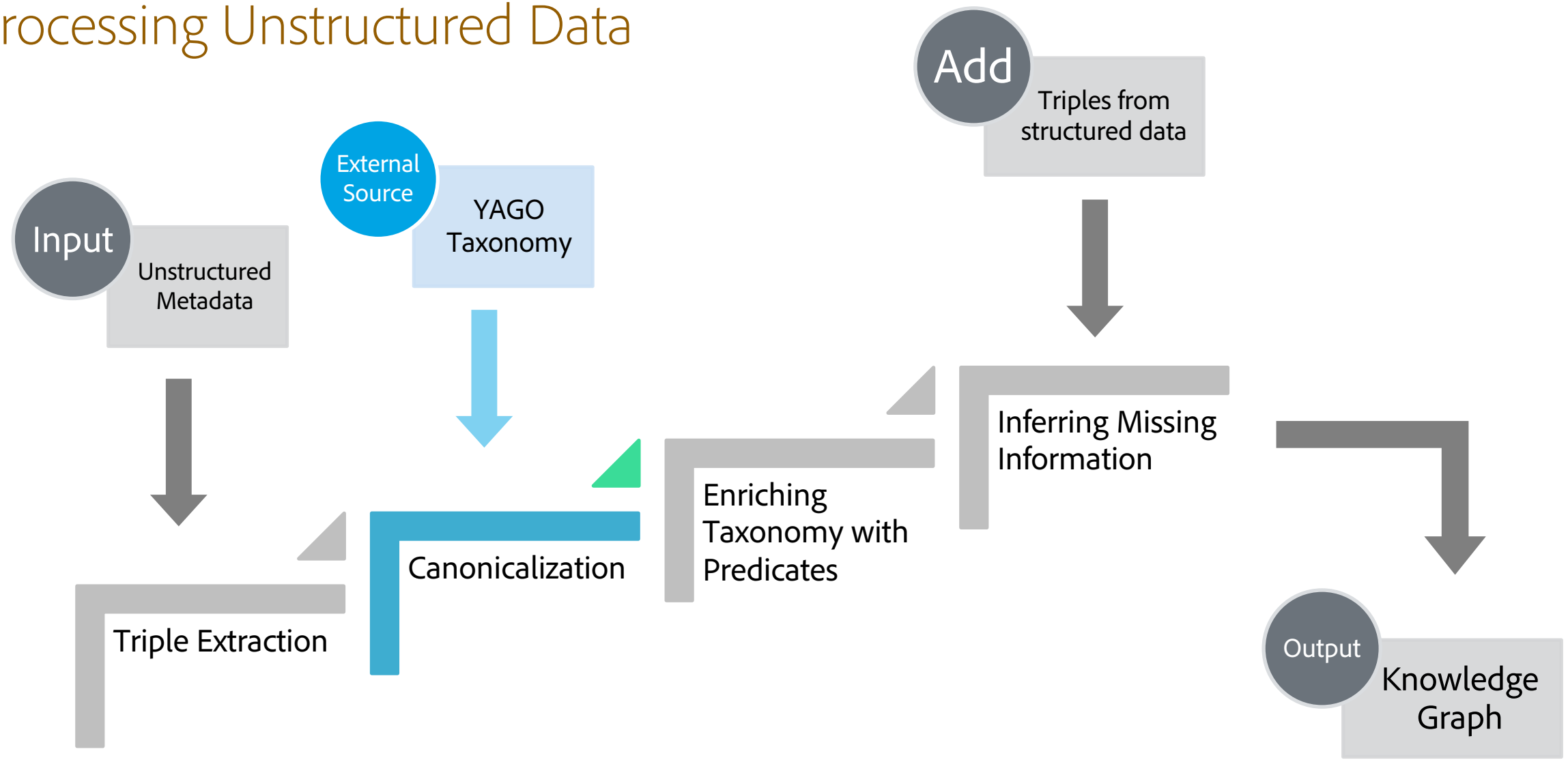
Nandalal Bose had a close relationship with Gandhi and shared many of Gandhi's ideals. Nandalal Bose was the only artist patronized by the leader, Gandhi often insisted Gandhi had no time for art.

[2]

SUBJECT	PREDICATE	OBJECT	CONFIDENCE SCORE
Nandalal Bose	had a close relationship with	Gandhi	0.940
Nandalal Bose	was	the only artist	0.815
Gandhi	often insisted	Gandhi	0.279
Gandhi	had no time for	art	0.749

[1] Raghunathan, Karthik, et al. "A multi-pass sieve for coreference resolution." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.
[2] Fader, Anthony, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

Processing Unstructured Data



Canonicalization

ENTITY MAPPING

Edit Distance based mapping

Nandalal Bose → **http://yago-knowledge.org/resource/Nandalal_Bose**

Indian Painting → **http://yago-knowledge.org/resource/Indian_Paintings**

RELATION MAPPING

Statistical Distance between Word-Set [1,2,3]

CONSTRAINT: Consistent NER tags for domain and range of relation

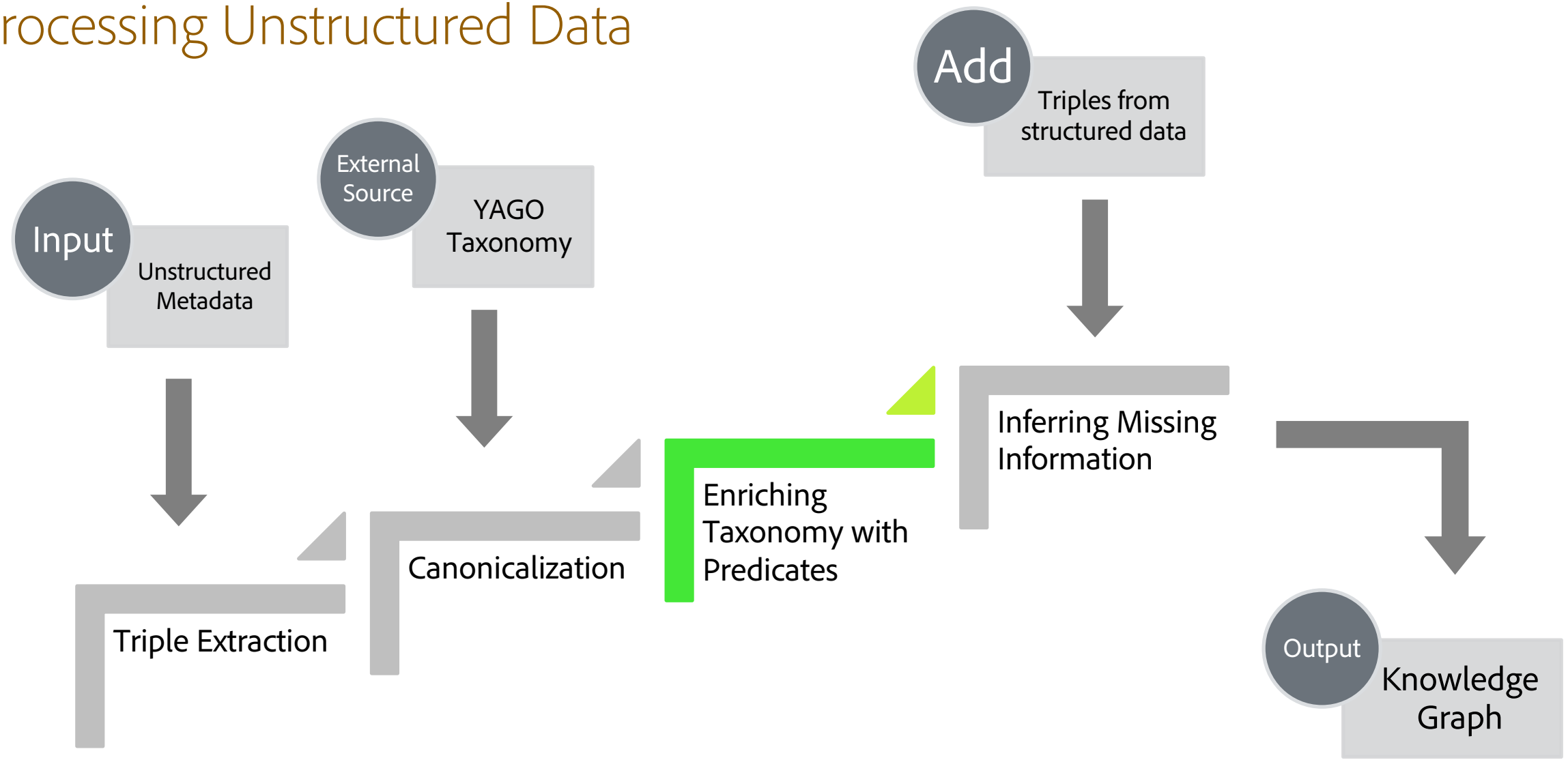
<PERSON> was born in <LOCATION> → **wasBornIn**

<PERSON> was born in <DATE> → **wasBornOnDate**

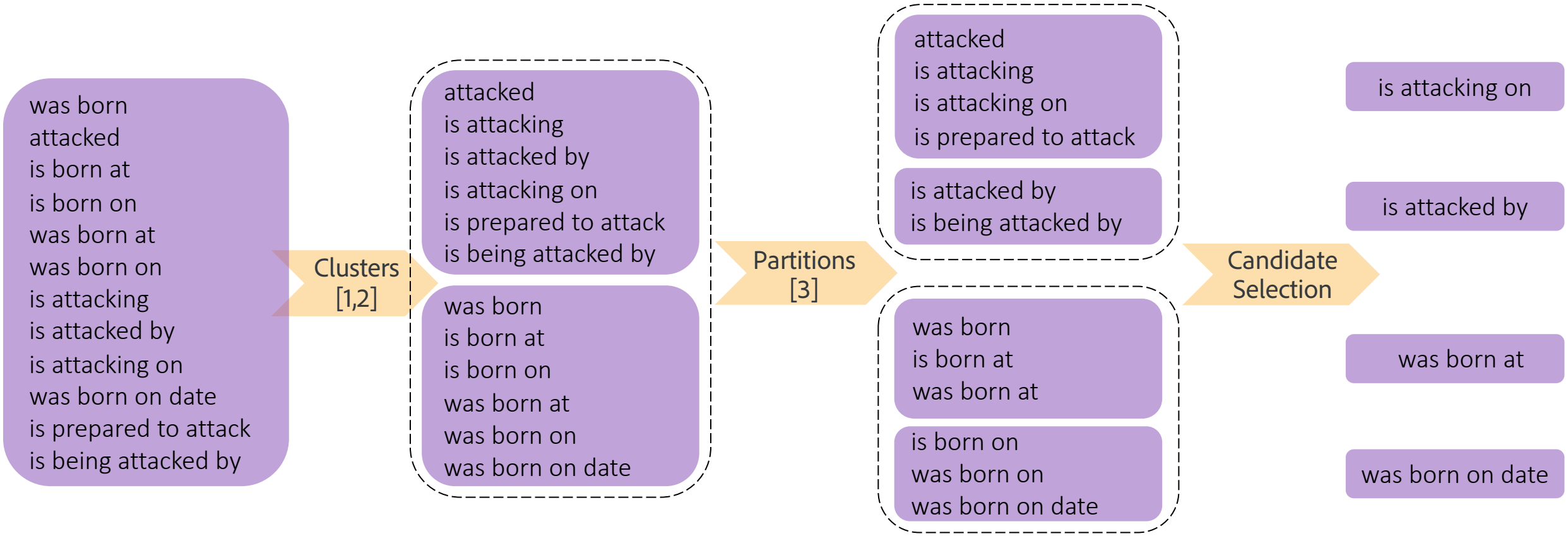
[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
[2] Miller, George A., et al. "Introduction to WordNet: An on-line lexical database." *International journal of lexicography* 3.4 (1990): 235-244.
[3] Nakashole, Ndapandula, Gerhard Weikum, and Fabian Suchanek. "PATTY: A taxonomy of relational patterns with semantic types." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

Adobe

Processing Unstructured Data



Relation Clustering

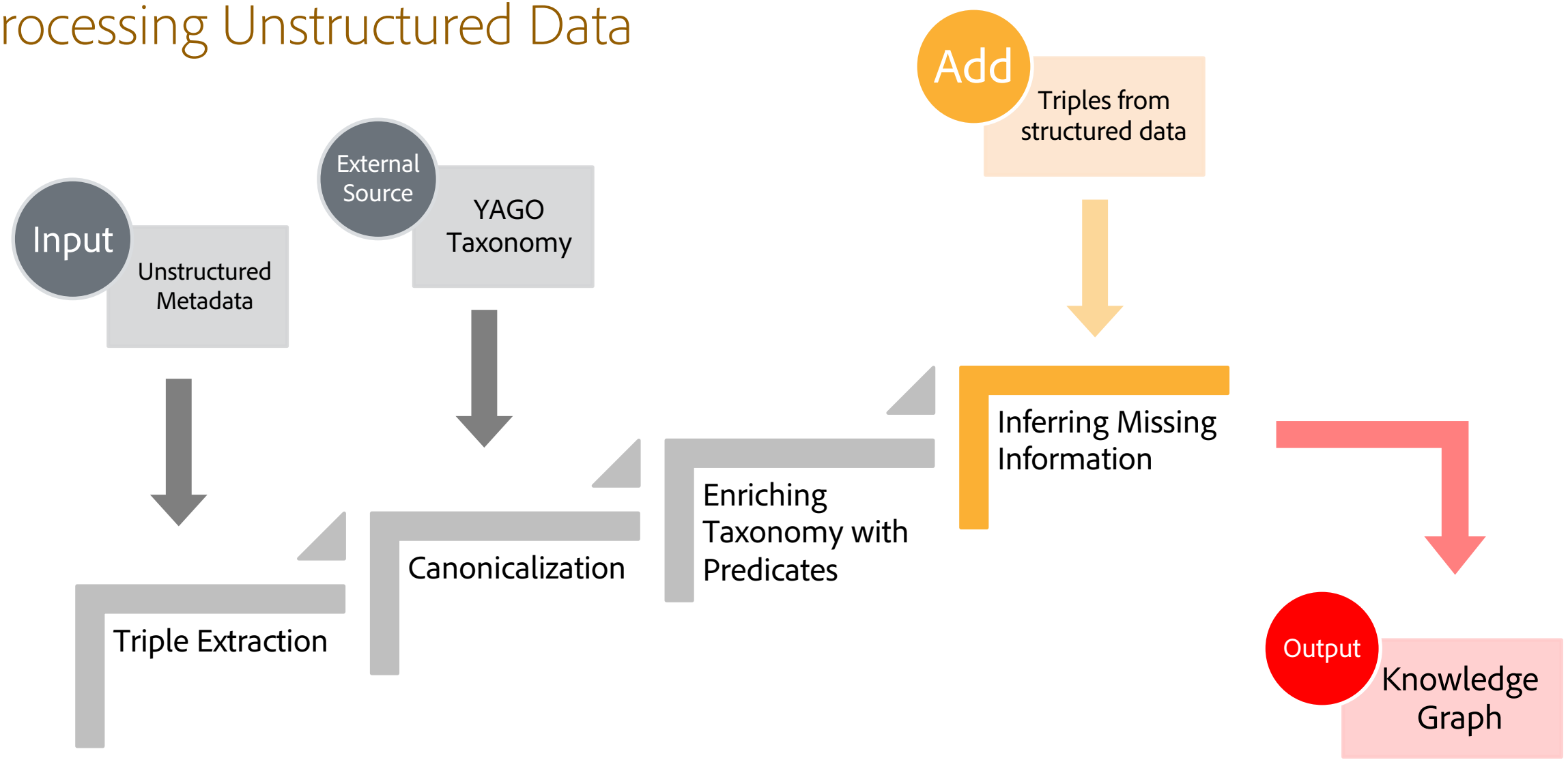


[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

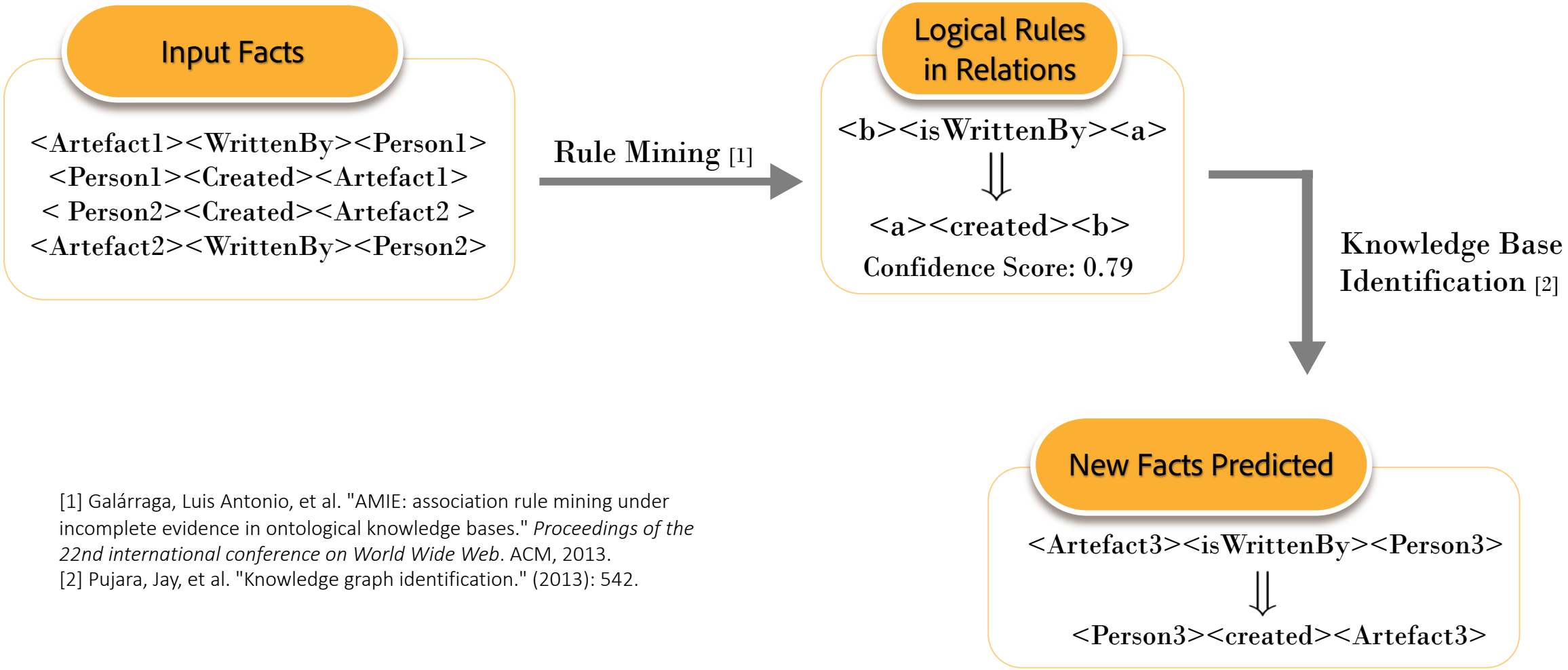
[2] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370.

Processing Unstructured Data

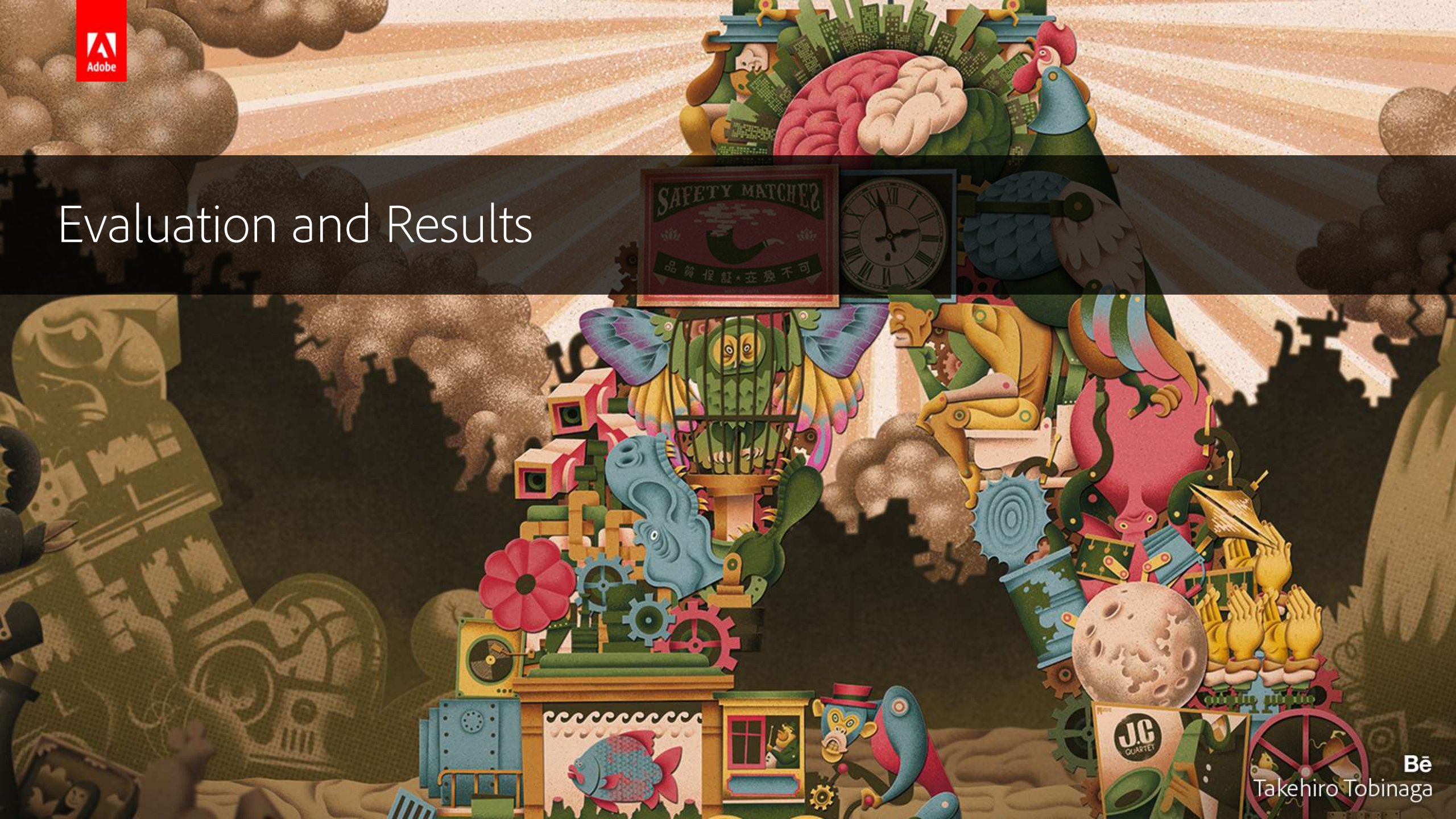


Enrichment



[1] Galárraga, Luis Antonio, et al. "AMIE: association rule mining under incomplete evidence in ontological knowledge bases." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.
[2] Pujara, Jay, et al. "Knowledge graph identification." (2013): 542.

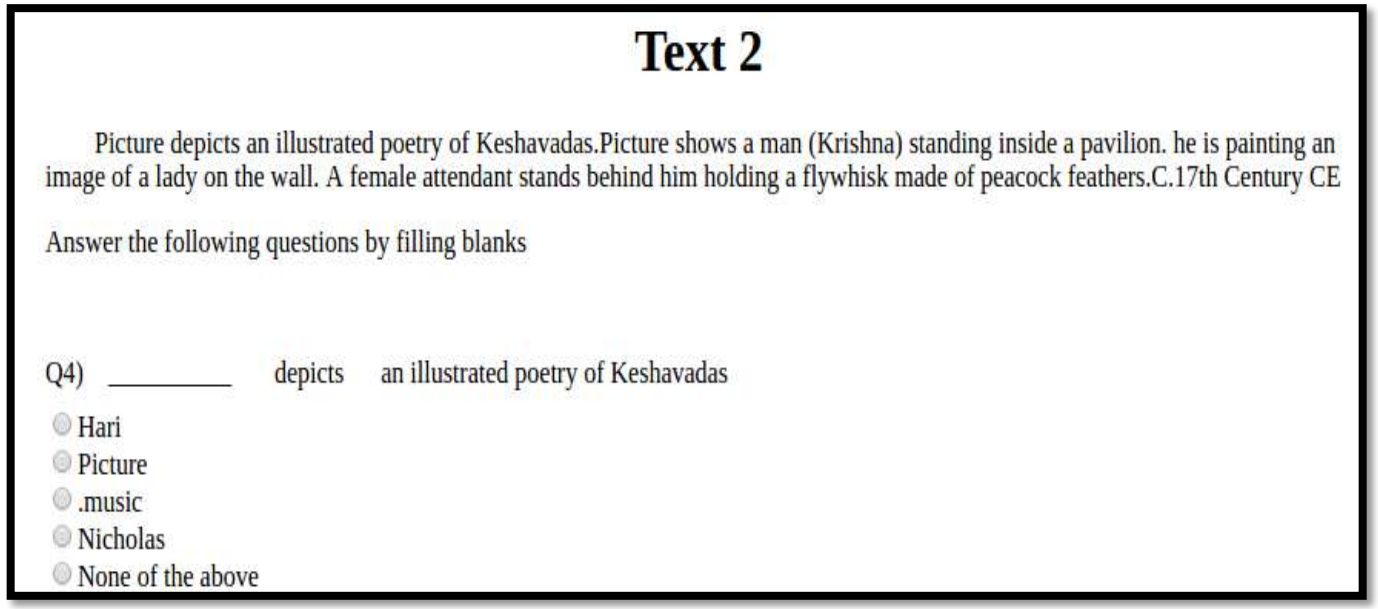
Evaluation and Results



Takehiro Tobinaga

Evaluation using Amazon MTurk

- Evaluated facts generated from the unstructured information of artifacts
- Each human evaluator evaluates information of 3 artifacts with 3 facts each
- Evaluation of a fact involves filling the blank with multiple choices provided.
- Each fact is checked by 3 evaluators



Text 2

Picture depicts an illustrated poetry of Keshavadas.Picture shows a man (Krishna) standing inside a pavilion. he is painting an image of a lady on the wall. A female attendant stands behind him holding a flywhisk made of peacock feathers.C.17th Century CE

Answer the following questions by filling blanks

Q4) _____ depicts an illustrated poetry of Keshavadas

- Hari
- Picture
- .music
- Nicholas
- None of the above

Evaluation Results

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

CultKB Exploration

Total number of Facts added to the Knowledge Graph - 1,407, 090

From structured text – 847,547

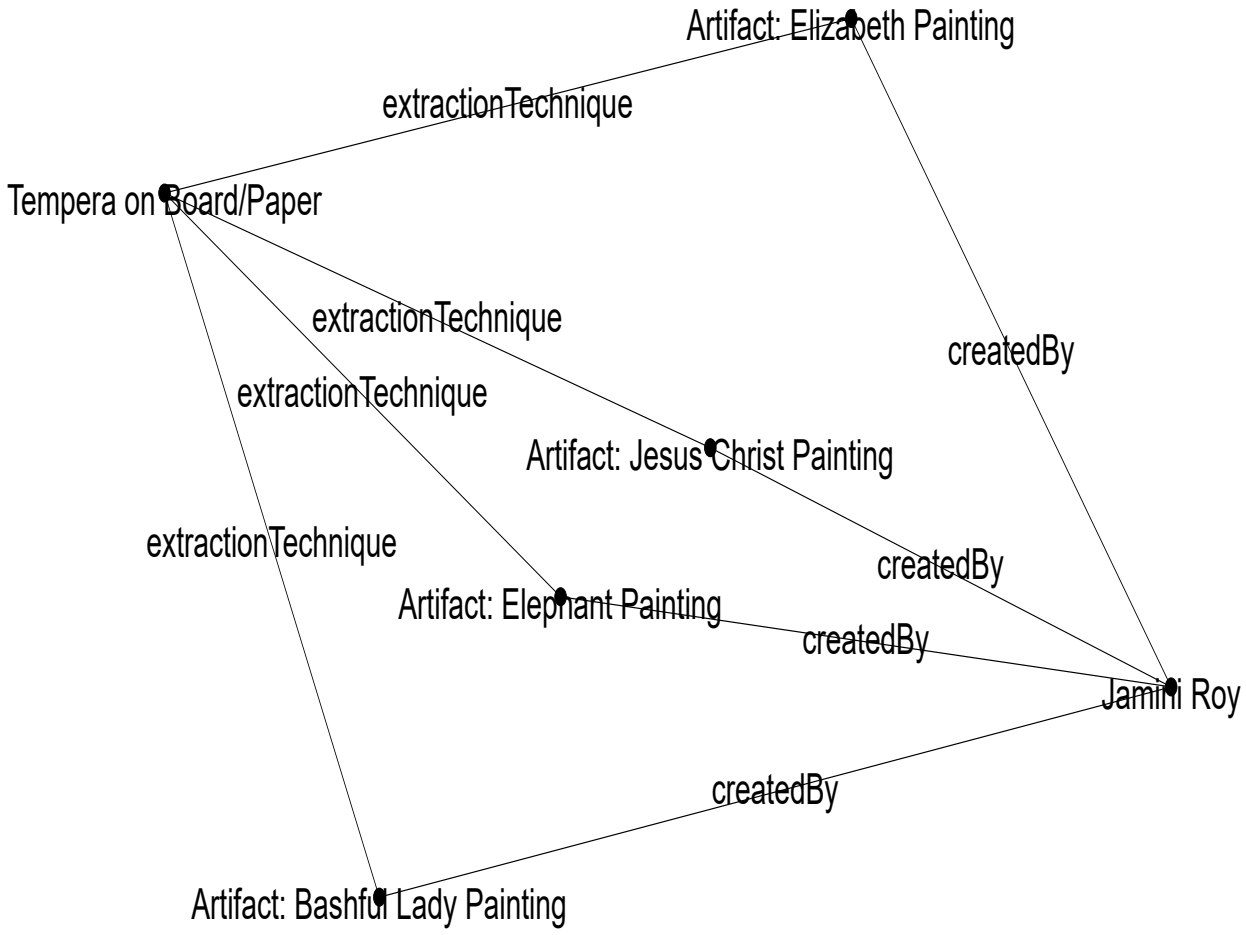
After canonicalization – 3615

After relation clustering – 147,176

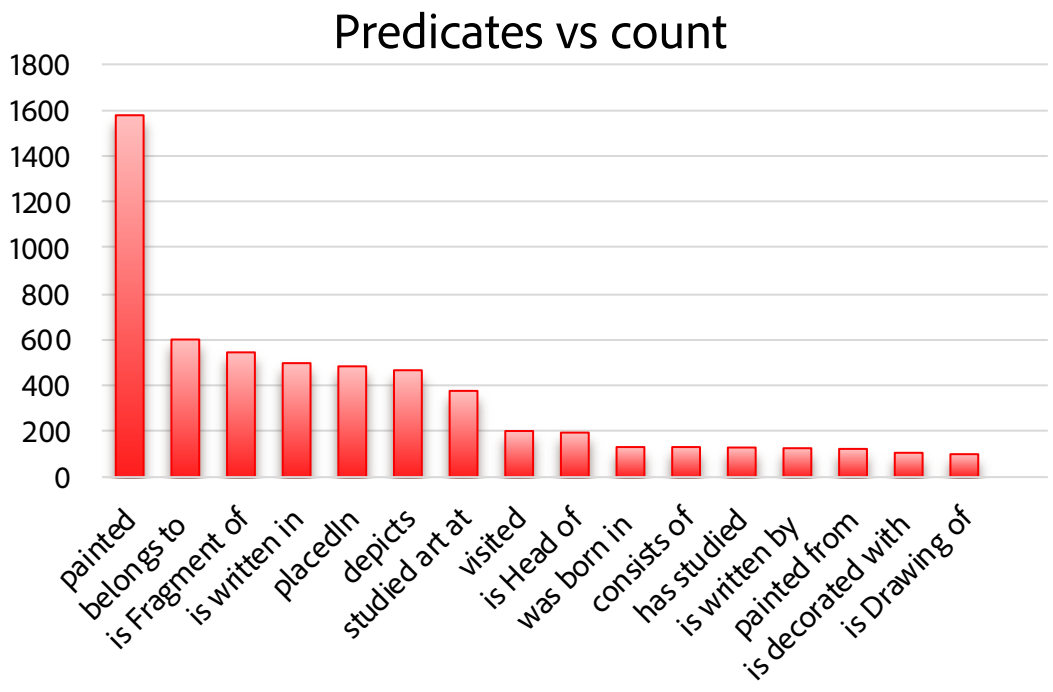
After enrichment – 408,753

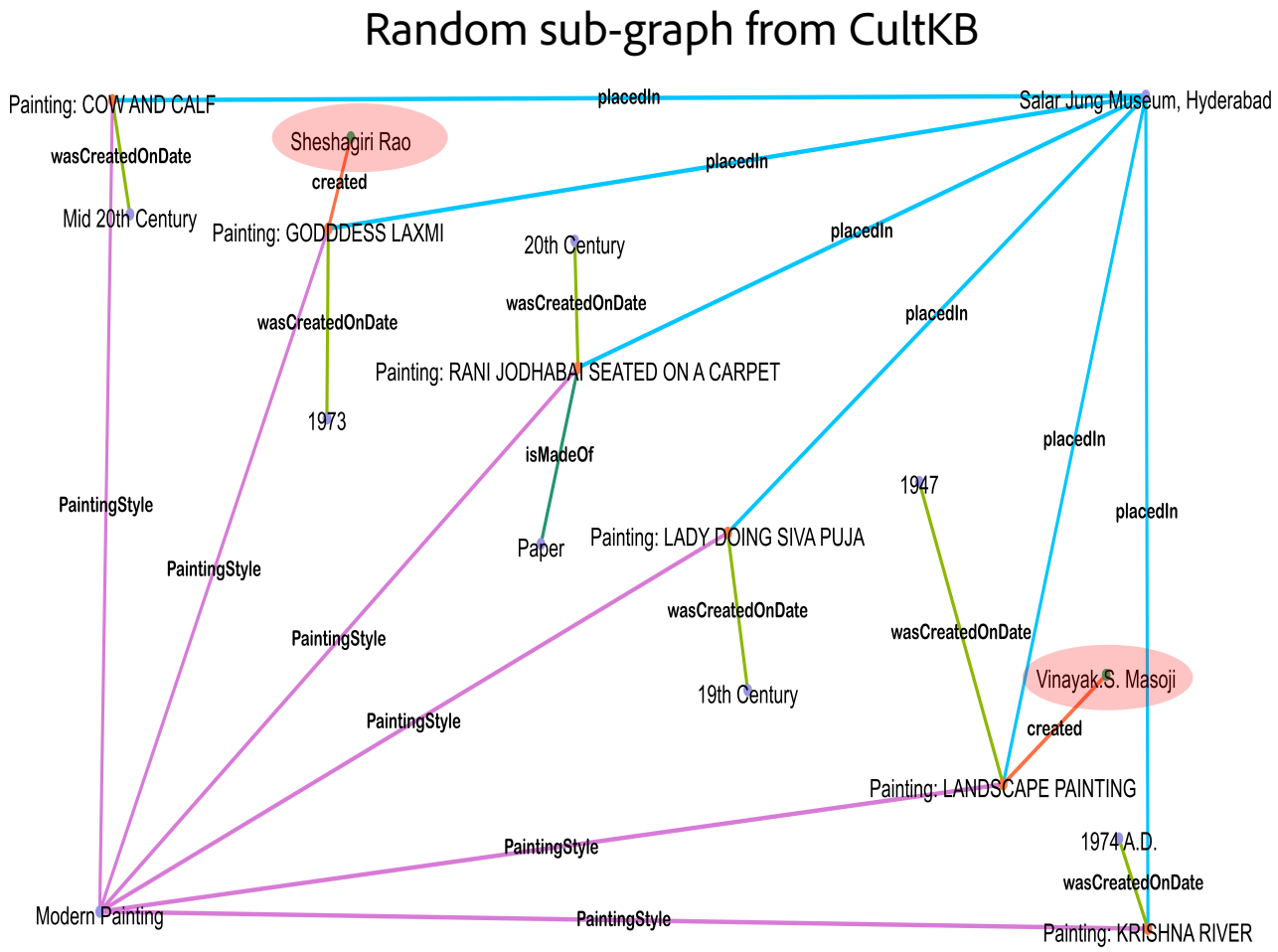
CultKB Exploration

Result of the query 'Tempera Images by Jamini' ' from CultKB



CultKB Exploration





Take Away

- Presented a technique to generate knowledge graph from a limited domain
- Initial results are promising
- Lesser studied space
  - Can be extended to organize knowledge of a low-resource domains

WE ARE HIRING!

adoberesearchjobs@adobe.com

MAKE IT AN EXPERIENCE