UPX: A New XML Representation for Annotated Datasets of Online Handwriting Data

Mudit Agrawal, Kalika Bali and Sriganesh Madhvanath
HP Labs, Bangalore, India
mudit.a,kalika,srig@hp.com

Louis Vuurpijl
NICI, Nijmegen, The Netherlands
vuurpijl@nici.ru.nl

Abstract

This paper introduces our efforts to create UPX, an XML-based successor to the venerable UNIPEN format for the representation of annotated datasets of online handwriting data. In the first part of the paper, shortcomings of the UNIPEN format are discussed and the goals of UPX are outlined. Prior work related to UPX in the form of the recently proposed hwDataset representation is presented. The second part of the paper summarizes the status of the UPX effort, in particular, experiments to map UNIPEN elements to hwDataset and InkML and identify potential issues with migrating existing UNIPEN data to UPX. This is work in progress, and we invite participation from the handwriting recognition research community and industry to make UPX a reality.

1. Introduction

Linguistic Resources are critical for the development of any human language technology, and handwriting recognition is no exception. The UNIPEN consortium [5] formed in the early 1990s, was one of the first to address the need for standard linguistic resources for online handwriting recognition. The eponymous representation for digital ink as well as its annotation has become a *de facto* standard for online handwriting corpora, and has been used for significant data collection efforts in recent times [6, 7].

The UNIPEN representation employs ASCII flat files to store handwriting data and associated annotation. This brings with it the advantages of simplicity, ease of viewing and editing using a simple text editor. The representation is also extensible in that it allows the definition of additional keywords to describe additional attributes of the data or the writers.

However UNIPEN suffers from some shortcomings. In particular,

- UNIPEN is unstructured. There is no way of organizing the information in semantically well-categorized classes such as dataset information, writer definitions, label sources, or annotation hierarchy. Instead, UNIPEN provides a number of keywords that can be specified in any order.
- UNIPEN is not strict. Moreover, many relevant aspects of the data collection process and of the data itself, are described in the UNIPEN *.COMMENT* expressions. Also, keywords like *.SETUP* often contain information that cannot be automatically extracted, such as information about writers, recording device, software, form layout, etcetera.
- UNIPEN has a scope problem. Given that the order in which keywords are entered is not fixed, the scope of UNIPEN expressions is defined as follows: Any coordinate that is specified in UNIPEN, is described by the context of preceding UNIPEN tags. Any UNIPEN tags that are specified below other tags, are not valid for these tags.
- The focus of the original UNIPEN effort was the recognition of cursive English text, and support for non-Latin scripts, and for modalities such as drawings and math is limited at best.

This paper describes our efforts to define UPX, an XML-based successor to UNIPEN which addresses the shortcomings of UNIPEN, while providing a path for migrating existing UNIPEN databases to the new representation. In addition, the UPX effort is an attempt to create the first standard representation for handwriting datasets that (i) supports all scripts and allows semantic interpretation of the writing at various user-defined logical levels by multiple annotators, (ii) captures information about script, writing style, quality of writing and truth, (iii) captures information about writers and the data capture environment, (iv) supports automatic generation of annotation using recognizers, and subsequent manual validation processes, (v) keeps handwriting data separate from its semantic interpretations and (vi) supports planned as well as casual data collection.

The format builds on our previous work on the hwDataset representation and InkML, a draft standard for the representation of digital ink from the World Wide Web Consortium. The section that follows provides an introduction to the hwDataset format and describes enhancements in the most recent release (version 0.5, Dec 04). The section also lists research issues with hwDataset and briefly describes the annotation tool that supports this representation.

The third section describes the status of our efforts to evaluate hwDataset from the perspective of supporting migration of existing UNIPEN data. Some conclusions and current directions are presented in the final section.

2. hwDataset

hwDataset [2, 3] was proposed recently as an XML representation for the annotation of handwriting data that is inspired by the UNIPEN standard. XML is a natural choice for the representation of annotation because of its hierarchical nature and extensibility [1, 4]. The hwDataset representation in turn makes use of Digital Ink Markup Language (InkML) for the representation of the digital ink being annotated.

2.1 InkML

Digital ink refers to a series of pen positions and optional attributes (related to time-stamp, pen pressure, pen tilt and so forth) captured from a suitable pen input device. Recognition of handwriting captured as digital ink is known as online handwriting recognition. Today there are literally thousands of different digital ink aware devices available ranging from standalone digitizing tablets, to PDAs, Tablet-PCs and mobile phones, and proprietary devices for different vertical markets - supporting different proprietary representations of digital ink.

Digital Ink Markup Language (InkML) [8] from the World Wide Web Consortium (W3C) is a emerging standard for the platform and device-independent representation of digital ink. InkML markup is designed to support the input, storage and processing of handwriting, gestures, sketches, music and other notational languages in ink-aware applications, independent of platform. InkML also provides a common format for the exchange of ink data between components such as handwriting and gesture recognizers, signature verifiers, and other ink-aware modules. Although InkML provides many proper features when it comes to the specification of digital ink, it lacks certain more advanced aspects necessary for annotation.

Fortunately InkML provides means for application-specific extensions. By virtue of being an XML-based language, it allows users to easily add specific information to ink files to suit the needs of the application at hand. In this sense, hwDataset may be thought of as an application-specific extension of InkML (Figure 1).

Figure 1. Conceptual relationship between hwDataset and InkML

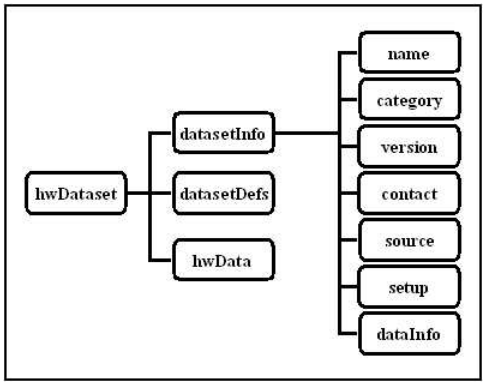

Figure 2. The *datasetInfo* element captures metadata about the dataset

The hwDataset representation includes a set of XML elements for detailed annotation of handwriting. The hwDataset element is the root of the document and captures metadata about the dataset as part of the *datasetInfo* element, various definitions as a part of *datasetDefs*, and hierarchical annotation of handwritten data as a part of one or more *hwData* elements. These elements are described briefly in the following subsections.

2.2 datasetInfo

The *datasetInfo* element (Figure 2) captures metadata related to the dataset as a whole. It contains the following elements: (a) *name* - name for referring to the dataset, (b) *category* - type of dataset captured using UNIPEN-style codes, (c) *version* - version number and/or date-stamp of dataset publication, (d) *contact* - contact info for dataset related queries, (e) *source* - the source of collected data, (f) *setup* - physical conditions in which the data was collected,

and (g) *dataInfo* - information about the data.

The *dataInfo* element in turn contains the following sub-elements: (a) *contentDesc* - general description of content of dataset, e.g. scripts, writing styles etc., (b) *numWriters* - number of writers contributing data, (c) *quality* - overall assessment of quality of handwritten data captured in dataset, (d) *style* - overall writing style of data, (e) *truthRef* - reference to file containing transcription of the reference text.

2.3 datasetDefs

The *datasetDefs* element captures information about different writers and sources of labels (annotation) represented in the dataset, and provides the means for referring to them later in the document. It contains the following elements:

- *writerDefs* - declarations of writers as a sequence of *writer* elements
- *labelSrcDefs* - declarations of sources of annotation (human or machine) as a sequence of *labelSrc* elements
- *annotationDefs* - definitions of various annotation schemes used in the dataset as a sequence of *annotationScheme* elements

Each *writer* element in turn contains three sub-elements: (a) *personal* - captures personal information such as *hand*(left/right handedness), *educationLevel*(highest level of education), *gender*, *profession*, *region*(native region) and *dateOfBirth*, (b) *skillDevice* - level of familiarity with the writing device, (c) *skillScript* - level of skill with each script present in the dataset, in turn described in terms of *style*, *usageFreq* and *proficiency*.

Each *labelSrc* element contains the following sub-elements: (a) *name* - name of the human/automated source of labels, (b) *source* - organization that label source represents, (c) *contact* - contact details of label source, and (d) *desc* - descriptive details.

In addition, an attribute *labelTypes* describes the categories of labels (e.g. truth, quality, script, style) generated by the given source and their character encoding (e.g. UNICODE).

Each *annotationScheme* element specifies the user-defined hierarchy of annotation such as PAGE, PARAGRAPH, LINE, WORD, CHAR by means of a series of *annotationLevel* elements.

2.4 hwData

The hwDataset document may contain one or more *hwData* elements corresponding to different writing trials, or different fields of writing captured from a writer in a single trial. These instances may be distinguished using the *id* attribute. Each *hwData* element follows one of the annotation schemes defined earlier. It contains one or more $H(i)$ elements, where i refers to an appropriate level of the annotation hierarchy defined by the user as a part of the specification of the *annotationScheme* element (Figure 3). Each $H(i)$ contains one or more label elements that capture annotation information at that level. In addition, $H(i)$ may in turn contain either one or more $H(i+1)$ elements, or *hwTraces*, the leaf element of the hierarchy that refers to digital ink traces represented using InkML

hwData also includes a *uiInfo* element that describes the writing area or field used to capture ink.

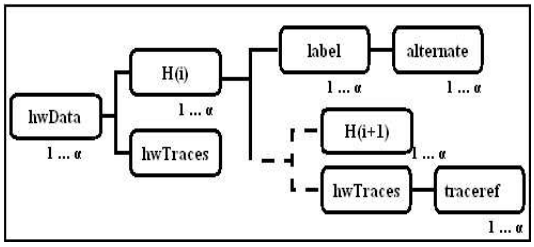

Figure 3. *hwData* **element showing hierarchical organization of annotation**

label The *label* element is the chief mechanism for annotation of handwriting data in hwDataset. Any number of *label* elements may be associated with a particular $H(i)$ element, and each element can be used to capture exact time of annotation with alternative choices of label with confidence values if any. Although primarily intended to describe the truth value of a particular set of ink traces, it may also be used for describing other characteristics such as writing style, quality and script. The timestamp can be used to generate the history of annotation of a particular unit of writing, spanning different label sources. The alternates can be used to facilitate the process of semi-automatic annotation by prompting automatically generated options for human validation.

Formally, the attributes of label are (a) *id* - identification of label, (b) *labelSrcRef* - a reference to a label source defined earlier, (c) *labelType* - type of label (e.g. truth, quality, script, style), and (d) *timestamp* - time of the act of annotation.

The hwDataset representation attempts to satisfy some core requirements for the creation of annotated handwriting datasets in different languages. Script-independence is achieved by supporting different encoding standards for the truth values. The representation supports semantic interpretation of the writing at various user-defined logical levels and captures information about script, style, quality at these levels. In addition, these attributes may also be associated

with the dataset as a whole, or with specific writers. Attributes have closed sets of values wherever possible.

2.5 Handwriting Annotation Tool

HWAT (Handwriting Annotation Tool) is a graphical tool for the annotation of online handwriting data that natively supports the InkML and hwDataset representations. While the tool is designed to read and write hwDataset documents, it is also capable of importing digital ink in input formats such as InkML, UNIPEN, and simple ASCII encodings of trace data. The tool supports input and output, viewing, editing and annotation of hwDataset documents at different levels of a user-defined annotation hierarchy. The tool is supplemented by a library of basic functions that can be used to access and extract handwriting data from hwDataset documents based on user-defined criteria.

The tool implements an open and extensible architecture using plug-ins for different operations such as segmentation and recognition of units at different levels of the annotation hierarchy. Segmentation plug-ins are implemented for common hierarchical levels such as strokes, words, and lines. The tool also allows multiple plug-ins for the same operation (for example, line segmentation) and selection of a specific plug-in at the beginning of the annotation session. This allows for customization and dynamic selection of these modules. In addition, word recognition plug-ins may be used to partially or fully automate the generation of ground-truth for handwriting data. Since all the plug-ins for a given class of operations return results in standard formats, they are handled within the tool in a consistent manner. Sample plug-ins are provided along with the tool, and new plug-ins may be written in C++.

2.6 Current Status

The first complete version of hwDataset and the HWAT tool was presented in November 2004 [3]. Since then, advances have been made to resolve some of the open issues with the format, and address completeness of the representation. Some of the specific improvements are discussed below.

Multiple annotation hierarchies: A general digital ink document may contain text, mathematical equations, figures and so on. Each of these categories of data in general requires a different annotation hierarchy. To support such scenarios, the *annotationDefs* element now allows the definition of multiple distinct hierarchy schemes, each with distinct annotation levels. Each *hwData* block refers to one hierarchy scheme from among those defined. However, once defined for the *hwData* block, the semantics associated with different levels (H1, H2 etc.) within the block is fixed.

Distribution of dataset across multiple documents: A complex element such as *datasetDefs* can now refer to a similar element in another dataset file using the *href* attribute. This allows shared information to be represented once and referred to elsewhere. For example, instead of repeating information in every file, the *writerDefs* or *labelSrcDefs* elements can refer to definitions of writers or labelsources respectively stored in a common document.

User interface elements: Different digitizers and pen-aware devices assume different positions of origin (e.g. top-left, bottom-left, middle etc.) which is central to the interpretation of digital ink. Moreover, different horizontal and vertical reference lines may characterize the writing area used for handwriting data collection. The *uiInfo* element, a sub-element of *hwData* supports the representation of such attributes of the input field.

With any representation, there is a clear trade off between flexibility and completeness of the representation on the one hand, and its complexity on the other (and that of tools that have to support the representation). Similar issues exist with hwDataset. An open research question is support for heterogeneous hierarchies, wherein each node can have children of different types. One can imagine many contexts where this may be needed, for example, to describe a handwritten document that may be decomposed into writing and drawing dominated subregions, which in turn may be similarly decomposed.

Similarly, the distribution of the dataset across documents also raises research questions. The present model is that when reference is used, the element may not be defined locally. However, other models are possible for supporting common information. Some of the open issues are those around granularity of the information that is shared, what happens when the same information is present both locally and in a shared manner, and so forth.

These questions are now being studied in the context of the UPX effort.

3. UPX: A Status Report

Whereas hwDataset was created primarily to support new data collection, the starting point for the UPX effort has been to assess the validity of hwDataset for storing existing handwriting data repositories. We have performed a case study which entailed the transformation of a relevant collection of UNIPEN data into hwDataset. The results of this case study will be used as the basis of the definition of UPX.

UNIPEN has been the result of a large effort in which numerous institutes and commercial organizations have provided data. We may conclude that because of the heterogeneity of the data (different writers, different recording

conditions, different languages and writing setup), UNIPEN provides an excellent test case for the assessment of hwDataset. Actually, it was one of the goals of the UNIPEN collection efforts to provide such variety.

Since UNIPEN and hwDataset have very similar goals, they are functionally quite similar though they might accomplish the same ends differently. For instance, in UNIPEN, sharing of common information such as writer information is accomplished by means of .INCLUDE statements. Different "views" of the dataset can also be supported by keeping digital ink in separate UNIPEN files and including them. In hwDataset, sharing of common information as well as digital ink across views is accomplished by means of references.

We have attempted to map at a granular level UNIPEN keywords to hwDataset or InkML elements. While a 1:1 mapping exists for most keywords, we have observed some differences, such as those described below.

Timestamp: Detailed recording of timing information is a critical part of any representation of digital ink. The most general approach is a *time channel* which allows for detailed recording of timing information for each sample point within a trace. For devices with uniform sampling rate, timestamps may be used rather than a time channel. The absolute time, or timestamp relative to a reference time at the beginning and the end of the stroke may be recorded. UNIPEN supports time channels, whereas InkML supports both timestamps and time channels.

Trace ranges: traceRef is the primary mechanism provided in InkML for referring to digital ink for the purpose of annotation. A *traceRef* may refer to either a single *trace* or a *traceGroup* by name. In the former instance, the *from* and *to* attributes of *traceRef* may be used to indicate a segment of the trace. In the latter instance, these attributes may be used to indicate a contiguous range of traces within a *traceGroup*. In UNIPEN a contiguous range of traces can be specified using just the trace indices, which is very useful for annotating large chunks of digital ink. In InkML, this is only currently supported within a *traceGroup*, and there is currently no way to index traces across *traceGroups*. However since InkML is still being evolved, it is our hope that such observations, and other results from this mapping exercise can be fed back to the InkML effort in order to improve the support in InkML for annotated datasets of handwriting and other ink modalities such as drawings and math.

4. Conclusions

In this paper, we have presented a status update on efforts to define UPX, a new XML representation for digital ink and its annotation, to succeed UNIPEN, the present de facto standard. This effort will build on other efforts to define Digital Ink Markup Language (InkML) from W3C, and the recently published hwDataset representation and tools, while addressing new realities - for example, the fact that research on (pure) handwriting recognition has shifted toward free writing conditions with multiple modes, including drawing, sketching and gestures. UPX, with support for heterogeneous hierarchies will provide a structured solution for addressing handwriting databases of the future, while accommodating ones from the past, such as those collected using the UNIPEN representation. As a first step in this direction, a conversion of UNIPEN data and format to hwDataset is being attempted, and the resulting observations and issues are being worked on, as part of the UPX effort and in collaboration with the InkML community. This is work in progress, and we invite participation from the handwriting recognition research community and industry to make UPX a reality.

References

[1] A. P. Lenaghan, R. R. Malyan. XPEN: An XML Based Format for Distributed Online Handwriting Recognition . In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1270–1274, 2003.

[2] A. S. Bhaskarabhatla and S. Madhvanath. An XML Representation for Annotated Handwriting Datasets for Online Handwriting Recognition. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, May 2004.

[3] A. S. Bhaskarabhatla, M. Pavan Kumar, A. Balasubramanian, C. Jawahar, and S. Madhvanath. Representation and Annotation of Online Handwritten Data. In *Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR-9)*, Tokyo, Japan, October 2004.

[4] K. Franke, L. Schomaker, C. Veenhuis, L. Vuurpijl, M. van Erp, and I. Guyon. WANDA: A Common Ground for Forensic Handwriting Examination and Writer Identification. *ENFHEX news - Bulletin of the European Network of Forensic Handwriting Experts*, (1/04):23–47, 2004.

[5] International Unipen Foundation. The UNIPEN Project. *http://www.unipen.org*, 1994.

[6] M. Nakagawa and K. Matsumoto. Collection of On-line Handwritten Japanese Character Pattern Databases and their Analysis. *International Journal on Document Analysis and Recognition*, 7(1), 2004.

[7] C. Viard-Gaudin, P. M. Lallican, P. Binter, and S. Knerr. The IRESTE On/Off (IRONOFF) Dual Handwriting Database. In *Proc. Intl. Conf. Document Analysis and Recognition, ICDAR'99*, pages 455–458, Bangalore, India, 1999.

[8] W3C Multi-modal Interaction Working Group. Ink Markup Language (InkML). *http://www.w3.org/2002/mmi/ink*, 2003.