An XML Representation for Annotated Handwriting Datasets for Online Handwriting Recognition

Ajay S Bhaskarabhatla and Sriganesh Madhvanath

Hewlett-Packard Labs
Bangalore, India
{ajay.b, srig}@hp.com

Abstract

In this paper, we briefly descibe an XML representation for annotation of online handwriting data to support the development and evaluation of handwriting recognition algorithms, that is based on the emerging Digital Ink Markup Language (InkML) draft standard from W3C. In particular, we describe how the XML representation we have defined attempts to address issues of (i) support for different scripts, (ii) partial automation of labeling using recognition engines, (iii) planned as well as casual capture of handwriting data and (iv) semantic annotation of handwriting data at various levels such as character, word and phrase. The representation keeps the raw handwriting data (described by InkML) separate from its semantic interpretations. We also compare and contrast the XML representation with the extant UNIPEN representation for annotation of handwriting data.

1. Introduction

Online Handwriting Recognition refers to the interpretation of handwriting captured as "digital ink" from a suitable pen input device. Digital ink consists of a series of pen positions and optional attributes such as pen tilt and pressure sampled at (typically) uniform intervals of time. Annotated datasets of handwriting covering a variety of writing styles are essential for the development and evaluation of modern data-driven handwriting recognition engines.

Early standards for digital ink such as ITU-T 150 (1988) and Jot (1992) focused on the representation of digital ink and did not address the issue of annotation of handwriting data for handwriting recognition research and development. The lack of a common annotation standard resulted in duplication of data collection efforts for each research effort, and made systematic evaluation and comparison of different recognition algorithms difficult. This issue was first addressed by the UNIPEN consortium (The UNIPEN Consortium, 1994). The UNIPEN representation employed ASCII flat files to store handwriting data and associated annotation. The focus of the UNIPEN effort was on the recognition of cursive English, and the members of the consortium collected and annotated large amounts of handwriting data in the UNIPEN format. However, there have been attempts at creating datasets using the same standard in other languages such as Japanese (Kanji) and Arabic. UNIPEN also made available a set of tools to create annotated datasets from captured handwriting data (Guyon et al., 1994).

1.1. Need for a new standard

While research in online handwriting recognition in the context of Roman and many Oriental scripts has continued unbroken for over three decades and resulted in several commercial engines, the same cannot be said for the majority of the world's scripts especially in developing countries. The lack of significant and easily available linguistic resources in the form of annotated datasets of handwriting has been one of the obstacles to research in these scripts. It is clear that many of these resources need to be created, and the creation of such handwriting databases in different scripts calls for a standard representation that is independent of script and allows semantic interpretation of the writing at various user-defined logical levels (e.g. Word, Character). The representation should capture information about script, writing style, quality of writing and truth. It should also capture information about writers and the data capture environment. It should support automatic generation of annotation using recognizers, and subsequent manual validation processes. It should keep handwriting data separate from its semantic interpretations and it should support planned as well as casual data collection.

In this paper, we describe an XML representation for the annotation of handwriting data which addresses these requirements. XML is a natural choice for the representation of annotation because of its naturally hierarchical nature. The representation makes use of an underlying XML representation of the raw handwriting data called Digital Ink Markup Language, a standard being developed by the W3C for the description of digital ink (W3C, 2003).

1.2. Digital Ink Markup Language

Digital Ink Markup Language or InkML for short, has been mooted to provide a platform independent data format for representing ink entered with an electronic pen or stylus. The markup is designed to support the input, storage and processing of handwriting, gestures, sketches, music and other notational languages in Web-based applications. It 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. InkML provides means for application-specific extensions. By virtue of being an XML-based language, users may easily add application-specific information to ink files to suit the needs of the application at hand.

The current InkML specification defines a set of core primitive elements useful for any ink application. The *trace* is the basic element used to record the trajectory of the pen as the user writes digital ink. Details of the input device

(digitizer), digitizer channels, and coordinate system comprise the context in which ink is recorded and are captured in Context element. The *brush* element captures certain attributes of the pen during ink capture. The *traceRef* and *traceRefGroup* elements provide the basis for semantic labelling of groups of traces. One of the attributes of *traceRefGroup* is *contentCategory*, which may be used to describe at a rudimentary level the category of content that the traces represent; e.g., "Text/English", "Drawing", "Math", "Music". These elements are considered as core elements since they are useful for most pen-based applications.

The XML representation described in this paper may be thought of an application-specific extension of the "core" InkML for the creation of annotated datasets of handwriting data.

2. Representation for Annotated Handwriting Datasets

Our representation for annotated datasets of handwriting is called *hwDataset*, and it includes several elements for detailed annotation of handwriting, some of which are derived from the traceRefGroup element of the core InkML. The hwDataset element is the root of the XML document and captures meta-data about the dataset under *datasetInfo*, various definitions as part of *datasetDefs*, and hierarchical annotation of handwriting data under *hwData* (Figure 1). These elements are described briefly in the following subsections.

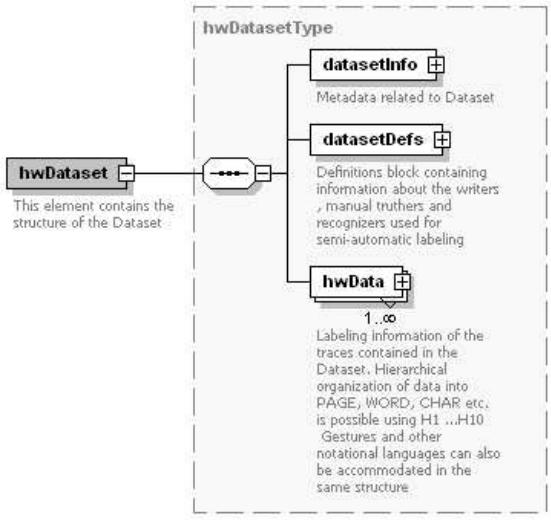

Figure 1: The document root element (hwDataset) and its sub-elements

2.0.1. datasetInfo

The *datasetInfo* element (Figure 2) captures metadata related to the dataset as a whole. It contains the following elements:
- *name* - name for referring to the dataset
- *category* - type of dataset
- *version* - version number and/or datestamp of publication
- *contact* - contact info for dataset-related queries
- *source* - source of collected data
- *setup* - physical conditions in which data was collected
- *dataInfo* - information about the data

The DataInfo element in turn contains the following subelements:
- *script* - language/script captured in dataset
- *quality* - quality of handwriting data captured in dataset
- *truth* - truth of what is captured
- *methodology* - design of data and collection procedure
- *annotationScheme* - description of annotation scheme

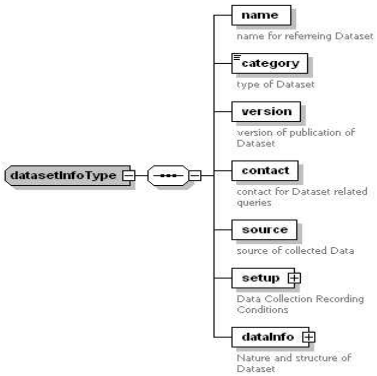

Figure 2: Element capturing metadata about the dataset

2.0.2. datasetDefs

The datasetDefs element (Figure 3) 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 as a sequence of labelSrc elements

The *writer* element contains the following elements:
date - date when writing occurred (meant to be a coarse description as opposed to the trace timestamps in the core InkML)
personal
- *hand* - left/right handedness
- *gender* - gender
- *age* - age at the time of capture
- *skill* - level of skill with script
- *style* - predominant writing style
- *region* - native region

The *labelSrcDefs* element contains the following elements:
- *name* - name of the human/automated source of labels
- *source* - organization that this label source represents
- *time* - date and time of annotation
- *contact* - contact details of label source
- *labelTypes* is an attribute and it describes the categories (e.g. truth, quality, script, style, etc) and encodings (e.g. UNICODE) of labels produced by the label source.

The above provides a mechanism for representing the writing of different writers in the same dataset, as well as multiple sources and categories of annotation for the same handwriting data. An algorithm for script identification might be used as a source of script labels, while a human annotator may provide labels for truth as well as script, style and quality of writing. Of course, the representation can also accommodate multiple label sources for the same category of label information, e.g. a recognition engine for truth labels and a human annotator for their validation.

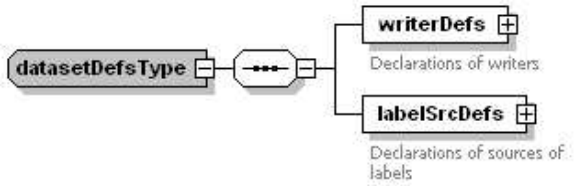

Figure 3: Element capturing dataset definitions

2.0.3. hwData

The *hwData* element allows hierarchical organization of data and annotation. It typically contains the root of the annotation hierarchy defined by the user, denoted by the element H1 (Figure 4). Each level of hierarchy H(i) contains a label element that captures annotation information at that level. H(i) also contains either one or more H(i+1) elements or *hwTraces*, the leaf elements of the hierarchy that refer to raw ink traces represented using InkML (Figure 5).

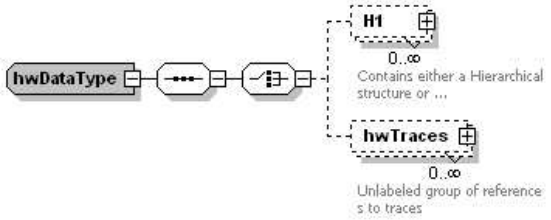

Figure 4: Element capturing annotation

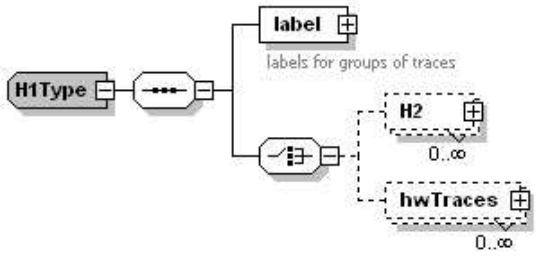

Figure 5: Nested hierarchy of annotation

The H(n) elements are meant to be used to indicate the structural makeup of handwriting, and assigned meaningful names such as PARAGRAPH and WORD using the corresponding attributes of the *hwData* element.

The *label* element (Figure 6) at each level can be used to capture alternative choices of label with confidence values if any, and the timestamp of annotation. 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 spanning different label sources of a particular unit of writing. The alternates can be used to facilitate the process of manual validation by prompting options for human validation. Formally, the attributes of *label* are
id - identification of label
labelSrcRef - reference to label source defined earlier. This holds good for sub-levels of the current level except where explicitly overridden
category - category of label (e.g. truth, quality, script, style, etc)
timestamp - time of the act of annotation

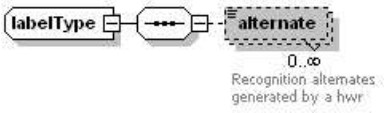

Figure 6: Label element with *alternates*

2.1. Use scenario

The hwDataset representation is meant to be used for annotation of handwriting from both planned data collection as well as casual capture from an ink application such as “ink chat”. Once the handwriting data has been captured as an InkML document and a suitable hierarchy for annotation defined (e.g., paragraph, word, character), a hwDataset document can be generated in which annotation is organized according to the hierarchy. In case multiple writers are involved, they can be defined in the definitions section and referred to in any H(n) element. Multiple sources of annotation can then be applied in stages. For example, one or more handwriting recognition engines may be used to generate truth labels (with multiple alternatives) at all of the different hierarchical levels, and the labels manually validated and corrected by a human. The hwDataset document will refer to the raw ink trace elements in the InkML document only in the hwTraces element at the bottom of the hierarchy. Thus annotation is separated from the raw ink, and this arrangement allows multiple hwDataset documents containing annotation to refer to the same ink data. The representation also supports multiple hwData blocks of annotation distinguished by their *trialId* attribute - a feature designed to support planned data collection.

2.2. Comparison with UNIPEN

The hwDataset representation is to a large extent inspired by the UNIPEN standard. However, there are some important differences between the two. hwDataset is an XML representation (currently instantiated as a schema) unlike UNIPEN which uses a custom text format. Unlike UNIPEN, hwDataset contains only annotation and does not

contain any information about the raw ink data or the digitizer used. These are left to the core InkML schema. This allows the separation of ink and annotation which are combined in the same document in UNIPEN. hwDataset does not include support for evaluation of recognition engines, although it does support their use in the annotation process. As a consequence, UNIPEN elements relating to alphabet used, inline lexicons, characterization of dataset as being training, test or adaptation set and recognition results have been dropped from hwDataset. Also, the data design and wordlists used for planned data collection are meant to be described in separate documents and only referred to in the hwDataset document.

3. Conclusions and Future Directions

In this paper, an XML representation for hierarchical annotation of online handwriting to support handwriting recognition is described. The representation also supports annotation of other aspects of handwriting such as writing style, quality and script, and accommodates multiple writers and annotation sources. The representation builds upon Digital Ink Markup Language (InkML), a draft specification of digital ink being developed by W3C.

The representation is in a preliminary stage, and we invite comments and suggestions from those engaged in handwriting recognition research and the creation of linguistic resources. We are also currently engaged in prototyping annotation tools based on the representation in the hope that their use for real-world data collection will provide valuable feedback for improving the representation further, especially when used for a variety of scripts.

4. References

Guyon, Isabelle, Lambert Schomaker, Rejean Plamondon, Mark Liberman, and Stan Janet, 1994. Unipen project of online data exchange and recognizer benchmarks. In *Proc. 12th Int'l. Conf. on Pattern Recognition, ICPR'94*. Jerusalem, Israel.

The UNIPEN Consortium, 1994. *UNIPEN 1.0 Format Definition*. http://www.unipen.org.

W3C, Multimodal Interaction Working Group, 2003. *Ink Markup Language (InkML)*. http://www.w3.org/2002/mmi/ink.

5. hwDataset Document

This is an example of typical hwDataset file

```
<hwDataset>
<datasetInfo datasetId="ID018484">
  <name>English Word Dataset</name>
  <category>2</category>
  <version>Ver1.0 Published Oct20,2003</version>
  <contact>Ajay <Ajay.B@hp.com></contact>
  <source>Indian Institute of Science</source>
  <setup>Writing on a iPAQ PocketPC placed on a horizontal surface while seated</setup>
  <dataInfo>
    <script>English/Roman</script>
    <style>mixed</style>
    <truth>http://hpl.hp.com/truth.htm</truth>
    <methodology>100 most frequent English words selected from CIIL text corpus written 5 times each writer</methodology>
    <annotationScheme>http://hpl.hp.com/scheme.htm
</annotationScheme>
  </dataInfo>
</datasetInfo>
<datasetDefs>
  <writerDefs>
  <writer id="ID018485">
    <date>19800813</date>
    <personal>
      <hand>left</hand>
      <gender>male</gender>
      <age>23</age>
      <skill>poor</skill>
      <style>print</style>
      <region>india/bihar/patna</region>
    </personal>
  </writer>
  </writerDefs>
  <labelSrcDefs>
    <labelSource id="ID018486" type="machine">
    <name>English Recognizer v1.0</name>
    <source>HP Labs India</source>
    <time>20031020 12:10:23</time>
    <contact>deepu@hp.com</contact>
    <labelTypes>
      <labelType encoding="UNICODE">truth
</labelType>
      <labelType>quality</labelType>
    </labelTypes>
    </labelSource>
  </labelSrcDefs>
</datasetDefs>
  <hwData trialId="0" H1="PHRASE" H2="WORD">
    <H1 id="ID000005" writerRef="writer reference">
    <label id="ID000006" labelSrcRef="label source reference" category=quality>
      <alternate>OK</alternate>
    </label>
    <label id="ID000007" labelSrcRef="label source reference" category=truth>
      <alternate number=0>Hello Ink </alternate>
      <alternate number=1>Hallo Ink </alternate>
    </label>
    <H2 id="ID000008">
    <label id="ID000007" category=truth>
      <alternate number=0 score=0.90> Hello
</alternate>
      <alternate number=1 score=0.80> Hallo
</alternate>
    </label>
      <hwTraces>
        <traceref xpath="reference to trace" />
    ...
      </hwTraces>
    </H2>
  </hwData>
</hwDataset>
```