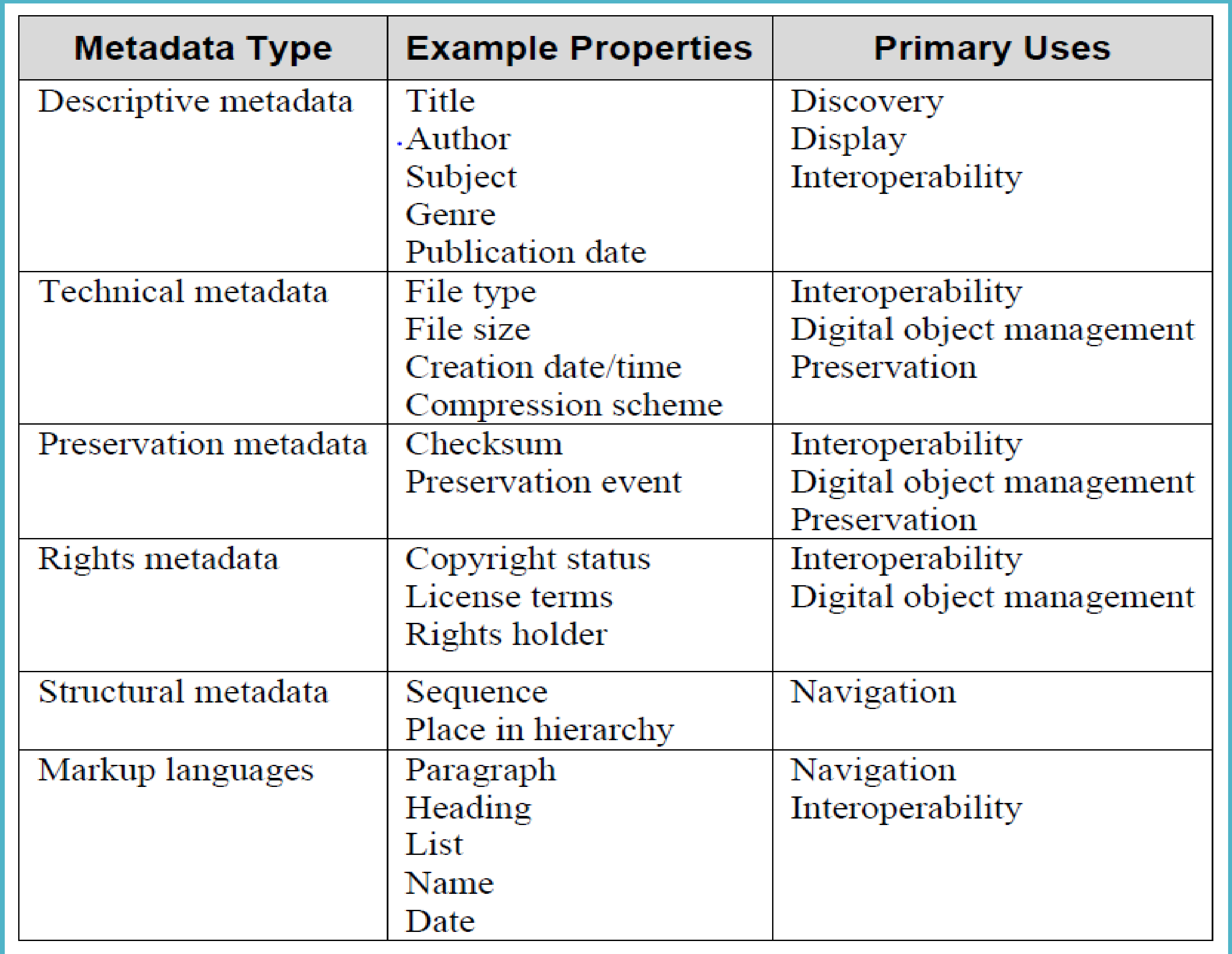
Metadata defenition and generation

Bi-weekly Colloquium
06/08/2021
Online

What is metadata: data about data

Metadata is data that provides information about other data.



Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Why it is important to generate metadata:

An increasing number of research and industrial initiatives have focused on publishing Linked Open Data, but little attention has been provided to help consumers to better understand existing data sets.

How is Metadata Generated?

Descriptive metadata: **humans** create that metadata

For other pieces of metadata, such as background information about an author or performance, Value-added and interpretive information, such as summaries or subjects: by **experts**

Purpose-built metadata entry systems: widely accessible **tools** such as spreadsheets

metadata creation interfaces have become increasingly sophisticated, with user-friendly designs.

Creating metadata through automated processes,

Automated transcription of speech from audio and video technology

Facial recognition technology for video and still images is improving quickly.

For textual resources, latent semantic analysis and topic modeling allow for semi-supervised generation of topics relevant to the analyzed texts.

Part-of-speech and named-entity recognition technologies

Automated image annotation, using algorithms to identify objects in photographs

Creating Metadata for book knowledge graph:

Creator:

Maryam Mohammadi
https://orcid.org/0000-0003-4850-8068
Email:
marygmhm@gmail.com

Date Published:
April 5, 2021

Tags/kewords:
Genre information for book
Knowledge graph
book recommendation

Url :
https://doi.org/10.17605/OSF.IO/CJDP8

File Format:
RDF serialization formats(N-Quads)

Sample size:
total statement in the book knowledge graph is 7858239.

Citation:
Maryam. 2021. “Books Knowledge Graph.” OSF. April 4. doi:10.17605/OSF.IO/CJDP8.

License
CC0 1.0 Universal

Dataset Descriptions: HCLS Community Profile

1. Number of triples in the dataset:

```
SELECT (COUNT(*) AS ?triples)

{ ?s ?p ?o  }
```

7858239

2. number of unique, typed entities in the dataset: 1736308

```
SELECT (COUNT(DISTINCT ?s) AS ?entities)

{ ?s a [] }
```

3. the number of unique subjects in the dataset: 1736316

```
SELECT (COUNT(DISTINCT ?s) AS ?distinctSubjects)

{  ?s ?p ?o   }
```

4. the number of unique properties in the dataset: 25

```
SELECT (COUNT(DISTINCT ?p) AS ?distinctProperties)

{ ?s ?p ?o }
```

hCLS

5. the number of unique objects in the dataset: 1475103

```
SELECT (COUNT(DISTINCT ?o ) AS ?distinctObjects)

{  ?s ?p ?o  FILTER(!isLiteral(?o)) }
```

6. the number of unique classes in the dataset: 10

```
SELECT (COUNT(DISTINCT ?o) AS ?distinctClasses)

{ ?s a ?o }
```

hCLS

7. the number of unique literals in the dataset: 277874

```
SELECT (COUNT(DISTINCT ?o) AS ?distinctLiterals)

{  ?s ?p ?o  filter(isLiteral(?o)) }
```

8. the number of graphs in the dataset: 1

```
SELECT (COUNT(DISTINCT ?g ) AS ?graphs)

{ GRAPH ?g { ?s ?p ?o }}
```

ABSTAT

ABSTAT: Ontology-driven Linked Data Summaries with Pattern Minimalization(2016)

a **general** linked data summarisation framework

It is based on an ontology-based ABstraction model and on the computation of STATistics.

Through this framework a data consumer should be able to answer to questions such as:

what types of **resources** are described in the data set?

What **properties** are used to describe the resources?

What types of resources are linked and by means of what properties?

How many resources have a certain type and how frequent is the use
of a given property?

ABSTAT

ABSTAT framework takes as input a data set and an ontology (used by the data set) and returns a data summary.

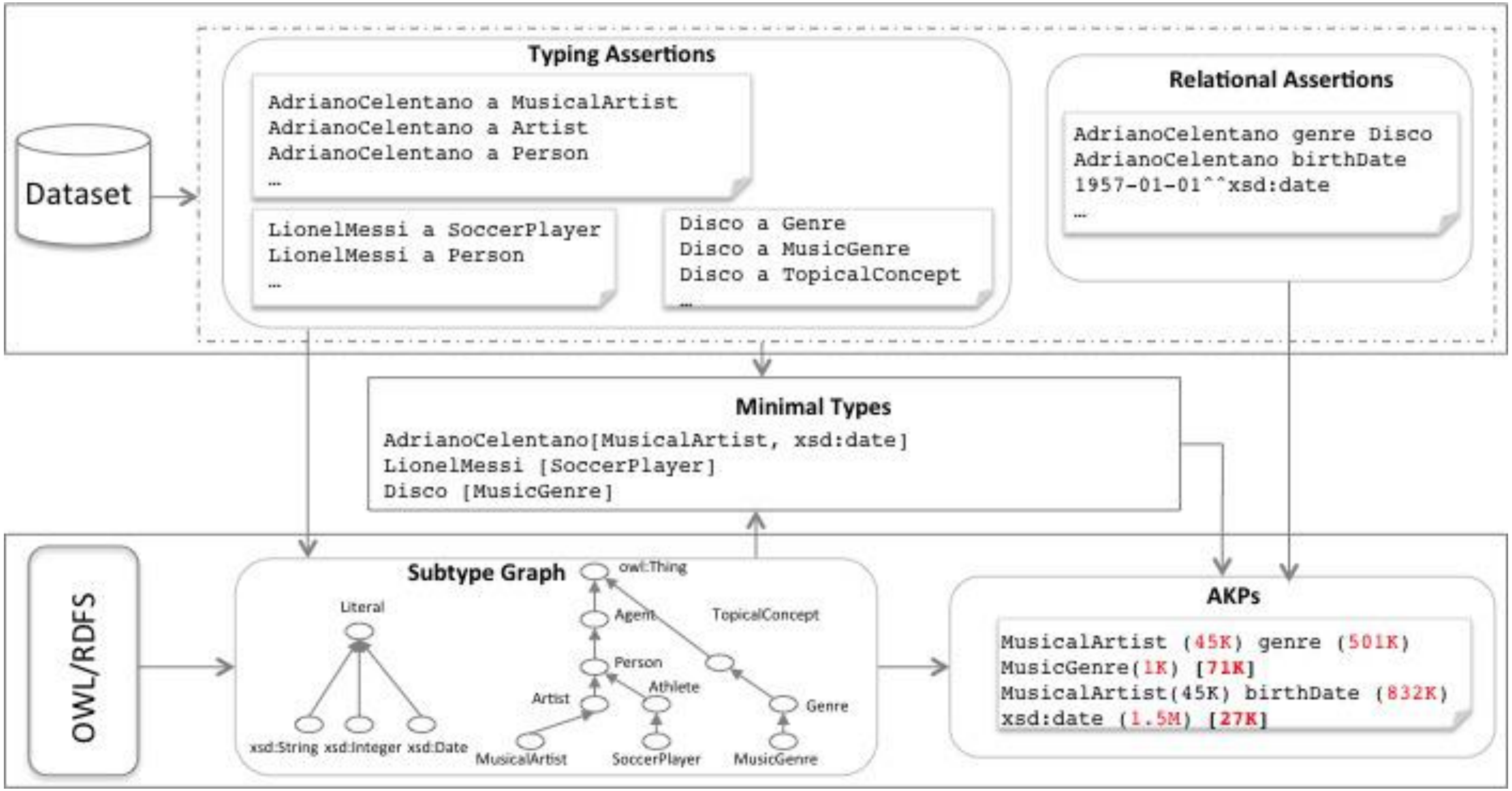
The summary is exported in RDF to support query and navigation.

key feature of approach :

the extraction of Abstract Knowledge Patterns (AKPs) from the data with the help of the data ontology (this work considers OWL and RDFS ontologies).

AKPs are triples having the form <subjectType, pred, objectT ype>, which represent the occurrence of triples <sub, pred, obj> in the data, such that subjectType is a minimal type of sub and objectType is a minimal type of obj.

ABSTAT



Typing Assertions

```
AdrianoCelentano a MusicalArtist
AdrianoCelentano a Artist
AdrianoCelentano a Person
…
```

```
LionelMessi a SoccerPlayer
LionelMessi a Person
…
```

```
Disco a Genre
Disco a MusicGenre
Disco a TopicalConcept
…
```

Relational Assertions

```
AdrianoCelentano genre Disco
AdrianoCelentano birthDate
1957-01-01^^xsd:date
…
```

Minimal Types

```
AdrianoCelentano[MusicalArtist, xsd:date]
LionelMessi [SoccerPlayer]
Disco [MusicGenre]
```

Subtype Graph

AKPs

```
MusicalArtist (45K) genre (501K)
MusicGenre(1K) [71K]
MusicalArtist(45K) birthDate (832K)
xsd:date (1.5M) [27K]
```

Example:

Myke wants to create *MyMusicNow*, a new mobile app with music data. He wants to exploit one data set from the LOD cloud to semantically link news coming from twitter accounts of music artists.

What is the semantic structure of the DBpedia or LinkedBrainz data?

With using ABSTAT:

1) How are music artists/groups/songs described?
2) How many instances are covered?
3) Is there any incongruence in the data?

http://abstat.disco.unimib.it/

ABSTAT

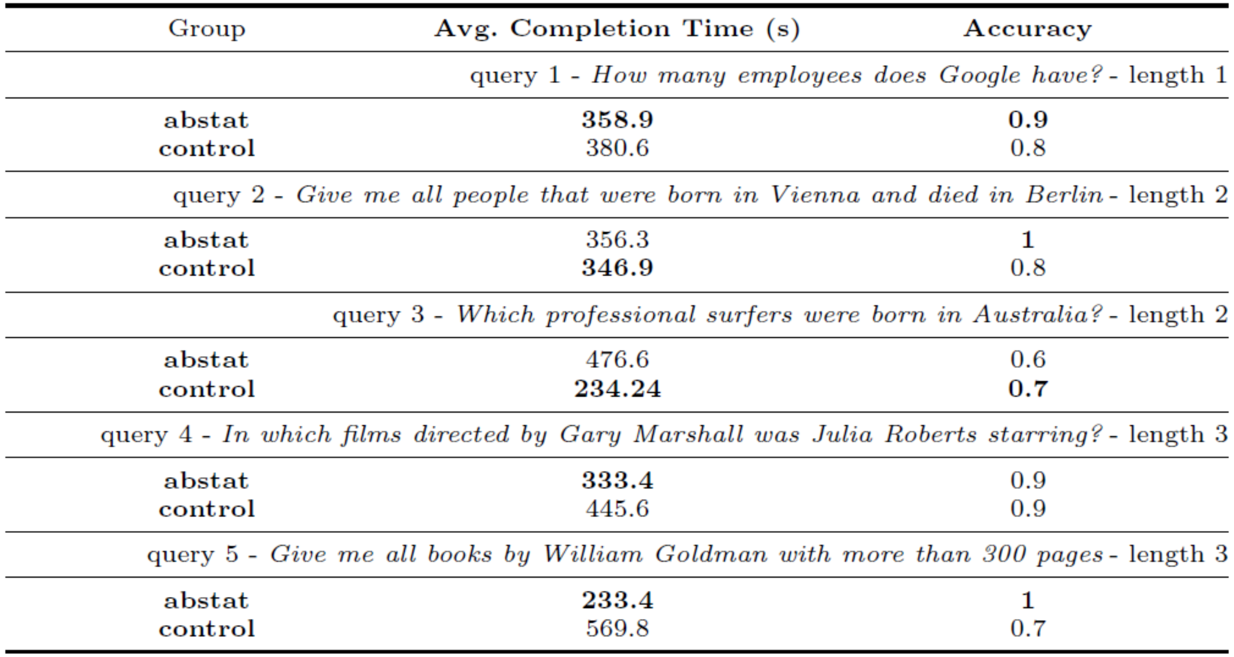
Experimental Evaluation

measure the compactness of ABSTAT summaries and compare the number of their paterns to the number of patterns extracted by **Loupe**

reduction rate, defined as the ratio between the number of patterns in a summary and the number of assertions from which the summary has been extracted.

user study to evaluate if the exploration of the summaries can help users in **query formulation tasks**

Table 3. Results of the user study.



Group	Avg. Completion Time (s)	Accuracy
query 1 - *How many employees does Google have?* - length 1		
abstat	**358.9**	**0.9**
control	380.6	0.8
query 2 - *Give me all people that were born in Vienna and died in Berlin* - length 2		
abstat	356.3	**1**
control	**346.9**	0.8
query 3 - *Which professional surfers were born in Australia?* - length 2		
abstat	476.6	0.6
control	**234.24**	**0.7**
query 4 - *In which films directed by Gary Marshall was Julia Roberts starring?* - length 3		
abstat	**333.4**	0.9
control	445.6	0.9
query 5 - *Give me all books by William Goldman with more than 300 pages* - length 3		
abstat	**233.4**	**1**
control	569.8	0.7

Graph summarisation

Graph Summarization Methods and Applications: A Survey

Its purpose is to extract concise and meaningful information from a graph, representing their content as faithfully as possible.

Main objective

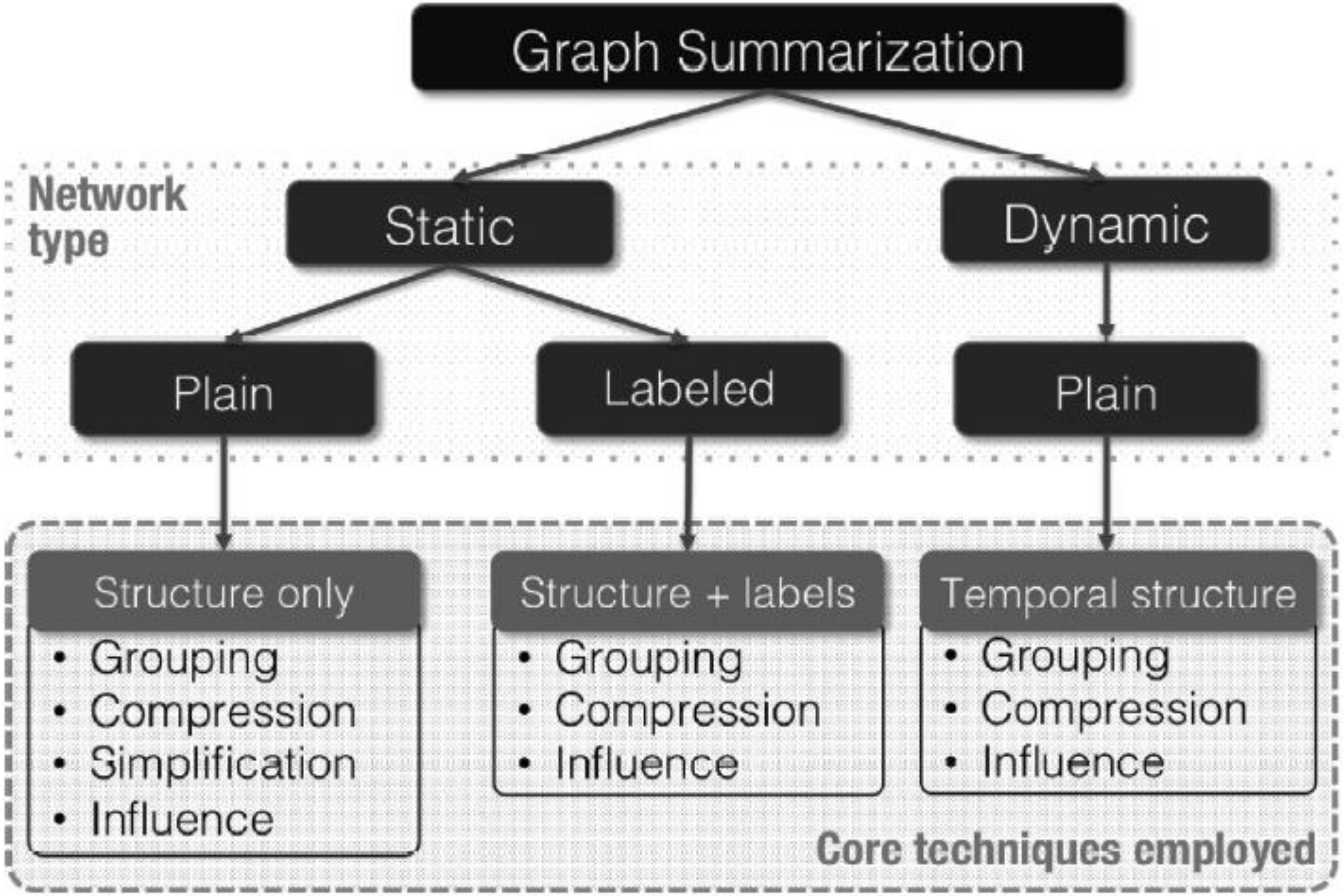
query efficiency,
data size reduction,
static or temporal pattern discovery,
visualization and interactive large-scale visual analytics,
influence analysis and understanding,
entity resolution,
and privacy preservation.

Graph summarisation

Types of Graph Summaries Input:

Static or dynamic.

Homogeneous or heterogeneous.



Graph summarisation

Output: Summary type.

1. supergraph:

Consists of super nodes

or collections of original nodes, and super edges between them.

2. sparsified graph:

which has fewer nodes and/or edges than the original network.

3. A list of (static or temporal) structures or influence propagations, which are seen independently instead of in the form of a single summary graph.

(a) flat, with nodes simply grouped into super nodes, or

(b) (b) hierarchical, with multiple levels of abstraction.

Non-overlapping or overlapping nodes.

Each original node belongs only to one summary element (e.g., supernode, subgraph).

Graph summarisation

Grouping-Based Methods

Node-Grouping Methods.

Node clustering-based methods.

After applying clustering algorithm

(i) mapping all the nodes that belong to the same cluster/community to a super node

(ii) linking them with super edges with weight equal to the sum of the cross-cluster edges or the sum of the weights of the original edges

Node aggregation-based methods:

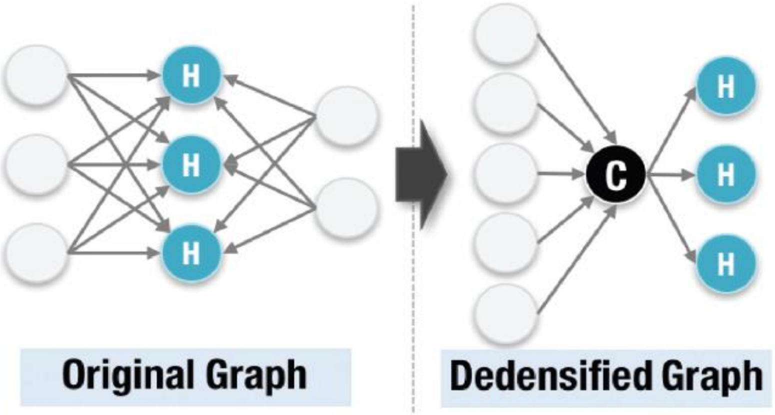
to merge nodes with similar relationships to other entities (structurally equivalent nodes) such that approximation error is minimized and compression is maximized.

Edge-Grouping Methods

aggregate edges into compressor or virtual nodes to reduce the number of edges in a graph accelerating query processing

assumption is: high-degree nodes are surrounded by redundant information that can be synthesized and eliminated

dedensification provides exact answers due to its losslessness and does not suffer from the space/time tradeoff of graph indexing



Graph summarisation

Bit Compression-Based Methods

Bit compression is a common technique in data mining. In graph summarization, the goal of these approaches is to minimize the number of bits needed to describe the input graph

Goal:

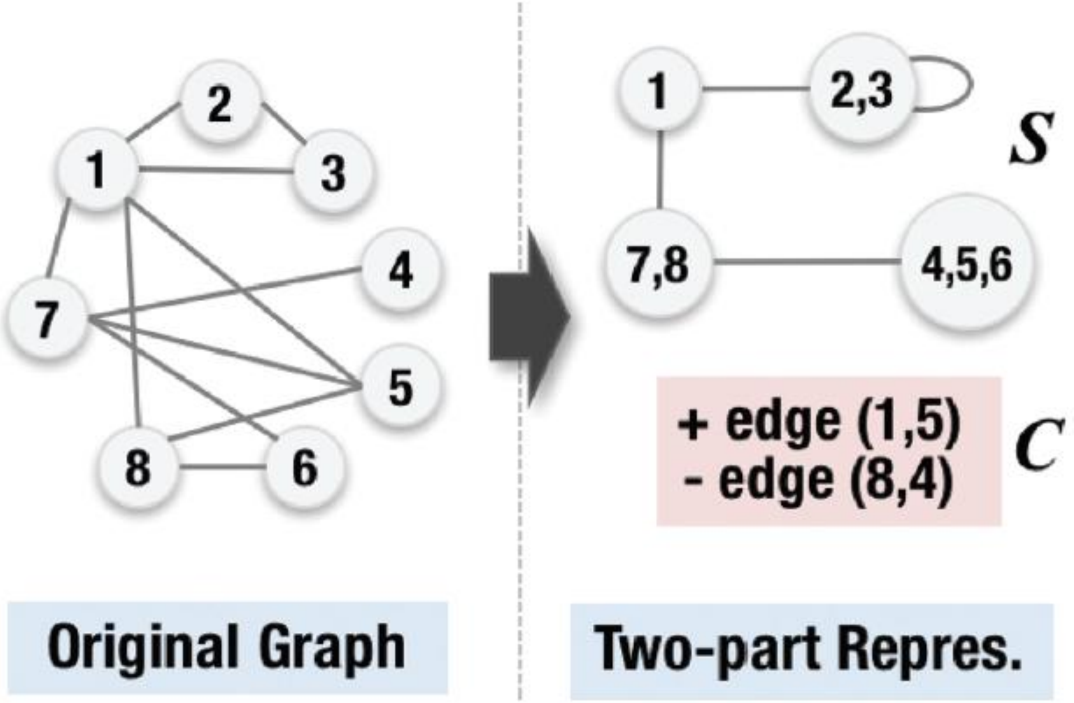
to minimize the description of the given graph G and the model class M in terms of bits:

$$minL(G, M) = L(M) + L(G|M),$$

two-part Minimum Description Length (MDL)

greedy heuristic algorithm which iteratively combines node pairs that give the maximum cost reduction into super nodes

$$cost(R) = |E_S| + |C|$$



Graph summarisation

vocabulary-based summarization of graphs

identifying cliques and near-cliques, stars, chains, and (near-) bipartite cores.

$$minL(G, M) = L(M) + L(G|M),$$

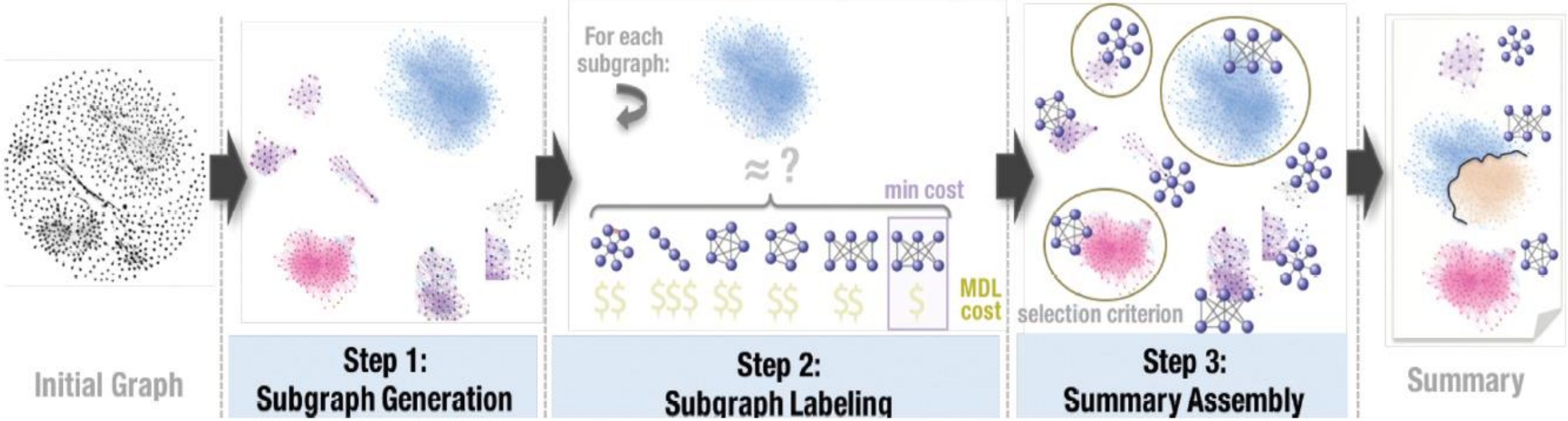


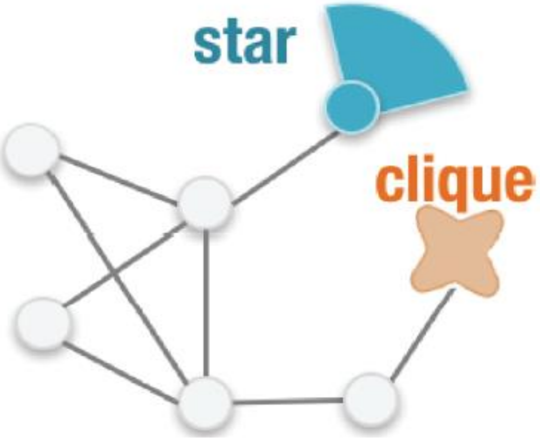
Fig. 5. VoG (Koutra et al. 2014a). Overview of vocabulary-based graph summarization.

Graph summarisation

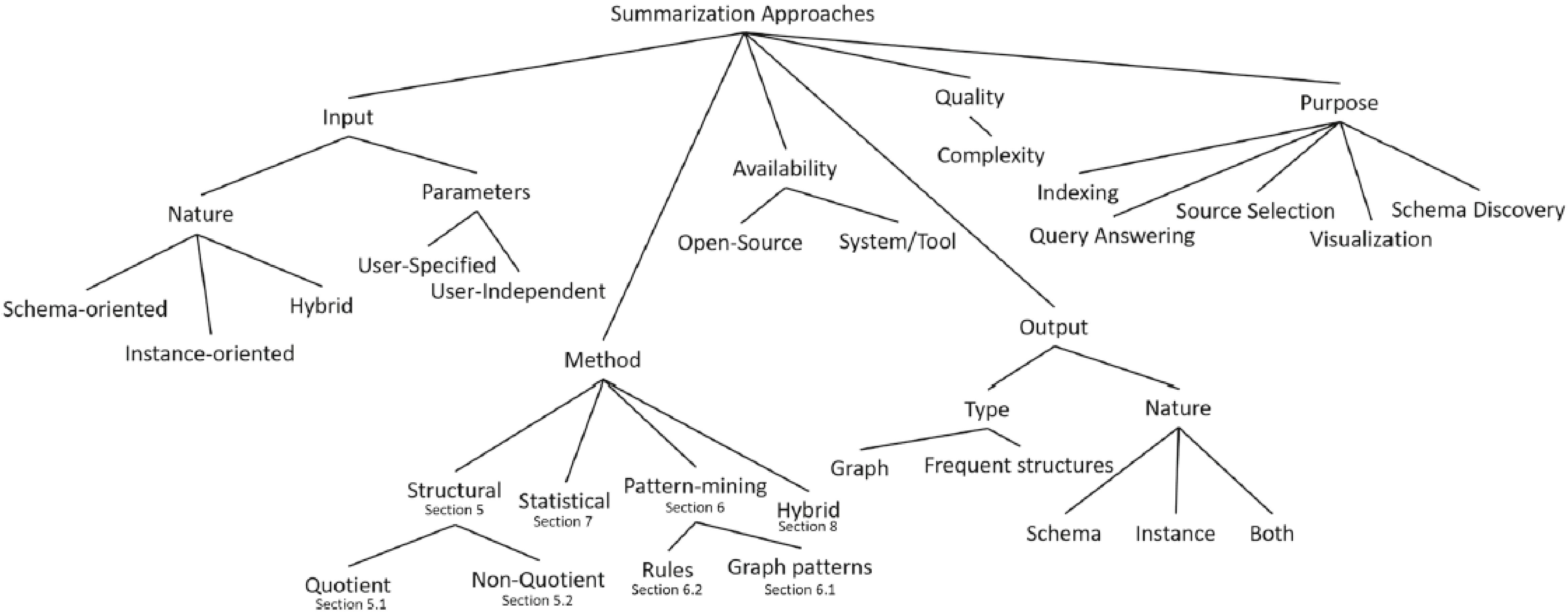
Motif simplification:

replaces common links and common subgraphs, like stars and cliques, with compact glyphs to help visualize and simplify the complex relationships between entities and attributes.

This approach uses exact pattern discovery algorithms



RDF graph summarization



Thank you.