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/

Abstract. Event data, often stored in the form of event logs, serve as the starting point for process mining and other evidence-based process improvements. However, event data in logs are often tainted by noise, errors, and missing data. Recently, a novel body of research has emerged, with the aim to address and analyze a class of anomalies known as *uncertainty*—imprecisions quantified with meta-information in the event log. This paper illustrates an extension of the XES data standard capable of representing uncertain event data. Such an extension enables input, output, and manipulation of uncertain data, as well as analysis through the process discovery and conformance checking approaches available in literature.

Keywords: Event Data · Uncertainty · XES Standard · Process Mining · Business Process Management.

1 Introduction

Through the last decades, the increase in the availability of data generated by the execution of processes has enabled the development of the set of disciplines known as process sciences. These fields of science aim to analyze data accounting for the process perspective—the flow of events belonging to a process case.

Uncertain event data is a newly-emerging class of anomalous event data. Uncertain data consists of events that have been logged with a quantified measure of uncertainty affecting the recorded information. Sources of uncertainty include noise, human error, or limitations of the information system supporting the process. Such imprecisions affecting the event data are either recorded in an information system with the data itself or reconstructed in a subsequent processing step, often with the aid of domain knowledge provided by process experts. Recently, the possible types of uncertain data have been classified in a taxonomy, and effective process mining algorithms for uncertain event data have been introduced [7,9]. However, the data standards currently in use within the process

---

science community do not support uncertain event logs. A very popular event data standard is XES (eXtensible Event Stream) [11,1]. As the name suggest, this standard has been designed to flexibly allow for extensions; in the recent past, many such extensions have been proposed, to support communications, messages and signals [5], usage and performance of hardware resources [6], and privacy-preserving data transmission [10]. This paper contributes to the field of process science by describing an XES extension which allows the representation of uncertain data, enabling XES-compatible tools to manipulate uncertain logs. Furthermore, our extension is implemented through the meta-attribute structure already supported by XES, making uncertain data retroactively readable by existing tools.

The remainder of the paper is structured as follows. Section 2 formally describes uncertain event data. Section 3 introduces an extension to the XES standard capable of representing uncertain event data. Lastly, Section 4 concludes the paper.

2   Uncertain Event Data

In order to more clearly visualize the structure of the attributes in uncertain events, let us consider the following process instance, which is a simplified version of actually occurring anomalies, e.g., in the processes of the healthcare domain. An elderly patient enrolls in a clinical trial for an experimental treatment against myeloproliferative neoplasms, a class of blood cancers. This enrollment includes a lab exam and a visit with a specialist; then, the treatment can begin. The lab exam, performed on the 8th of July, finds a low level of platelets in the blood of the patient (event e_2), a condition known as thrombocytopenia (TP). During the visit on the 10th of July, the patient reports an episode of night sweats on the night of the 5th of July, prior to the lab exam (event e_1). The medic notes this but also hypothesizes that it might not be a symptom, since it can be caused either by the condition or by external factors (such as very warm weather). The medic also reads the medical records of the patient and sees that, shortly prior to the lab exam, the patient was undergoing a heparin treatment (a blood-thinning medication) to prevent blood clots. The thrombocytopenia, detected by the lab exam, can then be either primary (caused by the blood cancer) or secondary (caused by other factors, such as a concomitant condition). Finally, the medic finds an enlargement of the spleen (splenomegaly) in the patient (event e_3). It is unclear when this condition has developed: it might have appeared at any moment prior to that point. These events are collected and recorded in the trace shown in Table 1 within the hospital's information system.

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 than one possible activity label, either $PrTP$ or $SecTP$ (primary or secondary thrombocytopenia, respectively). Lastly, event e_3 has an uncertain timestamp, and might have happened at any point in time between the 4th and 10th of July. These uncertain attributes do not

Table 1: The uncertain trace of an instance of healthcare process used as a running example. 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*	

describe the probability of the possible outcomes, and we refer to such situation as *strong uncertainty*.

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 $\mu = 7$. When probability is available, such attributes are affected by *weak uncertainty*.

Let us now describe a data standard extension able to represent strong and weak uncertainty, enabling the analysis of uncertain data with process science techniques.

3 An XES Standard Extension Proposal

The XES standard is designed to effectively represent and transfer event data, thanks to the descriptors extended from the XML language. Additionally, XES has been designed for flexibility: its descriptors, containers, and datatypes can be extended to define new types of information.

Figure 1 describes an extension of the XES standard able to represent uncertain data as described in the previous section and illustrated in the running example of Table 1.

This proposed extension can represent all scenarios of uncertain data shown in Section 2. As a consequence, it enables XES-compliant software to import and export uncertain event data, and it allows uncertainty-aware process mining tools to implement process discovery and conformance checking approaches on uncertain data, as described in the literature.

An example of a tool able to exploit this extended XES representation to manage and analyze uncertain event data is the PROVED project[1], which offers process mining and data visualization techniques capable of handling uncertain event data [8].

It is important however to emphasize the fact that the use of the extension described here is not limited to the PROVED tool. There exist multiple tools able to support the XES standard, such as ProM [3], bupaR [4], and PM4Py [2]. Each of these tools is able to edit attributes, meta-attributes and values in a XES

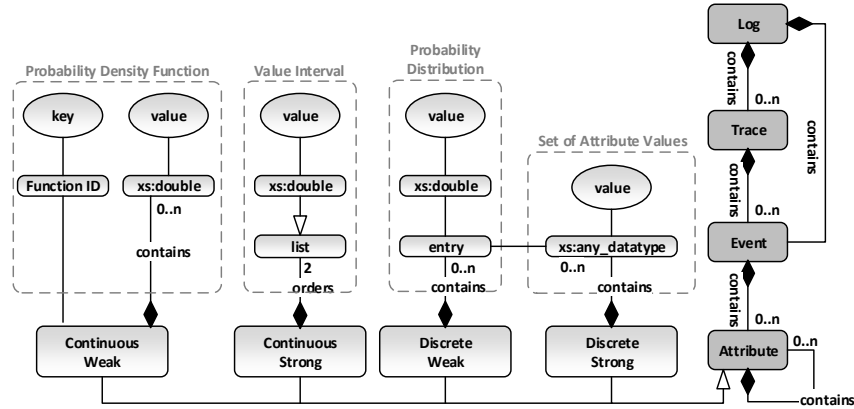---

[1] https://github.com/proved-py/

Fig. 1: UML diagram illustrating an extension of the XES standard capable of representing uncertain data.

event log, and is then capable to record uncertain attributes on process traces. In summary, while uncertainty-aware analysis techniques are only available on a narrow selection of tools (such as PROVED), this extension benefits any tool that supports XES as one of its input/output data standards.

A set of synthetic uncertain event logs is publicly available for download[2]. In the same folder, it is possible to find the additional document (part of the BPM Resource track submission) explaining more in detail how our extension proposal models uncertain event data[3].

4   Conclusion

Recent literature in the rapidly-growing field of process mining shows how descriptions of specific data anomalies can be extracted from information systems or obtained through domain knowledge. Anomalies labeled by such descriptions characterize uncertain event data, and there exist process mining algorithms able to exploit this meta-information to gain insights about the process with a precisely bounded reliability. A fundamental part of these data analysis approaches is however needed: formats for data representation and transmission. In this paper, we described an extension of the XES data standard which enables representation of such uncertain data, and that allows uncertain event to be read and

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

[3] https://github.com/proved-py/proved-core/blob/An_XES_Extension_for_Uncertain_Event_Data/data/uncertainty_XES_standard.pdf. A version of this document is reproduced in Appendix A.

written by existing XES-compliant software. This, in turn, empowers process mining researchers and practitioners to build analysis techniques that account for data uncertainty, and that can thus be more trustworthy and reliable.

References

1. van der Aalst, W.M.P., Günther, C., Bose, J., Carmona, J., Dumas, M., van Geffen, F., Goel, S., Guzzo, A., Khalaf, R., Kuhn, R., et al.: 1849–2016—IEEE Standard for eXtensible Event Stream (XES) for achieving interoperability in event logs and event streams. No. IEEE Std 1849TM-2016 (Sep 2016), http://hdl.handle.net/2117/341493
2. Berti, A., van Zelst, S.J., van der Aalst, W.M.P.: Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science. In: ICPM Demo Track (CEUR 2374). p. 13–16. CEUR-WS.org (2019)
3. van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: The ProM framework: A new era in process mining tool support. In: Ciardo, G., Darondeau, P. (eds.) Applications and Theory of Petri Nets 2005, 26th International Conference, ICATPN 2005, Miami, USA, June 20-25, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3536, pp. 444–454. Springer (2005)
4. Janssenswillen, G., Depaire, B.: bupaR: Business process analysis in R. In: Clarisó, R., Leopold, H., Mendling, J., van der Aalst, W.M.P., Kumar, A., Pentland, B.T., Weske, M. (eds.) Proceedings of the 15th International Conference on Business Process Management (BPM 2017), Barcelona, Spain, September 13, 2017. CEUR Workshop Proceedings, vol. 1920. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1920/BPM_2017_paper_193.pdf
5. Leemans, M., Liu, C.: XES Software Communication Extension. XES Working Group pp. 1–5 (2017)
6. Leemans, M., Liu, C.: XES Software Telemetry Extension. XES Working Group pp. 1–7 (2017)
7. Pegoraro, M., van der Aalst, W.M.P.: Mining Uncertain Event Data in Process Mining. In: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019. pp. 89–96. IEEE (2019)
8. Pegoraro, M., Uysal, M.S., van der Aalst, W.M.P.: PROVED: A Tool for Graph Representation and Analysis of Uncertain Event Data. In: Buchs, D., Carmona, J. (eds.) Application and Theory of Petri Nets and Concurrency. pp. 476–486. Springer International Publishing, Cham (2021)
9. Pegoraro, M., Uysal, M.S., van der Aalst, W.M.P.: Conformance checking over uncertain event data. Information Systems **102**, 101810 (2021). https://doi.org/https://doi.org/10.1016/j.is.2021.101810
10. Rafiei, M., van der Aalst, W.M.P.: Privacy-preserving data publishing in process mining. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings. Lecture Notes in Business Information Processing, vol. 392, pp. 122–138. Springer (2020)
11. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: XES, XESame, and ProM 6. In: Soffer, P., Proper, E. (eds.) Information Systems Evolution - CAiSE Forum 2010, Hammamet, Tunisia, June 7-9, 2010, Selected Extended Papers. Lecture Notes in Business Information Processing, vol. 72, pp. 60–75. Springer (2010)

A   XES Standard for Uncertain Event 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 2: 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 2 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 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 3.

Table 3: 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 2 and σ_2 in Table 3 are denoted as:

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

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

The attribute domains are[4]:

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

---

[4] 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).

Examples of uncertain attributes are:

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

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.

```
1   <trace>
2   <string key="concept:name" value="ID192"/>
3   <event>
4   <string key="concept:name" value="NightSweats"/>
5   <date key="time:timestamp" value="2011-07-05T12:00:00+00:00"/>
6   <container key="uncertainty:discrete_strong">
7   <bool key="uncertainty:indeterminacy" value="true"/>
8   </container>
9   </event>
10  <event>
11  <string key="concept:name" value="PrTP"/>
12  <date key="time:timestamp" value="2011-07-08T12:00:00+00:00"/>
13  <container key="uncertainty:discrete_strong">
14  <string key="concept:name" value="PrTP"/>
15  <string key="concept:name" value="SecTP"/>
16  </container>
17  </event>
18  <event>
19  <string key="concept:name" value="Splenomeg"/>
20  <date key="time:timestamp" value="2011-07-07T12:00:00+00:00"/>
21  <list key="uncertainty:continuous_strong">
22  <date key="time:timestamp" value="2011-07-04T12:00:00+00:00"/>
23  <date key="time:timestamp" value="2011-07-10T12:00:00+00:00"/>
24  </list>
25  </event>
26  </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 3, contained in the following code snippet.

```
1   <trace>
2   <string key="concept:name"  value="ID192"/>
3   <event>
4   <string key="concept:name"  value="NightSweats"/>
5   <date key="time:timestamp"  value="2011−07−05 T12:00:00+00:00"/>
6   <container key="uncertainty:discrete_weak">
7   <container key="uncertainty:entry">
8   <bool key="uncertainty:indeterminacy"  value="true"/>
9   <double key="uncertainty:probability"  value="0.25"/>
10  </container>
11  </container>
12  </event>
13  <event>
14  <string key="concept:name"  value="PrTP"/>
15  <date key="time:timestamp"  value="2011−07−08 T12:00:00+00:00"/>
16  <container key="uncertainty:discrete_weak">
17  <container key="uncertainty:entry">
18  <string key="concept:name"  value="PrTP"/>
19  <double key="uncertainty:probability"  value="0.90"/>
20  </container>
21  <container key="uncertainty:entry">
22  <string key="concept:name"  value="SecTP"/>
23  <double key="uncertainty:probability"  value="0.10"/>
24  </container>
25  </container>
26  </event>
27  <event>
28  <string key="concept:name"  value="Splenomeg"/>
29  <date key="time:timestamp"  value="2011−07−07 T12:00:00+00:00"/>
30  <container key="uncertainty:continuous_weak">
31  <string key="uncertainty:density_function"  value="GAUSSIAN"/>
32  <list key="uncertainty:function_parameters">
33  <double key="parameter_mean"  value="7"/>
34  <double key="parameter_stddev"  value="1"/>
35  </list>
36  </container>
37  </event>
38  </trace>
```

A set of synthetic uncertain event logs is publicly available for download[5].

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