A Scalable Approach to Incrementally Building Knowledge Graphs

Gleb Gawriljuk (KIT), Andreas Harth (KIT), Craig A. Knoblock (USC), Pedro Szekely (USC)

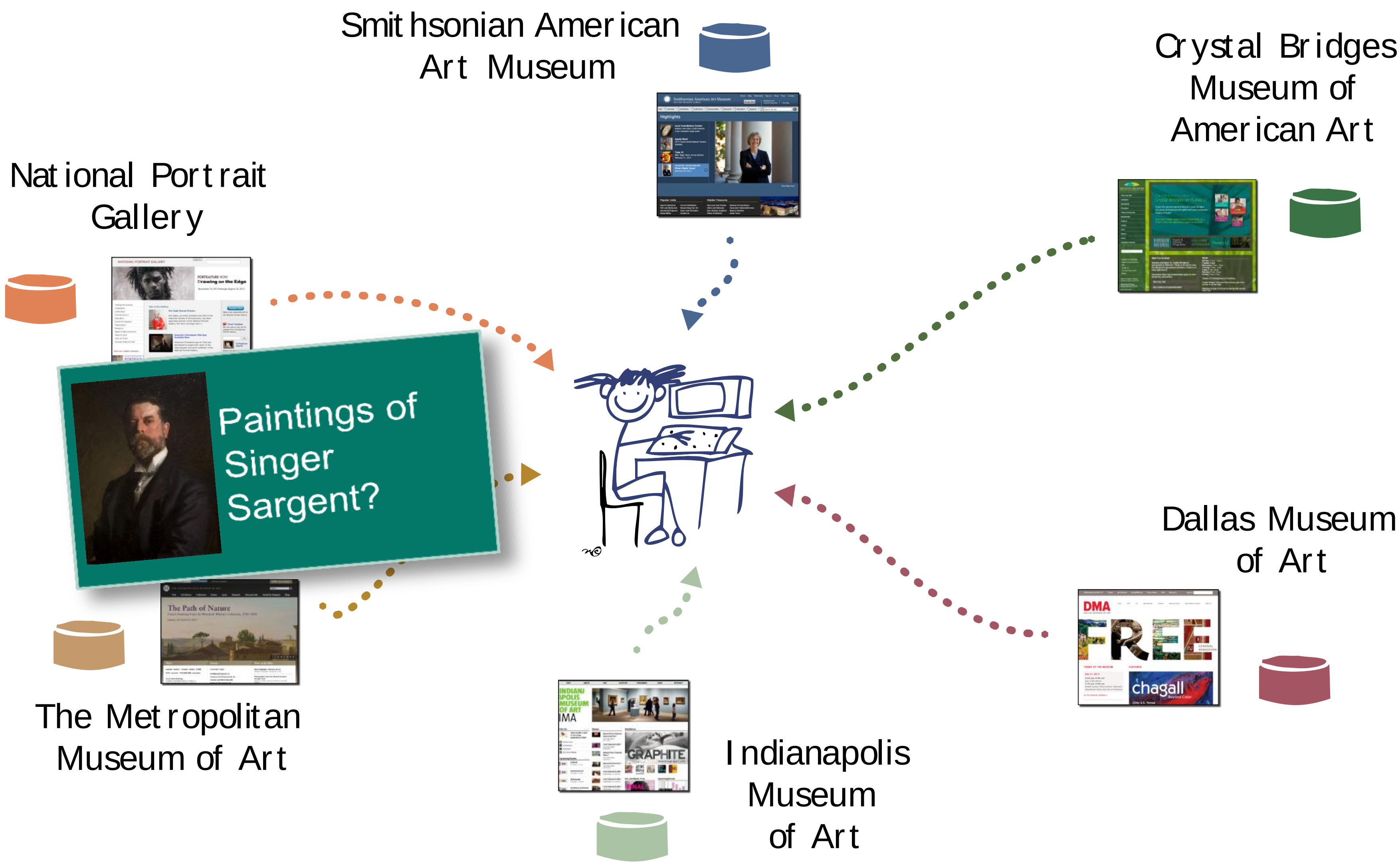
INSTITUTE AIFB, CHAIRS OF KNOWLEDGE MANAGEMENT AND WEB SCIENCE

http://www.imageduplicator.com/main.php?decade=70&year=79&work_id=1042

Outline

- **Motivation**
- Overview of Approach
- Building and Extending a Knowledge Graph
- Evaluation
- Conclusion

Current State of Cultural Heritage Data: Get Info from Web Pages



Problem

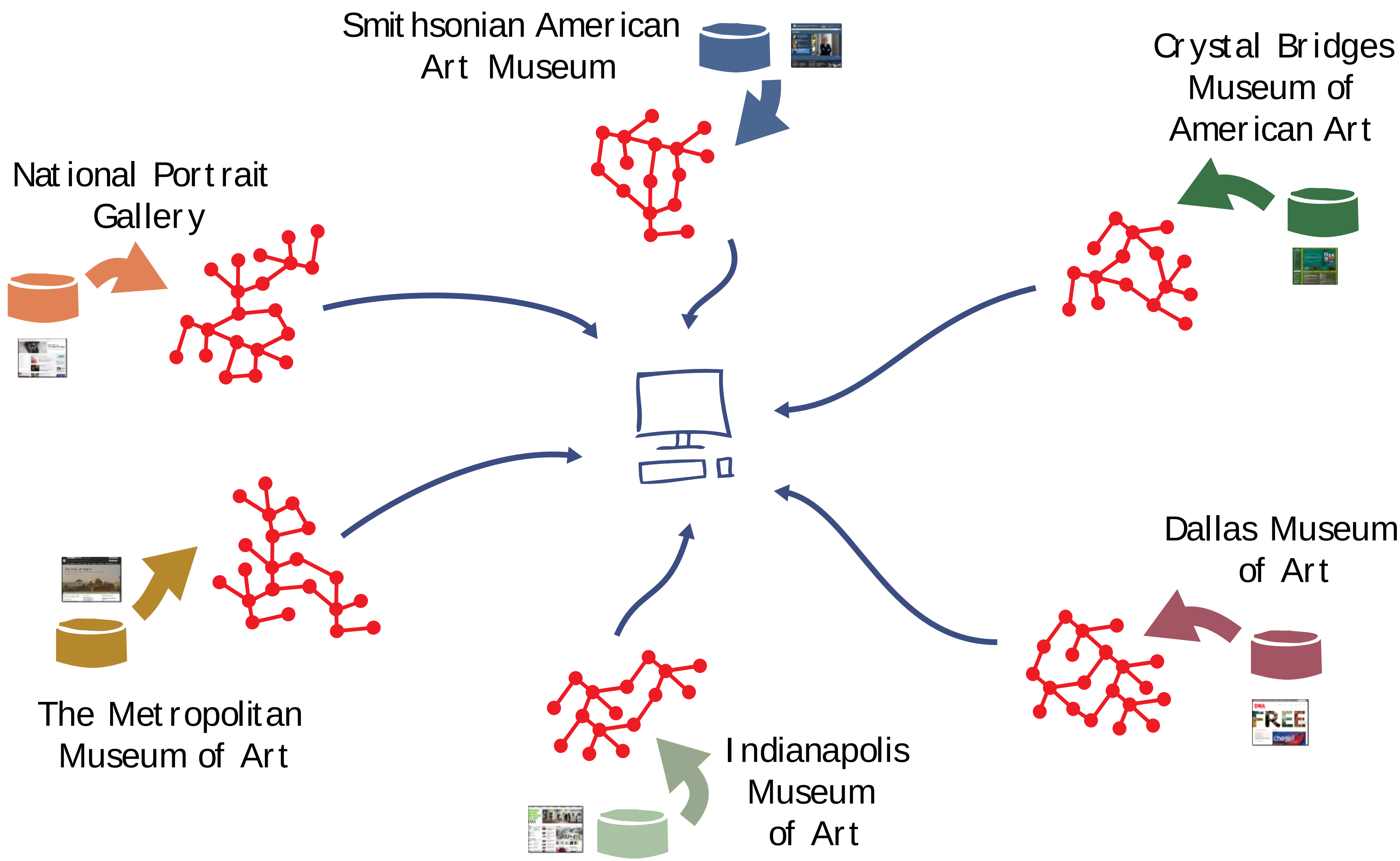
web pages are machine processable,
but not machine understandable

impractical for building applications using the data

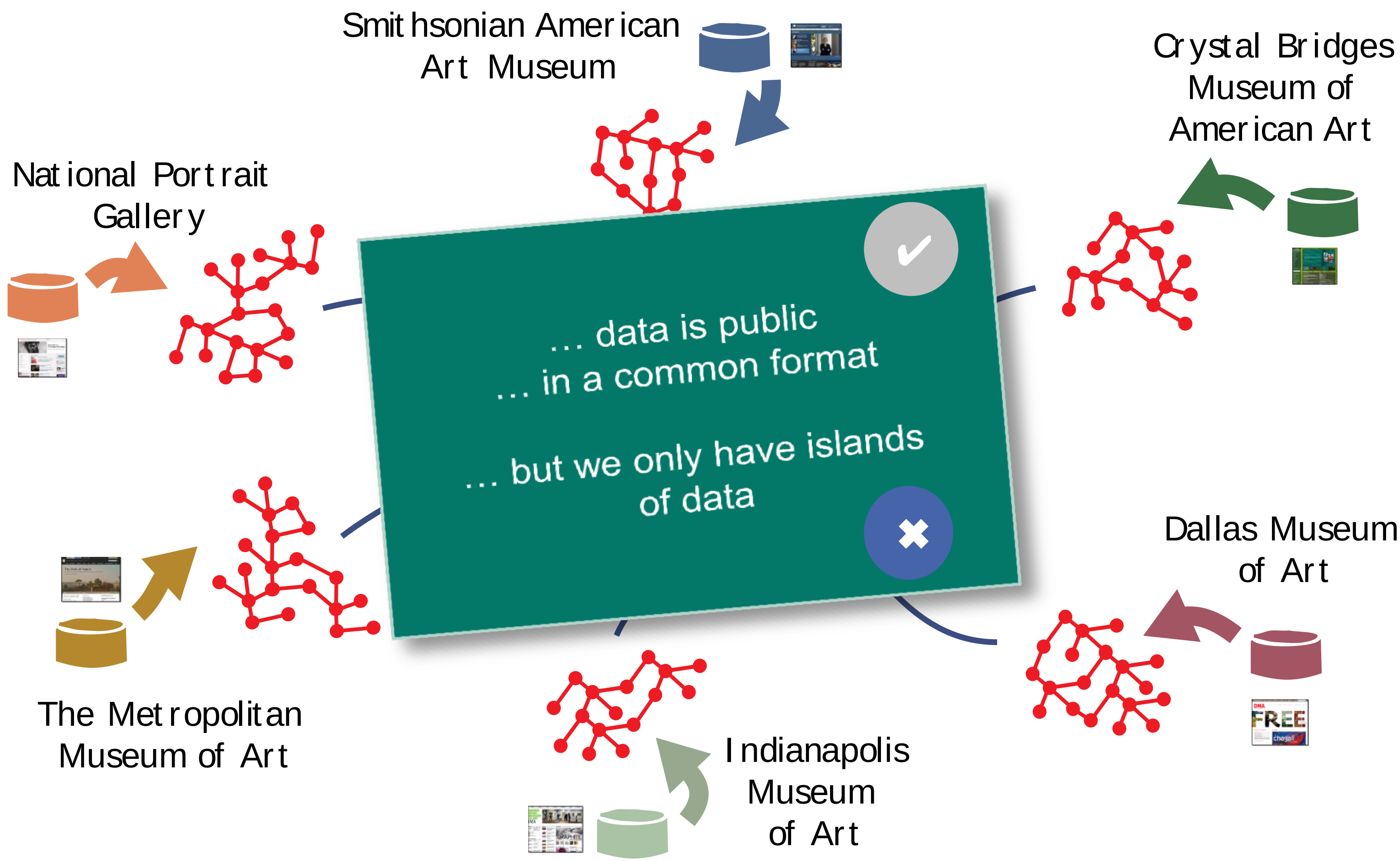
Solution

publish the data as Linked Open Data

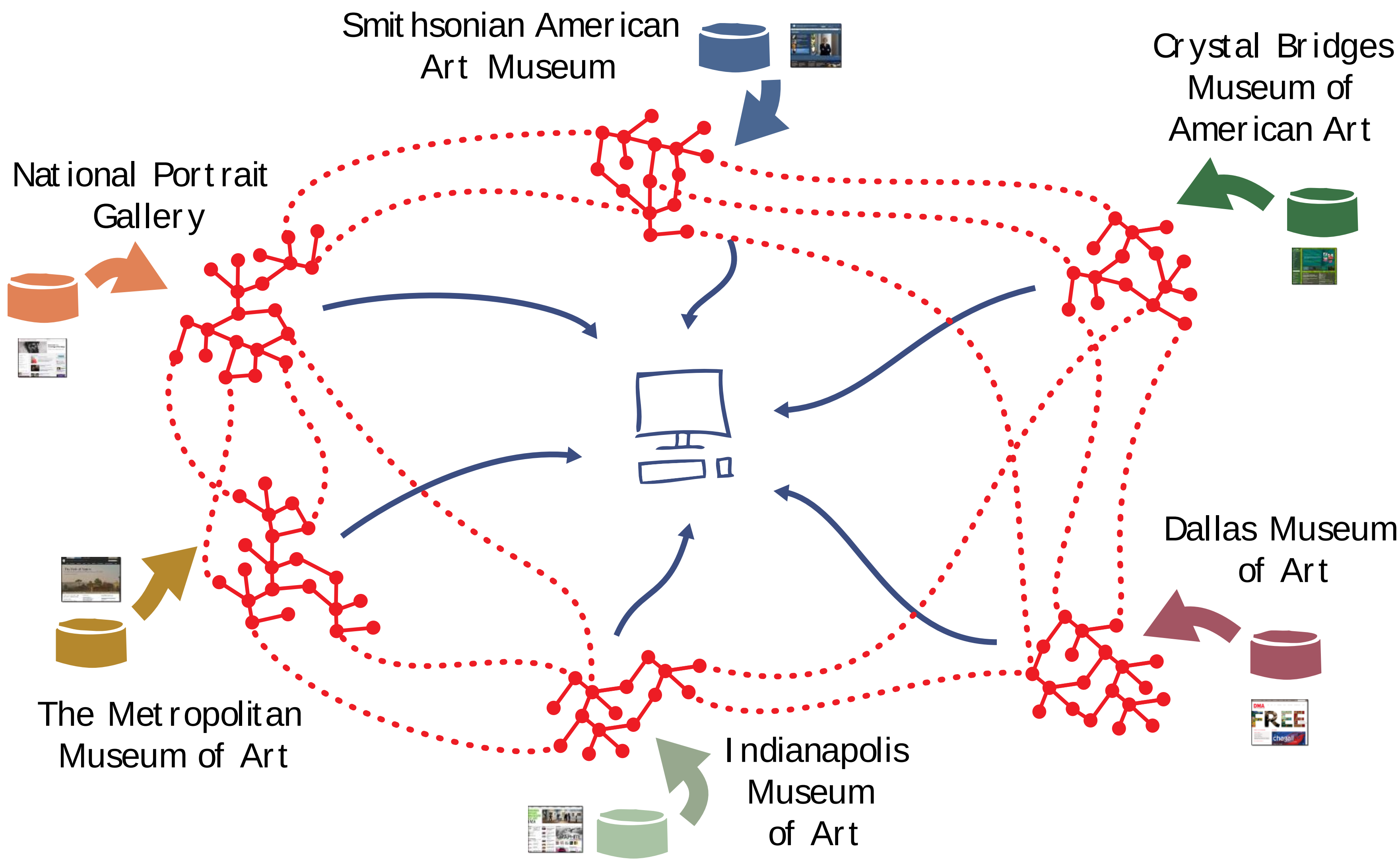
Cultural Heritage "Linked" Open Data



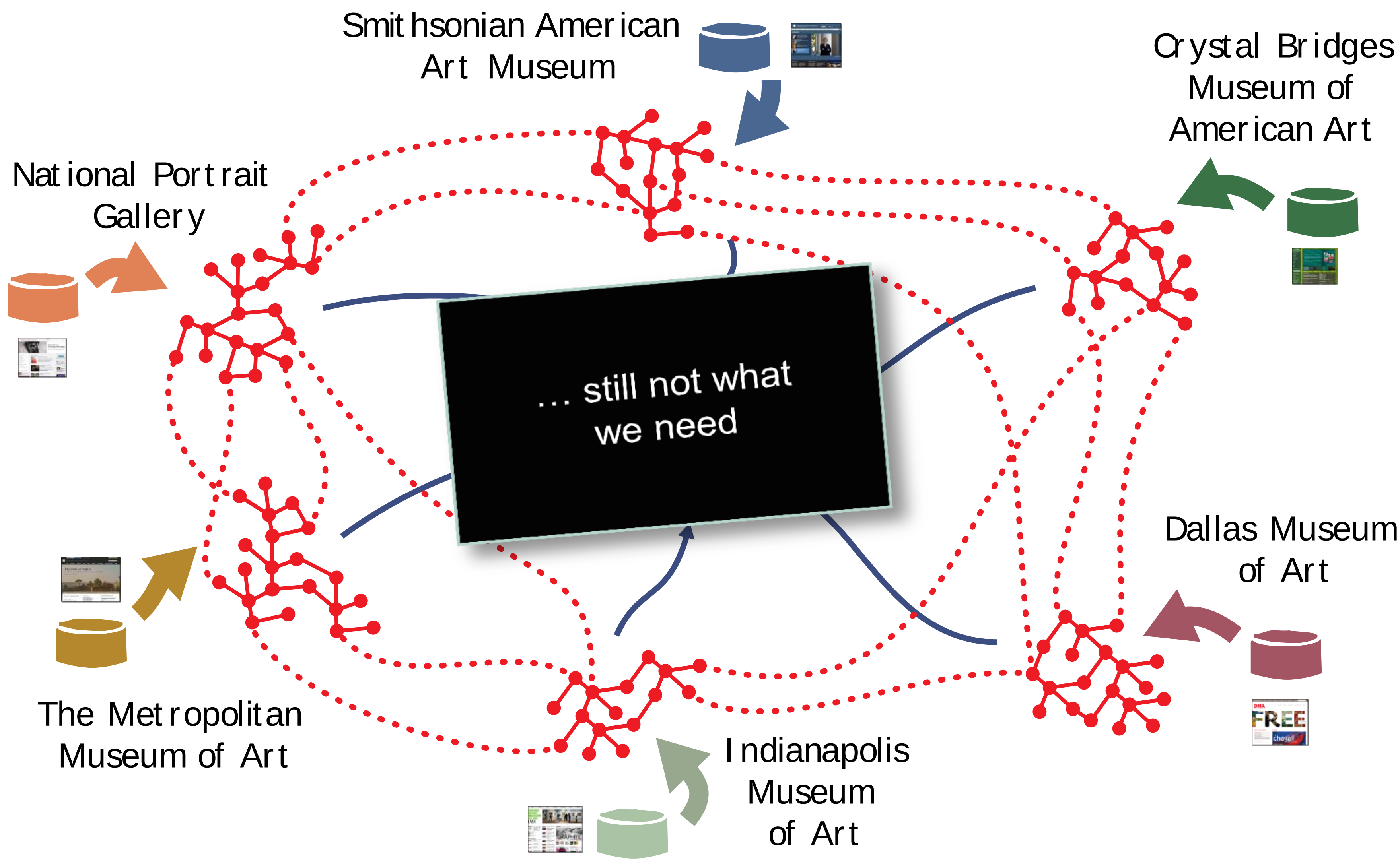
Cultural Heritage “Linked” Open Data



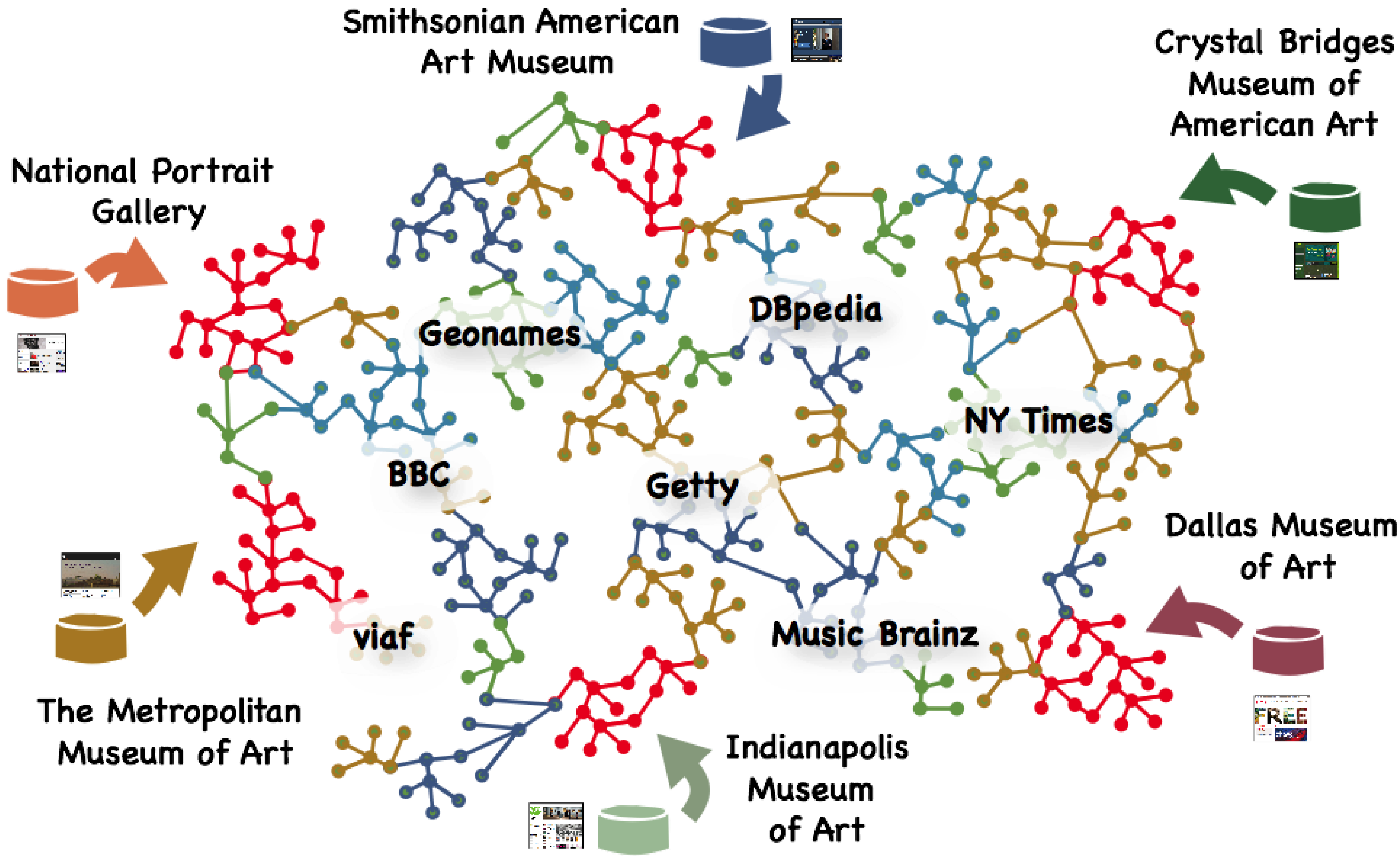
Cultural Heritage Linked Open Data



Cultural Heritage Linked Open Data



Linked Open Data



Integrated Querying based on owl:sameAs Links

http://d-nb.info/gnd/118547739

http://id.loc.gov/authorities/names/n50019335

http://viaf.org/viaf/12466780

http://dbpedia.org/resource/John_Singer_Sargent

http://www.wikidata.org/entity/Q155626

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia: <http://dbpedia.org/resource/>

SELECT ?object ?title ?picture WHERE {
  dbpedia:John_Singer_Sargent foaf:made ?object .
  ?object dc:title ?title .
  ?object foaf:depiction ?picture .
}
```

Link traversal query processing with reasoning yields 16 results out of ~2900 paintings on data from the LOD cloud as of Jan 2015

Steps to Create Linked Data

- Select ontologies

  … that define classes and properties for our data (e.g., DC, FOAF, CIDOC CRM…)

- Convert data to RDF

  … from the museum database to the ontologies

- Identify links to other Linked Data datasets

  … to other museums and Linked Data hubs

Outline

- Motivation
- **Overview of Approach**
- Building and Extending a Knowledge Graph
- Evaluation
- Conclusion

Goal: Integrate Artist Descriptions

- Getty Union List of Artist Names (ULAN): 109,415 artists
- Smithsonian American Art Museum (SAAM): 8,407 artists
- DBpedia: 1,176,759 people
- The Virtual International Authority File (VIAF): 16,244,546 people

- Goal: consolidate the data into a knowledge graph of artists

$$D_S^1 \qquad D_S^2 \qquad D_S^3 \qquad D_S^n$$

$$D_{KG}^o = \emptyset \qquad D_{KG}^1 \qquad D_{KG}^2 \qquad D_{KG}^3 \qquad D_{KG}^n$$

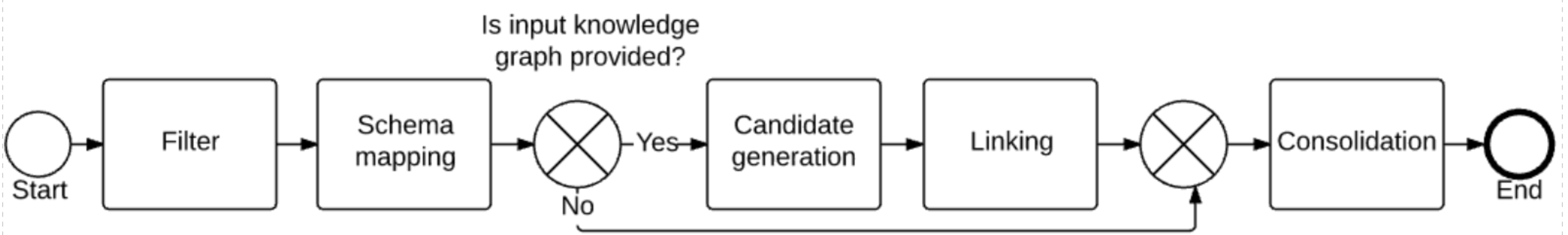
Challenge: Scalability

- Object consolidation requires to compute the similarity of each entity with each other entity
- Impractical with our data size
  - DBpedia ~1.2m people (~900 MB), VIAF ~16.2m people (67 GB)

- How to reduce the number of pair-wise comparisons?

Overview of Approach

1. Filter
2. Schema mapping
3. Candidate generation
4. Linking
5. Consolidation



Outline

- Motivation
- Overview of Approach
- **Building and Extending a Knowledge Graph**
- Evaluation
- Conclusion

1. Filter

- We are interested in artists, but the data sources contain information about many more things

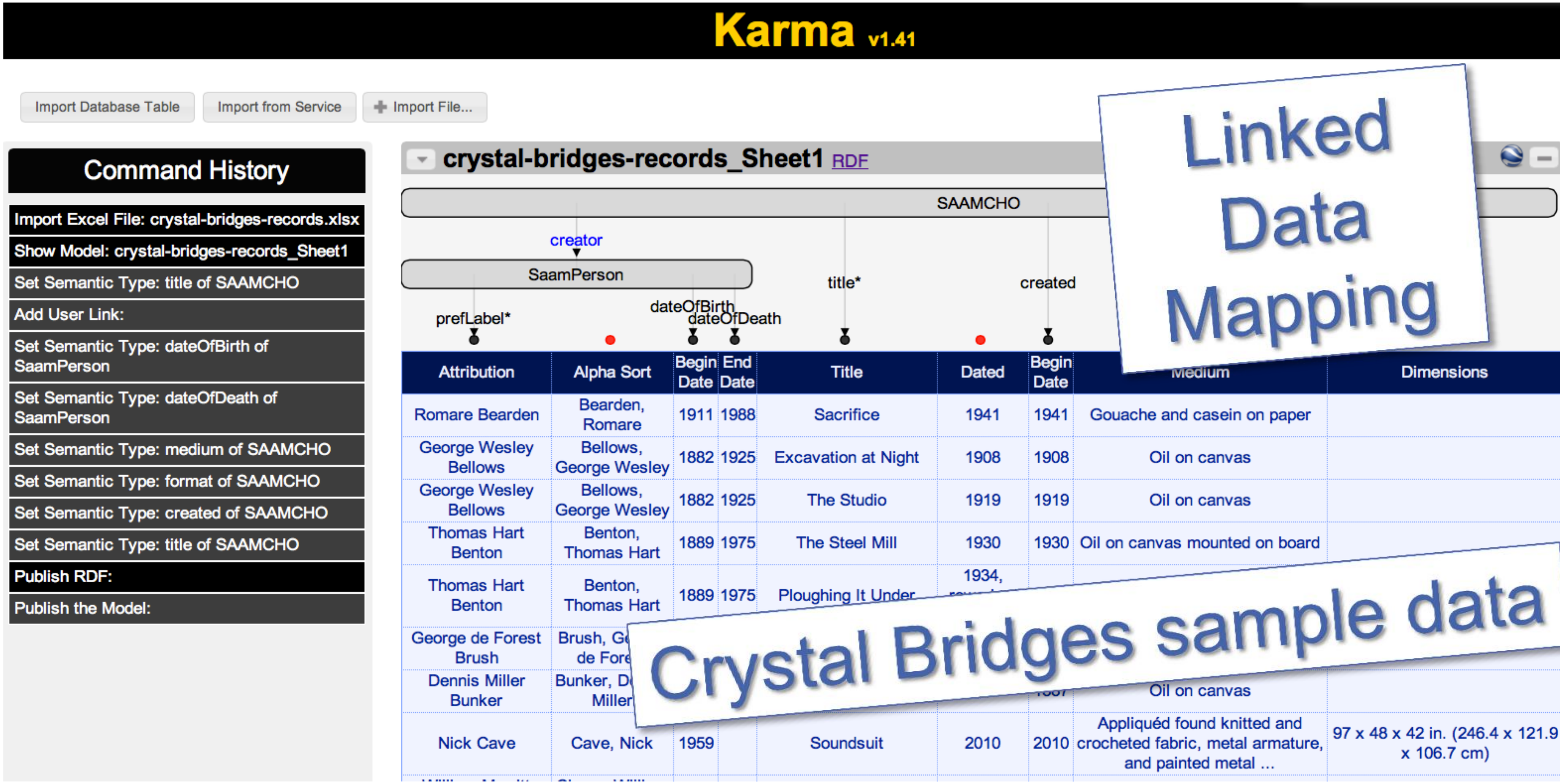
- In the filter step, we select all artists from DBpedia and VIAF via SPARQL queries
- We use a streaming query processor (Linked Data-Fu) to run a query that selects only people from the data and thus reduce the amount of data we have to process further

2. Schema Mapping

- We use the Karma tool to map the person descriptions in different ontologies to terms from schema.org

Property name	ULAN	SAAM	DBpedia	VIAF	Knowledge graph
name	X	X	X	X	X
alternateName		X		X	X
givenName		X	X	X	X
familyName		X	X	X	X
gender	X				X
nationality	X				X
birthDate	X	X	X	X	X
deathDate	X	X	X	X	X
birthPlace	X	X	X		X
deathPlace	X	X	X		X
description	X	X	X	X	X

Karma in Action



3. Candidate Generation

- MinHash/LSH operates over an n-gram representation of the name values, and hashes similar entities into the same cluster, based on the Jaccard similarity between the two sets of n-grams representing the two entities
- MinHash/LSH recall/precision performance depends on the number of use d minhashes m and the number of items in the generated hashes I
- LSH threshold t can be approximates as $t = \frac{1}{i}^{\frac{1}{m}}$

t	34%	40%	46%	52%
Precision	14.12%	15.69%	23.12%	26.96%
Recall	100%	99.50%	99.50%	98.01%

- We apply the MinHash/LSH with a low threshold of 46% to achieve high recall
- A low threshold leads to a low precision which we tolerate because the precision will be increased in the linking step

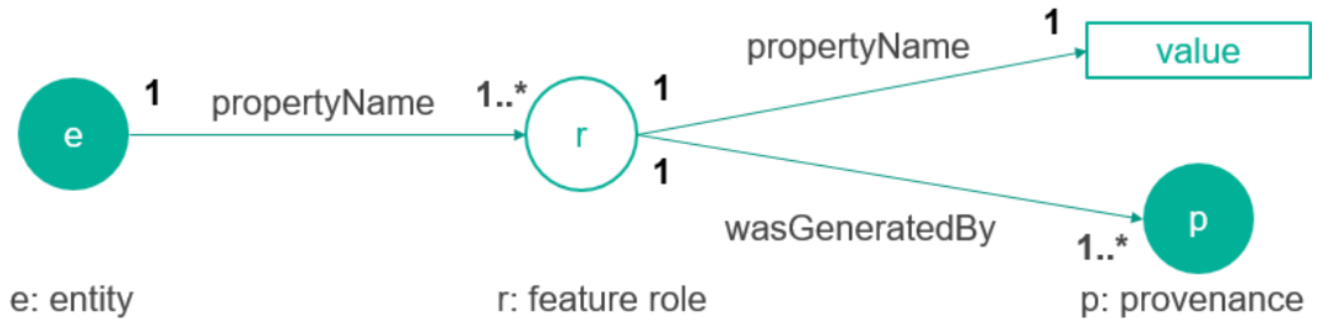
4. Linking

- Computes similarity based on matching functions on the found candidates
- When comparing people entities, we can define a matching function to
  - first check the similarity of the names and then
  - remove candidates with a different birth year

- Birth year might remove correct candidates (e.g., candidate “Pietro Aquila” has birth year “1592” in ULAN but “1650” in SAAM)

5. Consolidation

- Merge data from different sources while keeping provenance using the PROV ontology
- We use an n-ary representation to be able to keep provenance information within the triple data model (binary predicates)



Outline

- Motivation
- Overview of Approach
- Building and Extending a Knowledge Graph
- **Evaluation**
- Conclusion

Runtime Performance Results

- 161,465 artists consolidated from four data sources, based on 17,539,125 entities processed (link to dataset in paper)
- 4 AMD Opteron 62xx class 2GHz CPU cores and 32 GB RAM

Step	ULAN	SAAM	DBpedia	VIAF
Candidate generation	-	00:15:59	01:55:14	29:58:26
Linking	-	00:01:37	01:11:22	55:02:13
Consolidation	00:02:12	00:04:49	00:23:20	156:34:12
Total	00:02:12	00:22:25	03:29:56	229:00:39

Quality Evaluation

Karlsruhe Institute of Technology

- We manually build up a ground truth of links for the alphabetically first 200 artist entities which are represented in each of the four data sources and measured recall and precision

			T	Recall	Precision
ULAN	Initial KG		-	-	-
SAAM	Candidate generation		200	100%	10.87%
	Linking	Hybrid-Jaccard	200	100%	92.59%
		birth year	187	93.50%	100%
DBpedia	Candidate generation		199	98.50%	12.83%
	Linking	Hybrid-Jaccard	197	98.00%	91.59%
		birth year	187	93.50%	100%
VIAF	Candidate generation		271	97.84%	18.80%
	Linking	Hybrid-Jaccard	271	97.12%	77.36%
		birth year	262	93.53%	96.30%

Most links are correct

Only few links are missing

Outline

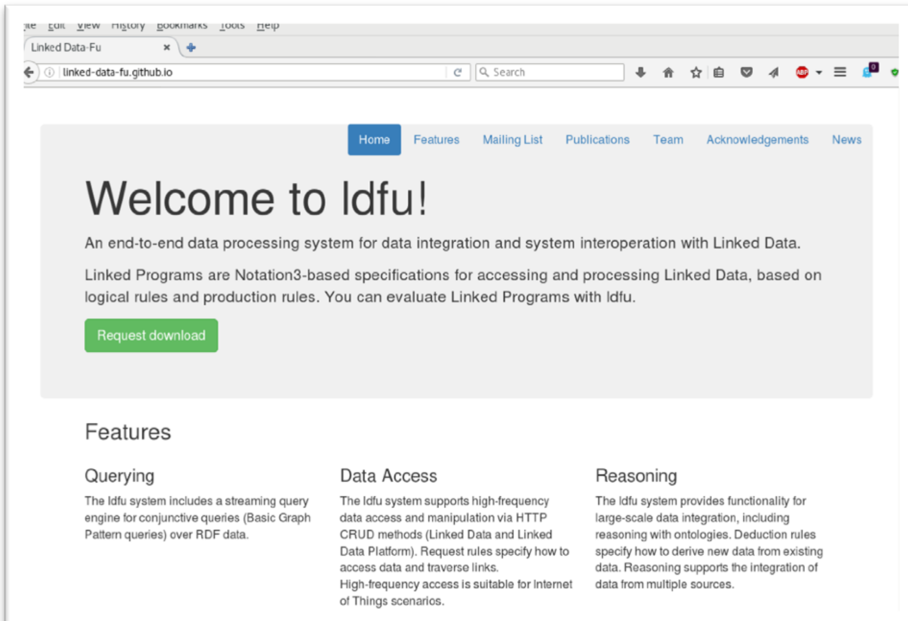
- Motivation
- Overview of Approach
- Building and Extending a Knowledge Graph
- Evaluation
- **Conclusion**

Conclusion

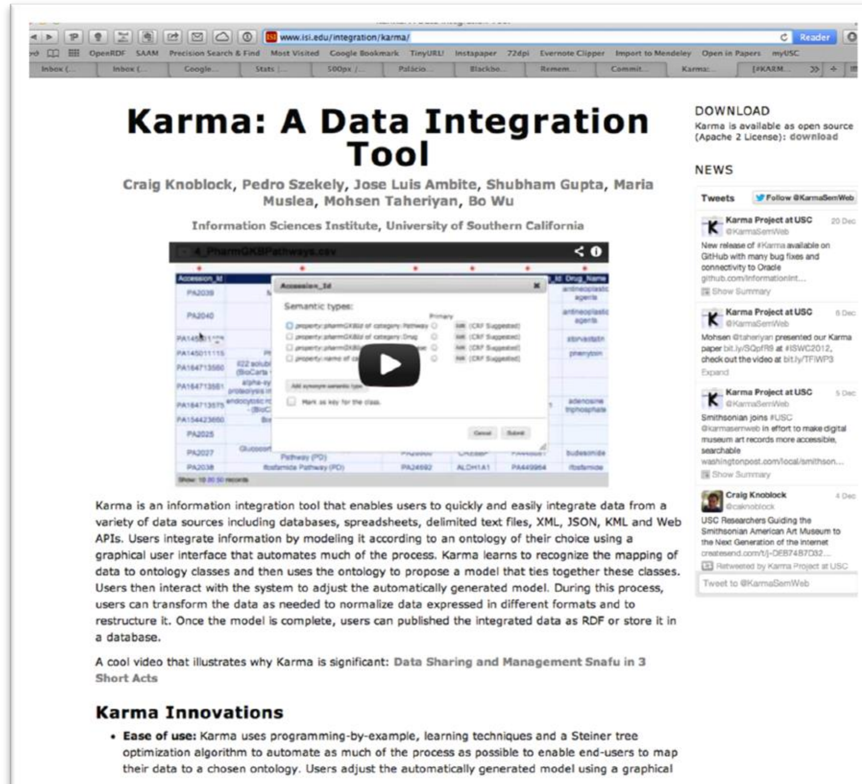
- We have addressed the problem of efficiently building a consolidated knowledge graph out of multiple large data sources
- We have used the MinHash/LSH algorithm to identify candidate links to address the scalability challenge
- The approach can be used on different entity types and different datasets with minimal changes
- More elaborate matching functions could be used in conjunction with our approach
- We provide the used software as open source

Links

http://linked-data-fu.github.io/



http://www.isi.edu/integration/karma/



American Art Collaborative

- Amon Carter Museum of American Art
- Archives of American Art, Smithsonian Institution
- Autry Museum of the American West
- Colby College Museum of Art
- Crystal Bridges Museum of American Art
- Dallas Museum of Art (DMA)
- Indianapolis Museum of Art (IMA)
- Thomas Gilcrease Institute of American History and Art
- National Portrait Gallery, Smithsonian Institution
- National Museum of Wildlife Art
- Princeton University Art Museum
- Smithsonian American Art Museum (SAAM)
- Walters Art Gallery
- Yale Center for British Art

http://americanartcollaborative.org/about/members-of-the-american-art-collaborative/