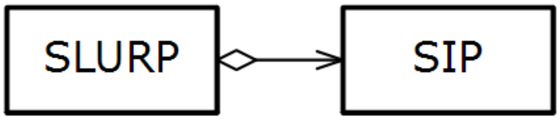
XML FOR STOCHASTIC INFORMATION PACKETS (SIPS)

Proposed Standard

Version 0.3

Marc Thibault
ProbabilityManagement.org

7 September 2013



Contents

1 INTRODUCTION

The SIP XML format encapsulates an array of sample values and related metadata. I've implemented and tested this in Excel and an Excel workbook with code and tests is available.

The XML tag is <SIP>. The content is the value array formatted as a comma-separated values (CSV) string.

A collection of SIPs is encapsulated in a SLURP. Its tag is <SLURP>.

Each has required and optional standard attributes, and arbitrary attributes can be added to meet specific requirements. As is the norm with XML, any attributes that aren't recognized by a particular application are to be ignored.

In object-oriented terms, a particular SIP is an instance of the Sample Distribution class, and the XML string is a serialization of the instance state.

A SLURP can also be seen as a set whose members are SIPs. SIPs and SLURPs can also be seen as entities with attributes and relationships. It all depends on your database bias.

One of the advantages of XML is that translation to a database concept is reasonably easy. In the Excel examples accompanying this proposal, simple key:value structures suffice.

2 SIP STANDARD ATTRIBUTES

name	Required. A text string identifying the SIP, usually unique in context.
count	Required. The number of samples.
type	Required. The string "CSV".
ver	Required. The SIP XML format version. The string "1.0.0".
csvr	Required. Number of digits to the right of the decimal for CSV conversion.

3 SIP FORMAT

<SIP name="$$" count="##" type="CSV"
ver="1.0.0" csvr="2" ... > *CSV Encoding* </SIP>

Example:

```
<SIP name="SIP1"
        count="1000"
        type="CSV"
        origin="smpro.ca"
        ver="1.0.0"
        csvr="2"    >
        0.02, 0.06, 0.14, ..,0.14, 0.01, 0.08, 0.01, 0.0
        5, 0.004, 0.12, 0.15, 0.05, 0.02, 0.17, 0.07, 0.
        17, 0.02
</SIP>
```

4 SIP COLLECTIONS: SLURPS

A SIP Collection is a text document or string containing one or more SIP XML distribution strings encapsulated by a SLURP.

The attributes are collection metadata. The tag is <SLURP>. Text prior to the line starting with the <SLURP tag is not defined by this standard.

Two attributes are required: *name* and *coherent*.

Each enclosed SIP element must begin on a new line.

Rather than containing the component SIPs, the *src="URL"* attribute can be used to refer to the sources of SIPs, to be retrieved as needed.

5 SLURP ATTRIBUTES

name — Can be any string, should be a unique identifier in context.

coherent must be either “true” or “false”. If true, applications should preserve the collection’s coherence.

5.1 SLURP FORMAT EXAMPLE

```
<SLURP name=”LMCO536” coherent=”true”
       about=”Simulation B, Scenario 6” >
       <SIP name= …
       <SIP name= …
       …
</SLURP>
```

6 COMMON OPTIONAL ATTRIBUTES

about A description of the SIP or SLURP.

avg The average or mean of the SIP sample values before they’re encoded into the string.

dataver A number or date indicating the currency of the data in a SIP or SLURP.

min The SIP minimum sample value.

max The SIP maximum sample value.

offset An offset factor to be applied to a SIP encoded value to get the sample value. The ‘b’ in ax+b. Default is 0.

origin An arbitrary text string should say something about the provenance of a SIP or SLURP.

scale A scale factor to be applied to a SIP encoded value to get the sample value. The ‘a’ in ax+b. Default is 1.

src A URL linking to an authoritative copy of the SIP or SLURP. If the SIP or

	SLURP body is empty, this can be used to get the data.
units	A text string for the SIP data measurement units e.g. “Can$”.

7 DOMAIN SPECIFIC ATTRIBUTES

As much as possible, SIP consumers shouldn’t have to develop different decoders for different sources of data. As much as possible, the attributes needed for specific application domains should be standardized before too much fragmentation happens.

A process for proposing and agreeing on such standards needs to be put into place. Probabilitymanagement.org is the appropriate organization to establish the relevant resources.

8 SLURPS OF SLURPS

This standard must also treat hierarchical structures of SLURPs. An example is a group of coherent variables, each with a SLURP holding a time series of SIPs.

Given that we already have SLURPs being treated as files or text streams, the upper levels could be directories or folders in tree structures (let the file system do the work). Rather than complicating the XML with nested, giant SLURPs, a directory structure encapsulated in a common ZIP file can serve the requirement and compress the data as well.

9 METADATA

SIPs and SLURPs will generally be accompanied by administrative metadata, especially provenance. Where it’s provided outside of a SLURP, the human-readable form will be in a file called *provenance.txt* and the machine-readable in a file called *library.txt*.

10 FUTURE

1. Standards relating to the content of *provenance* and *library* files – almost certainly domain-dependent.
2. *Open Stochastic Information Resource* – an architectural and management standard for a managed repository of SIPs and SLURPs. Borrows a lot from the trusted digital repository described by the *Open Archival Information System* standard.

11 PROPOSED REVISIONS

11.1 VERSION 0.3

11.2 CSVR ATTRIBUTE

15 July 2013 Marc Thibault

Add *csvr* as a SIP required attribute. Conversion to CSV format needs to limit the number of digits to the right of the decimal. Application platforms usually have a limit on the number of characters in a string. This will avoid wasting characters on unneeded precision.

There's no reasonable default.

The .*5 rounding rule should be to round up. E.g. with *csvr="2"* the value 2.345 will encode as "2.35".