汉文古籍集外字描述规范

汉文古籍（以下简称古籍）在文字、版式、内容等方面与现代文献和民国文献有显著的区别，数字化的难度也最高。现有研究成果多是对数字图书馆，涉及全部文献类型，其中只有部分成果适用于古籍。现行的字符集标准《信息技术 中文编码字符集》（GB 18030-2005）收录了70244个汉字，unicode7.0收录CJK文字74566个，而古籍使用的文字量远大于现代文献和民国文献，还有大量的异体字、俗体字、避讳字等，均无法满足古籍数字化的用字需求。同时还存在大量字符集中未包含的汉字，即集外字。集外字描述是集外字处理方法之一,也是古籍文本化的重点和难点，直接影响古籍文本数据的质量和可用性。

本标准的规范内容包含对古籍文献特性、古籍集外字特性分析、给出集外字拆分规则和表意描述序列（Ideographic Description Sequence，IDS）应用等。