## dKAIROS Schema Data Format v3.0

Last Modified: November 1, 2022

The following are JSON-LD reserved keywords used by the KAIROS Schema Data Format (SDF). See [JSON-LD v1.1 keyword documentation](https://www.w3.org/TR/json-ld/#syntax-tokens-and-keywords) for details.

* "**@id**": The id of an object that uniquely identifies it with an IRI. Essentially, this is how JSON-LD links to other object definitions: via their ids. It is REQUIRED in top-level documents, events, participants, entities, and relations objects.
  + **IMPORTANT**: With the exception of document-level IDs, all IDs must follow the following naming convention:
    - prefix:<KE\_type>/<unique-5-digit-number>/<anything>, where KE\_type is one of *Events*, *Entities*, *Relations*, and *Participants*,, and /<anything> is optional.
    - The scope of the unique numbers is within the entire schema library.
  + IDs must not have whitespace or certain punctuation (%, \, ", ^, [, ], {, }, |, <, >, and `), and cannot be null.
    - In addition, the following keywords are subject to the same punctuation rules: child, entity, instanceOf, outlinks, wd\_node, reference, relationObject, and relationSubject.

The following are program-level JSON terms within the KAIROS SDF context, as defined by the KAIROS Schema Format (KSF) working group. These terms cannot be overridden by performer context. In many cases, a term may be applied equally to more than one type of element (e.g., an event and a participant).

* "**sdfVersion**": Defined at the document root level (sibling to context and TA1 events/entities), this is a REQUIRED string that identifies the target version of the Schema Data Format. The format of the version string is x.y.z where x, y, and z are numerical digits. Generally, the major version will change with each program phase, and the minor version will change with each quizlet/evaluation. The system will support specific versions of the SDF and is not expected to support every intermediate version.
* "**version**": a REQUIRED performer-specified version defined at the document level. Akin to the "filename" of the submission that, at a minimum, unambiguously indicates the performer. It may also indicate algorithm versions or author in a succinct way.
* "**events**": An array of JSON objects (with an underlying graph structure), each of which contains a definition for an event node in the graph. It is defined at the document root level and is REQUIRED at the document level.
* "**name**": A *human-readable* English label or handle, typically only a word or two (e.g., "transport," "vaccinate," or "patient") It is REQUIRED for events and entities, and recommended in relations objects to convey the meaning of the relation to humans.
  + Note that in entities and events objects, for TA2 it is a handle for use by assessors, and should be the extracted text value or a meaningful human-readable handle (i.e., "alt-text") for non-text extractions.
* "**description**": An English description, typically several words, a phrase, or even 1-2 sentences. REQUIRED in events.
* "**comment**": Provides human-readable detail. This is typically used in human-generated JSON files, but is valid in machine-generated files. Can be a string or an array of strings.
* **"instanceOf"**: When an event is an instance of another event, i.e., when reusing a low-level subevent, this optional keyword specifies the @id of the original *template* event. This allows TA1 to preserve context that a subevent is reused.
* "**privateData**": A map of performer-specific properties that can be associated with any object (e.g., events, relations, temporal, etc.). The keyword may be used freely, but is not designed to store large amounts of data.
* "**modality**": Indicates additional properties for events, entity fillers (values), and relations, in accordance with LDC guidance. For events, valid values are generic, hedged, irrealis, and negated. For argument fillers, generic, negated and hedged are allowed. For relations, only negated and hedged are allowed. Multiple modality strings are defined as an array of elements.
* "**isSchema**": If set to true, it indicates that the event is a meaningful semantic unit such as a schema or an event complex. If it is set to false *or missing*, the event is not a meaningful semantic unit, such as logical gate containers and primitives.TA2 systems should only instantiate events where isSchema is true.
* "**importance**": An array of objects, each of which represents the probability of the parent event given the child event, entity, or relation, i.e., P(parent|child). It is intended to be useful for schema matching. The field can be defined either within an event object or at the document level. If this field is defined in an event where instanceOf is set, its values override equivalent parent-child values in the template event.
* "**likelihood**": An array of objects, each of which represents the probability of the parent event given the child event, entity, or relation, i.e., P(child|parent). It is intended to be useful for event prediction. The field can be defined either within an event object or at the document level. If this field is defined in an event where instanceOf is set, its values override equivalent parent-child values in the template event.
* "**probParent**": An @id reference to a parent event.
* "**probChild**": An @id reference to a child event, entity, or relation. The child can be either a direct child or a descendant, but it must belong to the same schema as the parent.
* "**probability**": A float value between 0 and 1 that represents the probability in an importance or likelihood object.
* "**children**": An array of strings, each of which is an @id reference to a child event node, an element in the events array.
* "**repeatable**": When set to true, indicates that the event can occur multiple times. Repeatable events must be expanded by TA2 as distinct events at instantiation time.
* "**outlinks**": One or more event @id references, usually to *sibling* event nodes of the node specified by children. It indicates the next event node(s) to be processed in the sequence.
* "**children\_gate**": In an event node, indicates the logical processing semantics of the node's children. REQUIRED when the children keyword is specified. Expected values are "and", "or", and "xor" (case-insensitive), which have the following operational definitions:
  + **AND**: all children are expected to occur (e.g., "prep bomb").
  + **OR**: at least one child is expected to occur (e.g., "covid symptoms").
  + **XOR**: exactly one child is expected to occur (e.g., "manslaughter").

**NOTE:** It is up to TA2 teams to decide how to use the schema for instantiation and prediction. The logical operators are meant to describe the prototypical case.

* "**wd\_node**": A q-node (or p-node) from Wikidata. A q-node with QID Q1234 should be expressed as an @id in the form "wd:Q1234". A p-node with QID P1234 should be expressed as an @id in the form "wdt:P1234". The wd and wdt prefixes are defined in the KAIROS context file. In TA1 output, It is REQUIRED in objects in the entities and relations arrays, as well as in primitive events (events objects with no children). It is optional in participants.
  + To specify multiple wd\_nodes, instead of "wd:Q1234 | wd:Q4321", create a list of wd\_nodes: ["wd:Q1234", "wd:Q4321"], which has "OR" semantics. There is no way to specify AND wd\_nodes.
  + The number of wd\_labels and wd\_descriptions must match the number of wd\_nodes, but blank wd\_labels are ignored.
  + NOTE: In the event that a DWD node does not have a Wikidata equivalent, you can refer to it via "dwd:Q1234". The dwd prefix is also be defined in KAIROS context, but can be overridden by performers.
* "**wd\_label**": The Wikidata label of the wd\_node defined at the same level (typically, an object in the entities, relations, or events array), REQUIRED when wd\_node is present. Given that the label is set at a particular moment in time, it can become out of date as Wikidata evolves.
* "**wd\_description**": This is the Wikidata description of the corresponding wd\_node. It is REQUIRED in entities, events, participants, and relations when wd\_node is provided. If the description is null or missing in Wikidata, then include an empty string.
* "**participants**": An ordered JSON array of participant arguments, REQUIRED in event nodes with no children.
* "**roleName**": REQUIRED in participant objects. It is essentially the name/label of the slot, but connotes the semantic or linguistic role that is filled by the slot in the event. Guidance on how to specify role names can be obtained from the Cross-Program Ontology Group (XPO).
* "**entity**": REQUIRED in participant objects, this keyword links to the @id of an object in the entities array with a wd\_node, or—in the case of event arguments—a valid @id of an object in the events array. In this sense, entity can be thought of as a variable in the TA1 schema, which is filled by TA2 based on Graph G.
* "**templateParticipant**": REQUIRED in participant objects when instanceOf is set and forbidden otherwise. This keyword links to the @id of the corresponding object in the participants array of the template even. In this sense, templateParticipant can be thought of as a parameter in a function that is being filled by an argument in the calling function.
* "**entities**": An array of objects, each of which defines an entity in the schema or a real-world entity from documents or human Graph G. Each object in the array has a REQUIRED @id , name, wd\_node, wd\_label, and wd\_description for the entity. It can also specify a reference or aka. Entities can be defined in events objects or at the document level. The former is preferred to limit the scope of an entity to the containing event and its descendants.
* "**relations**": An array of objects that collectively specify entity-entity or event-event relations, ultimately establishing *subject* – *predicate* – *object* triples. Each object in the array is REQUIRED to contain an @id, wd\_node, relationObject, relationSubject, ta1ref, wd\_label, and wd\_description. Relations can be defined in events objects or at the document level. Relations defined at the top level can include temporal relations, equivalence relations between entities/events of different subevents in the hierarchy, or any other relation.
  + Relation scope should be as limited as possible, so it is preferable to define them in an event if they only apply to the event and its descendants.
  + outlinks should be used where possible instead of temporal relations.
  + **IMPORTANT**: NIST has requested that performers use a subset of Wikidata qnodes for temporal relations:
    - before (Q79030196), partial coincidence (Q65560376), and contains (P4330).
  + **NOTE**: It is the responsibility of the performer to ensure that there are no cycles in temporal relations: cycles between relations are NOT detected, other than to ensure that they do not contain the same relationSubject and relationObject.
* "**relationSubject**": The single *subject* of a relation, typically a reference to an entity or event @id. It is REQUIRED in each relations array element.
* "**relationObject**": The single o*bject* of a relation, typically a reference to an entity or event @id. It is REQUIRED in each relations array element.

## Alphabetical Keyword Quick Reference

| **Keyword** | **Location in document** | **Required?** | **Type / Notes** |
| --- | --- | --- | --- |
| @id | document, entities, events, participants, relations | Yes | Must follow a prescribed format, be a valid URI, and be unique |
| children | events | No | array of @ids of child event nodes |
| children\_gate | non-primitive events | Yes when children present | "and", "or", or "xor" |
| comment | any object | No | string or array of strings |
| description | events | Yes | string |
| entities | document, events | Yes, at the document level | array of objects |
| entity | participants | Yes | @id of entity or event |
| events | document | Yes | array of objects |
| importance | events | No | array of objects |
| instanceOf | events | No | @id of the reused subevent |
| isSchema | events | No | Set to true if event is a meaningful semantic unit |
| likelihood | events | No | array of objects |
| modality | events, relations | No | string subject to controlled vocab |
| name | entities, events, relations | Yes, except only encouraged for relations | string |
| outlinks | events | No | array of event @ids |
| participants | events | No | array of objects |
| privateData | any object | No | anything |
| probability | importance, likelihood | Yes | float between 0 and 1, inclusive |
| probChild | importance, likelihood | Yes | event, entity, or relation @id |
| probParent | importance, likelihood | Yes | event @id |
| relationObject | relations | Yes | @id of entity or event |
| relations | document, events, instances | No | array of objects |
| relationSubject | relations | Yes | @id of entity or event |
| repeatable | events | No | boolean |
| roleName | participants | Yes | string |
| sdfVersion | document | Yes | specially formatted string |
| templateParticipant | participants | Yes, when isSchema is set | @id of corresponding participant in template event |
| version | document level | Yes | specially formatted string |
| wd\_description | entities, events, participants, relations | Yes, when wd\_node is present | string or array of strings, straight from Wikidata |
| wd\_label | entities, events, participants, relations | Yes, when wd\_node is present | string or array of strings, straight from Wikidata |
| wd\_node | entities, events, participants, relations | Yes, in primitive events and when linked by entity | @id or array of @ids, e.g., "wd:Q1234" |

## Appendix

The following fields existed in SDF v2.2 but are currently unused. They might be re-added in the future.

* "**aka**": a synonym or string substitution for an event or entity. Specifying multiple synonyms is achieved by putting them in an array. It can be used by TA2 in schema matching, but is not used in evaluation or by assessors.
* "**reference**": an external URI reference (or references, if placed in an array) to other datasets, knowledge bases, etc., e.g., VerbNet.
* "**minDuration**": TA1 can set this to express a suggested notional minimum duration for the event that may be used by TA2 for schema matching.
* "**maxDuration**": TA1 can set this to express a suggested notional maximum duration for the event that may be used by TA2 for schema matching.
* "achieves": Allowed in events. TA1 defines its semantics for TA2.
* "requires": Allowed in events. TA1 defines its semantics for TA2.
* "**goal**": Allowed in events. TA1 defines its semantics for TA2.

| **Keyword** | **Location in document** | **Required?** | **Type / Notes** |
| --- | --- | --- | --- |
| achieves | event | No | string |
| aka | events, entities | No | string or array of strings |
| goal | events | No | string |
| maxDuration | events | No | xsd:duration |
| minDuration | events | No | xsd:duration |
| reference | entities, events, participants, relations | No | @id or array of @ids of reference to external source(s) (not Wikidata) |
| requires | events | No | string |