電子檔案長期保存面面觀

柯皓仁教授兼臺師大圖書館館長
國立臺灣師範大學圖書資訊學研究所
中華圖書資訊館際合作協會理事長

內容大綱

- 何謂數位保存
- OAIS參考模型
- PREMIS與保存性詮釋資料
- 結語

何謂數位保存

數位保存

- 數位保存(digital preservation) 為一連串有系統、有管理的行動，以達成下列兩項目的(Research Libraries Group, 2002)：
  - 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

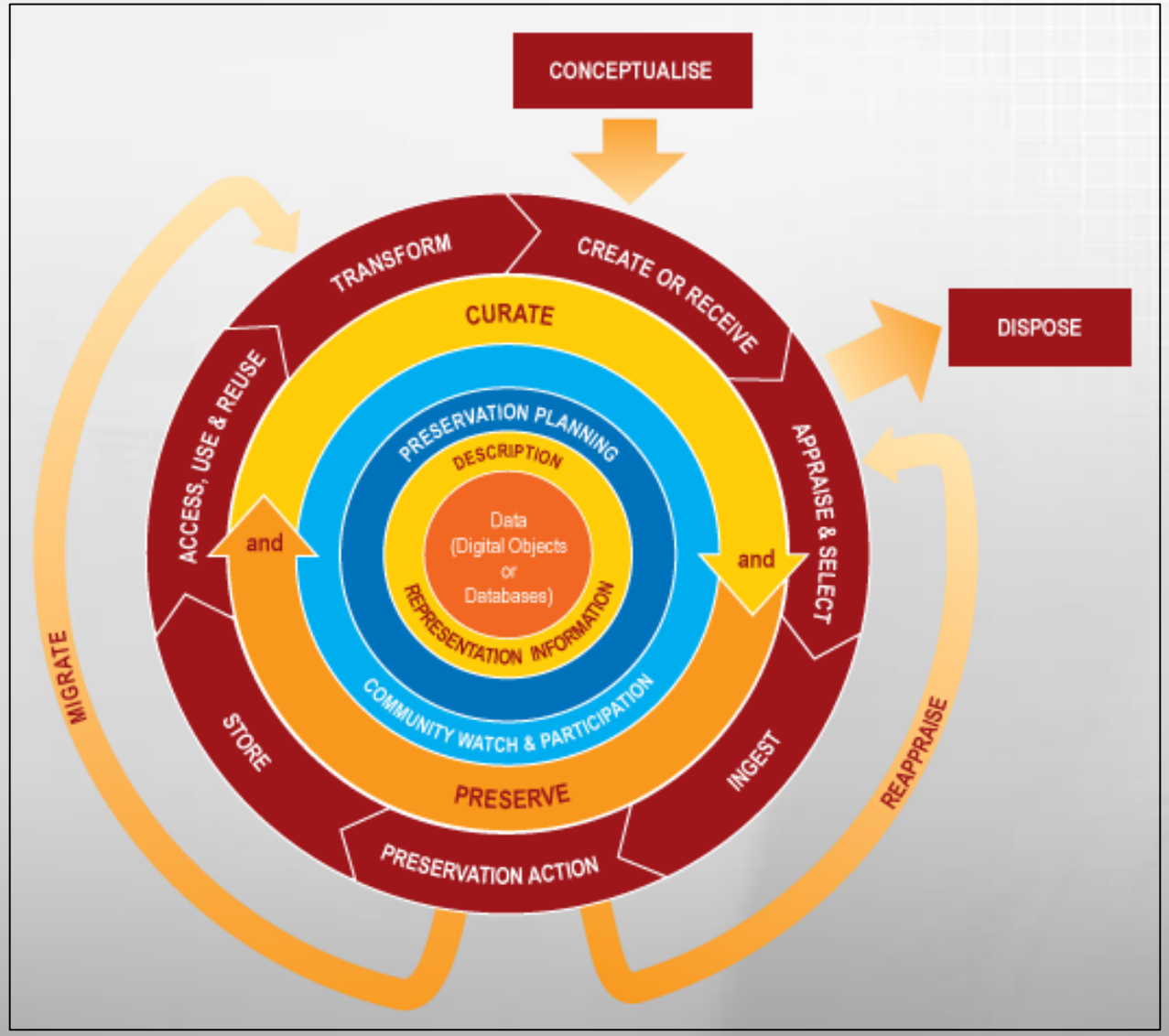
何謂「保存」

- 保存資料：位元(bit-level)等級保存
  - Refresh、Backup、Offsite-backup、Checksum…
- 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，甚至是文化與智能資產
- 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**

數位度用生命週期


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

數位物件保存概念

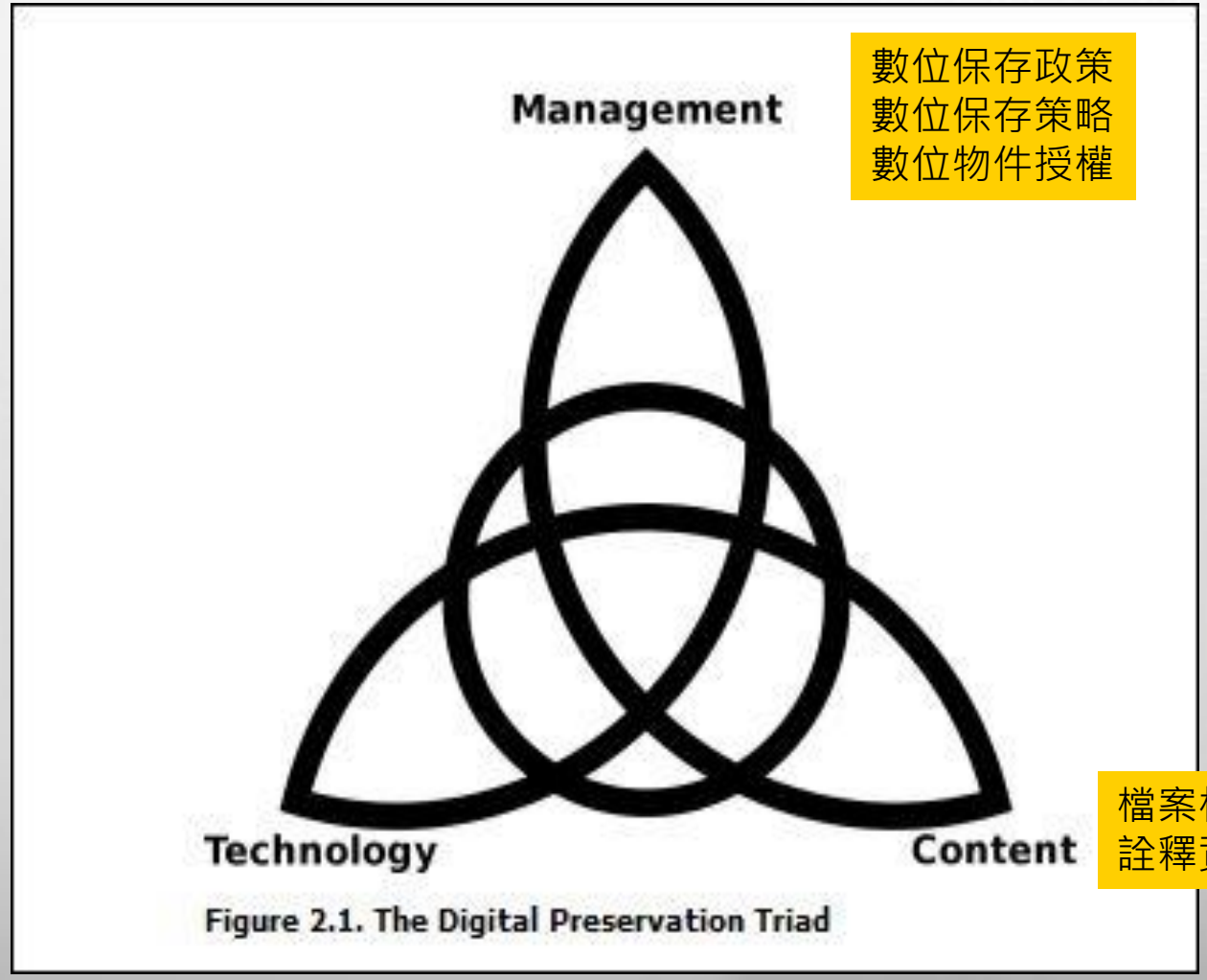


Figure 2.1. The Digital Preservation Triad

數位保存政策
數位保存策略
數位物件授權

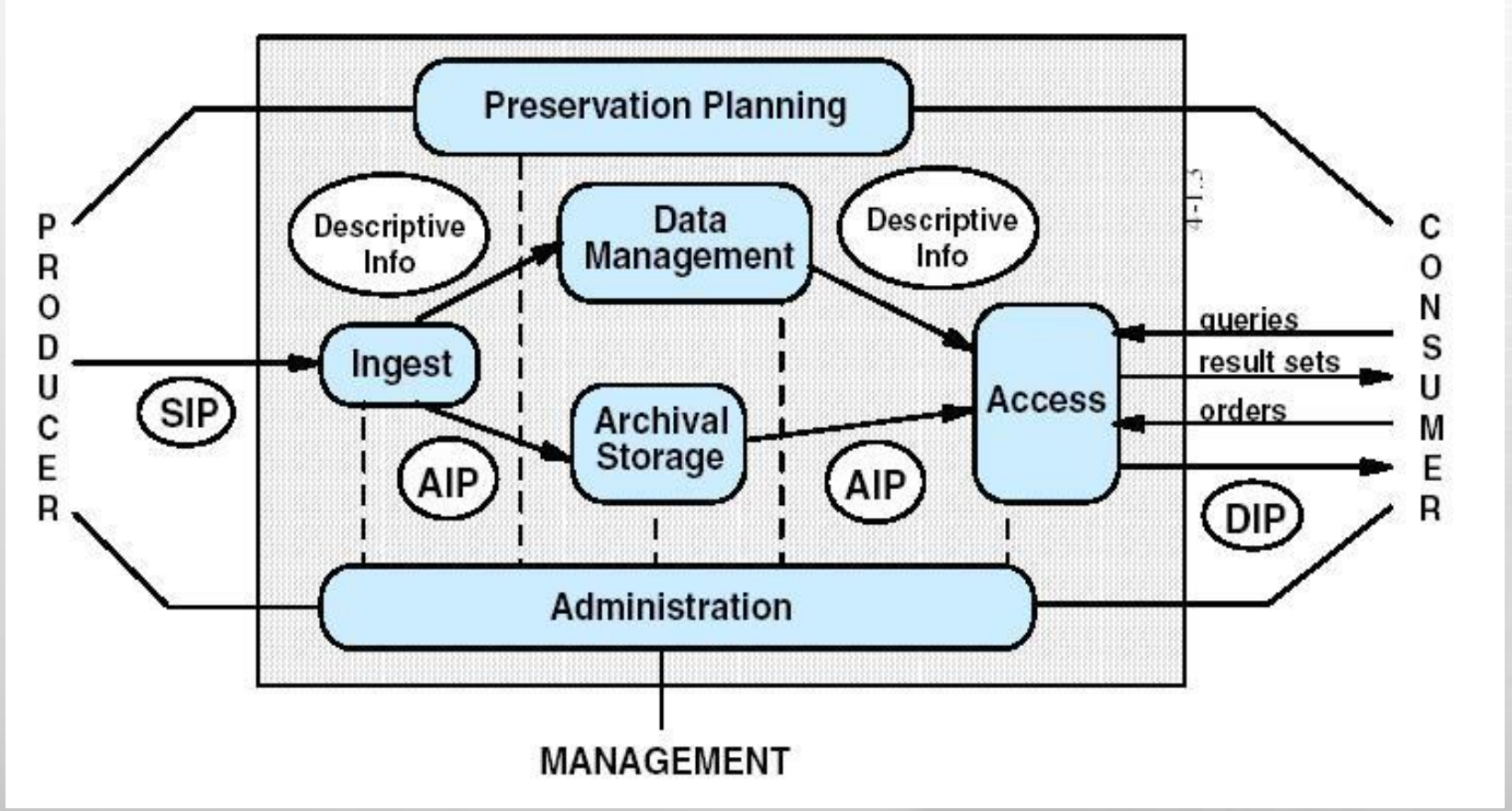
檔案格式、結構
詮釋資料

OAIS参考模型

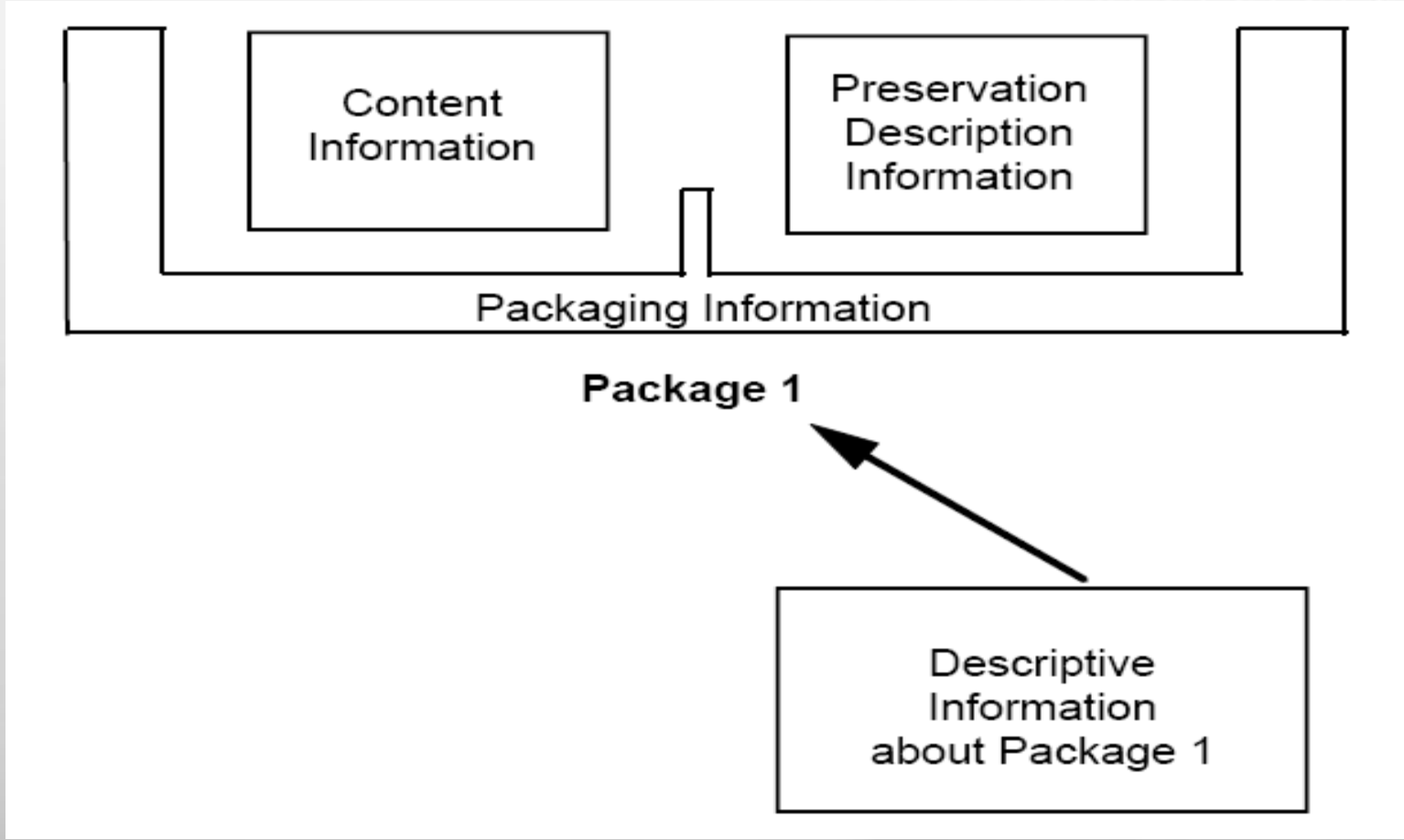
OAIS參考模型

- OAIS = Open Access Information System
- 一開始由美國國家檔案與文件署所主導，後來轉由美國太空資訊系統諮詢委員會發展
- 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
- OAIS參考模式已於2001年通過ISO標準
  - **ISO 14721:2012**
- OAIS 參考模式分成資訊模型及功能模型

OAIS Functional Model



Information Package Concepts and Relationships



PDI = 參考資訊、來源資訊、脈絡資訊、固定性資訊、取用權限資訊

PREMIS與保存性詮釋資料

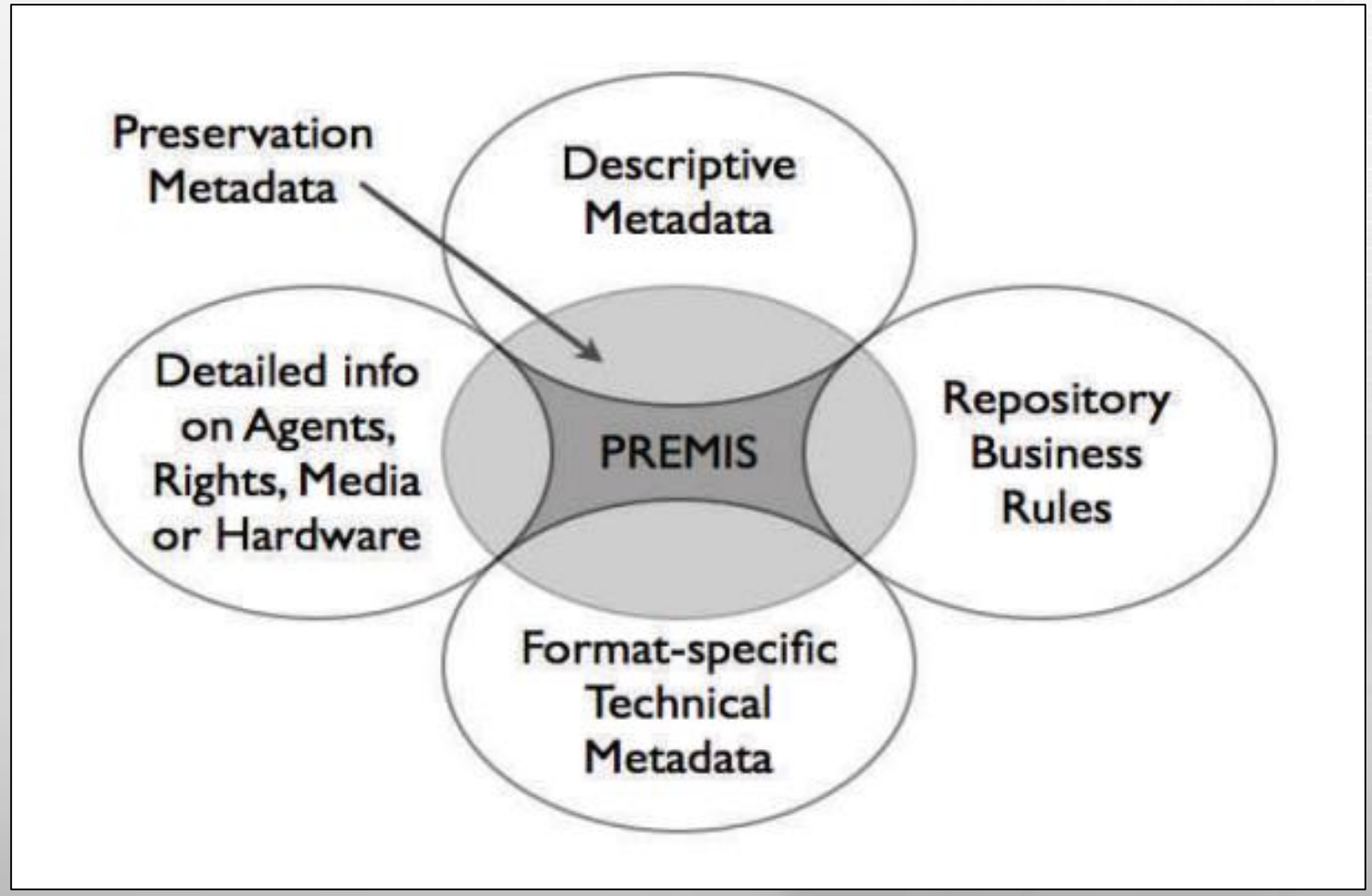
整體概念

- 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digital provenance (數位起源)，包含保管鍊(chain of custody)和授權變更歷史(authorized change history)

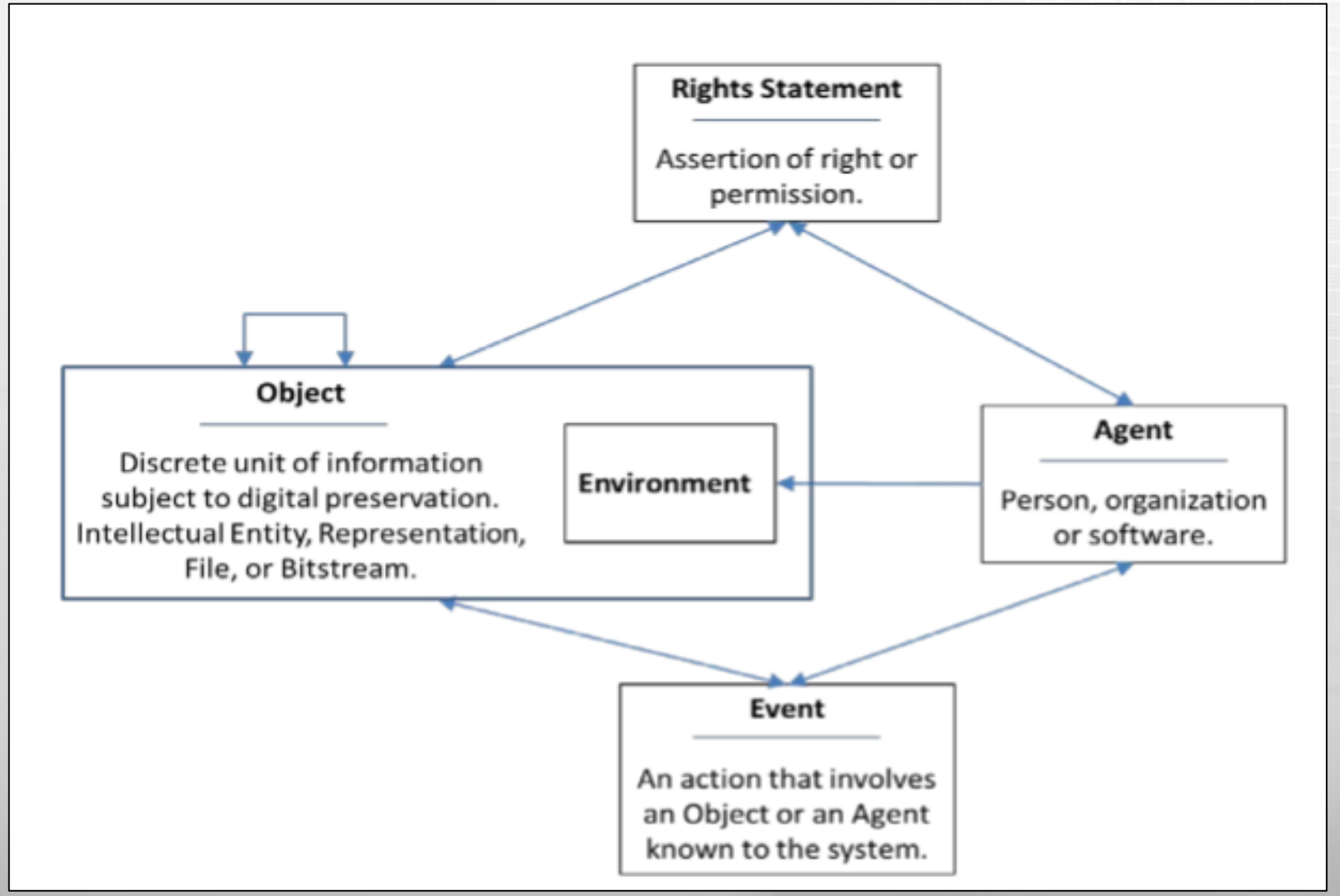
PREMIS

- 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性詮釋資料元素，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



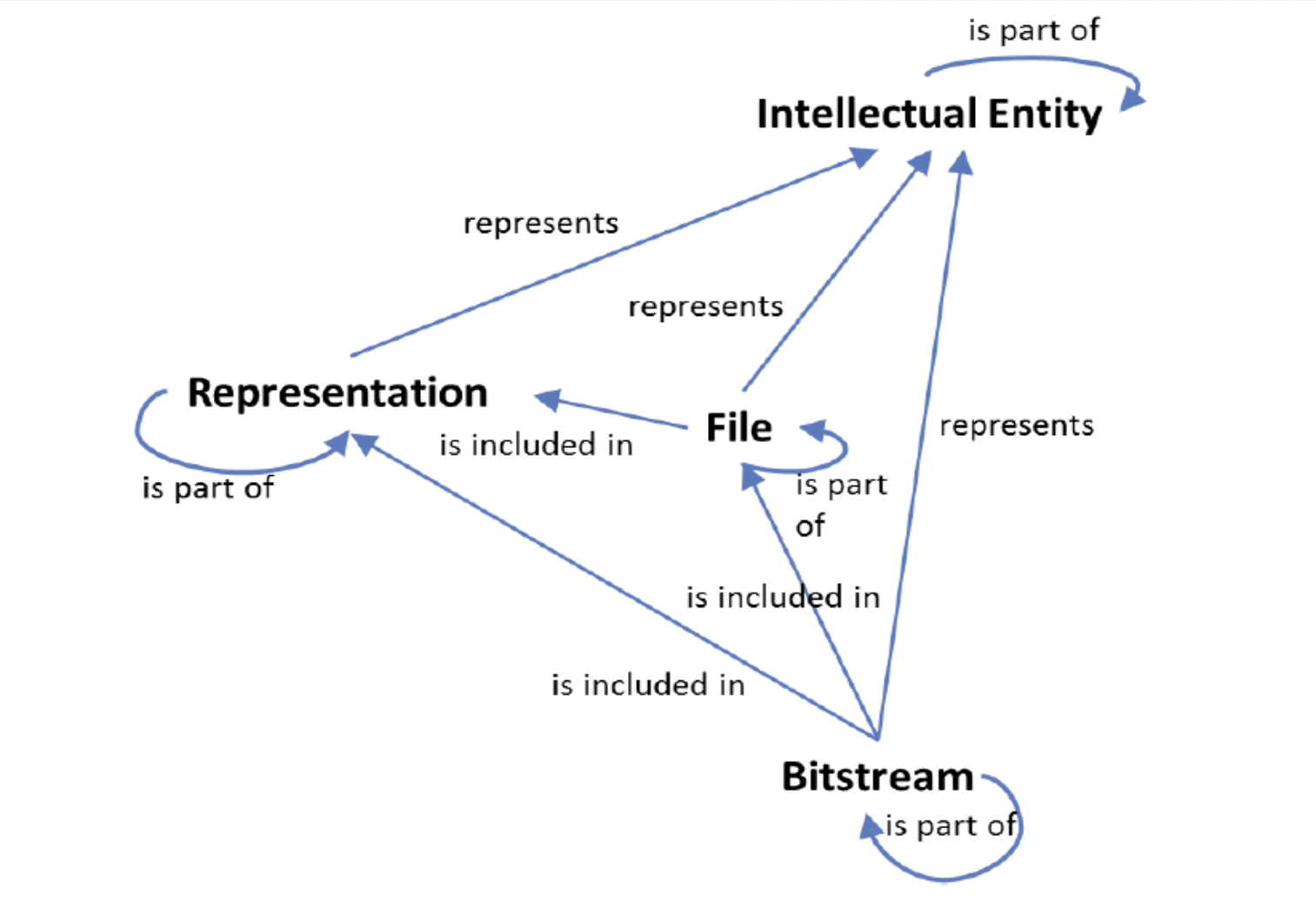
PREMIS的五種實體



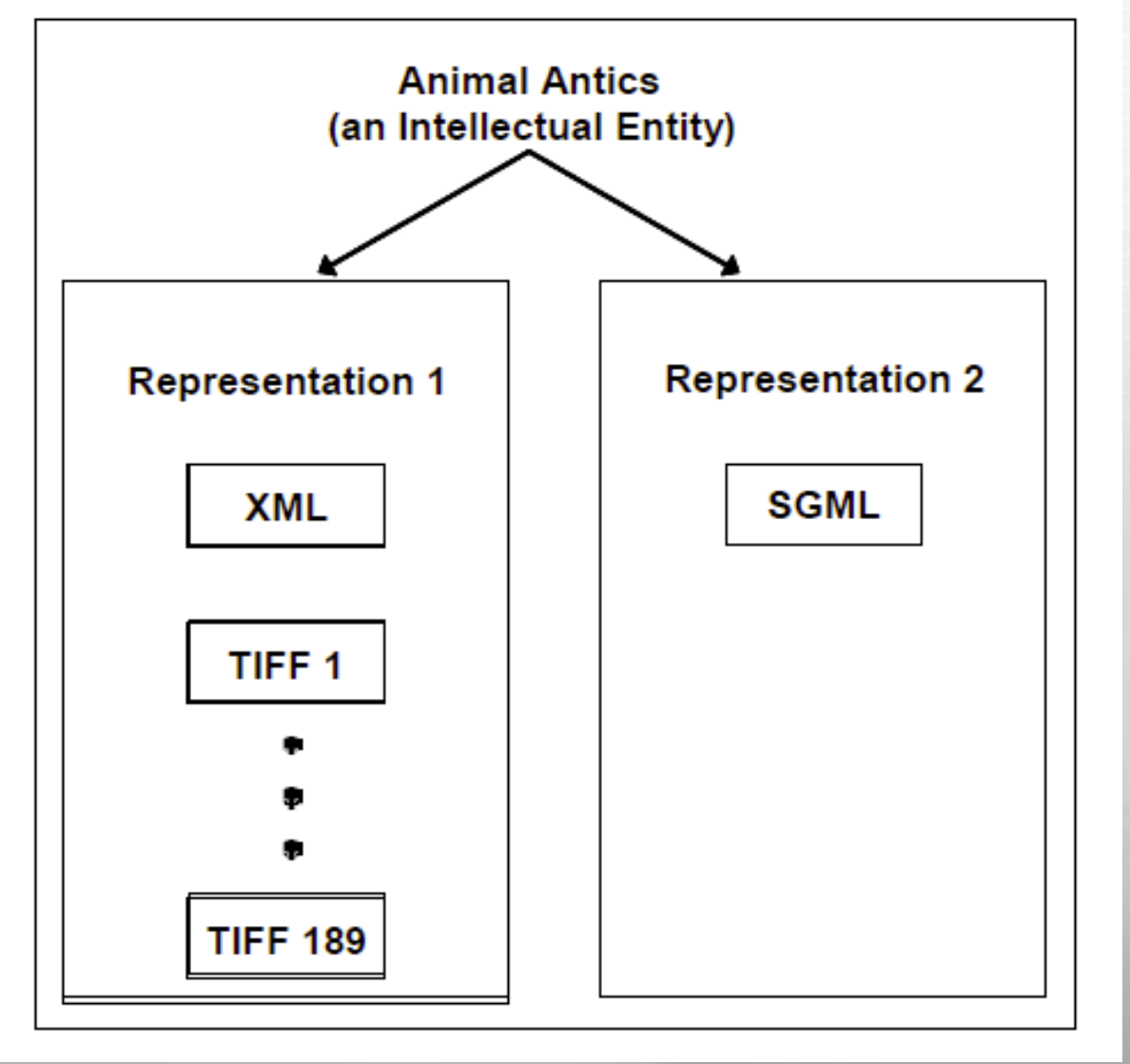
PREMIS實體

- 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 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. The *objectIdentifier* is repeatable in order to allow both repository-assigned and externally-assigned identifiers to be recorded. See "Creation/Maintenance" note above. 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. 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)**

有幾項需要被記錄的資訊

- PREMIS強調盡量能自動填入，但有些資訊必須想辦法記錄
  - Inhibitor: 可能導致無法取用、使用、轉置的特性
    - Password protection and encryption
    - Inhibitor type, target (the actions that are inhibited), and key (password or other mechanism to bypass the inhibitor)
  - Digital provenance:保管鍊和變更歷史
  - Significant properties：保存過程中必須保留下來的特性(字形？背景？格式？Look and feel…？)
  - Rights：智財權狀態、授權條款、特殊權限

結語

- 文化記憶機構應儘速制定數位保存政策
- 文化記憶機構應建構符合 OAIS 參考模型之數位保存系統並導入 PREMIS
- 提升國內文化記憶機構對數位保存的認知及對數位保存相關工作的熟悉度
  - 可用性、可識別性、可理解性、不變性、可行性、可呈現性、真實性等
  - 內容單位歸內容單位、技術單位歸技術單位？
- 建立全國數位保存機制