The hOCR Embedded OCR Workflow and Output Format

Thomas Breuel (editor)

- December 2007 - initial release
- March 2010 - bug fixes, clarifications

1 Rationale

The purpose of this document is to define an open standard for representing OCR results. The goal is to reuse as much existing technology as possible, and to arrive at a representation that makes it easy to reuse OCR results.

2 Getting Started

This document describes many tags and a lot of information that can be output. However, getting started with hOCR is easy: you only need to output the tags and information you actually want to. For example, just outputting **ocr_line** tags with bounding boxes is already very useful for many applications. Just start simple and add more output information as the need arises.

3 Terminology and Representation

This document describes a representation of various aspects of OCR output in an XML-like format. That is, we define as set of tags containing text and other tags, together with attributes of those tags. However, since the content we are representing is formatted text, However, we are not actually using a new XML for the representation; instead embed the representation in XHTML (or HTML) because XHTML and XHTML processing already define many aspects of OCR output representation that would otherwise need additional, separate and ad-hoc definitions. These aspects include:

- standard representations for common logical structuring elements, including section headings, citations, tables, emphasis, line breaks, quotations, citations, and preformatted text
- standard representations for fonts, embedded images, embedded vector graphics, tables, languages, writing direction, colors
- standard representations for geometric layout and positioning
- output files that are understood without any further modification by widely used