# 數位保存的趨勢與實務

柯皓仁
國立臺灣師範大學圖書資訊學研究所、圖書館

# 內容大綱

- 概論
- 管理：數位保存政策
- 管理 / 科技：OAIS 參考模型概述
- 科技：檔案格式與數位保存策略
- 科技：PREMIS與保存用詮釋資料
- 管理 / 科技：LOCKSS
- 結語

# 概論

# 數位物件保存

- 數位物件保存(digital preservation)又稱數位資料保存或數位保存，可定義為一連串主動、有系統、有管理的行動，以達成下列兩項目的
  - 數位物件位元串流(bit stream)和詮釋資料(metadata)的長期維護，以利重現原始文件適當的擬真版本；
    - 維持數位物件的語意、來源資訊、真實性、數位物件間的關聯性，以及創建與使用物件的脈絡資訊
  - 不因時間流逝和科技演進而能持續取用數位物件內容。
- 長期主動管理數位物件，以確保數位物件可被持續取用

(Research Libraries Group, 2002)

# 數位保存的迷思概念

- 數位物件保存不僅僅是備份和復原
- 數位物件保存不僅僅是有關取用
- 數位物件保存不是事後才有的想法
  - 數位物件保存必須呼應典藏機構的使命、目標、任務、文化、典藏內容
  - 及早期考量數位物件保存的相關議題，例如檔案格式的選擇、關鍵文件的取得、詮釋資料關鍵關係的描述

(Corrado & Moulaison, 2014)

# 何謂「保存」

- 保存資料：位元(bit-level)等級保存
  - Refresh、Backup、Offsite-backup、Checksum…
  - 資料的不變性 (fixity)
  - 儲存媒體的正常運作

(Digital Preservation Coalition, 2015)

- 保存資料意義：確保長期可被使用
  - Migration、Emulation、Alternative Software
- 保存資料的可信賴度：確保數位物件的真實性(authenticity)與完整性(integrity)
  - 運用詮釋資料(metadata)記錄創建與操作物件過程中的脈絡資訊
  - 運用運用資料完整性技術和維持稽核紀錄(audit trail)等方式保持數位物件的可信賴度

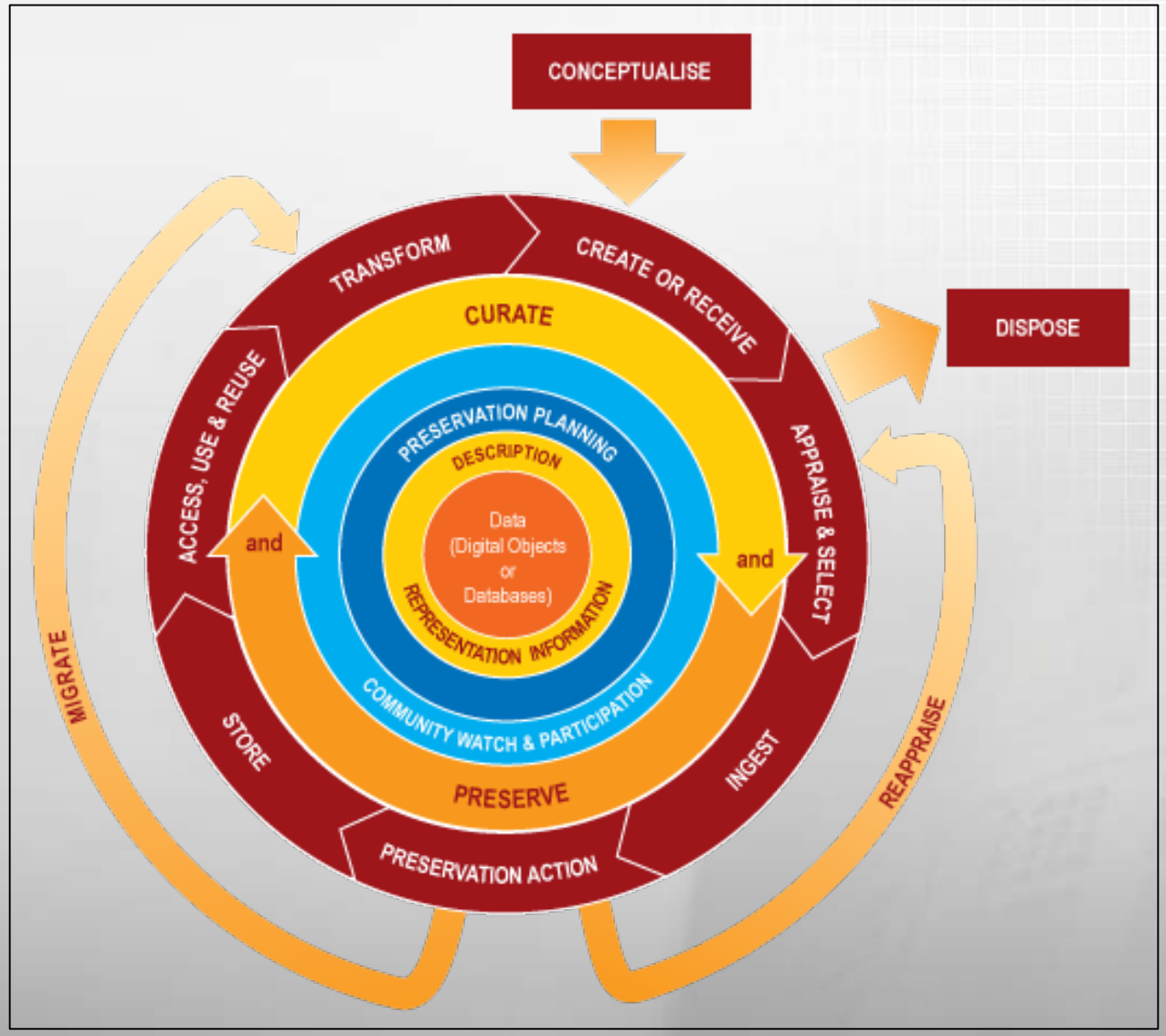
# 數位庋用 (Digital Curation)

- 在數位研究資料的生命週期中對其進行維護、保存與加值。而數位研究資料範圍廣泛，可以是政府資訊、科學資料、文化與智能資產
- 數位庋用生命週期
  - 全生命週期活動：包含資源的描述與表徵、保存規劃、社群注視與參與、庋用與**保存**
  - 循序行動：概念化、創建或接收、評價與挑選、**攝入**、**數位物件保存行動**、儲存、取用、利用與再利用、轉換
  - 偶發行動：棄置、再評價、**轉置**

(Abbott, 2009)

(Digital Curation Centre, 2008a, 2008b)

# 數位度用生命週期


CONCEPTUALISE
DISPOSE
TRANSFORM
CREATE OR RECEIVE
ACCESS, USE & REUSE
APPRAISE & SELECT
CURATE
PRESERVATION PLANNING
DESCRIPTION
Data
(Digital Objects
or
Databases)
REPRESENTATION INFORMATION
COMMUNITY WATCH & PARTICIPATION
and
and
MIGRATE
STORE
PRESERVE
INGEST
REAPPRAISE
PRESERVATION ACTION

# 數位保存概念

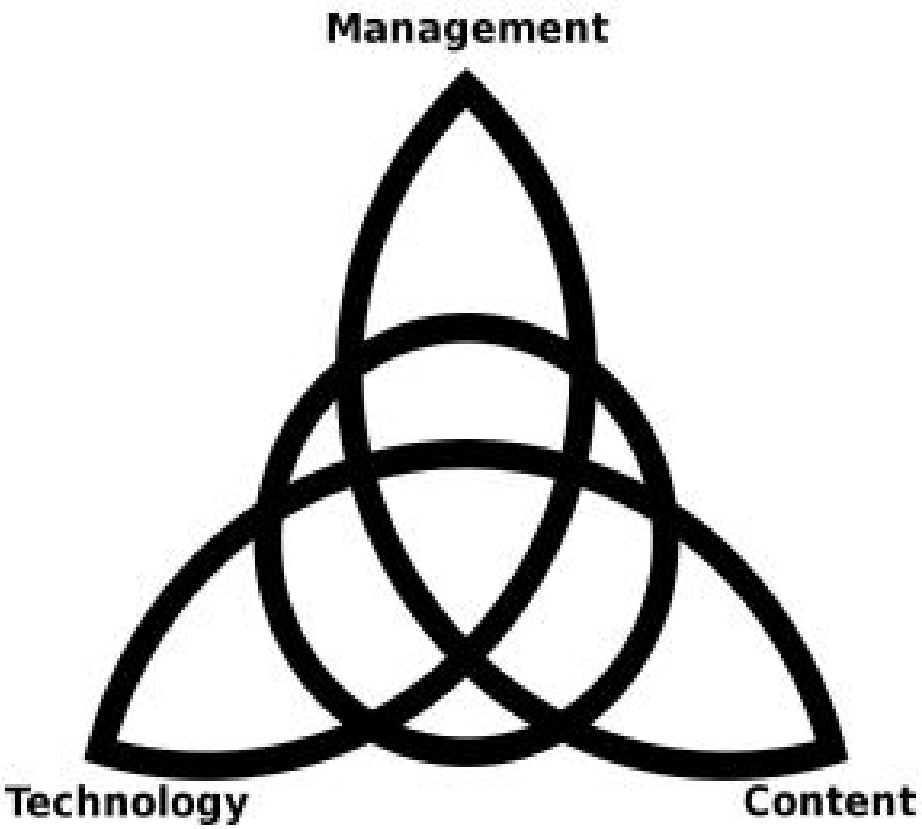



Figure 2.1. The Digital Preservation Triad

(Corrado & Moulaison, 2014)

# 數位物件保存概念 (續)

- 管理：數位物件政策與相關文件的建立、人力和財務等資源的投入與照管、推廣宣傳與永續經營，智慧財產權相關議題處理，並包含科技與內容環節的規劃
- 科技：建置和維運可信賴的數位物件保存系統、保存性詮釋資料、檔案格式等
- 內容：即需要永久保存的數位物件

# 管理：數位保存政策

# 數位保存政策

- 反映機構任務的高階文件，用以引導機構整體行動方案、準則和最佳實務的建立和執行
- 案例
  - Digital Preservation Policies Study (JISC)
  - Digital Preservation Policies: Guidance for Archives (The National Archives)
  - Analysis of Current Digital Preservation Policies: Archives, Libraries and Museums (Madeline Sheldon, Library of Congress)
  - British Library Digital preservation strategy 2013-2016 (British Library)

# 數位保存政策涵蓋內容

- Access and Use
- Accessioning and Ingest
- Audit
- Bibliography
- Collaboration
- Content Scope
- Glossary/Terminology
- Mandates
- Metadata or Documentation
- Policy/Strategy Review
- Preservation Model/ Strategy
- Preservation Planning
- Rights and Restriction Management
- Roles and Responsibilities
- Security Management
- Selection/Appraisal
- Staff Training/Education
- Storage, Duplication, and Backup
- Sustainability Planning

# 管理 / 科技：OAIS 參考模型概述

(Consultative Committee for Space Data Systems, 2012)

# 簡介

- OAIS (Open Access Information System) Reference Model
- 被數位保存專業工作者廣泛接受做為數位儲存庫的高階模型
- 描述如何為特定使用社群保存數位物件，涵蓋從將數位物件存入系統，到其為使用者所運用，以及其間的持續保存與管理活動
- OAIS RM為一抽象模型(abstract model)，情境中立(context neutral)、與軟硬體技術及實作無關

# 發展歷史

- 由美國太空資料系統諮詢委員會(Consultative Committee for Space Data Systems, CCSDS)發展
  - 原始目的是為保存太空資料，但整個發展過程亦開放(Open)給其他領域的專家學者共同討論
- 目的為發展長久保存與資訊取用所需具備的功能，提供一個參考架構和概念，便於設計出可應用於各種類型的資源且更符合數位資源永久保存的系統
- 2003年通過ISO標準，最新版為2012年版本
  - **ISO 14721:2012**
- OAIS 參考模式分成資訊模型及功能模型

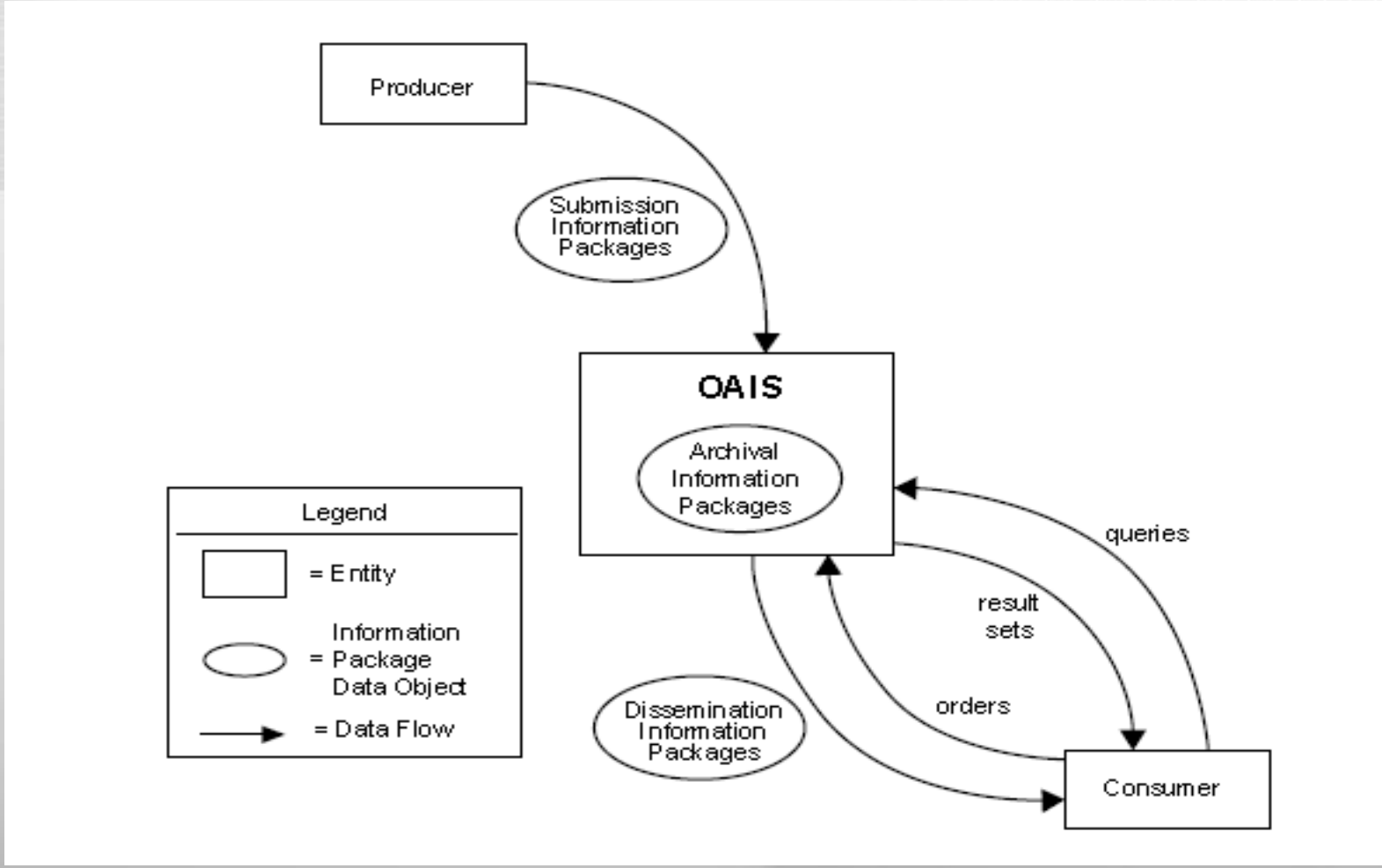
# OAIS整體概念

- 描述數位物件的結構、呈現與軟硬體環境
  - 由於資訊科技發展迅速，克服資訊軟硬體的汰舊換新成了數位物件是否能夠在未來使用的關鍵。
  - OAIS認為在保存數位物件時就需要描述其檔案結構、呈現方式，以及目前的軟硬體環境等，以利未來資訊軟硬體環境無法呈現時，能參考這些描述資訊，使用諸如轉移、模擬等策略以呈現數位物件。

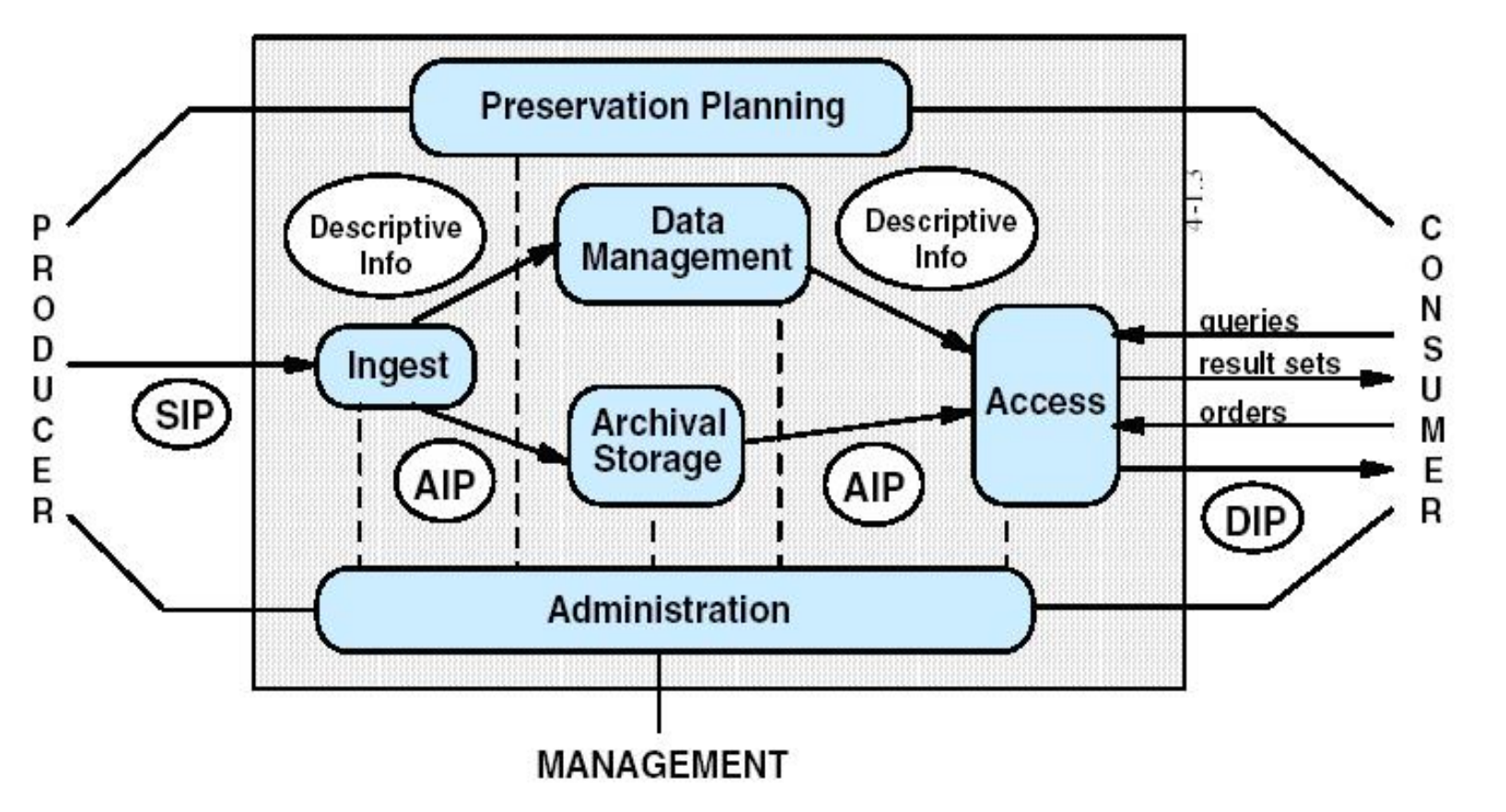
# OAIS整體概念 (續)

- 資訊封包 (information package)
  - 將數位物件及描述資訊封包在一起對長久保存是非常重要的，對一個數位物件而言，當描述資訊無法取得時，那麼可能永遠無法使用這個數位物件
  - OAIS資訊封包
    - 提交資訊封包(Submission Information Package)
    - 典藏資訊封包(Archival Information Package)
    - 傳送資訊封包(Dissemination Information Package)
- 基於於OAIS參考模式的長久保存系統間，數位物件及其相關資訊易於互通與交換
- OAIS RM 包含了名詞解釋(terminology)、資訊模型(information model)、功能模型(functional model)、必要責任(mandatory responsibility)

# OAIS High-Level External Interaction



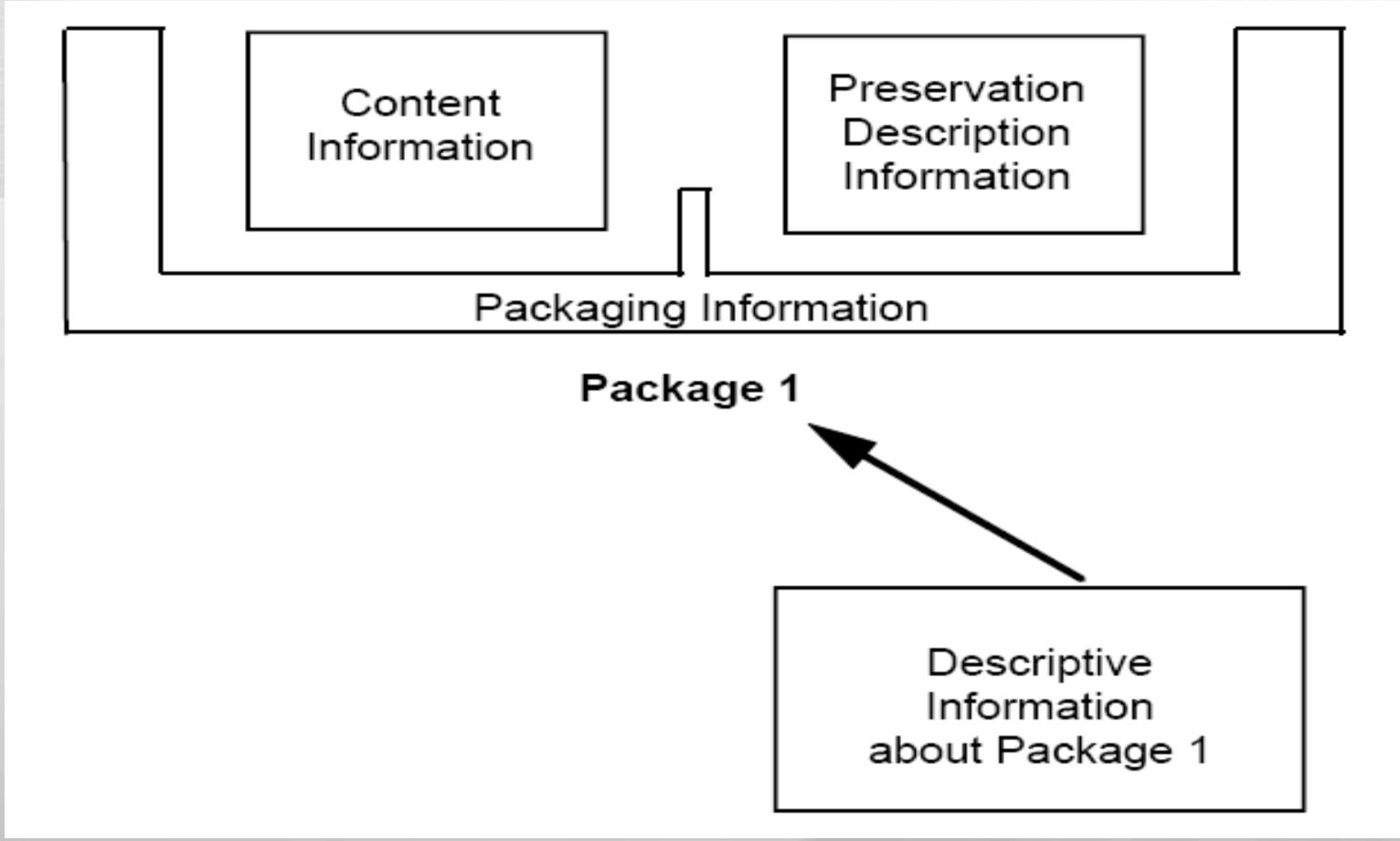
# OAIS Functional Model



# 資訊封包 (Information Package)

- 資訊封包是一個由內容資訊(Content Information, CI)和保存描述資訊(Preservation Description Information, PDI)所組成的概念型容器(conceptual container)
  - CI 和 PDI 皆為可選擇性(optional)的使用
  - 保存描述資料即是保存用詮釋資料
- 每個資訊封包都有封包資訊 (Packaging Information)用以區隔和辨識 CI和 PDI
- 資訊封包藉由描述資訊(Descriptive Information)而被發現

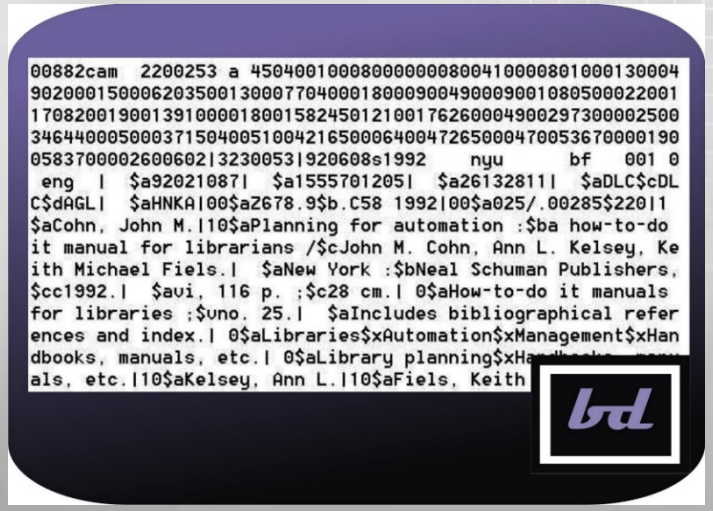
# 資訊封包


Content Information
Preservation Description Information
Packaging Information
Package 1
Descriptive Information about Package 1

# 資訊封包 (續)

- 內容是數位保存的標的
  - 內容資料物件(Content Data Object, CDO)：位元串流
  - 表徵資訊 (Representation Information, RI)：讓CDO得以被(人或軟體)解讀並呈現的資訊
    - ISO 2709 format: 24-bit record label + directory (field tag – field length – starting field position) + data fields + record separator

- 英文文件 + 英漢字典 + 文法規則
- PDF文件 + PDF標準文件
- 軟體本身可能就擁有RI的功能




Source

# 保存用詮釋資料

- 參考資訊 (Reference Information)
  - 描述一或多個用以辨識內容資訊(CI)的識別機制與識別號
  - 外部系統可利用這些識別機制與識別號獨一無二地連結、使用特定的內容資訊
  - 在OAIS中，可能和描述資訊(DI)的內容有若干重複
- 脈絡資訊 (Context Information)
  - 描述內容資訊(CI)和環境(實體或虛擬)的關係
  - 為何創建特定內容資訊，和其他內容資訊間的關係

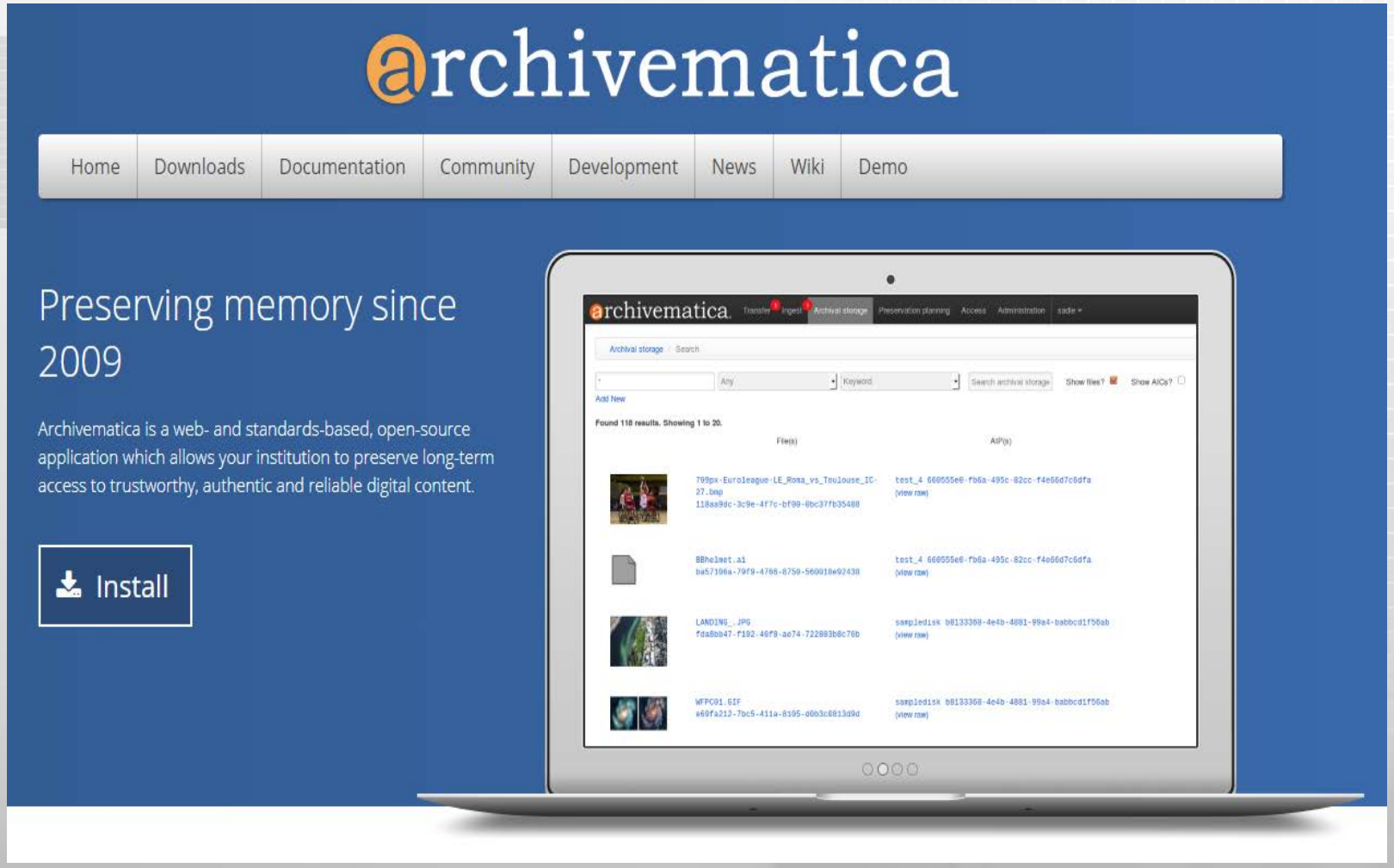
# 保存用詮釋資料 (續)

- 來源資訊 (Provenance Information)
  - 描述內容資訊的歷史
  - 包含內容資訊的來源或出處、創建後所做的任何更修、創建後的保管單位、審計追蹤(audit trail)
  - 提供內容物件真實性(authenticity)的若干保證
  - 可視為脈絡資訊的一種
- 不變性資訊 (Fixity Information)
  - 做為資料完整性(data integrity)檢核之用
  - checksum / encryption key

# 保存用詮釋資料 (續)

- 取用權利資訊 (Access Rights Information)
  - 與內容物件相關的取用限制
  - 包含法律框架、授權條款、取用控制
  - 與內容產生者/所有權人商訂之取用及傳播條件，通常記載在資訊提交合約(Submission Agreement)，載明OAIS在進行內容資訊保存條件，以及終端使用者的使用條件
- OAIS並未說明保存用詮釋資料的元素集或資料結構
- PDI: Preservation Metadata → PREMIS
  (Preservation Metadata: Implementation Strategies)

# Archivmatica



# 科技：檔案格式與數位保存策略

# 適合數位保存的檔案格式要件

- 資訊揭露程度：檔案格式是否具備完整規格說明和可驗證技術完整性的工具，且該等規格說明和工具能為創建與維護數位物件者所取用。
- 採用程度：檔案格式是否已經被主要的數位物件創建者、傳播者和使用者所採用。
- 通透程度：檔案格式的內部表徵方式是否可讓基本工具進行直接分析，甚至讓使用者可直接用文字編輯器閱讀檔案內容。

(Arms & Fleischhauer, 2005)

# 適合數位保存的檔案格式要件(續)

- 自我描述性：檔案格式的詮釋資料是存放於該檔案內部，而非獨立存在。
- 外在依存性：檔案格式是否必須依賴特定的硬體、軟體、作業系統才能夠使用。
- 專利的影響：檔案格式是否受專利保護。
- 技術保護機制：檔案格式是否經過加密等機制保護導致其內容不易解讀或轉置成其他格式。

# 檔案格式辨識與盤點工具

- 檔案格式辨識工具：DROID、JHOVE
- 檔案格式盤點

# DROID

- Digital Record Object Identification (DROID)
- 英國國家檔案館 (UK National Archives) 開發之批次自動辨識檔案格式的工具
- 搭配PRONOM檔案格式註冊中心，存放各種檔案格式的資訊和特徵簽名

# PRONOM

The National Archives

Search our website

About us | Education | Records | Information management | Archives sector

You are here: Home > Information management > Our projects and work > Digital preservation > PRONOM > Search by format > Details: Summary

The technical registry
PRONOM

Welcome : About | Add an entry
Search | Help | Information resources

Details: File format summary

Help : detailed report on file format

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

## Details for: OpenDocument Text 1.0

Save as... XML | CSV  Print

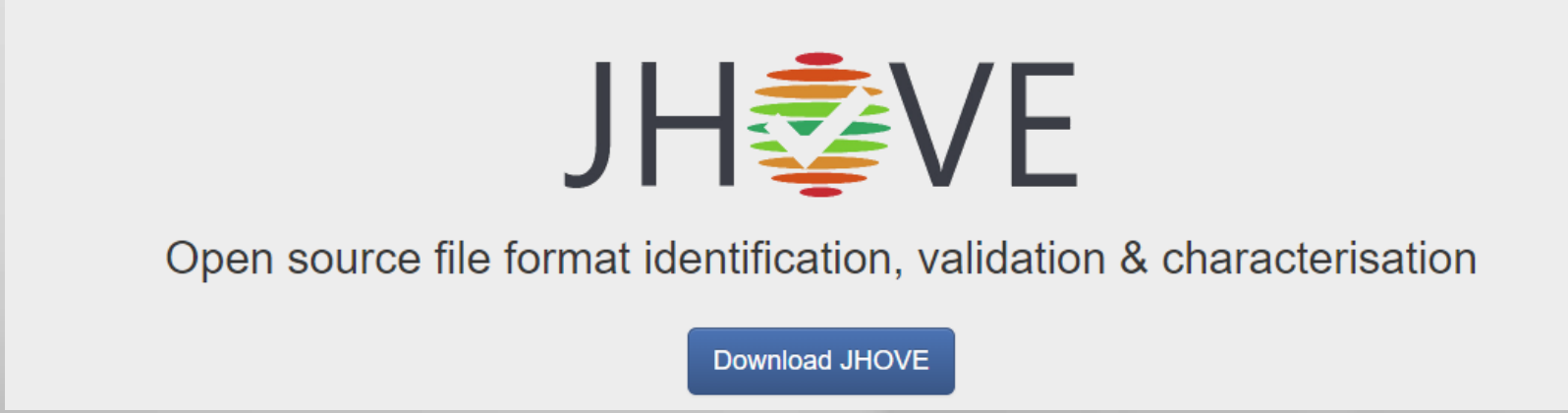
Go to: Summary | Documentation > | Signatures > | Compression > | Character encoding > | Rights > | Reference files > Properties >

### Summary

| | |
|---|---|
| **Name** | OpenDocument Text |
| **Version** | 1.0 |
| **Other names** | ODF Text (1.0) |
| **Identifiers** | MIME: application/vnd.oasis.opendocument.text<br>PUID: fmt/136 |
| **Family** | |
| **Classification** | Word Processor |
| **Disclosure** | Full |
| **Description** | OpenDocument (ODF) format is a generic format for office documents, such as text, spreadsheets, presentations, drawings, and databases. It was developed by OASIS, based on the original OpenOffice document format. The format is an open standard, and is supported by a variety of office software packages. ODF is an XML format, which may comprise a single XML document, or a collection of subdocuments within a package. The package format is the commonly used form, and comprises a standard ZIP package containing the XML content and any associated binary data, together with a manifest which lists the package content. An ODF text document contains word-processed text. The content is contained in a 'content' xml subdocument. Additional subdocuments contain style information, document metadata, and application-specific settings. |

# JHOVE

- JSTOR/Harvard Object Validation Environment
- JSTOR/Harvard共同開發的物件確認環境
- 可擴充軟體框架，用以辨識、確認、描述物件特性



# 數位保存策略

- 更新(refreshing)：更新儲存媒體
- 轉置(migration)：物件格式的轉換
- 技術保存(technology preservation)：維持取用數位物件所需的軟硬體
- 模擬(emulation)：任天堂模擬器、Bluestacks、虛擬機(virtual machine)
- 標準化(standardization)：數位物件的標準格式
- 封裝(encapsulation)：數位物件、軟硬體、說明文件的時光膠囊
- 重製系統(redundancy)：異地備援、容錯/高可用性系統、鏡射站

(歐陽崇榮，2006)

# 科技：PREMIS與保存用詮釋資料

(PREMIS Editorial Committee, 2015)

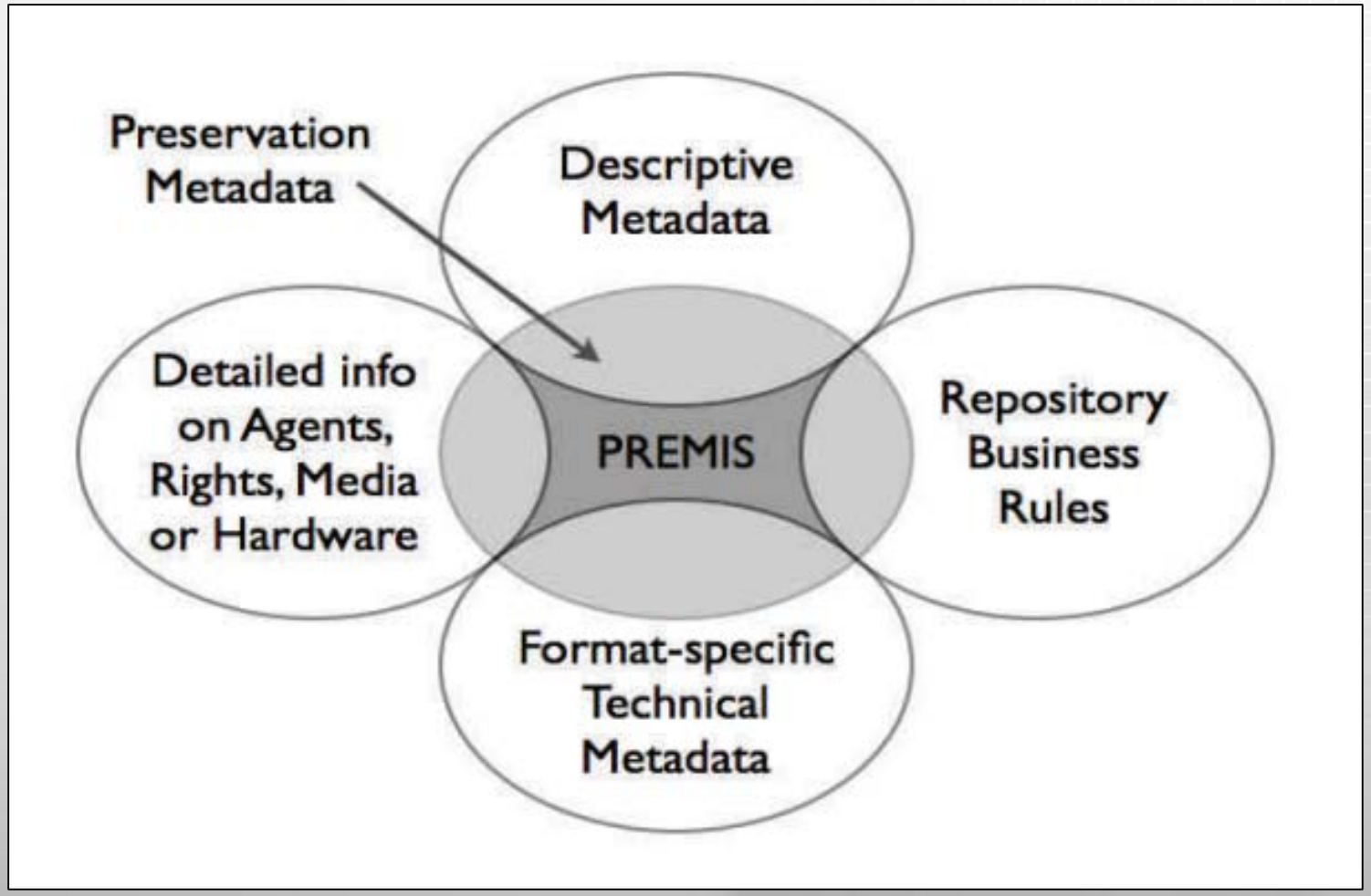
# 整體概念

- PREMIS – **PRE**servation **M**etadata: **I**mplementation **S**trategies
- 保存性詮釋資料：儲存庫用來支持數位物件保存程序的資訊
  - 確保物件在不知情的情況下被修改：Checksum
  - 檔案儲存媒體老舊：儲存媒體形式與年限、最後一次更新日期
  - 原始檔案格式與軟硬體環境，以利實踐保存策略
  - 數位物件保存行動可能更改原始資源或其呈現方式，讓資源真實性存疑。必須記錄資源的數位起源(digital provenance)，包含保管鍊(chain of custody)和授權變更歷史(authorized change history)

# PREMIS

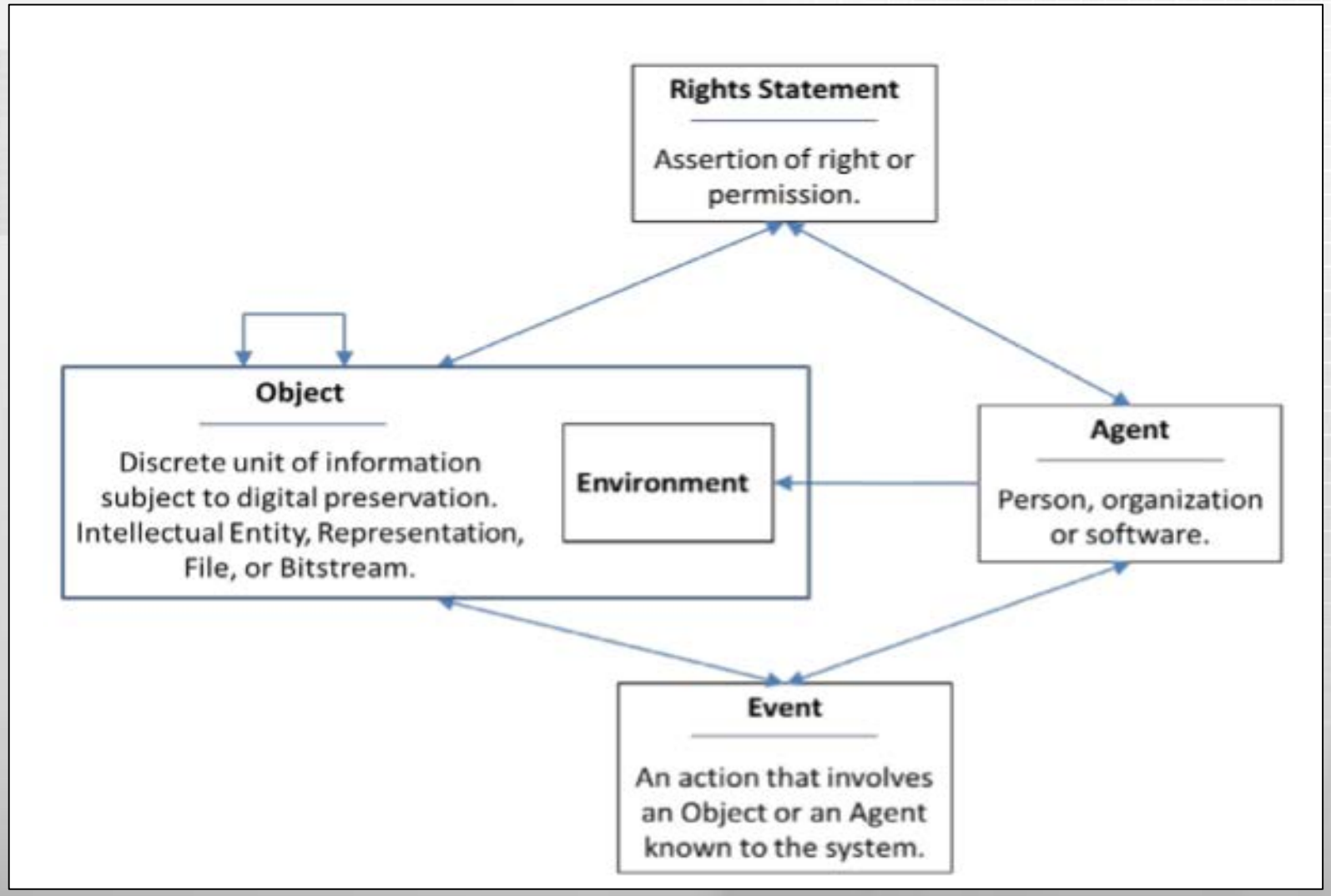
- 資料字典 (data dictionary) – Narrative description of PM – Schema
- 大多數數位物件保存環境系統所需要的保存用詮釋資料元素，有些詮釋資料是刻意被排除：
  - 與特定格式相關的詮釋資料，例如僅與特定檔案格式或僅與特定資料種類相關
  - 與特定實作方式相關的詮釋資料和企業規則，例如特定數位物件保存系統的政策或實務。
  - 描述性詮釋資料
  - 關於特定儲存媒體或硬體相關的資訊
  - 除了辨識需求之外的有關代理者的資訊
  - 除了直接影響數位物件保存功能外之權利與權限相關資訊

# PREMIS與保存性詮釋資料




(Caplan, 2009)

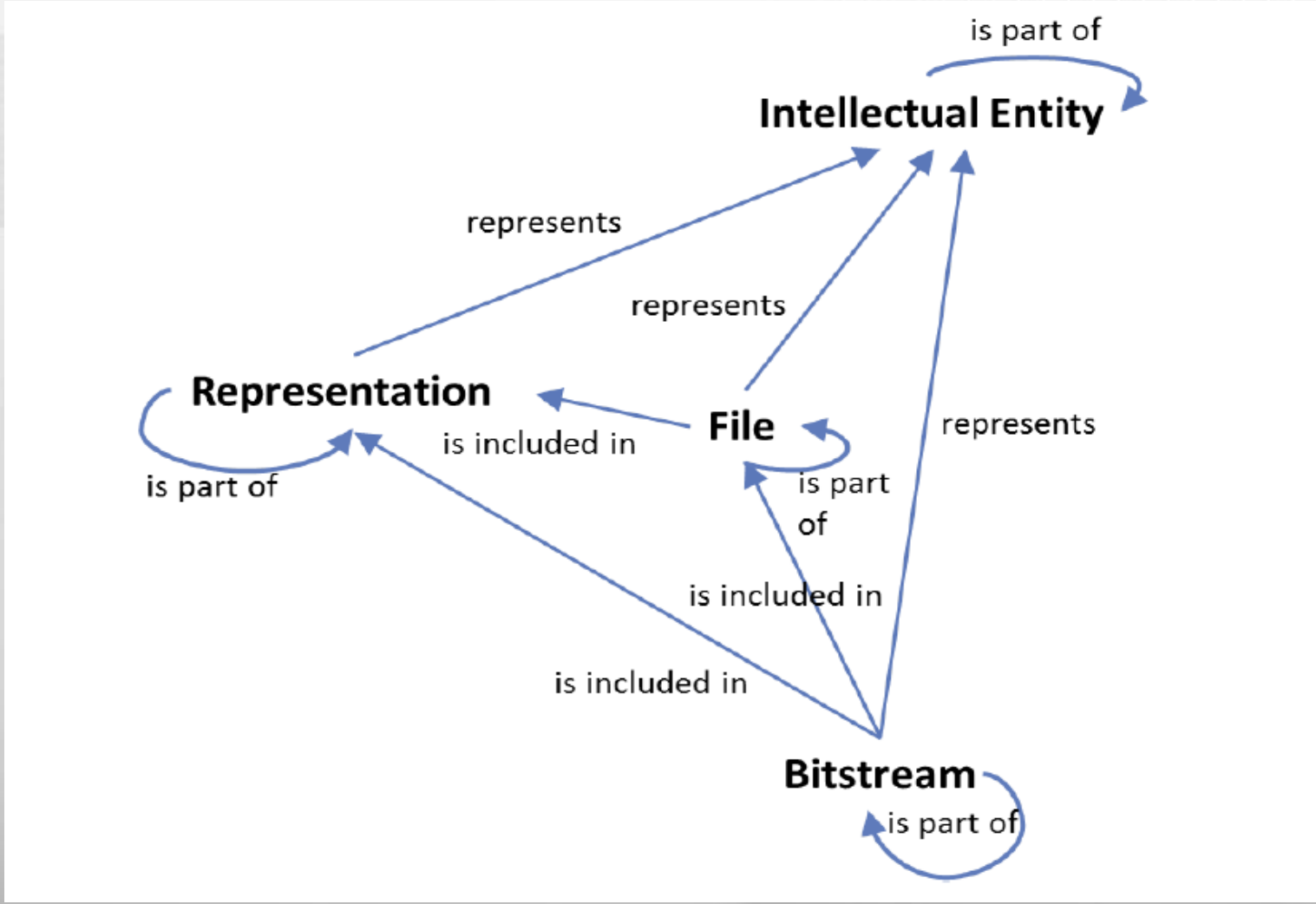
# PREMIS的五種實體



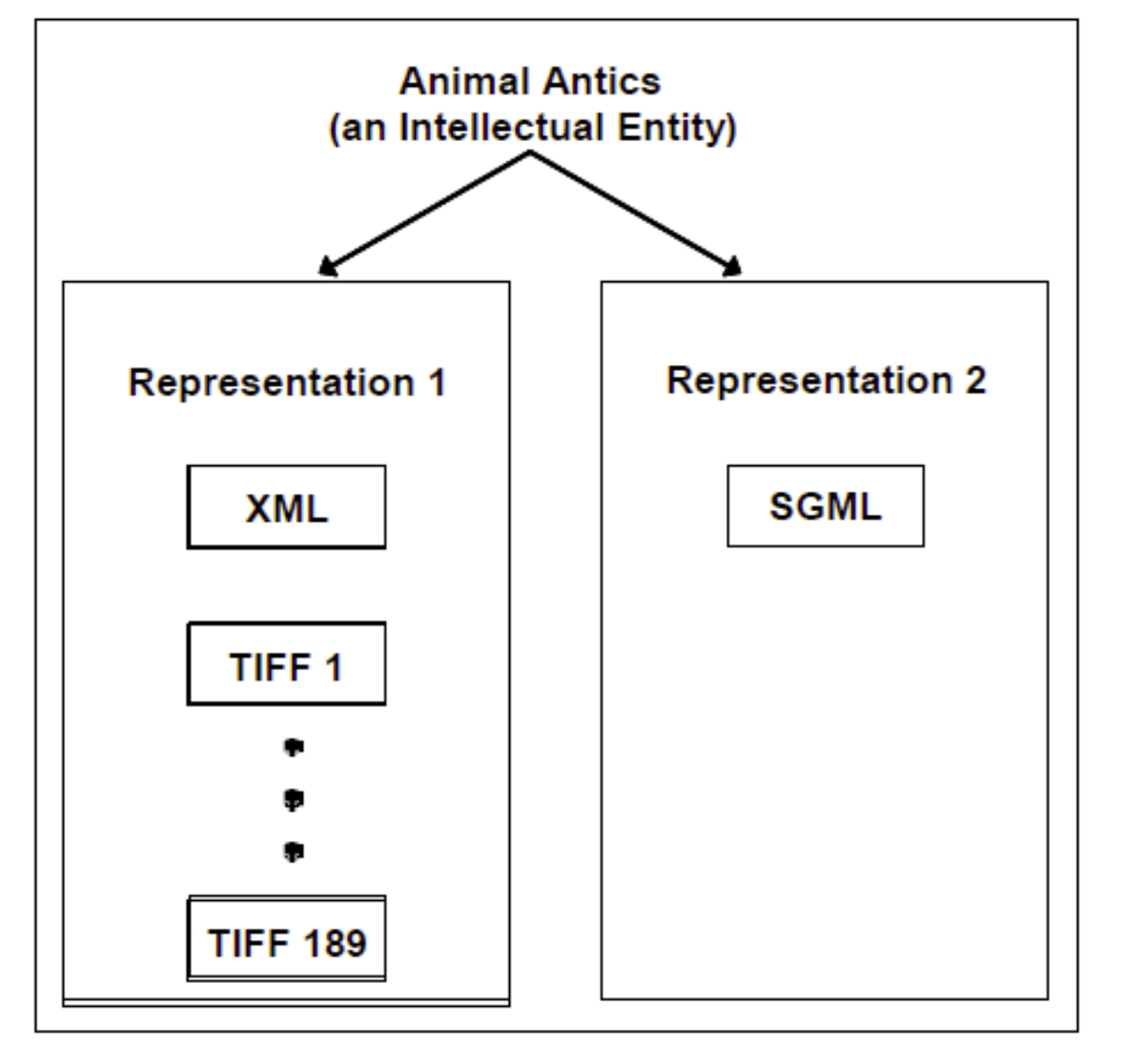
# PREMIS實體

- 物件(Object)：需要數位保存的獨立資訊單元
  - 與數位保存程序相關的環境(Environment)亦屬於物件實體
- 環境(Environment)：支持數位物件的軟硬體科技
- 事件(Event)：數位物件長久保存**儲存庫**所知曉之關係或影響至少一個物件或代理人的活動
- 代理者(Agent)：與物件生命中的事件或權利相關的人、組織、或軟體程式/系統
- 權利描述(Right Statement)：與物件或代理人有關的權利或許可

# 不同層級的物件


is part of
Intellectual Entity
represents
represents
Representation
File
represents
is included in
is part of
is part of
is included in
is included in
Bitstream
is part of

# 知識實體、表徵、檔案、位元流



# PREMIS 實體的範例

- 物件：AVI格式的”Welcome to U”
  - ”Welcome to U” – 知識實體
  - Single AVI file – 表徵、檔案
  - Audio bits / video bits – 位元流
- 事件：轉置、版本更新、備份、CheckSum
  - 儲存庫依照事件的重要性決定是否要記錄
  - 會修改物件內容的事件一定要記錄
  - 如備份的事件可以放在 system log 或 audit trail

# PREMIS 實體的範例 (續)

- 代理人
  - 擁有或給予權利者
  - 執行、認可、引發事件者
  - 藉由事件或權利宣告而產生物件或對物件採取行動者
- 權利描述
  - 至少要記錄：儲存庫憑藉著何種權利或許可而能對物件執行轉置等長久保存的工作

# 容器與子單元

- 容器(Container)：這種語意單元本身並不擁有值，而是用來將相關的語意單元聚合成群(這些相關的語意單元就稱為子單元(subunit)

```
1.1 objectIdentifier (M, R)
    1.1.1 objectIdentifierType (M, NR)
    1.1.2 objectIdentifierValue (M, NR)
```

  - objectIdentifierType：例如ISBN, DOI, URI
  - M: Mandatory, R: Repeatable
- 擴充容器(Extension Container)：這種特殊容器不含有子單元，其目的是用來記錄非PREMIS的詮釋資料

# 物件的語意單元

- 物件ID
- 不變性資訊
- 物件大小
- 物件格式
- 物件原始名稱
- 創建相關資訊
- 抑制器相關資訊
- 重要性質相關資訊
- 環境相關資訊
- 媒體種類與存放位置
- 數位簽章資訊
- 與其他物件或實體的關係

# 物件的語意單元 (續)

1.1 objectIdentifier (M, R)
- 1.1.1 objectIdentifierType(M, NR)
- 1.1.2 objectIdentifierValue(M, NR)

1.2 objectCategory (M, NR)
1.3 preservationLevel (O, R)
1.4 significantProperties (O, R)
1.5 objectCharacteristics (M, R)
1.6 originalName (O, NR)
1.7 storage (O, R)
1.8 signatureInformation (O, R)
1.9 environmentFunction (O, R)
1.10 environmentDesignation (O, R)
1.11 environmentRegistry (O, R)
1.12 environmentExtension (O, R)
1.13 relationship (O, R)
1.14 linkingEventIdentifier (O, R)
1.15 linkingRightsStatementIdentifier (O, R)

# objectIdentifier的定義

| Semantic unit | 1.1 objectIdentifier | | |
|---|---|---|---|
| Semantic components | 1.1.1 objectIdentifierType<br>1.1.2 objectIdentifierValue | | |
| Definition | A designation used to identify the Object uniquely within the preservation repository system in which it is stored. | | |
| Rationale | Each Object held in the preservation repository must have a unique identifier to allow other entities to refer to it and to relate it to descriptive, technical, and other metadata unambiguously. | | |
| Data constraint | Container | | |
| Object category | Intellectual Entity / Representation | File | Bitstream |
| Applicability | Applicable | Applicable | Applicable |
| Repeatability | Repeatable | Repeatable | Repeatable |
| Obligation | Mandatory | Mandatory | Mandatory |
| Creation / Maintenance notes | An identifier may be created by the repository system at the time of ingest, or it may be created or assigned outside of the repository and submitted with an object as metadata. Similarly, identifiers can be generated automatically or manually. | | |
| Usage notes | The *objectIdentifier* is mandatory for all Objects stored.<br>The *objectIdentifier* is repeatable in order to allow both repository-assigned and externally-assigned identifiers to be recorded. See "Creation/Maintenance" note above.<br>Primary identifiers must be unique within the repository. They may be preexisting, and in use in other digital object management systems. Ideally, secondary identifiers should also be unique but sometimes this is not possible (e.g., if the values are inherited from a legacy system which did not enforce this or only identified items at a higher level). Identifiers for each item must be sufficient to identify the item uniquely at the appropriate level of aggregation. For example, an Intellectual Entity that represents all books in the same edition could use an ISBN but this would be insufficient to identify a particular copy of that book.<br>A preservation repository needs to know both the type of object identifier and the value. If the value itself contains the identifier type (e.g., "oai:lib.uchicago.edu:1"), the identifier type does not need to be recorded explicitly. Similarly, if the repository uses only one type of identifier, the type can be assumed and does not need to be recorded. | | |

# 事件的語意單元

- 事件ID **(eventIdentifier)**
- 事件種類 (創建、攝入、轉置等) **(eventType)**
- 事件發生的日期和時間 **(eventDateTime)**
- 事件的詳細敘述 **(eventDetailInformation)**
- 事件的結果**(eventOutcomeInformation)**
- 事件結果的詳細敘述**(eventOutcomeDetail)**
- 與事件相關的代理者和其所扮演的角色 **(linkingAgentIdentifier)**
- 與事件相關的物件和其所扮演的角色 **(linkingObjectIdentifier)**

# 代理者的語意單元

- 代理者
  - 代理者ID **(agentIdentifier)**
  - 代理者名稱 **(agentName)**
  - 代理者種類(人、組織、軟體) **(agentType)**
  - 代理者版本 **(agentVersion)**
  - 代理者備註 **(agentNote)**
  - 代理者擴充容器 **(agentExtension)**
  - 與代理者相關的事件 **(linkingEventIdentifier)**
  - 與代理者相關的權利描述 **(linkingRightsStatementIdentifier)**
  - 與代理者相關的環境和其所扮演的角色 **(linkingAgentIdentifier)**

# 權利的語意單元

- 權利
  - 權利描述的ID **(rightStatementIdentifier)**
  - 權利性質(著作權法、授權、法令) **(rightsBasis)**
    - 根據權利性質Copyright, License, Statute選用適當的語意單元
  - 有關權利性質的詳細敘述
  - 權利描述所允許的行動 **(act)**
  - 行動的限制 **(restriction)**
  - 允許的條款、有效時間 **(rightsGranted)**
  - 權利描述應用的物件與角色 **(linkingObjectIdentifier)**
  - 相關的代理者與角色 **(linkAgentIdentifier)**

# 管理 / 科技： LOCKSS

(Reich & Rosenthal, 2001)

# 想解決的問題

- 想解決的問題：學術文獻的保存與長期取用
  - 未來的科學家需要取用學術文獻以便教學與研究
  - 現在與未來的圖書館館員需要便宜、強健、能管控的機制，以確保其服務社群能長期取用重要文獻
  - 現在與未來的出版社需要確保其出版品僅能為授權且身分認可的讀者使用
- 解決方案的需求
  - 內容必須原汁原味保存 (bit-level preservation)
  - 必須能取用內容
  - 必須能剖析與瞭解內容

# 儲存庫(Repositories)模型

- 集中式模型 – OAIS
  - 重責大任落於一身 – 保存內容、軟硬體、技術人力
  - 通常以暗黑典藏(dark archives)形式處理
- 分散式模型 – LOCKSS
  - 責任分攤
    - 每個螺絲釘負責一部分
    - 軟硬體不用太高檔、太複雜
  - 以取用內容為優先考量
  - LOCKSS亦宣稱每個Box符合OAIS

# LOCKSS 起源

- 1999年由史丹佛大學圖書館發起的一個電子資源、電子期刊為主的數位典藏計畫
- 現行圖書館對於電子資料庫 / 期刊大部分只能擁有Access 的權利
- 主要是為了解決電子出版物的收集和永久性保存問題
  - 確保電子期刊、資料庫能得到良好的保存
  - 是一個便宜、技術上可靠的點對點(peer-to-peer)軟體來備份資料
  - 與各圖書館、出版社攜手合作，為下個世代保存電子資料/資訊

# LOCKSS 起源 (續)

- Lots Of Copies Keep Stuff Safe
- 保護遺留下來的東西不應該用鎖和保險箱，以至於他們無法被看見。反而應該製造很多備份，這麼一來萬一出了什麼意外，我們還有很多備份 – *美國第三任總統 湯瑪斯傑佛遜*

"…let us save what remains not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident."

*-Thomas Jefferson, Feb. 18, 1791*

# LOCKSS網站

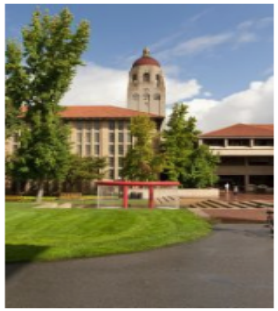
LOCKSS LOTS OF COPIES KEEP STUFF SAFE

Search

About LOCKSS | The Community | News & Media | Technical Support | How to Join | Contact Us

LOCKSS

## Lots Of Copies Keep Stuff Safe



**The LOCKSS Program**, based at Stanford University Libraries, provides libraries and publishers with award-winning, low-cost, open source digital preservation tools to preserve and provide access to persistent and authoritative digital content.

**LIBRARIES KEEP WHAT THEY BUY.** Librarians need to know that the digital content they acquire today will not disappear when they cancel subscriptions, and that their electronic collections can be preserved and accessed by readers far into the future.

**PUBLISHERS PRESERVE WHAT THEY PUBLISH.** Publishers need to know that the integrity of their web-based content will remain unchanged and available in perpetuity—even if their own website is no longer available.

**READERS HAVE CONTINUAL ACCESS.** The LOCKSS Program was formed by the library community in response to these needs and concerns.

Read more about LOCKSS collaborative digital preservation, post cancellation and perpetual access.

9/2017

**An Andrew W. Mellon Foundation** grant enables systems to be outfitted with award-winning LOCKSS digital preservation capabilities.

**Private LOCKSS Networks.** Communities of all sizes are using Private LOCKSS Networks to preserve their own cultural heritage materials. Join one, or build your own!

# LOCKSS 原理

- 紙本媒體的永久保存：把全球圖書館看成是一個分散式系統，紙本媒體一旦出版、被圖書館典藏，就很難消除它在世上曾經存在的痕跡
- LOCKSS原理
  - 圖書館扮演Web Cache的角色 (安裝LOCKSS BOX)
  - 使用者利用其圖書館Web Cache取用授權使用的數位資源 (無論出版社是否仍然存在)
  - Cache間會***慢慢***地***對話***以偵測遺失或損毀的數位資源，必要時透過出版社或其他Cache補足或修復數位資源

**緩慢、決心、長壽**

# 從使用者角度來看

- 使用者取用學術文獻的需求
  - 檢索 – 準確的檢索結果與後續連結 (discovery)
  - 超連結 – 不要隨便404 (delivery)
- No discovery, no delivery
- 在LOCKSS裡，Cache扮演Proxy的角色
  - 固定時間間隔到出版社網站去爬取資料，並預載新出版的學術文獻
  - 提供跨出版品查找的功能

# 從圖書館角度來看

- Lease 不等於 Purchase-and-Own
- LOCKSS有點類似Purchase-and-Own
  - 天下沒有免費的午餐
  - 好在LOCKSS就算不是免費，也不算貴
- LOCKSS扮演代理伺服器的角色
  - 但Cache不會清掉，因此可以長期使用文獻內容
- 萬一自己的LOCKSS掛點
  - 可以向其他圖書館的LOCKSS索取 – 類似館際互借
  - 但前提時必須要證明自己有資格索取，以避免自由載入(free load)

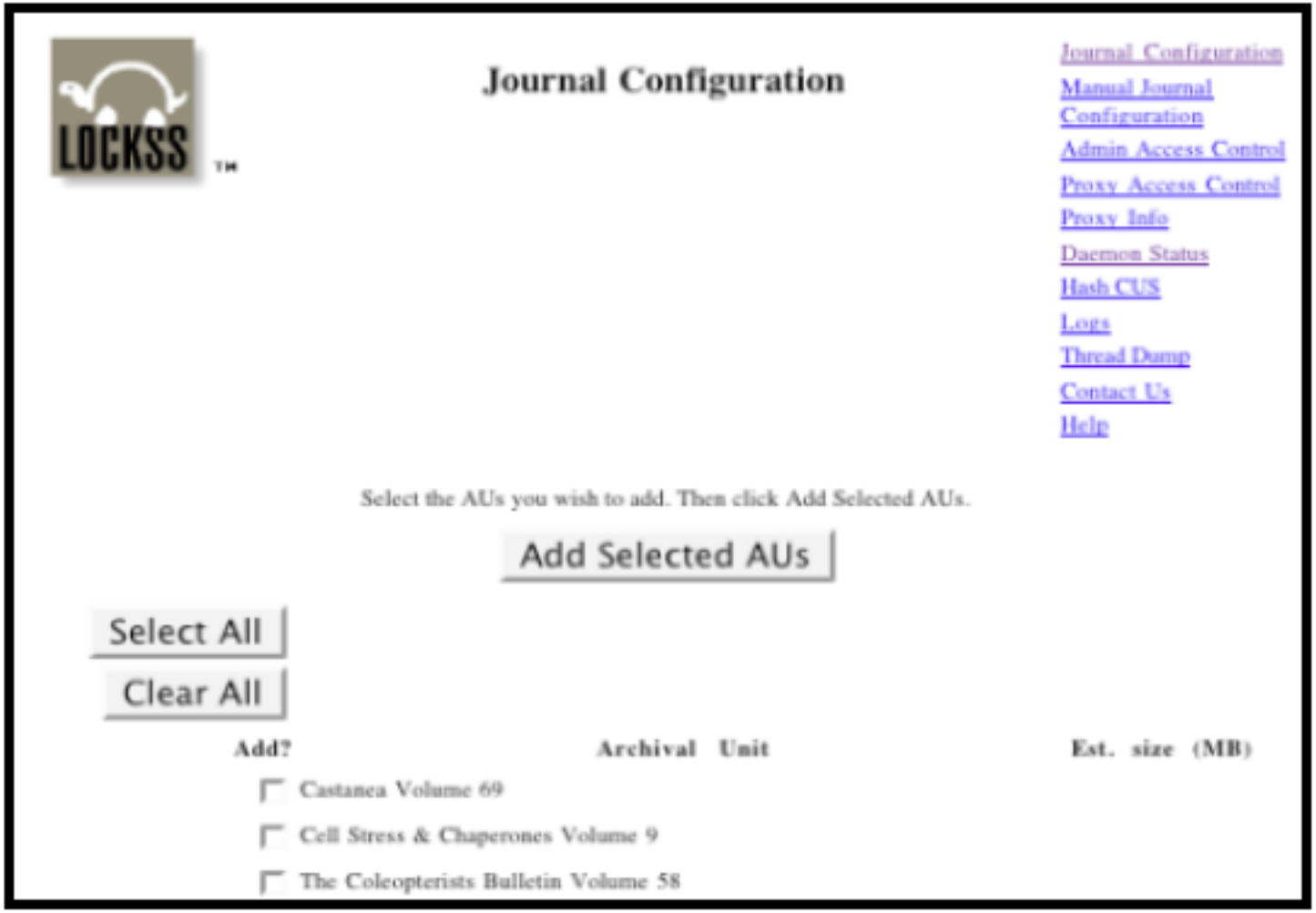
# 從出版商角度來看

- 出版商在乎品牌與形象，LOCKSS可以確保資料的運作正常，且能提供紀錄資料
- 出版商在乎未經授權的下載、複製、傳遞
- 出版商參加LOCKSS的優點
  - 善盡永久保存的責任，且讓圖書館承擔永久保存的費用
  - 出版商也可成為Cache，在協助修復過程中可偵測侵權者
  - 在正常情形下，使用者透過Cache向出版商下載資料，不影響使用(統計)

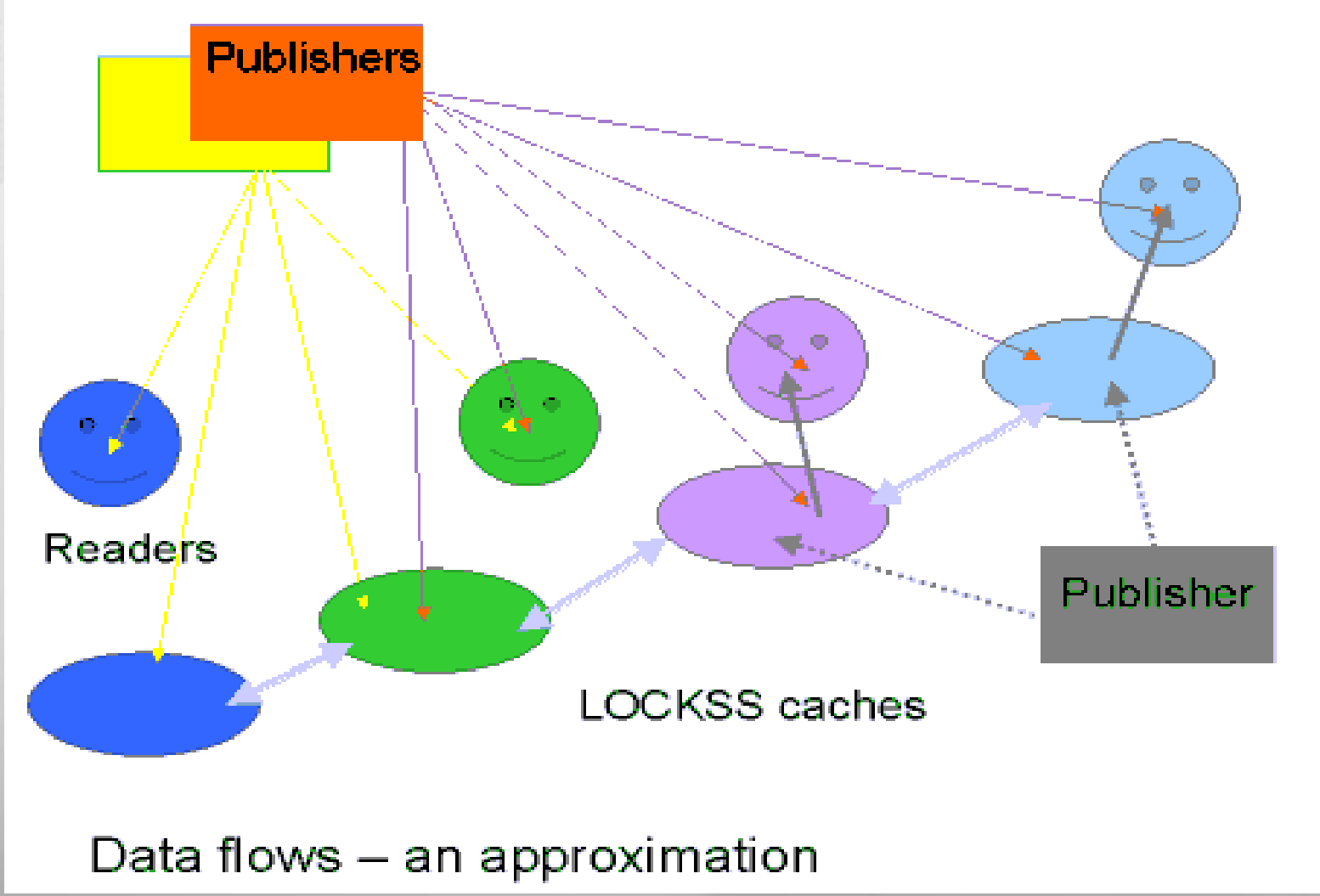
# LOCKSS如何保存內容

- LOCKSS介面
- LOCKSS資料流示意圖
- 資料攝入
- 資料保存
- 傳播
- 格式轉置

# LOCKSS 介面


LOCKSS
Journal Configuration
Journal Configuration
Manual Journal Configuration
Admin Access Control
Proxy Access Control
Proxy Info
Daemon Status
Hash CUS
Logs
Thread Dump
Contact Us
Help
Select the AUs you wish to add. Then click Add Selected AUs.
Add Selected AUs
Select All
Clear All
Add?
Archival Unit
Est. size (MB)
Castanea Volume 69
Cell Stress & Chaperones Volume 9
The Coleopterists Bulletin Volume 58

# 資料流示意圖

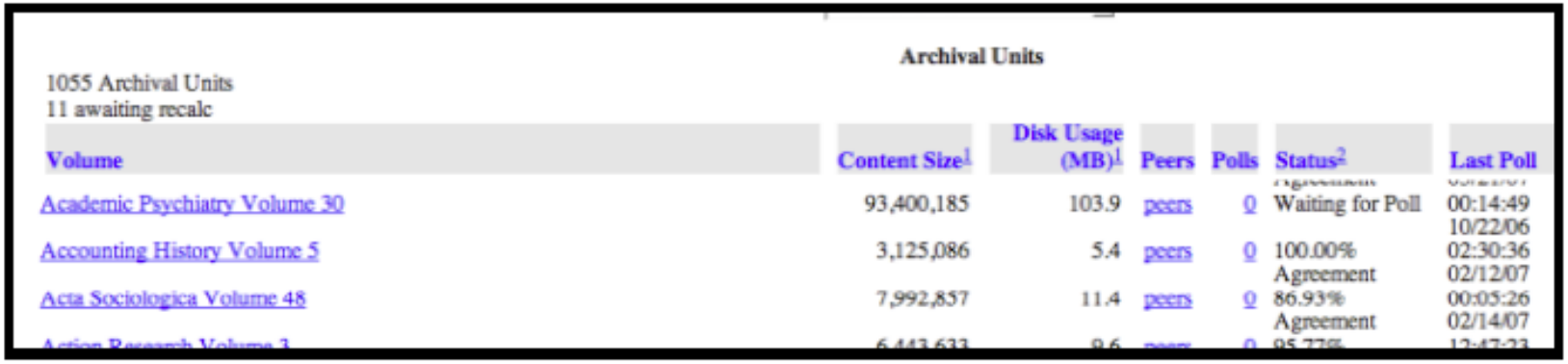

Publishers
Readers
LOCKSS caches
Publisher
Data flows – an approximation

# 資料攝入 (Ingest)

- 利用 Web Crawler 到出版商網站爬取內容
  - 出版商必須允許LOCKSS爬資料 (manifest page)
  - 為每個出版商建置設定檔
    - manifest page的位置、網站架構、爬取的時間、一次可爬取的量
  - 每個Box個別到出版商網頁爬資料，再透過Poll機制建立權威版本

# 資料保存

- 主動稽核內容正確性與修復損壞內容
  - LOCKSS peer-to-peer polling protocol
  - 使用雜湊演算法確保內容一致
  - 發覺損壞 (投票機制)，則從出版社或其他Boxes取得正確內容
  - 一般來說，同一份文件至少存放在七個Boxes (重複係數)



Archival Units

1055 Archival Units
11 awaiting recalc

| Volume | Content Size[1] | Disk Usage (MB)[1] | Peers | Polls | Status[2] | Last Poll |
|---|---|---|---|---|---|---|
| Academic Psychiatry Volume 30 | 93,400,185 | 103.9 | peers | 0 | Waiting for Poll | 00:14:49 10/22/06 |
| Accounting History Volume 5 | 3,125,086 | 5.4 | peers | 0 | 100.00% Agreement | 02:30:36 02/12/07 |
| Acta Sociologica Volume 48 | 7,992,857 | 11.4 | peers | 0 | 86.93% Agreement | 00:05:26 02/14/07 |
| Action Research Volume 3 | 6,443,633 | 9.6 | peers | 0 | 95.77% | 12:47:23 |

# 資料保存 (續)

- 打擊壞東西蓄意破壞
  - 壞東西必須滲透一定數量的Boxes方能影響結果
  - 聲望系統 – 投票勝的一方，其平均聲望必須在一定程度以上

# 傳播 (Dissemination)

- 提供機構使用者持續取用Box所存放的內容
- 兩種傳播方式
  - Transparent Proxy (通透性代理伺服器)
    - 整合機構內的Proxy基礎建設
    - 使用者不須知道LOCKSS Box的存在
  - Serving Content
    - 獨立可接受使用者Request的系統
    - 使用者需要知道LOCKSS Box的存在，或透過OpenURL Resolver

# 格式轉置 (Format Migration)

- LOCKSS只儲存原始(出版商)格式，不轉置成正規化格式
- 不去管格式轉置的問題，直到瀏覽器無法呈現為止 (Migration on Access)
  - 此時Box會動態產生一份瀏覽器可呈現的格式 (a temporary access copy)，並傳遞給使用者
  - 優點
    - 保存原汁原味、Overhead較低、轉換時採用最新科技

# LOCKSS運作模式

- 公共網絡
- 私有網絡
- CLOCKSS

# 公共網絡

- 用來保存全球圖書館社群感興趣要保存的數位物件(如電子期刊)
  - 由史丹佛大學統轄管理
  - 圖書館通常加入Public LOCKSS Network來保存他們有訂購的電子期刊 + OA 期刊
  - 平均重複係數為 40

# 私有網絡

- Private LOCKSS Network (PLN)
- 讓區域圖書館、檔案社群建立並管理其所欲保存的物件
- 通常用來保存特藏、數位化影像、區域網站
- 通常會有7-15參與機構
- 有自己的管理、資金、館藏發展、取用政策
- 範例：州政府公開紀錄(PeDALS)、社會科學資料(DataPass)、文化資產(MetaArchive)、電子期刊與電子書(CLOCKSS)

# CLOCKSS

- 2006年開展的前導計畫，由出版社、學術與研究圖書館、Stanford UL LOCKSS合作
- 學術性數位出版品的長期保存
- 社群治理
  - 商業出版社、大學出版社、學會出版社、圖書館
  - CLOCKSS 董事會建立政策與實施程序
- 圖書館保管與管理 (custody and stewardship)
  - 由全球、跨地域政治法律邊界的圖書館合作
  - 目前應該有超過100個圖書館參與
- 出版社共同參與
- 如何運作 (今天不細講)

# CLOCKSS (續)

- 取用 (Access) – 當觸發事件(trigger event)發生時，由董事會決定將內容開放(給每一個人)
  - 出版社停止運作
  - 出版社停止出版一期刊
  - 出版社不再提供過刊
  - 出版社平台產生悲慘性的持續當機
- 內容
  - 專論、研討會論文集、資料集、期刊
  - 可保存原始檔(如inDesign、ePub)或呈現檔(如PDF)
- 永續運作 – 會員會費(便宜)、捐贈

# 結語

# 結語

- 數位保存是近幾年學術圖書館發展趨勢或環境掃描的重要議題(ACRL Research Planning and Review Committee, 2012, 2014, 2015, 2016)
- 學術文獻、數位典藏、研究資料的數位保存
- 以聯盟、集體的力量建構我國數位保存基礎建設

# 參考文獻


- Abbott, D. (2008). What is Digital Curation? *DCC Briefing Papers: Introduction to Curation*. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation.
- ACRL Research Planning and Review Committee (2012). *2012 top ten trends in academic libraries*. Retrieved from http://crln.acrl.org/content/73/6/311.full.pdf.
- ACRL Research Planning and Review Committee (2014). *Top trends in academic libraries*. Retrieved from http://crln.acrl.org/content/75/6/294.full.pdf.
- ACRL Research Planning and Review Committee (2015). *Environmental Scan 2015*. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan15.pdf.
- ACRL Research Planning and Review Committee (2016). 2016 top trends in academic libraries. Retrieved from http://crln.acrl.org/content/77/6/274.
  Arms, C. R. & Fleischhauer, C. (2005). Digital formats: Factors for sustainability, functionality, and quality. IS&T Archiving 2005 Conference, Washington, D.C. Retrieved from http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf.
- Caplan, P. (2009). Understanding PREMIS. Retrieved from https://www.loc.gov/standards/premis/understanding-premis.pdf.

# 參考文獻 (續)


- Consultative Committee for Space Data Systems. (2012). Reference Model for an Open Archiving Information System (OAIS). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf.
- Corrado, E. M. & Moulaison, H. L. (2014). *Digital preservation for libraries, archives, & museums*. Rowman & Littlefield, Maryland.
- Digital Curation Centre (2008a). What is digital curation? Available at: http://www.dcc.ac.uk/digital-curation/what-digital-curation.
- Digital Curation Centre (2008b). The DCC curation lifecycle model. Available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model.
- Digital Preservation Coalition (2015). *Digital preservation handbook*, 2nd Edition. Retrieved from http://handbook.dpconline.org/.
- PREMIS Editorial Committee (2015). PREMIS Data Dictionary for Preservation Metadata. Retrieved from http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.
- Reich, V. & Rosenthal, D. S. H (2001). LOCKSS: A Permanent Web Publishing and Access System. D-Lib Magazine, 7(6). Retrieved from http://www.dlib.org/dlib/june01/reich/06reich.html.
- 歐陽崇榮(2006)。**數位資訊保存策略**。臺北市：文華。