# Efficiently Representing Uncertainty as Probability Distributions:

Stochastic Information Packets (SIPs) and Stochastic Library Units

With Relationship Preserved (SLURPs)

**Lone Star Analysis**

**25 August 2015**

# Table of Contents

## Overview

This paper discusses two means for efficiently representing uncertainty as probability distributions: Stochastic Information Packets (SIPs) and Stochastic Library Units with Relationship Preserved (SLURPs).

Strings of numbers representing uncertainty and probability distributions have been used at least since 1991 (Dembo, 1991). In 2005, the use of number strings (SIPs and SLURPs) was extended to drive interactive simulations for high-level decision-makers at Royal Dutch Shell (Savage, Scholtes, & Zweidler, 2006). Subsequently, the discipline of probability management was formalized. The approach is further described in *The Flaw of Averages, Why We Underestimate Risk in the Face of Uncertainty (*Savage, 2009) and *Calculating Uncertainty: Probability Management with SIP Math* (Thibault, 2013).

SIPs enable legacy and future simulation models to communicate with each other. SIPs advance the modeling of uncertainty in three fundamental ways:

- Actionable – SIPs may be used directly in calculations involving uncertainty on numerous platforms
- Additive – SIPs allow uncertainties to be aggregated across platforms across the enterprise
- Auditable - Uncertainties are represented as unambiguous data with provenance

Beyond modeling uncertainty, the SIP (as a vector array) makes data from one database/simulation easily accessible to other databases/simulations, thereby facilitating the movement of potentially large and unstructured data between distributed simulators. SIPs/SLURPS can accommodate both “big data” and data as small as a single number. Furthermore, conversion between the three file types (XLSX, CSV, and XML) is facilitated with ease. SIP/SLURP formats have been successfully used with a wide range of software and simulation types, including R, MATLAB, Autobox, and other proprietary software.

The SIP standard is open, neutral, and not tied to any particular format or firm. It is sponsored by Probability Management (PM), a non-profit. There is no fee or license to use SIPs, and the standard is freely available at www.probabilitymanagement.org.

## Standard Specification

The purpose of the specification is to define standards for probability distributions as auditable and transportable data. The two standards defined herein are the Stochastic Information Packet (SIP) and the Stochastic Library Unit with Relationships Preserved (SLURP). The standard defines a simple, adaptable data architecture that makes it easy to create and use SIP libraries by piggybacking on common data formats: CSV, XML, and XLSX (Excel Worksheets). The open SIPmath™ 2.0 Standard may be downloaded from the Probability Management web site http://probabilitymanagement.org/library/SIP-Standard-Version2.pdf.

While the standard was created to support simulations and analysis dealing with uncertainty (“Stochastic Processes”) any data can be archived using the specification. The features of storing a string or table of numbers, or even a single value with all of the descriptive information is valuable. The SIP provides a way to deliver units, the name of the variable, data provenance, and other information. Even when the data delivered in a SIP doesn't "seem" stochastic, it provides a useful way to create open interfaces among simulation tools and organizations. Because information in one format can easily and reliably be translated to another format, the cost and barriers to information sharing are reduced.

The SIP provenance described below is the "data about the data." This is one of the most valuable features of the specification standard. It provides a way to communicate important information about the origins, vintage, and details of the data. There are two fields in the specification for providing "data about the data," one is the "about" field and the other is "provenance" field.

## SIP Details

The Stochastic Information Packet (SIP) represents a probability or frequency distribution as a data structure that holds an array of values and metadata. In the current standard, the values are realizations of the possible outcomes of an uncertain variable. The array for a probability distribution is composed so that the default probability of each element is 1/N where N is the number of elements in the array. The key benefit of using SIPs is that they are actionable, in that they may be used in calculations. If X is a random variable represented by SIP(X), and F(X) is a function of X, then SIP(F(X))=F(SIP(X)). That is, the function, F, is applied sequentially to each element of SIP(X). This means in effect that SIPs and the arithmetic, relational, and logical operators comprise a group.

**SIP Standard Attributes**

| Name | Description |
|---|---|
| name | Required. A text string identifying the SIP, usually unique in context |
| count | Required. The number of samples |
| type | Required. The format type |
| ver | Required. The format version |

**Common Optional Attributes**

| Name | Description |
|---|---|
| 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 |
| csvr | The number of digits to the right of the decimal for CSV conversion |
| dataver | A number or date indicating the currency of the data in a SIP or SLURP |
| dims | The dimensions of a multidimensional SIP |
| hbin | The bin width of a histogram of the SIP |
| hmin | The minimum value in a histogram of the SIP |
| hnum | The number of bins in a histogram of the SIP |
| hvalN | The value in the Nth bin in a histogram of the SIP |
| max | The SIP maximum sample value |
| min | The SIP minimum 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 institution or project that produced a SIP or SLURP |
| provenance | Information about the source and authority of the data |
| Ptile | The (P/100) percentile |
| 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. |
| units | A text string for the SIP data measurement units e.g. "Can$" for Canadian dollars |

## SLURP Details

A coherent set of SIPs that preserve statistical relationships between uncertainties is known as a Stochastic Library Unit with Relationships Preserved (SLURP). Two or more SIPs are coherent if the values of their corresponding samples are in some way interdependent. For calculations with these SIPs to be valid, the alignment of the samples must be preserved; if one of the SIPs is permuted, the others must be permuted by the same permutation index to preserve coherence. In this respect, the importance of the SLURP is that any SIP calculated with arithmetic, relational, or logical operations on SIPs in a given SLURP will also be coherent with that SLURP. Two attributes are required: name and coherent; one is optional: count.

**SLURP Standard Attributes**

| Name | Description |
|---|---|
| name | Can be any string, should be a unique identifier in context. |
| coherent | Must be either "true" or "false". If false, the coherence of the included SIPs is not assured. |
| count | Optional. The number of SIPs in the SLURP. |

## Example

The Standard's annexes provide examples of all three formats: XML, XLXS, and CSV. Below is an example of how SIPs are used with Excel (XLXS) format. It is intended to show how to construct a SIP using an Excel spreadsheet.

Imagine a restaurant manager who wants to understand the wait for a table. Our manager is too smart to rely on an average. If many patrons have a no wait, their zero minute delay gets averaged with people who may wait a long time. Also, she is too smart to just measure how often customers must wait more than some goal. If they all wait one second less than the limit, that's probably not really good. So, our imaginary manager decides to record all the waiting times, and she wisely chooses to record those waits in a SIP. She has her choice of using XML tagged data, an Excel spreadsheet, or a CSV file. It doesn't matter what she chooses because software can easily change from one file format to another. Like millions of other people, our manager has a copy of Microsoft Excel on her laptop, so she decides to use that format. To record her data as a SIP she only needs to include five things:

- name - She needs a text string identifying the SIP
- count - She needs to record the number of samples of waiting times
- type - She records the format type, in this case Excel
- ver - She makes note of the version of the format

Of course the actual data (the collected waiting times) is included in the SIP as well.

The SIP standard provides the option to record some other things too, and she decides she will also include:

- about - She decides to make some comments about her wait records and her restaurant
- units - She wants to make sure people know she decided to use minutes (not seconds) so she uses the name “Minutes”
- provenance - She thinks she will make some other records like this one, so she wants to record some information about the source and history (the “pedigree”) of her data

She decides to start with one evening. There are 34 parties that night, so she records how long each of them waited to be seated from about 5 PM until about 8 PM. Her SIP is located in the figure below.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | SIP | | units | Minutes |
| 5 | | | about | This SIP describes the wait time from arrival to being seated in Franchise 448 |
| 6 | | | type | Excel |
| 7 | | | ver | 1.0 |
| 8 | | | count | 34 |
| 9 | | | provenance | These data were recorded on August 16, 2014. This was a Saturday evening. The times were recorded by the Franchise unit manager |
| 10 | | | name | Wait time |
| 11 | | | | |
| 12 | | | | 0 |
| 13 | | | | 2 |
| 14 | | | | 3 |
| 15 | | | | 7 |
| 16 | | | | 7 |
| 17 | | | | 5 |
| 18 | | | | 4 |
| 19 | | | | 2 |
| 20 | | | | 0 |
| 21 | | | | 0 |
| 22 | | | | 0 |
| 23 | | | | 3 |
| 24 | | | | 3 |
| 25 | | | | 4 |
| 26 | | | | 6 |
| 27 | | | | 6 |
| 28 | | | | 9 |
| 29 | | | | 1 |
| 30 | | | | 1 |
| 31 | | | | 0 |
| 32 | | | | 0 |
| 33 | | | | 2 |
| 34 | | | | 5 |
| 35 | | | | 3 |
| 36 | | | | 7 |

**Figure 1 – Example SIP in Excel (xlxs) format**

When she shares the SIP with her team, it generates a lot of comment. Some people wonder if the longer waiting times were caused by a rush. They wish time had been recorded. Some people think longer waiting times tended to be associated with larger groups. The servers think the long waiting times hurt their tips, and think people are less likely to order multiple courses when they have to wait. So the following week, the manager decides to track all these things from about 5 PM until about 8 PM. She records what happens with each party that comes to the restaurant. Because each party's information is recorded in order, she has *preserved the relationships* between each entry. When SIPs are organized in a manner that preserves relationships, they can become SLURPs. Her SLURP is located in the figure below.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | SLURP | | name | Waiting Data | | | | | |
| 4 | | | count | 5 | | | | | |
| 5 | | | coherent | TRUE | | | | | |
| 6 | | | | | | | | | |
| 7 | | | | | | | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | SIP | | units | Minutes | PDT | People | Courses | Percent | |
| 15 | | | about | This SIP de | The time of arrvial, Pacific Daylight Time | | | | |
| 16 | | | type | Excel | Excel | Excel | Excel | Excel | |
| 17 | | | ver | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 18 | | | count | 39 | 39 | 39 | 39 | 39 | |
| 19 | | | provenance | These data were recorded on August 23, 2014. This was a Saturday evening. The times were recorded by the Franchise unit manager | | | | | |
| 20 | | | name | Wait time | Arrival | Number | Courses | Tip | |
| 21 | | | | | | | | | |
| 22 | | | | 0 | 17:45 | 2 | 2 | 17 | |
| 23 | | | | 0 | 17:48 | 1 | 1 | 15 | |
| 24 | | | | 3 | 17:55 | 2 | 2 | 13 | |
| 25 | | | | 7 | 17:55 | 5 | 3 | 9 | |
| 26 | | | | 5 | 17:55 | 2 | 4 | 22 | |
| 27 | | | | 5 | 17:56 | 3 | 3 | 13 | |
| 28 | | | | 4 | 17:57 | 4 | 2 | 14 | |
| 29 | | | | 2 | 17:59 | 5 | 2 | 18 | |
| 30 | | | | 0 | 18:04 | 2 | 3 | 17 | |
| 31 | | | | 1 | 18:10 | 2 | 3 | 20 | |
| 32 | | | | 0 | 18:11 | 2 | 3 | 21 | |
| 33 | | | | 2 | 18:12 | 1 | 2 | 14 | |
| 34 | | | | 2 | 18:15 | 2 | 4 | 17 | |
| 35 | | | | 4 | 18:16 | 3 | 4 | 18 | |
| 36 | | | | 5 | 18:16 | 5 | 1 | 0 | |
| 37 | | | | 7 | 18:16 | 4 | 1 | 15 | |
| 38 | | | | 6 | 18:16 | 2 | 2 | 12 | |

**Figure 1 – Example SLURP in Excel (xlxs) format**

There are only two "about" records. The first one, in cell D15 is the same as shown in the SIP example. We can't see all that information because cell E15 has blocked it. We don't have "about" for the other data records, but that's not a problem, since the SIP specification says this information is optional. The same thing is true for provenance. We might have it for several of the SIPs, but in this case, our manager decided it was only needed for one of them. Now the manager has data recorded in a format that is easy to share and easy to compare. She can compare the waiting times on 16 August and 23 August. She can test to see what relationships exist in her SLURP. If she records the same kind of information next week, those can easily be compared in many ways.

At some point, the restaurant information may be examined by a sophisticated data scientist. The data science professional may decide the SLURPS are not really stochastic distributions. There may be artifacts of periodicity, or some seasonal trends. But the preserved coherence, and the "data about the data" means the data can be processed with advanced analysis methods. Also, the restaurant manager does not need to be a data scientist to provide useful data.

## SIP/SLURP Interface to TruNav℠

TruNav℠ is Lone Star's 5th generation enhanced decision support tool designed to address virtually any problem, and particularly those problems with significant analytical and organizational complexity.

In order to accommodate SIPs and SLURPs, TruNav℠ provides configurable input nodes, which are one-dimensional, unsorted arrays. Sampling randomly from the SIP is recommended. But, the SIP could be sampled in order. In cases where the SIP size is less than the number of Monte Carlo iterations, different combinations of alternatives are available: one could oversample the SIP randomly or fill the SIP to match the number of Monte Carlo iterations using interpolation and then sample randomly. A similar approach may be used if one desires to sample the SIP in order, i.e. oversample the SIP in order or first fill the SIP using interpolation and then sample in order. Again, the random sampling approach is recommended. As with SIPs, these recommendations apply when SLURPs are used. Just recall coherence needs to be preserved when sampling from more than one SIP.

The output data exports from TruNav℠ are currently a standardized .xls format that can be quickly adapted to generic SIPs and SLURPs formatting as required. Automated transfer of formatting between TruNav℠ native data formatting (which is more useful for internal data manipulation and transfer) and SIP / SLURP data formatting (which is more appropriate for external transfer of data sets) is a proposed future tool development.

## Summary

Interfacing multiple modeling environments together requires two things. First, that the transfer protocol of data between the modeling environments is standardized, and second, that the variables of interest are known and well defined. In order to address these requirements, it is recommended that data be transitioned to standardized SIP and/or SLURP format. In addition, it is recommended that definition of data classes expected to be transferred between modeling environments be jointly researched and specified.

## References

Scenario Optimization, Ron S. Dembo, Annals of Operations Research, 1991, Volume 30, Issue 1, pp 63-80. http://link.springer.com/article/10.1007%2FBF02204809

Probability Management, Sam Savage, Stefan Scholtes and Daniel Zweidler, OR/MS Today, February 2006, Volume 33 Number 1. http://www.lionhrtpub.com/orms/orms-2-06/frprobability.html

The Flaw of Averages, Why we Underestimate Risk in the Face of Uncertainty, Sam Savage, John Wiley 2009.

Calculating Uncertainty: Probability Management with SIP Math, John Marc Thibault, 2013.