Restructuration physique et logique de documents électroniques textuels

Jean-Luc BLOECHLE, Rolf INGOLD

Département d'Informatique, Université de Fribourg, CH-1700 Fribourg, Suisse.

Mots-clés : PDF, OCD, XML, structure physique, structure logique, modèle de document

Keywords : PDF, OCD, XML, physical structure, logical structure, document model

Résumé : La reconstruction des structures physiques et logiques de documents électroniques reste une problématique ouverte. Cet article présente une approche flexible et efficace permettant de régénérer de telles structures à partir de documents PDF. Une brève introduction présente tout d'abord le format PDF, ses atouts ainsi que ses défauts. Les principaux travaux dans le domaine de la restructuration de documents électroniques sont présentés. Un système complet de rétro-ingénierie du format PDF est ensuite exposé, celui-ci est basé sur une représentation intermédiaire appelée le document canonique, et permettant d'exprimer la structure physique tout en conservant l'apparence originale du document. L'étape finale de notre système d'analyse, la restructuration logique, est particulièrement mise en évidence. L'article conclut en exposant les travaux actuels et les éventuels améliorations futures.

Abstract : Physical and logical structure recovering from electronic documents is still an open issue. In this paper, we propose a flexible and efficient approach for recovering document structures from PDF files. After a brief introduction of the PDF format and its major features, we report about different existing works for PDF content extraction and analysis. To overcome the weaknesses of these systems, we propose a new analysis strategy, based on an intermediate representation, called canonical document, which enables representing physical structures in a canonical way. This paper then describes the PDF reverse engineering workflow and focuses on the document logical restructuring. Finally, the paper concludes with potential future improvements.

1 Introduction

Depuis sa publication en 1993, le format PDF de Adobe Systems est devenu le format standard pour l'échange et l'archivage de documents électroniques textuels et graphiques. En effet, le format PDF permet de restituer fidèlement l'apparence d'un document électronique quelconque aussi bien sur un écran que sur une imprimante. D'après Adobe Systems Incorporation, plus de 200 millions de documents PDF sont disponibles sur le web. Le format PDF peut être considéré comme un format universelle dans le sens où il est capable de reproduire toute information imprimable telle que du texte, des graphiques, des images, etc. Dans l'article "Why PDF is Everywhere" [1], McKinley met en évidence les points forts de ce format pour la gestion de documents et la recherche d'information. Le format PDF est d'ailleurs reconnu par les industries et gouvernements du monde entier. Dernièrement, un standard ISO a même été développé par l'organisation internationale pour la standardisation dans le but de spécifier un format PDF épuré nommé PDF/A et destiné à l'archivage à long terme.
Malgré toutes les qualités précitées, le format PDF n'est de loin pas parfait. En réalité, la spécification PDF a été définie afin de pouvoir reproduire tout document imprimable fidèlement et ceci au détriment de sa structure interne. Bien que les récentes spécifications du format PDF permettent d'incorporer des méta-données au contenu, la plupart des imprimantes PDF actuelles n'utilisent pas de telles possibilités. En conséquence, beaucoup de caractéristiques intéressantes liées aux structures du document sont perdues, alors qu'elles existaient au moment de l'édition. Cette perte d'information limite grandement la réutilisation de documents PDF, par exemple, la réédition ou le reformatage sont impossible, tandis que même des opérations aussi simple que copier/coller sont compromises.

Figure 1 : trois types de segmentation textuelle originale de documents PDF.

Il est intéressant de constater que la segmentation textuelle originale des documents PDF est totalement imprévisible. Aucune segmentation en mots ou unités lexicales n'est assurée par le format PDF, puisqu'il a pour unique but un rendu correct de telles entités, laissant leur représentation

interne au bon vouloir de l'imprimante PDF. La Figure 1 expose justement trois extraits de journaux au format PDF ayant des segmentations textuelles très diverses.
Au niveau logique, la séquence des blocs de texte n'est également pas assurée. De ce fait, la sélection ou l'exportation de texte avec Adobe Acrobat peut engendrer quelques surprises comme le présente la Figure 2.


Is online piracy unstoppable?

By Steve Lohr

The recording industry's long-running battle against online music piracy has come to resemble one of those "whack-a-mole" arcade games, in which the player hammers one rubber rodent's head with a mallet only to see another pop up nearby. Conk one, and up pops another, and so on. Three years ago, the music industry sued Napster, the first popular music file-sharing network on the Internet.

That sent Napster reeling, but other networks for trading copyrighted music — Kazaa, Grokster, Morpheus and others — sprang up. Last week, in the latest swing of the hammer, the Record-ing Industry Association of America filed 261 lawsuits against individual file sharers, which will surely make some of their estimated 60 million compatriots think twice — for now.

News Analysis

Earth Station Five, a company based in the West Bank, surfaced recently with claims of being at war with the industry association. It promises the latest in anonymous Internet file sharing. Its motto: "Resistance is futile."

Since Gutenberg's printing press, new technologies for creating, copying and distributing information have eroded the power of those in control of various media. In the last century, the pattern held true, for example, when recorded music became popular in the early 1900's, radio in the 1920's and cable television in recent years. But the heritage and design of the Internet present a particularly disruptive technology. Today's global network had its origins in the research culture of academia with its ethos of freely sharing information. And by design, the Internet turns every user in every living room into a mass distributor of just about anything that can be digitized.

See MUSIC, Page 7

Musicians have conflicting views on song-swapping. Page 9

Figure 2 : une sélection multicolonnes erronée ne respectant pas l'ordre de lecture.

2 Taxonomie des méthodes existantes pour l'analyse de PDF

Un nombre restreint de travaux et recherches ont été accomplis [2] afin d'exploiter le contenu des documents PDF, d'en extraire les structures physiques et logiques, et d'en dériver certaines annotations.
L'analyse de l'image du document bénéficie de méthodes qui ont mûri durant ces dernières décénies, de telles méthodes peuvent également être appliquées à des documents synthétique, sans bruits et imprimés en haute résolution [3], afin de retrouver le contenu et les structures originales de documents électroniques. Tandis que l'analyse direct du contenu électronique du document [4] profite de techniques partiellement dérivées de celles de l'analyse d'image. Ces méthodes récentes utilisent les primitives internes des document PDF [5]. Dans [6, 7], nous avons proposé de mélanger les deux méthodologies afin de pouvoir analyser tout type de PDF.
L'analyse du contenu électronique est à son tour composée de méthodes extensives et de restructuration. Les premières analysent le contenu du document afin de reconstituer les structures originales et y ajouter des annotations (tags PDF) sans réorganisation des primitives du document électronique. Ces techniques ont été appliquées avec des résultats intéressants dans plusieurs travaux [8, 9, 10]. L'objectif des techniques de restructuration est de représenter le document électronique en utilisant un

format différent du PDF, par exemple XML, pour permettre d'accéder facilement à l'information. Le cas le plus intéressant de restructuration est celui de la ré-ingénierie, qui vise à réorganiser le contenu du document en fonction des structures découvertes [11, 12, 13, 14, 15]. La conversion est un cas particulier de restructuration dans lequel aucune structure n'est extraite, le fichier PDF étant simplement transformé dans un format plus facile à manier [2].

3 Format canonique et restructuration physique

Le format canonique est un format développé au sein de notre groupe de recherche préservant fidèlement l'apparence d'un document électronique tout en y incorporant ses structures physiques. Le processus permettant de générer un tel document est le suivant : le contenu d'un fichier PDF est tout d'abord extrait par XED [7], puis la restructuration physique du document au format canonique est effectuée en utilisant une approche hybride. La restructuration physique a pour but de segmenter l'information textuelle en paragraphes homogènes composés de lignes elles-mêmes composées d'unités lexicales. L'algorithme de restructuration est divisé en trois phases :

- pré-traitement : normalisation, cristallisation, tri;
- phase ascendante : lexicalisation, linéarisation, fusion en blocs, fusion rétroactive, post-linéarisation;
- phase descendante : détection de changement d'interligne, détection de changement d'alignement.

Figure 3 : texte PDF brut à gauche et document canonique à droite

Toutes les étapes de l'algorithme utilisent des seuils dynamiques, relatifs à la taille de la police courante, permettant de fusionner ou segmenter le texte avec précision. La recherche des seuils a été faite empiriquement, tout d'abord par une estimation a priori de leurs valeurs, puis par un affinage minutieux sur un corpus éclectique de documents PDF. Quatre seuils ont été nécessaires au bon fonctionnement de l'algorithme: un seuil pour la fusion des caractères en mots, un seuil pour la fusion des mots en

lignes, un seuil pour la fusion des lignes en blocs de texte, et finalement un seuil plus général appelé seuil de précision (utile pour des tests d'alignement ou d'interligne par exemple). Une présentation détaillée de l'algorithme a déjà été présentée dans [2] et [16]. La Figure 3 ci-dessous présente un extrait de texte PDF brut à gauche, puis sa version segmentée à droite.
L'extraction de la structure physique a été appliquée sur trois documents différents, dont deux à structures complexes, les résultats obtenus sont exposés sur le Tableau 1.

Document title	number of text blocks	number of errors	correctness
French newspaper - Le Monde 2009/01/16	1827	35	98.08%
Swiss newspaper - La Liberté 2009/01/15	1410	16	98.87%
E-book - Alice's Adventures in Wonderland	1108	7	99.37%

Tableau 1 : résultats de l'extraction de la structure physique sur trois documents.

4 OCD, un formalisme XML optimisé pour le contenu physique

Le stockage permanent d'un document canonique au format OCD (Optimized Canonical Document) [16] permet à la fois de représenter la structure physique et de garantir la reproduction fidèle de ce document. Le format OCD est une description XML compacte et simple permettant le stockage permanent d'un document au format canonique sur un support physique. Son but n'est pas de concurrencer un quelconque autre format, mais bien de conserver un document structuré tout en préservant son aspect visuel d'origine, et cela d'une manière simple et synthétique. L'accès aux informations d'un tel format doit être facilité au maximum.

```
<block x="96.78" y="97.8">
 <line>
  <token tm="13.5" font-id="1" cs="0" ts="0">
43 48 41 50 54 45 52</token>
  <token ws=".3328"/>
  <token>49 49</token>
  <token>3a</token>
  <token/>
  <token>54 68 65</token>
  <token/>
  <token>50 6f 6f 6c</token>
  <token/>
  <token>6f 66</token>
  <token/>
  <token>54 65 61 72 73</token>
 </line>
</block>
```

Figure 4 : un extrait du format canonique représenté en OCD.

OCD supporte trois sortes de primitives graphiques : texte, image, et graphique vectoriel. Chaque primitive textuelle, graphique ou image y est décrite relativement à un état graphique de la page courante. Ainsi, un attribut est déclaré uniquement si celui-ci a changé de valeur relativement à l'état graphique qui lui-même mis à jour avec la nouvelle valeur de l'attribut. Les représentations des primitives utilisent des descriptions synthétiques. Les images sont compressées au formats JPG ou PNG puis insérées dans le document XML sous forme de flux hexadécimal. Les graphiques utilisent une description similaire à SVG, des coordonnées relatives sont employées à l'intérieur d'un même graphique. La représentation du texte bénéficie grandement du regroupement homogène des entités textuelles du format canonique permettant ainsi une description très réduite. Les primitives textuelles utilisent les largeurs de caractère de la fonte courante ainsi que des opérateurs d'espacement de caractère, de mot et d'interligne (cf. Figure 4). Le positionnement de chaque caractère est de ce fait respecté avec précision et cela avec un minimum d'espace disque. Finalement le fichier XML résultant est compressé en suivant le standard GZIP.
Ainsi, bien que OCD soit basé sur une représentation XML, sa taille est extrêmement réduite. Le Tableau 2 montre en effet que, par rapport au format PDF, notre format OCD permet de substantielles réductions de tailles de fichiers sur des documents textuels. Le tableau compare également notre format de fichiers aux formats XPS (le format de Microsoft) et XCD (ou XCDF, notre ancien format de stockage de documents canoniques [2]).

Document title	PDF	XPS	OCD	XCD
Aesop's Fables - 93 pages	243 KB	441 KB	91.9 KB	7'014 KB
Around the World in 80 Days - 339 pages	766 KB	1'912 KB	422 KB	36'045 KB
The Odyssey - 550 pages	1280 KB	3082 KB	850 KB	63'693 KB
The Last of the Mohicans - 698 pages	1'605 KB	3'944 KB	924 KB	80'896 KB
Ulysses - 1305 pages	2'953 KB	7'334 KB	1'743 KB	-

Tableau 2 : évaluation du format OCD par rapport à PDF, XPS, et XCD.

5 Dolores : un outil interactif pour la restructuration logique

A partir d'un document au format canonique, Dolores [17] (Document Logical Restructuring) permet de régénérer une structure logique par apprentissage interactif incrémental. L'utilisateur créé un modèle par interaction, apprentissage et correction. Il peut ensuite l'appliquer à d'autre documents d'une même classe et améliorer ce même modèle grâce

à l'apprentissage incrémental (cf. Figure 5). Trois phase principales peuvent être mise en évidence dans ce processus : l'extraction des caractéristiques, l'étiquetage logique et l'apprentissage.

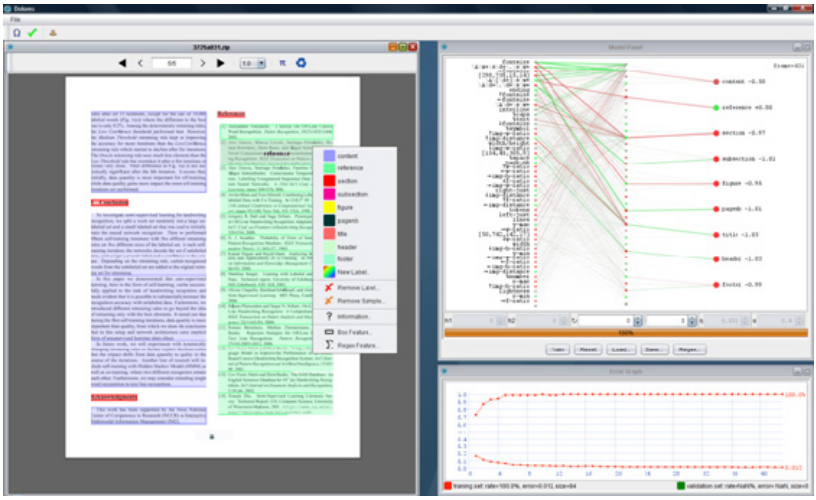

Figure 5 : Capture d'écran de Dolores, à gauche le document étiqueté, à droite le modèle.

6 Extraction des caractéristiques

L'extraction des caractéristiques est une tâche primordiale préalable à la phase d'apprentissage du système. Le choix des caractéristiques extraites, leur nombre, leur pertinence a un impact direct sur la création du modèle de document et donc sur les résultats de la classification. Dolores extrait un ensemble de caractéristiques de natures diverses sur chaque bloc textuel : géométriques, typographiques, topologiques.
Les caractéristiques extraites sur chaque bloc textuel sont les suivantes: coordonnée x/y, largeur, hauteur, rapport largeur/hauteur, taille de la fonte, interligne, luminosité de la fonte, écart type de la justification à gauche/droite, nombre de mots, nombre de lignes, pourcentage de majuscules, pourcentage de symboles, pourcentage de mots, pourcentage de nombres, pourcentage d'espaces, présence d'un caractère de ponctuation en fin de bloc, numéro de page, distance aux blocs textuels voisins (supérieur/inférieur/droite/gauche), tailles des fontes des blocs textuels voisins, rapports de la fonte courante aux tailles des fontes des blocs textuels voisins, rapport de la largeur du bloc courant aux blocs textuels voisins, distance aux images voisines, rapport de largeur du bloc courant aux images voisines.

Deux autres classes de caractéristiques sont également prises en compte : les régions et les expressions régulières. Concernant les régions, l'intersection des surfaces des blocs de texte (boîte englobante) d'une même classe est calculée, si celle-ci n'est pas nulle, la boîte englobante résultante est ajoutée comme caractéristique au modèle. La valeur de la caractéristique est le pourcentage de recouvrement de la surface d'intersection avec le bloc de texte courant. Concernant les expressions régulières, le principe est le même, une expression régulière est générée pour chaque échantillon (bloc de texte), l'expression régulière commune à chaque classe est recherchée, en cas de succès, celle-ci est ajoutée au caractéristiques du modèle.

6.1 L'étiquetage logique

La figure 5 montre l'interface de Dolores. L'étiquetage logique y est effectué d'une manière interactive. En effet, l'utilisateur peut ajouter ou supprimer des étiquettes lorsque bon lui semble. Le système d'apprentissage ajoute dans le modèle tout nouveau bloc de texte étiqueté. Une phase d'entraînement est ensuite instantanément effectuée, les blocs de texte sont alors étiquetés à la volée. L'action de l'utilisateur (l'étiquetage) est directement suivi de la mise à jour du modèle et reflété au travers de l'interface. L'utilisateur voit les erreurs d'étiquetage et corrige celles-ci de manière itérative. L'utilisateur peut étiqueter un bloc par l'intermédiaire du menu contextuel de la souris, ou alors directement en cliquant sur celui-ci si la classification actuelle est adéquate. De plus, dans le cas ou tous les blocs de texte d'une page sont correctement étiqueté, l'utilisateur peut insérer ceux-ci en vrac en allant dans le menu contextuel et en cliquant sur "étiqueter page" (ce menu ne peut apparaître que lorsque le pointeur de souris est à l'extérieur de tout bloc de texte et que l'utilisateur clique sur le bouton droit).
L'interface fournit des informations cruciales à l'utilisateur, lui permettant d'effectuer son étiquetage aisément et rapidement. Par exemple, la classe (l'étiquette logique) attribuée à chaque bloc de texte par le modèle est représentée par une surface rectangulaire colorée et semi-transparente (la couleur étant définie au préalable par l'utilisateur). Chaque bloc de texte contenu dans l'ensemble d'entraînement est encadré par un rectangle englobant dont la couleur correspond à celle de son étiquetage. Une barre horizontale est également affichée en-bas de chaque bloc de texte, son pourcentage de remplissage exprime le taux de confidence de l'étiquette attribuée par le modèle. Ainsi un taux de confidence bas indique qu'il est préférable de continuer à étiqueter la classe correspondante. Finalement, lorsque l'utilisateur passe sur un bloc de texte, celui-ci est mis en évidence par la superposition d'une surface rectangulaire grise semi-transparente, son étiquette logique s'affiche au centre de celui-ci, le code

couleur pouvant parfois s'avérer insuffisant (s'il y a beaucoup de classes par exemple).

6.2 Modèle et apprentissage

L'apprentissage est géré par un perceptron multicouches. Le modèle de document comprend à la fois l'ensemble des échantillons étiquetés (blocs de texte) ainsi que les données définissant le réseau de neurone. Une interface simple et conviviale implique que l'apprentissage soit totalement automatisé et instantané. De ce fait, la topologie du réseau est dynamique, elle s'adapte automatiquement au nombre d'entrées et de sorties. Le réseau contient une couche cachée. La couche d'entrée est totalement connectée à la couche cachée tandis que chaque neurone de la couche de sortie est connecté à quatre neurones de la couche cachée. Ceci assure à chaque neurone de sortie un nombre égale de neurones caché et évite que ceux-ci soit accaparés par un autre neurone de sortie dont la probabilité a priori est beaucoup plus élevée. Sans entrer dans les détails, l'algorithme d'entraînement du réseau est une rétro-propagation stochastique avec moment d'inertie. Le taux d'apprentissage diminue en fonction de l'erreur en sortie d'un neurone. Ces caractéristiques assurent un apprentissage convergeant et rapide, tout en minimisant le risque de stagner dans des minima locaux. Actuellement, l'apprentissage s'arrête lorsque le taux de reconnaissance est de 100% sur un minimum de 30 cycles consécutif (avec une borne temporel).
L'affichage du réseau neuronal met en évidence la force des pondérations ainsi que la pertinence de chaque caractéristique d'entrée par rapport à l'ensemble des classes ou alors pour une classe donnée (en pointant un neurone de sortie avec le curseur de la souris). Ceci permet à l'utilisateur d'appréhender d'un seul regard les caractéristiques discriminantes du réseau dans sa globalité ou pour chaque classe séparément. L'interface du réseau de neurone offre également la possibilité de désactiver un neurone d'entrée, afin de voir son impact sur le modèle. Un graphe d'erreur est affiché en dessous du réseau de neurones, il contient la courbe d'erreur ainsi que le taux de reconnaissance sur l'ensemble d'apprentissage et éventuellement sur un ensemble de validation/test. Enfin, il est possible de sauvegarder et d'ouvrir les modèles afin de les appliquer sur d'autres documents, ou éventuellement de les améliorer.

7 Conclusion

Cette article présente un système complet d'analyse de documents électroniques textuels. A partir d'un document PDF, ou tout autre document électronique textuel imprimable, le système extrait toutes les données textes, images et graphiques. Une restructuration physique est

ensuite effectuée sur le document, le résultat est alors sauvegardé au format OCD. L'étape de restructuration logique est assurée par Dolores, un outil interactif pour l'apprentissage incrémental de modèles de documents. Actuellement, seul les étiquettes logiques sont supportées par le modèle. La reconstruction de la hiérarchie fait partie des travaux futurs. Tandis que l'étude approfondie de la génération des modèles, ainsi que l'impact des divers paramètres d'apprentissage sur le taux de reconnaissance sont en cours d'évaluation. Le résultat de la restructuration logique d'un document peut finalement être conservé directement dans le format canonique au moyen de liens internes et sauvegardé sur disque grâce à un format étendant OCD nommé OCDL.
Le développement d'un processus complet permettant la réutilisation de contenus PDF est une gageure qui ne saurait être mise de côté, en effet, un tel processus permet de réactiver le cycle de vie des documents électroniques.

8 Références bibliographiques

McKinley, T. Why PDF is Everywhere. Inform, the journal of AIIM, 11(8), 1997.

Bloechle, J.-L., Rigamonti, M., Hadjar, K., Lalanne, D. and Ingold, R. XCDF: A canonical and structured document format. In 7th International Workshop, DAS'06, pages 141-152, Nelson, New Zealand, February 2006. Springer-Verlag.

Hadjar, K. and Ingold, R. Arabic Newspaper Page Segmentation. In Proceedings of the Seventh international Conference on Document Analysis and Recognition - Volume 2 (August 03 - 06, 2003). ICDAR. IEEE Computer Society, Washington, DC, 895.

Paknad, M.D. and Ayers, R.M., Method and apparatus for identifying words described in a portable electronic document, U.S. Patent 5,832,530, 1998.

Rigamonti, M., Bloechle, J.-L., Hadjar, K., Lalanne, D. and Ingold, R. Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering. ICDAR'05, 2005, pp. 1050-1054.

Hadjar, K., Rigamonti, M., Lalanne, D. and Ingold, R. Xed: a new tool for eXtracting hidden structures from Electronic Documents. DIAL'04, 2004, pp. 212-221.

Rigamonti, M., Hadjar, K., Lalanne, D. and Ingold, R. Xed: un outil pour l'extraction et l'analyse de documents PDF, CIFED'04, 2004, pp. 85-90.

Bagley, S.R., Brailsford, D.F. and Hardy, M.R.B. Creating reusable well-structured PDF as a sequence of component object graphic (COG) elements. DocEng'03, 2003, pp. 58-67.

Hardy, M.R., Brailford, D. and Thomas, P.L. Creating Structured PDF Files Using XML Templates, DocEng'04, 2004, pp. 99-108.

Lovegrove, W.S. and Brailsford, D.F. Document analysis of PDF files: methods, results and implications. Electronic Publishing, 1995, pp. 207-220.

Anjewierden, A. AIDAS: Incremental logical structure discovery in PDF document. ICDAR'01, 2001, pp. 374-377.

Chao, H. and Fan, J., Capturing the Layout of electronic Documents for Reuse in Variable Data. ICDAR'05, 2005, pp. 940-944.

Dejan, H. and Meunier, J.L.,A System for Converting PDF Documents into Structured XML Format. DAS'06, 2006, pp. 129-140.

Futrelle, R.P., Shap, M., Cieslick, C. and Grimes, A.E. Extraction, layout analysis and classification of diagrams in PDF documents. ICDAR'03, 2003, pp. 1007-1012.

Rahman, F. and Alam, H. Conversion of PDF documents into HTML: a case study of document image analysis. Asilomar CSS'03, 2003, pp. 87-91.

Bloechle, J.-L., Lalanne, D. and Ingold, R. OCD: An Optimized and Canonical Document Format. In 10th International Conference on Document Analysis and Recognition, ICDAR'09, Barcelona, Spain, July 2009, pp. 236-240.

Bloechle, J.-L., Pugin, C. and Ingold., R. Dolores: An Interactive and Class-Free Approach for Document Logical Restructuring. In 8th International Workshop, DAS'08, pages 644-652, Nara, Japan, September 2008.