



(19) 中華民國智慧財產局

(12) 發明說明書公開本

(11) 公開編號：TW 201535984 A

(43) 公開日：中華民國 104 (2015) 年 09 月 16 日

(21) 申請案號：104102959

(22) 申請日：中華民國 104 (2015) 年 01 月 29 日

(51) Int. Cl. : H03M7/30 (2006.01)

(30) 優先權：2014/01/29 美國 61/933,104

(71) 申請人：瑞里肯分析公司 (美國) RELICAN ANALYTICS, INC. (US)  
美國(72) 發明人：古德溫 安德魯 GOODWIN, ANDREW J. (US) ; 費雪 馬修 FISHER, MATTHEW  
P. (US)

(74) 代理人：黃志揚

申請實體審查：無 申請專利範圍項數：18 項 圖式數：3 共 54 頁

(54) 名稱

最佳化資料壓縮器與方法

OPTIMIZED DATA CONDENSER AND METHOD

(57) 摘要

本發明揭示一種提供將數字、字母、字、片語與其他標記無損壓縮成資料物件值以減少檔案大小之壓縮器與方法。該資料壓縮器與方法將資料歸類為個別資料物件或資料物件群組，且區別重複(例如循環)的詞條時。一參考庫係根據歸類的資料量而最佳化以減少儲存需求。該歸類的資料指定一獨特值，用以植入該參考檔案。一輸出檔案是由利用該參考庫以達成最佳無損壓縮的該資料壓縮器建立。一資料復原器與方法提供將諸如數字、字母、字、片語與其他標記的壓縮資料物件復原成未壓縮資料，以供有效率與正確使用而沒有資料物件損失。

A data condenser and method provides lossless condensation of numbers, letters, words, phrases, and other indicia to data object values which results in reduction of file size. The data condenser and method classifies data as individual data objects or groups of data objects and distinguishes terms which repeat (e.g. recur). A reference library is optimized according to the quantity of classified data to minimize storage requirements. The classified data is assigned a unique value which populates the reference file. An output file is created by the data condenser using the reference library to achieve optimal lossless condensation. A data reverter and method provides for reversion of condensed data objects such as numbers, letters, words, phrases and other indicia to uncondensed data objects for efficient and accurate use without loss of data objects.

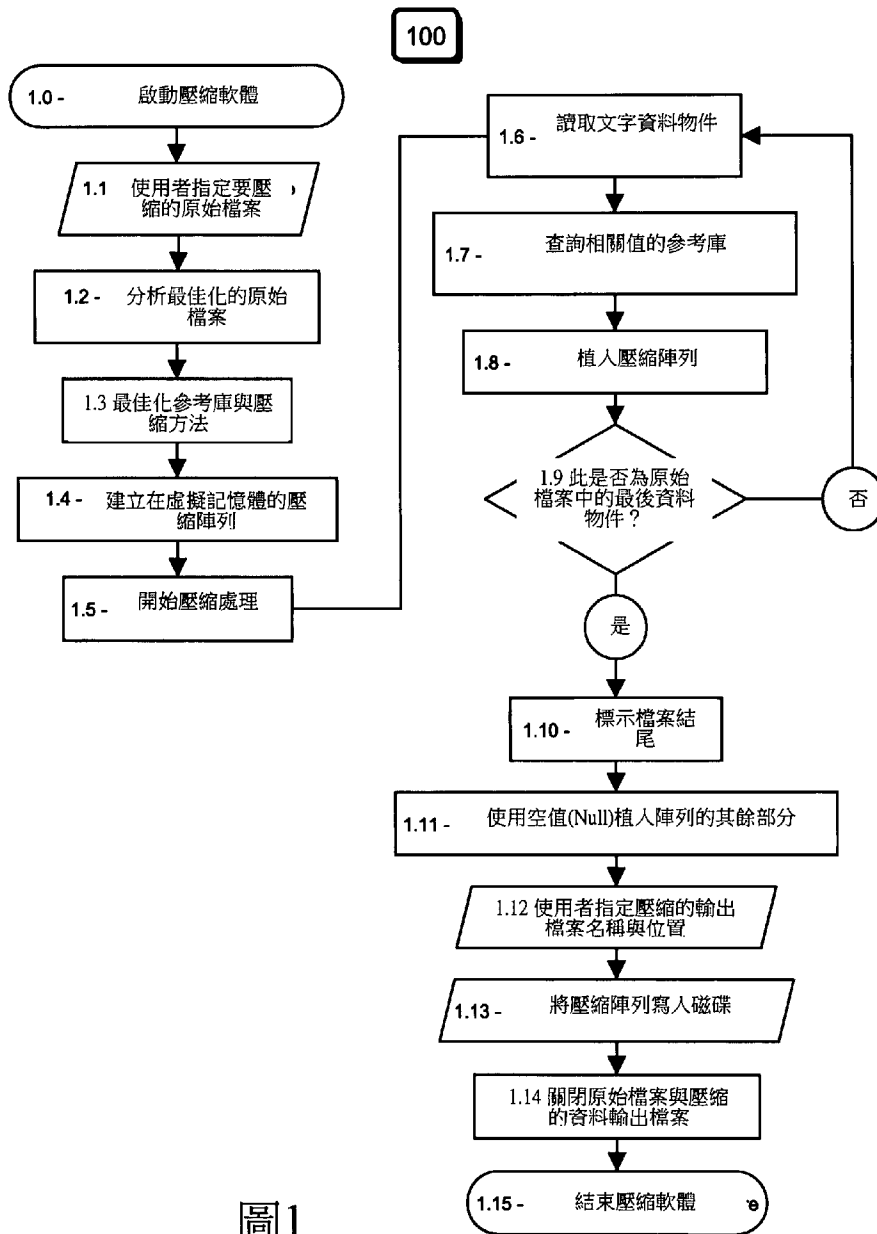


圖1

201535984

## 發明摘要

※ 申請案號：

104/102959

※ 申請日：

104.1.29

※IPC 分類：

H03M 7/30

(2006.01)

## 【發明名稱】(中文/英文)

最佳化資料壓縮器與方法 /OPTIMIZED DATA  
CONDENSER AND METHOD

## 【中文】

本發明揭示一種提供將數字、字母、字、片語與其他標記無損壓縮成資料物件值以減少檔案大小之壓縮器與方法。該資料壓縮器與方法將資料歸類為個別資料物件或資料物件群組，且區別重複(例如循環)的詞條時。一參考庫係根據歸類的資料量而最佳化以減少儲存需求。該歸類的資料指定一獨特值，用以植入該參考檔案。一輸出檔案是由利用該參考庫以達成最佳無損壓縮的該資料壓縮器建立。一資料復原器與方法提供將諸如數字、字母、字、片語與其他標記的壓縮資料物件復原成未壓縮資料，以供有效率與正確使用而沒有資料物件損失。

## 【英文】

A data condenser and method provides lossless condensation of numbers, letters, words, phrases, and other indicia to data object values which results in reduction of file size. The data condenser and method classifies data as individual data objects or groups of data objects and distinguishes terms which repeat (e.g. recur). A reference library is optimized according to the quantity of classified data to minimize storage requirements. The classified data is assigned a unique value which populates the reference file. An

output file is created by the data condenser using the reference library to achieve optimal lossless condensation. A data reverter and method provides for reversion of condensed data objects such as numbers, letters, words, phrases and other indicia to uncondensed data objects for efficient and accurate use without loss of data objects.

**【代表圖】**

**【本案指定代表圖】**：第（ 1 ）圖。

**【本代表圖之符號簡單說明】**：

100 壓縮方法

**【本案若有化學式時，請揭示最能顯示發明特徵的化學式】**：

# 發明專利說明書

(本說明書格式、順序，請勿任意更動)

## 【發明名稱】(中文/英文)

最佳化資料壓縮器與方法 /OPTIMIZED DATA  
CONDENSER AND METHOD

## 【技術領域】

本發明有關一種資料壓縮器與方法，尤指一種提供將數字、字母、字、片語與其他標記無損壓縮成資料物件值以減少檔案大小之壓縮器與方法。

## 【先前技術】

資料儲存是研究人員與其他人員面對解譯兆位元組、千兆位元組與更大範圍為單位的巨大基因體及其他資料集之令人畏怯性挑戰的中心議題。諸如研究與臨床設備(本地與遠端兩者)間資料的協作與重要分享需要處理極大資料集流且需要無損資料壓縮的策略，以能有效率與正確比較諸如記錄檔案或參考基因體，該等策略在稽核與報告目的方面是共同需要。資料儲存及其安全轉移已藉由傳送實體硬碟機至本地或遠端地點來達成，其通常是難以處理、昂貴與逐漸無法承受的。

隨手可得的研究型工具存在於有關大型資料檔案的選擇性壓縮、編碼、檢索、註解、映射、與比對之演算法與方法之領域的其他應用及出版，其包括(但未侷限於)：重複性序列群集、文字、二進位文字影像、與資料庫。

高產量下一代序列技術的快速進展產生正以指數率累積的基因體資料集(整個基因組與人口類型兩者)。公眾可用的基

因體資料集典型是利用純文字檔來儲存，增加對數位儲存、分析、與可靠傳輸的負擔。儲存、分享、分析、或下載大型資料檔案(諸如遺傳基因資訊或其他大型資料對於許多機構是麻煩且幾乎不可能，尤其在缺乏高速網際網路存取世界的這類部分。現階段，大型儲存地點將花費數百萬元在儲存設備，且大資料量傳輸對伺服器與網際電腦網造成負擔。對於沒有資料損失或惡化而能儲存、分享、傳輸大型資料集的解決方案存在著急迫性技術需要。對本技術而言，此充其量是麻煩，且在許多情況，對於諸如機構的許多實體是幾乎不可能，尤其是研究中心、醫院及其他。

#### 【發明內容】

一最佳化資料壓縮器與方法減少儲存媒體的負荷，諸如(但未侷限於)磁碟、固態、或任何形式的儲存媒體(固定或可攜式)、連同有關資料儲存的資料轉移與額外負荷，諸如源自下一代整個基因組序列。輸入檔案可包括(沒有侷限於)BLAST 格式、記錄/錯誤檔案、資料庫、與其他普通檔案格式。

最佳化資料壓縮器與方法將諸如數字、字母、字、片語與其他標記的資料(亦已知為資料物件)壓縮成一或多個值，該等值是利用手動定義及/或自動最佳化的十六進位、二五六進位(Base-256)或其他 NBase 值所定義。NBase(Nb)定義為一可變的數值基底系統，其具有二基底單位(二進位)的最小表示，且沒有上數值基底限制。如需要，此新的壓縮十六進位、Base-256 或其他 NBase 檔案的整個復原可被解碼以使資料返回最初格式，沒有任何資料損失。為了清楚，壓縮處理不是如在某些 ZIP 文檔所見的壓縮，而是導致壓縮的檔案大小的壓縮。

一參考庫包括一或多個參考檔案。每個參考檔案包括資料及/或資料群組，其結合用來與識別的資料物件形成關聯性的指定獨特十六進位、Base-256、或其他 NBase 碼。此識別的資料物件、以及不在參考庫中找到的複製(即是重複)與其他資料物件(其係指定一相關的獨特碼)係加入在參考庫內含的最初參考檔案的最佳化版本。

有關指定編碼之產生的十六進位、Base-256 或其他 NBase 值是以格式化參考檔案儲存，其中一資料物件是與一獨特十六進位、Base-256、或其他 NBase 值有關聯性，可經由一簡單的請求來設置(例如 X、Y 軸)。壓縮的檔案有能力採用水平與垂直兩者來擴展，不像需要增加頁以擴展的已定義且限制欄的一文字文件或資料庫。壓縮的檔案採用無損格式儲存。此允許在壓縮的檔案中將資料物件指定為十六進位、Base-256 或其他 NBase 值。特定的使用與一般的使用版本可藉由修改參考檔案來設計，同時核心引擎保持相同。基於檔案大小以選擇結構化格式可進一步最佳化檔案儲存與處理。

壓縮的檔案有效地隱藏嵌入資料物件的真實本質，由於用來代表檔案中資料物件的各種不同十六進位、Base-256、或其他 NBase 值不能夠容易地辨別，藉此提供自然狀態加密。即使透過使用電腦分析區別出十六進位、Base-256、或其他 NBase 值，但沒有方法告訴其代表哪些資料物件，或如果已應用替代的排列順序。雖然壓縮的檔案能以 X、Y 軸建立，且預設是從左至右植入，但沒有限制資料壓縮器與方法從中央開始且逆時

針植入或任何其他起始地方及/或植入方法。

在本發明的一態樣中，一種用於減少資料儲存與處理需求之無損資料壓縮器包括：一資料記憶體，其包含具有要壓縮的資料物件之一原始檔案；一預先決定的參考檔案，其具有預先決定的資料物件與預先決定的資料物件群組；一進階參考檔案，其具有演算法選定資料物件與演算法選定資料物件群組，其從該該原始檔案與該預先決定的參考檔案之至少一者取得；一壓縮檔案，其包含壓縮的資料物件與壓縮的資料物件群組；及一陣列檔案，其包含該壓縮資料物件與該壓縮資料物件群組的陣列；及一壓縮處理器，其操作用以至少部分基於該等資料物件的計數以選擇至少一數值基底編碼系統，至少部分基於該等資料物件的計數以選擇該陣列檔案的一壓縮深度，至少部分基於該等資料物件的計數以限制該至少一數值基底編碼系統，分析在該原始檔案中的該等資料物件以找出一或多個循環資料物件與循環資料物件群組，將一第一獨特碼指定給在該預先決定的參考檔案中的該等資料物件與資料物件群組，將一第二獨特碼指定給在進階參考檔案中的一選定循環與一選定非循環資料物件與資料物件群組，將該等編碼資料物件與該等編碼資料物件群組置放在該壓縮檔案，且在該陣列檔案中建立該等壓縮資料物件的陣列，其中該無損資料壓縮器允許有效率與正確無損資料儲存與轉移，其中該壓縮處理器具有指定給預先決定的資料物件與資料物件群組的一 Base-256 編碼系統；一 Base-16 編碼系統，其指定給非預先決定的資料物件與資料物件群組；及該記憶體陣列，其是由編碼的資料物件與資料物件群組之總數的平方根定義，其係四捨五入(如需要)成一整數，

且空值(Null)資料物件填滿該陣列(如需要)。

在本發明的另一態樣中，一種用於執行無損壓縮運算之裝置包括：一資料源記憶體，其具有包含資料物件的輸入資料之至少一資料檔案；一壓縮處理器，其構成從一單資料物件區別資料物件群組，且區別重複超過一選定臨界值之一重複資料物件群組；及一參考庫，其包含一自訂檔案，該自訂檔案包含預先定義的資料物件與預先決定的資料物件群組，其中該壓縮處理器進一步構成將一獨特碼指定給該重覆個體與一非重覆資料物件與資料物件群組，以提供一壓縮的檔案。

在本發明的另一態樣中，一種用於減少資料儲存與處理需求之無損資料壓縮器包括：一資料源記憶體，其具有包含資料物件的輸入資料之至少一資料檔案；一壓縮處理器；一壓縮記憶體，其具有資料物件群組，其每一者根據在該資料源記憶體中的資料物件之重複模式以包含一或多個資料物件；一壓縮處理器，其操作用以基於在該資料源記憶體中的資料物件與資料物件群組數目以選擇複數個數值基底編碼系統之至少一者；一自訂參考庫記憶體，其具有預先定義的資料物件與先定義的資料物件群組，其指定來自該等選定數值基底編碼系統之至少一者的獨特碼；及一格式化參考庫記憶體，其具有循環與非循環資料物件與資料物件群組，其指定來自該等數值基底編碼系統、連同該等選定編碼預先定義的資料物件群組之至少一者的獨特碼；及該格式化參考庫記憶體，其具有由編碼資料物件與資料物件群組的總數之平方根所定義的一記憶體陣列，其係四捨五入(如需要)成一整數，且空值(Null)資料物件填滿該陣列

(如需要)，且該記憶體陣列壓縮深度係基於在該格式化參考庫中的編碼資料物件與資料物件群組之總數的計數。

在本發明的另一態樣中，一種用於減少資料儲存與處理需求之無損資料壓縮器包括：一資料記憶體，其具有至少一原始檔案，用以儲存輸入資料，包含要壓縮的複數個循環與非循環資料物件；一介面記憶體，其操作用以接收及讀取來自輸入資料的資料物件；一參考庫，用以儲存有關對應資料物件的編碼值；一壓縮處理器，其連接該參考庫以最佳化該輸入資料；及一壓縮記憶體，其連接用以接收最佳化的輸入資料且提供壓縮的輸出資料。

在本發明的另一態樣中，一種無損資料壓縮系統包括：一壓縮處理器，其構成基於資料物件群組的重複模式數目而從複數個基底編碼系統選擇至少一數值基底系統，且將來自該等選定數值基底編碼系統之至少一者的獨特碼指定給每個資料物件群組；及一壓縮記憶體，其具有編碼的資料物件群組以提供壓縮的資料儲存與資料轉移。

在本發明的另一態樣中，一無損資料壓縮系統包括壓縮處理器，其構成將獨特碼指定給含有資料物件的重複模式之資料物件群組；及一記憶體陣列，用以儲存具有一壓縮深度之編碼資料物件群組，該壓縮深度係選定以減少該記憶陣列的儲存需求。

在本發明的另一態樣中，一種用以將壓縮的資料物件復原

成未壓縮形式之裝置包括：一資料記憶體，其包含具有壓縮資料物件之一壓縮檔案；用於未壓縮資料物件至少一目的模式；一參考檔案，其具有編碼的循環與非循環壓縮資料物件、與識別至少一或多個數值編碼系統的資訊、與有關該壓縮檔案的壓縮深度(如適用)；及一復原處理器，其操作用以選擇要復原的該等壓縮資料物件之至少一部分；選擇用於未壓縮資料物件的目的模式；及連結該壓縮檔案與該參考檔案，以將該等壓縮資料物件復原成未壓縮的資料物件，以供有效率與正確使用而沒有資料物件損失。

在本發明的另一態樣中，一種用於無損資料壓縮諸如數字、字母、字、片語與其他標記的資料物件之方法，以允許有效率與正確可靠資料評估、資料儲存與資料轉移，該方法包括：取得要壓縮的資料物件集；將該資料物件集置放在一原始檔案；進行該原始檔案的分析，其包含計數資料物件的總數；讀取該原始檔案且建立資料物件群組，其每一者包含一或多個資料物件，以回應該原始檔案內之資料物件的重複模式；計數資料物件群組的總數；基於資料物件群組的選定量以選擇一或多個數值編碼系統；建立預先定義的資料物件群組之一或多個庫參考檔案；將來自該選定編碼系統的一第一獨特碼指定給每個預先定義的資料物件群組；識別資料物件的個體循環群組，以從非循環資料物件群組區別出；將來自該選定編碼系統之一第二獨特碼指定給每個循環與非循環資料物件群組，其不是預先定義的資料物件群組；將編碼資料物件群組置放在一或多個庫參考檔案內；讀取該原始檔案且計數編碼資料物件群組的總數；選擇記憶體陣列的壓縮深度；形成由選定編碼資料物件群

組的總數之平方根所定義的記憶體陣列且四捨五入(如需要)以建立一整數；增加空值編碼資料物件以填滿該陣列；及將該陣列置放在儲存媒體上。

在本發明的另一態樣中，一種壓縮資料以提供減小資料儲存與處理之方法包括：讀取一資料檔案，其包含複數個資料物件；識別循環資料物件群組，其每一者包含一或多個資料物件，以從非循環資料物件群組區別出；將一程式碼指定給每個循環與非循環資料物件群組；格式化最佳大小的參考庫；及將該等編碼的資料物件群組置放在用於資料儲存與處理的該參考庫。

在本發明的另一態樣中，一種用於無損資料復原諸如數字、字母、字、片語與其他標記的壓縮資料物件之方法包括：取得壓縮資料物件的壓縮檔案；選擇要復原的壓縮資料物件之至少一部分；選擇復原資料物件的目的模式；取得一參考檔案，其具有指定給循環與非循環壓縮資料物件的程式碼，該參考檔案具有識別至少一或多個數值編碼系統及有關該壓縮檔案的壓縮深度(如適用)的資訊；及基於對參考檔案的連結來讀取壓縮檔案，以將該等壓縮資料物件復原成未壓縮的資料物件，以供有效率與正確使用而沒有損失資料物件。

#### **【圖式簡單說明】**

圖 1 為本發明的資料檔案壓縮方法之一具體實施例的流程圖。

圖 2 為本發明之用以將一壓縮檔案壓縮成其最初未壓縮狀態的方法之一具體實施例的流程圖。

圖 3 為本發明之一最佳化資料壓縮器系統之示范性具體實施例的系統圖。

### 【實施方式】

在本說明書中，特定用詞是為了簡潔、清楚與理解而使用。沒有不必要的限制，遠超過先前技術的需求，因為此用詞只為描述目的與寬義解釋。在本說明書描述的不同系統與方法可單獨或結合其他系統與方法使用。各種不同等效物、替代選擇與修改是可能，且在文後申請專利範圍的範疇內。文後申請專利範圍中的每個限制為援用美國專利法（35 U.S.C.）第 112 條第 6 項的詮釋，只有術語「構件」或「步驟」為明示地詳述在個別限制。

在下列實施方式中，將參考形成其部分的附圖，且其經由可實施的說明特殊具體實施例顯示。這些具體實施例經充份詳細描述，以使熟諳此技者能夠實施具體實施例，且應瞭解，其他具體實施例可利用，且可達成邏輯、機械、電氣及其他變化，不致悖離具體實施例的範疇。因此，下列實施方式不認為是限制本發明的範疇。

本發明的系統與方法利用無損資料壓縮(造成減少檔案大小)將單一數字、字母、字、片語與其他標記壓縮成十六進位、Base-256、或其他 NBase 值，其為形成一或多個壓縮檔案的不可思議序列技術。本發明的系統與方法利用無損壓縮檔案保有其從其讀取及寫入的能力。本發明的系統與方法利用無損壓縮檔案復原成其先前最初未壓縮的狀態，供使用在無法與壓縮檔案(或陣列檔案)互動的應用方面。因此，本發明的系統與方法

利用裝置與系統：(a)明顯減少儲存需求，特別是針對大型資料檔案；(b)允許較快資料傳輸，特別是大型文字或資料庫檔案；(c)允許包含其他資料(例如，電子醫療紀錄、臨床詮釋資料(Metadata)、相關稽核記錄、等等)，具最小的檔案大小增加；(d)允許快速分析及修改壓縮的檔案；(e)提供自然狀態加密；及/或(f)作為進一步開發以能夠實施以 GPU 為主的處理與數學/科學分析的平台。較小的壓縮資料檔案大小允許大型檔案時常儲存在特殊化系統，且現在更廣泛儲存在較小的容量系統，且亦減少通訊中的網路流量或需求頻寬。

術語「資料物件」按其廣義來說可用來包括電子信號所攜帶文數字(即是數位)值所代表的最小單組件。每個資料物件指定一或多個十六進位、Base-256 或其他 NBase(即是數位)值。

術語「壓縮深度」用來定義 Base-16、Base-256 或其他 NBase 資料物件數目，其經壓縮(儲存)在陣列檔案中。壓縮深度必須為一正整數+i 值，且系統可在轉換成一壓縮形式時，將其設定成大於或等於 1( $\geq 1$ )的預設值，或手動設定成大於或等於 1( $\geq 1$ )的值。

在本說明書描述的最佳化壓縮器的技術效益或技術功效在於，壓縮的檔案大小是非常小於在目前壓縮系統中所達成最初檔案大小的部分，且操作明顯不同。某些明顯實例在於此不僅有用，且實際必須包括(但未侷限於)遺傳學。

大多數的目前電腦採用二進位數字(即是「位元」)，亦已

知為「Base-2」。某些電腦使用 8 位元集(即是「位元組」)運算，其每個具有 256 個值。當使用 ADCII(統一碼)格式時，介於 0 至 127 值通常有關特殊字元，同時，介於 128 至 255 值通常有關特殊詞條，諸如來自外國語言的重音字元。Windows 視窗系統中的「記事本」係每字元或空間使用一位元組的記憶體。在本說明書揭露的最佳化資料壓縮器與方法可使用各種不同數值基底系統，諸如十六進位、Base-256 或其他 NBase，其可手動或自動選擇供最佳化，且未侷限於其內含實例中提供的 Base-16(如前述十六進位)。

如果採用二進位 Base-2 系統，每個資料物件可包括用以定義值之一位元或複數個位元，如數個位元或 8 位元集情況(即為 1 位元組)表示。十六進位是 Base-16 系統，且每個單獨十六進位值等於 4 位元的二進位碼。例如(且未侷限於)：

1 個十六進位值 = 16 個可能的資料物件或 4 位元

2 個十六進位值 = 256 個可能的資料物件或 8 位元/1 位元組

3 個十六進位值 = 4096 個可能的資料物件或 12 位元

4 個十六進位值 = 65536 個可能的資料物件或 16 位元/2

位元組

5 個十六進位值 = 1048576 個可能的資料物件或 20 位元

6 個十六進位值 = 16777218 個可能的資料物件或 24 位元

/3 位元組

7 個十六進位值 = 268435456 個可能的資料物件或 28 位元

8 個十六進位值 = 4294967296 個可能的資料物件或 32 位

元/4 位元組

採用十六進位空間的四個一組的任一位元(1/4)來表示同樣的二進位資料物件數目。

在另一具體實施例中，如果採用二進位 Base-2 系統，每個資料物件可包括用以定義值的 1 位元或複數個位元，如數個位元或 8 位元集情況（即為 1 位元組）表示。Duocentosexapentagesimal 是 Base-256 系統且每個單獨值等於 8 位元的二進位碼。例如且(未侷限於)：

1 個二五六進位值 = 256 個可能的資料物件或 8 位元/1 個位元組

2 個二五六進位值 = 65536 個可能的資料物件或 16 位元/2 位元組

3 個二五六進位值 = 16777216 個可能的資料物件或 24 位元/3 位元組

4 個二五六進位值 = 4294967296 個可能的資料物件或 32 位元/4 位元組

在另一具體實施例中，如果採用二進位 Base-2 系統，每個資料物件可包括用以定義值的 1 位元或複數個位元，如數個位元或在 8 位元集情況(即 1 位元組)表示。NBase(Base-N)是一可變基底系統(Base-N)，其中 N 定義為大於或等於 1( $\geq 1$ )的任何正整數(+i)，且每個單獨值為大於或等於二進位碼 1( $\geq 1$ )位元的一正整數(+i)。資料物件可藉由壓縮深度加以壓縮，其中壓縮深度為大於或等於 1( $\geq 1$ )的一正整數(+i)。

1 NBase 值  $\geq +i$  個資料物件，且  $\geq +i$  Packdepth 中的  $+i$  個位元

圖 1 和 2 為分別描述用以將資料檔案壓縮成實質較小的大小之壓縮方法(100)的示范性具體實施例、與用以將該較小的大小檔案復原成沒有資料損失(即是無損)的最初資料檔案大小之復原方法(200)的流程圖。在圖 3 說明的最佳化資料壓縮器系統(300)通常為一計算系統，其可包括(但未侷限於)桌上型電腦、手持式裝置、多處理器系統、個人數位助理(PDA, Personal Digital Assistant)、膝上型電腦、網路電腦、雲端伺服器、迷你電腦、主機電腦、等等。壓縮器系統(300)包括介面(301)；一或多個處理器(302)；及一記憶體(303)，其耦接處理器(302)。處理器(302)從記憶體(303)載入及執行軟體。當壓縮器系統(300)執行軟體時，模組(304)指使處理(302)操作，如在本說明書描述的壓縮方法(100)或復原方法(200)。

介面(301)可包括滑鼠；鍵盤；語音輸入裝置；觸控輸入裝置，用於從使用者接收手勢；運動輸入裝置，用於偵測非觸控手勢與使用者的其他運動；及其他相容輸入裝置與相關處理元件，其能夠從使用者接收使用者輸入。輸出裝置(諸如，一視訊顯示器或圖形顯示器)可顯示進一步有關如在本說明書揭露的系統與方法之具體實施例的界面。喇叭、印表機、觸覺回饋裝置及其他類型輸出裝置亦可包括在使用者介面(301)。此外，介面(301)可使系統(300)溝通其他裝置，諸如網頁伺服器與外部資料庫。介面(301)可促成各種不同網路與協定類型內的多重通

訊，包括有線網路(例如，區域網路(LAN, Local Area Network)、電纜、等等)與無線網路(諸如無線 LAN(WLAN, Wireless LAN)、蜂巢式網路或衛星。為了此目的，介面(301)可包括一或多個連接埠，用於彼此連接某些計算系統或連接另一伺服器電腦。

處理器(302)可為一單處理單元或某些單單元，其全部可包括多重計算單元。處理器(302)可以採用一或多個微處理器、微電腦、微控制器、數位信號處理器、中央處理單元、狀態機器、邏輯電路、及/或基於操作指令以操縱信號的任何裝置加以實施。除了其他能力，處理器(302)構成取得及執行在記憶體(303)中儲存的電腦可讀指令與資料。

記憶體(303)可包括在技術中已知的任何電腦可讀媒體，其包括(例如)揮發性記憶，諸如，靜態隨意存取記憶體(SRAM, Static Random Access Memory)與動態隨機存取記憶體(DRAM, Dynamic Random Access Memory)；及/或非揮發性記憶體，諸如，唯讀記憶體(ROM, Read Only Memory)、可抹消可編程 ROM、快閃記憶體、硬碟、光碟與磁帶。記憶體(303)亦包括模組(304)與資料(305)。

除其他以外，模組(304)包括副程式、程式、物件、組件、資料結構、等等，其可執行特定工作、或實施特定資料類型。模組(304)更包括介面模組(306)、壓縮模組(307)、參考庫自訂模組(308)、視覺檢視器模組(309)、復原模組(310)、與可能其他模組(如需要)。其他模組可包括程式，其為補充系統(300)的應用程式，例如，作業系統程式。另一方面，除其他以外，資料(305)

記憶體當作儲存庫使用，用以儲存經由該等模組(304)之一或多者處理、接收、及產生的資料。資料(305)記憶體包括輸入資料(311)，該輸入資料可包括一原始檔案；參考庫(312)資料，其是由參考檔案(314)組成，諸如一預先決定的參考檔案及/或一進階參考檔案；及輸出資料(313)，諸如在壓縮檔案及/或陣列檔案內。輸出資料(313)包括執行該等前述模組之一或多者的結果所產生的資料。

針對壓縮，系統(300)係與輸入資料(311)有關聯。輸入資料(311)諸如原始檔案)可包括(但未侷限於)文字、影像、試算表或資料庫格式，且可利用介面模組(306)選擇，以及經由壓縮模組(307)提供給系統(300)供壓縮。壓縮模組(307)將可選擇性啟動參考庫自訂模組(308)以分析輸入資料(311)且最佳化該參考庫(312)。壓縮模組(307)然後使用參考庫(312)將輸入資料(311)壓縮成輸出資料(313)。介面模組(306)傳回當作資料(305)的壓縮檔案(或陣列檔案)的輸出資料(313)，其可在記憶體(303)中從固定或可攜式儲存媒體取得。

針對復原，系統(300)係與輸入資料(311)有關聯，諸如一壓縮檔案。在此情況的輸入資料是在如前面詳述處理中產生的一壓縮檔案(或陣列檔案)。輸入資料(311)是由介面模組(306)選擇，且經由復原模組(310)提供給系統(300)供復原。介面模組(306)將提供一目的模式選項以復原在檔案或螢幕。如果選擇檔案，介面模組(306)將啟動復原模組(310)。復原模組(310)將使用來自參考庫(312)的一參考檔案，將輸入資料(311)復原成輸出資料(313)。介面模組(306)傳回當作資料(305)檔案的輸出資料

(313)，其可包括(但未侷限於)字、影像、試算表或資料庫格式，且可在記憶體(303)中從固定或可攜式儲存媒體取得。如果選擇螢幕，介面模組(306)將啟動復原模組(310)。復原模組(310)將使用來自參考庫(312)的一參考檔案，將輸入資料(311)復原成輸出資料(313)。介面模組(306)將輸出資料(313)傳回給視覺檢視器模組(309)，其能以視覺未壓縮可視格式，利用介面(301)模組將輸出資料(313)顯示在電腦監視器或其他裝置。

連同壓縮模組(307)，處理器(302)讀取在原始檔案中的全部格式的資料，諸如字元、數字、空間、等等，且從其他類型資料(例如字元、數字、等等)區別空白。連同壓縮模組(307)，處理器(302)然後決定在整個原始檔案內重複(例如循環)資料物件的次數。

基於此資料，處理器(302)編輯原始檔案或產生一進階參考檔案，且將此參考檔案加入參考庫。循環一次以上的資料物件將會加入進階參考檔案。使用者可手動定義或由壓縮模組(307)自動決定要增加以致資料物件必須出現的次數(臨界)。

例如，當使用 ASCII(統一碼)格式時，資料壓縮器可決定超過普通 128 個值用於特定字元，20 個詞條具有重複 20 次的 5 個字元，15 個詞條具有重複 15 次的 10 個字元，且 5 個詞條具有重複 5 次的 50 個字元。在此處理期間，資料壓縮器決定有總數 40 個內部結構符合超過 128 個 ASCII 格式普通特殊字元的臨界需求。

下列提供利用十六進位方法具體實施例而於資料檔案中發現的 5、10 和 50 個字元之最佳化的實例：

5 個詞條資料物件(20)，其中「r」代表重複次數。

詞條 01 - r150	詞條 02 - r100	詞條 03 - r50	詞條 04 - r80
詞條 05 - r150	詞條 06 - r120	詞條 07 - r170	詞條 08 - r90
詞條 09 - r 70	詞條 10 - r140	詞條 11 - r130	詞條 12 - r50
詞條 13 - r90	詞條 14 - r 70	詞條 15 - r210	詞條 16 - r180
詞條 17 - r160	詞條 18 - r210	詞條 19 - r330	詞條 20 - r 60

每個詞條是以 1：1 的詞條資料物件寫入參考庫(312)一次，且將代表超過臨界計數的循環詞條。每個詞條資料物件將需要一識別十六進位資料物件加上一指標十六進位資料物件。此將需要(20x5)+20 個十六進位資料物件，或總數 120 個十六進位資料物件。

10 個字元詞條(15)，其中「r」代表重複次數。

詞條 01 - r150	詞條 02 - r100	詞條 03 - r50
詞條 04 - r80	詞條 05 - r150	詞條 06 - r120
詞條 07 - r170	詞條 08 - r90	詞條 09 - r70
詞條 10 - r140	詞條 11 - r130	詞條 12 - r50

詞條 13 - r90      詞條 14 - r70      詞條 15 - r210

每個詞條是以 1：1 的詞條資料物件寫入參考庫(312)一次，且將代表超過臨界計數的循環詞條。每個詞條將需要一識別十六進位資料物件加上一指標十六進位資料物件。此將需要  $(15 \times 10) + 15$  個十六進位資料物件，或總數 165 個十六進位資料物件。

50 個字元詞條(5)，其中「r」代表重複次數。

詞條 01 - r 70      詞條 02 - r50      詞條 01 - r90      詞條  
01 - r80      詞條 01 - r 60

每個詞條是以 1：1 的詞條資料物件寫入參考庫(312)一次，且將代表超過臨界計數的循環詞條。每個詞條將需要一識別十六進位資料物件加上一指標十六進位資料物件。此將需要  $(5 \times 50) + 5$  (位元)十六進位資料物件、或總數 255 個十六進位資料物件。

對於循環最佳化侷限於如前述的 5、10 和 50 個字元的詞條，寫入參考庫(312)的整個內容於供應給虛擬記憶體字首檔案的總計 540 個十六進位資料物件是  $120 + 165 + 255$ 。此說明的循環處理如此使用在壓縮資料內含 540 個十六進位資料物件來取代 52,500 個字元。

在不是使用臨界最佳化(諸如藉由將循環處理限制在 5、10 和 50 個字元的詞條)的情況，資料壓縮器與方法可感測兩或多

個字元的所有循環，其達成明顯減少所需使用十六進位(或其他 NBase)資料物件總數，且仍提供整個檔案的無損壓縮。

每個識別、非重複單字元或字元群組係與單十六進位(或其他 NBase)資料物件有關聯。在拼錯字的情況，系統會檢查主旨字的字元直到無法再繼續為止；然後將單字元壓縮成十六進位(或其他 NBase)資料物件，而不是壓縮整個字。此允許補償內部資料物件錯誤與不完全，而沒有軟體缺陷或錯誤輸出。

在完成循環處理，資料壓縮器與方法執行參考庫(312)所內含壓縮資料的第二計數，其通常為由於循環而從初始計數明顯減少的數字。

一最佳化參考庫可使用供儲存透過循環所識別實體的壓縮檔案，且其可最佳化以減少用於儲存壓縮資料的容量需求。如果使用一平方壓縮陣列檔案(例如)，資料壓縮器與方法決定壓縮資料之總計數的平方根。如果平方根包括一數字分數，平方根數字將調整到下一更高整數以避免資料損失。例如，456,840 的平方根是 675.8 x 675.8。不過，資料壓縮器與方法係四捨五入成 676 x 676(經過四捨五入的下一最接近整數值)。陣列檔案其後植入來自壓縮檔案的所有十六進位、Base-256 或其他 NBase 資料物件，且在檔案結尾前，其餘空間填滿非值(例如，空值(Null))資料物件。

在十六進位具體實施例中，資料壓縮器與方法提供(但未侷限於)由個別十六進位值數目所指定的某些預先決定的十六

進位資料物件。為了簡單描述「每個十六進位物件(HPO, Hex Per Object)」，可基於壓縮的資料以自動及/或手動選擇此格式，以進一步最佳化資料處理與儲存。

十六進位資料物件實例	相對的十進位	每個十六進位物件	最大值
B	(十進位 11)	1 hpo	16
1C	(十進位 28)	2 hpo	256
F7A	(十進位 3962)	3 hpo	4,096
4AF6	(十進位 19190)	4 hpo	65,536
FFA75	(十進位 1047157)	5 hpo	1,048,576
A7B9E5	(十進位 10992101)	6 hpo	16,777,216
…等等			

資料壓縮器與方法將如此藉由選擇進一步最佳化的結構化格式以壓縮儲存需求。例如，利用透過 6 hpo 格式提供的 16,777,216 值可能遠無法有效率地最佳化。如果假設英文語言具有約 130,000 個普通字，包括時態變化版本(ing、ed、est … 等等)，使用 4 hpo 格式(65,536 值)是太小。然而，資料壓縮器與方法可將 5 hpo 格式 1,048,576 值選擇為最佳化格式。如另一實例所示，外顯子/內含子(Exon/Intron)基因體序列在每 1,000,000 個字元約 1,000,000 個位元的資料質量平衡，可擴及約 45,000,000 至 260,000,000 個字元長度。在基因體序列中，只有八個可能字元，加上一空值/非讀取(N)、與檔案結尾。不需要使用 2 hpo、4 hpo、或 6 hpo 格式。然而，提供 16 個值的 1 hpo 格式是極其充份。在基因體序列中只有總數八個字元，高度循環模式會發生，且最佳化因此可決定及直接提供給預先決定的參考檔案(314)或參考庫(312)。染色體(16)(例如)包括約

88,827,255 遺傳因子基對。若採用文字檔存放，可能需要約 88,830,625 個位元，包括相關的檔案資訊。當藉由資料壓縮器與方法壓縮時，檔案大小會自動調整到沒有基對損失的  $9425 \times 9424$ 。選擇 1 hpo 格式造成檔案減少到 2,683,293 個位元，此約為最初大小的 3%。

資料壓縮器與方法可提供含有預先定義的資料物件之一或多個預先決定的參考檔案。資料壓縮器與方法使在每個原始檔案中的識別個別資料物件與在預先決定的參考檔案中的預先決定的資料物件產生關聯性，藉此形成一進階參考檔案。可使用加入新壓縮檔案的此新資料物件，將位於進階參考檔案的新資料物件指定獨特的十六進位(或其他 NBase)值。在任何情況，使用資料壓縮器與方法將數位指標加入壓縮檔案中的每個資料物件。例如，一預先決定的參考檔案可包括 130,000 個普通英文字，其使用提供 1,048,576 個值的 5 hpo 格式。另一預先決定的參考檔案可包括完整染色體的外顯子/內含子基因體序列，其使用 1 hpo 格式以提供 16 個值。另一預先決定的參考檔案可包括一 FastQ 多變格式，其使用 3 hpo 格式提供 4096 個值。另一預先決定的參考檔案可包括普通語言，其使用全球每個普通語言的變體。5 hpo 或其以上格式可對預先決定的參考檔案提供最佳化。換行控制字元、空值區域(諸如檔案結尾或列結尾、細胞符號標記、欄、列、定位鍵、等等)可內建在共同及/或 FastQ 庫。

預先決定的參考檔案可結構化，以使多重資料物件合併，以便能由一單十六進位(或 NBase)資料物件指定。因此，可合

併來自相關的兩(或多個)但不是相同滑尺量度的值。當分析基因體資料時，例如，一基對「A(腺嘌呤)」可結合在特定連結與位置的允許誤差範圍內之「#」的 Phred 分數，以建立能由一單十六進位(或 NBase)資料物件識別的一值。

資料壓縮器與方法將一獨特十六進位(或 NBase)物件碼指定給該等壓縮檔案之每一者(例如預先決定及/或進階的參考檔案)。如一實例所示，「O say can you see by the dawn's early light」資料可如下列所示來指定十六進位資料物件碼值：

讀	碼(2 hpo)	寫
O	F8	0,0
Space	49	0,1
s	16	0,2
a	04	0,3
y	1C	0,4
Space	49	0,5
c	06	0,6
a	04	0,7
n	11	0,8
Space	49	0,9
y	1C	0,10
o	12	0,11
u	18	0,12
Space	49	0,13
s	16	0,14
e	08	0,15
e	08	0,16
Space	49	0,17
b	05	0,18
y	1C	0,19
Space	49	0,20
t	17	0,21
h	0B	0,22
e	08	0,23
Space	49	0,24
d	07	0,25
a	04	0,26
w	1A	0,27
n	11	0,28
'	44	0,29
s	16	0,30
Space	49	0,31

讀	碼(2 hpo)	寫
e	08	0,32
a	04	0,33
r	15	0,34
l	0F	0,35
y	1C	0,36
Space	49	0,37
l	0F	0,38
i	0C	0,39
g	0A	0,40
h	0B	0,41
t	17	0,42
,	45	1,0...

資料壓縮器與方法透過將獨特十六進位、Base-256、或其他 NBase 值指定給在參考庫中儲存的壓縮資料(例如壓縮檔案、陣列檔案、等等)以提供固有的加密。接收者必須知道此十六進位、Base-256、或其他 NBase 值以使用壓縮的資料檔案。

資料壓縮器與方法可藉由建立壓縮資料檔案內含資料的交替起始點、位置變化、錯誤資料物件、空白、與非線性寫入模式以提高加密。

此外，資料壓縮器與方法可利用一習知的加密技術，允許選擇目前日期、目前時間、一關鍵字或其他一般文字/數值輸入來建立一次性密鑰以加密資料記錄。當然，資料的接收者亦必須具有對密鑰的存取以開啟及利用資料。資料壓縮器與方法亦允許使用者建立其自己的演算法或提供附加的安全性。例如，一壓縮的資料檔案可使用另一壓縮或標準的資料檔案進行加密。

資料壓縮器與方法將運作在一習知的電腦處理單元

(CPU, Computer Processing Unit), 但亦可運作於其他處理單元, 諸如(例如)一圖形處理單元(GPU, Graphical Processing Unit)。

資料壓縮器與方法提供的固有或特別加密將支援標準的進階先出(FIFO, First-in-first-out), 但不必然使用此格式來操作。其可從檔案的任意區域開始, 且其後採用非線性方式在不同座標間跳躍。

在另一具體實施例中, 壓縮的資料檔案(例如, 壓縮檔案、陣列檔案、等等)可從 FastQ 文字檔案的內容直接值入。對於每個 FastQ 資料封包, 在第 1 列每個單字元(例如, 數字或字母)會先被編碼, 接著是列結尾符號。每個 FastQ 資料封包的第 2 列與第 4 列組合產生雙字母碼壓縮成十六進位(或 NBase)資料物件, 接著列結尾(EoL)符號, 同時忽略第 3 列。當完成壓縮所有 FastQ 資料封包時, 加入檔案結尾符號。其次, 處理器(302)將每個單字母或雙字母碼轉換成十六進位(或 NBase)資料物件, 使得每個壓縮的資料檔案元件構成單十六進位(或 NBase)資料物件。壓縮資料檔案的其餘部分填入無值(例如, 空值(Null))十六進位(或 NBase)資料物件直到檔案結尾符(EoF)。

在復原處理中, 壓縮檔案(或陣列檔案)中的十六進位(或 NBase)資料物件會被復原成原始檔案中的最初資料。在前述的替代具體實施例中, 處理器(302)用來將每個十六進位(或 NBase)資料物件轉換成單或雙字母未壓縮資料物件。每個復原的 FastQ 資料封包目前會是兩列。對於每個 FastQ 資料封包: (a) 所有十六進位(或 NBase)資料物件會被讀取直到列結尾為止; (b)

每個十六進位(或 NBase)資料物件會被轉換成一單字母，且使用文字植入資料檔案；(c)文字 FastQ 封包列 1 會被寫入一未壓縮的資料物件；(d)所有十六進位(或 NBase)資料物件會被讀取直到列結尾(EoL)為止；(e)FastQ 封包列 2 會被讀入一未壓縮的資料檔案(每個元件將有兩文字字元)；(f)每個未壓縮資料檔案元件的第一字元會被寫入一文字列(FastQ 封包列 2)；(g)具有 + 字元的一文字列(FastQ 封包列 3)會被寫入；及(h)每個壓縮資料檔案元件的第二字元會被寫入一文字列(FastQ 封包列 4)。檔案結尾(EoF)符號與所有無值十六進位資料物件會被忽略。產生的文字檔案不再被壓縮且可被儲存。

資料壓縮器與方法可採用具有不同數字系統的電腦，諸如(但未侷限於)八進位(Base-8)、十進位(Base-10)、十四進位(Base-14)、十五進位(Base-15)、十六進位(Base-16)、二四進位(Base-24)、三二進位(Base-32)、三六進位(Base-36)、六四進位(Base-64)與二五六進位(Base-256)。

某些 Base-2 電腦使用 Base-16(十六進位)碼來表示電腦記憶體位址。此十六進位記號法提供二進位碼值的友善化表示。此十六進位使用十六個劃分區域符號，在多數情況，符號 0-9 表示零至九的數值，且 A、B、C、D、E 和 F(或者，a-f)分別代表十至十五的數值。每個十六進位數字代表半位元組，其為一位元組(8 位元)的一半。位元組值範圍從 0 至 255(十進位)，但可更方便表示為範圍從 00 至 FF 的兩個十六進位數字。

在 Base-256 具體實施例中，一具有 256 個值之預先決定(例

如，預先定義)的參考檔案是用來轉換及壓縮可能在最初資料原始檔案中發生的資料物件。在此具體實施例中，所有普通字元、文數字與符號連同控制字元全包括在內，諸如(但未侷限於)換行鍵、定位鍵、檔案結尾與空值(Null)。然後加入長詞條，其是由英文語言中最普遍使用的多重字元字定義。預先決定(例如預先定義)的參考檔案可構成用於任何語言及/或主題，且未侷限於英語文字。

下列實例示意說明使用 Base-256 預先決定的參考檔案。

供轉換的原始檔案：

-- The big brown rabbit likes green leafy carrots; he eats them  
all in the garden. --

下列某些字是普通字，且已在預先決定的參考檔案中指定一 Base-256 值。

The = 130 Likes = 131 he = 132 them = 133 all = 134  
green = 135 brown = 136 big = 137

預先決定的參考檔案亦包括所有普通英語文數字字元與符號、連同控制字元。

資料壓縮器然後將最初原始檔案轉換成壓縮的資料物件。最初原始檔案包括 81 個資料物件(文數字字元、符號、控制字元)。壓縮最初資料物件將可利用下列方法來減少數目。

‘the’ 的兩使用 = 從總數 6 個資料物件減少成 2 個

‘big’ = 從 3 個減少到 1 個

‘brown’ = 從 5 個減少到 1 個

‘likes’ = 從 5 個減少到 1 個

‘green’ = 從 5 個減少到 1 個

‘he’ = 從 2 個減少到 1 個

‘them’ = 從 4 個減少到 1 個

‘all’ = 從 3 個減少到 1 個

總減少的資料物件(6+3+5+5+5+2+4+3)33，產生 9 個壓縮的資料物件。

(81 個最初資料物件，包括列結尾與檔案結尾 - 33)+9 個結果壓縮的資料物件。

壓縮的資料物件總數 = 57 個總資料物件。

陣列檔案然後使用總數 57 個資料物件計數定義，而不是 81 個資料物，平方根可決定且四捨五入成最近整數，且植入陣列檔案，如前述段落的定義。

利用 NBase 具體實施例來壓縮係使用可變方法完成，依單資料物件或許多資料物件循環(long read)的最佳化選擇可連同針對單最初原始檔案或許多最初檔案的最佳化方法來選擇。

當建立最佳化參考庫時，short read NBase 只利用單文數字字元、符號、與控制字元。

在下列實例中，short read NBase 將用來最佳化一單最初原始檔案。

供轉換的最初原始檔案：

```
-- AACCTGGTccccTTTCCGGAAAANNNNNCCccttgaaaa --
```

short read 最佳化檢查檔案中的資料物件且決定是否有包括列結尾與檔案結尾變數的總數 42 個資料物件。NBase 然後決定使用的資料物件是否為 A、C、T、G、a、c、t、g、N、列結尾、與檔案結尾。在最初原始檔案中有總數 11 個獨特資料物件，因此 NBase 壓縮檔案將定義為 Base-11 集，且只需使用 11 個獨特資料物件。陣列檔案係如前段落描述而建立。

long read NBase 亦讀取使用的所有單使用文數字字元、符號、與控制字元，且然後讀取所有字及/或片語，以尋找填滿於壓縮檔案中剩餘空間的多數循環。一旦完成，新的壓縮檔案用來決定用於建立在前段落描述陣列檔案的正式資料物件總數。

NBase 多重檔案最佳化係同前段落的描述，唯一需要注意的是，其係讀取一個以上的檔案來決定最佳化的壓縮檔案，而不是讀取一單最初原始檔案來定義最佳化的壓縮檔案。

在 Base-256(或 NBase)具體實施例中，壓縮深度是 Base-256(及/或 NBase)指定值的數目，其壓縮成由前段落描述

的方法所寫入的一壓縮陣列檔案。壓縮深度必須是一正整數  $+i$ ，其大於或等於  $1(\geq 1)$ 。

在下面的實例中，[DO] 中的詞條將會指定為 Base-256(或 NBase)資料物件的實例，其係與來自一預先決定的參考檔案的值有關聯。引用在()內的詞條將認為保持在由壓縮深度所定義的一單壓縮資料物件中。EoL 將保有列結尾的資料物件參考，且 EoF 將保有檔案結尾的資料物件參考。

壓縮深度 1：

([DO])([DO])([DO])([DO])([DO])([DO])([DO])([DO])([DO])([DO])([DO])([DO])([EoL])([EoF])

在前述實例中，具有供壓縮含有 EoL 和 EoF 的 13 個資料物件。因此，陣列檔案將定義為 13 的平方根，使得一陣列檔案將採用前段描述的方法來建立。

壓縮深度 2：

([DO][DO])([DO][DO])([DO][DO])([DO][DO])([DO][DO])([DO][DO])([DO][EoL])([EoF][NULL])

在前述實例中，具有供壓縮含有 EoL 和 EoF 的 7 個資料物件。因此，陣列檔案將定義為 7 的平方根，使得一陣列檔案將採用前段描述的方法來建立。

壓縮深度 3：

```
((DO)(DO)(DO))((DO)(DO)(DO))((DO)(DO)(DO))((DO)(DO)(EoL))((EoF)(NULL)(NULL))
```

在前述實例中，具有供壓縮含有 EoL 和 EoF 的 5 個資料物件。因此，陣列檔案將定義為 5 的平方根，使得一陣列檔案將採用前段描述的方法來建立。

壓縮深度 4：

```
((DO)(DO)(DO)(DO))((DO)(DO)(DO)(DO))((DO)(DO)(DO)(EoL))((EoF)(NULL)(NULL)(NULL))
```

在前述實例中，具有供壓縮含有 EoL 和 EoF 的 4 個資料物件。因此，陣列檔案將定義為 4 的平方根，使得一陣列檔案將採用前段描述的方法來建立。

壓縮深度可預先設定或手動定義在前述兩具體實施例中。對於所有意圖與目的，前述實例的壓縮深度可設定成 13 且陣列檔案將會是一(1)壓縮的資料物件，其大小為在該單陣列檔案內部儲存的 13 個 Base-256(或 NBase)值。

圖 1 示意說明一種壓縮資料檔案之壓縮方法(100)。如在應用的不同地方描述，系統與方法可利用執行於儲存媒體中儲存的電腦程式碼之一處理器達成。

處理器可包括一微處理器；及其他電路，其取回及執行來自儲存系統的軟體。處理器可實施於一單處理裝置內，但亦可

分散於協同執行程式指令的多重處理裝置或子系統。處理器的實例包括通用目的之中央處理單元、應用特殊處理器、與邏輯裝置、以及任何其他類型的處理裝置、處理裝置的組合、或其變化。

儲存裝置可包括處理器可讀且可儲存軟體的任何儲存媒體。儲存系統可包括能以任何方法或技術實施之用於儲存資訊的揮發性與非揮發性、可移除與非可移除媒體，諸如電腦可讀指令、資料結構、程式模組、或其他資料。儲存系統能以一單儲存裝置實施，但亦可實施於多重儲存裝置或子系統。儲存系統可更包括附加元件，諸如能夠溝通處理器的一控制器。

儲存媒體的實例包括一隨機存取記憶體、唯讀記憶體、磁碟、光碟、快閃記憶體、虛擬記憶體、與非虛擬記憶體、磁組、磁帶、磁碟儲存器或其他磁儲存裝置、或能用來儲存想要的資訊且可由指令執行系統存取的任何其他媒體、以及其任何組合或變化、或任何其他類型的儲存媒體。

步驟 1.0 包括啟動壓縮軟體。壓縮原始檔案的程式碼可隨著其所在的不同作業系統平台而執行、運行、及/或啟動。執行時，壓縮軟體會在壓縮處理過程將適當組件載入主動記憶體 (RAM, Random Access Memory)，且將輔助組件載入固定或可攜式儲存媒體以供必要時使用。

步驟 1.1 包括選擇要壓縮的原始檔案，以下視為原始檔案。透過壓縮軟體功能選單以選擇原始檔案。此可包括(但未

侷限於)瀏覽功能選單對話方塊盒、托曳檔案到壓縮軟體、及/或開啟與選擇檔案選項。此載入要壓縮的檔案且為了壓縮而建立對原始檔案的連結。

步驟 1.2 包括分析原始檔案。原始檔案分析可包括下列各項之一或多著，且未侷限於所列出的實例：

- 計數存在於原始檔案的所有資料物件
- 評估內部格式(例如試算表、FastQ 檔案、等等)
- 循環資料物件識別(例如定義群組)
- 預先決定/繼承的資料物件識別(例如，從預先決定的參考檔案)
- 預先決定/手動輸入的演算法(例如，數值基底編碼系統)

當藉由預設或手動決定的方法而符合所有標準時，整個這些資料物件(例如，一或多個 ASCII 文字字元)將會被設定且已知，以允許處理前往下一步驟進行壓縮。

步驟 1.3 開始預先決定的參考檔案與壓縮檔案的選擇性最佳化。基於原始檔案分析，選擇性存在以藉由使十六進位、Base-256、或其他 NBase 值與資料物件產生關聯性、與選擇陣列檔案的壓縮深度而不是預設值 3，以自訂預先決定的參考檔案(與壓縮檔案)。最佳化壓縮的自訂方法可手動指定或自動選擇。

步驟 1.4 有關建立陣列檔案以保持壓縮的資料。全部已知

的資料物件是一整數。資料物件的平方根係經由導出。此用來決定陣列檔案的 X 和 Y 軸的參數。由於多數的平方根包括使用小數位，故總壓縮資料物件的導出平方根是經過四捨五入成最近的整數，使得陣列檔案始終足夠大以處理原始檔案的壓縮，但對於小於資料物件的一額外列而言，從不會浪費超過額外需求的空間。假設資料物件的平方根係四捨五入成最近的整數，壓縮的陣列檔案針對 X 和 Y 軸兩者係定義為整數值。此準備用於形成陣列處理的虛擬記憶體。

步驟 1.5 開始壓縮處理。開始原始檔案壓縮是從處理確認開始。

步驟 1.6 有關將估計的資料物件讀入原始檔案供處理。讀取壓縮資料物件能以一或多個方法發生，但未侷限於以下列出的實例：

- 一次一資料物件
- 基於步驟 1.3 執行的最佳化分析
- 預先決定/繼承的結構

資料物件會被讀入用於壓縮處理的主動記憶體。

步驟 1.7 有關處理原始檔案，以取得針對估計的資料物件及/或資料物件群組所定義的十六進位、Base-256 或其他 NBase 值。主動記憶體中的資料物件是與預先決定的參考檔案相比較，該預先決定的參考檔案保有針對最初原始檔案所內含資料物件的指定十六進位、Base-256 或其他 NBase 值指派。

步驟 1.8 有關使用編碼的資料植入壓縮檔案。寫入十六進位、Base-256 或其他 NBase 值能以改變的方法寫入壓縮檔案，但未侷限於以下列出的實例：

- 左至右，上至下(預設值)
- 從中心逆時針
- 自訂設計與指定的模式；及/或
- 碎形方法(Fractal Methodology)
- 資料流
- 基於步驟 1.3 執行的最佳化分析

步驟 1.9 有關轉換週期。除非原始檔案沒有可讀取的額外資料物件，否則將持續植入壓縮檔案。如果存在額外資料物件，將持續植入壓縮檔案(參見步驟 1.6)，且處理再次執行步驟 1.7 和 1.8。如果沒有額外資料物件，處理將前往步驟 1.10，終止重複處理。當此發生時，處理轉移至檔案完成程序。

步驟 1.10 有關標示壓縮檔案結尾。十六進位、Base-256 或其他 NBase 值指定成檔案結尾符號的結尾。此十六進位、Base-256 或其他 NBase 值從不與任何其他十六進位、Base-256 或其他 NBase 值指派一起共用，但能以針對各類不同理由所指定的實際指定十六進位、Base-256 或其他 NBase 值指派來改變。此十六進位、Base-256 或其他 NBase 值在一資料檔案中只使用一次。

步驟 1.11 有關針對空值(Null)而使用十六進位、Base-256

或其他 NBase 值植入在壓縮檔案中的剩餘檔案空間。空值(Null)亦已知為空白空間且類似檔案結尾，然而，因為檔案結尾只可當作一結束符號使用一次；存在虛擬記憶體中的任何剩餘空間必須按照壓縮檔案的結構需求來植入。十六進位、Base-256 或其他 NBase 值指定成空值(Null)。此值從不與其他十六進位、Base-256 或其他 NBase 值指派一起共用，但能以針對各類不同理由所指定的實際指定十六進位、Base-256 或其他 NBase 值指派來改變。

步驟 1.12 有關命名壓縮檔案。壓縮輸出檔案可儲存在具有壓縮檔案所需足夠寫入空間的任何固定或可攜式儲存媒體。此時可決定壓縮檔案的名稱。此命名沒有特別需求且可特別或按照外部決定的標準來決定。

步驟 1.13 將壓縮檔案寫入供儲存的磁碟。在主動記憶體的壓縮檔案寫入固定或可攜式儲存媒體的儲存區，且現在稱為壓縮的資料輸出檔案。壓縮檔案可視需要從固定或可攜式儲存媒體存取，無需修改其新的壓縮狀態。

步驟 1.14 關閉原始檔案與壓縮檔案。從主動記憶體釋放原始檔案與壓縮檔案。

步驟 1.15 關閉壓縮軟體。從主動記憶體釋放壓縮軟體。刪除在軟體執行時使用的任何暫時資料檔案。

圖 2 示意說明將壓縮的資料轉變成最初未壓縮狀態的復原

方法(200)。

步驟 2.0 有關啟動復原程式。復原壓縮檔案的程式碼係基於存在的作業系統平台來執行、運行、及/或開始。如此，復原軟體在復原處理期間視需要將適當的組件載入主動記憶體 (RAM)，且將輔助組件載入固定或可攜式儲存媒體。復原使用應考慮對於快速檢視在虛擬環境的某些壓縮檔案、或復原整個原始檔案的需要，以供傳輸給沒有不明確壓縮資料檔案技術可用的系統或實體。

步驟 2.1 有關透過復原軟體功能選單以選擇壓縮檔案(例如，壓縮的檔案或復原檔案)。透過復原軟體功能選單以選擇壓縮檔案。此可包括(但未侷限於)瀏覽功能選單對話方塊盒、拖曳壓縮檔案至復原軟體、及/或開啟與選擇壓縮檔案選項。此載入要復原的壓縮檔案，且針對復原處理以建立對壓縮檔案的連結。

步驟 2.2 有關要復原的壓縮資料檔案量。使用者可指定小於 100%要復原壓縮檔案的特定量。此數量可為選定的百分比、列號或其他未列出的方法。

步驟 2.3 有關選擇輸出方法。使用者可指定輸出資料檔案(例如目的或虛擬檔案)的輸出、或視覺媒體的檢視視窗。輸出資料檔案可儲存在輸出資料檔案所需足夠寫入空間的任何固定或可攜式媒體。此時可決定輸出資料檔案的名稱。命名沒有特定的需求且可特別或依據外部決定的標準來決定。

步驟 2.4 有關讀取壓縮檔案。壓縮檔案是使用下列方法之一者來讀取，但未侷限於以下列出的實例。

- 左至右、上至下(預設值)
- 從中心逆時針
- 演算法指定模式
- 碎片方法(Fractal Methodology)
- 如參考庫所決定
- 資料流

步驟 2.5 有關從壓縮檔案讀取十六進位、Base-256 或其他 NBase 值。從壓縮檔案讀取十六進位、Base-256 或其他 NBase 值。

步驟 2.6 有關參考進階參考檔案與在壓縮檔案中所指定的壓縮深度。十六進位、Base-256 或其他 NBase 值是從參考庫取得。參考庫利用有關最初原始檔案的指定資料物件來連結十六進位、Base-256 或其他 NBase 值。

步驟 2.7 有關將復原的資料物件寫入在記憶體中的一目的資料檔案。復原的資料物件然後採用下列方法之一或多者來植入一目的檔案，但未侷限於以下列出的實例。

- 復原預先決定/手動輸入的演算法
  - 識別循環結構實體
  - 特殊格式
  - 如參考庫所決定

- 資料流

目的檔案存在主動記憶體(RAM)。植入目的檔案將會持續直到在壓縮檔案中發生檔案結尾(EoF)十六進位、Base-256 或其他 NBase 值(1.11)為止。目的檔案將採用與先前壓縮的最初原始檔案相同的方法來格式化，此最初原始檔案可為下列類型之一的檔案，但未侷限於以下列出的實例：

- Word、Excel、文字檔案
- 資料庫檔案
- 電子病歷卡
- 記錄檔案
- FASTQ/基因體檔案
- 二進位檔案

步驟 2.8 有關復原週期。一寫入週期定義為復原軟體必須將資訊寫入主動記憶體、或固定或可攜式儲存媒體的一單實例。如果系統限制調適將在步驟 2.2 的使用者指定量、或整個目的檔案寫入主動記憶體(RAM)，復原軟體將認識到只需要一單寫入週期，且植入目的檔案將持續直到壓縮檔案沒有可讀取的額外十六進位、Base-256 或其他 NBase 值為止。如果系統限制不會調適將在步驟 2.2 的使用者指定量或整個目的檔案寫入主動記憶體(RAM)，復原程式將計算大於 1 的寫入週期量，其將允許植入部分的目的檔案。決定應計算多大的寫入週期可為(但未侷限於)下列方法之一：

- 利用在壓縮資料檔案中的列結尾
- 利用壓縮檔案的分數或百分比

- 使用可用隨機存取記憶體(RAM)的分數或百分比

步驟 2.9 有關決定復原軟體是否已到達其最終寫入週期的結束。如果已到達最終寫入週期的結束，復原軟體前往步驟 2.10。如果未到達最終寫入週期的結束，復原軟體前往步驟 2.9.5。

步驟 2.9.5 有關將目的檔案寫入在固定或可攜式儲存媒體中的一暫時資料檔案。暫時資料檔案特徵係遵循如步驟 2.7 所述目的檔案的相同規格與限制。

步驟 2.10 有關較早選擇資料檔案輸出或檢視器輸出。(步驟 2.3)該處理是在輸出給資料檔案或檢視器間的此點處分歧。

步驟 2.11 有關完成將目的檔案寫入固定或可攜式儲存媒體。涉及復原處理的暫時資料檔案與在主動記憶體中的目的檔案會被組合且寫入固定或可攜式儲存媒體。目的檔案目前是以資料檔案存在，其可視需要從固定或可攜式儲存媒體存取。

步驟 2.12 關閉原始與輸出資料檔案。從主動記憶體釋放原始與輸出資料檔案。從固定或可攜式儲存媒體刪除涉及復原處理的暫時資料檔案。

步驟 2.13 有關如果已選定視覺檢視(步驟 2.3)則建立檢視視窗。在主動記憶體中建立能夠檢視目的檔案的處理。

步驟 2.14 有關將復原的目的檔案輸出給在步驟 2.13 建立的檢視器視窗。為了在檢視器視窗中檢視，組合及準備涉及復原處理的暫時資料檔案與主動記憶體中的目的檔案。此經傳輸顯示在視覺媒介體供檢驗、診斷/檢閱內容。

步驟 2.15 關閉原始目的檔案與檢視器視窗。從主動記憶體釋放目的檔案與檢視視窗。從固定或可攜式儲存媒體刪除涉及復原處理的暫時資料檔案。

步驟 2.16 關閉復原軟體。從主動記憶體釋放復原軟體。刪除使用於軟體執行的任何暫時檔案。

在前述中，某些詞條是為了簡潔、清楚與瞭解而使用。其意謂超越先前技術的需求而沒有不必要的限制，因為這類詞條用於描述目的且寬廣解釋。本說明書描述的不同結構、系統、與方法步驟可單獨或結合其他結構、系統與方法步驟一起使用。可預期地係，可有各種不同的等效物、替代選擇與修改，且在文後申請專例範圍的範疇內。

#### 【符號說明】

100	壓縮方法
200	復原方法
300	系統
301	介面
302	處理器
303	記憶體
304	模組

- 305 資料
- 306 介面模組
- 307 壓縮模組
- 308 參考庫自訂模組
- 309 檢視器模組
- 310 復原模組
- 311 輸入資料
- 312 參考庫
- 313 輸出資料
- 314 參考檔案

**【生物材料寄存】**

國內寄存資訊【請依寄存機構、日期、號碼順序註記】

國外寄存資訊【請依寄存國家、機構、日期、號碼順序註記】

**【序列表】** (請換頁單獨記載)

# 申請專利範圍

## 【第1項】

一種用於減少資料儲存與處理需求之無損資料壓縮器，其包括：

一資料記憶體，其包含：

一原始檔案，其具有要壓縮的資料物件；

一預先決定的參考檔案，其具有預先決定的資料物件與預先決定的資料物件群組；

一進階參考檔案，其具有演算法選定資料物件與算術選定資料物件群組，其從該原始檔案與該預先決定的參考檔案之至少一者取得；

一壓縮檔案，其包含壓縮的資料物件與壓縮的資料物件群組；及

一陣列檔案，其包含該等壓縮資料物件與該等壓縮資料物件群組的陣列；及

一壓縮處理器，其實施用以：

至少部分基於該等資料物件之計數以選擇至少一數值基底編碼系統；

至少部分基於該等資料物件之計數以選擇該陣列檔案的壓縮深度；

至少部分基於該等資料物件之計數以限制該至少一數值基底編碼系統；

分析在該原始檔案中的資料物件，以找出一或多個循環資料物件與循環資料物件群組；

將一第一獨特碼指定給在該預先決定的參考檔案中的該等資料物件與資料物件群組；

將一第二獨特碼指定給在該進階參考檔案中的一選定循環與一選定非循環資料物件與資料物件群組；

將該等編碼資料物件與該等編碼的資料物件群組置放在該壓縮檔案；及

在該陣列檔案中建立該等壓縮資料物件的陣列，

其中該無損資料壓縮器允許有效率與正確無損資料儲存與轉移。

**【第 2 項】**

如請求項 1 所述之無損資料壓縮器，其中該壓縮處理器具有一 Base-256 編碼系統，其指定給預先決定的資料物件與資料物件群組；及一 Base-16 編碼系統，其指定給非預先決定的資料物件與資料物件群組。

**【第 3 項】**

如請求項 1 所述之無損資料壓縮器，其中該記憶體陣列是由編碼資料物件與資料物件群組的總數之平方根加以定義，其係四捨五入(如需要)成一整數，且空值資料物件填滿該陣列(如需要)。

**【第 4 項】**

一種用於執行無損壓縮運算之裝置，其包括：

一資料源記憶體，其具有包含資料物件的輸入資料之至少一資料檔案；

一壓縮處理器，其構成用以從一單資料物件區別資料物件的群組，且用以區別一重複的資料物件群組，其重複超過一選定臨界值；及

一參考庫，其包括一自訂的檔案，其包含預先定義的資料物件與預先決定的資料物件群組，

其中該壓縮處理器更構成用以將一獨特碼指定給該重覆與非重覆資料物件與資料物件群組，以提供一壓縮的檔案。

**【第5項】**

如請求項 4 所述之裝置，其中該壓縮處理器將一 Base-256 編碼系統指定給預先定義的資料物件與資料物件群組，且將一 Base-16 編碼系統指定給非預先定義的資料物件與資料物件群組。

**【第6項】**

如請求項 4 所述之無損資料壓縮器，其中該記憶體陣列是由編碼的資料物件與編碼的資料物件之總數的平方根加以定義，其係四捨五入(如需要)成一整數，且空值資料物件填滿該陣列(如需要)。

**【第7項】**

一種用於減少資料儲存與處理需求之無損資料壓縮器，其包括：

一資料源記憶體，其具有包含資料物件的輸入資料之至少一資料檔案；

一壓縮記憶體，其具有資料物件群組，其每一者包含根據該資料源記憶體中資料物件的重複模式之一或多個資料物件；

一壓縮處理器，其操作用以基於在該資料源記憶體中的資料物件與資料物件群組數目，以選擇複數個數值基底編碼系統之至少一者；

一自訂參考庫記憶體，其具有預先定義的資料物件與預先定義的資料物件群組，其指定來自該等選定數值基底編碼系統之至少一者的一獨特碼；及

一格式化參考庫記憶體，其具有循環與非循環資料物件與

資料物件群組，其指定來自該等數值基底編碼系統之至少一者的一獨特碼；連同選定編碼預先定義的資料物件群組；及

該格式化參考庫記憶體具有由編碼的資料物件與資料物件群組之總數的平分根所定義的一記憶體陣列，其係四捨五入(如需要)成一整數，具空值資料物件填滿該陣列(如需要)，且該記憶體陣列壓縮深度係基於在該格式化的參考庫中的編碼資料物件與資料物件群組之總數的計數。

#### 【第 8 項】

一種用於減少資料儲存與處理需求之無損資料壓縮器，其包括：

一資料記憶體，其具有用以儲存輸入資料之至少一原始檔案，該輸入資料包含要壓縮的複數個循環與非循環資料物件；

一介面記憶體，其操作用以接收及讀取來自輸入資料的資料物件；

一參考庫，其儲存有關該等對應資料物件的編碼值；

一壓縮處理器，其連接該參考庫以最佳化該輸入資料；及

一壓縮記憶體，其連接用以接收最佳化輸入資料且提供壓縮的輸出資料。

#### 【第 9 項】

如請求項 8 所述之無損資料壓縮器，其中該壓縮處理器係基於最佳的壓縮檔案大小以指定一數值基底編碼系統。

#### 【第 10 項】

如請求項 9 所述之無損資料壓縮器，其中該壓縮處理器將一 Base-256 編碼系統指定給該參考庫資料物件，且將一 Base-16 編碼系統指定給資料記憶體資料物件。

#### 【第 11 項】

如請求項 9 所述之無損資料壓縮器，其中該壓縮處理器係至少部分基於資料物件群組的重複模式數目以指定一數值基底編碼系統。

**【第 12 項】**

如請求項 8 所述之無損資料壓縮器，其中該壓縮記憶體具有一選定用以減少儲存需求的壓縮深度。

**【第 13 項】**

一種無損資料壓縮系統，其包括：

一壓縮處理器，其構成基於資料物件群組的重複模式數目，從複數個數值基底編碼系統選擇至少一數值基底編碼系統，且將來自該等選定數值基底編碼系統之至少一者的一獨特碼指定給每個資料物件群組；及

一壓縮記憶體，其具有編碼資料物件群組，以提供壓縮的資料儲存與資料轉移。

**【第 14 項】**

一種無損資料壓縮系統，其包括：

一壓縮處理器，其構成將獨特碼指定給含有資料物件的重複模式之資料物件群組；及

一記憶體陣列，用以儲存具有一壓縮深度之編碼資料物件群組，該壓縮深度係選定以減少該記憶陣列的儲存需求。

**【第 15 項】**

一種用以將壓縮的資料物件復原成未壓縮形式之裝置，其包括：

一資料記憶體，其包含：

一壓縮檔案，其具有未壓縮的資料物件；

未壓縮的資料物件之至少一目的模式；

一參考檔案，其具有編碼的循環與非循環壓縮資料物件、與識別至少一或多個數值編碼系統的資訊、及有關該壓縮檔案的壓縮深度(如適用)；及

一復原處理器，其操作用以：

選擇用以復原的該壓縮資料物件之至少一部分；

選擇用於未壓縮資料物件之目的模式；及

連結該壓縮檔案與該參考檔案，以將該等壓縮資料物件復原成未壓縮資料物件，以供有效率與正確使用而沒有損失該等資料物件。

#### 【第 16 項】

一種用於諸如數字、字母、字、片語及其他標記之資料物件的無損資料壓縮之方法，以允許有效率及正確可靠資料評估、資料儲存與資料轉移，其包括：

取得要壓縮之資料物件的集合；

將整個資料物件置放在一原始檔案；

進行該原始檔案的分析，包含計數資料物件的總數；

讀取該原始檔案且建立資料物件群組，其每一者包含一或多個資料物件，以回應該原始檔案內的資料物件之重複模式；

計數資料物件的總數；

基於選定的資料物件群組數量，以選擇一或多個數值基底編碼系統；

建立預先定義的資料物件群組之一或多個庫參考檔案；

將來自該選定編碼系統之一第一獨特碼指定給每一預先定義的資料物件群組；

識別資料物件的個體重複群組，以從非循環資料物件群組區別出；

將來自該選定編碼系統之一第二獨特碼指定給每個循環與非循環資料物件群組，該等資料物件群組不是預先定義的資料物件；

將編碼的資料物件群組置放在一或多個庫參考檔案；

讀取該原始檔案及計數編碼資料物件群組的總數；

選擇一記憶體陣列的壓縮深度；

形成由選定資料物件群組之總數的平方根所定義的該記憶體陣列，且經過四捨五入(如需要)以建立一整數；

增加空值(Null)編碼資料物件以填滿該陣列；及

將該陣列置放在儲存媒體。

#### 【第 17 項】

一種壓縮資料以提供減少的資料儲存與處理之方法，其包括：

讀取一資料檔案，其包含複數個資料物件；

識別含有一或多個資料物件的循環資料物件群組，以從非循環資料物件群組區別出；

將一編碼指定給每個循環與非循環資料物件群組；

格式化最佳大小的參考庫；及

將該等編碼的資料物件群組置放在用於資料儲存與處理的該參考庫。

#### 【第 18 項】

一種用於無損資料復原諸如數字、字母、字、片語與其他標記的壓縮資料物件之方法，其包括：

取得壓縮資料物件的壓縮檔案；

選擇要復原的該壓縮資料物件之至少一部分；

選擇復原資料物件之一目的模式；

取得一參考檔案，其具有指定給循環與非循環壓縮資料物件的編碼，該參考檔案具有識別至少一或多個數值編碼系統的資訊、與有關該壓縮檔案的壓縮深度(如適用)；及

基於對該參考檔案的連結以讀取該壓縮檔案，以將該壓縮的資料物件復原成未壓縮的資料物件，以供有效率與正確使用而沒有資料物件損失。

## 圖式

1 / 3

100

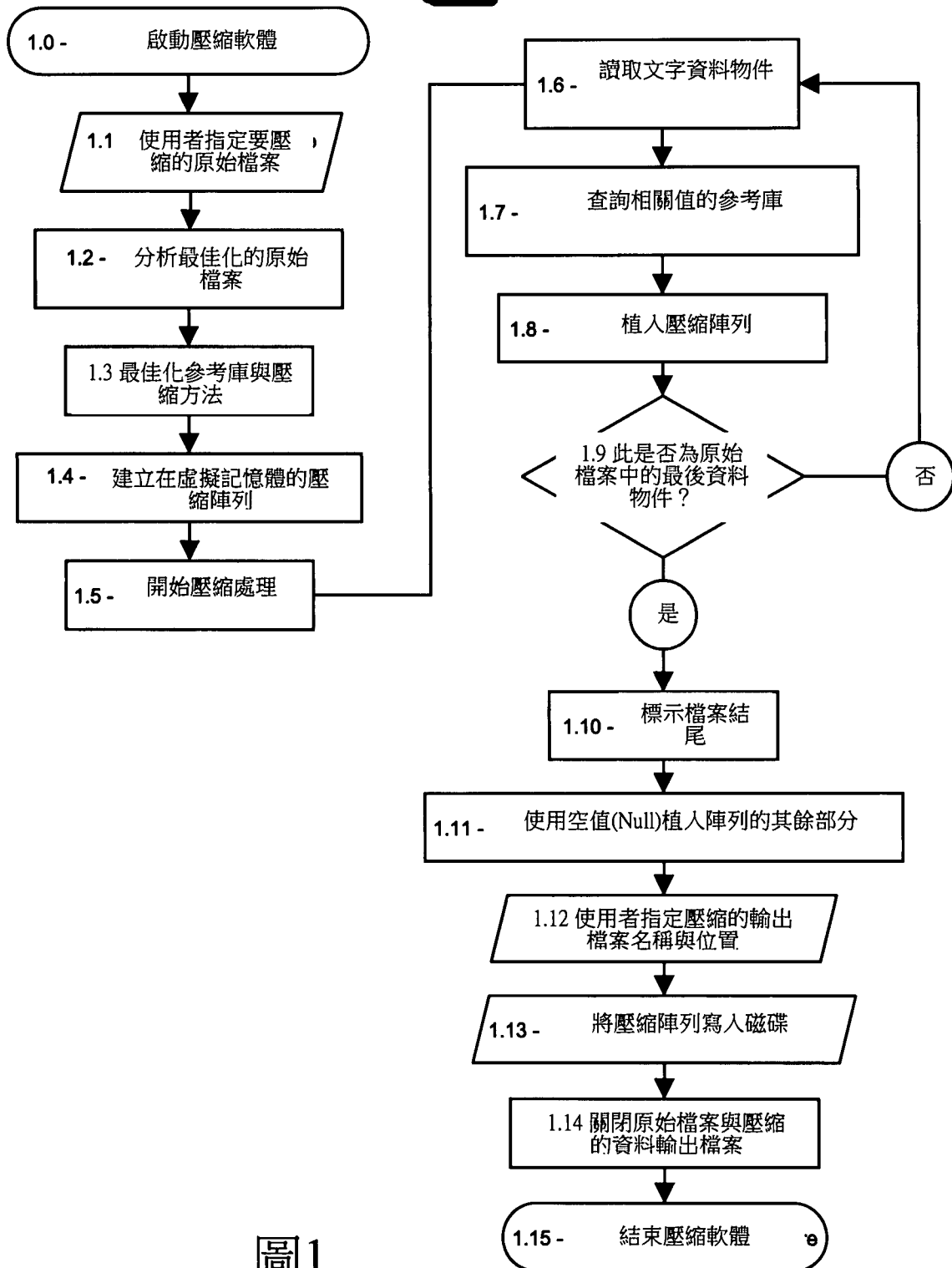


圖 1

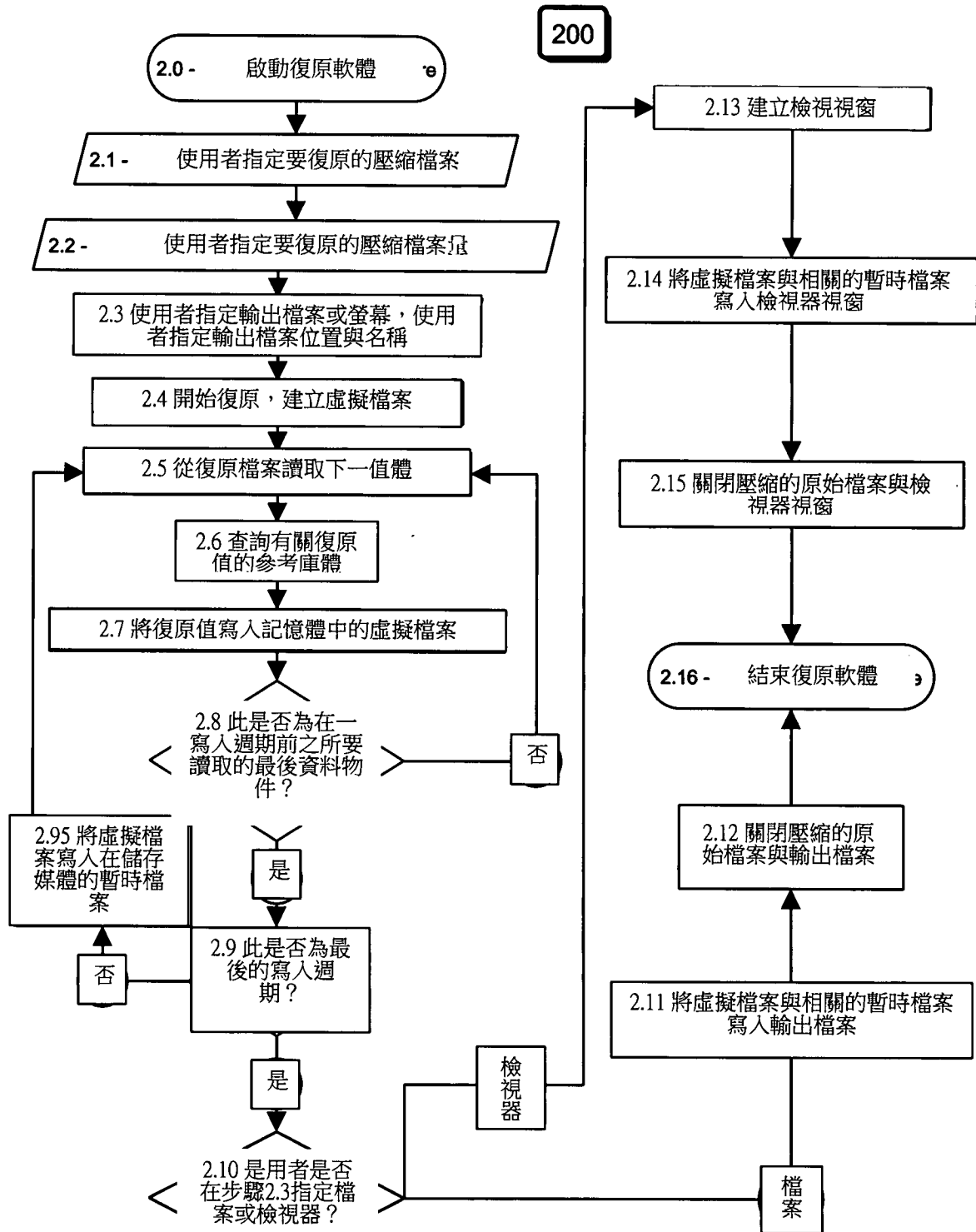


圖2

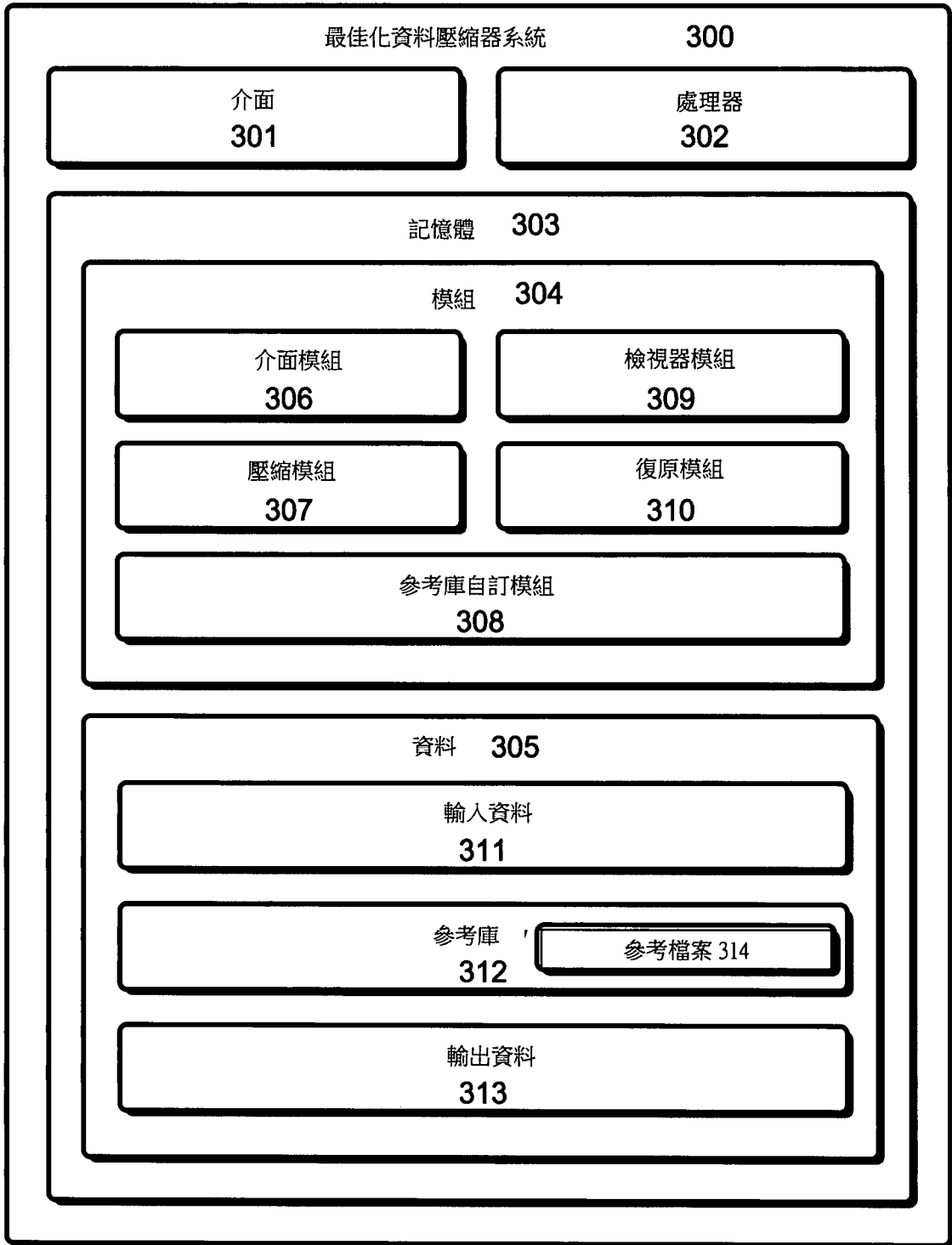


圖3