Addendum for "An XES Extension for Uncertain Event Data"*

Marco Pegoraro ✉[0000−0002−8997−7517], Merih Seran Uysal[0000−0003−1115−6601], and Wil M.P. van der Aalst[0000−0002−0955−6940]

Chair of Process and Data Science (PADS)
Department of Computer Science, RWTH Aachen University, Aachen, Germany
{pegoraro,uysal,wvdaalst}@pads.rwth-aachen.de
http://www.pads.rwth-aachen.de/

This document clarifies the content of the paper "An XES Extension for Uncertain Event Data", with more details about the modeling of uncertain attributes in the XES standard and examples of uncertain data. In order to more clearly visualize the structure of the attributes in uncertain events, we are going to illustrate them with two examples of uncertain traces.

Table 1: The uncertain trace of an example of healthcare process. For the sake of clarity, we have further simplified the notation in the timestamps column by showing only the day of the month.

Case ID	Event ID	Timestamp	Activity	Indeterminacy
ID192	e_1	5	*NightSweats*	?
ID192	e_2	8	*PrTP*, *SecTP*	
ID192	e_3	4–10	*Splenomeg*	

Table 1 illustrates our first example. In this trace, the rightmost column refers to event indeterminacy: in this case, e_1 has been recorded, but it might not have occurred in reality, and is marked with a "?" symbol. Event e_2 has more then one possible activity labels, either *PrTP* or *SecTP*. Lastly, event e_3 has an uncertain timestamp, and might have happened at any point in time between the 4th and 10th of July.

In some cases, uncertain events have probability values associated with them. In the example described above, suppose the medic estimates that there is a high chance (90%) that the thrombocytopenia is primary (caused by the cancer). Furthermore, if the splenomegaly is suspected to have developed three days prior to the visit, which takes place on the 10th of July, the timestamp of event e_3 may be described through a Gaussian curve with $\mu = 7$. Lastly, the probability

* We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research interactions.

that the event e_1 has been recorded but did not occur in reality may be known (for example, it may be 25%).

Assigning such probabilities to data results in the trace shown in Table 2.

Table 2: A trace where uncertain event attributes are labeled with probabilities. In this case, we also have an indeterminate event.

Case ID	Event ID	Timestamp	Activity	Indeterminacy
ID348	e_4	5	*NightSweats*	? : 25%
ID348	e_5	8	*PrTP:* 90%, *SecTP:* 10%	
ID348	e_6	$\mathcal{N}(7,1)$	*Splenomeg*	

Let us now formally define uncertain attributes, events, traces, and logs.

Definition 1 (Uncertain attributes). *Let $\mathbb{U}$ be the* universe of attribute domains. *Let the set $\mathcal{D} \in \mathbb{U}$ be an* attribute domain. *Any $\mathcal{D} \in \mathbb{U}$ is a discrete set or a totally ordered set. A* strongly uncertain attribute *of domain $\mathcal{D}$ is a subset $d \subseteq \mathcal{D}$ if $\mathcal{D}$ is a discrete set, and it is a closed interval $d = [d_{min}, d_{max}]$ with $d_{min} \in \mathcal{D}$ and $d_{max} \in \mathcal{D}$ otherwise. We denote with $S_\mathcal{D}$ the set of all such strongly uncertain attributes of domain $\mathcal{D}$. A* weakly uncertain attribute *$f_\mathcal{D}$ of domain $\mathcal{D}$ is a function $f_\mathcal{D} \colon \mathcal{D} \not\to [0,1]$ such that $\sum_{x \in \mathcal{D}} f_\mathcal{D}(x) \leq 1$ if $\mathcal{D}$ is finite, $\int_{-\infty}^{\infty} f_\mathcal{D}(x)dx \leq 1$ otherwise. We denote with $W_\mathcal{D}$ the set of all such weakly uncertain attributes of domain $\mathcal{D}$. We collectively denote with $\mathcal{U}_\mathcal{D} = S_\mathcal{D} \cup W_\mathcal{D}$ the set of* uncertain attributes *of domain $\mathcal{D}$.*

Definition 2 (Uncertain events). *Let $\mathbb{U}_I$ be the* universe of event identifiers. *Let $\mathbb{U}_C$ be the* universe of case identifiers. *Let $A \in \mathbb{U}$ be the discrete domain of all the* activity identifiers. *Let $T \in \mathbb{U}$ be the totally ordered domain of all the* timestamp identifiers. *Let $O = \{?\} \in \mathbb{U}$, where the "?" symbol is a placeholder denoting* event indeterminacy. *The* universe of uncertain events *is denoted with $\mathcal{E} = \mathbb{U}_I \times \mathbb{U}_C \times \mathcal{U}_A \times \mathcal{U}_T \times \mathcal{U}_O$.*

Definition 3 (Uncertain traces and logs). *$\sigma \subsetneq \mathcal{E}$ is an* uncertain trace *if all the event identifiers in σ are unique and all events in σ share the same case identifier $c \in \mathbb{U}_C$. $\mathcal{T}$ denotes the universe of uncertain traces. $L \subsetneq \mathcal{T}$ is an* uncertain log *if all the event identifiers in L are unique.*

In the notation of Definitions 1, 2 and 3, the traces σ_1 in Table 1 and σ_2 in Table 2 are denoted as:

$$\begin{aligned}\sigma_1 = \{&(e_1, \text{ID192}, \{\textit{NightSweats}\}, [5,5], \{?\}),\\ &(e_2, \text{ID192}, \{\textit{PrTP}, \textit{SecTP}\}, [8,8], \varnothing),\\ &(e_3, \text{ID192}, \{\textit{Splenomeg}\}, [4,10], \varnothing)\}\end{aligned}$$

$$\begin{aligned}\sigma_2 = \{&(e_1, \text{ID348}, \{\textit{NightSweats}\}, [5,5], \{(?, 0.25)\}),\\ &(e_2, \text{ID348}, \{(\textit{PrTP}, 0.85), (\textit{SecTP}, 0.15)\}, [8,8], \varnothing),\\ &(e_3, \text{ID348}, \{\textit{Splenomeg}\}, \mathcal{N}(7,1), \varnothing)\}\end{aligned}$$

The attribute domains are[1]:

$$\begin{aligned} A =& \{NightSweats, PrTP, SecTP, Splenomeg\} \\ T =& \mathbb{N} \\ O =& \{?\} \end{aligned}$$

Examples of uncertain attributes are:

$$\begin{aligned} S_A =& \{\{PrTP, SecTP\}, \{NightSweats, PrTP\}, \{Splenomeg, PrTP, SecTP\}, \dots\} \\ S_T =& \{[5,5], [8,8], [4,10], [1,1], [10,12], [10,16], \dots\} \\ S_O =& \{\varnothing, \{?\}\} \\ W_A =& \{\{(PrTP, 0.85), (SecTP, 0.15)\}, \{(NightSweats, 0.90)\}, \\ & \{(Splenomeg, 0.70), (PrTP, 0.20), (SecTP, 0.10)\}, \dots\} \\ W_T =& \{\mathcal{N}(7,1), U(4,10), \Gamma(3,2), \dots\} \\ W_O =& \{\{(?, 0.25)\}, \{(?, 0.05)\}, \{(?, 0.90)\}, \dots\} \end{aligned}$$

Note that, while the most usual case would involve label attribute values with a complete probability distribution (probabilities summing to 1), here we allow for a sum ≤ 1, to enable maximum flexibility in uncertain data representation.

This mathematical framework allows to represent events with uncertain attributes, both strongly and weakly uncertain, and both in the discrete and continuous domains. We will now see how to represent such events in the XES standard.

In this extension, discrete strongly uncertain attributes are represented by a container of data with any type: this represents a set of arbitrary objects, which are the possible values of the uncertain attribute. In the totally ordered case, the uncertain attribute is modeled by a list of two sorted values. Such values represent the extremes of an interval in which the values of the uncertain attribute can range. The following code snippet contains the full representation of the trace in Table 1.

```
<trace>
 <string key="concept:name" value="ID192"/>
 <event>
  <string key="concept:name" value="NightSweats"/>
  <date key="time:timestamp" value="2011-07-05T12:00:00+00:00"/>
  <container key="uncertainty:discrete_strong">
   <bool key="uncertainty:indeterminacy" value="true"/>
  </container>
 </event>
 <event>
  <string key="concept:name" value="PrTP"/>
  <date key="time:timestamp" value="2011-07-08T12:00:00+00:00"/>
  <container key="uncertainty:discrete_strong">
```

[1] Here, we defined the timestamp domain as the set $\mathbb{N}$ of natural numbers. The usual mathematical notation is unwieldy and unsuitable to represent complete timestamps as normally read and represented by humans; however, it is easy to see how a precise date and time can be represented by an integer without loss of information through conventions such as the Unix time (seconds since the Epoch, or fractions thereof).

```
    <string key="concept:name" value="PrTP"/>
    <string key="concept:name" value="SecTP"/>
   </container>
  </event>
  <event>
   <string key="concept:name" value="Splenomeg"/>
   <date key="time:timestamp" value="2011-07-07T12:00:00+00:00"/>
   <list key="uncertainty:continuous_strong">
    <date key="time:timestamp" value="2011-07-04T12:00:00+00:00"/>
    <date key="time:timestamp" value="2011-07-10T12:00:00+00:00"/>
   </list>
  </event>
</trace>
```

Weak uncertainty is also modeled by our extension. In this scenario, the discrete attributes are represented by a container of `uncertainty:entry` objects, which are pairs constituted by an attribute value and its probability. Lastly, the totally ordered case is described by a probability function, which is identified by a key and a set of parameters. We can see an example of these in the representation of the trace in Table 2, contained in the following code snippet.

```
<trace>
 <string key="concept:name" value="ID192"/>
 <event>
  <string key="concept:name" value="NightSweats"/>
  <date key="time:timestamp" value="2011-07-05T12:00:00+00:00"/>
  <container key="uncertainty:discrete_weak">
   <container key="uncertainty:entry">
    <bool key="uncertainty:indeterminacy" value="true"/>
    <double key="uncertainty:probability" value="0.25"/>
   </container>
  </container>
 </event>
 <event>
  <string key="concept:name" value="PrTP"/>
  <date key="time:timestamp" value="2011-07-08T12:00:00+00:00"/>
  <container key="uncertainty:discrete_weak">
   <container key="uncertainty:entry">
    <string key="concept:name" value="PrTP"/>
    <double key="uncertainty:probability" value="0.90"/>
   </container>
   <container key="uncertainty:entry">
    <string key="concept:name" value="SecTP"/>
    <double key="uncertainty:probability" value="0.10"/>
   </container>
  </container>
 </event>
 <event>
  <string key="concept:name" value="Splenomeg"/>
  <date key="time:timestamp" value="2011-07-07T12:00:00+00:00"/>
  <container key="uncertainty:continuous_weak">
   <string key="uncertainty:density_function" value="GAUSSIAN"/>
   <list key="uncertainty:function_parameters">
    <double key="parameter_mean" value="7"/>
    <double key="parameter_stddev" value="1"/>
   </list>
  </container>
 </event>
</trace>
```

A set of synthetic uncertain event logs is publicly available for download[2].

[2] https://github.com/proved-py/proved-core/tree/An_XES_Extension_for_Uncertain_Event_Data/data