**VAST 2024 MC1 Data Description**

This document introduces data model and serialization format used to store MC1 challenge problem:

Graph Description:

* Directed multi-graph, allowing multiple edges between nodes.
* 215 nodes
* 16231 edges
* There is one large connected component in the MC1 knowledge graph.
* Possible node types are: {'Entity.Organization.NGO', 'Entity.Organization.Company', 'Entity.Location.Region', 'Entity.Organization.GovernmentOrg', 'Entity.Organization.LogisticsCompany', 'Entity.Organization', 'Entity.Commodity', 'Entity.Person', 'Entity.Organization.FishingCompany'}
* Possible edge types are: {'Event.Aid', 'Event.Fishing', 'Event.Transaction', 'Event.Fishing.OverFishing', 'Event.Fishing.SustainableFishing', 'Event.Convicted', 'Event.Applaud', 'Event.CertificateIssued', 'Event.Criticize', 'Event.Owns.PartiallyOwns', 'Event.Communication.Conference', 'Event.Invest', 'Event.CertificateIssued.Summons'}
* The graph format is a JSON format generated by Python’s [network.node\_link\_data()](https://networkx.org/documentation/stable/reference/readwrite/generated/networkx.readwrite.json_graph.node_link_data.html) function. It can likewise be loaded to a networkx object using the corresponding [node\_link\_graph()](https://networkx.org/documentation/stable/reference/readwrite/generated/networkx.readwrite.json_graph.node_link_graph.html#networkx.readwrite.json_graph.node_link_graph) function. The root-level JSON object consists of graph-level properties specifying that it is directed and a multigraph, a “nodes” key which holds the list of nodes, and a “links” key which holds the list of edges.

**Node attributes:**

For the MC1 knowledge graph, ‘id’, and ‘type’ are the mandatory attributes for all the nodes.

* **id** – The unique identifier of the node and the label of the node
* **type** – The type of node.
* **dob** – The person’s date of birth
* **country** – The country associated with the entity.

**Edge attributes:**

For MC1 knowledge graph ‘id’ and ‘type’ are the only fields available. The network serialization format also stores ‘key’ field to support multi-edges but are not used in MC1.

**Metadata attributes:**

Every single node and edge should have each of these fields.

* **\_last\_edited\_by** – The name of the user who last modified this node/edge
* **\_last\_edited\_date** – The last time this node/edge was modified
* **\_date\_added** – The day on which this node/edge was first added to the graph
* **\_raw\_source** – The news source from which the information was originally obtained
* **\_algorithm** – The method in which the information was obtained (for this mini-challenge, either automatically imported from pre-existing databases or manually updated by FishEye analysts)
* **\_articleid** – The natural language description of the extracted knowledge graph node/edges from news organizations, government reports, etc.

Articles:

* The challenge includes a zip archive containing 338 text files.
* Each text file contains the raw text of a news article or news article excerpt.
* The name of the file corresponds to the \_raw\_source attribute

NOTE: The MC1 Knowledge graph may contain information (nodes, edges, or attributes) that are not present in the articles provided. Likewise, there will be information present in the articles that is not present in the graph. Not all raw source documents are provided, and some text provided is excerpted from a longer article. Participants should assume the information in the MC1 knowledge graph was generated from other sources and should not assume it was the result of LLM hallucination.