

# Garantie d'accès à long terme : La validation du PDF par JHOVE ?

---

Yvonne Friese (ZBW - centre d'Information Leibniz pour l'Economie), vient de publier un article<sup>1</sup> qui remet en cause la fiabilité de la validation des fichiers au format PDF par JHOVE.

Dans sa publication, l'auteure expose des limites de JHOVE<sup>2</sup> lors de la validation de fichiers PDF. La plus récente des failles de ce logiciel libre, a été révélée par Olaf Drummer, Chairman du PDF Association, au hackaton du PDF organisé par l'OPF (Open Preservation Foundation) à Hambourg les 1 et 2 septembre 2014. Concrètement, l'arborescence des pages d'un fichier PDF a pour but de faciliter la navigation dans ce fichier. Les pages peuvent être disposées sous forme d'un tableau ou sous forme d'un arbre (avec des nœuds et des feuilles). Si cette dernière est la plus souvent utilisée et recommandée par la norme, elle n'est pas obligatoire, mais JHOVE rejette la première et retourne une erreur. Dans ce cas, le message d'erreur est à ignorer, étant donné que ce n'est pas un problème et que ça ne présente aucun risque pour la pérennité du document.

Ce n'est qu'un exemple parmi tant d'autres. En effet, des tests réalisés au CINES<sup>3</sup> ont permis de révéler la faiblesse de JHOVE dans la validation de fichiers PDF/A destinés à l'archivage.. Yvonne Friese présente quelques points sur lesquels JHOVE trébuche. S'il est vrai que les PDF/A-1 et PDF/A-2 sont basés sur les versions de PDF 1.4 et 1.7, il n'en demeure pas moins qu'il existe quelques différences entre les versions normées et les versions standards. Un fichier PDF/A-1b interdit les éléments suivants :

- L'embarquement d'un fichier média (car nécessite un lecteur externe)
- La compression LZW
- Le cryptage
- La transparence
- Javascript, les fichiers exécutables, des commandes et des liens externes

Par ailleurs, un fichier PDF/A exige :

- Les données embarquées complètes. Ceci inclut les polices, les images et les métadonnées XMP
- La reproduction sans ambiguïté des polices et des couleurs

---

<sup>1</sup> <http://www.pdfa.org/2014/12/ensuring-long-term-access-pdf-validation-with-jhove/>

<sup>2</sup> URL:<http://jhove.sourceforge.net/>

<sup>3</sup> Centre Informatique National de l'Enseignement Supérieur (Montpellier, France) - [www.cines.fr](http://www.cines.fr)

En mentionnant les résultats de JHOVE lors de la validation de fichiers issus de la suite Isartor, YF confirme ceux que les experts formats du CINES ont obtenus au cours de tests récents réalisés dans le cadre d'une étude sur le PDF. En effet, sur plus de 200 fichiers contenant des erreurs, JHOVE n'a pu identifier qu'un seul comme invalide. Cette statistique n'est pas du tout rassurante.

Les résultats catastrophiques de JHOVE dans la validation du PDF/A sont dus au fait que son analyse concerne la structure générale d'un fichier PDF et non les spécificités du PDF/A. Ils suffisent pour remettre en cause, peut-être pas l'ensemble de ses modules<sup>4</sup>, mais celui qui contrôle les fichiers PDF. Par conséquent, l'avenir de JHOVE est mitigé. Pourtant c'est l'un des rares validateurs multi-formats open source que de nombreuses institutions et établissements de mémoire ont utilisé jusqu'à présent. Cette situation interpelle fortement la communauté archivistique numérique. L'étude PDF partie 3 du CINES, qui sera bientôt publiée, sera sans doute un bon guide pour cette problématique.

---

<sup>4</sup> AIFF, GIF, JPEG, JPEG 2000, PDF, TIFF, WAVE