TECHNICAL SPECIFICATION

SPÉCIFICATION TECHNIQUE

TECHNISCHE SPEZIFIKATION

CEN/TS 15873

March 2009

ICS 03.240; 35.240.60

English Version

Postal Services - Open Standard Interface - Address Data File Format for OCR/VCS Dictionary Generation

Services postaux - Interface de standard ouvert - Format de fichiers de données d'adresses pour la génération du dictionnaire OCR/VCS

Postalische Dienstleistungen - Offene Normschnittstelle - Adressdateiformat für die Generierung von Wörterbüchern in OCR/Videocodier-Systemen

This Technical Specification (CEN/TS) was approved by CEN on 1 March 2009 for provisional application.

The period of validity of this CEN/TS is limited initially to three years. After two years the members of CEN will be requested to submit their comments, particularly on the question whether the CEN/TS can be converted into a European Standard.

CEN members are required to announce the existence of this CEN/TS in the same way as for an EN and to make the CEN/TS available promptly at national level in an appropriate form. It is permissible to keep conflicting national standards in force (in parallel to the CEN/TS) until the final decision about the possible conversion of the CEN/TS into an EN is reached.

CEN members are the national standards bodies of Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland and United Kingdom.

EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

Management Centre: Avenue Marnix 17, B-1000 Brussels

Ref. No. CEN/TS 15873:2009: E

Contents

Page

Foreword

This document (CEN/TS 15873:2009) has been prepared by Technical Committee CEN/TC 331 "Postal Services", the secretariat of which is held by NEN.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. CEN [and/or CENELEC] shall not be held responsible for identifying any or all such patent rights.

According to the CEN/CENELEC Internal Regulations, the national standards organizations of the following countries are bound to announce this Technical Specification: Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland and the United Kingdom.

NOTE This document has been prepared by experts from CEN/TC 331 and UPU, in the framework of the Memorandum of Understanding between UPU and CEN.

1 Introduction

In initial meetings of CEN/TC331/WG3 interfaces which will benefit from standardization have been identified and agreed on. Candidates for Open Interface standardization are:

- interface between the image handler and automatic address readers or video coding places;
- interface from machine control to Barcode Printers;
- interface from machine control to Barcode Reader / Verifier;
- interface between scanner, image handler and machine control;
- file format of Sort Plan;
- MIS Interface (Statistics);
- file format of Address data files.

The new intended standard deals with the file format of Address Data Files.

OCR results and video coder inputs have to be verified against the "real" existing addresses in order to reach high recognition rates combined with low error rates. For that purpose postal operators provide postal address directories to the OCR/VCS suppliers. Usually different postal operators use different file formats for these (source) directories. In typical postal automation systems these files will be processed by directory generation software which creates application specific loadable data. This data – usually referred to as "operational directory" – is heavily compressed and contains access tables tailored for the specific reading software. Usually different OCR/VCS suppliers use different operational directory formats.

This standard shall define a common Address Data File format for postal address directories to be provided from the postal operators to the OCR/VCS suppliers.

This Address Data File format shall be designed to hold all information necessary to support address reading and video coding software including data required for special recognition tasks e.g. forwarding applications.

2 Scope and purpose

2.1 Scope

This document defines a file format for the generation of postal address directories. It is designed to hold all information necessary to support address reading software including data required for forwarding applications. In typical postal automation systems these files will be processed by directory generation software which creates application specific loadable data. This data – usually referred to as operational directory – is heavily compressed and contains access tables tailored for the specific reading software.

Not in the scope of this document are topics external to file like compression, checksums, the interface for transmission to the supplier, modification permissions, error handling on inconsistent data and undo in updates.

2.2 Purpose

The format has been designed with the following requirements in mind:

- must be able to hold the following data:
  - addresses composed of address components (including aliases and range-data);
  - person and organization names;
  - address codes typically used as sort codes;
  - links between addresses e.g. for use in forwarding;
- should not restrict character encoding;
- easily customizable for specific applications;
- should allow complete as well as incremental updates, i.e. change-only data;
- it must be possible to split data in multiple files for better handling.

The ideas behind this format are as follows:

- The format is based on XML.
- The basic XML structure is general. Project (the term project is used throughout this document to describe a specific application such as address data for a specific country or postal organization) specifics are coded as attributes. This should make it easier to build project independent parsers and tools.
- Address data can be structured hierarchically. An address component appearing in a lot of addresses shall be written once as parent node in all addresses it is used in the XML address tree.
- Beyond the pure address data, there are general as well as optional project specific attributes on the level of address components and string parts.
- In favour of faster parser execution and smaller file sizes the names of XML elements appearing very often are short strings.