The 1G Data Model*

Olaf Hartig
ohartig@amazon.com
Amazon Web Services

In this document we formalize the 1G data model by defining the notion of a 1G dataset. For this definition we assume three countably infinite sets: the set $\mathcal{S}$ of all strings, the set $\mathcal{V}$ of all values, and a set $\mathcal{N}$ of so-called *1G elements*. While every string is a value (i.e., $\mathcal{S} \subset \mathcal{V}$), the sets $\mathcal{N}$ and $\mathcal{V}$ are disjoint.

The 1G elements in $\mathcal{N}$ have identity and represent the basic building blocks of 1G datasets, where different elements may assume different roles within such a dataset. In particular, we shall see that there can be 1G elements in a dataset that represent so-called *property statements* which resemble the notion of properties in LPGs or RDF triples with literal objects. Other 1G elements within a dataset may represent so-called *relationship statements* which resemble edges in LPGs or RDF triples that have a resource as object. Besides the option to use 1G elements as statements, they may also be used simply as nodes in graphs, as well as to represent graphs themselves, where such graphs are containers of statements.

Now we are ready to define the notion of a *1G dataset* formally.

Definition 0.1. A **1G dataset** D is a tuple $(N, \phi_{\mathsf{ps}}, \phi_{\mathsf{rs}}, \phi_{\mathsf{ms}})$ where:

- $N \subset \mathcal{N}$ is a finite set of 1G elements that is partitioned into the following subsets, which are pairwise disjoint:[1]
  - $N_{\mathsf{sn}} \subseteq N$ is the set of so-called *simple nodes* of D
  - $N_{\mathsf{ps}} \subseteq N$ is the set of so-called *property statements* of D
  - $N_{\mathsf{rs}} \subseteq N$ is the set of so-called *relationship statements* of D
  - $N_{\mathsf{ms}} \subseteq N$ is the set of so-called *membership statements* of D
  - $N_{\mathsf{g}} \subseteq N$ is the set of so-called *1G graphs* of D
- $\phi_{\mathsf{ps}} : N_{\mathsf{ps}} \to N \times (N_{\mathsf{sn}} \cup \mathcal{S}) \times \mathcal{V}$ is a function that maps every property statement to its three components,
- $\phi_{\mathsf{rs}} : N_{\mathsf{rs}} \to N \times (N_{\mathsf{sn}} \cup \mathcal{S}) \times N$ is a function that maps every relationship statement to its three components,
- $\phi_{\mathsf{ms}} : N_{\mathsf{ms}} \to (N_{\mathsf{ps}} \cup N_{\mathsf{rs}}) \times N_{\mathsf{g}}$ is a function that maps every membership statement to its two components.

Example 0.2. As an initial example, consider a 1G dataset $D = (N, \phi_{\mathsf{ps}}, \phi_{\mathsf{rs}}, \phi_{\mathsf{ms}})$ that contains a single 1G graph g that consists of two property statements, st_1 and st_2, and one relationship statement, st'. Suppose

$$\phi_{\mathsf{ps}}(st_1) = (n, \texttt{"name"}, \texttt{"Bob"}) \text{ and}$$
$$\phi_{\mathsf{ps}}(st_2) = (n', \texttt{"name"}, \texttt{"Alice"}),$$

where n and n' are 1G elements that represent persons named Bob and Alice, respectively. The relationship statement st' may then capture a sibling relationship between these persons:

$$\phi_{\mathsf{rs}}(st') = (n, \texttt{"is sibling of"}, n').$$

The fact that each of these three statements is contained in the 1G graph g of the dataset is captured by means of three membership statements, st'_1, st'_2, and st'', with:

$$\phi_{\mathsf{ms}}(st'_1) = (st_1, g),$$
$$\phi_{\mathsf{ms}}(st'_2) = (st_2, g), \text{ and}$$
$$\phi_{\mathsf{ms}}(st'') = (st', g).$$

The subsets of 1G elements used in this dataset are $N_{\mathsf{sn}} = \{n, n'\}$, $N_{\mathsf{ps}} = \{st_1, st_2\}$, $N_{\mathsf{rs}} = \{st'\}$, $N_{\mathsf{ms}} = \{st'_1, st'_2, st''\}$, and $N_{\mathsf{g}} = \{g\}$.

While our definition distinguishes property statements and relationship statements of a 1G dataset $D = (N, \phi_{\mathsf{ps}}, \phi_{\mathsf{rs}}, \phi_{\mathsf{ms}})$, we sometimes want to refer to the union of these two sets, which we then call the *1G statements* of D. Moreover, for every such statement $st \in (N_{\mathsf{ps}} \cup N_{\mathsf{rs}}) \subseteq N$ with $\phi_{\mathsf{ps}}(st) = (s, p, o)$, respectively $\phi_{\mathsf{rs}}(st) = (s, p, o)$, we call s the *subject* of st, p is the *predicate* of st, and o is the *object* of st. Similarly, for every membership statement $st \in N_{\mathsf{ms}} \subseteq N$ with $\phi_{\mathsf{ms}}(st) = (st', g)$, we call st' and g the *subject* and the *object* of st, respectively. Note that these notions of subject, predicate, and object are dataset dependent (because they depend on the functions ϕ_{ps}, ϕ_{rs}, and ϕ_{ms}) and, thus, should be used only in contexts in which it is clear what dataset is considered.

TODO: write proper text to highlight the following aspects of the definition (perhaps as part of the examples):

- the subjects of statements may be other statements, same for the objects of relationship statements; hence, relationships between relationships, or between properties; also meta-properties
- the subjects of statements may be graphs, same for the objects of relationship statements; i.e., metadata about graphs
- there may be multiple relationship statements with the same subject, predicate, and object (i.e., true multigraphs, as LPGs), but they may have different property statements (or even relationship statements) about them
- there may be multiple property statements with the same subject, predicate, and object, which is not possible in LPGs, nor in RDF (at least not within the same RDF graph); each of these identical property statements may also have different other statements about them
- the same statement may be in multiple graphs of the dataset
- there may be statements that are not directly in a graph, but indirectly (namely, by being subject or object of another statement that is in the graph), in which case they are considered as *not asserted* in that graph (i.e., like quotes triples in RDF-star)
- hence, there may be statements in $(N_{\mathsf{ps}} \cup N_{\mathsf{rs}})$ that are not asserted in any of the graphs of the dataset
- additionally, membership statements are a bit special; discuss them here as well; emphasize that they do not need to be captured physically as extra statements but, instead, by some more space-efficient means (e.g., a special data structure)

*This document presents work in progress. It will be published properly once completed.

[1] Note that each of these subsets may be empty.

Irrespective of how 1G elements are used in specific 1G datasets, for some of them (not all), their identity is captured not only implicitly by the fact that each of them is a distinct element but they are also associated with an explicit identifier in the form of an IRI. To capture these identifiers formally, we assume a partial function $iri : \mathcal{N} \nrightarrow \mathcal{I}$ that is injective, where $\mathcal{I}$ denotes the countably infinite set of all IRIs, which is disjoint from both $\mathcal{N}$ and $\mathcal{S}$.

TODO: write proper text to highlight the following aspects of the definition (perhaps as part of the examples):

- while a 1G statement st may have an IRI $iri(st) = u$ (if $st \in \text{dom}(iri)$), note that this IRI is an identifier of the statement, which is something different from the IRI of the 1G element (simple node) that the statement may have as its predicate
- 1G elements may assume different roles in different datasets (e.g., a 1G element that is used as a property statement in one dataset may be used as a 1G graph in another dataset). While an alternative would be to capture the roles of 1G elements globally (i.e., independent of 1G datasets), but this alternative is problematic when trying to import data in which a particular IRI is used for a different purpose than anticipated by the global definition. More precisely, this alternative has the flaw that it does not allow 1G graphs and 1G statements to have arbitrary IRIs because the function iri is (and should be) injective; hence, a 1G element with a particular IRI cannot be used as a graph in one dataset, as a statement in another dataset, and as an "ordinary" element in a third dataset.

TODO: write text to introduce (and to motivate) the following definition

Definition 0.3. Let $D = (N, \phi_{\text{ps}}, \phi_{\text{rs}}, \phi_{\text{ms}})$ and $D' = (N', \phi'_{\text{ps}}, \phi'_{\text{rs}}, \phi'_{\text{ms}})$ be 1G datasets. D and D' are **equivalent** if there exists a bijective function $n2n : N \to N'$ that has the following properties:

(1) $N'_{\text{sn}} = \{n2n(n) \mid n \in N_{\text{sn}}\}$ and $N'_{\text{g}} = \{n2n(g) \mid g \in N_{\text{g}}\}$.

(2) For every property statement st of D, with $\phi_{\text{ps}}(st) = (s, p, o)$, it holds that $n2n(st)$ is a property statement of D' such that
$$\phi'_{\text{ps}}\big(n2n(st)\big) = \big(n2n(s), n2n(p), o\big).$$

(3) For every relationship statement st of D with $\phi_{\text{rs}}(st) = (s, p, o)$, it holds that $n2n(st)$ is a relationship statement of D' s.t.
$$\phi'_{\text{rs}}\big(n2n(st)\big) = \big(n2n(s), n2n(p), n2n(o)\big).$$

(4) For each membership statement st of D with $\phi_{\text{ms}}(st) = (st', g)$, it holds that $n2n(st)$ is a membership statement of D' s.t.
$$\phi'_{\text{ms}}\big(n2n(st)\big) = \big(n2n(st'), n2n(g)\big).$$

TODO: motivate coherence (we want to avoid "dangling" elements in 1G datasets; i.e., elements in N that are not used within any graph of the dataset))

TODO: motivate the notion of constituents which is needed for the actual definition that comes afterwards

Definition 0.4. Let $D = (N, \phi_{\text{ps}}, \phi_{\text{rs}}, \phi_{\text{ms}})$ be a 1G dataset and g be a 1G graph of D. A 1G element $n \in N$ is a **constituent** of g if any of the following is true:

(1) n is a 1G statement of D, i.e., $n \in (N_{\text{ps}} \cup N_{\text{rs}}) \subseteq N$, and there exists a membership statement st of D s.t. $\phi_{\text{ms}}(st) = (n, g)$;
(2) there exists a property statement st of D such that st is a constituent of g and n is the subject or the predicate of st;
(3) there exists a relationship statement st of D such that st is a constituent of g and n is the subject, predicate, or object of st;
(4) there exists a membership statement st of D such that st is a constituent of g and n is the subject or the object of st.

TODO: some more text here

Definition 0.5. A 1G dataset $D = (N, \phi_{\text{ps}}, \phi_{\text{rs}}, \phi_{\text{ms}})$ is **coherent** if every 1G element $n \in N$ has at least one of the following properties:

(1) n is a membership statement of D, or
(2) n is a 1G graph of D, or
(3) there exists a 1G graph g of D such that n is a constituent of g.

TODO: examples, etc

REFERENCES

[1] Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan Reutter, and Domagoj Vrgoc. 2017. Foundations of Modern Query Languages for Graph Databases. *ACM Comp. Surv.* 50, 5 (2017).

[2] Olaf Hartig. 2019. Foundations to Query Labeled Property Graphs using SPARQL*. In *Proceedings of the 1st International Workshop on Approaches for Making Data Interoperable (AMAR).*

[3] Olaf Hartig, Pierre-Antoine Champin, Gregg Kellogg, and Andy Seaborne (Eds.). 2021. *RDF-star and SPARQL-star.* W3C Community Group Report, online at https://www.w3.org/2021/12/rdf-star.html.