Recensement de corpus de fichiers utilisables							en date du 15/12/2020
Code	**Nom**	**Organisation**	**URL**	**Type**	**Volumétrie**	**Licence**	**Repérage des fichiers valides ou non**
CF_01	Format corpus	OPF	https://github.com/openpreserve/format-corpus	Tous		CC0	
CF_02	veraPDF-Corpus	OPF, DualLabs	https://github.com/veraPDF/veraPDF-corpus	PDF		CC-BY 4.0	Oui
CF_03	Isartor	PDF Association	https://www.pdfa.org/resource/isartor-test-suite/	PDF/A		Freely downloadable and usable without restriction	Oui
CF_04	Bavaria	PDFLib	https://github.com/bfosupport/pdfa-testsuite	PDF/A		Creative Commons Public License	
CF_05	Jpylyzer test Suite	OPF	https://github.com/openpreserve/jpylyzer-test-files	JP2000		CC-BY	Oui
CF_06	Google image test suite		https://code.google.com/archive/p/imagetestsuite/	Images : TIFF, PNG, GIF, JPEG	143 Mo	CC-BY 3.0	

CF_07	PNGSuite		http://www.schaik.com/pngsuite/	PNG		Permission to use, copy, modify and distribute these images for any purpose and without fee is hereby granted.	Oui
CF_08	govdocs1		https://digitalcorpora.org/corpora/files	Tous	1 million de fichiers	For any use. They all came from US Government websites. Distribution is unlimited.	Non
CF_09	XML Conformance Test Suites	W3C	https://www.w3.org/XML/Test/	XML	zip d'un à 2 Mo. De l'ordre de plusieurs milliers de fichiers test	W3C Software License	Oui
CF_10	1000 .gov PDF Dataset Download	LOC	https://s3.us-east-2.amazonaws.com/lclabspublicdata/lcwa_gov_pdf_README.txt	PDF	1000 fichiers; 673.5 MB zip file (BagIT)	https://www.loc.gov/collections/legislative-branch-web-archive/about-this-collection/rights-and-access/	Fichier CSV avec des métadonnées techniques (extraites par Tika)
CF_11	1000 .gov Audio Dataset Download	LOC	https://s3.us-east-2.amazonaws.com/lclabspublicdata/lcwa_gov_audio_README.txt	Audio	1000 fichiers; 4.6 GB zip file (BagIT)	https://www.loc.gov/collections/legislative-branch-web-archive/about-this-collection/rights-and-access/	Fichier CSV avec des métadonnées techniques (extraites par Tika)
CF_12	CR du Grand Débat		https://granddebat.fr/pag	Tous	??	Licence ouverte (c	Non