Common Outage Data Format, version 1.0

USC/ISI Technical Report ISI-TR-729

Last updated: 2018-10-29

Alberto Dainotti[1], John Heidemann[2], Alistair King[1], Ramakrishna Padmanabhan[1,3], Yuri Pradkin[2]

1: CAIDA; 2: USC/Information Sciences Institute; 3: University of Maryland

Abstract: *This document defines a data format for exchanging information about Internet outages. It specifies the semantics data about network outages, and two syntaxes that can be used to represent this information. This format is designed to support reports from Internet outage detection systems such as Trinocular, Thunderping, and IODA.*

1. Introduction and Goals

This document defines a data format for exchanging information about Internet outages. With multiple Internet outage detection systems in place (for example, Trinocular [Quan13c], Thunderping [Schulman11a], IODA [Dainotti17a], etc.), a common format facilitates data interchange and comparisons.

This outage format documents "cooked" data, representing the processed conclusions of a sensor or sensors, possibly merged and post-processed to remove noise or incorrect detections. We expect that there will be other "raw" formats that expose more details about a given sensor. Standardizing raw formats is outside the scope of this document.

We define the data interchange *semantics* (the logical contents) separate from its *syntax* (a particular encoding). We expect that all data will follow the same semantics, although some implementations may leave some fields at default values. We expect that there will be multiple syntaxes to allow for different workflow and libraries while optimizing slightly differently (for example, greater self-description vs. greater efficiency).

We anticipate that APIs for exchanging near-real-time outage reports may be built on this data interchange format in the future. At current this format concerns only static (stored) data.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [Bradner97a].

2. Outage Data Semantics

An outage dataset consists of two parts, the metadata and outage events.

2.1. Outage Metadata

Outage metadata contains information about this outage dataset as a whole.

Most outage metadata fields are OPTIONAL. Unless otherwise specified, each field is unicode text to be interpreted by a human.

Defined fields:

- FormatVersion (REQUIRED): the string "1.0". Future versions that are upwards compatible will start with "1.".
- SoftwareVersion: uninterpreted text identifying the software generating this file. (Example: icmptrain-20180101)
- VantagePointLocation: either one location or a list of locations. (Example: Los Angeles, CA, USA.)
- OutageLocationType: an enumeration defining the semantics of the Location field in outage records. The value *"block"* indicates locations are 32-bit, hexadecimal numbers with the low-8 bits set indicating /24 IPv4 address blocks. The value "*prefix"* indicates location is a text-format IPv4 or IPv6 prefix in traditional textual notation (192.0.2/24, or 2001:db8::/32). The value "*external*" indicates Location is an opaque key to an external table. No other values are currently defined. (Example: block.)
- StartTime: of the entire dataset, in UTC. (Example: 1514764800)
- EndTime: of the entire dataset, in UTC. (Example: 1522540800)
- Notes: unstructured information about this data collection. (Example: "Data is missing for Feb. 1 due to a failure of the collection machines.")
- DatasetProvider: name of the organization generating the data (example: USC/ISI)
- DatasetName: a long-term, canonical name for the dataset. (Example: internet_outage_adaptive_a31all-20180101)
- Anonymization: textual description of anonymization, if any. (Example: none.)
- Other fields may be added in the future.

These fields are standardized, but data providers may add other fields. All unknown fields should be ignored.

2.2. Outage Events

The bulk of outage data are outage events, each identifying the Status (up or down) of a particular Location (a network or geographic region) for a particular duration.

While we call them "outage events", records indicate if the Location is down or up for that period.

Outage record fields are either mandatory or optional, as given below. For ordered encodings, they must be provided in this order and optional fields may only be omitted from the end of the list.

Defined fields:

- Location (REQUIRED): The location of the entity with a given status. Should be treated as opaque information as unicode text. In some datasets, Location will be a key to an external table with more information (for example, defining geographic regions when OutageLocationType is "external"). In other datasets, the Location itself has direct meaning (for example, a hex encoded /24 IPv4 prefix, with OutageLocationType "block":).
- Start (REQUIRED): The time this event begins, in UTC.
- Duration (REQUIRED): The duration of the event, in seconds. A duration of 0 indicates actual duration is unknown.
- Uncertainty (REQUIRED): Uncertainty about the duration, if any, in seconds. Duration may be up to this much shorter than is specified. Uncertainty may be zero for when duration is known.
- Status (REQUIRED): 0 to indicate the Location is down, 1 for up, or a negative value for some other state. Will always be an integer in the range [-127..127].
- StatusDetail (OPTIONAL): Additional detail about the status. A small integer [-127..127] that indicates a measurement-specific condition.
- Fraction (OPTIONAL): The fraction of the entity that is up, given as a decimal value between 0 and 1.
- DeltaDown (OPTIONAL): compared to the previous period of observation, the fraction of entities that transitioned to down
- DeltaUp (OPTIONAL): compared to the previous period of observation, the fraction of entities that transitioned to up
- Confidence (OPTIONAL): A numerical estimate of the confidence of this reading, given as a decimal value between 0 and 1.

3. Data Syntax and Encoding

We define multiple encodings to accommodate different sets of implementation optimizations. The intent is that all encodings are easily re-encoded into each other and are 100% semantically equivalent.

In general, all text in encodings is in UTF-8 format. By default, all times are Unix-epoch times (seconds since January 1, 1970) written in seconds, for the UTC timezone. (In principle, some formats could

specify a different time format or require timezone to be explicit, but currently all MUST use the Unix epoch and UTC to promote interoperability.)

3.1. Tab Separated Encoding

For tab separated format, the Outage Metadata and Outage Events are placed in two separate files.

Outage Metadata is written in JSON format as described in the Outage Metadata section of JSON Encoding listed below.

Outage events are written as a particular kind of tab-separated data (Fsdb format [Heidemann08e]):

1. The first line MUST be a header, consisting of the text "`#fsdb -F t block start duration uncertainty downup detail fraction confidence`". If optional fields are omitted from the data, they MUST be omitted from this header.
2. Each line output consists of each field, separated by a single tab, terminated by a newline.
3. Any lines that begin with the character "#" (in the first column only) are to be ignored.

A sample event file, taken from [LANDER-A31]:

```
#fsdb -F t block start duration uncertainty downup
01000400        1514766051      1256835 2062    1
01000400        1516022886      33673   2523    0
01000400        1516056559      6544564 216     1
01000500        1514765721      1257153 2077    1
01000500        1516022874      34545   2497    0
01000500        1516057419      6543635 660     1
01000600        1514766217      1256836 2062    1
01000600        1516023053      33319   2628    0
01000600        1516056372      6543034 660     1
01000600        1522599406      1883    660     -1
# ...
```

Note that this example omits the optional fields (detail, fraction, confidence), and the last record indicates an error status (in this case, a period that is unmeasurable).

3.2. JSON Encoding

As with the tab-separated format, the Outage Metadata and Outage Events are placed in two separate files as described below.

In both files, field names MUST be encoded in `snake_case`. For example, the "OutageLocationType" field defined in the Outage Metadata section would be encoded as `outage_location_type`. Required fields MUST be included, and MUST NOT be set to `null`. Optional fields SHOULD be included and set to `null`, but MAY be omitted, this provides backwards compatibility when new optional fields are added in the future. Any unrecognized fields MUST be ignored by a parser.

3.2.1. Outage Metadata

An Outage Metadata file contains a single JSON object with fields as defined in the Outage Metadata section above. In addition to the above general requirements, the Outage Metadata JSON encoding SHOULD be pretty-printed to improve human-readability.

A sample metadata file:

```
{
  format_version: "1.0",
  software_version: "icmptrain-20180101",
  vantage_point_location: ["Los Angeles, CA, USA", "San Diego, CA,
USA"]
  outage_location_type: "block",
  start_time: 1514764800,
  end_time: 1522540800,
  notes: "Data is missing for Feb. 1 due to a failure of the
collection machines.",
  dataset_provider: "USC/ISI",
  dataset_name: "internet_outage_adaptive_a31all-20180101",
  anonymization: null
}
```

3.2.2. Outage Events

An Outage Events file contains many JSON objects, one per event. Each line MUST be separated by a single newline character (‘\n’). In order to simplify parsing and reduce storage cost, the event JSON SHOULD NOT be pretty-printed. That is, the encoded JSON for a single event SHOULD NOT contain any newline characters.

A sample event file (encoding the same content as the sample given in the tab-separated format section):

```
{"location":"01000400","start":1514766051,"duration":1256835,"uncertai
nty":2062,"status":1}
{"location":"01000400","start":1516022886,"duration":33673,"uncertaint
y":2523,"status":0}
{"location":"01000400","start":1516056559,"duration":6544564,"uncertai
nty":216,"status":1}
{"location":"01000500","start":1514765721,"duration":1257153,"uncertai
nty":2077,"status":1}
{"location":"01000500","start":1516022874,"duration":34545,"uncertaint
y":2497,"status":0}
{"location":"01000500","start":1516057419,"duration":6543635,"uncertai
nty":660,"status":1}
{"location":"01000600","start":1514766217,"duration":1256836,"uncertai
nty":2062,"status":1}
```

```
{"location":"01000600","start":1516023053,"duration":33319,"uncertaint
y":2628,"status":0}
{"location":"01000600","start":1516056372,"duration":6543034,"uncertai
nty":660,"status":1}
{"location":"01000600","start":1522599406,"duration":1883,"uncertainty
":660,"status":-1}
```


In order to minimize the storage and interchange cost of such JSON-encoded Outage Events, the resulting file should be compressed. The sample above requires 909 bytes to store uncompressed, whereas it only requires 237 bytes after normal gzip compression (compared to the 232 bytes required to store a similarly compressed version of the tab-separated sample above).

4. Acknowledgements

John Heidemann and Yuri Pradkin's work on this document is supported by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via contract number 70RSAT18CB0000014 (DIVOICE), and by the Air Force Research Laboratory under agreement number FA8750-18-2-0280 (LACANIC). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

5. References

[Bradner97a] S. Bradner. Key words for use in RFCs to Indicate Requirement Levels. RFCN. 2119, Internet Request For Comments, March, 1997. <ftp://ftp.rfc-editor.org/in-notes/rfc2119.txt>.

[Dainotti17a] Alberto Dainotti, kc claffy, Alistair King, Vasco Asturiano, Karyn Benson, Marina Fomenkov, Brad Huffaker, Young Hyun, Ken Keys, Ryan Koga, Alex Ma, Chiara Orsini, and Josh Polterock. IODA: Internet Outage Detection & Analysis. Talk at CAIDA Active Internet Measurement Workshop (AIMS). March, 2017. <http://www.caida.org/publications/presentations/2017/ioda_aims/ioda_aims.pdf>.

[Heidemann08e] John Heidemann. Fsdb, the flatfile streaming database. Web page https://www.isi.edu/johnh/SOFTWARE/FSDB/. October, 2008.

[LANDER-A31] USC/LANDER Project. Internet Outage Dataset, PREDICT ID: USC-LANDER/internet_outage_adaptive_a31all-20180101. Provided at http://ant.isi.edu/datasets.

[Quan13c] Lin Quan, John Heidemann, and Yuri Pradkin. Trinocular: Understanding Internet Reliability Through Adaptive Probing. In *Proceedings of the ACM SIGCOMM Conference*, pp. 255-266. Hong Kong, China, ACM. August, 2013. <http://doi.acm.org/10.1145/2486001.2486017>, <https://www.isi.edu/%7ejohnh/PAPERS/Quan13c.html>.

[Schulman11a] Aaron Schulman and Neil Spring. Pingin' in the Rain. In *Proceedings of the ACM Internet Measurement Conference*, pp. 19-25. Berlin, Germany, ACM. November, 2011. <https://doi.org/10.1145/2068816.2068819>,