

出國報告（出國類別：研究）

113年度交通部出國專題研究計畫  
短期氣候模式對於東亞梅雨季預報能力  
之評估研究——人工智慧資料科學技術  
於氣象資料整備的應用

服務機關：交通部中央氣象署

姓名職稱：吳蕙如技正

派赴國家/地區：美國

出國期間：113年9月1日至11月29日

報告日期：114年1月17日

## 摘要

氣象與氣候領域大量收集各式觀測、衛星、雷達資料，並產生天氣、氣候預報數值產品，已有數十年的時間，累積起極大量的氣象及氣候資料，並且隨著觀測技術、設備的進步，高速運算能力的增加，衛星、雷達、數值模式產品的解析度持續倍數提升，資料量亦以級數增加中。近年隨著機器學習及人工智慧等資料科學方法的發展，對於氣象及氣候資料的應用需求快速增加，如何整備資料、提供合宜的服務成為單位的課題，因此本次研究以氣象相關組織、單位的數項資料科學方法應用情形、對人工智慧就緒資料的要求、提供的資料服務與其背後的要求標準及系統架構為主題，並於此報告中記錄本次行程學習到的內容以及心得建議。

# 目次

摘要	2
目次	3
一、 目的	5
二、 過程	6
(一)、 氣象單位對人工智慧、機器學習等資料科學方法的運用	7
1、 氣象資訊服務引入聊天機器人 (CHATBOT)	8
2、 防減災工作運用 AI 與群眾互動	11
3、 數位雙生 (DIGITAL TWINS) 技術運用	14
4、 關於可信任的 AI 驅動天氣預報	18
(二)、 人工智慧就緒(氣象)資料 (AI READY DATA) 整備	19
1、 人工智慧就緒資料分級	20
2、 人工智慧就緒資料檢核清單	21
3、 人工智慧研究轉型為作業	30
(三)、 氣象資料庫	32
1、 FAIR 資料原則 (FAIR DATA PRINCIPLES)	32
2、 NOAA 的資料與資訊管理政策	38
3、 開放資料(OPEN DATA)	39
4、 NCEI 資料服務 – ONESTOP 及其他	44
5、 NCCF (NESDIS COMMON CLOUD FRAMEWORK)	51
三、 心得與建議	55
(一)、 關於對人工智慧、機器學習等資料科學方法的運用	55

(二)、 關於人工智慧就緒(氣象)資料整備 AI READY DATA	57
(三)、 關於氣象資料庫	59
<u>四、 參考文獻</u>	<u>61</u>
<u>五、 附錄</u>	<u>63</u>



# 一、 目的

近年來臺灣深受乾旱及洪澇事件影響，政府及社會各界對於應用天氣及氣候預報資訊在防災減災的需求日益增加，近年拜電腦算力及以資料科學為核心的人工智慧 (AI) 技術發展迅速，如何應用 AI 加速氣象、氣候測報技術以滿足社會、國家對災防的需求成為主要的課題。因此本次規劃至美國氣象單位訪問，了解天氣與氣候預報在人工智慧、機器學習等技術的發展與應用現況，亦瞭解資料驅動人工智慧和機器學習模型所需之相關資料匯集、整備工作 (AI ready data) 及服務的提供，如資料處理、品管、儲存與供應方式等，後續將納入本署國家氣象資料庫及資料供應服務等之資料經緯 (data fabric) 規劃中，以增加資料取得之便利性、加速預報研究及作業發展。

本計畫藉由拜訪美國國家海洋暨大氣總署 (National Oceanic and Atmospheric Administration, NOAA) 所屬之人工智慧中心 (NOAA Center for Artificial Intelligence, NCAI) 及國家環境資訊中心 (National Centers for Environmental Information, NCEI)，並與氣候及天氣資料整備、系統及服務設計之相關人員、學者進行交流與學習，瞭解美國引入資料科學方法，和資料服務中的資料整備作法，以期應用於規畫本署國家氣象資料庫及對內外部的資料供應服務，提供具有完整性、正確性與一致性的觀測與網格測預報資料服務，降低跨領域應用進入的門檻。使我國氣候監測與預報作業和研發的資料取得複雜度降低，進而有助於滿足社會大眾對氣象資訊及防災、減災日漸升高的需求。

## 二、 過程

梅雨季的降雨是臺灣在颱風季之前最重要的水資源來源，臺灣平均每年約有4至5道鋒面影響臺灣地區，有時甚至帶來強降水現象，此強降水往往會造成局部地區的暴洪、淹水、山崩、土石流等，造成重大災情與人民的損失。具有預報能力的預報資訊對於防災、農業、水資源管理等許多部門有重要的社會和經濟價值，倘若能有效提供準確且具有預報能力的梅雨季預報產品，除了可讓防災單位決策者可提前規畫與準備，讓災害損失降到最低，也可同時運用於水資源管理，提供水情調度之所需。

近十年來，隨著對於季內至季節尺度可預報度來源的逐漸瞭解和數值模式不斷的改進，使得模式季內至季節尺度預報技術逐漸提升。雖然科學家對於發展模式已付出了相當大的努力，但是仍然有許多科學問題還無法解決，模式於季內至季節尺度的預報能力相對於天氣尺度仍是相對較弱，預報水準仍不符合預報資訊使用者的需求，為提升預報技術，開發人工智慧降尺度或偏差修正等統計後處理技術，發展具有應用價值之短期氣候監測預報產品為未來值得努力的方向，配合人工智慧方法所做的資料整備則是相關發展的前置工作。

本計畫至國外先進氣象研究機構訪問，了解天氣與氣候預報在人工智慧、機器學習、資料探勘等技術的發展現況，並評估如何運用資料科學方法改善對東亞梅雨季的預報能力。同時，瞭解資料驅動所需相關資料匯集、整備工作 (AI ready data)，如資料品管，儲存與供應等，納入本署國家氣象資料庫及資料供應服務等之資料經緯 (data fabric) 規劃中，以增加資料取得便利性、加速預報研究及作業發展。主要為瞭解可能改善預報能力之資料科學方法，以及學習資料整備作法，可應用於規劃國家氣象資料庫及對內、外部的資料供應服務，增加資料取得便利性、加速預報研究及作業發展。因112年12月美方合作單位人員來署時，於交流活動中表示可能安排訪問國家環境資訊中心，考量瞭解美國政府單位相關業務運作於本署業務推動更有助益，爰計畫執行書研究機構調整為前往國家環境資訊中心 (National Centers for Environmental Information, NCEI)，並於3月獲交通部同意辦理。後因擬參訪單位美國國家環境資訊中心分4地辦公，該單位為釐清、協調受訪地點及人員持續討論，雙方討論決議仍以美國科羅拉多州波德為參訪地點，惟仍需時辦理行政流程，美方窗口建議延後於113年9月1日至11月29日前往，於113年5月申請延期出國，同月獲交通部同意回函，後於113年9月正式出發。

本次參訪行程整體停留在美國科羅拉多州波德 (Boulder, Colorado)，主要到美國國

家海洋暨大氣總署 (National Oceanic and Atmospheric Administration, NOAA) 位於美國商務部 (US Department of Commerce) Boulder Labs 的分部，拜訪在該處作為機構內部人工智慧 (AI) 技術節點的 NOAA 人工智慧中心 (NOAA Center for Artificial Intelligence, NCAI) (圖1左)、及做為資料節點的國家環境資訊中心 (NCEI) (圖1右) 相關成員，就該氣象單位對人工智慧、機器學習等資料科學方法的運用、人工智慧就緒(氣象)資料整備 AI Ready Data、氣象資料庫等主題進行了解，其中並參與6th NOAA AI Workshop (圖2、主題及議程請詳附錄)，各主題之學習所得整理、記錄如後：



圖1、NCAI 及 NCEI 的重點工作分別是機關業務的技術及資料節點



圖2、6th NOAA AI Workshop 實體活動會場 Rembrandt Yard 外部及內部影像

## (一)、 氣象單位對人工智慧、機器學習等資料科學方法的運用

人工智慧的快速發展為包括氣象科學在內的各個領域帶來了全新變革，AI 正在改變科學家收集、分析和解釋資料的方式，不過在氣象領域主要仍停留在研究發展階段，雖然部分技術已在相關環境中得到驗證，並展現出發展潛力，但要完整進入作業應用仍有距離，還需要付出大量努力。AI 在地球科學領域最引人注目的一點是 AI 看來能夠處理領域內多年來監測環境變化和研究自然災害積累的大量資料集，透由機器

學習和深度學習進行複雜、耗時的氣候建模，並且產出的結果正在逐步提高準確性和效率。

近幾年流行的通用生成式 AI 模型，如用於交談寫作的 ChatGPT 和用於圖像創建的 Midjourney，這些模型透過揭露不同資料集間的關係和歷史資料中的訊息模式，展示出令人興奮的潛力。AI 透過分析廣泛的資料集、簡化任務和揭露隱藏的關係，看起來將會徹底改變氣象科學方法，促進曾經無法實現的發展。

未來除了為特定主題專門訓練的 AI 模型，或用通用模型整合多樣化資料集各種作法，也要考慮如何不斷學習與更新以適應新的資料和環境變化、如何監控 AI 行為以確保在實際應用中的準確性和可靠性、如何符合倫理與安全要求以確保技術的使用不會對社會和環境造成負面影響 (圖3、Ethical AI)。

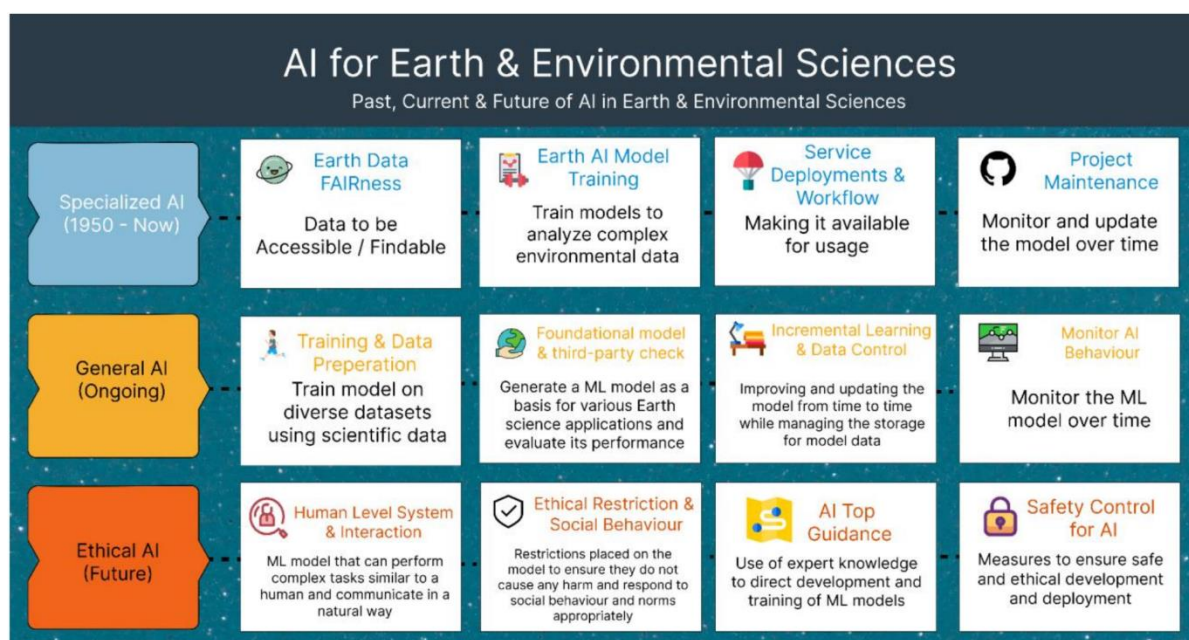


圖3、地球科學領域 AI 應用的過去、現在與未來

## 1、氣象資訊服務引入聊天機器人 (Chatbot)

### Geospatial Chatbot

NOAA 對於引入生成式 AI 提供服務正在進行內部嘗試。其中一個項目與 Amazon 合作，以 Bedrock 知識庫提供支援的生成式 AI 地理空間聊天機器人測試 (圖4)，希望能作為科學家的乾旱研究助手。為建立相應檢索增強生成 (Retrieval-Augmented



Generation, RAG) 模型，目前輸入了 National Weather Service 近年的年度乾旱報告 (圖5左)，藉由模型解析全美各地各年月的乾旱資訊，呈現介面設計在 Amazon Location Service 視覺化平臺，搭配上 Prompt Engineering，提供用戶查詢 (圖5右)，使用者可以輸入各種查詢條件，包含地理區域、年度、乾旱程度、影響人口等，例如：“Which part of the Pacific Northwest were in extreme drought?”、“Can you draw circles around those localities base on the population.” 平臺不只以文字來提供報告中的符合資訊，更在地圖介面呈現被詢問的地理區域及標記相應的位置及範圍，抽取、組合報告內容回應各種應用性查詢。

生成式 AI 被用來回應各項問題查詢是趨勢，但對政府部門的首要挑戰項目則是一致的：讓系統學會「不知為不知」；“If data is not included, it should answer don't know.” 因為必須要對提供的資訊負責，在能夠完全確定系統不會以 AI 幻覺回應出貌似事實的虛假或誤導性資訊之前，服務只能在內部進行測試，不適宜對外服務。同時，發展者也注意到在地球科學領域，image-to-text translator，以圖形作為輸入，讓系統解讀圖形，然後能依圖形內容回答問題的後續發展需求。

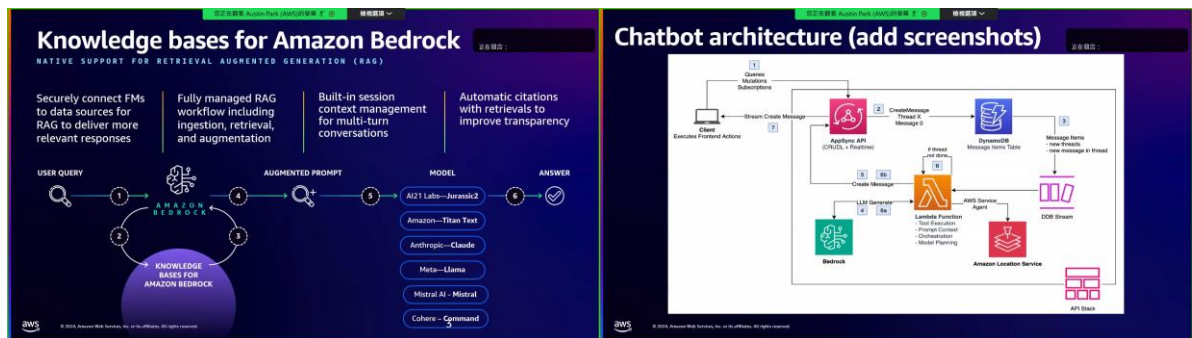


圖4、NOAA geospatial chatbot 使用的 Amazon Bedrock 技術及系統架構圖

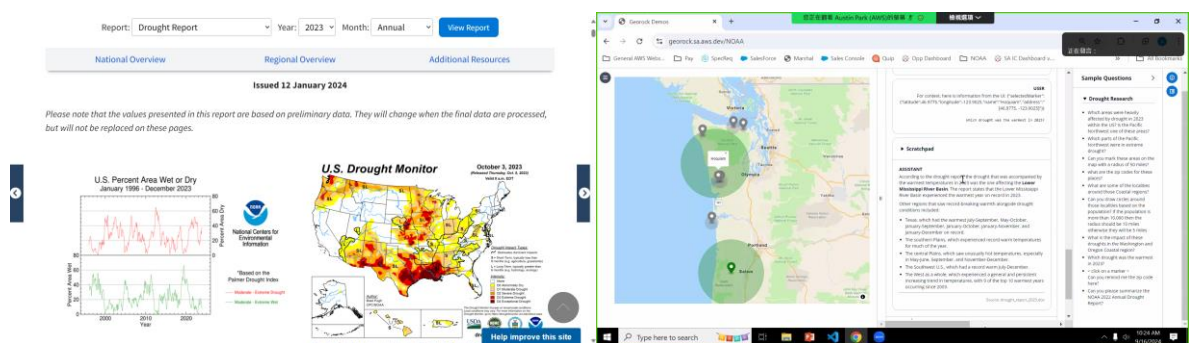


圖5、Annual Drought Report 2023 (公開資訊)及 NOAA geospatial chatbot 介面(發展中)

另一個案例中，歐洲中期天氣預報中心 (European Centre for Medium-Range Weather Forecasts, ECMWF) 為了讓使用者有效檢索其大量的文件、數百 PetaByte 的資料、查找數千種不同的計量變數、減少使用者在眾多選項中的尋找困難，嘗試運用聊天機器人解決問題，發展了 ECMWF Chatbot，允許使用者以自然語言描述個人感興趣的資訊，利用對話方式透過介面互動，Chatbot 在適當的時候會詢問更多細節，最後結合適當的氣象圖資、從 ECMWF 知識庫、或從 GitHub 儲存庫中檢索的文本答案來回答問題。

Chat@ECMWF 實作在 ChatGPT 上，採兩大方向處理查詢，一類做法運用檢索增強生成 (RAG) 來處理使用者查詢，引用各項文件與說明，並將其向在向量資料庫中編入索引，大型語言模型 (Large Language Model, LLM) 再用此來提供說明和提供相關信息 (圖6)。另一類做法則建立使用者訪問行為的推理和行動 (Reason & Act, ReAct)，交付 LLM 的任務是從預定義的集合中選擇最合適的工具，並確定該工具的相關參數 (圖7)。ReAct 的工具包含 API，用於檢索例如指定日期在給定位置的點預報資訊。

RAG 及 ReAct 作法通過使用 LLM 的功能，解析內容：若詢問知識性內容，則導向 RAG，以文件內容回應；若詢問天氣等資訊，則分析被詢問的地點，而後將使用者的請求轉換為 ReAct API 的輸入參數，進行查詢以獲得所需的資料數值。並且，雖然文件主要以英文寫成，系統可善加利用 ChatGPT 的多語言功能來回答使用者來自 ECMWF 各成員國不同語言的問題，例如：“What is the role of ECMWF?”、“I’m studying cyclone Katrina. How can I download related data from the CDS?”、“Plot me a map of wind over France.”。

整體的對話操作上，以 LLM 解讀使用者問題後，運用 prompt engineering 進行答覆者的立場設定、及素材選用範圍限制，在需要呼叫 API、查詢即時資料的項目上，則篩選出合適的工具、以合適的參數進行查詢，再交由 LLM 進行回覆。等於是在使用者的提問之上，於後臺發展出一系列與 LLM 的對話標準，來妥適的提供回答。但也因為這樣的設計，減緩了回答的速度，也增加了花費 (ChatGPT 使用費用)；同時，LLM 的幻覺仍是無法避免，並且不論進行了多少避免不當答覆的調整，在 LLM 模型版本更新後，因為幻覺答覆特性改變，全部訓練、調整必須重新來過，這是運用生成式 AI 仍待解決的問題。

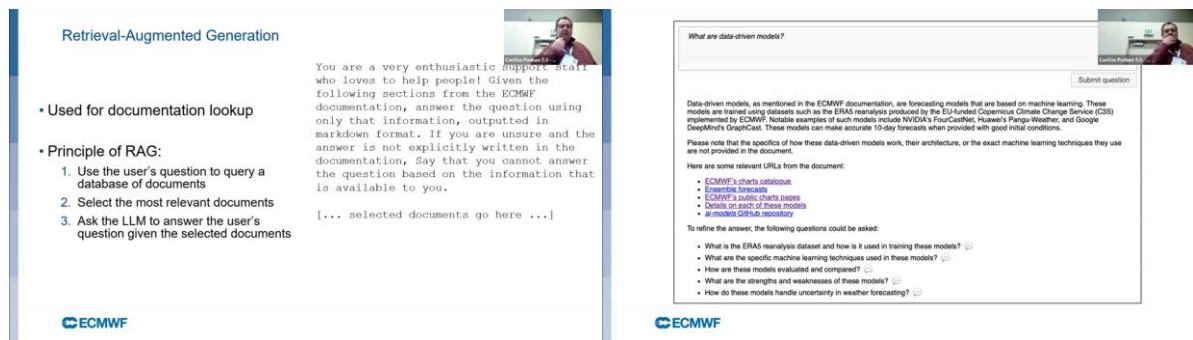


圖6、Chat@ECMWF 運用檢索增強生成 (RAG)來處理使用者查詢的原則與範例

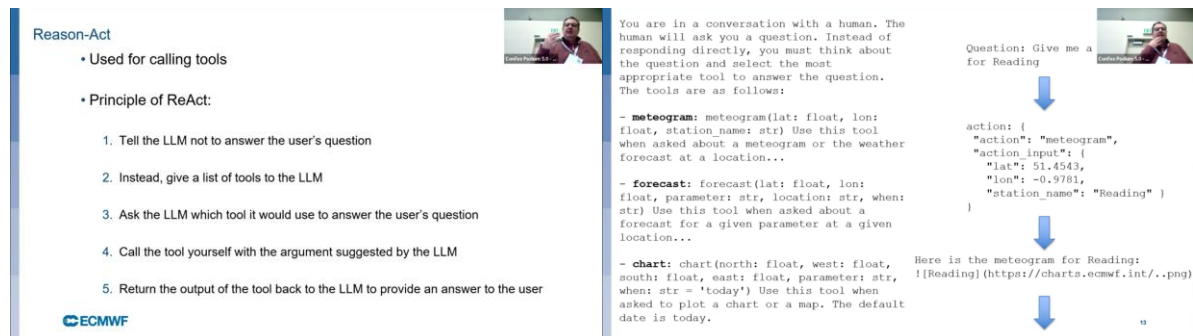


圖7、Chat@ECMWF 運用行為推理和行動 (ReAct)來處理使用者查詢的原則與範例

## 2、防減災工作運用 AI 與群眾互動

### NWS Translation

NOAA 為了實現其「天氣就緒國家」(Weather Ready Nation) 的戰略願景，針對 2500 萬英語能力有限的美國民眾，提供多語言防災訊息發布，自2023年10月開始，國家氣象服務 (National Weather Service, NWS) 上架實驗性的翻譯網站 (<https://www.weather.gov/translate/>)，使用人工智慧技術將預報、警告翻譯成西班牙文和中文。提供服務的目的是“在危險風暴來臨之前，以多種語言及時提供天氣警報，有助於確保每個人都能獲得可能挽救生命的資訊”。(圖8、圖9)

過去 30 年來，NOAA 的 NWS 一直以人工翻譯方式來提供西班牙文天氣預報和警告，但經過幾年來的系列測試專案，NWS 預報員訓練人工智慧軟體使用西班牙文和簡體中文的天氣、氣候、水文和防災術語，並規劃將在未來添加更多語言。利用 AI 技術的進步，這個語言翻譯專案將提高對英語水準有限的弱勢群體的服務公平性，通過提供多種語言的天氣預報和警告，NWS 提高社區和個人的準備程度和復原力，保護更多人免受美國惡劣天氣風暴的影響。



為了完成此一翻譯服務，NWS 與機器學習公司合作進行大型語言模型訓練，科學家用天氣術語訓練語言模型，確保翻譯符合該語言文化和可理解性，例如，代表守視或注意的“watch”，在一般翻譯工具中會被翻譯為“鐘、錶”，代表陣雨的“shower”，則會被譯作“淋浴”等。

過往的手動翻譯受限於人力，難以即時對發生的劇烈、危害天氣發布不同語言的警告訊息。引入 AI 工具後，國家颶風中心風暴將產品翻譯為西班牙文所需的時間從 1 小時縮短為不到 10 分鐘(或更短)，使預報員能將時間放在預報和決策支援服務的核心職責上。這些自動翻譯的產出的成果，符合通用警報協定格式，可提供給媒體運用，已在社群媒體上以多種語言自動發佈。

因為有與一般大型語言模型不完全相同的需求，NOAA 氣象服務翻譯專案並未直接使用市面上 LLM 服務，NOAA 選擇在其作為製作和傳播天氣警告和警報的關鍵平臺—進階天氣互動式處理系統 (Advanced Weather Interactive Processing System, AWIPS) 上，引入 AI 自動語言翻譯專案工具，增加語言支援服務。由翻譯試點專案構建的服務已經進入驗證階段，在保持高翻譯準確性的同時，藉由系統運作功能的整合，可以為工作人員節省通知製作到發出的關鍵時間，並維持訊息的一致性，長期目標是為提供更多英語能力有限的民眾易於取得的天氣資訊，開發更多種語言的自動化、跨語言一致性 NWS 產品即時翻譯。

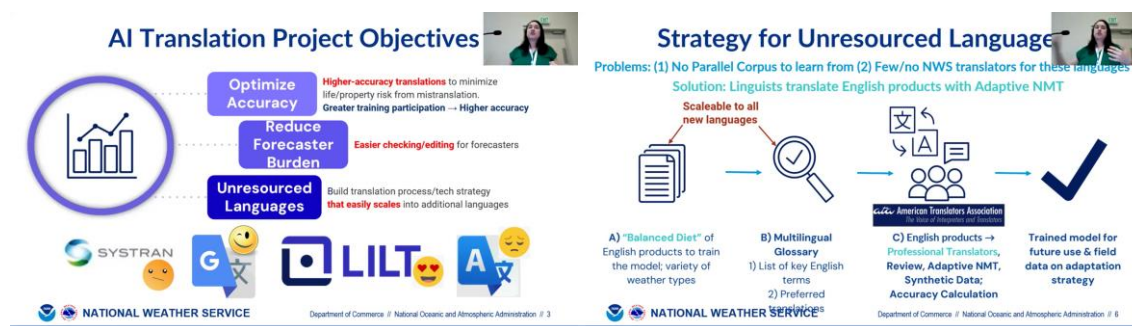


圖8、NWS Translation 的專案目標及策略

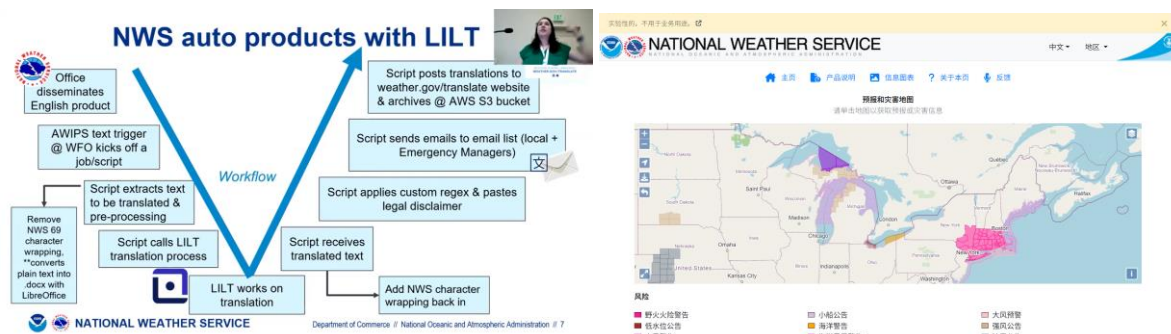


圖9、NWS Translation 的工作流程及服務網站介面



## 群眾外包資訊收集

當季節性的強降雨使土壤含水飽和、河川和溪流水位升高，或當天氣警特報發布，需要評估進行道路封閉、發生淹水或土石流的風險時，因為美國幅員遼闊，測量洪水水位等的儀器難以隨處佈建，各地社區內面對的氣候風險究竟是什麼？當社區的積淹水風險如何被監測成為議題，考慮以人為中心的 AI 應用 (human-centered AI)、找到可靠、有效率又簡易的風險溝通方式，群眾外包資訊收集 (Crowdsourced Data) 變成美國民間提出的解方。

例如，ISeeChange 組織結合群眾外包，由民眾回饋積淹水訊息的平臺設計，再以 AI 結合官方天氣預報的內容，一方面可以透過即時回報之訊息了解社區何處發生狀況，另一方面則透過將兩類資料融合、分析、處理，可以了解氣候風險所在，回饋給地方政府和公用事業公司，依資訊來進行緊急疏散、救援資源部署，或維護、管理和保護、強化基礎設施，應對和防止極端氣候變化引起的損害。(圖10)

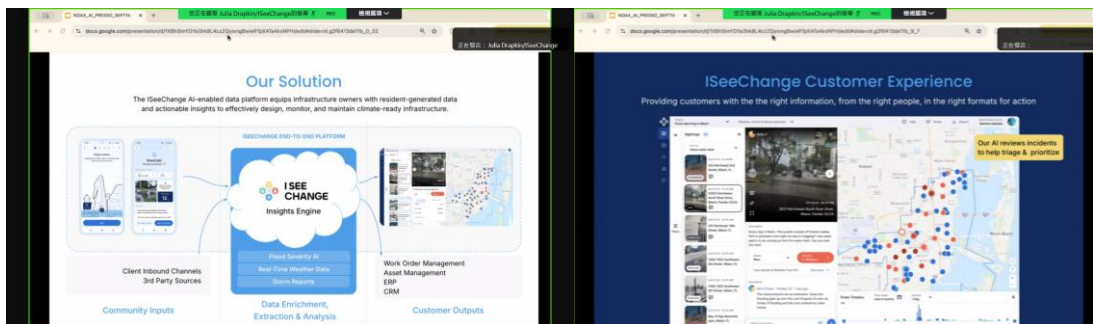


圖10、ISeeChange 提出的解決方案概念圖及提供民眾與客戶訊息的介面

而 Amir Behzadan 教授與團隊則在 BluPix app 計畫中，針對水位量測的設施分布有限，無法完整掌握各社區淹水深度資訊的困境，提案以市區隨處可見的“Stop Signs”為參考標準，結合 google 街景圖、對群眾提供的淹水影像進行即時 AI 分析，對比交通標誌的固定尺寸及設置標準，可以快速估算出各地方的積淹水高度，即時傳達給民眾、也提供縣市政府安排處理與救援。(圖11)

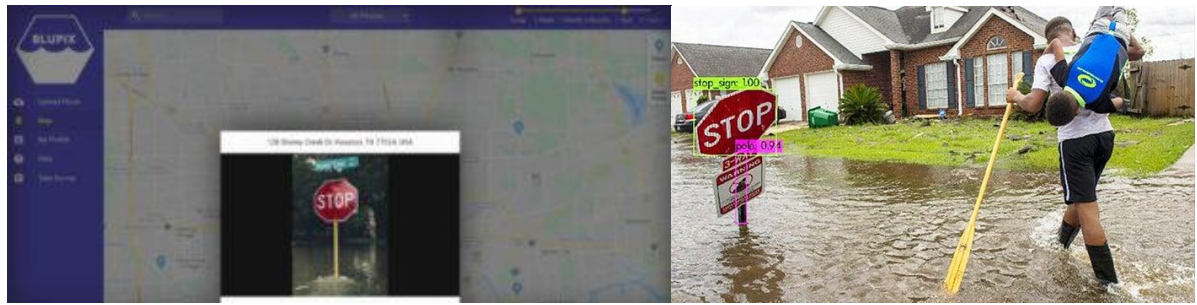


圖11、BluPix 的 crowdsourced 資訊回報呈現地圖介面及淹水高度計算

### 3、數位雙生 (Digital Twins) 技術運用

#### Washington DC 熱島虛擬實境

案例之一是 Urban Heat Island of Washington DC Virtual Reality (VR) Experience (華盛頓特區城市熱島的虛擬實境體驗)，由美國 NIHHS (National Integrated Heat Health Information System，國家綜合熱健康資訊系統) 與 NOAA 團隊 (NOAA Global Systems Laboratory 全球系統實驗室 + Office of Education / Science on a Sphere，教育辦公室/球體科學 + NESDIS Viz Lab，可視化實驗室) 合作，使用數年前在美國華盛頓特區收集的城市熱島資料和 NOAA Geo photos 等素材創建，專案贏得 American Council for Technology and Industry Advisory Council (美國技術委員會和工業諮詢委員會) 授予的 Innovation Champion Award (創新冠軍獎)，透過運用新技術媒介來拓展新受眾，促進民眾對氣候變遷的了解。(圖12)

虛擬實境的展示中讓民眾看到城市中不同場域的熱島效應，也看到都市不同地區內因樹木覆蓋率不同而產生的高溫不平等 (heat inequities)，以期讓民眾了解解決城市高溫問題的重要性以及綠色基礎設施和其他城市環境降溫措施的潛在好處。

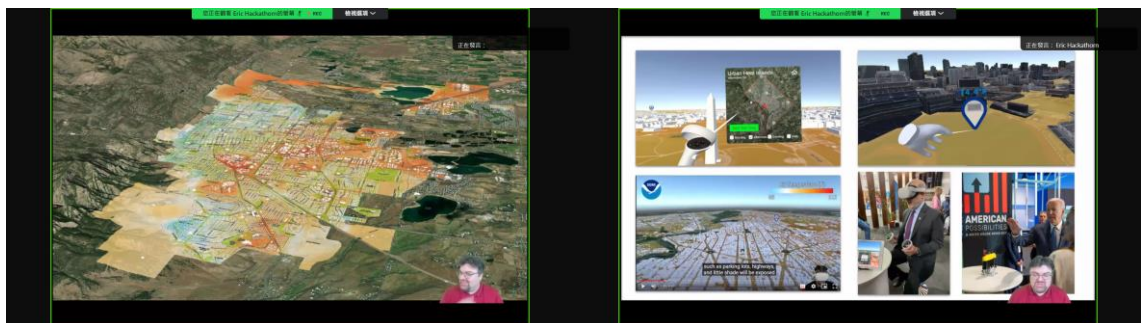


圖12、NOAA Urban Heat Island of Washington, DC 及 Virtual Reality (VR)體驗介面

## NOAA Digital Twin for Earth Observations

另一個正在進行中的項目，是由 NOAA's Satellite and Information Service (NESDIS) 主導的 Digital Twin for Earth Observations Using Artificial Intelligence (EO-DT) 探索型計畫 (圖13)，與三家公司簽訂了共計500萬美元的合約 (Lockheed Martin Space, 290萬、Orion, 150萬、Science and Technology, 60萬)，進行2年期的工作，透過合作的公司夥伴評估地球觀測數位雙生技術的成熟度，及運用數位雙生技術增強處理、監測、品質控制、整合、融合和同化環境觀測的能力，評估將其引入 NOAA 的系統和作業中的可行性，主要著眼在簡化衛星資料地面處理和透過應用程式向使用者傳播的能力。預計將來自多個來源的不同觀測資料融合到地球系統的即時數位副本 (digital replication) 中，用來對氣候、天氣和生態系統預測作模型的建立、進行 what/if 情境規劃，以用於決策和根因分析。

系統的發展主要在建立後端資料處理、即時互動的分析及可調整的資料視覺呈現。資料處理包含將資料格式化、標準化、建立儲存規則及應用 AI/ML 技術的資料異常檢測演算法，並設計可靈活操作的平臺，透過即時互動選擇觀測資料種類及參數、地理區域、時間範圍和層場項目，並快速呈現，未來將進一步通過標準介面達成與其他計畫的資料交互運用性。(圖14)

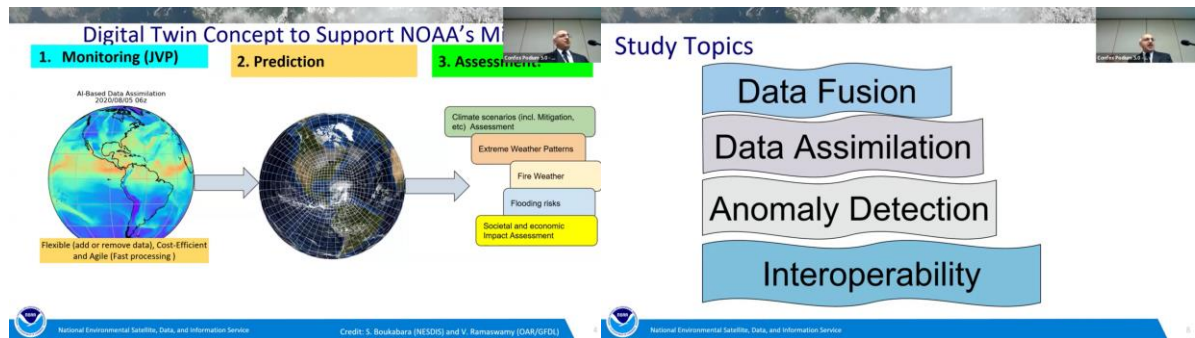


圖13、NESDIS 宣告的 EO-DT 計畫重點項目



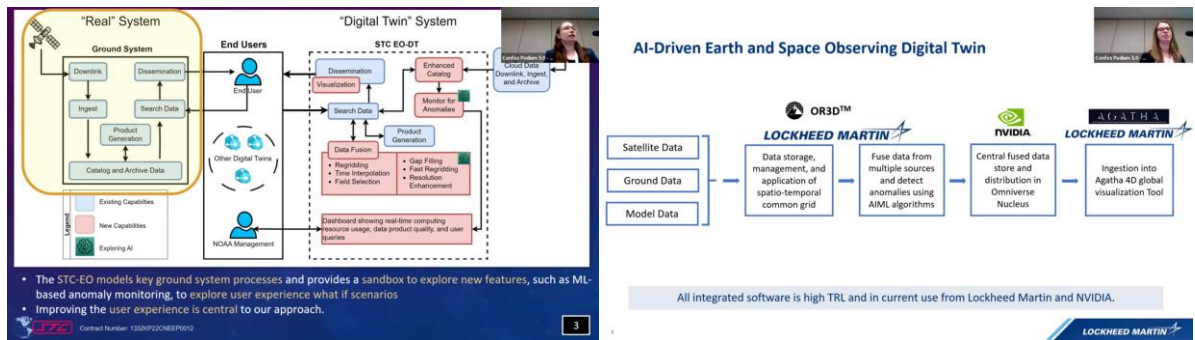


圖14、EO-DT 計畫其中2間廠商所提出的系統架構

## NASA Earth System Digital Twins

美國國家航空暨太空總署 (National Aeronautics and Space Administration, NASA) 發展的地球系統數位雙生(ESDT) 則是規畫為一個動態、互動式的資訊系統，首先，它需要盡可能準確和及時地提供地球系統過去和當前狀態的數位副本 – What now?，其次，需要能夠在假設下基於當前副本計算未來狀態的預測 -- What next?，第三，應提供在不同影響評估/假設下試算各種可能場景的能力 -- What if?。換句話說，ESDT 通過不斷擷取新的觀測資料並利用多個互連模型、機器學習以及高速計算和視覺化功能，要達到能提供地球系統的現在、下一步和可能性假設樣態。(圖15)

近年地球領域數位雙生的發展，源於幾項進展的融合，其一是已經連續收集了50多年、大量的不同資料，隨著格式與解析度變遷，越來越難以整合、有效利用這些資料；同時，由於氣候變遷的影響，許多非傳統的研究分析人員和預測人員對資料中可能藏有的資訊發生興趣；再者，隨著計算能力和視覺化呈現的進步以及機器學習技術的發展，從大量資料中萃取相關訊息及複雜模型運作皆已成為可能。因此，NASA 著手構建直觀的數位雙生互動式框架，目的是讓具有不同技術水準的使用者能夠輕鬆運用大量目標資料、以及相關地球系統和人類活動工具、模型，支持他們分析和視覺化這些資訊，並支援決策或政策制定。通過提供不同解析度的呈現以及標準化的工具和軟體技術，同一數位雙生副本可以滿足不同使用者在各種(空間和時間)解析度，和各種應用(科學、經濟、政策等)下的需求。

構建數位雙生系統的第一個挑戰是確定數位雙生應該是全域的 (global) 還是區域 (local) 的，例如，氣候或天氣等領域將需要全球性的數位雙生功能，而生物多樣性等科學領域可能是區域化的。同時要決定是要提供多重領域 (multi-domain) 還是主題性的 (thematic) 的數位雙生系統，多個主題性的 ESDT (例如空氣品質、野火、水文)可以聯合

起來，為其他 ESDT 提供輸入資料。整體而言，NASA 想像未來的數位雙生網絡能夠根據需要進行連接或聯合，這部分帶來了交互運用性 (interoperability) 的重要挑戰，包括需要從一開始就訂定這些系統的標準和資訊交換協定，每個單獨的數位雙生在內部構建時可具有充分的靈活性，但需要建立標準的輸入和輸出介面，才能與其他數位雙生相容。構建數位雙生系統的另一個挑戰是決定如何合併各種不同規模和類型的原始資料、分析資料，並自各種解決方案作出挑選，包括資料立方 (Data Cubes)、資料湖 (Data Lakes)、指標或隨選計算。還有一個挑戰則是將使用的視覺化的設計，包含對後臺的資訊取用及更新頻率，這也取決於使用者的需求。(圖16)

最後則是 AI 工具和技術因應不同類型的數位雙生系統作開發或改進，包括更好的物理模型、機器學習結果的可解釋性、或因果關係的呈現。還有，ESDT 模組需要可靠的不確定性量化方法，來驗證資料同化結果到評估 ML 模型的準確性和權衡這些系統支援的決策價值。

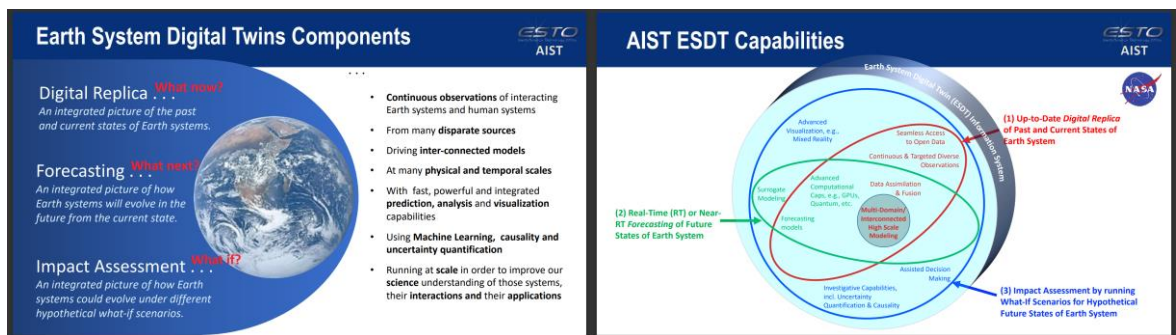


圖15、地球系統數位雙生 (ESDT) 的功能需求

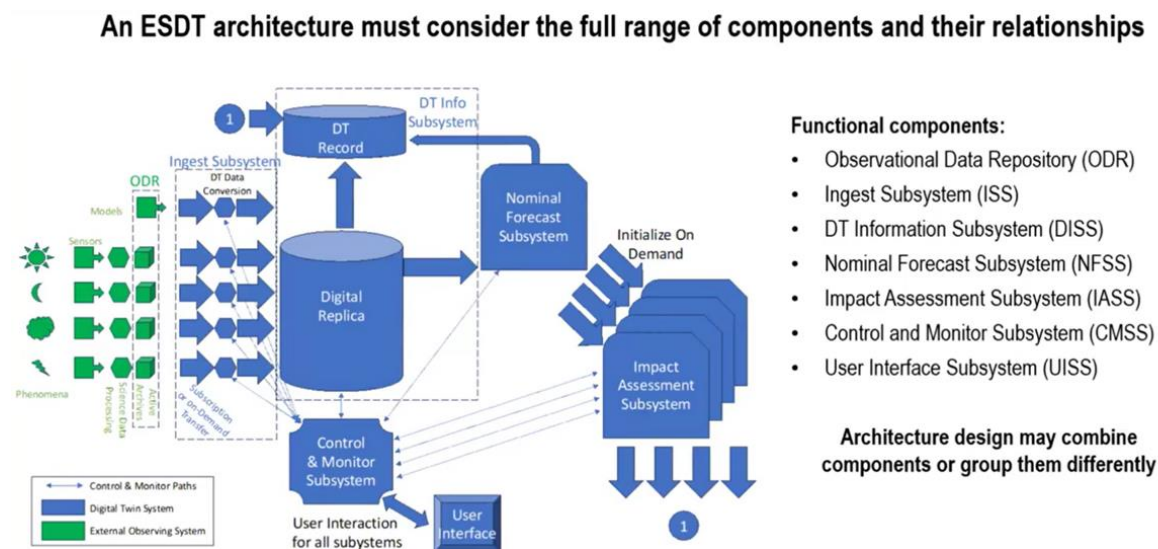


圖16、地球系統數位雙生 (ESDT) 的模組架構概念圖

## 4、關於可信任的 AI 驅動天氣預報

在過去的幾年裡，出現了純粹以 AI 驅動的全球氣象預報模型，這些模型已顯示出與 NWP 模型預報結果競爭的能力。目前，這些基於 AI 的模型仍處於概念驗證階段，要成為作業使用的產品仍需要許多評估。但全新的 AI 模型每 2-3 個月出現一次，並且功能迅速增強，又，許多 AI 模型的運算速度比傳統 NWP 模型快幾個數量級，並且可以在有限的計算資源上運行，低計算成本可以創建非常大的系集，而這些集合分佈可以更好地預測出包含稀有和極端事件的可能性，從而成為傳統 NWP 模型的強大競爭。

然而，這些基於 AI 的天氣預報模型尚未經過氣象界的嚴格測試，它們對業務預報員的效用尚不清楚。其中的問題包含，AI 系統被視為“黑箱”，因為無法解釋 AI 如何做出預測。而即使目前預報看似準確，但尚未進行有效的偏差與不確定性量測，且隨著時間和情境變化，AI 系統的產出是否能維持可被信任的準確度仍屬未知。

在模型的評估方面，AI 的模型與 NWP 模型具有截然不同的特徵，無法直接應用原為 NWP 模型開發的評估程式，例如，NWP 模型及其物理方程式保證了輸出的某些特性，例如各層場之間的動態耦合，但基於 AI 的天氣預報模型並沒有這些特性，因此需要另行開發合適的測試，找出其不確定性的特性，了解模型在哪些環境中和哪些應用上更合適，但因為現有 AI 模型的可解釋性受限，相關工作面臨挑戰。(圖17)

為了解決黑箱問題，研究人員著手提出方法來提高 AI 模型的解釋性；例如，開發視覺化工具和控制界面，讓模型決策方式透明，使用者能夠與 AI 系統達成共同的情境理解、確定模型計算結果的可重複性、使用獨立的資料驗證其可靠性等，從而提高使用者對 AI 系統的信任。並在 AI 開發過程中，邀請使用者等利害關係人共同參與設計和測試，並根據回饋意見進行調整。

同時，研究人員開發新的測量工具和方法，例如建立視覺化工具和解釋模型輸出不確定性的介面，以能更準確的量測與呈現不確定性，並依結果進行風險溝通。進一步再設定判斷與調校的標準，進行並進行實證評估，確保 AI 系統的效能和目的符合設計者的期待，亦即在特定情境下的產出有符合要求的可靠性和有效性，藉此建立使用者的完整信任。

總結來說，基於 AI 的天氣預報模型具有與傳統 NWP 模型不同的特徵，因此面臨多重挑戰，包括黑箱問題、應用情境的限制性、信任與標準的缺乏以及不明確的不確

定性等。為了解決這些問題，研究人員計畫提高 AI 模型解釋性、讓使用者共同參與開發了解應用情境、改進信任測量方法、進行風險與不確定性溝通、設置標準等，可以明確作出模型的特性等評估，為模型開發人員提供回饋、為預報員等使用者提供產品運用的指南。(圖18)

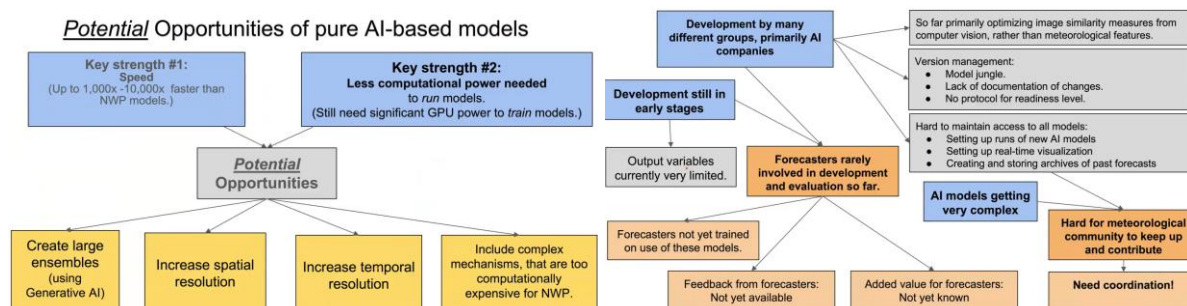


圖17、AI 天氣預報模型的機會與挑戰

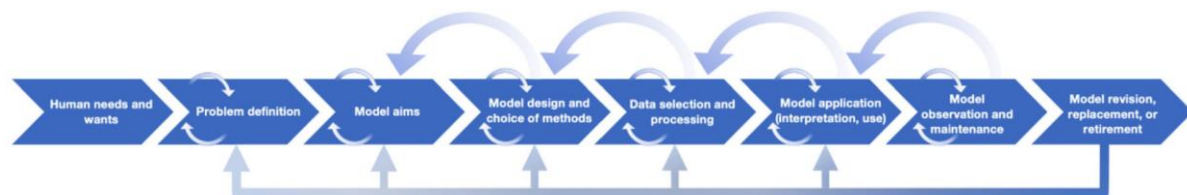


圖18、人工智慧/機器學習模型生產週期

## (二)、人工智慧就緒(氣象)資料 (AI Ready Data) 整備

新興的人工智慧和機器學習 (AI/ML) 建立了許多類應用模型，並逐漸被引用來應對天氣分析、模擬和氣候變遷的挑戰，這些 AI 應用模型需要大量資料作為輸入，而氣象單位存有大量的歷史資料可以提供運用，並預計未來資料量將呈現指數級增長。然而，這些資料的狀態對於研究人員是合用的嗎？

能否有效應用機器學習方法和從資料中提取有意義的知識，取決於精準對應所期望解決的科學問題之輸入、訓練和測試資料。這些用於機器學習、人工智慧的資料集通常很龐大，並需要相關資料專家承擔大量的處理與協助工作後才能“就緒”(符合使用要求)。依據調查，目前使用各類環境資料進行 AI 研究的人員有至少半數需要花費研究期間的50%以上的時間來處理資料，也就是說，整體而言，環境領域的 AI 相關研究，耗費了總體約50%的時間，只為了將資料準備為研究所需的輸入，其中的工作包括在缺乏對資料的認識及完整說明文件的狀態下嘗試瞭解資料內容、進行資料清洗、



處理格式轉換、依需求標記資料等工作，若資料處理人員對資料內容的解讀或處理錯誤，則後續所有步驟產出的品質都將受到影響，甚至可能完全失去價值。

## 1、 人工智慧就緒資料分級

“You’re not AI-ready until your data is.”

收集原始資料，而後清理這些資料，並將其處理為可被 AI 程式/模型應用的格式，即所謂“AI Ready”，也就是將資訊清理和解析成結構化格式。而在前端資料收集期間完成這項工作是最有效率的，因為資料提供者擁有對資料的專業了解，較可能有效率地處理資料，而大多數研究人員在開始使用資料時對其不熟悉。

NCAI 對資料的人工智慧就緒情形，依據品質、存取方式、文件化等項目作了以下初步分級 (下表及圖19)：

分級	資料品質	存取方式	文件化
<b>Level 0: 非人工智慧就緒 (Not AI-Ready)</b>	在資料發佈前沒有進行內容一致性處理。	僅應請求或通過訂購系統提供資料。	沒有可用的資料字典，或採用非機器可讀格式 (例如：pdf)。
<b>Level 1: 初步 (Minimal)</b>	以人工方式檢查內容一致性。	僅有非程式方式的取用選項，例如檔案下載。	有機器可讀格式的資料字典 (例如：csv、XML、JSON)。
<b>Level 2: 中階 (Intermediate)</b>	部分一致性檢查是自動化的，部分是結果文檔。	提供數種資料取用選項，包括一種以上用程式方式取得的方法 (例如：批次檔案下載、API 或雲端服務)。	資料字典符合機器可讀的詮釋資料 (metadata) 標準。
<b>Level 3: 最佳 (Optimal)</b>	全自動的內部一致性檢查和報告，並考慮領域資料集之	有多種資料取用選項 (下載、API、雲端、資料即服務	資料字典符合機器可讀的詮釋資料標準，參數並與其他



	間的外部一致性。	等)。	領域資料集或跨領域資料標準進行協調。
--	----------	-----	--------------------

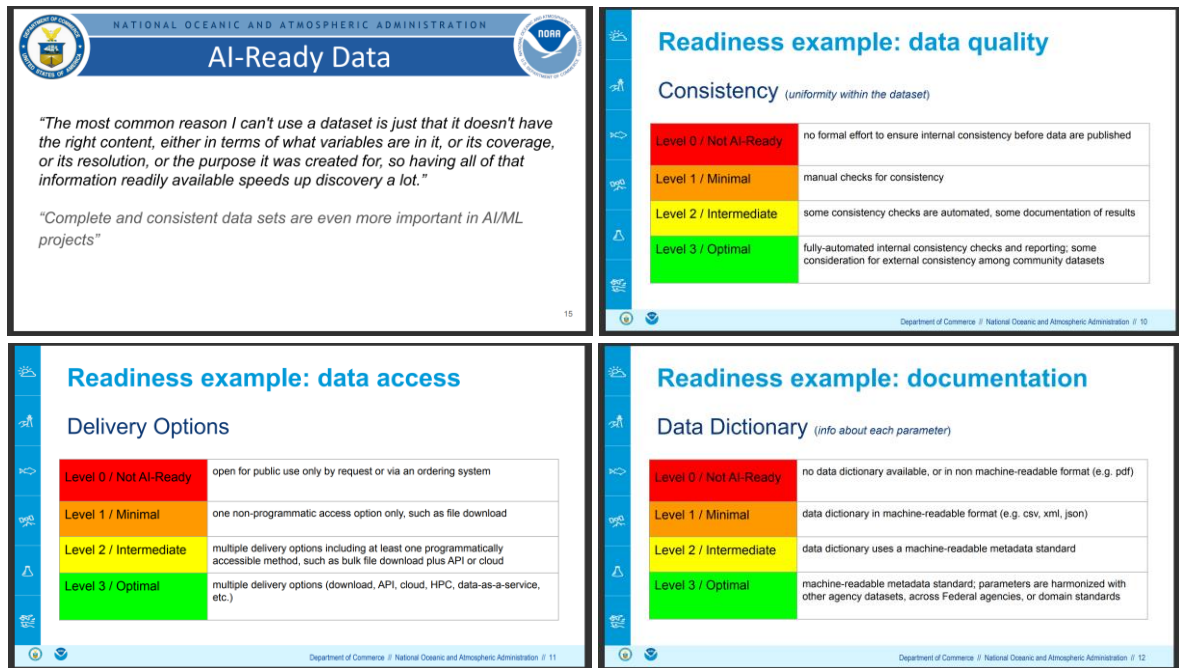


圖19、AI Readiness Matrix

## 2、人工智慧就緒資料檢核清單

Earth Science Information Partners (ESIP)是由 NASA、NOAA 和 USGS (United States Geological Survey) 等單位的地球科學資訊專業人員組成的非營利組織，主要功能是進行跨學科合作、分享技術知識並與資料使用者互動，目的是支援地球科學資料的創新使用和管理，讓所有需要地球科學資料的人可以隨時隨地通過 ESIP 推出的解決方案運用資料。

AI 和 ML 正大量被應用在解決地球科學研究問題，但在資料使用上“garbage in, garbage out”是不會改變的，而隨著處理的資料解析度提高、資料量增大，資料未妥善完成前處理以致後續分析成為無效產出的資源損失也會相應增加，若能確保資料可用於 AI，意味著能夠提高研究的效率、減少資料被錯誤解讀的情形，並且資料集可以被共用、重用。

ESIP 的資料就緒小組 (Data Readiness Cluster) 向 100 多名跨部門的 AI/ML 研究人

員、地球科學家、資料產製者、資料科學家，諮詢了其認為影響資料使用及處理方便性的各種因素，整理出四個主要分類：資料準備 (Data Preparation)、內容品質 (Data Quality)、說明文件 (Data Documentation)、資料取用 (Data Access)，並依據調查結果，推出了一份人工智慧就緒資料檢核清單 (AI-readiness Checklist) (圖20)，內容包含 AI 資料準備與標準分析應用的各項關鍵要素，並轉化為簡要的問題清單，目的為幫助 AI 工作者藉由填答其中各項目，完整評估可能用於 AI 應用的資料集仍需進行的資料處理投資及工作量。

### **資料準備 (Data Preparation)**

沒有人喜歡雜亂無章的資料集，尤其是必須耗費大量時間精力清理它的時候，因此檢核清單第一部分側重於可能加快後續資料分析的項目，評估資料集完成某些預處理步驟的程度，同時，在準備過程中完成的內容將成為詮釋資料中的標籤。

- 是否已填入缺失值 (null values/gaps) ?
- 是否已識別出異常值 (outliers) ?
- 資料是單一來源還是整合多個來源 ?
- 資料是否已網格化 (gridded，亦即在空間和時間上規則取樣(sampling)) ?
  - 已轉換為標準網格或是採用原始取樣/解析度 ?
- 資料是否已確認並標記 (labeling) 目標？亦即是否可以直接將其用作監督學習 (supervised learning) 的訓練資料集 (training dataset) ?

### **內容品質 (Data Quality)**

從資料的完整性，各部分處理、格式的一致性，對內容不確定性的量化，再到供應的即時性，資料的品質對任何應用程式都很重要，因為它直接影響機器學習演算法的回應，從而影響其預測結果的有效性或相關性。

- 資料集有多大？根據資料類型的不同，可能包含總資料量 (total data volume)、資料維度 (dimensionality)、圖形數量、檔案數量、資料表行數 (table rows)、圖形大小等。
  - 是原始資料還是衍生/後處理的產品？
  - 是觀測結果還是模式輸出？

- 資料是否更新 (updated) ?
  - 資料更新的即時性 (timeliness) 及更新頻率 (update frequency) 如何? (近即時、日、週、月、年、數年)
  - 即時提供的(原始)資料 (preliminary data) 與與稍後更新的完整資料 (later update of the full record) , 在品質上是否存在差異? 是怎樣的差異? 花多少時間更新 (delay between different stages) ?
  - 新版資料是否完全取代 (supersede) 舊版資料?
- 是否有關於資料的空間和時間解析度的定量資訊?
  - 資料的時間解析度為何? 是否符合要求?
  - 資料的空間解析度為何? 是否符合要求?
  - 資料已經過降低/提高解析度處理, 還是原始的? 如果經過處理, 原始資料是否提供? 是否可用?
- 資料集內各資料使用的單位 (units)、型態 (data types)、參數名稱 (parameter names) 是否一致 (consistency) ?
  - 是否有自動化流程來監控單位、型態、參數名稱的一致性?
- 資料集內容是否有已知誤差 (bias) /不確定性 (uncertainty) ? 描述誤差/不確定性型態。
  - 是否已檢驗、量測誤差/不確定性? 量測方式為何?
  - 是否有誤差/不確定性的定量衡量標準?
  - 各資料的誤差/不確定性是否能被追蹤/校驗?
  - 各資料的誤差/不確定性是否有工具能減少?
  - 是否有可提供的資料品質報告?
- 是否追蹤記錄了資料履歷 (provenance) ?
- 是否有資料完整性 (completeness) 檢查, 如 checksum 或其他?
  - 是否已採取措施確保完整性?
- 各項資料處理與紀錄內容是否經過內外部審查?

### **說明文件 (Data Documentation)**

關於如何使用資料的說明訊息, 應提供有關資料內容、詮釋資料、各參數的定義、代碼表、資料典/運用的資料標準、資料格式、標籤、資料來源、資料範例、資料處理工具、資料授權/許可...等各項資訊。另外, 資料標準化與完整說明文件是相輔

相成的，研究人員處理資料時除了依賴文件、清單做為地圖指引，也能期待資料以一致的形式、格式等呈現。

- 資料集是否提供有詮釋資料？
- 詮釋資料是否遵循領域資料標準 (community/domain standard)？依據哪個標準？
  - 詮釋資料是否為機器可讀 (machine-readable)？
  - 詮釋資料是否包含有關空間和時間範圍 (spatial and temporal extent) 的詳細資訊？
- 是否有全面的資料典/代碼手冊 (comprehensive data dictionary/codebook)來描述每一項資料參數？
  - 參數是否遵循定義的資料標準？
  - 資料典是否標準化？
  - 資料典是機器可讀的嗎？
  - 參數名稱是否使用常用詞彙或依循資料交換標準，例如 NIEM (National Information Exchange Model)？
- 資料集是否具有唯一的永久識別標記，例如 DOI (Digital Object Identifier)？
- 是否提供有領域專家的聯繫資訊？
- 是否設有用戶回饋和提出建議的機制？
- 是否有範例程式 (example codes) /使用註記 (notebooks) /工具包 (toolkits) 來幫助運用資料？
- 是否有提供明確的資料使用授權 (license/usage rights)？
  - 授權是否以標準化且機器可讀方式呈現(例如 Creative Commons)？
- 資料集是否已被使用於 AI 或 ML？若有，成果/報告的連結為何？
- 是否列出關於建議和不建議的資料使用方式？

### **資料取用 (Data Access)**

資料集以各種格式和方式提供使用，包括雲端儲存、批次檔案下載、應用程式程式設計介面 (API) 供應，多樣性的供應方式是 FAIR 要求 (詳後續章節) 的關鍵。資料取用的相關設計亦有層次，包含規格化存放資料，搭配完整的說明，也包括使用授權管理和安全保護。資料取用作為清單的最後一步，實作整個流程的主要目標：AI 就緒資

料成為人們可以隨時使用的資訊。

- 主要的檔案格式 (file formats) 是什麼？
  - 有哪幾種不同的檔案格式？
  - 它是機器可讀的嗎？
  - 它是否至少以一種開放的非專有格式 (open, non-proprietary format) 提供？
  - 有支援檔案讀取或轉換 (data format conversion) 的工具/服務嗎？
- 資料如何取用？
  - 取用資料需要身分認證 (authentication) 嗎？例如申請使用帳號。
  - 直接下載檔案 (direct file downloading) 還是訂購 (ordering)？
  - 有提供 API 或網頁服務嗎？
  - 若有 API，是否遵循開放的標準協定 (open standard protocol)？
- 對於限制使用的資料 (restricted data)，是否已進行授權管理，並對資料隱私和安全進行適當的保護？
  - 是否對資料進行了聚合處理以降低顆粒度 (aggregated to reduce granularity)？
  - 資料是否匿名化 (anonymized) / 去識別化 (de-identified)？
  - 授權使用者 (authorized users) 是否可以安全地訪問 (secure access) 完整資料集？

ESIP Data Readiness Cluster 提出之最新的人工智慧就緒資料檢核清單於 ESIP 線上資源中發布，並附如圖20。

ESIP AI-readiness Checklist			
Dataset Name	_____	Version 1.0, last updated 7/13/2023	
Dataset Version	_____		
Dataset Location/Link	_____		
Assessor Name	_____		
Assessor Email	_____		
	<b>Responses</b>	<b>Definitions</b>	<b>Usability Score</b>
Is this raw data or a derived/processed data product?	(please select option)		
Is this observational data, simulation/model output, or synthetic data?	(please select option)		
Is the data single-source or aggregated from several sources?	(please select option)		
<b>Data Quality</b>			
Will the dataset be updated?	(please select option)		
If the data will be updated, how often will it be updated?	(please select option)		
If the data will be updated, choose the update frequency that best describes the dataset.	(please select option)		
Will there be different stages of the update (e.g., updated with preliminary data first and replaced by a later update of the full dataset)?	(please select option)		
If yes, what is the delay between different stages?	(short answer)		
Should the new version of the dataset supersede the current version?	(please select option)		
Please provide an explanation for "Other" response.	(short answer)		
Is there any documentation about the completeness of the dataset?	(please select option)	<a href="#">Completeness definition</a>	
If there is documentation regarding the completeness of the dataset, please provide a link to report/document.	(link to report/document)		
How complete is the dataset compared to the expected spatial coverage?	(please select option)		
How complete is the dataset compared to the expected temporal coverage?	(please select option)		
Is this dataset self-consistent in that its units, data types, and parameter names do not change over time and space?	(please select option)		
Is this dataset's units, data types, and parameter names consistent with similar data collections?	(please select option)		
Are there processes to monitor for units, data types, and parameter consistency?	(please select option)	<a href="#">Consistency definition</a>	
If there are processes to monitor consistency, what measures are taken?	(short answer)		
Is there known bias in the dataset?	(please select option)	<a href="#">Bias definition</a>	
If there is known bias in the dataset, please provide more information.	(short answer)		
Have measures been taken to examine bias?	(please select option)		
If yes, what measures were used?	(short answer)		
Is the bias metrological traceable?	(short answer)		
Is there reported bias in the data?	(please select option)		
Link to the report/document on the bias.	(optional link)		
Link to tools available to reduce bias.	(optional link)		
Link to a bias-corrected or bias-reduced version of the dataset.	(optional link)		
Is there quantitative information about data resolution in space and time?	(please select option)		
Are there published data quality procedures or reports?	(please select option)		
If there is published quality information, please provide the link to the information.	(link)		
Is the provenance of the dataset tracked and documented?	(please select option)		
Are there checksums / other checks for data integrity?	(please select option)		
What is the size of the dataset?	(short answer)		
<i>(depending on the resource, this might be total data volume, dimensionality, number of images, data files, table rows, image size, etc.)</i>			
<b>Data Documentation</b>			
Does the dataset metadata follow a community/domain standard or convention?	(please select option)	<a href="#">Dataset metadata definition</a>	
If the metadata follows a community/domain standard, which standard is it?	(short text)		
Is the dataset metadata machine-readable?	(please select option)		
Does it include details on the spatial and temporal extent?	(please select option)		
Is there a comprehensive data dictionary/codebook that describes what each element of the dataset means? parameters?	(please select option)	<a href="#">Data dictionary/cookbook definition</a>	
Is the data dictionary machine-readable?	(please select option)		
Do the parameters follow a defined standard?	(please select option)		
If the parameters follow a defined standard, which standard is it?	(short text)		
Are parameters crosswalked in an ontology or common vocabulary (e.g. NIEM)?	(please select option)	<a href="#">Ontology definition</a>	
Does the dataset have a unique persistent identifier, e.g. DOI?	(please select option)	<a href="#">Identifier definition</a>	
If the dataset has a unique persistent identifier, please record the identifier.	(short text)		
Is there contact information for subject-matter experts?	(please select option)		
Is there a mechanism for user feedback and suggestions?	(please select option)		
Are there example codes/notebooks/toolkits available showing how the data can be used?	(please select option)		
What is the license for the data?	(short text)		
Is the license standardized and machine-readable (e.g. Creative Commons)?	(please select option)		
Has this dataset already been used in AI or ML activities?	(please select option)		
If the dataset already has been used in AI or ML activities, please provide a link to publications/reports	(link)		
Are there recommendations on the intended use of the data, and uses that are not recommended?	(please select option)		
<b>Data Access</b>			
What is/are the major file formats?	(short text)	<a href="#">Formats definition</a>	
Is this format machine-readable?	(please select option)		
Is the data available in at least one open, non-proprietary format?	(please select option)		
Are there tools/services to support data format conversion?	(please select option)		
If there are tools/services, please provide the link.	(link)		
Does data access require authentication (e.g., a registered user account)?	(please select option)		
Can the file be accessed via direct file downloading or ordering?	(please select option)	<a href="#">Delivery options definition</a>	
Is there an Application Programming Interface (API) or web service to access the data?	(please select option)		
If there is an API, does the API follow an open standard protocol (e.g., OGC)?	(please select option)		
If there is an API, is there documentation for the API?	(please select option)		
If there is documentation for the API, please provide the link.	(link)		
Is the data available publicly via cloud services?	(please select option)		
For restricted data, have measures been taken to provide some access while still applying appropriate protection for privacy and security?	(please select option)	<a href="#">Privacy and security definition</a>	
Has the data been aggregated to reduce granularity?	(please select option)		
Has the data been anonymized / de-identified?	(please select option)		
Is there secure access to the full dataset for authorized users?	(please select option)		
<b>Data Preparation</b>			
Have null values/gaps been filled?	(please select option)		
Have outliers been identified?	(please select option)		
Is the data gridded (regularly sampled in time and space)?	(please select option)		
If the data is gridded, was it transformed from a different original sampling?	(please select option)		
If the data is resampled from the original sampling, is the data also available at the original sampling?	(please select option)		
Are there associated targets or labels for supervised learning techniques (i.e., can this be used as a training dataset for supervised learning)?	(please select option)		
If there are associated targets/labels, are community labeling standards implemented (e.g., STAC label extension, ESA AIREO)?: (short text)	(short text)		

圖20、ESIP Data Readiness Cluster AI-ready data checklist

## AI 就緒資料集 NCAI Pilot Project: TC PRIMED

預報員極度依賴各種來源的大量觀測資料與預報產品來了解熱帶氣旋的狀態並發佈預報，這些資料與產品至少包含最佳路徑、環境診斷和再分析資料，且除了對於短

期預報應用非常寶貴，氣候研究人員也進行長期收集用於熱帶氣旋研究。為了研究熱帶氣旋並使用衛星觀測及預報模式，研究人員通常會整合觀測資料、加值產品及模式預報等的不同來源。資料整合涉及不同的觀測、加值和預報產品源進行資料獲取、搭配和相互校驗，運用同質來源資料作環境診斷、以及標準化這些不同的產品以進行分析。

由 NCAI 主導的 TC PRIMED (Tropical Cyclone PRecipitation, Infrared, Microwave, and Environmental Dataset) 是一個鎖定熱帶氣旋相關資料的計畫 (圖21)，主要是為了改善大量資料預處理負擔、提供全球熱帶氣旋的長期數位資料集；透過將產品整合到單一儲存庫中來改善不同資料來源的問題，為研究人員提供一個以分析熱帶氣旋為目的的資料集，以促進和支援研究，加速對熱帶氣旋對流和降水結構、強度和環境之間關係的理解。計畫目前已整理了從 1987 年到 2023 年，超過 1.6 TB 的資料，囊括全球3,533 個熱帶氣旋，各項資料包含：

- NOAA National Weather Service National Hurricane Center and Central Pacific Hurricane Center and Department of Defense Joint Typhoon Warning Center global tropical cyclone characteristics from the best-track database
- Multi-agency inter-calibrated, multi-channel microwave brightness temperatures
- Goddard Profiling Algorithm (GPROF) retrieved rainfall
- Geostationary satellite infrared imagery
- ECMWF fifth-generation reanalysis (ERA5) product fields and derived environmental diagnostics
- TRMM/GPM precipitation radar variables

TC PRIMED 蒐羅了三十多年的熱帶氣旋相關資料，依據 AI 就緒資料要求進行整理、標記就緒程度，搭配如何讀取、繪製和分析資料的步驟、範例，幫助使用者熟悉大氣科學界的常用檔格式，提供以人工智慧方法應用 TC PRIME 資料的複雜範例；藉此做法可以減少使用者準備資料的時間、提供資料集的使用指導及工具，可使未來的研究受益，增進我們對熱帶氣旋的理解並改進預報。TC PRIMED 團隊已於2024年9月在 NOAA Open Data Dissemination 上推出了資料集版本 (<https://noaa-nesdis-tcprimed-pds.s3.amazonaws.com/index.html#v01r01/>)，開放下載使用。



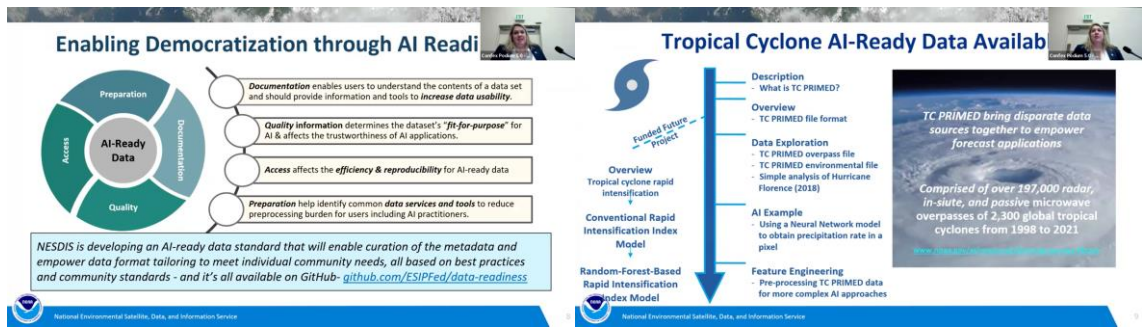


圖21、TC PRIMED AI 就緒資料集的發展要求及進程

**關於 NIEM (National Information Exchange Model)**

NIEM 是建立資訊交換的基礎，提供有關模型使用的規則和方法，以及可以重複被每個人使用的標準化 Information Exchange Development Lifecycle。也包括支援運用 NIEM 資料治理訓練、工具、及技術援助。NIEM 概念上資料標準，但是藉由建立介接的交互參照運行，並未要求各資料源改動原始資料以符合標準。

NIEM 包含共通詞彙 (common vocabulary) (圖22)、參考模型 (reference model)，用於加速溝通在不同的公、私組織間的訊息交換。參考模型中的資料元素就像字典中的單詞，NIEM 可被視為一本包含專業術語、定義、關係和格式的字典，這些術語、定義、關係和格式與資訊在單一系統中的儲存方式無關。NIEM 模型包括使用的組織、社群普遍同意的核心元素。例如，NIEM 核心中的常見元素包括 “person”、 “location”、 “item”、 “organization” 和 “activity”。

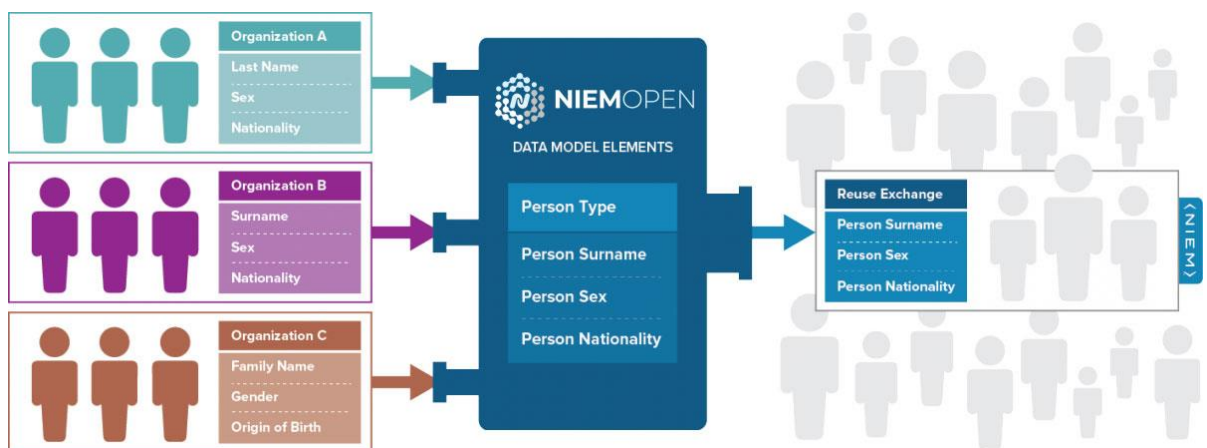


圖22、NIEM 的共通詞彙概念示意

NIEM 包含著資料標準的想法，例如我說 “vessel”，你說 “boat”，他說 “ship”，她說 “transportance”，其中代表的意思可能一樣，但是各應用系統不會自動把這些詞彙



當作相同的意思，需要對詞彙的共同理解，或將它們關聯起來的方法。NIEM 的發展就是希望以確保資訊在各個社群中具有相同一致的含義來溝通系統，但不是從新建立各系統，而是提供一致、可重用和可重複的通用詞彙、標準化的定義和訊息交換的流程，各組織能夠節省時間、金錢和資源，相互聯繫，並專注於解決問題、降低風險以及推進任務。

NIEM 模型中所有資料元素都遵循 NIEM 的標準命名規則。為防止命名衝突，每個資料元素或類型都在 NIEM Naming and Design Rules (NDR)中宣告，同時允許各組織對資料中含有概念進行適當治理，例如，不同的組織對“address”有不同的定義。NIEM 模型因此由兩層相關的詞彙組成：使用 NIEM 的所有社群普遍同意的核心元素，以及各個 NIEM 領域內部一致的特定元素。此部分於我國政府資料標準中的共通標準及領域標準有類似設計。(圖23)

NIEM 模型並為資訊交換商定了術語、定義、關係和格式，其與資訊在單一系統中的儲存方式無關，目前以 XSD、JSON 和 Microsoft Excel 格式提供，程式為 NIEM 中的每個 namespace 提供 Uniform Resource Identifier mapping，讓使用者引用模型，及鏈結資料。使用 NIEM 時，各組織只需了解自己和 NIEM 的系統詞彙表，並完成參照，NIEM 模型會處理系統間的一致可重用交換，各組織不需再另外處理連接到不同組織系統的通用詞彙協商。

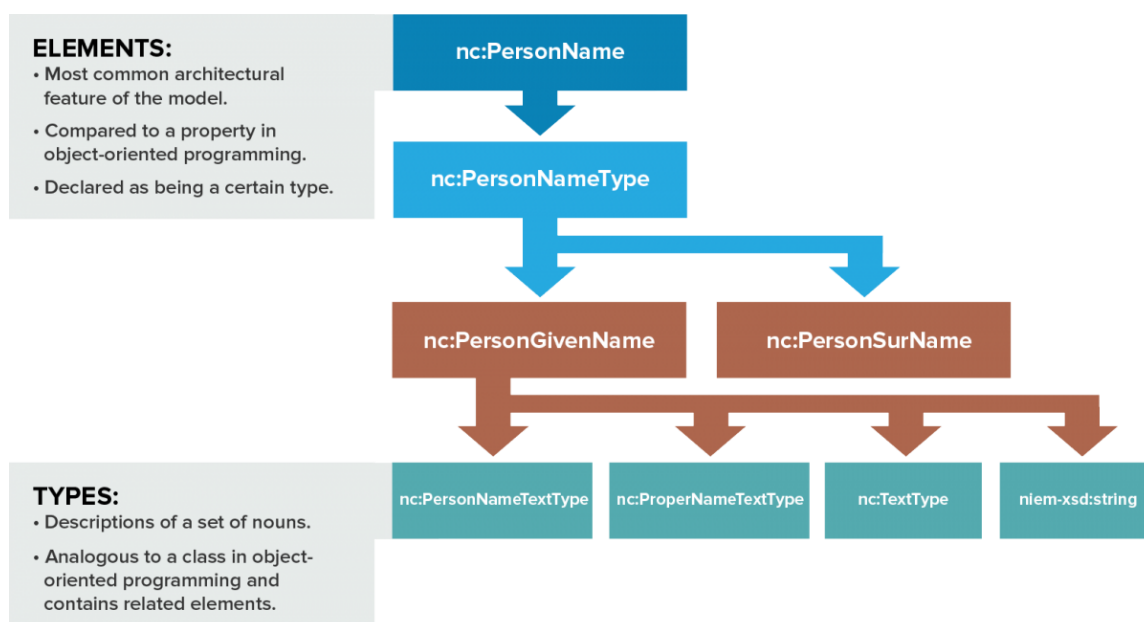


圖23、NIEM 的模型架構

### 3、 人工智慧研究轉型為作業

NOAA 內部依各專案的準備就緒級別 (Readiness Levels) 系統性的評估專案的成熟度，以對不同類型的研發專案之間的成熟度進行一致的比較，並選擇適合自研究轉型為作業、應用、產品或服務的項目，投入資源進行移轉。準備就緒級別一共分作9個等級。相關研究的團隊，可以依標準，評估專案符合的級別，及規劃上線為作業的各項前置準備工作。各階段工作整理如下表、簡化如流程圖24、原文內容如圖25。

階段	準備就緒級別	工作內容
研究	Level 1	基礎研究、實驗或理論工作，主要是獲得有關現象的基本基礎和可觀察事實的新知識，尚無任何特定應用或用途。基礎研究可以針對一些感興趣的廣泛領域，後續目標是一系列的未來應用。
	Level 2	應用研究，為獲取新知識而進行，主要針對一個具體、實際的目的或目標；進行應用研究可以是為了確定基礎研究結果的可能用途，或為了確定具體實現預定目標的新方法或途徑。
開發	Level 3	系統、流程、產品、服務或工具的概念驗證，屬於實驗開發的早期階段，包括可行性研究。
	Level 4	在實驗環境中通過評估的系統、子系統、流程、產品、服務或工具，屬於開發的中間階段。
	Level 5	以雛型設計通過評估的系統、子系統、流程、產品、服務或工具，屬於進入演示階段前的最後開發階段。
整合測試	Level 6	在相關或測試環境中演示並展現潛力的雛型系統、子系統、流程、產品、服務或工具。
	Level 7	在操作或其他接近真實的環境中演示的雛型系統、流程、產品、服務或工具，相關子系統元件已完成整合。
	Level 8	經過測試的最終系統、流程、產品、服務或工具，並可在使用者環境中依預期運作，並完成使用者培訓和相關文件、獲得上線許可。
部署	Level 9	系統、流程、產品、服務或工具已完成作業部署並進入常態運行。

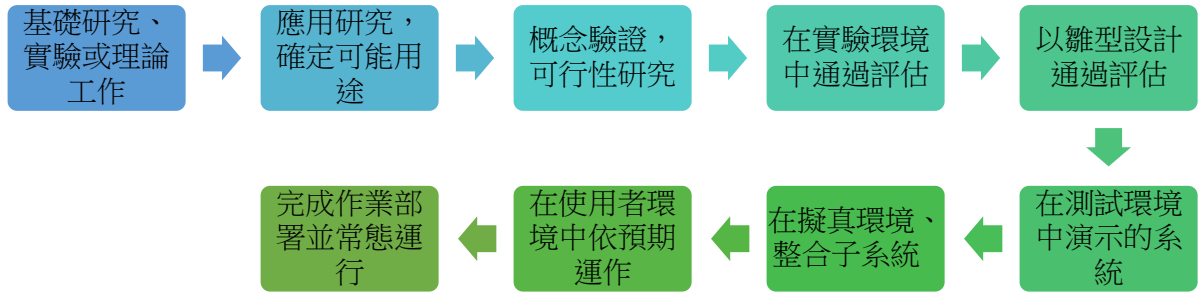
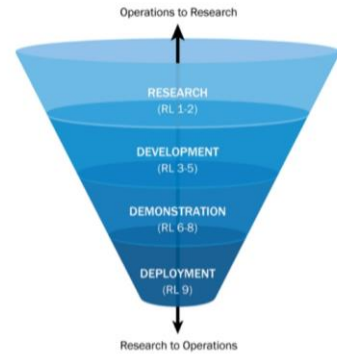


圖24、從研究到作業部署的各準備就緒級別

## NOAA READINESS LEVELS

Readiness Levels are a systematic project metric or measurement system that supports assessments of the maturity of R&D projects for transition from research to operation, application, commercial product or service, or other use and allows the consistent comparison of maturity between different types of R&D projects. NOAA's Policy on Readiness Levels can be found in NAO 216-115A.



There are 9 Readiness Levels as follows:

### Readiness Level 1

Basic research, experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. Basic research can be oriented or directed towards some broad fields of general interest, with the explicit goal of a range of future applications (OECD, 2015).

### Readiness Level 2

Applied research, original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific, practical aim or objective. Applied research is undertaken either to determine possible uses for the findings of basic research, or to determine new methods or ways of achieving specific and predetermined objectives (OECD, 2015).

### Readiness Level 3

Proof-of-concept for system, process, product, service, or tool; this can be considered an early phase of experimental development; feasibility studies may be included.

### Readiness Level 4

Successful evaluation of system, subsystem, process, product, service, or tool in a laboratory or other experimental environment; this can be considered an intermediate phase of development.

### Readiness Level 5

Successful evaluation of system, subsystem process, product, service, or tool in relevant environment through testing and prototyping; this can be considered the final stage of development before demonstration begins.

### Readiness Level 6

Demonstration of a prototype system, subsystem, process, product, service, or tool in relevant or test environment (potential demonstrated).

### Readiness Level 7

Prototype system, process, product, service or tool demonstrated in an operational or other relevant environment (functionality demonstrated in near-real world environment; subsystem components fully integrated into system).

### Readiness Level 8

Finalized system, process, product, service or tool tested, and shown to operate or function as expected within user's environment; user training and documentation completed; operator or user approval given.

### Readiness Level 9

System, process, product, service or tool deployed and used routinely.

圖25、NOAA Readiness Levels (NAO 216-105B DEFINITIONS 11)

目前，NOAA 旗下的各機構及各合作單位進行有潛力的天氣相關 AI/ML 研發工作，透過競爭 Joint Technology Transfer Initiative (JTTI) 計畫的資金補助，由 NOAA Weather Program Office 以評選 JTTI Award 型式，選擇出成熟的天氣研究專案，協助後續的開發、測試和評估，並建立移轉計畫，以期於3-5年內過渡轉換為 NOAA Weather Service 的作業(圖26)。JTTI 計畫的目標是加速成熟天氣研究進入 NWS 作業的過渡時程，解決專案與 NWS 作業要求間的差距，以改進預報產出，造福美國公眾。JTTI 計畫優先補助與 AI/ML 相關的項目包含：推進新觀測和資料同化技術的 AI/ML 技術，以支援統一預報系統 (Unified Forecast System，UFS)；對天氣模式資料進行後處理的 AI/ML 方法之開發、測試和評估。

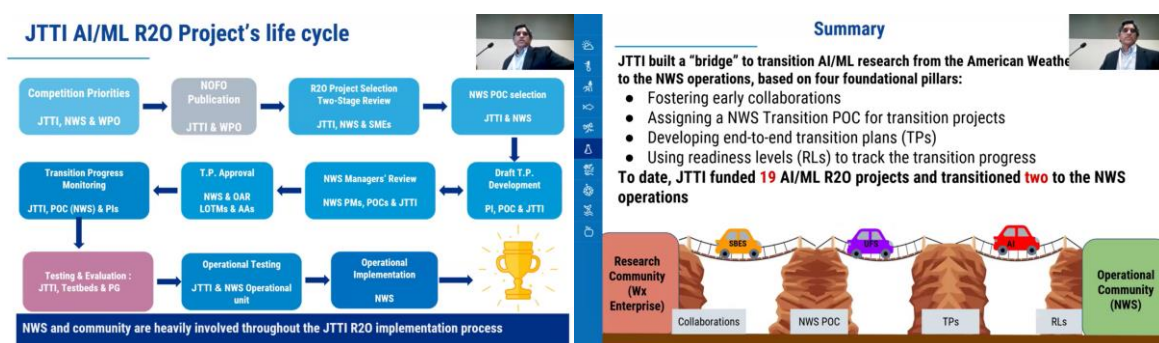


圖26、JTTI AI/ML Research to Operation 專案流程與成果

### (三)、 氣象資料庫

氣候環境變化的速度持續加快，公、私部門對氣象資料在社會、經濟層面上的需求持續上升，這使得有公信力且即時的氣象資訊服務愈形重要。需要持續精進服務內容、作法，來適應一直在增加的資料源和數量、不斷創新的資訊技術和永不停止的使用者新需求。

#### 1、 FAIR 資料原則 (FAIR data principles)

在各項的 NOAA 內、外部資料庫及各項資料服務了解過程中，反覆接收到宣告該資料服務依 FAIR 資料原則規劃或實作的訊息，該原則已成為資料服務設計應遵循的基本原則，其中要求提供之資料及服務符合 (1) 可查找性(Findable)、(2) 可取得性 (Accessible)、(3) 可交互運用性(Interoperable)及 (4) 可重用性(Reusable)。

相關原則是依循於2016年在 Science Data 上的一篇評論 “The FAIR Guiding Principles for scientific data management and stewardship”，作者們注意到各項作業、研究所得的資料並沒有發揮足夠效益，因為資料管理缺乏明確定義和標準、資料儲存缺乏統一性和整合性，以致資料的長期保管和重用存在挑戰，同時機器對資料格式、結構等的可運用性不足，在在影響資料運用和進一步的知識發現；因而提出了以下資料管理的 FAIR 原則 (圖27)：

- **可查找性 Findable：**

F1：資料/詮釋資料獲得全球通用、唯一、永久的識別碼。

F2：使用豐富的詮釋資料描述資料。(如 R1定義)

F3：詮釋資料清晰明確地包含其所描述資料的識別碼。

F4：在可搜尋之資源中註冊、索引詮釋資料。

- **可取得性 Accessible：**

A1：能以標準通信協定通過識別碼搜尋取得資料/詮釋資料。

A1.1：該協定是開放、自由運用及可普遍實作的。

A1.2：該協定允許在必要時執行身份驗證和授權程序。

A2：即使資料不再供應，仍應提供其詮釋資料。

- **可交互運用性 Interoperable：**

I1：資料/詮釋資料中的資訊使用正式、可取用、可分享且廣泛適用的語言表達。

I2：資料/詮釋資料的內容使用符合 FAIR 原則的詞彙。

I3：資料/詮釋資料應以正式方式引用其他資料/詮釋資料。

- **可重用性 Reusable：**

R1：資料/詮釋資料的各屬性應被準確和符合關聯地詳實描述。

R1.1：資料/詮釋資料以清晰易懂的使用授權發佈。

R1.2：資料/詮釋資料應包含詳細的來源與變更紀錄。

R1.3：資料/詮釋資料應符合社群相關標準。

## Box 2 | The FAIR Guiding Principles

### To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

### To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

### To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

### To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

圖27、FAIR 指導原則 (原始評論文章截圖)

### FAIR 資料原則的實作步驟

FAIR 實作會在資料、詮釋資料和資訊系統架構上進行。依據 FAIRCOOKBOOK 網站 (<https://w3id.org/faircookbook/FCB079>) 上的介紹，大致為以下步驟，詳如圖28：

- (1) 取得資料 (Get the data)：以 API 等方式取得開放或授權資料及其相關說明。
- (2) 確認領域模型 (Model the domain)：確定資料集中涉及的型態、了解用於資料表示的領域標準。
- (3) 選擇資料識別方式 (Select the identifier scheme)：建立用於識別資料集的方式，可以是新的識別方式或重用既有的方式。
- (4) 應用資料標準 (Apply data standards)：進行資料標準化驗證，或建立/修訂所需的資料標準，讓資料以領域標準指定的格式表示，以確保資料交互運用性。
- (5) 選擇資料詞彙表 (Choose data vocabularies)：深入瞭解資料內容，並選定要使用的既有資料詞彙，或為資料制定新的資料詞彙表供應用。
- (6) 轉換資料以實現資料交互運用性 (Transform data for interoperability)：建立不同資料集間資料標示及的資料詞彙對應，以便可以在多組資料詞彙和標示間作資料的交互運用。
- (7) 上架資料 (Host your data)：資料集準備就緒後，上架資料集並作適合資料搜尋



的優化，同時建立資料版本控制和資料格式管理。

(8) 分享資料 (Share your data)：資料集已符合 FAIR 資料原則，可通過合適的授權釋出分享。若資料內含敏感資訊，應在分享之前考慮去識別化。

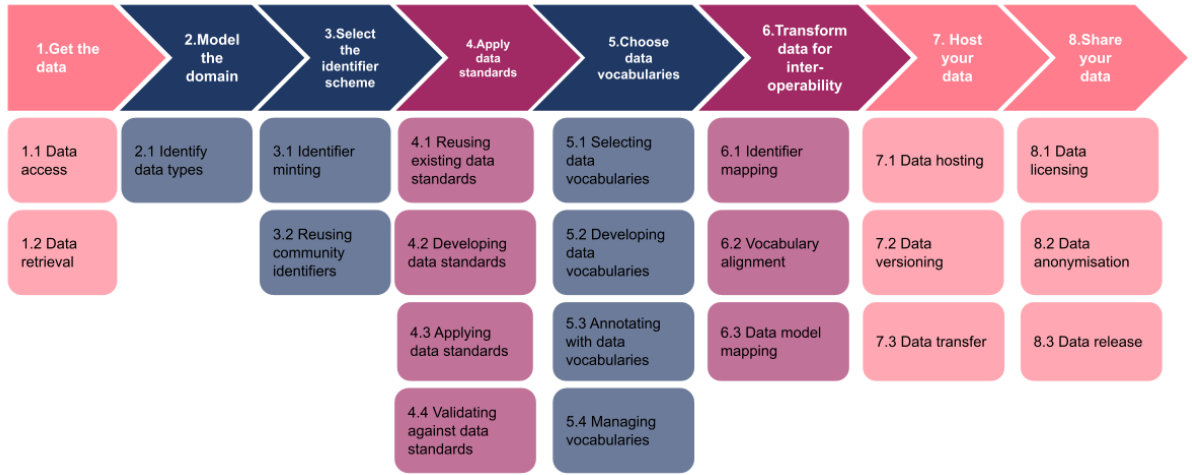
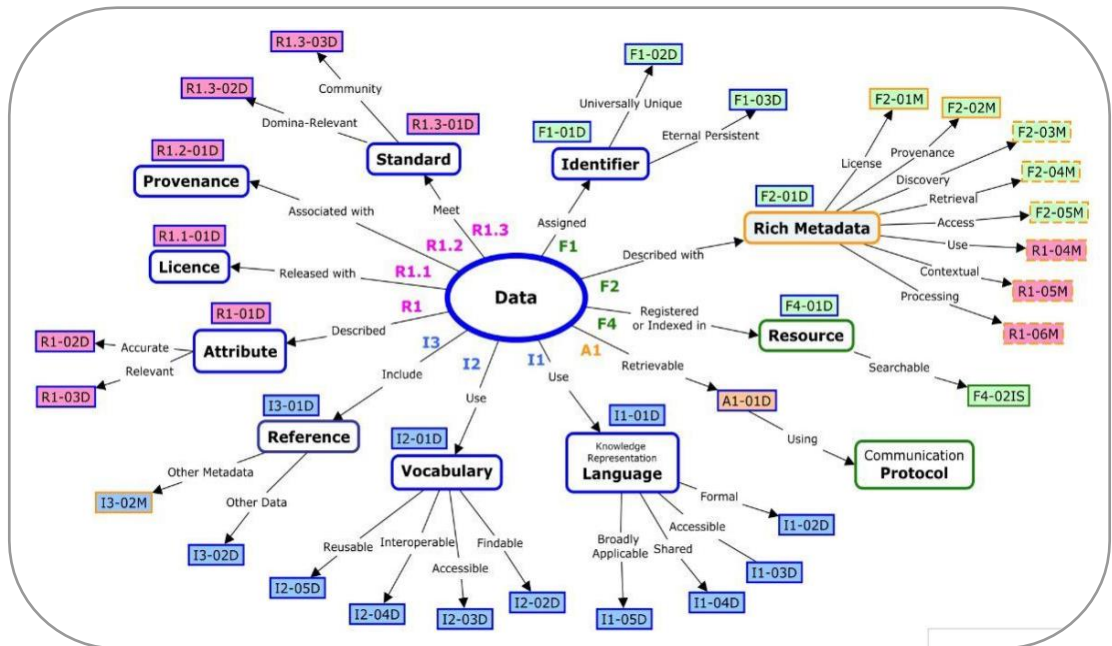


圖28、FAIR 資料原則實作步驟(FAIRCOOKBOOK 網站)

而各項實作細節則分別與資料、詮釋資料及系統架構有關，在 Workbook for dissecting the FAIR Guiding Principles 中剖析了工作進行的主要類別、核心概念、關鍵要素及運用的指標，最終可以轉化為資料服務是否符合 FAIR 資料原則的評量標準。(圖29)



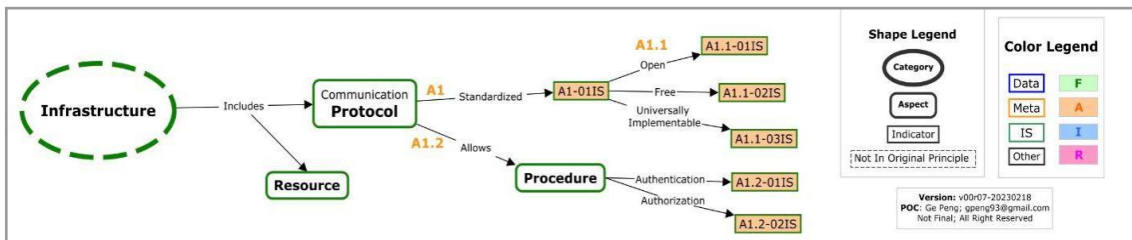
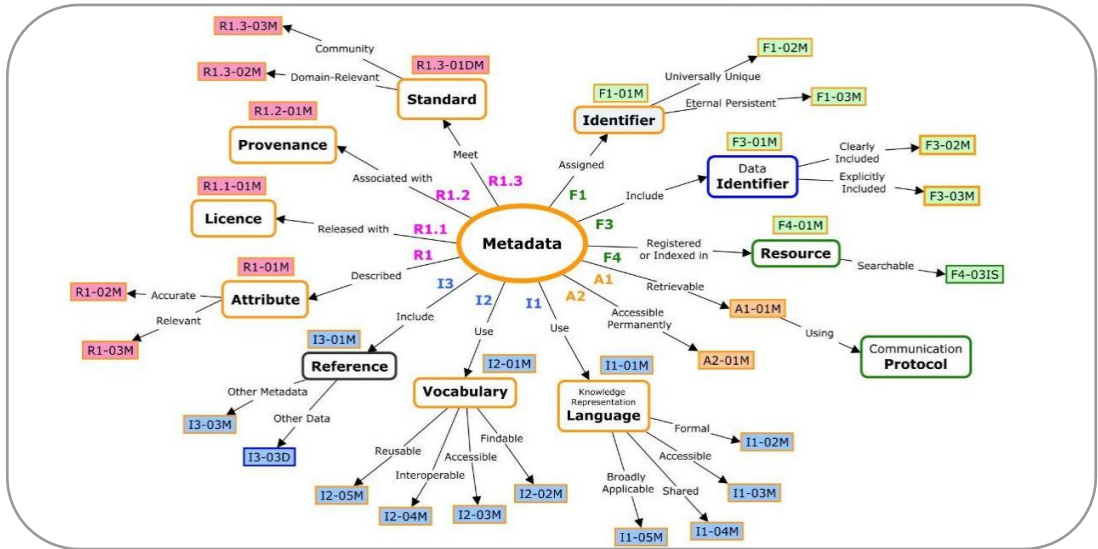


圖29、FAIR 資料原則對資料、詮釋資料及架構各要求的概念整理圖

### FAIR 資料實作—Copernicus Climate and Atmosphere Data Store

氣候和大氣資料倉儲(CADS)是 ECMWF 的哥白尼服務的骨幹軟體基礎設施，其中氣候資料倉儲(CDS)作為氣候變化服務的支援，大氣資料倉儲(ADS)則是大氣監測服務的共用後端元件。CDS 和 ADS 的組合為超過 25萬的註冊使用者提供服務，平均每天有超過 70萬筆處理請求，提供超過 120 TB 的資料。

CADS 軟體基礎架構設計為分散式系統和開放式框架，通過網頁介面提供對眾多資料集、應用程式和其他數位資訊目錄的單一入口點，可用的資料集則分佈在不同的供應者端。目前，CDS 目錄提供對141個資料集，包括再分析、季節性、預測和衛星資料，而 ADS 目錄提供了13個關於大氣成分、空氣品質再分析和預報等的項目。



# Enabling FAIR data output using the CDS

Edited by Benjamin Bellenger, Marco Zanarini, Angel Alcaz and many others  
ECMWF, Reading, UK

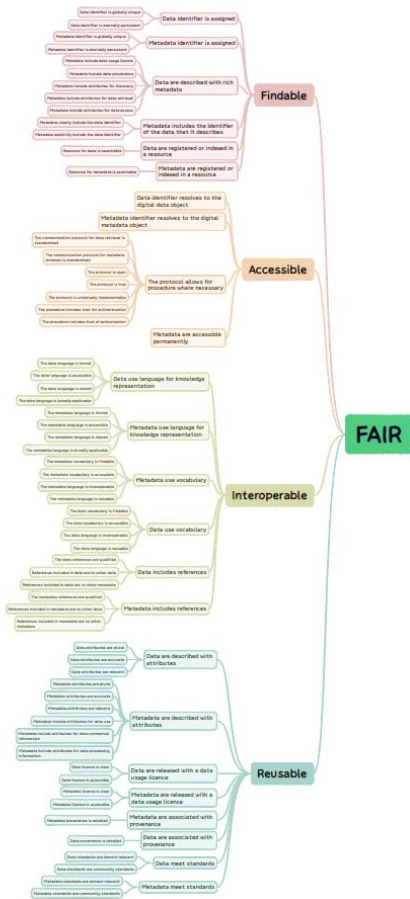
## Context

FAIR principles aim to measure the maturity of four key aspects that optimise data sharing across systems and disciplines: Findability, Accessibility, Interoperability, and Reusability.

Various organisations and data scientists have developed their own FAIR data and metadata assessment standards, based on their priorities, interpretations, and constraints, leading to inconsistencies.

In this poster, we use one<sup>(1)</sup> of those frameworks that may yield results slightly different from other FAIR evaluation frameworks.

## FAIR evaluation tree



## FAIR principles in action at Copernicus' CDS

At the ECMWF, we are committed to making our data as useful and accessible as possible. That's why we aim to publish all data in accordance with the FAIR principles. Data already distributed through the Copernicus Climate Data Store (CDS) exemplifies how FAIR principles can be applied to provide authoritative, traceable, and understandable climate data, enhancing data accessibility and reuse.

The CDS data is carefully curated and described using relevant metadata, that provides detailed information about the variables and parameters included in the data. For each variable, we provide a clear definition, specify the units, and include any relevant notes or caveats that users should be aware of, ensuring that the data can be used accurately and reliably.

The CDS uses DOIs (Digital Object Identifiers) to provide persistent, stable links to data, allowing users to easily find and access the data they need.

The CDS also use open, standardized formats for the data and provide API (Application Programming Interface) access, allowing users to easily integrate CDS data with other systems and applications.

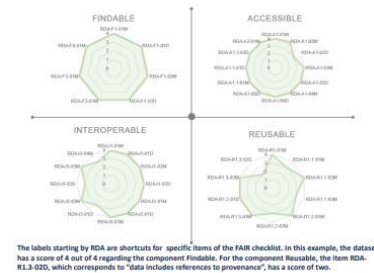
And the CDS provides clear licensing information, enabling users to freely reuse and repurpose the data for their own purposes.

By following the FAIR principles, the CDS is helping to make data more valuable and useful for a wide range of users, from meteorologists and researchers to policymakers and the general public.

## Core FAIR attributes and components



## Template for the visualisation of a FAIR assessment<sup>(1)</sup>



The labels starting with RDA are shortcuts for specific items of the FAIR checklist. In this example, the dataset has a score of 4 out of 4 regarding the component Reusable. For the component Reusable, the item RDA-R1.3-S2D, which corresponds to "Data includes references to provenance", has a score of two.

圖30、CADS 的 FAIR 資料供應說明

簡單和一致性是 CADS 架構的關鍵，可促進使用者探索、發現、檢索和使用可用資源，因此 CADS 建立了一個定義明確、嚴格的內容管理方法，並由自動部署工具和儲存在集中式版本控制儲存庫中的一組定義檔提供支援。定義檔依 FAIR 原則 (可查找、可取得、可交互運用、可重用) 要求，涵蓋資源定義的不同方面，從描述性資料、詮釋資料到互動式表單、說明文檔和品質控制，藉由多功能的內容管理和部署機制提供了靈活性，亦可以適應不同的標準，並促進與第三方平臺和服務的交互運用性，不斷發展和改進平臺及其內容。(圖30)

## 2、NOAA 的資料與資訊管理政策

NOAA 將資料視為戰略資產，並對其進行管理，以發揮 NOAA 在觀測、模式和研究資料的最大價值。以下節錄整理其 “Management of NOAA Data and Information” 之內容 (圖31)：

NOAA 的環境資料包括但不限於觀測之海洋、大氣、空間環境、太陽和固體地球的物理、化學、生物、地質和地球物理特性和狀態、衍生的相關資料、模式輸出或預測資料、相關文件和詮釋資料。

政策中規範所有 NOAA 資料都將按照 FAIR 原則進行管理，以確保其可查找、可取得、可交互運用及可重用，除非受法律、法規、合約等其他限制。NOAA 的資料管理涵蓋完整資料生命週期，包括以下階段：

- 計畫：規畫各專案、計畫或研究自起始至結束的資料管理方法和資源。
- 建立或收集：從 NOAA 和外部系統獲得資料，包括資料收集、購買、分析、整合和建立模型。
- 流程：處理資料以使其對使用者有用並產出產品。
- 保存：確定需要保留的資料與紀錄、選定保存方法，同步資料與相應紀錄的保存與後續存取。
- 取用：確保內部和公眾取用資料和詮釋資料的及時性，明確標記開放資料授權或其他版權限制。
- 處置：根據保存計劃刪除或銷毀資料。

各資料管理應包含以下項目：

- 描述：在資料收集和建立過程中，以詮釋資料記錄資料來源及收集過程。
- 追蹤和監控：在整個生命週期中追蹤資料和詮釋資料，並監控資料應用。
- 品質：確保資料的品質、客觀性、可用性和完整性。
- 安全性：保護資料免受未經授權的存取、損壞和盜竊。

NOAA 首席資料官和 NOAA 資料管理委員會協調，負責政策執行及與內部協商。

NOAA FORM 58- (4-04)

National Oceanic and Atmospheric Administration	NOAA Administrative Order 212-15B	
NOAA ADMINISTRATIVE ORDER SERIES	DATE OF ISSUANCE 11/16/2023	EFFECTIVE DATE 11/16/2023
SUBJECT MANAGEMENT OF NOAA DATA AND INFORMATION		

**SECTION 1. PURPOSE.**

This NOAA Administrative Order (NAO) establishes the Department of Commerce (DOC) / National Oceanic and Atmospheric Administration (NOAA) Data Management Policy. The purpose of the policy is to ensure data are treated as a strategic asset and managed to realize the maximum value from NOAA's investment in observations, modeling, and research. This policy provides high-level direction that guides decisions – and required actions regarding data and information management throughout NOAA – aligned with Federal law and data strategies. Further guidance on how this policy is to be implemented will be provided in a Data Management Handbook (the Handbook) that addresses the full data lifecycle for all domains of NOAA data and information.

**SECTION 2. SCOPE.**

This NAO applies to all NOAA environmental and administrative data – including data produced with NOAA funding via grants or cooperative agreements (or procured via contract) – and to the personnel and organizations that manage these data.

**SECTION 3. DEFINITIONS.**

The following are terms used in this NAO. Additional terms related to implementation will be defined in the Handbook.

- .01 Data – recorded information, regardless of form or the media on which the data is stored.<sup>1</sup>

<sup>1</sup> Definition from the Evidence Act, 44 U.S.C. § 3502 (16)



**SECTION 4. POLICY.**

- .01 All NOAA data will be managed following the FAIR principles to ensure it is Findable, Accessible, Interoperable, and Reusable,<sup>2</sup> except to the extent limited by law, regulation, policy (such as those applicable to personally identifiable information, protected critical infrastructure information, or proprietary trade information), contract, agreement, or security requirements, and accordingly labeled as controlled unclassified information.
- .02 Management of NOAA's data will be based upon an end-to-end data lifecycle that includes:
  - Plan: Plan and design the project, program, or research from onboarding to project closure, including methods and resources for data management.
  - Obtain: Acquire data from NOAA-owned and non-NOAA owned systems, including data collection and analysis, data integration, data buys, and models.
  - Process: Process data to make it useful to users and create products.
  - Preserve: Identify essential records, determine data required to be preserved, determine the method of preservation, align preservation with records retention schedules, and coordinate discovery and access to archived records.
  - Access: Ensure timely internal and public access to data and metadata as appropriate, with no or limited restrictions on use, clearly marked with an open data license or as controlled unclassified information.
  - Disposal: Removal of data based on the record retention schedule.
- .03 With cross-cutting components of:
  - Describe: During data collection and creation, document data and collection processes through metadata to ensure provenance and independent understandability.
  - Track and Monitor: Track data and metadata throughout the lifecycle and monitor application of data management principles.
  - Quality: Ensure the quality, objectivity, utility, and integrity of the data.
  - Security: Protect data from unauthorized access, corruption, and theft.

**SECTION 5. IMPLEMENTATION.**

Implementation will be described in the Handbook that provides detailed guidance on the above lifecycle components. The Handbook will be developed and maintained by the NOAA Data Governance Committee (DGC). The Handbook inherits the requirements and authority of this policy.

**SECTION 6. RESPONSIBILITIES.**

- .01 The NOAA Chief Data Officer (CDO) in coordination with the NOAA Assistant Chief Data Officers (ACDO) and the NOAA Data Governance Committee (DGC) shall be

<sup>2</sup> Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

- .02 Data Asset – collection of data elements or data sets that may be grouped together.<sup>3</sup>
- .03 Environmental Data – Data that includes, but is not limited to: 1) recorded and derived observations; 2) measurements of the physical, chemical, biological, geological, and geophysical properties and conditions of the oceans, atmosphere, space environment, sun, and solid earth; 3) correlative data, such as socio-economic data; 4) model outputs using or predicting data; and 5) related documentation and metadata (from NOAA 212-15A).
- .04 Administrative Data – Administrative data is derived from the operation or management of an organization or institution,<sup>4</sup> and is collected for the purposes of registration, transaction, and record keeping and often associated with the delivery of a service.<sup>5</sup> Administrative data differs from experimental, scientific, and observational data in that it is found (rather than systematically made) data and is not primarily collected for environmental research purposes. As such, it can be large, complex, and not necessarily collected in an organized manner that allows for linkages to other information or data.<sup>6</sup> Through the curation, enhancement, documentation, and accessibility of administrative data, NOAA can better understand how it achieves its mission and serves society.<sup>5</sup> This is especially important for the research and observational components of NOAA whose value in decision making and societal impacts is found elsewhere in the agency. Applications of administrative data allow for research in service delivery, user engagement, program management, social science, and economic valuation. Administrative data can include personally identifiable information or other sensitive, controlled unclassified information. Public data collection is also subject to the Paperwork Reduction Act. Data collection and warehousing will follow Federal and NOAA policies where applicable.
- .05 Information – Any communication or representation of knowledge such as facts, data, or opinions in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual forms.<sup>7</sup>
- .06 Data Lifecycle – The stages through which information passes – typically characterized as a creation or collection, processing, dissemination, use, storage, and disposition – to include destruction and deletion.<sup>8</sup>

<sup>3</sup> Definition from the Evidence Act, 44 U.S.C. § 3502 (17)

<sup>4</sup> Elias, P. (2014). *Administrative data. Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. Berlin: Sciendo Verlag, 47.

<sup>5</sup> Woolard, M. (2014). 3.1 administrative data: Problems and benefits, a perspective from the united kingdom). *Facing the future: European research infrastructures for the humanities and social sciences*, 49.

<sup>6</sup> Cunniff, R., Playford, C. J., Gayle, V., & Döbken, C. (2018). The role of administrative data in the big data revolution in social science research. *Social science research*, 59, 1-12.

<sup>7</sup> McGrath-Lone, L., Jay, M. A., Blackburn, R., Gordon, E., Zylberstein, A., Wijlans, L., & Gilbert, R. (2022). What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *International journal of population data science*, 7(1).

<sup>8</sup> As defined in OMB Circular A-130:

<https://oig.dhs.gov/eoarchive/archives/data/omb/memoranda/2013/m-13-12.pdf>

<sup>9</sup> OMB Circular A-130, Managing Information as a Strategic Resource

responsible for managing this Order in consultation with the appropriate NOAA Councils and Committees.

- .02 The DGC shall develop, approve, and maintain the Handbook associated with this Administrative Order.
- .03 The NOAA CDO and ACDOs through the DGC shall monitor compliance with and reporting on this policy and associated with the Handbook throughout the data management lifecycle.
- .04 The DGC will report at least annually on the implementation of this NAO to the NOAA Chief Information Officer Council and NOAA Observing Systems Council.

**SECTION 7. REFERENCES.**

The NOAA CDO will maintain a list of applicable authorities and references and will provide access to their electronic editions.

**SECTION 8. EFFECT ON OTHER ISSUANCES.**

This order supersedes NAO 212-15A, Management of Environmental Data and Information, effective November 4, 2010, which superseded NAO 212-15, Management of Environmental and Geospatial Data and Information, dated December 2, 2008.

An electronic copy of this Order will be posted in place of the superseded Order on the NOAA Office of the Chief Administrative Officer website under the NOAA Administrative Issuances Section.

Under Secretary of Commerce  
for Oceans and Atmosphere

Office of Primary Interest:  
NOAA Office of the Chief Information Officer (OCIO)

圖31、NAO 212-15B: Management of NOAA Data and Information(全文)

### 3、開放資料(Open Data)

#### NOOD—NOAA Open Data Dissemination

NOAA 的開放資料傳播 (NOOD) 計劃始於 2015 年，作為其 Big Data Partnership 規

劃的一環，是一項與3家雲端服務商 Amazon Web Services (AWS)、Microsoft Azure (Azure)和 Google Cloud Platform (GCP)的實驗性合作，透過3年合作協定，嘗試使用新興雲技術向更廣泛的受眾免費公開 NOAA 環境資料(圖32)。NOAA 每天自其眾多觀測和感測平臺(包括衛星、感測器、模型和預報)產出數 TB 的資料，包含地球系統從大氣到海洋的各方面資訊，相關資訊被政府、研究人員和私人企業用來瞭解從自然災害動態到氣候變化的一切。NODD 負責提供可取得資料的傳播平臺，並非資料倉儲，倉儲工作由 NCEI 承擔。

NODD 計畫履行美國政府“開放政府資料法案(The Open Government Data Act)”的要求——政府資料以開放為原則，應以開放格式完全公開 NOAA 資料，並提供匿名查詢、取用，將政府公共資料資產發佈為機器可讀、維護資料清單，讓公眾參與、自由使用。NODD 目前提供許多關鍵的 NOAA 業務資料集，可參考其網站：

<https://noaa.gov/nodd/datasets>。

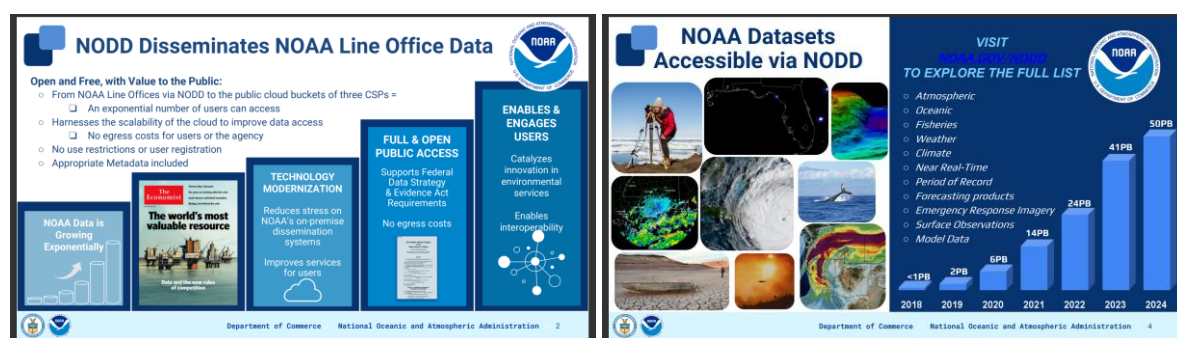


圖32、NODD 的重點工作、主要資料集和供應資料量

與各雲服務業者簽訂的 NODD 協定建立了三贏的公私合作夥伴關係，使雲服務提供者、NOAA 和資料使用者皆受益；NOAA 將自行維運開放資料的花費轉為長期、較低成本的合作，並獲得高效能的資料傳播平臺，雲服務業者藉由使用者就近使用付費雲計算資源進行資料分析應運獲利，資料使用者則能自高效能、低延遲的平臺免費取得環境資料。整體服務以需求為導向，由雲服務業者、公共和私營企業、資料科學家和研究人員參與，共同識別高價值、高影響力的資料集並確定其開放優先順序。

2015年 NODD 在 AWS 首先開放的是 NEXRAD 氣象雷達資料，這是受許多產業界使用者歡迎的資料集，用於分析、了解冰雹、雨、雪、龍捲風等風險，但當時的資料大小已達270 TB，且是以保存為目的的方式存放在 NCEI 的磁帶中，以致資料公開給大眾時，使用者取得費時費力。在 AWS 開始在雲端提供自1991年起的 NEXRAD 資料，並在其平臺同步提供資料處理的工具、服務及使用手冊之後，80%的資料取用需求自 NCEI 轉向 AWS，NCEI 的資料供應負載下降約60%，資料利用率(被取用次數)提高了



2.3倍，並且有64%被取用的 NEXRAD 資料留在 AWS 平臺上(圖33)。

至 2024 年，NODD 已託管超過 50 PB 資料在各雲端平臺，每日約有1.2 PB 的下載資料量，120~150萬次的資料取用 (包含直接及透過 API 存取)。並且依據分析，資料取用的尖峰值常是由於大規模取用單一資料集/產品，而非分散在多個資料集，受益於雲端服務的特性，這種規模的資料取用並未對 NODD 服務造成問題，若是傳統 FTP 或 HTTP 網站可能將面對連結數量達到上限或下載延遲的問題。高效能併行訪問和低延遲讓使用 NODD 服務的系統可以一次性取得大量資料檔案，為快速進行整個資料集 (包含歷史資料) 的完整分析創造了機會，且這些大量但短時的分析時常是在雲端平臺上進行，隨著時間的推移，NODD 的雲內和跨雲存取量呈上升趨勢。

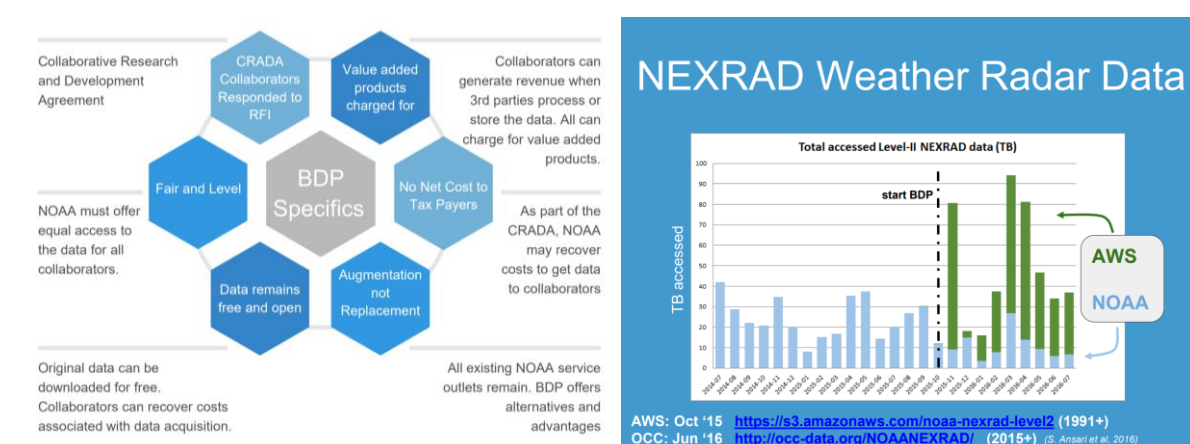


圖33、Big Data Project 的主要概念及 NEXRAD 資料於 AWS 上線後的供應變化

為了支援 NODD 在雲端的服務，資料的處理流程及背後的基礎架構是關鍵。NODD 依靠 NOAA 內部的許多來源來提供資料，資料源包含衛星、各項環境觀測、NOAA 內部運行的模型等，通過 NOAA 各系統收集、整理、存檔和傳播，這些資訊由 NOAA 內部管理，並將資料集分配提供給雲服務業者 (部分資料集為共同持有，部分則專屬單一業者供應)。

NOAA 通過 Weather and Climate Operational Supercomputing System (WCOSS)、Comprehensive Large Array Data Stewardship System (CLASS)、Production Distribution and Access (PDA) 系統、NCEP Production FTP Server (FTPPRD) 持續更新、供應資料，並使用高彈性、高效能和容錯的 Apache NiFi 部署將資料傳輸到三個雲服務業儲存中，流程中持續監控各資料源，並進行資料品管。每個資料集來源都建置備援及故障切換機制 (failover)，通過相關架構，可以快速有效處理 NODD 上游的中斷，維持資料可用性。

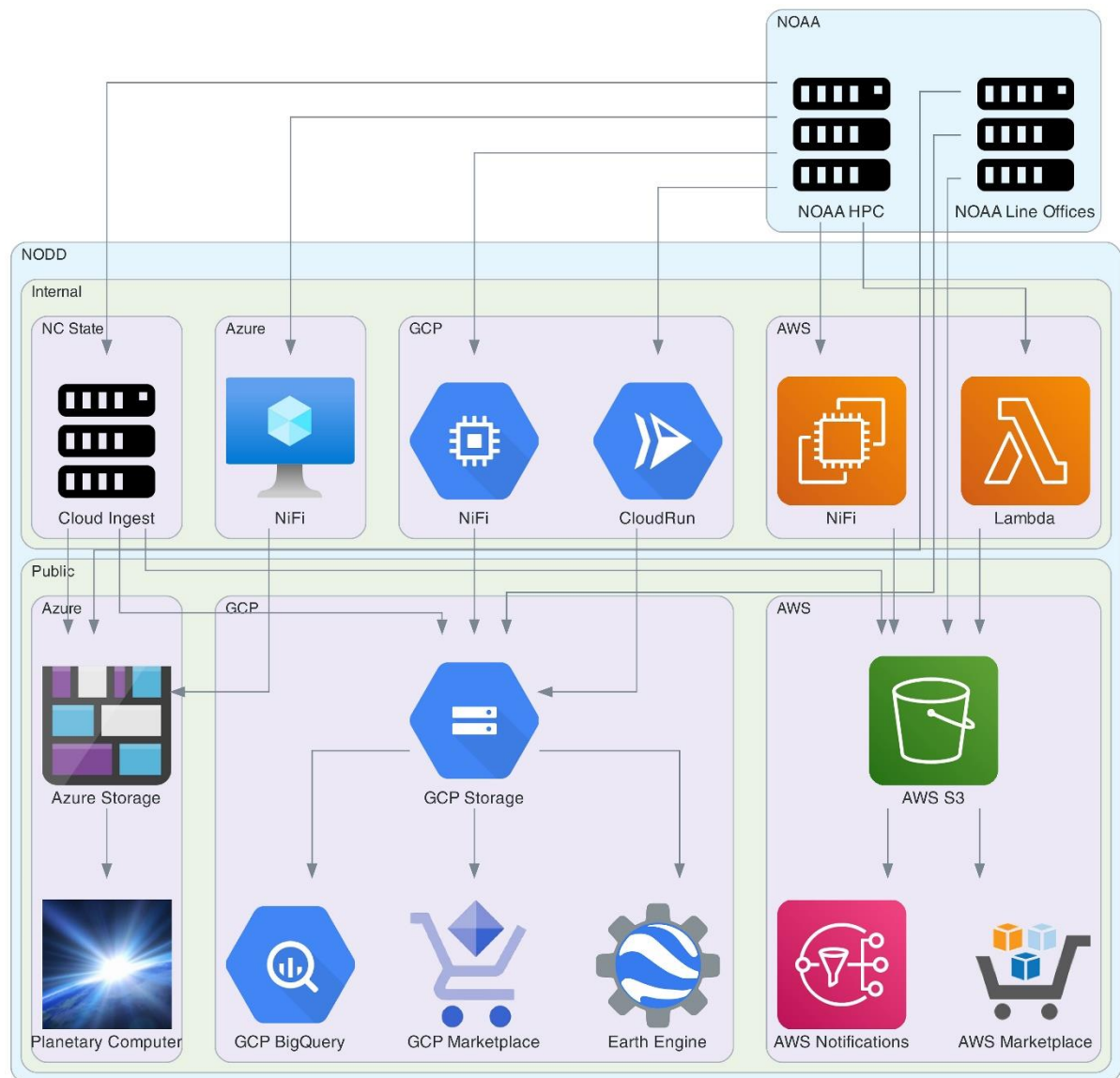


圖34、NODD 的產製、供應資料流程

雲端的 NODD 資料主要保存在儲存物件 (object storage) 中：AWS 上的 Simple Storage Service (S3)、Microsoft Azure 上的 Azure Blob Storage (ABS) 和 GCP 上的 Google Cloud Storage (GCS)。儲存物件是高效能資料供應服務的基礎層，在其上則由各雲服務平臺上提供許多功能，包括雲計算服務、支援事件驅動型工作流程的事件通知 (如 AWS Notification)、Google Earth Engine 和 Microsoft Planetary Computer 等。更進一步，雲服務業者建立自動流程，用事件驅動將部分資料轉換為更易於使用的雲原生格式 (如 Cloud Optimized Geo-Tiffs、Zarr 和 Parquet)，並提供相應的 STAC (SpatioTemporal Asset Catalog) API，可以擷取結構化資料的特定時間、空間範圍子集。NODD 使用者可以使用基本形式取得資料，例如通過程式設計方式和互動式的 Web 瀏覽器，也可進一步運用平臺功能。(圖34)

## **ECMWF Open Data**

ECMWF 的預報模式產品一直是全球領先的水準，但自其1975年成立以來，長期維持 close data 的政策，1995年之前，資料只提供成員國使用，到1995年開始商業銷售資料，2000年才首度釋出非核心資料並限制為研究使用。直到2019年底，隨著開放資料已經被歐盟的大多數成員國認可為最大化資料社會經濟效益的重要方式，理事會決議 ECMWF 的資料政策應轉為免費、開放，並將轉向開放資料的戰略方向調整，寫入“ECMWF STRATEGY 2021–2030”中，目前目標是在 2026 年底前實現完全開放的資料狀態。在2022年初，ECMWF 開始對公眾開放其部分核心資料，目前已公開其 Integrated Forecasting System (IFS) and Artificial Intelligence/Integrated Forecasting System (AIFS) 模式資料，模式解析度為 0.25 度。

隨著資料政策轉向，相關的內部挑戰隨之而來。首先是自2011年到2022年逐年上升的資料銷售收益，預期將開始下降並在2026年歸零 (仍會有遞送服務及相應收費)，ECMWF 資料銷售的所得，全數皆挹注為單位預算，在2022年貢獻為總預算的11%，此部分的資金填補將成為主要課題。隨之需要處理的包括資金短少後，內部資料增量所需要的資源如何取得；過往可以向付費用戶了解使用需求、情境等訊息，並用以調整模式發展、資料產製方向，未來如何向匿名的免費用戶收集；然後是 ECMWF 資料有上百種授權方案，如何調整及向使用者明確說明，及如何提供未來的大量使用者服務，(目前付費客戶450+，每年提出約45萬大小需求及問題。) 當然還有如何支應使用大幅增加所需的軟硬體更新需求。(圖35、圖36)

ECMWF 的開放資料，後端使用與商業銷售資料一致的系統進行供應，並套用同樣的資料命名規則，模式 GRIB 資料皆搭配相應的 index 資料。為避免資料取用的網路負載過度衝擊銷售資料的遞送，資料延遲1小時釋出，並在線上保留3天。2023年初的下載量約為每日9 TB，90萬次存取要求，每日00Z、12Z 資料供應時段的尖峰負載使 ECMWF 尋求雲端服務業者的解方，目前除了其自身的下載點外，也透過 Microsoft、Amazon、Google 和其他較小的獨立使用者 (例如 Open-meteo.de) 供應資料，以分散負載並擴大接觸的使用者，並且提供開放工具如 ecmwf-opendata Python package、線上分享 Jupyter 筆記協助資料使用。

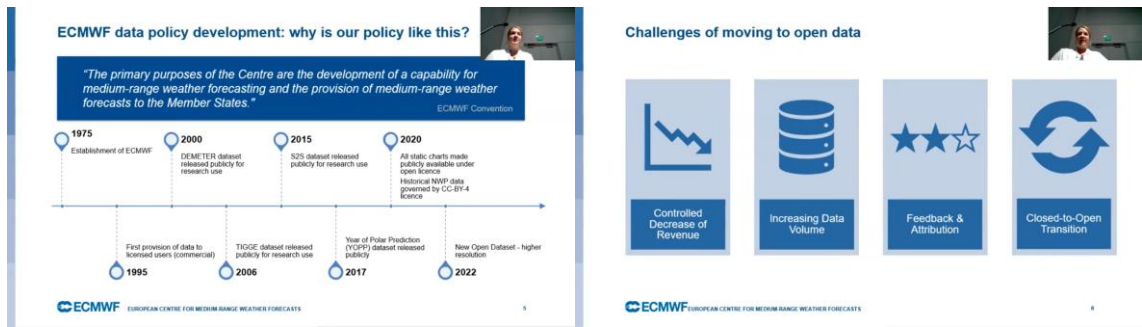


圖35、ECMWF 資料政策的演變、轉向開放資料後面臨的挑戰

## ECMWF's Open Data Timeline

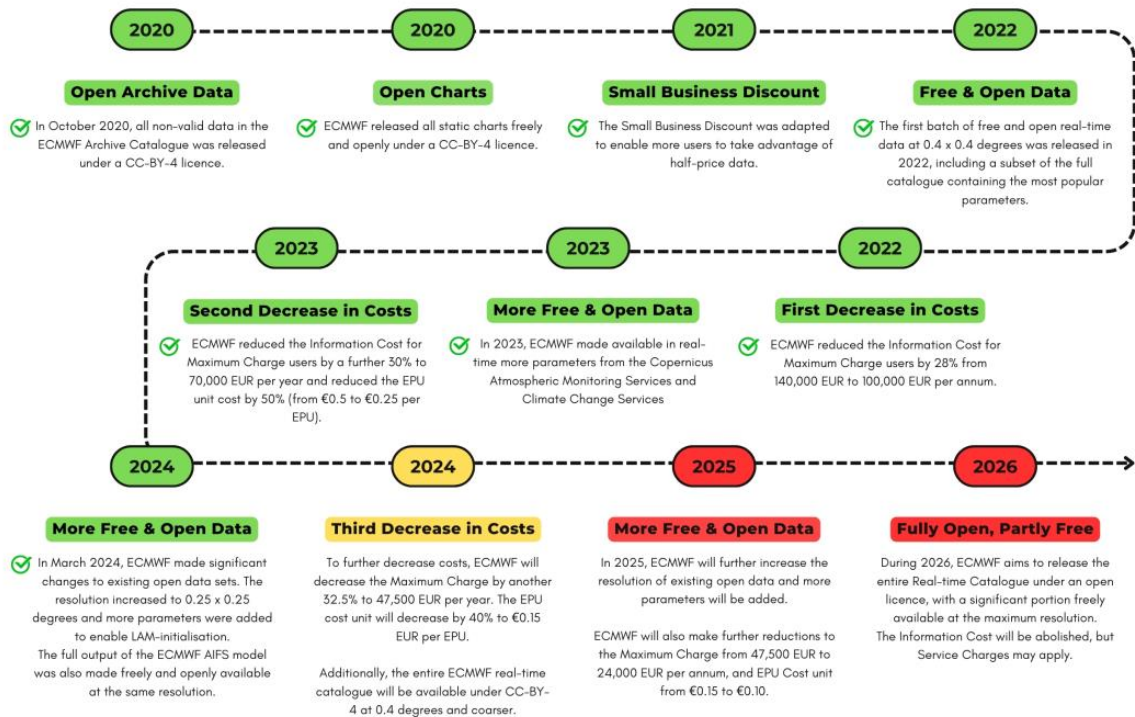


圖36、ECMWF 的資料開放時程 (紅色部分尚待批准)

### 4、NCEI 資料服務 – OneStop 及其他

NOAA 及其合作夥伴每年花費數十億美元收集各種環境資料，NCEI 則負責免費和開放提供大量資料及永久保存這些資料。因此，NCEI 的資料管理能力需要持續擴展並藉由與使用者互動調整，以滿足 NOAA 內外不斷增長、變化的需求，也成為提供符合透明度、完整性、高品質資料的供應單位。(圖37)



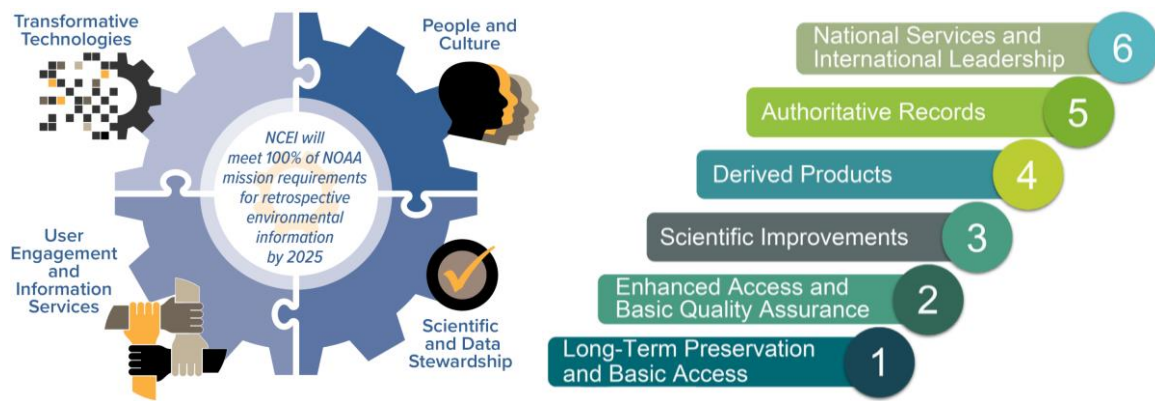


圖37、NCEI Aspiration、Tiers of Data Stewardship

### **NOAA OneStop**

OneStop (<https://data.noaa.gov/OneStop/>) 是 NOAA 將地球物理、海洋、海岸、天氣、氣候、自然災害和太空觀測資料集中的環境資料搜尋平臺，還提供多媒體預覽和直接資料下載。OneStop 於 2015 年提案，並於 2017 年底啟動，當時屬於 NOAA Big Data Project 的一環，目的是擴展環境資料獲取途徑並配合開放政府計劃，NCEI 收存的每個具有詮釋資料的資料集都可以在 OneStop 中找到，擴大潛在用戶的數量，提升 NOAA 資料的使用率。

OneStop 專案的目標是改進 NOAA 的資料搜尋和取得的服務框架，專案不只包含提供簡單、易用資料搜尋體驗的 Web 使用者介面，更要建立資料基礎設施 (儲存、串接、管理...)，解決資料格式的問題和實作詮釋資料，提高資料集搜尋結果的相關性，改進資料和詮釋資料管理。

OneStop 是提供民眾作為資料發現與取用 (data discovery and access) 的平臺，由幾種服務 (Services) 組成其整體架構，包含安全 (Security)、儲存 (Storage)、資料存取 (Data Access)、詮釋資料管理 (Metadata Management)、目錄 (Catalog) 及使用者介面 (User Interface) 服務。(圖38、圖39)

- 安全服務 (Security Service)：提供基於規則的資通安全相關服務，對所有資料和詮釋資料介面的存取進行身份驗證、授權和稽核，包含弱點掃描、日誌管理、防火牆、雙因素身份驗證、使用者權限控制等。
- 儲存服務 (Storage Service)：針對包含海洋、氣候、模型、衛星...等高優先順序資料集提供線上存取服務 (包含雲端資料取用)，對於僅可從磁帶獲得的資

料，則將使用者連接到舊有的資料取用服務點，以檢索選定的資料。

- 資料存取服務 (Data Access Service)：對提供下載或遞送的資料集進行子集設定、擷取、視覺化和依需求轉換格式的功能，包含資料伺服器、OPeNDAP 網路資料訪問協定、Live Access 網路伺服器、ArcGIS 地圖視覺化伺服器...等，及 FTP 和 HTTPS 等基本下載服務。
- 詮釋資料管理服務 (Metadata Management Service)：提供詮釋資料儲存庫的使用者介面，由資料管理員編輯、強化和管理供給搜尋介面的詮釋資料內容；包含詮釋資料編輯工具、支援由外單位收集記錄和索引到目錄服務中的工具、及用於評估詮釋資料記錄的完整性和嚴謹性的工具。
- 目錄服務 (Catalog Service)：提供工具和元件，通過標準介面管理、搜尋和供應整合 (collection-level) 或顆粒化 (granule-level) 的詮釋資料；包含社群廣泛使用搜尋服務、使用者查詢紀錄搜尋工具如 ElasticSearch、與搜尋引擎綁定的資料排名工具以優先呈現最相關的結果、能處理不同顆粒度詮釋資料的擷取/轉換/載入 (ETL) 工具。
- 使用者介面服務 (User Interface Service)：提供標準化的方式來瀏覽 OneStop 基礎設施提供的儲存和搜尋功能。包括：簡潔的用戶介面供使用者發現資料，帳戶管理元件以管理使用者檔案、保存搜尋和對資料集進行評分，資料集展示介面、進階(欄位)搜尋。

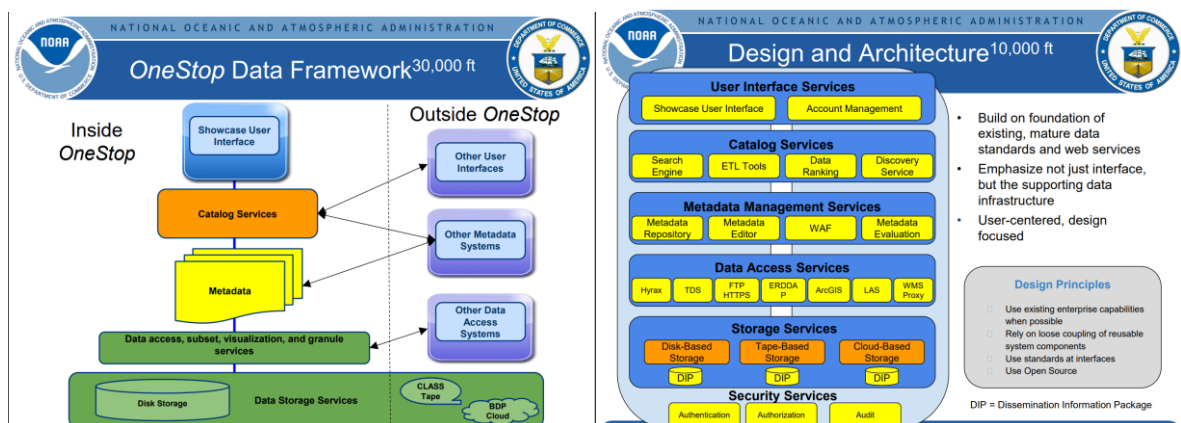


圖38、OneStop 的資料框架、設計架構

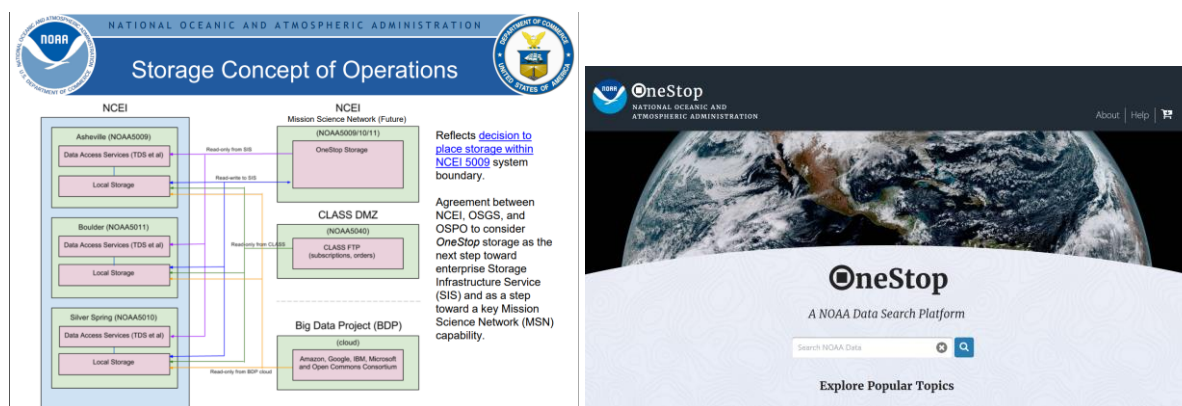


圖39、OneStop 的資料框架、設計架構、儲存規劃與服務網頁截圖

### OneStop Readiness

為了確保與提升品質，OneStop 定義了一套資料和詮釋資料就緒的衡量標準，資料管理員可以依據相應標準來提高 NOAA OneStop 架上資料和詮釋資料的可發現性、可取用性和理解性。就緒標準分為三個級別：銅、銀和金。

“銅”是最低的就緒級別，滿足 OneStop 資料合規性的最低要求：詮釋資料必須完全符合 ISO 19115-2 詮釋資料標準，達到銅級的資料就緒性符合 OneStop 索引的可發現性要求。OneStop 資料就緒的第二級是“銀”，除 ISO 標準外，需再符合2項要求，資料管理員有機會可以經由 OneStop ISO-Lite 範本來檢索細部資料，透由資料管理成熟度矩陣 (DSMM) 評估結果進行資料相關性排名演算，資料集檔案被轉換為機器可讀和方便分析的格式 (例如 NetCDF、HDF、GeoTIFF 等)。準備狀態的最後一個級別是“金”，在 OneStop 中提供最大的可發現性，詮釋資料需符合 ISO 標準、資料提供直接下載選項、內有參數或格點的資料應符合 OneStop ISO-Lite 標準、有 DSMM 評估、以及資料採用標準化、非專有的機器可讀格式。(圖40、圖41)

<b>Gold Readiness</b>	
All Required - Maximum Discoverability in <i>OneStop</i>	
●	ISO 19115-2/1 Compliant Collection Metadata Records
●	Data are online with direct download options
●	<i>OneStop</i> ISO-Lite compliant metadata records (for collections with granules)
●	Each collection shall have a DSMM Assessment
●	Data are in a standardized, non-proprietary machine-readable format

圖40、NOAA OneStop 詮釋資料就緒標準金級要求

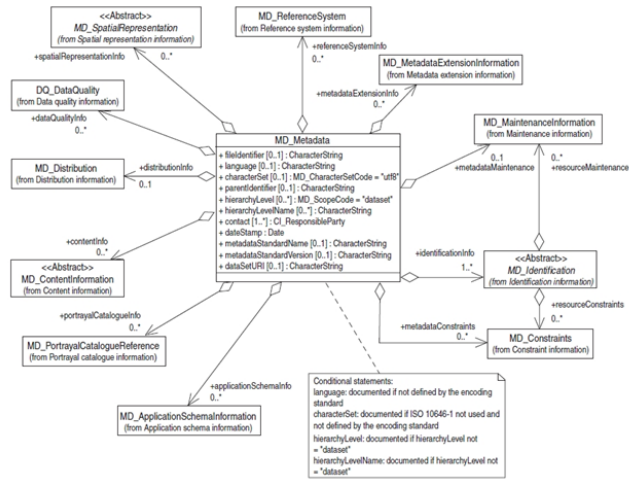
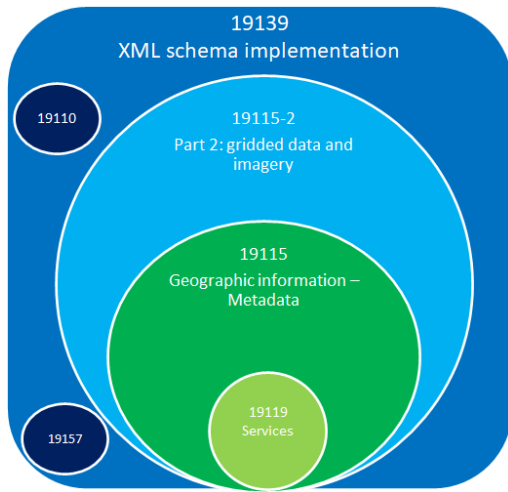


圖41、ISO 標準中關於詮釋資料的要求

### ATRAC - Advanced Tracking and Resource Tool for Archiving Collections

在透過 NOAA OneStop 等平臺供應資料之前，資料擁有者需先向 NCEI 註冊資料及加入倉儲系統。倉儲系統需確保對資料的完整紀錄，以便未來的使用者可以在沒有產製者協助的情況下充分運用資料，因此關於資料的專業資訊必須通過妥善的品質管理要求與資料一起存檔。ATRAC (<https://www.ncei.noaa.gov/archive/atrac/index.html>) 是 NCEI 用於進行資料倉儲的管理工具，提供申請資料倉儲 (archive)、在 Web 表單上為資料建立符合 ISO 標準的詮釋資料，查詢已建立的專案目前狀態和摘要資訊等。(圖42)

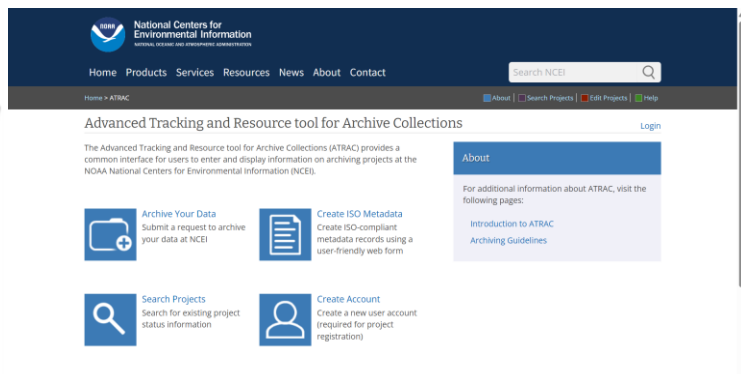
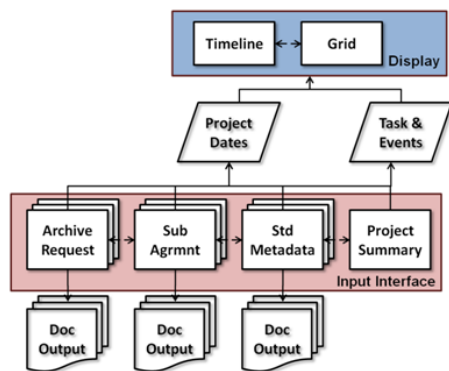


圖42、ATRAC 的功能流程與服務網頁截圖

在向 NCEI 提出存檔請求之前，資料擁有者需先評估資料的大小和複雜性 (如檔案建議切割為25-50GB)、提供資料樣本及完整說明或詮釋資料供評估，依標準進行格式、編碼轉換 (如 self-describing 的 NetCDF、HDF 等)、檔案命名的調整 (具可讀性並排除特殊字元)，規劃資料生命週期內 (研究或作業期間) 的資料轉移方式，運行期間端



到端的資料傳輸方式與品質確保等，並尋求必要的支援。(圖43)

**Request to Archive  
With The National Centers for Environmental Information  
For National Environmental Record Management System (Images and Publications)  
Collections  
Provided by NOAA/NCEI**

**2024-04-19**

This information will be used by NCEI to conduct an appraisal and make a decision on the request.

**1. Who is the primary point of contact for this request?**  
Jason Cooper  
NCDC/RSAD/AB  
Archivist  
828-271-4382  
jason.cooper@noaa.gov

**2. Name the organization or group responsible for creating the dataset.**  
DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce

**3. Provide an overview summarizing the scope of data you want to archive. Describe the outputs, data variables, including their measurement resolution and coverage.**  
This request encompasses all ~130 collections held in NCEI's National Environmental Records Management System, or NERMS. NERMS is a web application serving as the primary ingest and access point for NCEI-stewarded collections comprised of human-readable image formats such as PDF or JPG. These files may have been created by migrating a record from paper or microfilm format, but increasingly contain born-digital files. Collections can be publications, observational forms, or maps/charts. Observational forms and charts are the original data source for hundreds of millions of environmental data elements across four centuries.

**4. What is the time period covered by the dataset? (YYYY-MM-DD, YYYY-MM or YYYY)**  
From 1735-01-01  
Ongoing as continuous updates to the data record

**5. Edition or version number(s) of the dataset:**  
2.0

**6. Approximate date when the dataset was or will be released to the public:**  
2023-09-30

**7. Who are the expected users of the archived data? How will the archived data be used?**  
Data are used by a wide variety of customers, from researchers validating data in environmental datasets created by keying the data from these files, to attorneys, insurance industry, and historians. Files in image formats contain observations that can be transcribed and archived as structured datasets.

Printed 2024-11-14 - Verify Document Currency Before Use 1

---

Number of Data Files: 47750987

**Continuous Data: data volume rate for a continuous data production.**  
Total Data Volume Rate: 50MB per Month  
Data File Frequency: 10 per Month  
Data Production Start:

**18. Are later updates, revisions or replacement files anticipated? If so, explain the conditions for submitting these additional data to the archive.**  
No additional updates, revisions or replacement data are anticipated.

**19. Describe the server that will connect to the ingest server at NCEI for submitting the data.**  
Physical Location: Asheville, NC  
System Name: NERMS01-PROD.ncei.noaa.gov  
System Owner: DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce  
Additional Information:

**20. What are the possible methods for submitting the data to NCEI? Select all that apply.**  
1. FTP PUSH  
Push to ceph bucket

**21. Identify how you would like NCEI to distribute the data. Web access support depends on the resources available for the dataset.**  
1. User interface to order and stage data for download

**22. Will there be any distribution, usage, or other restrictions that apply to the data in the archive?**

Constraint Type	Description
Access	A small number of collections contain PII in the form of observer names, phone numbers and home addresses, which must be restricted from public access.

**23. Discuss the rationale for archiving the dataset and the anticipated benefits. Mention any risks associated with not archiving the dataset at NCEI.**  
While primary access will come from the NERMS web application and the underlying filesystem, this project will result in a better-structured collection of collections that can easily be used to repopulate the access filesystem in the event of catastrophic failure. Entire collections could be staged for bulk access.

**24. Are the data archived at another facility or are there plans to do so? Please explain.**  
No

**25. Is there an existing agreement or requirement driving this request to archive? Have you already contacted someone at NCEI?**  
These data are already archived at NCEI. This request completes modernization of NCEI's management of collections in image formats and establishes a more usable structure for the archive.

Printed 2024-11-14 - Verify Document Currency Before Use 3

**8. Has the dataset undergone user evaluation and/or an independent review process? Did NCEI participate in design reviews?**  
The data are already in the archive, and are also accessible via the NERMS application. NCEI has managed and provided access to these collections for >20 years.

**9. Describe the dataset's relationship to other archived datasets, such as earlier versions or related source data. If this is a new version, how does it improve upon the previous version(s)?**  
The vast majority of data in this submission is already in the archive as DSI 3900\_01, NCEI's Imaged Weather and Climate Document Repository. These files we uploaded to or included in the previous system, EV2. New data has been added as the migration from EV2 to NERMS is nearly complete. Once file checks are complete, NCEI may elect to remove DSI 3900\_01 from the archive.

**10. List the input datasets and ancillary information used to produce the data.**  
The collections come to NCEI from many sources including NOAA line offices, the US military, and the National Archives. Also included are publications created by NCEI and its predecessor Centers.

**11. List web pages and other links that provide information on the data.**  
<https://www.ncei.noaa.gov/nerms/Home>

**12. List the kinds of documents, metadata and code that are available for archiving. For example, data format specifications, user guides, algorithm documentation, metadata compliant with a standard such as ISO 19115, source code, platform/instrument metadata, data/process flow diagrams, etc.**  
1. Metadata records for collections in NERMS already exist, and can be found at <https://docs.google.com/spreadsheets/d/1gXcyZEf67a0xELdYs9ZAqlpzd0YMKVNGcBKSyqu-YEw/edit?usp=sharing>. They will need to be updated to show NERMS and Common Access as access points.

**13. Indicate the data file format(s).**  
1. PDF  
2. JPG  
3. TXT  
4. JSON

**14. Are the data files compressed?**  
gzip

**15. Provide details on how the files are named and how they are organized (e.g., file\_name\_pattern\_YYYYMM.tar in monthly aggregations).**  
The millions of granule files, individual PDFs and JPGs, will be packaged into tar files organized by collection. The vast majority of collections are >1GB, so we can split collections multiple tar files as defined by archive file size limitations. Each tar file will include json file of record metadata and limited collection information

**16. Explain how to access sample data files and/or a file listing for previewing. If it is not available now, when will it be available?**  
Granules are available at NERMS website. Sample tar files can be generated upon request.

**17. What is the total data volume to be submitted?**  
**Historic Data: all historic data or data submitted as a completed collection.**  
Total Data Volume: 12.5TB

Printed 2024-11-14 - Verify Document Currency Before Use 2

---

**26. Do you have a data management plan for your data?**  
No

**27. Have funds been allocated to archive the data at NCEI?**  
No

**28. Identify the affiliated research project, its sponsor, and any project/grant ID as applicable.**  
N/A

**29. Is there a desired deadline for NCEI to archive and provide access to the data?**  
Archive by: 2024-06-14  
Accessible by: 2024-09-30

**30. Add any other pertinent information for this request.**  
None

Printed 2024-11-14 - Verify Document Currency Before Use 4

圖43、NCEI 資料倉儲申請表單內容



## DSMM (Data Stewardship Maturity Matrix)

為資料集準備高品質的描述性資訊和詮釋資料是滿足透明度要求及建立產品信賴度的必要步驟，但需要跨領域知識和密切合作才能進行，並且資料集的有用性很大程度上取決於資料提供者和管理員評估資料品質、描述資料內容和展示資料應用性的能力。為了評估工作進行的品質，NCEI 使用資料管理成熟度矩陣 (DSMM)，作為衡量 NOAA OneStop 單一資料集管理成熟度的參考框架。(圖44、圖45)

DSMM 模型指定了九項關鍵管理屬性 (可保存性、可取得性、可用性、生產可持續性、資料品質保證、資料品質控制/監控、資料質量評估、透明度/可追溯性和資料完整性)，並為九個關鍵組成中的每一項定義了五個層級的成熟度量表(全無管理、有限管理、管理被定義但僅部分實作、管理被定義且完整實作、管理+控制+稽核)，用於評估管理實踐程度，並為改進資料品質的提供方向。藉由 DSMM 評估，可以一致、系統地管理和公佈資料的品質，增強資料的可查找性和可用性。

Document ID: NCDC-CICS-SMM\_0001  
Version: Rev. 1. 12/09/2014

**Dataset Name**

Maturity Level as of mm/dd/yyyy

### Stewardship Maturity Matrix for Digital Environmental Data Products

Maturity Scale	Preservability	Accessibility	Usability	Production Sustainability	Data Quality Assurance	Data Quality Control/Monitoring	Data Quality Assessment	Transparency /Traceability	Data Integrity
<b>Level 1 – Ad Hoc Not Managed</b>	Any storage location Data only	Not publicly available Person-to-person	Extensive product-specific knowledge required No documentation online	Ad Hoc or Not applicable No obligation or deliverable requirement	Data quality assurance (DQA) procedure unknown or none	None or Sampling unknown or spotty Analysis unknown or random in time	Algorithm/method/model theoretical basis assessed (method and results online)	Limited product information available Person-to-person	Unknown or no data ingest integrity check
<b>Level 2 - Minimal Managed Limited</b>	Non-designated repository Redundancy Limited archiving metadata	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable	Non-standard data format Limited documentation (e.g., user's guide) online	Short-term (individual) PI's commitment (grant obligations)	Ad Hoc and random DQA procedure not defined and documented	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 1+ Research product assessed (method and results online)	Product information available in literature	Data ingest integrity verifiable (e.g., checksum technology)
<b>Level 3 - Intermediate Managed Defined, Partially Implemented</b>	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving process standards	Level 2+ Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Community Standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Medium-term Institutional commitment (contractual deliverables with specs and schedule defined)	DQA procedure defined and documented and partially implemented	Level 2+ Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 2+ Operational product assessed (method and results online)	Algorithm/method/model Theoretical Basis Document (ATSD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 2+ Data archive integrity verifiable
<b>Level 4 - Advanced Managed Well-Defined, Fully Implemented</b>	Level 3+ Conforming to community archiving standards	Level 3+ Community-standard data services Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 3+ Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Long-term Institutional commitment Product improvement process in place	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 3+ Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 3+ Quality metadata assessed (method and results online) Limited quality assessment metadata	Level 3+ Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 3+ Data access integrity verifiable Conforming to community data integrity technology standard
<b>Level 5 - Optimal Level 4+ Measured, Controlled, Audit</b>	Level 4+ Archiving process performance controlled, measured, and audited Future archiving standard changes planned	Level 4+ Dissemination reports available online Future technology and standard changes planned	Level 4+ Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking	Level 4+ National or international commitment Changes for technology planned	Level 4+ DQA procedure monitored and reported Conforming to community quality metadata & standards External review	Level 4+ Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place	Level 4+ Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking	Level 4+ System information online Complete data provenance available online	Level 4+ Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported

圖44、NCEI Data Stewardship Maturity Matrix 評估表

作為 OneStop 資料就緒流程的一部分，NCEI 開發了工具、實施指南、工作流程和最佳實踐來協助 DSMM 的應用，並應用在管理和倉儲的資料集上，作為審查過程的一部分，幫助確定了現行狀態與組織程序或系統要求間的差距，也透過此流程持續提高

OneStop 資料集詮釋資料記錄的完整性和內容品質，強化資料的可查找性、可取得性、可交互運用性和可重用性的重要組成部分。



圖45、DSMM 資料流程圖、DSMM 遵循 CMMI(軟體能力成熟度)的層級架構

## 5、NCCF (NESDIS Common Cloud Framework)

NESDIS (National Environmental Satellite, Data, and Information Service) 是 NCEI 的上級單位，擁有一系列衛星、特定任務系統和資料集，隨著技術發展和更多任務的啟動，既有系統無法擴展以滿足 NESDIS 不斷增長的需求，在測試、評估後決定將資料管理業務向雲端遷移，著手構建、開發、部署 NESDIS 通用雲框架 (NCCF)。(圖46)

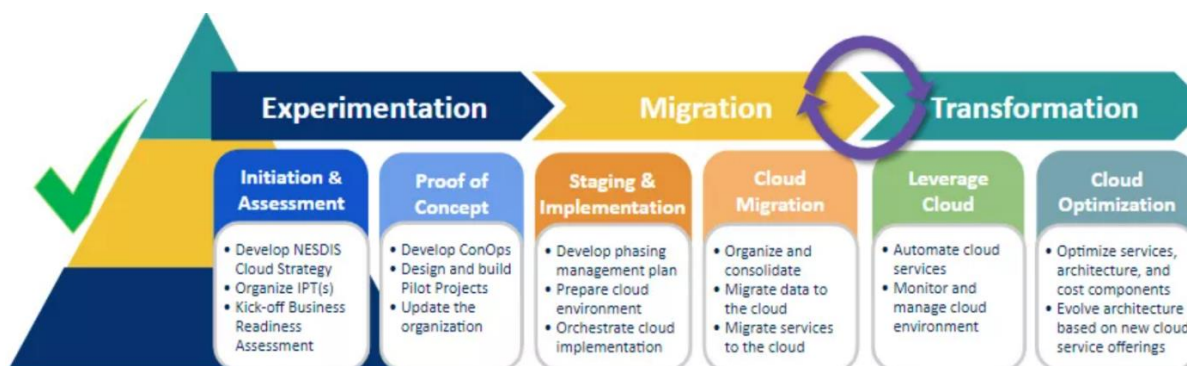


圖46、NESDIS Cloud Pilot Roadmap

計畫的緣起是在2017年，辦公室估計出在7年後，隨單位 line offices 分散多地的地面基礎設施維護、運營和維護資金將超過單位的可用資金。NESDIS 地面基礎設施由一系列定製的衛星和特定任務的資訊技術系統組成，結合資料處理、科學和管理工作流程，當時 NESDIS 維護了 100 多個關聯性鬆散的系統和子系統，每個都需要獨立的技术更新，但其中許多執行類似的功能。

NESDIS 亟需更具成本效益、可擴展性和效率的解決方案，並傾向於使用商業雲來實現這一目標，希望善用雲環境的可擴充性，同時利用遷移到雲作為系統通用功能調

整的催化劑。單位執行了一系列先導專案 (NESDIS Cloud Pilot , <https://www.nesdis.noaa.gov/nesdis-cloud-pilot>) , 確定向雲端移動的方向可行、產出服務功能雛型框架, 並評估出後續的進行時程及所需經費。(圖47)

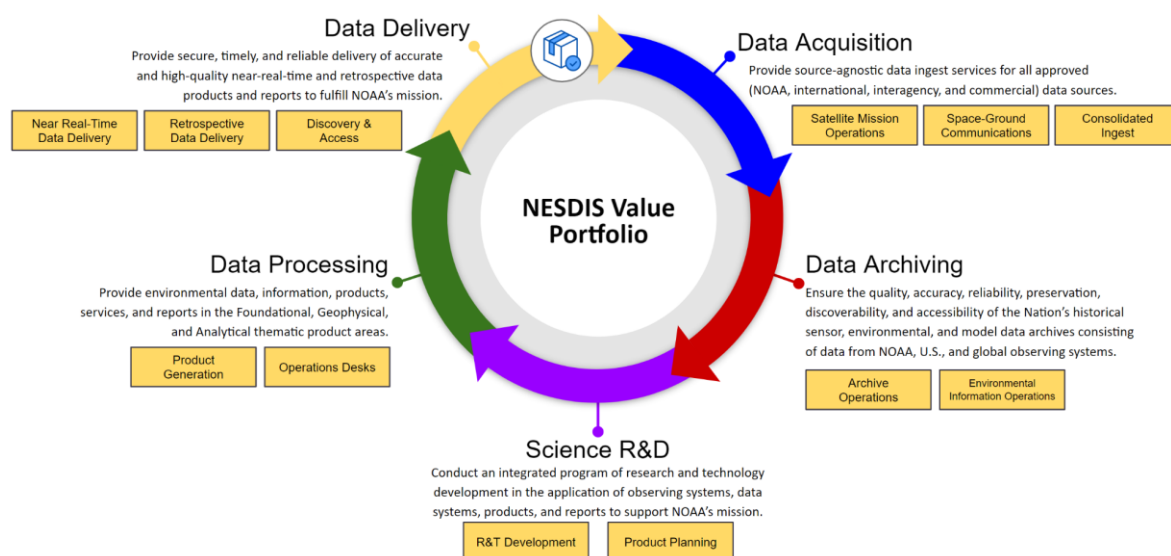


圖47、NCCF Operational Value Stream as a Continuous Operating Cycle

NESDIS 自2020年第3季開始正式建置 NESDIS 通用雲框架, 依據單位分析出的產品及服務提供 Operational Value Stream, 規劃出五個 NCCF 服務領域包括(資料)安全導入 (secure ingest)、計算和產品生成 (compute and product generation)、倉儲和儲存 (archive and storage)、科學沙箱 (science sandbox) 以及傳播和取用 (dissemination and access)。為了滿足 NESDIS 對資料高可用性的作業要求, 這些服務採用鬆散耦合 (loosely coupled)、可獨立擴展 (independently scalable), 並且是容錯 (fault-tolerant) 的架構, 一個區域的問題不會影響其他區域。同時, 為了未來可以在任何雲服務平臺上運行, 或跨平臺配置服務, NCCF 以基礎設施即程式碼 (Infrastructure as Code) 的形式構建。NCCF 在兩個雲端服務可用區域以雙作業配置 (dual processing configuration) 運行, 以確保單位期望的作業連續性。(圖48)

NESDIS 目前正在開發和部署 NCCF 五項核心服務, NCCF 使用的所有資料都通過一個導入點安全地引入, 先進行倉儲和儲存, 再用於產製產品或用於科研目的, 然後再發送或提供給使用者。其中, (資料)安全導入、計算和產品生成服務已上線, 倉儲和儲存服務正在進行作業準備, 科學沙箱以及傳播和取用服務則仍處於開發階段, 預計將於2026年完成所有核心服務功能的開發。NESDIS 已整理所有地端系統、資料和硬體的清單, 並確定優先順序, 於2024年提出一份正式遷移的計劃, 開始搬遷所屬包括 NCEI 的所有系統。

### **(資料)安全導入服務 (secure ingest service)**

此服務為通往 NCCF 的閘道，為 NESDIS、NOAA 和外部合作夥伴的資料提供了安全的單一入口，服務可在不影響效能的情況下自動擴展。相關工具可以根據不同資料源的風險評估，在檔案引入前執行安全檢查，如對每個檔案進行病毒掃描，及確認檔案中包含預期的資料。

### **計算和產品生成服務 (compute and product generation service)**

計算服務是一個可擴展、強大且容錯的資料處理環境，包括對產品生成功能的支援以及對資料的各式編輯、彙整、同化處理，由高效能計算 (HPC) 架構處理作為服務的主要骨幹，利用雲將分散的資料整合，可以縮短處理時間並降低成本。

產品生成服務建置了50多種演算法，生成200多種作業產品，提供對 NOAA 內部及重要客戶的服務，並仍引入更多的資料和處理能力，持續擴展中。後續擴充計劃包括增加自動化程式部署 (Continuous Improvement/Continuous Delivery, CI/CD pipeline)，將使演算法更安全、更快速地轉移到作業中。

### **倉儲和儲存服務 (archive and storage service)**

NCCF 以資料為最重心，倉儲和儲存服務是框架中最關鍵的服務——儲存、管理和保存 NOAA 資料。該服務集中儲存資料，確保資料不會損壞或丟失，並將其提供給所有其他服務。然後，它在雲中提供長期資料保存，並支援所有資料的持有管理和可存取性，從而在雲中實現倉儲功能。端到端的存檔和管理工作流程不僅可以在雲中完成，而且可以保持一致並提供增強的編目功能。

目前服務正在納入包括 NCEI 的檔案、管理和系統要求，預計 NOAA 資料和 NCEI 存檔和管理工作流程將遷移到 NCCF，並隨著倉儲和儲存服務的擴展而加速。未來所有資料都可供其他 NCCF 服務存取，包括即將推出的傳播和取用服務，該服務將通過公共介面提供公眾查找和取用資料。

## 科學沙箱服務 (science sandbox service)

沙箱服務為開發人員、科學家和監管員提供了一個協作平臺，可以在雲端創建新的演算法和程式碼。此平臺專為科學實驗和驗證而設計，所有合作夥伴都可以用 NOAA 的大量資料一起協同工作。使用者將通過虛擬桌面應用程式訪問科學沙箱中的程式碼、工具，並且無需下載重複資料，即可實作創新構想。服務將通過在內部編譯所有程式碼，利用 Operational Product Service 運行所有測試，然後推送到生產環境，從而支援基礎和應用研究轉換到作業的工作，NESDIS 辦公室預計規劃未來 NCCF 科學沙箱服務的有償使用。

## 傳播和取用服務 (dissemination and access service)

傳播和取用服務將為近乎即時的操作提供低延遲的資料分送，還將作為所有 NOAA 資料庫的官方公共取得位置。該服務將允許使用者以安全的方式搜尋、發現、定製和擷取資料，在滿足他們的需求同時保持資料完整性。該服務的公共取得部分確保現有服務的操作功能和介面，再提供更優化的 NOAA 資料可搜尋性並持續改善用戶體驗。

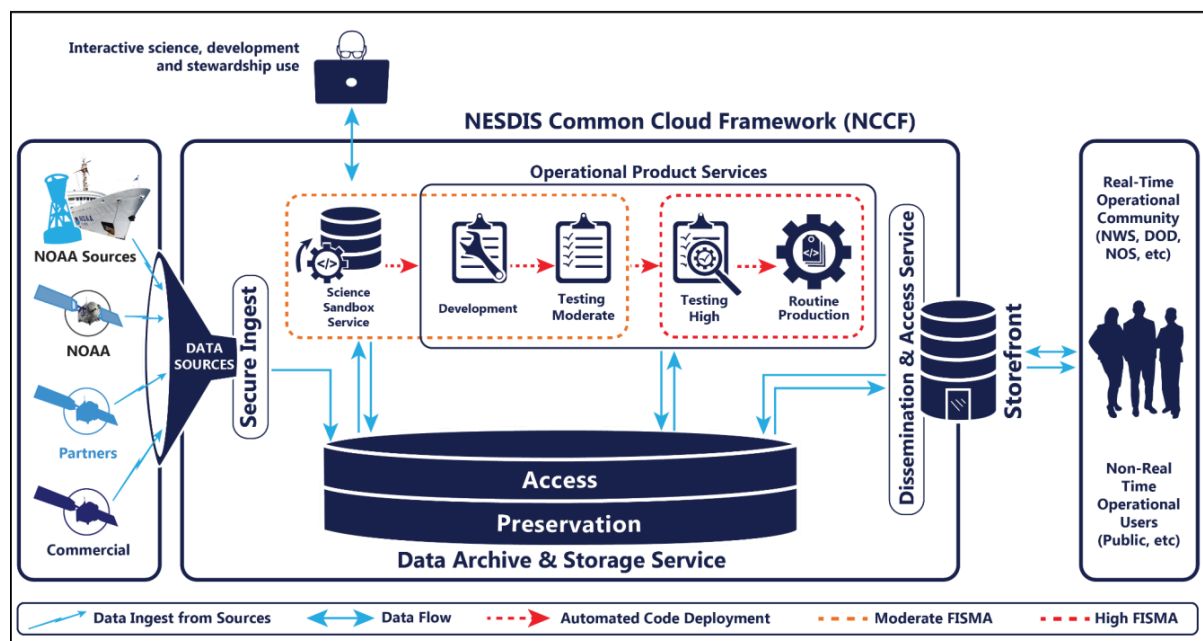


圖48、NESDIS Common Cloud Framework 功能架構圖



### 三、 心得與建議

針對氣象單位對人工智慧、機器學習等資料科學方法的運用、人工智慧就緒(氣象)資料整備、氣象資料庫等主題，總結學習心得與建議如下：

#### (一)、 關於對人工智慧、機器學習等資料科學方法的運用

ECMWF 的 Chatbot (<https://chat.ecmwf.int/>)，服務的範圍與單位目前規劃中的國家氣象資料庫 AI 諮詢櫃檯極為相似，將成為很重要的參考。該服務後端不只有我們原先設想到的運用 LLM 解析提問，再搭配 RAG 來限制回答範圍；為了確保答覆品質，ECMWF Chatbot 在將民眾提問轉到 OpenAI 時，進一步運用 prompt engineering 進行提問加工處理、設定對於答覆品質的要求與方式、依據得回的結果決定是否呼叫相關工具或 API、並確認呼叫時應輸入的參數等，把簡單提問發展成完整的一系列與 LLM 的對話，雖然推升了 OpenAI 使用費用，但較得以確保答覆精準。個人在介面上嘗試了以中文提出問題，得到頗為完整的結果，惟等待時間以分鐘計，回應有些緩慢，但這為單位後續氣象資料 AI 諮詢服務的發展提供了非常明確的參考方向。(圖49、圖50)

目前找到 ECMWF Chat 服務的測試程式原始碼已被公開在 github 上 (<https://github.com/ECMWFCode4Earth/ChatECMWF>)，將可以用作參考，而目前 ECMWF 針對後續發展的下一階段2年期280,000 EUR 的工作 Destination Earth's weather and climate chatbot 正在招商 (<https://ted.europa.eu/en/notice/-/detail/613372-2024>)，預計2025年2月會完成簽約，可以密切觀察後續發展。

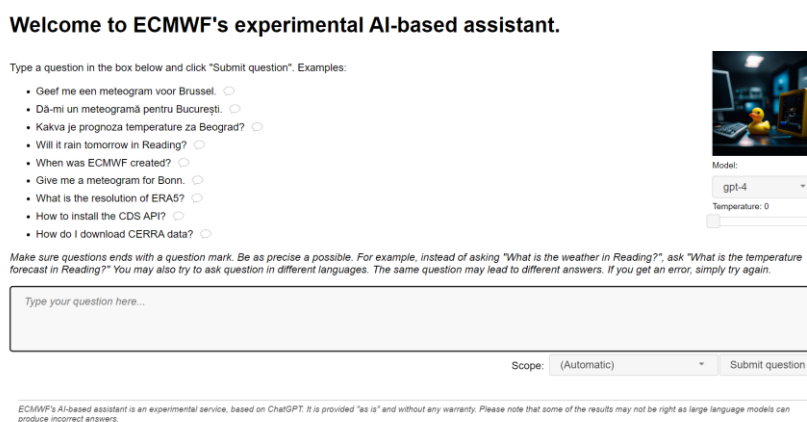


圖49、Chat@ECMWF 網頁介面

請告訴我ECMWF的開放資料政策。

Scope: (Automatic) ▾

Submit question

根據ECMWF的文件，以下是ECMWF的開放資料政策的相關資訊：

- **使用許可**：ECMWF的開放資料使用受到[Creative Commons CC-4.0-BY許可](#)和ECMWF [使用條款](#)的約束。這意味著資料可以被重新分發和商業使用，但需進行[適當的歸屬](#)。
- **資料可用性**：從2022年1月25日起，ECMWF提供了一系列的預報資料，這些資料基於高解析度預報（HRES – 9公里水平解析度）和集合預報（ENS – 18公里水平解析度），並以0.4 x 0.4度的解析度提供。
- **資料存取**：資料可以通過ECMWF的https服務和Microsoft Azure雲端獲取，未來可能也會通過其他雲端供應商提供。
- **技術支持**：ECMWF提供了一系列Jupyter notebooks，幫助熟悉Python和Jupyter的用戶探索開放資料集，並使用開源軟體庫重現開放預報圖表的繪圖。
- **用戶反饋**：ECMWF期待用戶的[反饋](#)，以幫助開放資料的提供不斷演進。

assistant

documentation

system

You are a very enthusiastic support staff who loves to help people!  
Given the following sections from the ECMWF documentation,  
answer the question using only that information,  
outputted in markdown format.

If you are unsure and the answer is not explicitly written in the  
documentation, Say that you cannot answer the question based on the information  
that is available to you.

CONTEXT:

Document 1: ECMWF moves further towards open data | ECMWF

Source: <https://www.ecmwf.int/en/about/media-centre/news/2022/ecmwf-moves-further-towards-open-data>

圖50、Chat@ECMWF 答問開放資料政策及後端處理內容截圖

NWS translation 工具在臺灣可以用於幫助在臺的外籍人士，氣象署現已部署有 AWIPS 工具，可以考慮進一步引進期新增的 AI 自動語言翻譯工具及訓練方法，未來在發布警特報等訊息時，能自動進行如英語、越南語、印尼語等翻譯，透由與現行中文訊息一致的方式進行發布、傳播，提供不熟悉中文的民眾運用，幫助相關人士的防災準備或應變。

再一點是 ECMWF Chat 的網頁上有如下宣告：“ECMWF's AI-based assistant is an experimental service, based on ChatGPT. It is provided "as is" and without any warranty. Please note that some of the results may not be right as large language models can produce incorrect answers.” NWS Translation 網頁上也宣告：“Experimental - Not for operational use.” (英文版)、“实验性的，不用于业务用途。” (該網頁僅提供簡體中文版) 對推出使用生成式 AI 語言模型這類實驗性的服務，類似這樣的宣告會減輕一些開發者的壓力，但畢竟仍是政府網域的網頁服務，如果因為語言模型的回應沒有被完善防備而出問題時，民眾的反應可能又是另外一回事。

結合 AI 等新興技術的數位雙生系統，可以被視為一個快速分析資訊、進行決策的機會。藉由將適當的自然變化和人類活動資訊整合進入地球數位雙生系統，然後藉由互動式及視覺化操作，設定空間和時間尺度，能監控和預測地球上的自然和人類活動，也快速進行各種假設情境的推演，提供作為對極端天氣和自然災害等的應對、決策參考。一個數位雙生系統應該包含有作為後端基礎的資料湖，有效率的計算工具和軟、硬體，能反映真實世界觀測和模擬的資料整合系統，能處理各種不同自然和人類活動資料，並定義前後端及系統間的資訊串接標準和協定，可互動式操作的使用者介面，能提供探索、模擬不同的假設情境，並能作結果的驗證及可靠度與不確定性的估計，方能將結果用於支援決策。

## (二)、 關於人工智慧就緒(氣象)資料整備 AI Ready Data

雖然 NOAA 已經與 ESIP 的許多資料夥伴合作建立了 AI Ready Data Checklist，但其實只是一份很長的書面清單，還沒有檢核工具，也沒有每一個項目的細部或分級標準。即使已經產出了一組 TC PRIMED 資料，但依其呈現出來的樣態，是收集了所有熱帶氣旋相關的資料，集中起來，依據便於 AI 使用的需求進行了格式等處理，然後在雲端新開一個資料空間，把所有資料提供在該處，但原先的資料源所提供的資料仍是維持原樣。又長出一組獨立資料的做法，似乎不是對國家氣象資料庫的期待方向。

想在本署國家氣象資料庫供應的 AI Ready Data，一方面希望能讓各方人員能獲得

清理過、格式與解析度皆合適、有完整說明，並且透過程式就能擷取所需資料子集，並快速引入到 AI 環境、模型上，不用再各自耗費大量時間進行資料前處理，可以加速研究進行。另一方面則是希望避免不同團隊重複處理、存放同一資料的不同子集或不同格式、解析度的版本，這是很浪費有限資源的，就如目前的署內的颱風資料庫就是從各種資料集挑選，獨立另存了一份颱風期間的觀測、預報圖資，也另做一套專有的資料庫及管理機制。

個人設想中，主題型資料集如颱風，若是改成以 **metadata** 儲存各颱風存在的日期區間、颱風位置、各種狀態資訊、以及各資料集的查詢位址連結或指標，需要各項資料時，則以存放之位址、指標，依據定位等資訊，完整下載或擷取所需範圍資料使用。如此做法，則不同主題資料可以包含相同一個資料集的全部或其中的不同子集，但使用存下來的同一組資料，節省儲存資源；未來各項應用主題的增加，也可以用新增相關 **metadata** 達成；而透過同一資料集在各種應用中的不同 **metadata**，也可以強化資料間的關聯，結合、比較不同應用，也許發現其他可能。看看中、長程的 AI Ready Data 路線規劃 (圖51)，應該有機會實現吧。

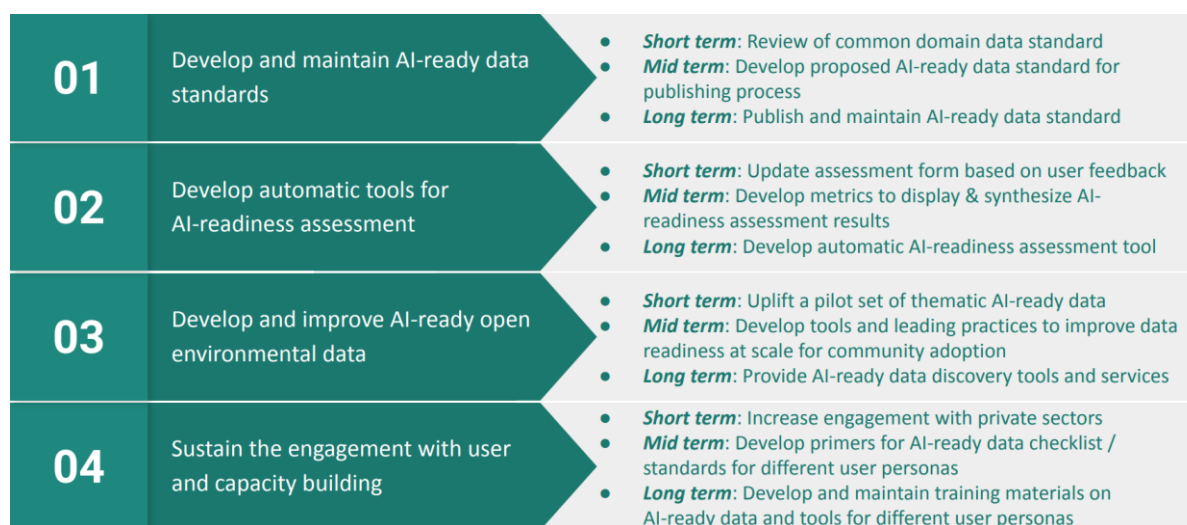


圖51、AI-Ready Data Community Roadmap

關於專案進入作業的 **Readiness Levels**，其實不是只針對 AI 相關工作，此分級可以用在各項資訊業務的開發上，從構想被提出、雛型設計、到測試驗證、移轉上線，可以明確的依此分級來判斷目前的進度，也知道後續尚有多少工作待進行，應能幫助進行資訊系統開發的工作規畫，訂出各階段的明確目標。

### (三)、 關於氣象資料庫

NODD 和 ECMWF opendata 服務皆強調其在設計時考慮了可查找性、可取得性、交互運用性和可重用性的 FAIR 原則，使用者可以通過各雲服務業者的搜尋介面找到所需資料，並可透由人類可讀和機器可讀格式的詮釋資料進行探索；然後資料可以通過直接下載，或雲服務業者的 API 取得；提供的資料及詮釋資料符合通用格式，也提供識別碼來增強資料的交互運用性與重用性。

使用雲端服務的高可用性、回應快速、彈性資源擴充、不占單位網路頻寬等優點很明顯，但費用將是問題，據說 NODD 目前是付費使用雲端服務，但合約費用並未查到，ECMWF 在去年是使用與本署同樣的 AWS Open Data Sponsorship Program，後續亦是未知數。氣象開放資料服務的成本，不論是在地端或是雲端運作，都將隨著資料下載量持續升高，營運資源的取得將是一大問題。

同時，隨著氣象資料的單一檔案越來越大，提供完整檔案的下載形式將無法長久，能夠協助僅提取所需參數、時空範圍的資料擷取 API 會是解方之一，本署除了已在發展之配合模式資料的 GRIB2API 之外，也可評估將雷達、衛星資料轉為已有資料擷取工具包的 GeoTIFF 格式，未來大量開放時可以有效節省耗用的網路頻寬。另外也聽說 NODD 在研究未來使用串流格式來供應資料，可以持續了解相關進展。

整體而言，氣象資料使用者需求一個簡單明瞭、可通過 Web 存取的資料平臺，該平臺應以標準化格式，及盡可能高的解析度提供全球範圍所有可用資料，並經常更新。該平臺應提供使用者自定義選項、資料層場、解析度等選擇，允許使用者動態分析資料，能同時應用多種資料類型，由平臺協助資料混合 (Data blending)、內插或調整解析度，並僅擷取所需的範圍，將資料引入分析，而不是要求使用者自行查找、下載、重新格式化、裁切資料子集再合併不同類型資料來執行用戶端分析。

完整、全面的天氣和氣候資料應開放供免費取得，藉由將資料集中起來，建立一站式商店 (one-stop shop)，消除需要訪問多資料源的混亂，並以規格化組織、呈現方式，降低資料集查找困難。平臺應提供一致性資料取用、下載機制，並允許彈性選擇資料項目、變數、時空範圍、解析度等。資料品質應進行控制，確保即時、穩定更新並作延遲問題管理，並在確保機密性、完整性與可用性的條件下，降低取得資料的安全設定等障礙。



平臺中資料應依通用範本提供有關資料內容說明、資料集結構 (包括資料樣本)、詮釋資料、資料源 (出處)、取得方法、應用程式介面 (API) 選項、合理的應用方式和推薦用途、使用範例、建議的替代產品，及針對品質進行分析並明確說明不確定性及使用警告。

像 NCCF 的規劃精神，將各種核心服務分類並進行功能型整合的想法，依設想首先可以用在整合目前單位內的數個資料供應服務相關系統上，拆解系統的功能，如資料接收、訊息遞送、檔案傳送、格式轉換、檔案儲存、到位監控...，把不同的功能模組化，並解降低彼此間的耦合關係，讓個別功能可獨立運作、擴展、更新，但以一致的 CI/CD 方式發展、佈署、管理。在之後則可以再逐步擴大整合的範圍。

至於像 NCCF 把全數作業往雲端移動，查到 NESDIS 5年花費加預算1億5千萬美金 (圖52)，還不一定能在2025年完成，可能要觀察 NESDIS 的後續發展與經驗分享再議。

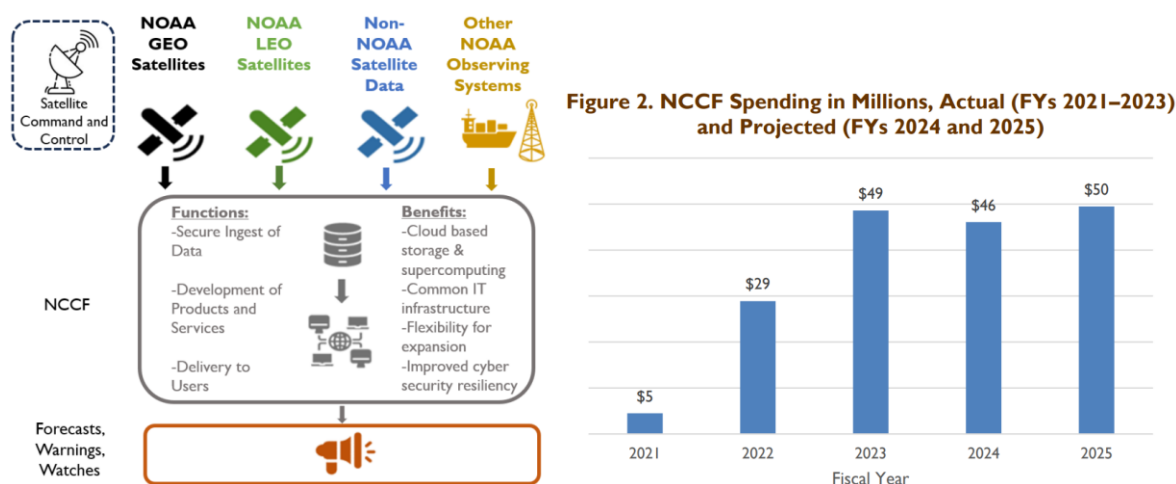


圖52、NESDIS 地面系統未來架構、NCCF 花費與預算

## 四、 參考文獻

對人工智慧、機器學習等資料科學方法的運用：

- Sun, Z., ten Brink, T., Carande, W. et al. Towards practical artificial intelligence in Earth sciences. Comput Geosci (2024). <https://doi.org/10.1007/s10596-024-10317-7>
- Rao, Yuhan, et al. Developing digital twins for Earth systems: purpose, requisites, and benefits. arXiv preprint arXiv:2306.11175 (2023).  
<https://doi.org/10.48550/arXiv.2306.11175>
- Ebert-Uphoff, I., Stewart, J. Q., and Radford, J. T., A Research Agenda for the Evaluation of AI-Based Weather Forecasting Models, EGU General Assembly Conference Abstracts, 13853, (2024). <https://doi.org/10.5194/egusphere-egu24-13853>

人工智慧就緒(氣象)資料整備：

- ESIP Data Readiness Cluster: Checklist to Examine AI-readiness for Open Environmental Datasets v.1.0. ESIP. Online resource. (2023)  
<https://doi.org/10.6084/m9.figshare.19983722.v1>
- Razin, M. N., C. J. Slocum, J. A. Knaff, P. J. Brown, and M. M. Bell, Tropical Cyclone Precipitation, Infrared, Microwave, and Environmental Dataset (TC PRIMED). Bull. Amer. Meteor. Soc., 104, E1980–E1998 (2023).  
<https://doi.org/10.1175/BAMS-D-21-0052.1>

氣象資料庫：

- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).  
<https://doi.org/10.1038/sdata.2016.18>
- Peng, G. Workbook for dissecting the FAIR Guiding Principles - Key Categories, Category-Specific Core Concepts, Focus Elements, and Harmonized Indicators. Version: v01r00-20230225 (2023). <https://doi.org/10.5281/zenodo.8057317>
- Denis S. Willett et al., NOAA Open Data Dissemination: Petabyte-scale Earth system data in the cloud. Sci. Adv.9, eadh0032(2023). <https://doi.org/10.1126/sciadv.adh0032>
- Peng, G. et al. Practical application of a data stewardship maturity matrix for the NOAA OneStop Project, vol. 18, no. 1, 2019, <https://doi.org/10.5334/dsj-2019-041>

- Brewer, M. J., A. Hollingshead, J. Dissen, N. Jones, and L. F. Webster: User Needs for Weather and Climate Information: 2019 NCEI Users' Conference. Bull. Amer. Meteor. Soc., 101, E645–E649 (2020). <https://doi.org/10.1175/BAMS-D-19-0323.1>

備註: 以上參考文獻僅列出有 doi 的相關文件，未列入各式單位工作會議、教育訓練、工作坊 (workshop) 簡報素材或演講內容，另有部分內容取材自相關網站，連結已附於內文中。

## 五、 附錄

### NOAA 第6屆人工智慧工作坊主題與議程資訊

#### **September 16-20, 2024: 6th NOAA AI Workshop on Leveraging Artificial Intelligence in Environmental Sciences**

資訊擷取自網頁：<https://www.noaa.gov/ai/events/6th-noaa-ai-workshop-2024>

This year’s workshop focuses on fostering collaborative development within the community around the theme of heat resilience.

We define “heat” broadly across different physical systems and its impacts on society and ecosystems. Examples that are relevant to the workshop include but are not limited to:

- Applying AI to advance prediction and attribution of extreme heat events across different domains of the earth systems (e.g., atmospheric and marine).
- Using AI techniques to quantify impacts of extreme heat events on human & natural systems (e.g., community and ecosystem health, drought and food security, wildfire, social equity).
- Leveraging AI tools to improve information services for heat resilience planning for the community and ecosystems (e.g., urban planning, coral reef conservation).
- Prototyping AI applications to enhance environmental observing systems for heat resilience (e.g., mission planning, observing network design, on-orbit computing).

This year’s NOAA AI Workshop will gather community members to discuss and explore collaborations on leveraging AI that align with NOAA’s strategic priorities. Participants will form community collaboration groups around selected use cases that can be sustained after the workshop via online collaboration platforms. We invite all interested scientists, AI practitioners, and application end users to join this community-driven collaboration development event.

Given available resources, the in-person component of the workshop will take place at up to three locations – College Park, MD, Boulder, CO, and Seattle, WA – simultaneously and connected via the video meeting platform. The event is open to the public with a fully

virtual participation option. Due to space limitations, priority for in-person participation will be given to participants who can commit to fully participate in the planned activities.

#### 6th NOAA Workshop on Leveraging AI in Environmental Sciences

[H] Hub Session [P] Plenary Session [T] Thematic Session [R] Tutorial & Demo

SEPTEMBER 16 • MONDAY			
8:00am – 8:30am	R	<b>Demo: Amazon Bedrock Haiku    Geospatial Chatbot</b> <i>Speakers: Rayette Toles-Abdullah, Austin Park</i>	Zoom
9:00am – 10:00am	R	<b>Tutorial: Machine Learning for Fire Radiative Power Modeling</b> <i>Speakers: Christina Kumler</i>	Zoom
10:30am – 11:30am	R	<b>Tutorial: Artificial Intelligence to Enhance Resilience: The use of ISeeChange crowdsourced data and AI to validate risk and prioritize action</b> <i>Speakers: Julia Kumari Drapkin, Erik Weber</i>	Zoom
12:00pm – 12:30pm	R	<b>Demo: Leveraging Virtual Reality for Public Outreach and Education on Washington D.C.'s Urban Heat Islands</b> <i>Speakers: Eric Hackathorn</i>	Zoom
2:00pm – 3:00pm	R	<b>Tutorial: Getting Started with AI Weather Prediction</b> <i>Speakers: Jacob Radford</i>	Zoom
SEPTEMBER 17 • TUESDAY			
7:30am – 9:00am	H	<b>MD Hub Session – Use Case Discussion</b> MD/DC Hub - Howard University Interdisciplinary Research Building (2201 Georgia Ave NW, Washington, DC 20059, USA)	
	P	<b>Plenary – Emerging Technologies for Heat Resilience</b> All Locations <i>Moderators: Eric A. Kihn, Rob Redmon, Douglas Rao</i>	
9:30am – 10:30am		<i>Speakers: Jennifer Mahoney, Monica Youngman, Michael C. Morgan, Thomas A. Oliver</i>	
	P	<b>Fireside Chat with FEMA</b> All Locations <i>Moderators: Rob Redmon, Douglas Rao</i>	
11:30am – 12:00pm		<i>Speakers: Michael C. Morgan, Derrick Hiebert</i>	
12:00pm – 1:30pm	T	<b>Modeling &amp; Observation Capability for Heat Resilience</b> All Locations <i>Moderators: Douglas Rao</i>	
		<i>Speakers: Ram Ramaswamy, William Collins, Jun Wang, Kim Martini, Jonathan Starfeldt, Deirdre Byrne</i>	
2:00pm – 3:30pm	H	<b>CO Hub Session - Use Case Discussion</b> CO Hub - Rembrandt Yard (1301 Spruce St, Boulder, CO 80302, USA)	
2:00pm – 3:30pm	H	<b>WA Hub Session - Use Case Discussion</b> WA Hub - eScience Institute (WRF Data Science Studio, UW Physics/Astronomy Tower, 6th Floor, Campus Box 351570, 3910 15th Ave NE, Seattle, WA 98195, USA)	
SEPTEMBER 18 • WEDNESDAY			
7:30am – 9:00am	H	<b>MD Hub Session - Use Case Discussion</b> MD/DC Hub - Howard University Interdisciplinary Research Building (2201 Georgia Ave NW, Washington, DC 20059, USA)	
	P	<b>Plenary – Equitable Climate Services for Heat Resilience</b> All Locations <i>Moderators: Sue Ellen Haupt</i>	
9:30am – 10:45am		<i>Speakers: Jaineey K. Bavishi, Benjamin Zaitchik, Zachary Iscol</i>	
12:00pm – 1:30pm	T	<b>Tools &amp; Information Services for Heat Resilience</b> All Locations <i>Speakers: Amy Quarkume, Monica Bozeman, Dev Niyogi, David Farham, Eric Hackathorn</i>	
2:00pm – 3:30pm	H	<b>CO Hub Session - Use Case Discussion</b> CO Hub - Rembrandt Yard (1301 Spruce St, Boulder, CO 80302, USA)	
2:00pm – 3:30pm	H	<b>WA Hub Session - Use Case Discussion</b> WA Hub - eScience Institute (WRF Data Science Studio, UW Physics/Astronomy Tower, 6th Floor, Campus Box 351570, 3910 15th Ave NE, Seattle, WA 98195, USA)	
SEPTEMBER 19 • THURSDAY			
7:30am – 9:00am	H	<b>MD Hub Session - Use Case Discussion</b> MD/DC Hub - Howard University Interdisciplinary Research Building (2201 Georgia Ave NW, Washington, DC 20059, USA)	
	P	<b>Plenary – Building AI-Ready Workforce for Heat Resilience</b> All Locations <i>Moderators: Terri Adams-Fuller</i>	
9:30am – 10:45am		<i>Speakers: Zachary Labe, Mariana Clare, Maria Molina, Maike Sonnewald, Steven Thur</i>	
12:00pm – 1:30pm	T	<b>Co-production of Knowledge and Technologies for Heat Resilience</b> All Locations <i>Moderators: Erin Moreland, Stacie Koslovsky</i>	
		<i>Speakers: Nate Harold, Patrick Johnson, Amir Behzadan, Elisabeth Cohen, Lorena Reyes, Oriana Chegwidan, Sara Cassidy</i>	
2:00pm – 3:30pm	H	<b>CO Hub Session - Use Case Discussion</b> CO Hub - Rembrandt Yard (1301 Spruce St, Boulder, CO 80302, USA)	
2:00pm – 3:30pm	H	<b>WA Hub Session - Use Case Discussion</b> WA Hub - eScience Institute (WRF Data Science Studio, UW Physics/Astronomy Tower, 6th Floor, Campus Box 351570, 3910 15th Ave NE, Seattle, WA 98195, USA)	
SEPTEMBER 20 • FRIDAY			
8:00am – 8:45am	R	<b>Demo: Climate Engine</b> Zoom <i>Speakers: Steve Ansari</i>	
11:00am – 12:00pm	R	<b>Tutorial: Scikit-Explain: A Python Toolkit for Explaining Traditional Machine Learning Models</b> Zoom <i>Speakers: Montgomery Flora</i>	
12:30pm – 1:00pm	R	<b>Demo: IBM Golden Retriever &amp; GenAI</b> Zoom <i>Speakers: Mohamed Ahmed</i>	

圖53、議程資訊