

檔案管理局

電子檔案保存管理機制委託服務案

電子檔案相關技術與實例

電子檔案保存工具

(1011130_01)

英福達科技股份有限公司謹呈

民國 101 年 11 月

目錄

壹、 序論	- 1 -
一、 電子檔案格式轉置之時機.....	- 1 -
二、 格式轉置注意事項.....	- 2 -
貳、 系統介紹	- 3 -
一、 轉置功能.....	- 3 -
二、 驗證功能.....	- 4 -
三、 工作管理功能.....	- 4 -
四、 其他功能.....	- 4 -
五、 設定功能.....	- 4 -
參、 系統安裝	- 5 -
一、 執行環境需求.....	- 5 -
二、 安裝程序.....	- 6 -
肆、 轉置功能	- 7 -
一、 封裝檔轉置.....	- 7 -
二、 封裝檔轉置報告.....	- 8 -
三、 檔案轉置.....	- 9 -

四、 檔案轉置報告	- 11 -
伍、 驗證功能	- 12 -
一、 JHOVE 格式介紹	- 12 -
二、 影像品質驗證	- 13 -
三、 視訊品質驗證	- 18 -
四、 其他驗證	- 20 -
陸、 其他功能	- 21 -
一、 影像辨識	- 21 -
二、 影像及視訊修補	- 24 -

壹、序論

電子檔案是以特定格式儲存之資料，必須經由應用程式方能予以呈現，例如：Microsoft Office 中的 Word 檔案，其中記載了各段落之文字格式，必須依賴能解讀 Word 格式的應用程式才能將其呈現與輸出；各種應用程式也必須執行在作業系統之上，某些作業系統更只能執行在特定硬體架構上，藉由格式轉置可將原依賴 Word 應用程式開啟的電子檔案轉置成其他格式，便可由其他應用程式開啟，即為電子檔案格式轉置。

一、電子檔案格式轉置之時機

- (一) 儲存的電子檔案格式為非開放格式。
- (二) 電子檔案格式可能過時或因普及率不高，應用程式廠商可能停止提供支援或服務。
- (三) 支援電子檔案之應用程式相依之作業系統或硬體架構無法繼續獲得。
- (四) 由於長期保存之需要，必須將特定格式之檔案轉置為適合長期保存之格式。
- (五) 配合法令規定。

二、格式轉置注意事項

- (一) 進行格式轉置前，應評估相關應用系統是否需要一併調整或更新。
- (二) 應建立格式轉置標準作業程序，並產出轉置作業相關文件，以利日後查核及作業改善之依據。
- (三) 轉置後之目的格式必須符合長期保存要求。
- (四) 轉置品質檢驗。
- (五) 應將轉置工具名稱、轉置工具版本、轉置參數、品質驗證結果及歷史版本資訊等轉置相關資訊記載於詮釋資料中，與轉置後檔案一併保存。
- (六) 格式轉置前，應評估相關檔案是否完整，並應進行格式辨識及掃毒程序。

為了避免重要的電子檔案遇到無法開啟的問題，則使用格式轉置進行資料保存，但於轉置過程中，資料可能因為格式的更新而有所損失或有所修改而失去原意(例如影像檔案 BMP 格式轉換為破壞性壓縮的 PNG 格式)，也可能導致珍貴資料的流失，且不斷轉置而越形嚴重。

因此檔案管理局委外開發「電子檔案保存工具」，係整合 OpenSource 軟體及國內大學技術轉移工具，主要提供電子檔案的轉置、驗證及其他功能(修補、OCR 辨識)，以利於電子檔案長期保存。

貳、系統介紹

電子檔案保存工具(Preserving Electronic Archives & Records Suite, PEARS)區分為簡易版及完整版 2 種版本。完整版提供「轉置」、「驗證」、「影像及視訊修補」、「文字辨識」及「條碼辨識」等功能；簡易版只提供「轉置」及「驗證」功能。

一、轉置功能

轉置電子檔案封裝檔或單一檔案(如：WDL、TIFF、DOC等格式)。電子檔案封裝檔的轉置過程中，檢測檔案是否有病毒、格式是否正確、封裝檔是否符合「文書及檔案管理電腦化作業規範」之格式規定及轉置檔案的驗證。



圖 1 電子檔案保存工具之轉置功能

二、 驗證功能

利用客觀的數學公式來檢測轉置後檔案的品質差異，以確保轉置後檔案與原檔案的內容格式的正確性。

三、 工作管理功能

提供工作停止及重新執行之動作。

四、 其他功能

提供 OCR 辨識、條碼辨識、老舊照片及影片修補功能。

五、 設定功能

可設定相關設定值，例如：驗證方法、轉置檔案儲存路徑或附加軟體位置。

參、系統安裝

一、執行環境需求

(一) 軟體環境需求

1. 作業系統：Windows XP 或 Windows 7(32 位元)。
2. 安裝 Microsoft .NET Framework 3.5 版本。
3. 安裝電子檔案檢測及瀏覽軟體。
4. 安裝 PDFCreator 1.2.2 版本。
5. 安裝 Adobe Reader 9.0 以上版本。
6. 安裝 WDL 閱讀軟體(DynaDoc Free Reader 2005 TC)。
7. 安裝 Microsoft Office 2003 以上版本。
8. 安裝 MCRInstaller。
9. 安裝 gs900w32。
10. 安裝 Combined-Community-Codec-Pack。
11. Microsoft Office Document Imaging 工具。
12. 安裝 GIMP 2.6.11。
13. 安裝 ImageMagick。
14. 安裝 JAVA JDK。
15. 安裝 Windows Installer 3.1。
16. 安裝 Microsoft SQL Server Compact 3.5。
17. 安裝 OpenOffice。

(二) 硬體需求

1. CPU：Pentium 4 以上。
2. 記憶體：1GB 以上。
3. 硬體空間：至少 10GB 以上。

二、安裝程序

(一) 下載方式

於電子檔案技術服務中心網站自行下載「電子檔案保存工具」必要安裝程式及主程式(網址：<http://erlp.archives.gov.tw>)。或申請寄送「電子檔案保存工具」安裝光碟。



圖 2 電子檔案技術服務中心首頁

肆、轉置功能



圖 3 電子檔案保存工具之轉置

一、封裝檔轉置

電子檔案封裝檔格式轉置(xml)，可設定電子檔案封裝檔(含附件)須轉置之格式項目，工具轉置完成後，將相關轉置資訊註記於電子檔案封裝檔格式轉置格式中。



圖 4 提供之封裝檔轉置格式



圖 5 封裝檔轉置功能畫面

二、封裝檔轉置報告

封裝檔轉置完畢後，可點選封裝檔轉置報告，可看到工作報告內容。

Microsoft Excel - 封裝檔轉置報告.xls

	A	B	C	D	E	F	G	H	I	J	K	L
1	封裝檔：	D:\Demo1\0970001476\0970001476.xml										
2	儲存位置：	C:\LTPPALTPPData\201110280201480\201110280201480										
3												
4			序號	原始檔案	原始檔案大小(KB)	轉置後檔名	轉置後檔案大小(KB)	轉置開始時間	轉置結束時間	花費時間(秒)	狀態	轉置後檔案儲存位置
5			1	lpage.wdl	71	lpage.pdf	100	2011/10/28 02:02	2011/10/28 02:02	0.277660075	成功	C:\LTPPALTPPData\201110280201480\201

圖 6 封裝檔轉置報告

三、檔案轉置

電子檔案保存工具提供單一檔案及批次檔案的格式轉置功能，電子檔案格式轉置項目如下。

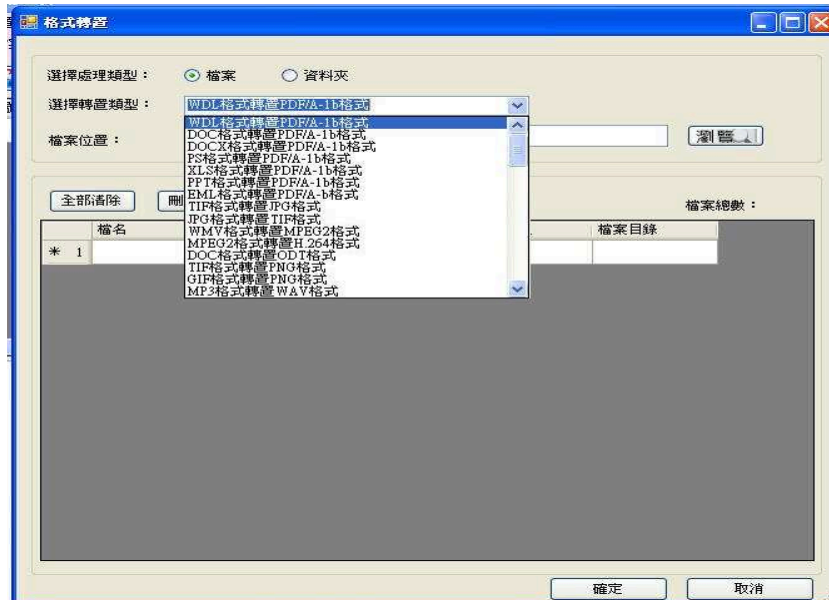


圖 7 格式轉置

(一) 文件類型格式轉置

1. WDL 格式轉置 PDF/A 格式。
2. DOC 格式轉置 PDF/A 格式。
3. DOCX 格式轉置 PDF/A 格式。
4. DOC 格式轉置 ODT 格式。
5. POSTSCRIPT 格式轉置 PDF/A 格式。
6. PPT 格式轉置 PDF/A 格式。
7. XLS 格式轉置 PDF/A 格式。
8. EML 格式轉置為 PDF/A 格式。

(二) 影像類型格式轉置

1. TIFF 格式轉置 JPEG 格式。
2. JPEG 格式轉置 TIFF 格式。
3. TIFF 格式轉置 PNG 格式。
4. GIF 格式轉置 PNG 格式。

(三) 視訊類型格式轉置

1. WMV 格式轉置 MPEG-2 格式。
2. MPEG-2 格式轉置 H.264 格式。

(四) 聲音類型格式轉置

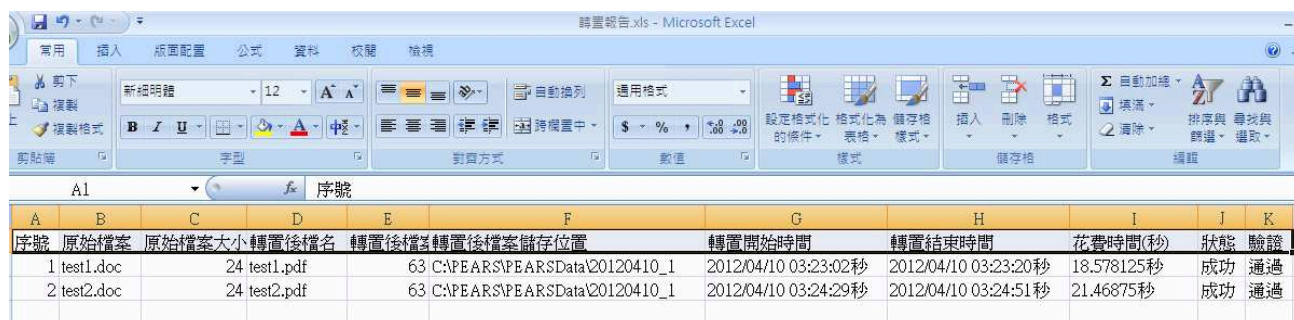
1. MP3 格式轉置為 WAV 格式。



圖 8 檔案轉置執行完畢畫面

四、檔案轉置報告

檔案轉置完畢後，可點選檔案轉置報告，可看到工作報告內容。



序號	原始檔案	原始檔案大小	轉置後檔名	轉置後檔案大小	轉置後檔案儲存位置	轉置開始時間	轉置結束時間	花費時間(秒)	狀態	驗證
1	test1.doc	24	test1.pdf	63	C:\PEARS\PEARSData\20120410_1	2012/04/10 03:23:02秒	2012/04/10 03:23:20秒	18.578125秒	成功	通過
2	test2.doc	24	test2.pdf	63	C:\PEARS\PEARSData\20120410_1	2012/04/10 03:24:29秒	2012/04/10 03:24:51秒	21.46875秒	成功	通過

圖 9 檔案轉置報告

伍、驗證功能

電子檔案保存工具整合格式檢測工具 Jhove，進行電子檔案格式確認，並整合清華大學所開發的影像視訊品質評估工具，可透過此工具針對轉置後檔案進行品質驗證，確保轉置品質符合要求。

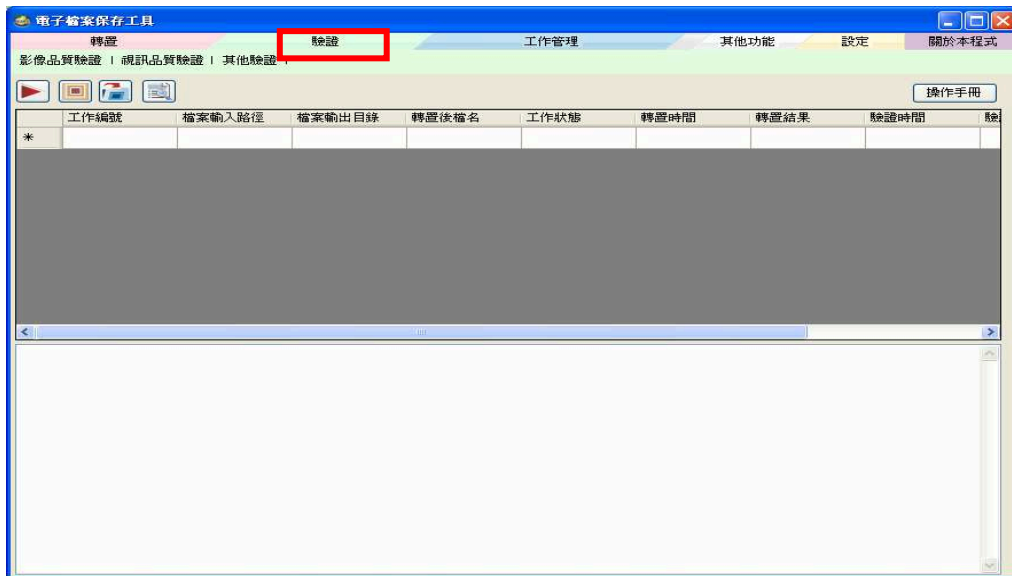


圖 10 電子檔案保存工具之驗證

一、JHOVE 格式介紹

由 JSTOR 與哈佛大學圖書館共同發展電子檔案型態描述 (Characterization) 工具，功能摘述如下：

- (一) 識別(Identification)：經由特徵比對，推定電子檔案格式。
- (二) 確認(Validation)：判斷電子檔案之格式完好(Well-formed) 且具有有效性(Valid)。
- (三) 提供 Module：AIFF、ASCII、Bytestream、GIF、HTML、JPEG、JPEG2000、PDF、TIFF、UTF-8、WAVE、XML

等格式辨識模組。

(四) 屬性擷取：格式、版本、作者、主題、產生時間、修改時間、字型、TrueType 字體、頁數。

二、影像品質驗證

常見的客觀評量方法有影像方面的均方差(Mean Square Error, MSE)、峰值信號雜訊比(Peak Signal to Noise Ratio, PSNR)、結構相似性指標(Structure Similarity, SSIM)、通用性影像品質指標(Universal Quality Index, UQI)及基於人類視覺系統之峰值信號雜訊比(HVS- Peak Signal to Noise Ratio, HVS-PSNR)的評估方式，其中最廣泛被使用的客觀評量方法是峰值信號雜訊比(Peak Signal to Noise Ratio, PSNR)。

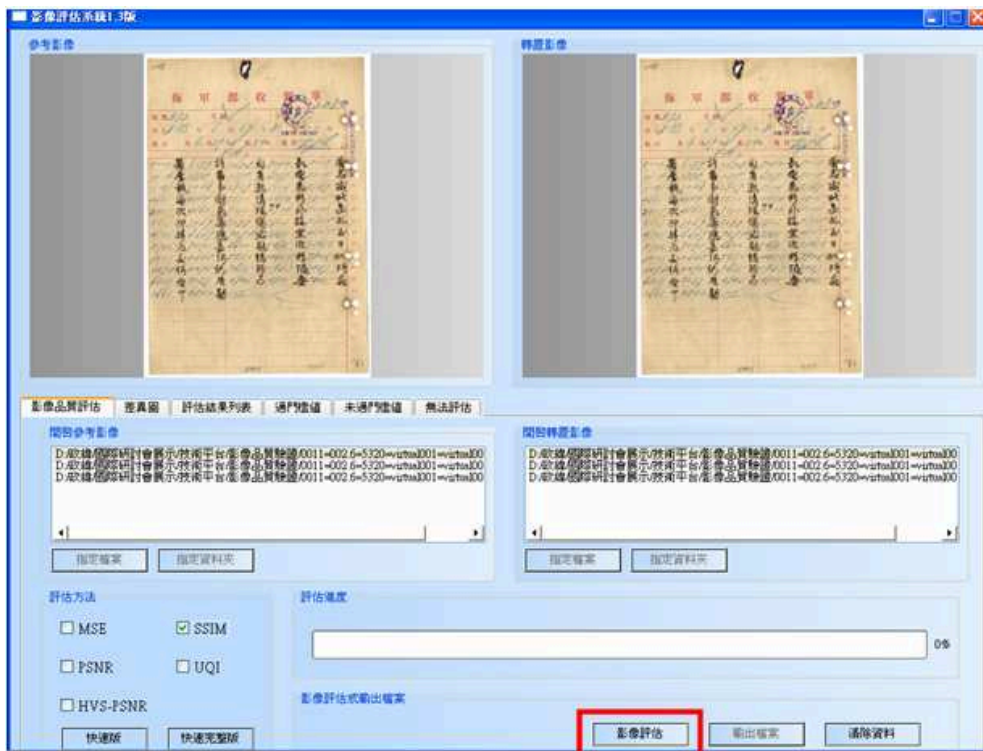


圖 11 影像品質評估

藉由影像品質評估的公式計算，訂定出評估分數的標準作為品質的界定，將轉置前與轉置後的圖像作評估並給於評估值，確認轉置後的影像是否達到一定的品質水準。

(一) 均方差(Mean Square Error, MSE)

均方差為一般影像評估常用的方式之一，求出來的 MSE 值越小，表示輸入影像與複製影像之間的差異少，也就代表品質較好。均方差主要計算方式是將輸入影像以及複製影像的每一個像素點相減，再將其差值的平方加總起來取平均值，取得 MSE 值，其計算公式如方程式下所示。

$$MSE = \frac{1}{MN} \sum_{i=1}^m \sum_{j=1}^n [f(i,j) - f'(i,j)]^2$$

(二) 峰值信號雜訊比(Peak Signal to Noise Ratio, PSNR)

PSNR 也就是峰值訊噪比，經常用作圖像壓縮等領域中信號重建質量的測量方法，它是利用影像信號的最大值與影像中雜訊的比值作為評估的標準，其計算公式如下方程式，其中 $(2^n - 1)^2$ 代表是表示圖像點顏色的最大數值，如果每個像素點用 8 位表示，那麼就是 255。在利用 PSNR 算出來的比值越大，代表複製影像與輸入影像越接近，影像品質良好。

$$PSNR = 10 \log \frac{(2^n - 1)^2}{MSE}$$

(三) 結構相似性指標(Structure Similarity, SSIM)

結構相似性指標係用於測量兩張影像間相似性之方法。SSIM 指標係改進以前提出的通用性影像品質指標 (UQI) 模型，而可以被視為一個完美的影像品質量測。從圖像結構角度，以亮度、對比度及反映場景中物體結構的屬性因素，分別就其屬性之平均值、標準差及變異數作為衡量指標。其最佳及最差的量化值介於 1 至 100 之間，若 SSIM=100 表示與原始檔案完全一致，若 SSIM>=60 代表品質可以接受，若 SSIM<=60，代表品質不可以接受，肉眼即可察覺到明顯的畫面劣化，因此被判定為無實際觀賞價值。

此一指標模型可以透過比較誤差敏感性理論 (Error Sensitivity Philosophy) 而得到瞭解。首先，誤差敏感性接近估計察覺錯誤確定影像降低的數量，考慮影像降低而察覺在架構性訊息變化的變化。其次，錯誤差敏感性理論由下而上的方式，相關功能類似早期 HVS 系統。而此模型是由上至下，模仿 HVS 的功能性，避開在以前部分提及臨界值的問題，因為它不倚賴臨界值確定察覺的變形數量。並降低相互作用問題，因此結構相似性指標主要在評估兩張影像間的結構變化。

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

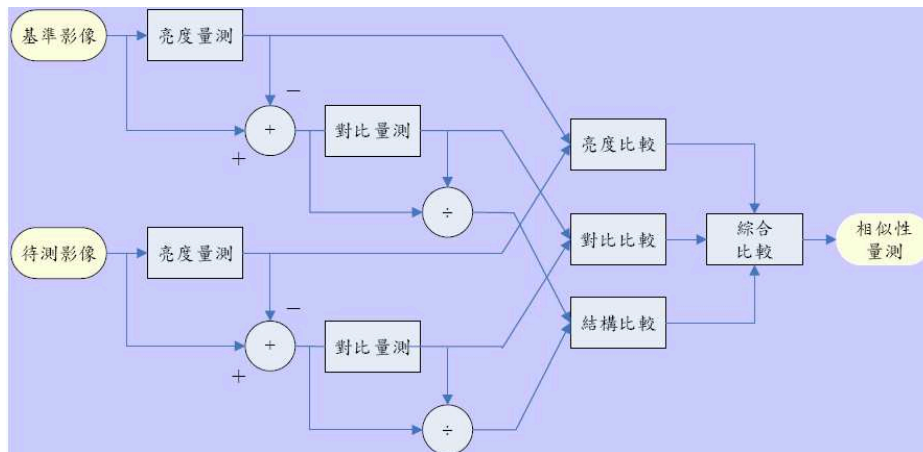


圖 12 結構相似性量測系統圖

傳統客觀品質指標是以統計的方法計算出有關於整張影像整體灰度值誤差的總和，而結構相似性指標不單僅是利用計算灰度值誤差的方法，並以模擬人眼視覺系統(HVS)，來進行影像品質評估結果，而由結構相似性指標計算結果卻可得到不同的指標值，並且得到符合人眼視覺的評估結果。

(四) 通用性影像品質指標(Universal Quality Index, UQI)

通用性影像品質指標，對於各類型影像處理均可適用，設計上是以任何影像失真模型均含有三種因素：相關係數降低、亮度改變及對比改變。

以相關係數降低、亮度改變及對比改變等因素來判定差異度。其最佳及最差的量化值介於 1 至 100 之間，若量化值 ≥ 60 ，代表品質可接受，若量化值 < 60 ，則代表品質不

可接受，因此可得到較具體的評估值。

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

(五) 基於人類視覺系統之峰值信號雜訊比(HVS- Peak Signal to Noise Ratio, HVS-PSNR)

HVS-PSNR 不同於傳統的 PSNR，因為 HVS-PSNR 加入了影像評估的前置處理，在計算每一個像素點的誤差值後，將其通過高斯低通濾波器，模擬人類視覺的低通特性，再計算影像失真的幅度，可以得到接近人眼所察覺到的影像評估值。HVS-PSNR 及 PSNR 一樣，評估值越大，代表轉置影像的影像品質越好。

三、視訊品質驗證

視訊品質驗證有六種視訊驗證方式，可利用視訊轉置前後的特徵、視覺心理學、線性色差、非線性色差、綜合轉置來作為驗證品質的方式。



圖 13 視訊品質驗證

(一) Full-Reference

一種參考原始與經處理過後視訊內容的視訊品質評估方法。這類的方法是假設在測量畫面品質的時候，同時也有原始視訊處理前的畫面可供參考。在這類方法中，主要偵測在時間上因 DCT 量化的變化結果造成的編碼誤差，同時也考慮的時間軸上的後遮蔽現象。此外，許多方法延伸了 SSIM 的做法，在其誤差模型中除了計算每張獨立畫面區塊中的平均值與變異數之外，同時也針對每個信號分別計算其 SSIM 的結果，之後再利用動量估測為結合的權重產生最後的結果。

(二) Reduced-Reference

一種參考部分原始與部分經處理過後視訊內容的視訊品質評估方法。有時候視訊品質的比較並沒有辦法完整得到原始視訊內容的資訊，例如在視訊串流系統中，我們不可能傳送完整的原始視訊內容給接收端要求其比較正在觀賞的視訊品質。因此，Reduced-Reference 的方法透過事先抽取部分原始視訊內容的特徵值，再將抽取出來的特徵值傳送給比較端後，比較端也同樣抽取相同的特徵值，再比較兩者特徵值得衰減情形比較視訊的品質的好壞。

(三) No-Reference

一種只參考經處理過後視訊內容的視訊品質評估方法。大部分的應用環境下，我們幾乎都沒有辦法得知原始視訊內容來比較視訊品質，例如觀看網路視訊的時候，我們不可能還有原始影片可以參考，或者是拿到一部 DVD 視訊時，我們也沒辦法得知其原始視訊內容，所以 No-Reference 利用結果影片中的誤差結果去推估視訊的品質好壞，不需要參考編碼端的原始影片，也因此 No-Reference 的方法最為廣泛應用在大部分的狀況下評估視訊品質。

四、 其他驗證

針對文字檔案類型格式轉置作業之品質驗證方式，係將文字類型之來源檔案及轉置後檔案，分別轉置為 JPEG 檔，再將兩個 JPEG 檔案進行影像結構相似性指標 (Structure Similarity, SSIM) 及通用性影像品質指標 (Universal Quality Index, UQI)，以檢測轉置前後的檔案品質。

檔案為文件類型 (例如：WDL 或 DOC) 時，則是使用其他驗證的功能來作驗證。

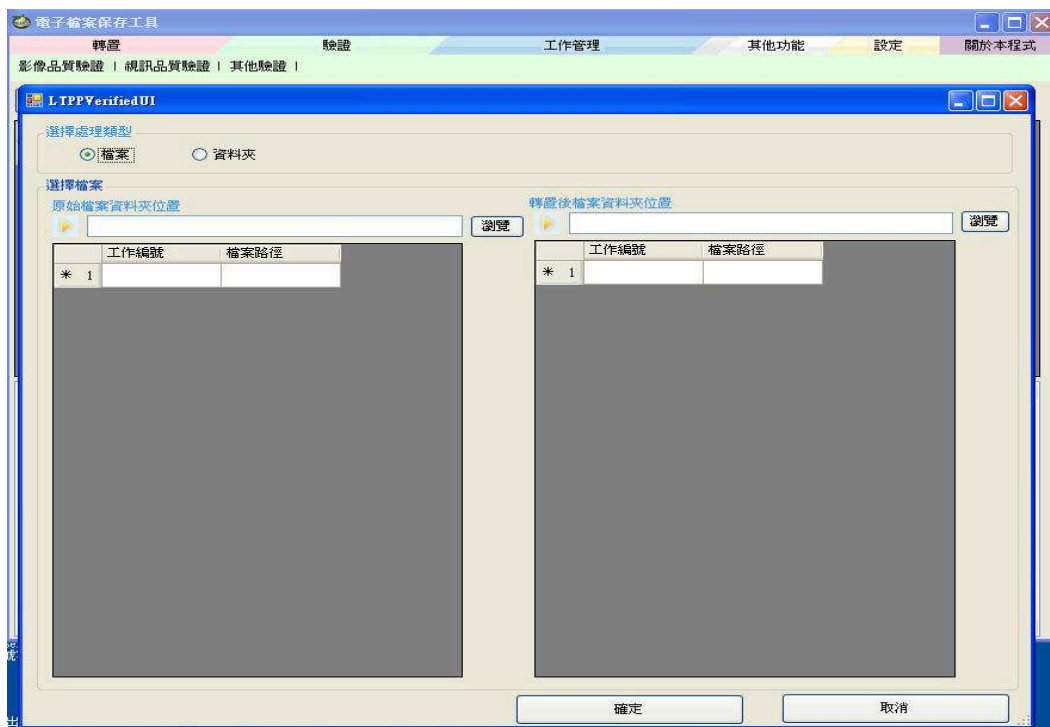


圖 14 其他驗證

陸、其他功能

一、影像辨識

本工具提供 2 種辨識功能分別為文字辨識(OCR)及條碼辨識，將文字儲存於文字檔中，以便使用者運用。「文字辨識(OCR)」是利用文字辨識軟體辨識及擷取 TIFF 格式圖片中的文字，「條碼辨識」是利用條碼辨識軟體辨識及擷取 TIFF 格式圖片中的條碼。

(一) 文字辨識(OCR)

「OCR」的原文為「Optical Character Recognition」，是指對文字資料的影像檔案進行分析處理，獲取文字及版面訊息的過程。主要功能是利用文字辨識(OCR)軟體辨識及擷取 TIFF 格式圖片中的文字，並將擷取內容另存於 TXT 純文字檔中，以便使用者可再次使用此文字內容。

文件或 word 文件掃描或轉成 PDF 格式的電子檔案，需要做文字內容編輯時，以往需要靠人力將內容重新逐一登打成文字才可將圖片上的文字轉化成純文字檔(.txt)，使用本工具之文字辨識(OCR)功能，可將此工作變得很簡單，輕而易舉的轉化成文字。

1. 內含文字影像檔。

	檔 號： /100602/
	保存年限：5年
裝	一、本局於98年11月11日召開「本局99年度委託研究及自行研究案研商會議」，並通過本組提列之99年度自行研究計畫「政府機關網站內容保存方式之研究」。
	二、惟因本局「99年度電子檔案生命週期管理機制委託服務案」已先行以本局網站試作網頁內容之保存，且考量網站內容保存事涉本會職掌之政府網站建置相關規範，刻值組織改造業務移交之際，為避免造成事權混淆，擬暫緩進行本項自行研究案。
	三、本案如奉 核可，擬請企劃組調整本局99年度自行研究項目，以上所擬，當否？
	敬請
	核示
訂	會辦單位：企劃組 承辦單位：檔案資訊組 後會：企劃組 決行

圖 15 內含文字影像檔

2. 文字辨識(OCR)轉置後

0990008395-P-0001.txt - 記事本	
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)	
檔號：/100602/ 保存年限：5年 一、本局於98年11月11日召開「本局99年度委託研究及自行研究案研商會議」，並通過本組提列之99年度自行研究計畫「政府機關網站內容保存方式之研究」 二、惟因本局「99年度電子檔案生命週期管理機制委託服務案」已先行以本局網站試作網頁內容之保存，且考量網站內容保存事涉本會職掌之政府網站建置相關規範，刻值組織改造業務移交之際，為避免造成事權混淆，擬暫緩進行本項自行研究案。 三、本案如奉核可，擬請企劃組調整本局99年度自行研究項目，以上所擬，當否？ 敬請 核示 會辦單位：企劃組 承辦單位復會 檔案資訊組企劃組 決行	

圖 16 OCR 轉置後

(二) 條碼辨識

主要功能是利用條碼辨識軟體辨識及擷取 TIFF 格式圖片中的條碼，並將擷取條碼內容另存於 TXT 純文字檔中，可直接辨識公文文號，以便使用者運用。

1. 條碼辨識前



圖 17 條碼辨識前

2. 條碼辨識後



圖 18 條碼辨識後

二、影像及視訊修補

本工具整合清華大學技術轉移的影像及視訊修補軟體，可提供老舊照片及老舊影片的修補。

(一) 老舊照片修補

早期的公文是以紙本的方式作保存，因時間及保存環境問題很可能造成紙張的損毀。電子檔案保存工具的老舊照片修補，可針對影像檔作修補的動作。

先將有部分損壞的文件先掃描成影像檔，再使用此功能將部分損壞的部分進行修補。雖修補的功效有限，但可維持檔案一定的完整度。

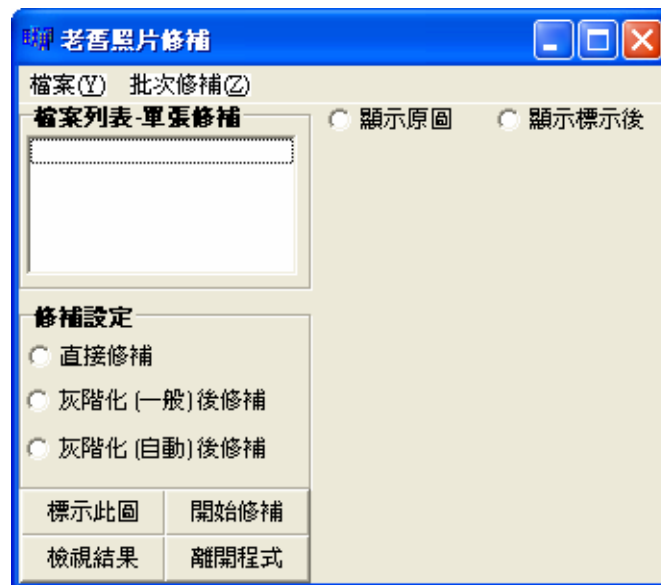


圖 19 老舊照片修補

1. 老舊照片修補前



圖 20 老舊照片修補前



圖 21 使用綠色筆刷標示欲修補的區域

2. 老舊照片修補後



圖 22 老舊照片修補後

(二) 老舊影片修補

視訊修補功能目前支援 WAV、AVI 及 MPEG-2 等類型的視訊檔，可針對受毀損影像進行修補工作，透過影片處理功能將影片分割為影格後，在影格中標示欲修補之區域進行修補，俟各原始影格均修補完成後，再透過影片處理功能將影格合併為影片，即完成影片修補。

1. 老舊影片修補前



圖 23 老舊影片修補前



圖 24 使用紅色筆刷標示欲修補的區域

2. 老舊影片修補後



圖 25 老舊影片修補後