

從自然語言到資訊組織的再思考

從生成式 AI 的最小單位，理解資訊組織的新挑戰與新機會

國立政治大學
圖書資訊與檔案學研究所兼圖書館副館長
陳世娟教授

資訊組織的現況

從圖書館到 AI 時代，
資訊組織沒有消失，
而是成為數位世界的重要基礎。



第一個改變：組織的對象改變

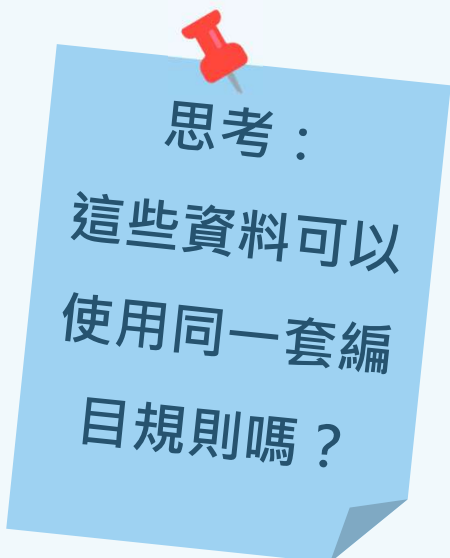
我們已經不再組織圖書，而是在組織世界

20年前_{的圖書館：}

- 圖書
- 期刊
- 地圖
- 光碟

今天圖書館的資料：

- 電子書
- 研究資料 (Data)
- ORCID
- DOI
- AI 生成內容
- 校史照片
- Facebook 貼文
- 網頁快照
- GitHub



思考：
這些資料可以
使用同一套編
目規則嗎？

資訊組織最大的改變

不是規則改了，而是**被組織的對象**已經完全改變。

資訊組織已從管理圖書，走向**管理各種數位資源與知識**。



第二個改變：管理關係

我們已經**不是管理書**，是在**管理關係**。搜尋「愛因斯坦」時，Google 提供的是：

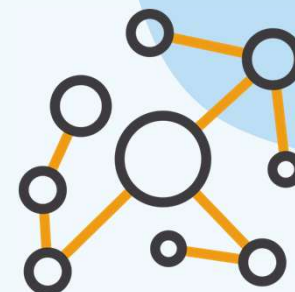
- 人物介紹
- 出生、死亡
- 配偶
- 論文
- 獲獎紀錄
- 相關人物

Google 不會告訴你**第幾排、第幾本書**；而是告訴你**愛因斯坦與哪些人、事、物有關**。

從書目到知識網路

過去

一本書 → 一筆書目



現在

人物 → 論文 → 事件 → 學校 → 照片 → 影片

全部建立關聯，這就是：

Linked Data

Knowledge Graph

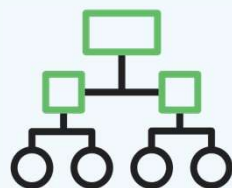
Wikidata

第三個改變：Google也是資訊組織

Google 是世界最大的資訊組織機構，每天都在：



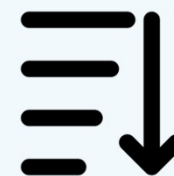
建立索引



分類資訊



建立關聯

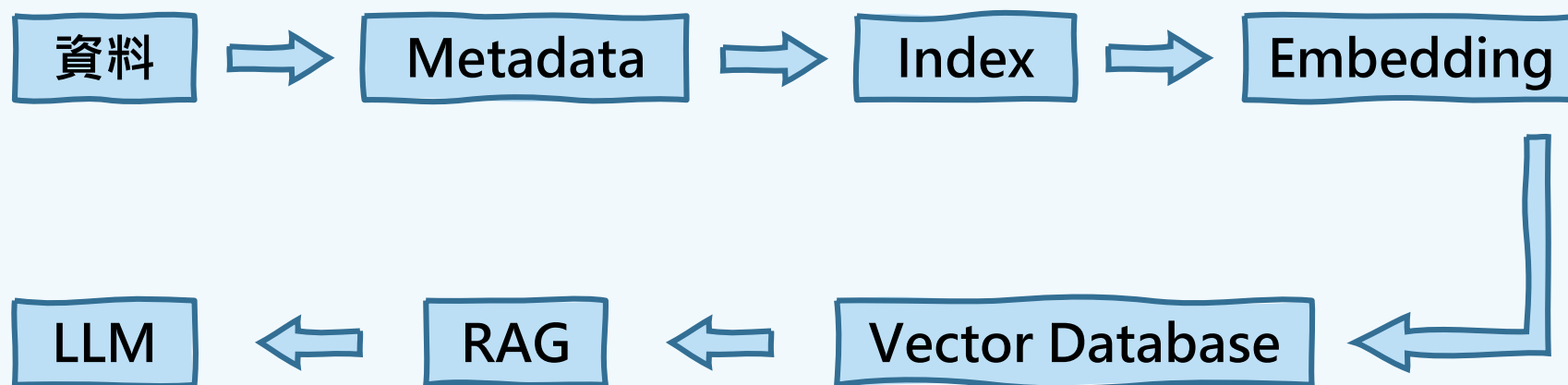


排序結果

不同的是圖書館使用**分類法**；Google 使用**演算法**，本質上都是資訊組織。

第四個改變：AI 更需要資訊組織

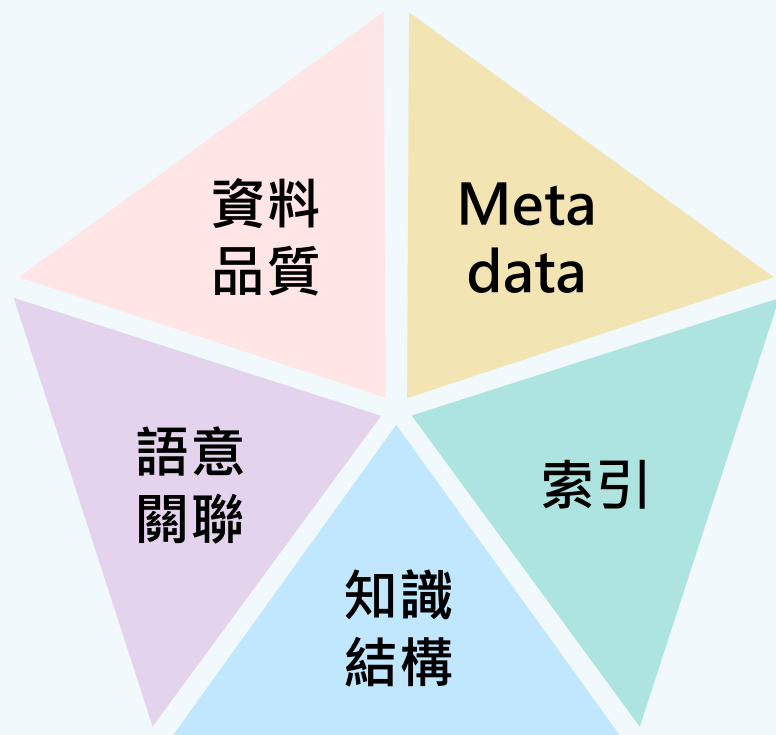
AI 沒有取代資訊組織，AI 的流程是：



AI 只是最後一個環節，前面的每一步都需要良好的資訊組織。

AI真正依賴的是資料整理

AI 的核心不是模型，而是資料，真正重要的是：



沒有良好的資訊組織，
就沒有高品質的 AI。

第五個改變：角色轉變

編目員沒有消失，只是換了一個名字。

20 年前

Catalog Librarian

現在

Metadata Librarian

未來

Knowledge Engineer、AI Knowledge Curator
、Ontology Designer

資訊組織人才的角色
正持續演變。

生活中的資訊組織

我們每天都在使用資訊組織，這些服務都建立在 **Metadata** 與**關聯資料** 之上。

Google Maps

- 地點
- 評論
- 照片
- 營業時間
- 交通資訊



Spotify

- 歌曲
- 歌手
- 曲風
- 情緒
- 推薦



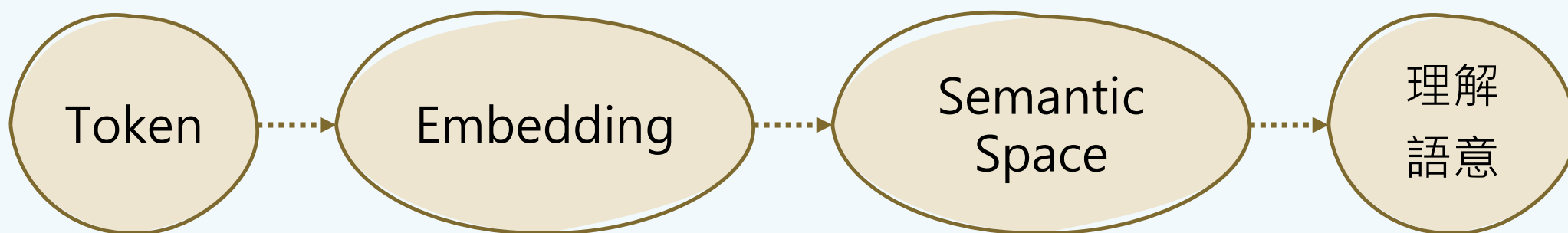
Netflix

- 類型
- 演員
- 導演
- 相似作品



ChatGPT 與資訊組織

ChatGPT 如何理解你的問題？



Token 只是開始，真正重要的是語意表示與知識組織

小結

資訊組織沒有消失，資訊組織已經從圖書館內部工作，變成整個數位社會的重要基礎設施，核心沒有改變，但



工具改變



資料型態改變

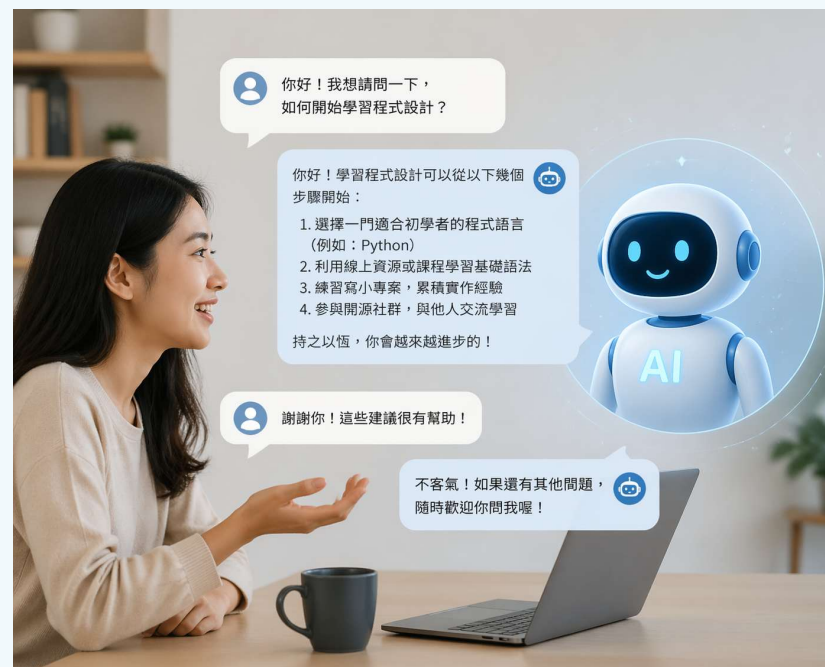


應用場景改變

Google、Wikipedia、Wikidata、ORCID、DOI、OpenAlex、Spotify、Netflix、ChatGPT，都建立在資訊組織的基礎之上。

你或許沒聽過 Token，但它正悄悄決定：

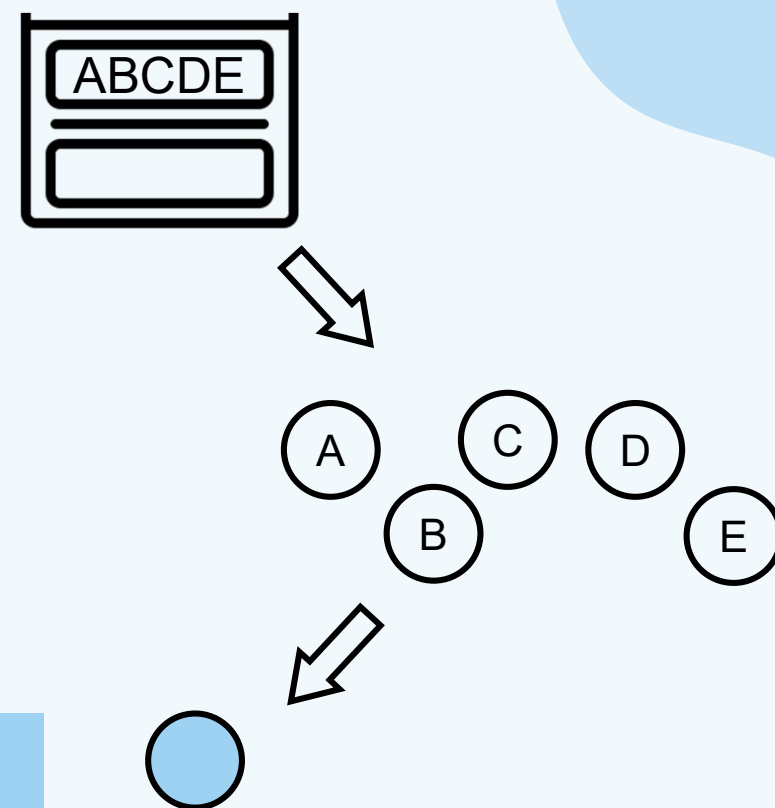
- 我們跟 AI 能說多少話
- 使用 AI 要花多少錢
- 一個流程能否順利跑完
- 模型能否完整理解文本



字粒宇宙：用比喻理解 Token

想像把一本長篇小說塞進一台超級翻譯機：

- 書頁不會被整頁閱讀
- 文字會先被切成一顆顆「字粒」
- 這些字粒，就是 token



Token 是 AI 處理語言時的基本顆粒

為什麼要理解 Token

越多 token，代表：

機器處理
更久

成本通常
更高

易碰到
長度限制

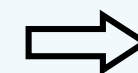
工作流程設
計須更精準

反之：

精準切割

巧妙重組

有效控制
輸入與輸出



- 提升效能
- 降低成本
- 穩定品質

今日 重點

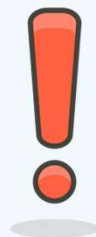
1. 什麼是 token
2. token 與自然語言的關係
3. token、關鍵字、權威控制的差異
4. token 如何影響圖書館應用
5. token 為什麼要花錢
6. 圖資界如何面對 AI token

AI 中的 Token 是什麼

在大型語言模型裡，token 是模型讀取與產生文字時的**最小單位**。

它可能是：

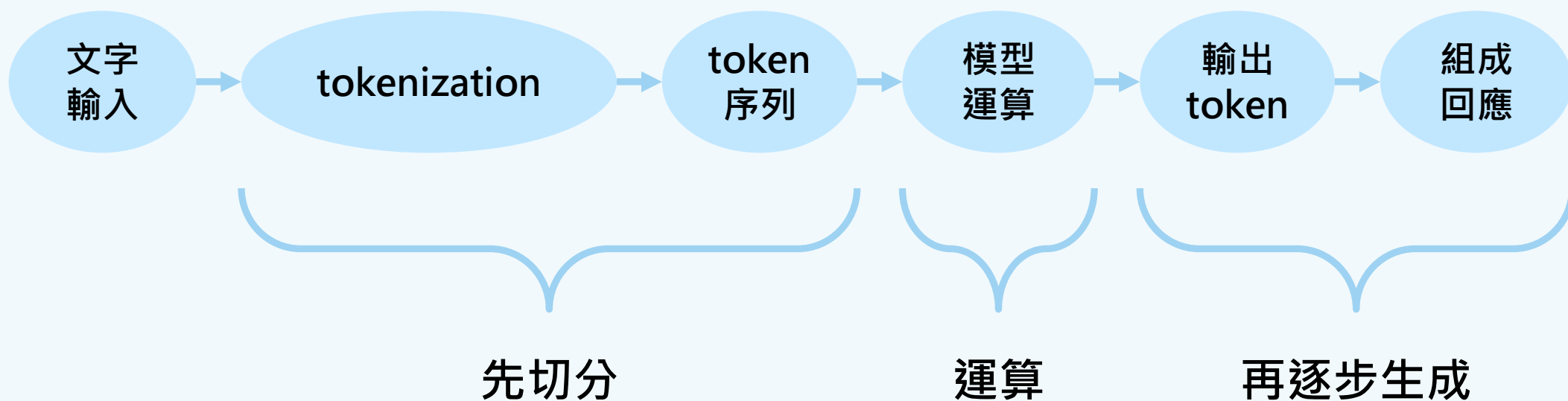
- 一個字母
- 一個字根
- 一部分單字
- 一個完整單字
- 中文中的單個漢字或字串片段



模型不是直接處理句子
，而是先處理 token

模型如何處理文字

AI 文本處理流程：



什麼是 Tokenization

tokenization = 把文本切成 token 的過程

在圖書資訊學中，可以類比為：

- 斷詞
- 文本前處理
- 主題分析前的切分
- 可分析文本單位的建立

圖資觀點下的 Token

從圖書資訊學角度看，token 可理解為：

- 機器的最小可處理資訊單位
- 類似索引前的切分單位
- 文本分析的基礎材料
- AI 進行資訊組織的起點

AI 概念	圖資對應概念
token	字詞切分單位 / 索引前單位
tokenization	斷詞 / 前處理
tokens 序列	可分析文本結構

Token 不等於語意單位



重要觀念

- token 是機器切分結果
- 不一定符合人的語意理解
- 不是天然的主題詞
- 不是現成的知識單位

token 是底層運算單位，不等於上層知識組織成果

Token 切分不是亂切

常見切分方法：

1. BPE (Byte Pair Encoding)
2. WordPiece
3. 其他子詞切分方法

特色：

- 常見字詞較可能保留完整
- 少見字詞較可能被拆開
- 切分依據偏向統計與頻率
- 目的是讓模型更容易計算

自然語言是什麼

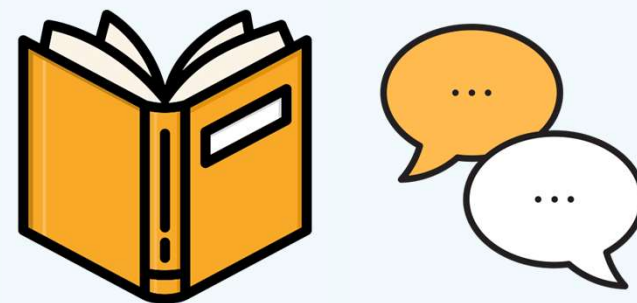
自然語言 (Natural Language) 是人類日常使用的語言形式

特性包括：

- 模糊
- 多義
- 不一致
- 依賴脈絡理解

在圖書館中常見於：

- 書名
- 摘要
- 目次
- 全文
- 讀者提問



圖書館長期在處理自然語言

圖書館面對的，不只是資料，而是語言問題：

- 同一概念有不同寫法
- 不同作者有不同表達方式
- 同一詞可能代表不同概念
- 書目描述與讀者查詢未必一致

這些問題，都會影響：

檢索

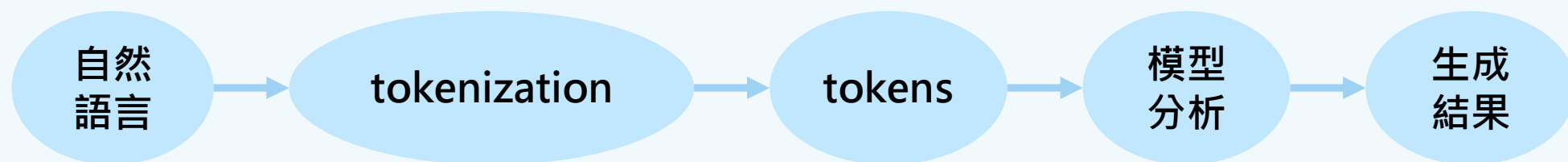
主題分析

知識組織

使用者理解

從自然語言到 Token

AI 面對自然語言時，第一步不是理解，而是切分。



範例：

「人工智慧在圖書館的應用」 → 人工 / 智慧 / 在 / 圖書館 / 的 / 應用

為什麼這件事重要

如果切分方式不同，可能影響：

- 模型看到的重點
- 關鍵概念是否被保留
- 上下文是否連貫
- 後續主題判斷與生成品質

自然語言如何被切成 token，會影響 AI 如何理解內容

什麼是關鍵字

關鍵字 (Keyword)

是從文本中抽取出來、能代表內容重點的詞彙

在圖資中常用於：

檢索

主題提示

內容辨識

初步分類

Token 與關鍵字的差異

Token 與 Keyword 並不相同：

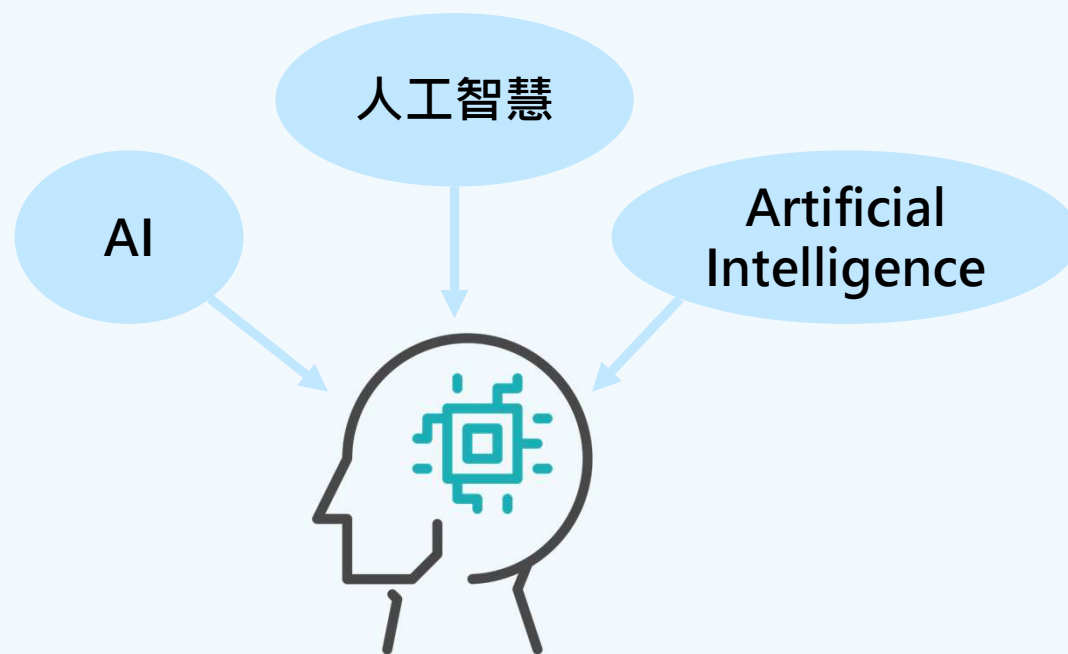
token	切分後的基本單位
keyword	從文本中挑出的重要詞

token 是原料；keyword 是挑出來的重點

關鍵字的問題

關鍵字檢索的常見限制：

- 同義詞問題
- 多義詞問題
- 表達不一致
- 語境不足



可能指同一概念，但字面並不一致

什麼是權威控制

權威控制 (Authority Control)

是透過標準形式來統一名稱與概念的作法

目的包括：

1. 建立一致表達
2. 提升檢索品質
3. 控制語意
4. 強化知識連結

權威控制在圖書館的重要性

權威控制可以處理：

- 作者名稱不同寫法
- 機構名稱變動
- 主題標目不一致
- 同概念多重表達

它的核心價值是：
把混亂語言整理成穩定知識

Token vs Keyword vs Authority Control

三者層次不同

層次	概念	功能
底層	token	切分與運算
中層	keyword	抽取重點
高層	authority control	統一概念與治理語意

token 負責讓機器能處理；authority control 負責讓知識能治理

整合流程：AI × LIS

從圖資角度整理成：

自然語言 → token → keyword → authority control

意義：

- 自然語言是原始材料
- token 是機器入口
- keyword 是中介表達
- 權威控制是知識治理

中文與英文 Token 的差異

英文

- 4 字元 $\approx$ 1 token
- 約 3/4 個英文單字

中文

- 1 個漢字 $\approx$ 1 ~ 1.5 個 token
- 實際會因字詞組合而異

結論：中文資料在某些情境下可能更耗 token

例子：中英 token 感受差異

原始詞句	被分解後
international	inter / national
圖書館	圖 / 書 / 館 or 其他片段組合

! token 切分方式，不一定符合直覺

Token 不等於字數

常見誤解：

- token $\neq$ 字數
- token $\neq$ 詞數
- token $\neq$ 句數

You can' t be perfect but you can be unique.

今天我共有8節課，你呢？

原因：

- ◆ 有些單字會拆成多個 token
- ◆ 有些字串會合併成一段
- ◆ 標點、數字、符號也可能佔 token

1 個 Token 大概多少字

★ 只能估算，不能精算 ★

英文：

- 約 4 個字元
- 約 3/4 個單字

中文：

- 約 1 個字 $\approx$ 1 ~ 1.5 token

⚠ 注意：這只是經驗值，真實情況仍需實測或工具計算

為什麼 Token 要花錢

因為每個 token 都需要計算資源

大型矩陣運算

GPU / TPU
叢集

電力消耗

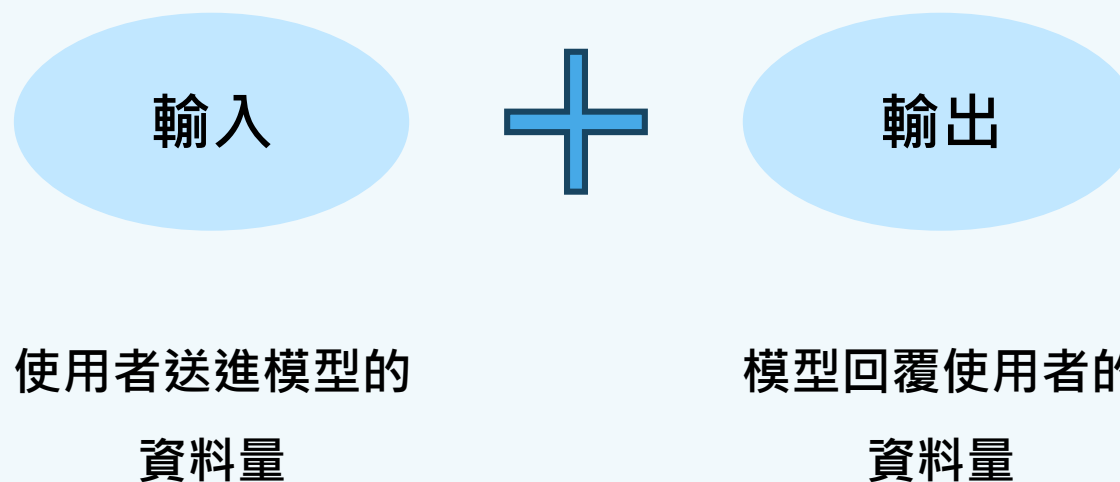
資料中心維護

網路頻寬

模型升級與
安全機制

token 收費是為雲端算力與基礎設施付費

Token 收費怎麼算



兩者皆會影響成本

ChatGPT / API 的 Token 計算方式

常見欄位	
prompt_tokens	輸入
completion_tokens	輸出
total_tokens	總數
用途	
• 計費 • 使用量統計 • 檢查是否超過模型上限	

Token 與上下文長度限制

模型有可處理資料量(文字長度)上限，當 token 太多時，可能出現：

- 內容被截斷
- 資訊遺漏
- 模型忽略前文
- 回應品質下降

token 不只影響費用，也影響模型能看到多少內容

圖書館資料為什麼容易遇到限制

圖書館文本常常很長：

書名

作者資料

摘要

目次

附註

主題描述

欄位規則

一旦全部送進模型，就可能同時增加：

- 成本
- 複雜度
- 超長風險



圖書館為什麼要關心 Token

只要使用生成式 AI，就一定會碰到 token，因為 token 會影響：

- 成本
- 速度
- 長度限制
- 輸出品質
- 流程設計

對圖書館來說，token 是技術問題，也是管理問題

圖書館可能的 AI 應用

應用場景包括：

自動書目摘要

主題標引建議

MARC 欄位輔助生成

問答系統

館藏內容分析

讀者服務輔助

共同點：都在處理文本，都會消耗 token

Token 對檢索品質的影響

token 切分方式可能影響：

- 關鍵概念是否保留
- 關鍵字抽取是否精準
- 主題判斷是否偏差
- 回應是否忽略重點

圖資語言可對照為：

底層切分品質，會影響上層檢索與主題品質



圖書館實務中的風險

常見風險包括：

- 一次丟太多資料，導致超長
- 欄位資訊太雜，影響模型聚焦
- 提示語不清楚，浪費 token
- 過度依賴模型輸出，忽略語意控制

AI 不是只要能用就好，還要能被良好設計與管理

實務建議：如何更聰明地用 Token

建議做法：

- 精簡輸入資料
- 分段處理長文本
- 控制輸出長度
- 明確設計 prompt
- 保留人工審核
- 結合權威控制與既有規範

token 不是只能被動承受，而是可以被策略性管理

Token 與人機協作

AI 擅長：

- 大量處理
- 初步整理
- 生成草稿
- 快速歸納

圖資專業擅長：

- 判斷語意
- 建立一致性
- 權威控制
- 品質把關

最理想的模式不是取代，而是人機協作

FAQ : AI 中的 Token 是什麼

AI 中的 token 是什麼？

- 在大型語言模型裡，token 是模型切分文字的最小單位。
- 它可能是一個字母、一部分單字，或在中文裡接近單個漢字。
- 模型先把輸入分成一串 token，再進行運算並依序產生新的 token 作為輸出。

FAQ：為什麼 AI 的 Token 要花錢

為什麼 AI 的 Token 要花錢？

原因包括：

1. **計算成本**：每個輸入或輸出 token 都要經過大量參數運算
2. **基礎設施維護**：需要資料中心、網路、電力、模型更新與安全機制
3. **服務營運成本**：供應商需回收研發與維運投入

用得越多，佔用計算資源越多，因此需要付費

FAQ : ChatGPT 的 Token 怎麼計算

ChatGPT 的 Token 怎麼計算？

OpenAI 常用統計方式：

`prompt_tokens` 、 `completion_tokens` 、 `total_tokens`

用途：

- 計費
- 使用量控管
- 模型長度限制檢查

輸入和輸出都會算

FAQ : 1 個 Token 大概多少字

1 個 Token 大概多少字？

- 英文經驗值：

1 token $\approx$ 4 字元 或 約 3/4 個單字

- 中文常見估法：

1 個漢字 $\approx$ 1 ~ 1.5 token

這只是粗估，實際情況仍會因文本內容而變化

本日重點整理

今天的主軸包含下列幾點：

自然語言

token

keyword

authority
control

代表意義：

- AI 從 token 進入文本
- LIS 從權威控制治理知識
- 二者需要整合，而不是互相取代

從圖資角度看 Token 的價值

Token 對圖資界的重要性在於：

- 是 AI 理解文本的入口
- 影響成本與流程
- 影響檢索與摘要品質
- 讓我們重新思考知識組織工作

理解 token 就是理解生成式 AI 的一個基礎邏輯

LIS 圖書館視角下的 token



token = 機器處理文本的最小單位

👍 類似於「索引詞切分單位」

中文 vs 英文的 token 切分

中文

1 字 ≈ 1 個 token

👍 沒有空格

範例：圖 / 書 / 館

英文

1 詞 ≈ 1-2 個 tokens

👍 長詞會被拆分

範例：International → inter / national

圖書館應用場景



書目摘要
自動擷取重點內容



主題標引
自動產生主題詞



MARC 欄位生成
輔助生成欄位內容



館員問答
支援自然語言查詢

token 多寡影響：👷 成本 ⚡ 速度 🎯 檢索準確度



結論

token 是 AI 的「字粒」，

也是生成式 AI 輸入、輸出的核心單位。

從圖書館的角度看，它不只是技術問題，

更牽涉自然語言、知識組織、成本管理與服務設計。

結語

在這個看似微小、卻牽動算力、成本與創意邊界的字粒宇宙中，
需要理解的不只是 AI 怎麼切字、模型怎麼計費，更是：

- 知識如何被處理
- 語意如何被組織
- 圖書館如何在 AI 時代持續發揮專業價值

讀懂 AI token，就是掌握生成式 AI 輸入、輸出的核心概念。



Thank you
for
listening!