Hypertexte : sommaire

Le terme hypertexte désigne un ensemble de textes reliés entre eux par des liens *(links)*. La présentation des textes et celle des liens s'effectuent avec un langage de balisage *(markup language)*. Le langage de balisage hypertexte le plus connu est le HTML. Le Web constitue donc le plus grand hypertexte. Celui-ci se subdivise en sous-domaines thématiques appelés sites web ou sites internet et qui se caractérisent par un nom de domaine commun, jusqu'à en arriver à une page internet qui à proprement parler est également un hypertexte, parce qu'elle est constituée d'une page HTML et d'un ensemble de ressources web référencées.
L'archivage est confronté à trois problèmes:
Premièrement, un hypertexte est structuré sous la forme d'un réseau. Si nous essayons de représenter les différentes pages web de manière linéaire ou hiérarchique, comme nous en avons l'habitude, nous perdons un aspect sémantique important. Nous devons donc archiver les pages et les liens ensemble. Le réseau de liens n'existe nulle part de manière externe, comme dans un système de classement, mais résulte au contraire implicitement de tous les liens dans toutes les pages et il est établi par exploration des pages web au moyen d'un robot d'indexation *(crawler)*.
Deuxièmement, seule l'interaction d'une page HTML et d'un ensemble de ressources web dans le navigateur constitue une page web. Celle-ci n'est nulle part entièrement disponible en tant que fichier source. C'est pourquoi l'archivage de pages web ou de sites internet entiers nécessite l'utilisation d'un logiciel qui simule le point de vue du navigateur ou qui recueille en tant que fichiers toutes les ressources nécessaires à la représentation dans le navigateur.
Troisièmement: Si nous suivons la différenciation entre objet et information de représentation dans l'OAIS, il est difficile de dire où trouver l'objet parce que les systèmes de gestion de contenus *(content management systems CMS)* actuels ne sauvegardent plus nulle part de véritables pages HTML, mais les assemblent depuis un ensemble de données seulement en cas de demande. L'aspect information de représentation est tout aussi problématique. Nous avons en effet dans le CMS un premier niveau dans lequel une page est présentée à partir d'objets d'information. Dans une deuxième étape, le serveur web complète cette page au moment même où celle-ci est générée. Dans la troisième étape de la représentation, le navigateur charge des ressources supplémentaires du serveur web et finalement il exécute du JavaScript intégré qui peut encore une fois charger des ressources et restituer la page dans sa forme d'origine et ensuite l'afficher.
L'absence de hiérarchie représente également un grand problème pour l'évaluation. On ne peut en effet évaluer les pages web que de manière individuelle ou les sites internet de manière intégrale.

Formats examinés

- HTML
- HTML5
- MHTML
- ARC
- WARC
- PDF-A-2 pour hypertexte

Recommandation

Bien que WARC soit le format instauré pour l'archivage web, il ne peut être recommandé comme format d'archivage à long terme en raison de la diversité potentielle de formats intégrés et de la difficulté de les migrer. Pour archiver un site internet entier, il est clairement recommandé à long terme de convertir en PDF/A . On peut convertir des pages isolées en PDF, en réalisant les liens au travers de la structure du classeur ou sauvegarder un site web entier en tant que fichier PDF/A-2 et les liens hypertextes renvoient alors de page PDF en page PDF.
Le HTML ou le HTML5 ne peuvent en fait être recommandés en tant que format d'archivage que pour des pages sans ressources externes intégrées, à la place du texte brut (plain text), avec la possibilité supplémentaire de fixer également la structure du texte et la mise en page. Dans le catalogue, les formats HTML figurent donc plutôt dans les formats textuels structurés.

Étude

Étude sur l'archivage web (en allemand)

Bibliographie

Manuel du groupe de travail Nestor: chapitre 17.9, Web-Archivierung zur Langzeiterhaltung von Internet-Dokumenten
http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_293.pdf

Contact
A propos
Impressum
Événements
Newsletter
RSS

ARC

Informations générales

Titre	ARC_IA, *Internet Archive ARC file format*
Catégorie	Hypertexte
Abréviation	ARC
Extension de fichier	.arc
Mime Type	application/x-internet-archive
Pronom PUID	x-fmt/219, fmt/410
Version	La désignation ARC a été utilisée dans les débuts de l'informatique pour différents formats d'archives de fichiers qui sont tous des précurseurs de TAR, PKARC et ZIP. ARC_IA désigne une variante spéciale qui a été utilisée par l'Internet Archive.

Description

ARC est un format de fichiers des années quatre-vingt pour sauvegarde compressée de différents fichiers dans un fichier. ARC ne pouvait à l'origine pas représenter les fichiers dans leur arborescence. L'*Internet Archive* a développé le format afin de pouvoir sauvegarder efficacement plusieurs ressources d'une page web.
Un fichier ARC contient la réponse HTTP complète *(response)* et le paquet de données *(payload)* transmis de toutes les pages explorées par le robot d'indexation (crawler) ainsi qu'un set de métadonnées pour le processus de crawling. Chaque bloc *(HTTP response*) est compressé de façon indépendante. Le fichier ARC résoud surtout le problème de la sauvegarde d'innombrables petits fichiers dont sont composés les contenus web. L'accès s'effectue au mieux par une base de données externe. Le fichier ARC ne possède pas sa propre rubrique d'indexation.
Ni HTTP *response* ni *payload* ne sont normalisés d'une quelconque manière dans le fichier ARC. Leur forme correspond exactement à ce qui a été envoyé du serveur web.

Evaluation

Ouverture du format : **3**
La spécification de ARC_IA est administrée par l'*Internet Archive*.

Licence libre : **3**
Il n'existe pas de patente connue pour ARC_IA.

Diffusion : **2**
WARC a aujourd'hui pris la relève d'ARC_IA. Cependant, de grandes parties de l'Internet Archives sont vraisemblablement encore basées sur des fichiers ARC.

Fonctionnalités : **2**
L'usage du format est fortement limité par l'absence de répertoire de fichiers.

Implémentation : **3**
L'*Internet Archive Wayback Machine* (« machine à remonter le temps ») peut travailler avec ARC_IA. Différentes solutions de moissonnage *(harvesting)* peuvent sauvegarder dans ce format, par exemple le produit open source Heritrix .

Densité de mémorisation : **2**
La densité de mémorisation est relativement élevée du fait de l'utilisation d'une compression.

Vérifiabilité : **3**
L'*Internet Archive Wayback Machine* peut être mise à contribution pour la vérification de format.

Bonnes pratiques : **1**
WARC a pris la relève du format et il ne peut donc plus être recommandé.

Perspectives : **1**
Aucune perspective n'est en vue.

Classe de formats : **X**
Il s'agit d'un format obsolète.

Conclusion

En cas de nouvelle exploration par robot d'indexation ou de moissonnage, il faut choisir WARC comme format d'archivage. Il ne faut pas absolument convertir des fonds d'archives existants d'ARC en WARC parce que cette opération implique de très gros efforts en raison de leur taille. La conversion ne changerait cependant rien aux pages HTML sous-jacentes ni aux ressources intégrées.

Références

Internet Archive: Mike Burner, Brewster Kahle « Arc File Format » September 15, 1996, Version 1.0
https://archive.org/web/researcher/ArcFileFormat.php
Library of Congress: Sustainability of Digital Formats - ARC_IA, Internet Archive ARC file format
http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml

Bibliographie

Voir WARC .

Articles connexes

WARC

HTML

Informations générales

Titre	Hypertext Markup Language; Langage de balisage hypertexte
Catégorie	Hypertexte
Abréviation	HTML
Extension de fichier	.html et .htm
Mime Type	text/html et application/xhtml+xml
Pronom PUID	fmt/96, fmt/97, fmt/98, fmt/99, fmt/100, fmt/102, fmt/103
Version	HTML 2.0 est la première version formellement spécifiée par le World Wide Web Consortium (W3C) en 1995. Les versions principales suivantes sont la 3.2 et la 4.0. En 2002 elles sont enrichies par XHTML et après une longue interruption dans le développement, elles sont remplacées fin 2014 par HTML5 .

Description

HTML est un langage de balisage basé sur du texte destiné à structurer des documents numériques. HTML utilise des balises plus ou moins explicites pour la structuration sémantique, le formatage exact du balisage revenant au logiciel d'affichage (navigateur) et à la mise en page (CSS). En plus de structurer, le HTML permet d'inclure des liens, des métadonnées et des images. Le HTML est une forme simplifiée du langage de balisage SGML développé spécialement pour le web.

Evaluation

Ouverture du format : **4**
La spécification HTML est administrée et développée par le *World Wide Web Consortium (W3C)*.

Licence libre : **3**
Il n'existe pas de patente connue pour le HTML.

Diffusion : **4**
Par sa diffusion sur le web, le format est l'un des plus répandus du monde informatique.

Fonctionnalités : **2**
Le HTML couvre complètement les fonctionnalités des données hypertextes. La séparation de la sémantique et de la forme a en revanche perdu de son importance avec le temps, bien qu'une réglementation claire existe avec la séparation HTML/CSS. Beaucoup de pages internet mettent l'accent sur la présentation graphique dans la fenêtre du navigateur.

Implémentation : **4**
Il existe plusieurs implémentations de navigateurs techniquement indépendantes les unes des autres pour HTML 2.0 jusqu'à 4.1.

Densité de mémorisation : **2**
À cause de l'utilisation de balises redondantes, la densité de mémorisation n'est pas très élevée.

Vérifiabilité : **3**
Dans les versions HTML plus récentes, on peut visualiser aussi bien la version que le codage des signes depuis l'en-tête du fichier (header) et les métadonnées. De ce point de vue, les versions plus anciennes sont peu fiables. Il existe divers validateurs HTML, mais en raison de la grande tolérance aux erreurs de la plupart des navigateurs, il existe également beaucoup de codes HTML invalides.

Bonnes pratiques : **1**
HTML n'est pas recommandé ni utilisé comme format d'archivage.

Perspectives : **1**
Seul HTML5 présente un potentiel en vue de développer un format adapté pour l'archivage.

Classe de formats : **A**
Il s'agit d'un des plus anciens formats du domaine de l'informatique

Conclusion

Du fait de la séparation de la structure des contenus et de la présentation formelle, le format répondrait en fait aux besoins de base de l'archivage numérique, puisqu'il rendrait l'archivage indépendant du développement du logiciel de visualisation. Malheureusement, le développement très rapide de la technologie des navigateurs a modifié entretemps le format et il s'avère que le contenu de beaucoup de pages internet n'est compréhensible que si la présentation graphique est correcte. Cela signifie que c'est dans une large mesure la rétrocompatibilité du navigateur qui garantit le fait que nous puissions regarder les anciennes pages internet dans leur apparence initiale et en comprendre le contenu.
La nécessité d'archiver séparément les ressources externes qui vont avec chaque page HTML a un effet très négatif et a mené aux formats conteneurs pour pages HTML tels que MHTML , ARC et WARC .

Références

Hypertext Markup Language - 2.0, September 22, 1995
https://www.w3.org/MarkUp/html-spec/html-spec_toc.html
HTML 3.2 Reference Specification, W3C Recommendation 14-Jan-1997
https://www.w3.org/TR/REC-html32
HTML 4.01 Specification, W3C Recommendation 24 December 1999
https://www.w3.org/TR/html4/
XHTMLâ„¢ 1.0 The Extensible HyperText Markup Language (Second Edition), A Reformulation of HTML 4 in XML 1.0, W3C Recommendation 26 January 2000, revised 1 August 2002
https://www.w3.org/TR/xhtml1/

Bibliographie

Ian S. Graham « The HTML SourceBook » New York, 1995
Rainer Klute « Das World Wide Web » Bonn Addison-Wesley, 1996
Erik Wilde « World Wide Web, Technische Grundlagen » Berlin Heidelberg, 1999

Articles connexes

HTML5
MHTML

Catalogue des formats de données d'archivage

version 6.0, juil. 2019

HTML5

Informations générales

Titre	*Hypertext Markup Language, Version 5*; Langage de balisage hypertexte, version 5
Catégorie	Hypertexte
Abréviation	HTML5
Extension de fichier	.htm, .html
Mime Type	text/html
Pronom PUID	fmt/471
Version	Après une longue interruption dans le développement, HTML5 a remplacé à la fin 2014 les anciennes versions HTML.

Description

HTML5 est un développement de HTML4 et apporte quelques avantages: meilleure sémantique, meilleur traitement des documents HTML mal conçus et surtout propre format audio et vidéo, de sorte qu'il n'y a plus besoin de module d'extension *(plug-in)*. La réduction des formats audiovisuels intégrés autorisés est un avantage certain pour l'archivage. Le nouveau stockage local *(local storage feature)* n'a en revanche pas d'importance directe pour l'archivage.

Evaluation

Ouverture du format : **4**
La spécification HTML5 est administrée et développée par le *World Wide Web Consortium (W3C)*.

Licence libre : **3**
Il n'existe pas de patente connue pour HTML5.

Diffusion : **2**
Le format est nouveau et il a été apparemment difficile de trouver un accord entre les milieux intéressés, à savoir les représentants des navigateurs et les prestataires de services internet.

Fonctionnalités : **3**
En plus de la fonctionnalité hypertexte, HTML5 possède les fonctionnalités web 2.0 grâce aux APIs de HTML5. Il convient cependant de relever que le web 2.0 est certes impensable sans JavaScript, mais après exécution des scripts correspondants, on obtient un véritable objet HTML5 (DOM) dans le navigateur.

Implémentation : **4**
Il existe plusieurs implémentations de navigateur techniquement indépendantes les unes des autres pour HTML5.

Densité de mémorisation : **2**
À cause de l'utilisation de balises redondantes, la densité de mémorisation est relativement faible.

Vérifiabilité : **3**
Dans HTML5, la version ainsi que le codage de caractères sont clairement reconnaissables. Il existe également déjà des validateurs. Cependant, il n'est pas très facile de faire une délimitation claire avec les anciennes versions HTML.

Bonnes pratiques : **1**
HTML5 ne peut pas encore être recommandé comme format d'archivage.

Perspectives : **2**
HTML5 présente un potentiel en vue de développer un format adapté pour l'archivage.

Classe de formats : **C**
Il s'agit d'une nouvelle définition avec un potentiel pour l'instant encore flou, mais pas d'un nouveau développement fondamental.

Conclusion

Un site web strictement normé d'après HTML5 contiendra lors de son archivage en plus des pages HTML5 d'autres ressources dans moins de formats différents que dans les versions précédentes de HTML . Plus précisément, la liste des formats autorisés devrait être exhaustive et connue. Comme le processus de normalisation justement dans ce domaine n'est pas encore terminé, il est impossible d'émettre un jugement définitif. Il est clair que le HTML5 ne remplace pas un format conteneur pour la sérialisation de pages individuelles ou de sites web entiers.
Le stockage local (local storage feature) souvent mentionné en relation avec HTML5 est une combinaison de cookies et de base de données côté navigateur, pas une sérialisation d'un objet DOM dans un fichier local et n'est de ce fait pas une alternative au MHTML ou formats analogues.
Il n'est pas défini s'il sera possible de convertir les outils des versions antérieures de HTML en HTML5, ni comment, car cela comprendrait également la possibilité de convertir des ressources incorporées. Il faut vraisemblablement envisager le passage en HTML5 par des adaptations dans les systèmes de gestion de contenus (CMS), plutôt que par de véritables migrations de fichiers HTML existants.

Références

HTML5, A vocabulary and associated APIs for HTML and XHTML, W3C Recommendation 28 October 2014
https://www.w3.org/TR/html5/

Bibliographie

Jürgen Wolf: HTML5 und CSS3; Das umfassende Handbuch, 25. Mai 2015, ISBN:978-3-8362-2885-5
Matthew MacDonald: HTML5; The Missing Manual, 2nd Edition, 3 January 2014, ISBN:978-1-4493-6326-0
Thomas A. Powell: HTML & CSS; The Complete Reference, Fifth Edition
www.dcpehvpm.org/E-Content/BCA/BCA-II/Web%20Technology/the-complete-reference-html-css-fifth-edition.pdf

Articles connexes

HTML
MHTML

MHTML

Informations générales

Titre	*MIME Encapsulation of Aggregate HTML Documents*; (Il n'y a pas de désignation française courante)
Catégorie	Hypertexte
Abréviation	MHTML
Extension de fichier	.mht, .mhtml
Mime Type	multipart/related
Pronom PUID	x-fmt/429
Version	Le format a été proposé par *l'Internet Engineering Task Force* dans le cadre de la RFC 2557. – Il présente une compatibilité binaire avec le codage des courriels MIME pour la transmission du HTML dans les messages électroniques.

Description

MHTML *(MIME Encapsulation of Aggregate HTML Documents)* est la tentative d'enregistrer localement dans un fichier une page web avec des ressources web incorporées distribuées. Comme expliqué dans l'introduction sur l' hypertexte , en règle générale un document HTML ne peut être présenté correctement que si le navigateur est en ligne et a accès à toutes les ressources incorporées. Avec le MHTML, toutes les ressources nécessaires à la présentation sont désormais enregistrées dans un fichier pour la navigation hors ligne. Cela signifie que lors de la conversion en MHTML, l'objet DOM actuel est sérialisé (sauvegardé) dans un fichier dans le navigateur. À cet effet, les contenus incorporés binaires passent par un encodage de type MIME, comme c'est le cas pour les messages électroniques.

Evaluation

Ouverture du format : **4**
La spécification MHTML est publiée dans le RFC 2557 de l'*Internet Engineering Task Force*.

Licence libre : **3**
Il n'existe pas de patente connue pour le MHTML.

Diffusion : **1**
Le format en tant que tel n'a pas connu une grande diffusion pour le stockage de pages web individuelles. Une grande diffusion indirecte est assurée par l'utilisation pour l'envoi de messages électroniques codés en HTML dans *Microsoft Outlook* et autres programmes de messagerie.

Fonctionnalités : **3**
MHTML permet de sauvegarder intégralement des pages internet individuelles avec toutes les ressources annexes (feuilles de style, images, etc.). Ça ne fonctionne de façon fiable que pour les pages web construites de manière simple. Il semble qu'il y ait des problèmes dès qu'on utilise JavaScript. MHTML ne peut pas représenter des sites internet entiers, c'est-à-dire avec des relations ou liens entre plusieurs pages.

Implémentation : **3**
Seuls *Microsoft Internet Explorer* et *Opera* prennent en charge le format MHTML pour l'enregistrement de pages web individuelles.

Densité de mémorisation : **1**
À cause de l'utilisation du codage Base64 pour contenus binaires d'une page web, la densité de mémorisation est faible.

Vérifiabilité : **2**
Il n'y a pas d'outil de validation connu. Les navigateurs ouvrent le format MHTML avec une grande tolérance aux erreurs et essaient si nécessaire de recourir aux URL de l'original.

Bonnes pratiques : **1**
MHTML ne peut pas être directement recommandé comme format d'archivage. Aucune utilisation dans ce sens n'est connue.

Perspectives : **2**
MHTML présente un certain potentiel en vue de développer un format adapté pour l'archivage.

Classe de formats : **D**
Il s'agit d'un développement de la norme MIME dont le potentiel est pour l'instant encore flou, mais pas d'un nouveau développement fondamental.

Conclusion

En s'appuyant sur le protocole MAIL, on adopte une solution éprouvée. Pour les pages internet (simples), cela résout le problème des ressources distribuées. Un fichier MHTML représente exactement un objet DOM dans le navigateur, mais il faut procéder autrement pour archiver des sites web entiers, car les hyperliens partant du fichier MHTML se rapportent à des ressources web. Comme critiqué dans d'autres cas (ARC et WARC), aucune normalisation de format n'est effectuée lors d'un enregistrement en tant que fichier MHTML non plus. Une animation flash reste une animation flash incorporée et exige un module d'extension *(plug in)* flash dans le navigateur lors de sa restitution.

Références

Internet Engineering Task Force RFC 2557
https://tools.ietf.org/html/rfc2557
Base64 décrit un processus pour le codage de données binaires, il fait partie de la norme MIME (Multipurpose Internet Mail Extensions)
https://de.wikipedia.org/wiki/Base64
Multipurpose Internet Mail Extensions (MIME)
https://tools.ietf.org/html/rfc2048

Bibliographie

A Simplified Guide to MIME
https://www.hunnysoft.com/mime/mime-guide.html

Articles connexes

HTML
HTML5

Contact
A propos
Impressum
Événements
Newsletter
RSS

PDF/A-2 pour hypertexte

On peut convertir en fichier PDF un document ou un objet DOM généré dans le navigateur lors du chargement d'un fichier HTML, exactement de la même manière qu'on peut l'imprimer. Lors de la conversion en PDF, toutes les ressources incorporées sont sauvegardées dans le fichier PDF et converties dans les formats audiovisuels correspondants selon les spécifications du convertisseur, particulièrement si on choisit comme format cible le PDF/A (par exemple, les images GIF seront converties en JPEG ou JPEG2000).
Chaque fournisseur résout la représentation des fonctionnalités hypertextes différemment. Le PDF utilise des liens (le PDF/A autorise des liens internes ou externes; les lecteurs PDF/A-1 ne sont pas censés exécuter des liens; les lecteurs PDF/A-2 devraient exécuter des liens internes, mais pas les liens externes, leur fonctionnement n'étant pas garanti).
Ainsi chaque page peut être reliée de la même manière. Il existe deux solutions différentes permettant de relier des pages web. La première approche établit un fichier PDF/A par page web et relie tous les fichiers PDF/A à un site internet complet. La deuxième approche consiste à sauvegarder dans le même fichier les pages web les unes derrière les autres dans l'ordre de l'exploration faite par le robot d'indexation. Ce faisant, un site web entier peut facilement dépasser la taille maximale d'un fichier PDF (8'388'607 objets, 10 Go; pour le PDF/A-1 la taille maximale est de 2 Go).
Les propriétés techniques du PDF/A-2 sont décrites en détail dans le chapitre sur les données textuelles . À cet endroit se trouve également une comparaison détaillée avec les versions 1 et 3. Il est nécessaire d'utiliser la version 2 comme format d'archivage parce que seule cette version autorise l'exécution de liens PDF internes et que la taille maximale des fichiers a été élevée à 2 Go.
Ci-après figure uniquement une brève évaluation du PDF/A-2 lorsqu'il est utilisé en tant que format d'archivage pour hypertexte .

Evaluation

Ouverture du format : **4**
PDF/A-2 est une norme ISO.

Licence libre : **3**
Le critère de licence libre est rempli, en particulier parce qu'il n'est pas possible d'utiliser tous les algorithmes de compression et que les polices sont encapsulées.

Diffusion : **3**
La diffusion du PDF/A-2 s'est beaucoup étendue ces dernières années dans les archives et le monde des affaires.

Fonctionnalités : **3**
En principe, l'aptitude du PDF/A-2 comme format d'archivage pour hypertexte dépendent de chaque page. On ne peut pas convertir en PDF/A-2 certains types de contenus incorporés de chaque page (par exemple des données vidéo).

Implémentation : **3**
Il existe d'une part des outils de création de documents PDF/A-2 qui sont capables de sauvegarder dans un fichier PDF non seulement des pages web, mais également des sites web entiers. D'autre part, il existe des solutions dédiées à l'archivage internet qui permettent aussi de sauvegarder en PDF/A-2 ou en PDF. Un fichier PDF peut être regardé par n'importe quel lecteur PDF.

Densité de mémorisation : **2**
Le PDF/A-2 sert de conteneur pour les diverses ressources HTML incorporées. Suivant la compression utilisée pour ces ressources, par exemple des images JPEG2000, la densité de mémorisation peut être relativement élevée. Un autre facteur dépend de la manière dont les différentes versions du site internet sont reliées entre elles.

Vérifiabilité : **4**
Les fichiers PDF/A-2 peuvent être reconnus par des logiciels de reconnaissance de formats. Il existe plusieurs validateurs pour ce format.

Bonnes pratiques : **3**
Le format est de plus en plus utilisé dans les archives et il est très bien accepté comme format d'archivage. Il ne revêt cependant jusqu'à maintenant qu'une faible importance dans le domaine de l'archivage web.

Perspectives : **4**
Le format a été développé en tant que format d'archivage et la suite de son développement bénéficie d'un suivi archivistique.

Classe de formats : **B**
Le format est actuellement en usage.

Conclusion

PDF/A en tant que format hypertexte présente dans tous les cas les avantages suivants: lors de la sauvegarde, la compression a lieu dans un format connu adapté pour l'archivage; tous les autres formats incorporés dans le HTML sont également incorporés de manière conforme au PDF/A; la fonctionnalité des hyperliens est conservée; il suffit d'un lecteur PDF pour regarder le fichier et il n'y a pas besoin de navigateur avec les modules d'extension (plug in) correspondants.
Il reste à tenir compte du fait qu'à la base de toute conversion de HTML en PDF se trouve un moteur de rendu HTML particulier donc un navigateur particulier. Le PDF archivé représente donc la vision du navigateur et pas la spécification universelle du document HTML. Cela signifie par exemple que l'élément de texte HTML abstrait « Titre 1 » sera représenté avec un certain formatage (corps, fonte de caractères et interligne).

WARC

Informations générales

Titre	Web ARChive file format
Catégorie	Hypertexte
Abréviation	WARC
Extension de fichier	.warc
Mime Type	application/warc
Pronom PUID	fmt/289
Version	Version actuelle: 2 (ISO-Standard 28500:2017) Version précédente: (ISO 28500:2009)

Description

Le format WARC *(Web ARChive)* est une extension du format ARC n'amène pas de concept fondamentalement nouveau Voir à ce sujet les réflexions sur ARC . Le format est publié en tant que norme ISO 28500:2017.

Evaluation

Ouverture du format : **4**
Le format a été développé par *Internet Archive* et la Bibliothèque nationale de France et il est disponible en tant que norme ISO 28500:2017.

Licence libre : **3**
Il n'existe pas de patente connue pour WARC.

Diffusion : **3**
La plupart des archives web comme *Internet Archive* utilisent aujourd'hui le format. La *Internet Memory Foundation* développe depuis 2012 un nouveau *Web Archive Repository*, qui doit être cependant compatible avec WARC.

Fonctionnalités : **3**
Le format possède de meilleures fonctionnalités par rapport à ARC .

Implémentation : **4**
L'*Internet Archive Wayback Machine* (« machine à remonter le temps ») peut travailler avec WARC. Différentes solutions de moissonnage *(harvesting)* peuvent sauvegarder dans ce format, par exemple le produit open source Heritrix .

Densité de mémorisation : **3**
La densité de mémorisation est relativement élevée du fait de l'utilisation d'une compression.

Vérifiabilité : **3**
L'*Internet Archive Wayback Machine* peut être mise à contribution pour la vérification de format.

Bonnes pratiques : **3**
WARC est en ce moment le format le plus répandu pour le moissonnage de sites web.

Perspectives : **2**
La perspective à long terme est floue parce qu'il semble que des problèmes d'échelle surviennent assez rapidement.

Classe de formats : **B**
Il s'agit d'un format bien établi.

Conclusion

Ni les fichiers ARC ni les fichiers WARC ne représentent une sérialisation des objets DOM constitués dans le navigateur. Pour regarder un contenu sauvegardé ARC ou WARC, un navigateur est tout aussi nécessaire qu'avant l'archivage. La quantité de formats utilisés ainsi que de langages de programmation et de scripts à interpréter n'a pas diminué. En revanche, le problème des ressources distribuées est résolu. Toutes les ressources nécessaires à un contenu web ou à un document sont rassemblées dans un fichier.

Références

ISO 28500:2017, Information and documentation — WARC file format
https://www.iso.org/standard/68004.html
WARC ISO 28500 Version 1 Latestdraft, 2008

https://archive.org/details/WARCISO28500Version1Latestdraft
WARC File Format Specifications (final draft)
http://archive-access.sourceforge.net/warc/WARC_ISO_28500_final_draft%20v018%20Zentveld%20080618.doc
Library of Congress: Sustainability of Digital Formats - WARC, Web ARChive file format
http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

Bibliographie

Internet Memory developed a new infrastructure with the ambition to reach « Web-scale »
http://internetmemory.org/en/index.php/News/workshop_at_the_iipc_2012_general_assembly_leveraging_web_archives_re
Stephan Strodl, Peter Paul Beran, Andreas Rauber: Migrating Content in WARC Files
https://publik.tuwien.ac.at/files/PubDat_181115.pdf

Articles connexes

ARC