Please find attached a zip file containing the format definitions and corresponding examples for better understanding. There are some additions to the old format but many things are still the same (i.e. ground-truth in the old format can be transformed to the new one but not necessarily vice versa). The most prominent change relates to the organisation of ground-truth data. There is now a root structure which links to ground-truth subsets and tracks any dependencies between them. The rationale for this is that we have to deal with many different types of ground-truth (according to the processing steps) which do not necessarily fit into the same structure. Moreover, the new format allows the coexistence of potentially different ground-truth views for the same task (e.g. different ground-truthing guidelines for segmentation). Last but not least, we are now able to reflect different paths for ground-truth production by defining dependencies (e.g. segmentation ground-truth from the deskewed image vs. segmentation ground-truth for the original). Also the page-content format has been extended. Among others it is now possible to express lines, words, glyphs, and the actual text. Furthermore, it is possible to define groups and reading order (including elements in unordered groups).

Perhaps you want to check the examples first. The gt-... file contains the aforementioned root structure and links to ground-truth subsets for binarisation (two views for binarisation - whole page, only binarised patches [this format was necessary to reduce the ground-truthing effort]), deskew, dewarping and page content (successor of the old format). Each format is defined in a separate XML Schema file and has its own namespace (e.g. <http://schema.primaresearch.org/gt/gts/deskew/2008-11-01>) which is also used to identify the used version. Linked images which are part of the ground-truth (e.g. the perfectly binarised page) are not included in the zip file.

Used file naming scheme:

Schema prefix (gt, ds, ...) - document ID (PRImA database)- timestamp - random number . suffix e.g.:

gt-00006664-20081107T11432018-0815.xml

ds-00006664-20081110T16252015-9876.xml

corresponding files e.g. the deskewed image should have the same file name apart from the suffix e.g.:

ds-00006664-20081110T16252015-9876.tif

The XML Schemas (including future versions) are maintained under this URL:

<http://schema.primaresearch.org/gt/>

The files contain annotations and at some point there will be a complete documentation. It is also planned to provide a DLL offering read and write functionality for other developers. But this might still take some time to be finalised and therefore we have to live with the XML Schemas and standard XML functions of the used programming languages for now.