DCMTK - Feature #1073

Character set conversion should warn or reject in certain cases

2023-04-14 10:06 - Marco Eichelberg

Status:	New		**Start date:**	2023-04-14
Priority:	Normal		**Due date:**	
Assignee:			**% Done:**	0%
Category:	Library and Apps		**Estimated time:**	0:00 hour
Target version:				
Module:	dcmdata		**Compiler:**	
Operating System:				

Description

There are certain cases in which a character set conversion in DCMTK (e.g. using the dcmconv tool) will not convert all strings to the new character set:

- the dataset is in implicit VR and an attribute containing an extended character set string is not in the data dictionary
- the dataset is in implicit VR and the attribute containing an extended character set string is contained in a sequence with defined length that is not in the data dictionary
- the dataset is in explicit VR and an attribute containing an extended character set string is coded as UN
- the dataset is in explicit VR and the attribute containing an extended character set string is contained in a sequence with defined length that is encoded as UN

In all cases the result is a dataset where some elements use the "new" character set and some elements remain encoded with the "old" character set.
When another application that has an updated data dictionary or is configured to convert UN to the proper VR reads and processes the file, it will encounter inconsistent character set encoding.

It would be desirable that

- we introduce a configuration flag that would cause the conversion to be rejected in these cases, with an error code, and the dataset remaining unmodified
- that in the case of current behaviour (no error) a warning is printed to the logger in these cases
- that dcmconv gets new command line options that allow the user to select between these two policies

Issue reported 2023-04-13 by Mathieu Maleterre <mathieu.malaterre@gmail.com>.

Related issues:

Blocked by DCMTK - Feature #870: Introduce separate class for VR=UN	**New**	**2019-02-08**

History

#1 - 2023-04-14 10:51 - Marco Eichelberg

One more case missing in the list above:

- the dataset is in explicit VR and an attribute containing an extended character set string is coded in a newly introduced VR unknown to the toolkit

#2 - 2023-04-14 16:53 - Marco Eichelberg

- Related to Feature #870: Introduce separate class for VR=UN added

#3 - 2023-04-14 16:54 - Marco Eichelberg

- Related to deleted (Feature #870: Introduce separate class for VR=UN)

#4 - 2023-04-14 16:54 - Marco Eichelberg

- Blocked by Feature #870: Introduce separate class for VR=UN added