

CONCERT 2019 年 學科館員沙盒計畫工作坊

開展更精準的圖書館資料服務 |

研究資料側寫分析： 理解研究者們的資料實踐

Profiling tools for understanding researchers' data
curation and management practices

台灣大學圖書資訊學系 助理教授 | 鄭瑋



研究資料側寫分析：理解研究者們的資料實踐

Part 1. | 淺談學術圖書館研究資料服務

台灣大學圖書資訊學系 助理教授 | 鄭瑋

研究資料為
學界廣泛接受的記錄事實性材料，
用來驗證研究結果。

Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.

一般來說，下列的「材料」不被視為研究資料 (US Federal gov defined)：

Preliminary analyses, [drafts of scientific papers](#), plans for future research, [peer reviews](#), communications with colleagues, [physical objects \(e.g., laboratory samples, plants, animals\)](#), personnel and medical information and similar information the disclosure of which would constitute a clearly [unwarranted invasion of personal privacy](#)

研究資料的不同型態-I

→ 觀測型資料 (Observational):

利用儀器或是觀察工具即時捕捉，獨一無二、無法複製或重新獲得的資料（非常珍貴）。

例| 太空遙測、街口人類觀察

→ 實驗型資料 (Experimental):

來自實驗室設備或受控條件下的數據；通常可重複（但這樣做可能很昂貴）例子包括基因序列，色譜圖，磁場讀數和光譜學。

研究資料的不同型態-II

→ 模擬型 (Simulation):

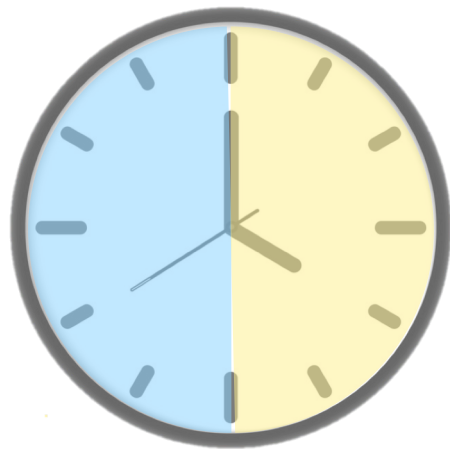
從模型生成的測試數據與 metadata, 其中模型本身比output 的數據更重要。

例 | 氣象預測資料、經濟模型等

→ 文獻與紀錄 (Literature and records) :

例 | 文獻、法庭資料、演講稿、作家手稿

資料生命週期 (以時鐘來比擬)



Stewardship

| 對資料的盡職治理能力

資料管治 (governance)、策管 (curation)、組織、資料保存、分享、再用…

Acumen

| 對情報快速且正確判斷的決策能力

數據探索(discover)、採集、清理、分析、運算思維、視覺化、…



學術圖書館員在 data stewardship 扮演舉足輕重的角色

- ◇ 支援廣大的 small science 校內讀者需求
- ◇ 因應「開放科學運動」的趨勢

名詞
解釋

團隊規模與資料量

Big science, big fund, big data
天文、基因、石油、核能…

Small science, small data
環境工程學、生態學、行為
科學、社會科學…

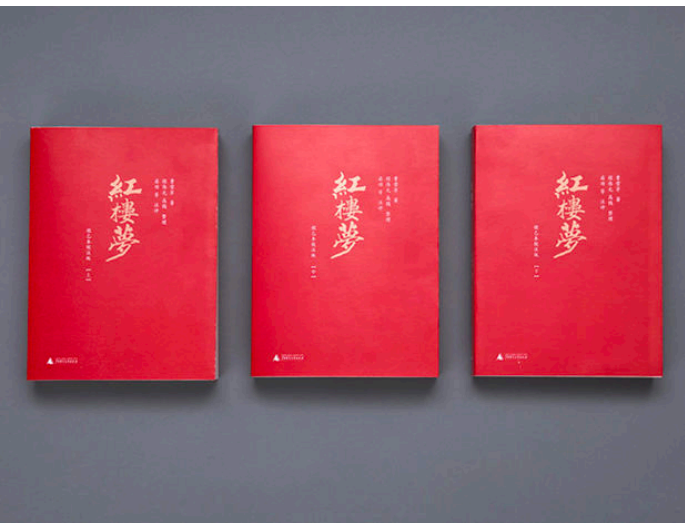
The long tail

學術圖書館員在 data stewardship 扮演舉足輕重的角色

- ◇ 支援廣大的 small science 校內讀者需求
- ◇ 因應「開放科學運動」的趨勢
- ◇ 因為圖書資訊學科特性，擁有跨領域的經驗
- ◇ 因應「政府資助獎勵計劃之成果開放相關法規」(open data mandates)
我國：2019 選案試推、2020 實施
- ◇ 學術圖書館出現新型態的分工：
資料館員 (data librarian)、GIS 館員、資料視覺化館員、資料策管人員
(data curator)、data champion (類似學科館員，專門 focus 在資料服務)

因為圖書資訊學科特性，擁有跨領域的經驗：

同樣的資料（文本），在不同領域有著不同實踐與用法



Dream of the Red Chamber (Text)

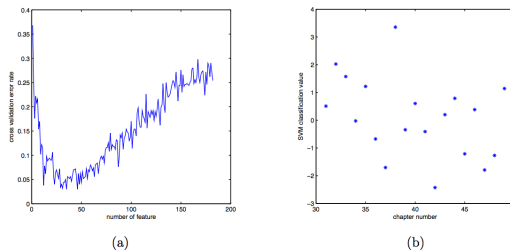


Figure 2. Experiment 2: (a) Mean cross validation error rate; (b) Values of SVM classifier on chapters 31-50. Note there is no chrono-divide.

紅樓夢之作者偵測（文字探勘）

Hu, X., Wang, Y., & Wu, Q. (2014). Multiple authors detection: a quantitative analysis of Dream of the Red Chamber. *Advances in Adaptive Data Analysis*, 6(04), 1450012.

紅樓夢中的庭審與冤獄（法律研究）

Xiaohuan Zhao (2011) Court Trials and Miscarriage of Justice in Dream of the Red Chamber, *Law & Literature*, 23:1, 129-156

例子：多倫多大學圖書館 資料視覺化館員--- 求才公告

- 協助與輔導校內師生資料視覺化的相關工具的使用
- 在多倫多大學圖書館建立 data visualization services

<https://iassistdata.org/resources/jobs/7233>

Data Visualization Librarian

Submitted by STulley on Mon, 2016-08-29 05:54

Posted to IASSIST on: 2016-08-26

Employer: University of Toronto

Employer URL: <https://www.utoronto.ca/>

Description:

The University of Toronto Libraries (UTL) is seeking a Data Visualization Librarian to develop programs and services to support the emerging data visualization needs of students and faculty at the University of Toronto.

Reporting to the Head of the Map and Data Library, the Data Visualization Librarian will engage with users across the university and work collegially within UTL to develop a strategic approach to this new area of library services, including conducting needs assessments to inform service, technical, and infrastructural requirements.

Other responsibilities are as follows:

- Supports students and faculty with their data visualization research and practice via consultations, and through teaching and instruction.
- In collaboration with other UTL librarians, develops and offers software instruction to support teaching, learning, and research at the University of Toronto
- Leads the writing and publishing of data visualization instructional tutorials, courses, and supplementary material
- In collaboration with other UTL librarians, contributes to the broad range of data services that support academic researchers.
- Collaborates with and contributes expertise to student and faculty data, statistical, and visualization projects in a range of disciplines
- Develops and reports on assessment measures for services
- Through various levels of engagement and partnerships, integrates data visualization services with existing services at the Map and Data Library and UTL
- Actively engages in conversations and forums around data visualization on a national and international basis

Qualifications:

Required:

- ALA-accredited Master of Library Science degree (or equivalent)
- Experience using data visualization tools such as D3.js, Tableau, Mathematica, and R.
- Proven understanding of best practices for creating effective visualizations

1. 給予讀者資料產製-分享-策管-再利用各階段的指引



2. 上述事項 → 研究資料管理計畫 (Data Management Plan)

服務舉隅



- **DMP、政策與法規的指導**
- **開設資料分析與視覺化相關訓練課程**
- **開設領域資料分析工具 (如 GIS)**
- 資料中心的推薦
- 知識產權、隱私、去識別化的指引
- **Metadata與描述資料撰寫的指引**
- **機構典藏資料庫的改/創建、管理與使用者服務**
- **機構/國家政策與法規的指導**
- 指引特殊資料的讀取與利用

Digital Scholarship Services Workshops

Digital Scholarship Services has offered workshops on a range of digital tools and methods since 2015. These sessions are available as on-demand workshops for courses and interested groups of faculty and students. These topics can also form the focus of a consultation. Contact us to schedule a session or appointment.

Past sessions have included:

Authoring Multimedia Narratives

An exploration of digital storytelling tools for creating collaborative, non-linear, and multimodal scholarship; tools covered include Scalar, TimelineJS, StoryMapJS, and Omeka.net

Building a Digital Portfolio with WordPress

A session on using Wordpress to create a portfolio representing your professional self

Creating Maps on the Web

An overview of web-based tools like CARTO and Google Earth for creating maps and exploring spatial relationships in data

Creating Online Digital Collections with Omeka

An introduction to Omeka.net, a free, web-based tool for organizing, describing, telling stories with, and sharing digital collections

Data Cleaning with OpenRefine

An overview of OpenRefine, a powerful tool for cleaning messy tabular data

Documenting Reproducible Research with Jupyter Notebooks

A session on Jupyter Notebooks, a platform for interactive computational research and data science

Introduction to Data Visualization with Tableau

An introduction to Tableau Public, a free data visualization tool for analyzing and illustrating patterns in data

Managing a Digital Project

An introduction to project management approaches and tools for individuals who are planning and implementing digital projects.

Collecting Web Data with Python Workshop

The basics of the Python programming language with a focus on collecting data from web sites and social media platforms.

數位敘事創作

用WordPress 創建 Digital portfolio

在Web上創建地圖

使用Omeka創建數位館藏

使用OpenRefine清理數據

使用Jupyter筆記本記錄實驗、支持研究「再現」

資料視覺化軟體 Tableau

專案管理 (project management) 軟體

使用Python Workshop收集Web數據





Workshops

例子：美國維吉尼亞大學圖書館
資料科學 workshop- 課程一覽

We offer training on data analysis, statistics, and computation and on Library resources and methods. In Fall 2019 we'll be offering multiple series of workshops, including a series in:

- [Using R](#)
- [Using Python](#)
- [Scientific Workflow Tools](#)
- [Tableau](#)
- [Online Surveys](#)
- [Writing Management Tools](#)
- [General Data Tools.](#)

大學圖書館館員如何開始準備 research data 相關的服務？

個人的觀察與經驗：

1. 從盤點現有人力，建立服務模式 (圖資學界也在努力)
2. 配套先走、法規後有(大學圖書館的角色舉足輕重)

真的想要做得更多：

1. 不妨從了解校內讀者需求開始。

Part 1 & Part 2小結：

浮現的問題：

- 學者與資料互動的經驗相當複雜，與諸多原因（學科訓練背景、研究興趣、獎勵計劃要求、合作夥伴要求）相關，並且需要進一步研究

最佳實踐：

- 制定與資料管理相關的工作流程，提供相對應的資訊服務將有助於減輕校內讀者的負擔。
- 進行上述的服務設計前，收集有關校內讀者的研究模式和資料特徵是不錯的開始。



動機：學者與資料互動的經驗相當複雜，與諸多原因（學科訓練背景、研究興趣、獎勵計劃要求、合作夥伴要求）相關，並且需要進一步研究

目的：理解研究者們的資料實踐

本日工作坊目標：介紹研究資料側寫分析工具 (profiling tool)

可能適用的情境：

- 建立大學圖書館中與科研資料相關的服務前，需要探索校內讀者的需求與研究特性。
- 變動、調整大學圖書館中與研究支援相關的服務前，可以從學者們的資料實務 (data related practices) 入手。
- 館方制定與資料活動 (如雲端儲存、隱私權、版權) 相關規章前，可以了解校內讀者的需求與研究特性。

研究資料側寫分析：理解研究者們的資料實踐

Part 2. | 研究資料側寫工具

台灣大學圖書資訊學系 助理教授 | 鄭瑋

資料側寫



圖書館或資訊
服務提供者



透過蒐集來的情報，去分析校內/國內讀者的需求、現況、挑戰，最後以側寫結果去發展新服務、爭取預算、改善現有服務。



讀者/ users

研究資料側寫工具-DCP & CCMF

側寫工具：一種調查方法，用以協助人員描繪未知研究標的。
目標是提供圖書館（或服務提供者）發展資料服務之策略與先後順序。

在研究資料的領域，被我分為：



面對單一學者之側寫（目標：建立該領域的 individual exemplar）
如 Data curation profiles (Purdue team)

→ 面對單一領域之側寫（目標：評估該學科之研究模式與 readiness）如
Community Capability Model Framework (UKOLN)

資料策管側寫工具 (Data Curation Profile)

- 常見縮寫為：DCP
- 由美國普渡大學與伊利諾大學香檳分校之資訊學院開發
- DCP 是一套完整的工具與資源（稱之為DCP Toolkit），館員或是研究者可利用此工具去側寫想要研究的對象（通常是個別的學者或是一個research group)
- DCP 的作者通常是由館員(訪談者)，但最終呈現的成果是研究對象的profile (DCP directory: <https://docs.lib.purdue.edu/dcp/>)

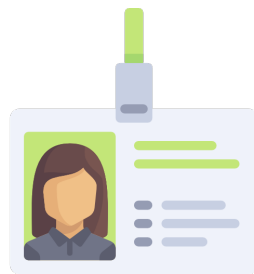
Data Curation Toolkits：給館員/訪談者使用的工具箱



訪談者操作手冊
Interviewer's Manual



訪談工作表單
Interview Worksheet



側寫檔模板
Profile Template



使用者指南
User Guide

訪談工作不需要完全 follow 工作表單，可以根據實際情況調整


Data Curation Profiles Toolkit

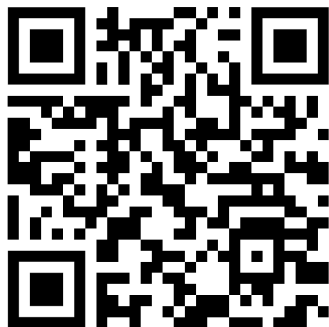
A **Data Curation Profile** is a resource for Library and Information Science professionals, Archivists, IT professionals, Data Managers, and others who want information about the specific data generated and used in research areas and sub-disciplines that may be published, shared, and preserved for re-use. Data Curation Profiles capture requirements for a specific data set generated by a single scientist or lab, based on their reported their needs and preferences for the data. More information about the Data Curation Profiles Project is available at <http://datacurationprofiles.org>

This collection contains materials related to **Data Curation Profile Toolkit**, providing a stable, DOI citable home for the Toolkit.

Follow

Data Curation Toolkits from 2010

-  PDF [The Data Curation Profiles Toolkit: User Guide](#), Jake Carlson, Version 1.0
-  PDF [The Data Curation Profiles Toolkit: Interviewer's Manual](#), Jake Carlson, Version 1.0
-  PDF [The Data Curation Profiles Toolkit: Interview Worksheet](#), Jake Carlson, Version 1.0
-  PDF [The Data Curation Profiles Toolkit: The Profile Template](#), Jake Carlson, Version 1.0



Three stages to conduct DCP

- 準備前置工作
- 進行訪談 (Interview)
- 將訪談結果寫成 “側寫檔” (Profile)

準備前置工作

1. 熟習 ToolKit、熟習 Manual、客製化符合館內需求的工作單 (Worksheet)
2. 熟習與專家晤談的技巧 (Try to interviewing a colleague first)
3. 依據不同機構的研究倫理規範，判定是否需要經過機構內倫理委員會之審查。（需要知情同意書嗎？）
4. 接洽並尋找校內學者、研究團體訪談。

進行訪談 (Interview)

- 訪談開始時，研究者手持「操作手冊」（Interviewer's manual），將Interview Worksheet 交給受訪者參考。
- 訪談進行中，研究者可利用且甚至宣讀操作手冊中的腳本，以確保訪談的信效度。



訪談工作表單 Interview Worksheet 範例問題

--	基本資料	--
模組 1	領域資料特性	Brief description of the data
模組 2	資料生命週期	Data stages (data volume, formats, ...)
模組 3	資料分享	Indicate what data you would be willing to share and with who: <ul style="list-style-type: none">• Who would you imagine would be interested in this data?• How would you imagine this data being used?• Would you place any conditions on sharing your data with the groups or people you have identified (such as requiring some form of acknowledgement, etc.)?
模組 4	資料取用	<ul style="list-style-type: none">• Have you ever deposited any of this data into a data repository?• Would you be willing to submit your data to a data repository?• Please prioritize your need for the following types of services for your data• ...
模組 5	資料典藏之選擇與利用	What preparations or actions (if any) would need to take place before your data could be ingested into a repository or otherwise transferred out of your direct control for curation purposes?
模組 6	資料描述與標準 (metadata)	Please explain briefly how this dataset is organized, and how it has been described (e.g. "detailed annotations", "a code book", "a data dictionary", "column headings in a spreadsheet", etc.).



訪談工作表單 Interview Worksheet 範例問題

模組 7	資料 (被) 探索	How do you imagine that people would find your data set?
模組 8	智財權與資料擁有權方面	<ul style="list-style-type: none">• Who is the owner of the data?• What are your funding sources for this research?
模組 9	資料產製工具	<ul style="list-style-type: none">• What tools – software or hardware – are used in generating the data?
模組 10	資料出版 (互操作性)	Please prioritize your need for the following types of services for your data: <ul style="list-style-type: none">• The ability to connect your data with publications or other outputs.• The ability to connect or merge your data with other data sets.
模組 11	資料之影響力	Beyond usage statistics, are there any other measurements or analytics that you would like to apply to your data?
模組 12	平時資料管理	<ul style="list-style-type: none">• What are the primary ways that you currently manage your data?• Please include the storage media(s) and any tools used in your management of the data.
模組 13	資料典藏需求	<ul style="list-style-type: none">• What are the most important parts of the data to preserve (manage and maintain over time)?• How long would your data set be useful or have value for you or others if it were to be preserved?

Try it yourself

→ 2 人一組，扮演：

1. 受訪者（學者）
2. 訪談者（館員）

請分配到受訪者的學員，回憶最近一次處理業務上「資料」的情景回答問題。歡迎您用大學專題或是碩博士論文的經驗作答。

請您使用手上單頁講義：

受訪學者請使用「worksheet」

訪談者請使用「interviewer's manual」

進行訪談 (Interview)

- 訪談開始時，研究者手持「操作手冊」 (Interviewer' s manual) ，將Interview Worksheet 交給受訪者參考。
- 訪談進行中，研究者可利用且甚至宣讀操作手冊中的腳本，以確保訪談的信效度。
- 盡量不要更動模組的順序，讓受訪者盡量回答完一個模組中的問題，再開始問下一組。
- DCP 推薦將13個模組拆成兩次進行

和同一位受訪者進行兩次 Interview

- 理由：避免受訪者疲勞；有些問題可以讓學者回去想一想，「發酵」後再回來，說不定能獲得更全面的回答。
- DCP 建議第一次的 interview 可以進行 模組 1, 2, 3, 4，然後第二次將剩餘模組訪完。

將訪談結果寫成 “側寫檔” (Profile)

- 將訪談內容轉換為逐字稿
- 將逐字稿排入側寫檔 template
- 若要將側寫檔分享給館外同仁或是出版至館訊、會議、期刊，要注意機構學術倫理相關的規範

案例舉隅：<https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1034&context=dc>

側寫檔模板

學科：生物化學/組織蛋白修飾



Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Methodology Development Part 1: Discovery	Microscopy images and descriptions of these images. Documentation of experiments and evolving methodologies Records of the components involved in the protocol (pdfs) Tabular information and photos of DNA gels about the fly stocks used in the experiments	Many small files are generated at this stage. They include: QPCR files: 5-6 mb each Excel files: 38 kb or so each TIFF files: 2-3 mb each	ND2 (proprietary) TIFFs MS Word PDFs QPCR files (proprietary) MS Excel MS Access MS OneNote	MS Word files are for documenting the development of the protocols and are versioned. Data in proprietary formats are kept as reference as they contain information that is lost in the transfer to open formats.
Methodology Development Part 2: Refinement	An optimized research method Annotated files	Similar to the previous stage: multiple small files.	In addition to the file formats listed in the previous stage, FCS files are generated (proprietary format)	Some of the flow cytometrics techniques are performed by a bioinformatics facility within the institution.
Data Collection	"Raw" sequence data Scripts for sequencing the data	Not discussed in detail, but referred to by the researcher as "very, very big"	Unknown but proprietary in nature	These data files are large enough so that they must be stored offsite. The sequencing scripts would be needed to

剛剛的模組2

對果蠅幼蟲的神經細胞做研究，細胞核失去組蛋白修飾複合物時，基因表達的變化

- **Data type**：記錄本（實驗、分析、影像版本、投稿版本等手寫記錄）、pdf、製圖
- **Sharing behavior**：在發表過程中，會同時將data發佈至database或其他repository。他人使用data需cite paper而非data本身。將data存進repository比較像是領域practice，無論有無資金。Raw data太龐雜對他人來說沒有意義，不值得分享，至少需經過初步清理。在領域中應是鼓勵分享，審查data也是出版程序之一，出版後也要存放至repository。
- **Needs**：高度仰賴實驗室筆記。目前沒有DMP，但未來若要提供，需要有人協助（因為目前整理data方式只供內需，而非對外給其他funding source）。Repository在對繳交data的格式等有嚴格的要求，通常需要外部資源協助。
- **Perspectives**：希望是paper能被cite才有實質利益，data被cite沒有（該領域近期應該也不會有改變）。雖然願意分享methodology，但又同時認為是最精華的部分，等研究發表後再分享方法。
- **Intended audiences**：同興趣的研究者，對early stage（發想方法，應用在其他生物體或是改變介質？）和later stage data（研究成果）
- **Data stage publicity**：研方、初步資料分析、研究成果
- **困難**：研究生的訓練（training, notebook protocol, documenting, practices...）

What can a Data Curation Profile be used for?

對受訪者來說:

- Provides a structure for conducting a data interview between an information professional and a researcher or research group.
- Provides a means for a researcher or a research group to thoughtfully consider their needs for their data beyond its immediate use.

對館方或學校來說:

- Can serve as a foundational document to guide the management and/or curation of a particular data set.
- Can be shared with staff providing data services and others to inform them and ensure that everyone is on the same page.

研究資料側寫分析：理解研究者們的資料實踐

Part 3. | 社群側寫工具-CCMF 研究案例分享

Case Study: Investigating community capability toward data-intensive research in social sciences

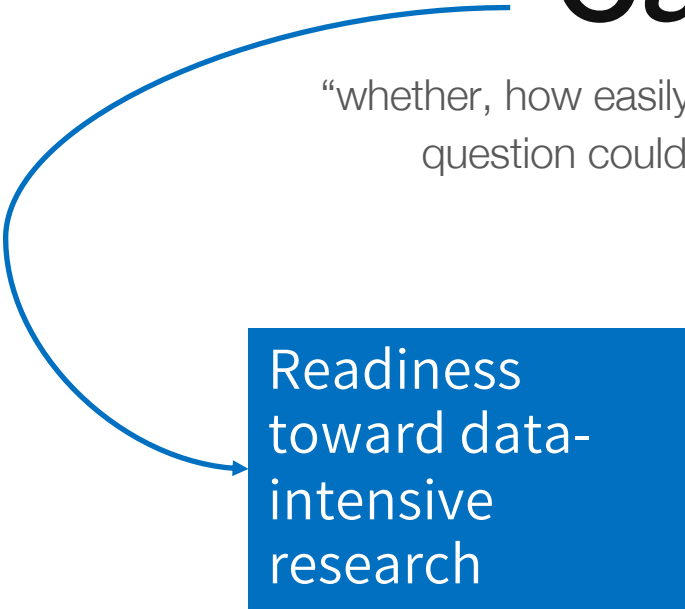
台灣大學圖書資訊學系 助理教授 | 鄭瑋

以社會科學為例

社群 Community Capability 能力

“whether, how easily, and how well the agent in question could accomplish a given task”

- e.g., 軟體
- e.g., 服務



Readiness
toward data-
intensive
research

(理解一個學科是否具有「資料集約」之特性，以及對於資料策管的「準備程度」的現況、包含缺口與領先處)



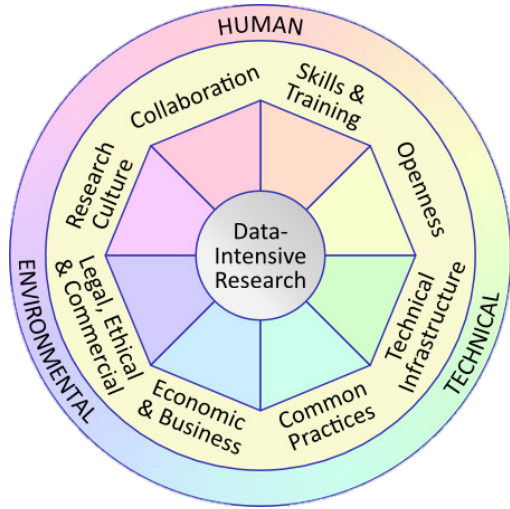
Indicate priority areas for change and investment. (排序優先發展事項與成本)



就可以做這些規劃了：

1. Develop roadmaps
2. Improve current status
3. Modify policies

How we started: CCMF (社群能力模型架構)



- Instrument: Community Capability Model Framework (CCMF) Profile tool
- Fourteen researchers have been completed the survey in October 2015.

Jeng, W. & Lyon, L. (2016). A report of data-intensive capability, institutional support, and data management practices in social sciences. *International Journal of Digital Curation (IDLC)*, 11 (1). doi: 10.2218/ijdc.v11i1.0

Eight Factors + One Data Profile in CCMF

Part 1:

Capability Factors

(55 close-ended questions)

1. Collaboration
2. Skills & Training
3. Openness
4. Technical Infrastructure
5. Common Practices
6. Economic & Business
7. Legal, Ethical & Commercial
8. Research Culture

Part 2:

Data Profile

(10 open-ended questions)

“How complex is your data?
Does it contain multiple
variables or attributes?”

Five levels

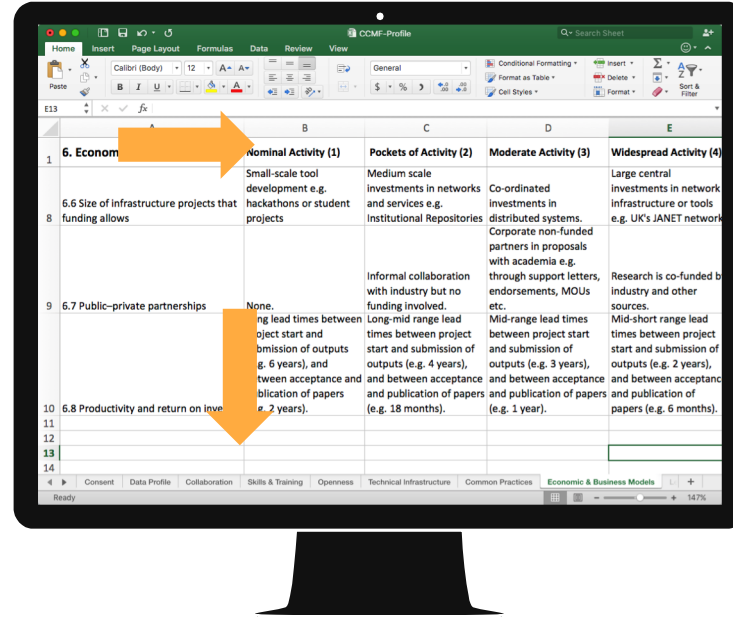
- 1: Nominal activity
- 2: Pockets of activity
- 3: Moderate activity,
- 4: Widespread activity
- 5: Complete engagement.

Try it yourself 工作表下載：
http://bit.ly/CONCERT_CCMF



白皮書下載：

<https://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf>



	B	C	D	E	
1	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	
8	6.6 Size of infrastructure projects that funding allows	Small-scale tool development e.g. hackathons or student projects	Medium scale investments in networks and services e.g. Institutional Repositories	Co-ordinated investments in distributed systems. Corporate non-funded partners in proposals with academia e.g. through support letters, endorsements, MOUs etc.	Large central investments in network infrastructure or tools e.g. UK's JANET network
9	6.7 Public-private partnerships	None.	Informal collaboration with industry but no funding involved.	Mid-range lead times between project start and submission of outputs (e.g. 3 years), and between acceptance and publication of papers (e.g. 1 year).	Mid-short range lead times between project start and submission of outputs (e.g. 2 years), and between acceptance and publication of papers (e.g. 6 months).
10	6.8 Productivity and return on investment	Long lead times between project start and submission of outputs (e.g. 6 years), and between acceptance and publication of papers (e.g. 2 years).	Long-mid range lead times between project start and submission of outputs (e.g. 4 years), and between acceptance and publication of papers (e.g. 18 months).	Mid-range lead times between project start and submission of outputs (e.g. 3 years), and between acceptance and publication of papers (e.g. 1 year).	Mid-short range lead times between project start and submission of outputs (e.g. 2 years), and between acceptance and publication of papers (e.g. 6 months).

Challenges found when conducting the profiles:



A Need for Tailoring the CCMF

- To speak the same research language
- To enhance comprehension and reduce error measurement

Previous exemplar:

- Applicability for agricultural science (Dr. Scott Brandt; Reported in Workshop 4, IDCC 2014)
- Customization in consultation with Agronomy faculty at Purdue University.

6. Economic & Business models (for data)	Nominal Activity (1)	Focused Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)
2	For this set of elements, describe scope and/or level of funding for the majority of your research:				
3	6.1 Duration of funding for research	One-off funding focused on quick returns e.g. 1-2 years	Funding focused on short term projects and quick returns e.g. 2-3 years	Larger term investments	Single phase thematic investments on a 2-7 year timescale. Multi-phase thematic investments in 5-20 year blocks which build a cumulative e.g. NSF DataONE Programme funding by international bodies and to lateral initiatives between national funders.
4	6.2 Geographic scale of funding for research	Projects funded internally (State or university).	Projects funded through grants from regional funders, agencies, etc.	Projects funded by national funders.	Projects funded by multiple national funders.
5	6.3 Size of research that funding allows	Small investigative projects	Small-sized projects	Mid-sized projects	Major investment infrastructure projects facilitate essential transitions to self-financing models. Collaborative development between international funders e.g. International Plant Diagnostic Network (IPDN).
6	6.4 Sustainability of funding for infrastructure	One-off investments with no commitment to sustainment e.g. start-up equipment	Multi phase projects to develop infrastructure e.g. enhance or enlarge equipment	Sustained multi-decade investments	Self-financing infrastructure, networks and services. Collaborative development between international funders e.g. International Plant Diagnostic Network (IPDN).
7	6.5 Geographic scale of funding for infrastructure	Projects funded internally	Investments by a single funding body at regional level.	Investments by a single funding body at national level.	Development at the national level by multiple funders e.g. NCCOR, Next Ag Ability Project.
8	6.6 Size of infrastructure projects that funding allows	Small-scale tool development e.g. student sensor network for built applications	CCAD/NG water balance sensor network for Wabash Monitors	Co-ordinated investments in distributed systems and services e.g. CCAD/NG water balance sensor network for Indiana.	Large central investments in network infrastructure investment e.g. PIACTS or tools e.g. Aquatic Database
9	6.7 Public-private partnerships	None.	Informal collaboration with industry but no funding involved.	Corporate non-funded partners in proposals with academia e.g. support letters, endorsements, MOUs.	Research is co-funded by investment partnerships industry and other sources. Established formal co-investment partnerships running long-term multi-phase projects.

Original: Modified or changed:

3.4 Openness of methodologies/workflows

(e.g. steps for preparing an interview or a focus group, how to run different statistical models on a software program)

Adding social-science-friendly descriptions

4.2 Tool support for data capture and collection

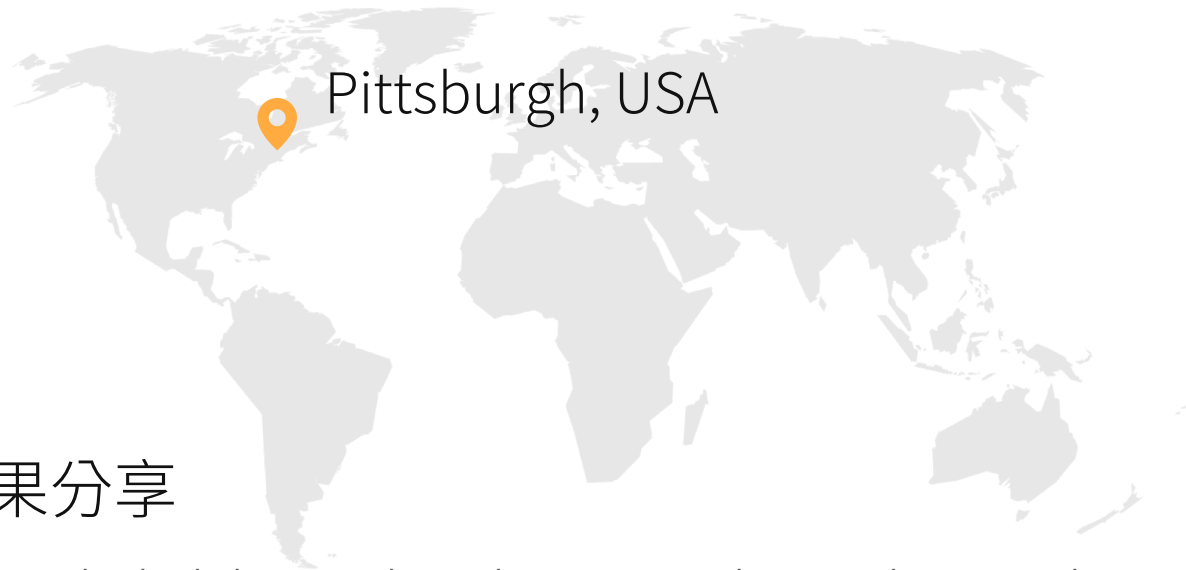
(e.g., screencasting tools, digital audio recorder, web scraping tools, survey software programmes such as Qualtrics, SurveyMonkey)

Adding discipline-tailored tools as examples

2.12 Data metrics and impact e.g., impact factors, altmetrics

(e.g., impact factors of research datasets, altmetrics of datasets such as the number of downloads)

Avoiding possible jargons, providing explanations of terminologies.



Pittsburgh, USA

CCMF 成果分享

Participants included doctoral students, postdoctoral researchers, and assistant professors from the Dept. of **Anthropology**, **Political Sciences**, and the **LIS** program at the University of Pittsburgh.

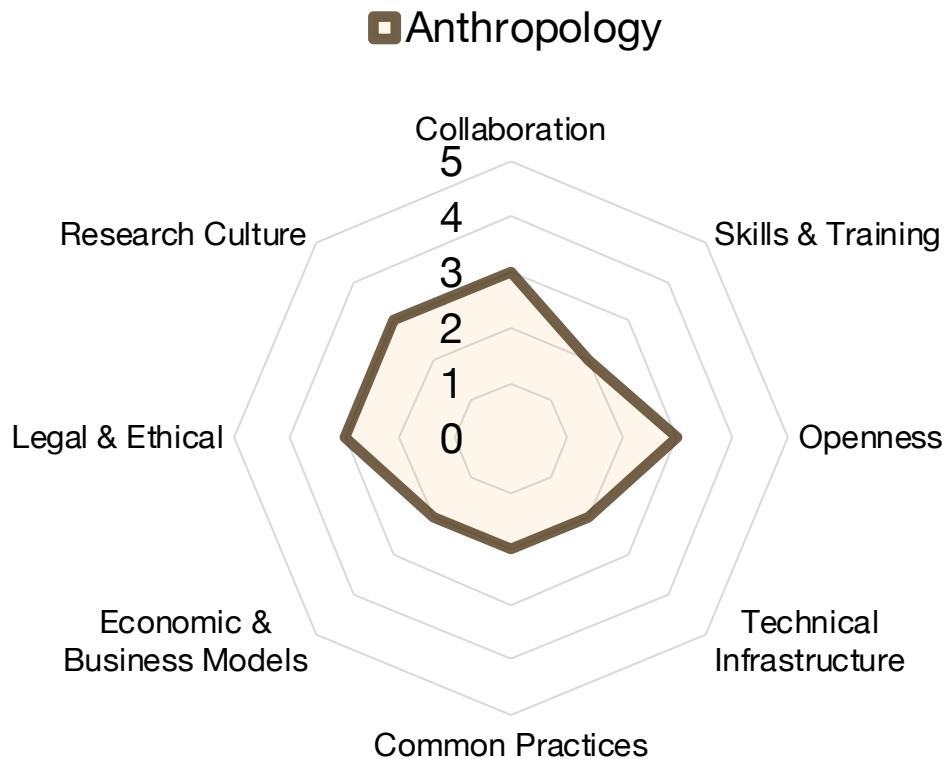
Table 3. Study participants (A = interviewed and mediated, B = self-assessed).

#	Approach	Position	Discipline	Self-Identified Research Topic
P01	A	Post-doctoral researcher	Anthropology	Cultural anthropology
P02	A	PhD student	Library and Information Science	Music metadata
P03	B	PhD student	Library and Information Science	Geospatial information systems (GIS) and accessibility
P04	B	PhD student	Library and Information Science	Information retrieval
P05	A	PhD student	Anthropology	Cultural anthropology, Legal Anthropology (child adoption)
P06	A	Assistant professor	Political Science	Comparative politics
P07	B	Post-doctoral researcher	Political Science	Area studies (South Asia)
P08	B	PhD student	Political Science	Comparative politics, political methodology
P09	B	PhD student	Anthropology	Archaeology
P10	B	Visiting scholar (assistant professor)	Library and Information Science	Public library management
P11	B	Post-doctoral researcher	Anthropology	Medical anthropology
P12	B	Assistant professor	Library and Information Science	Public library management
P13	B	Post-doctoral researcher	Library and Information Science	Information seeking behaviours
P14	B	PhD student	Political Science	International relations

Capability summary of social sciences

Anthropology scholars' ratings were relatively evenly distributed to each dimension.

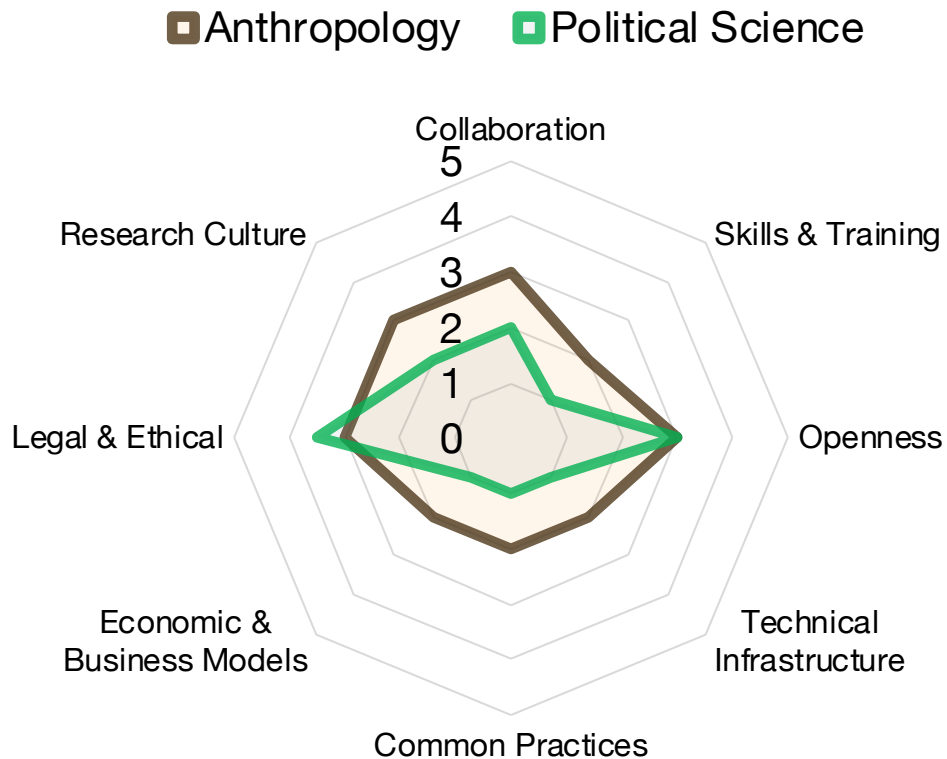
- 1: Nominal
- 2: Pockets of Activity
- 3: Moderate
- 4: Widespread
- 5: Complete Engagement.



Capability summary of social sciences

Political science scholars ranked Legal & Ethical and Openness as highest in development, whereas assigned relatively low scores to other dimensions.

- 1: Nominal
- 2: Pockets of Activity
- 3: Moderate
- 4: Widespread
- 5: Complete Engagement.



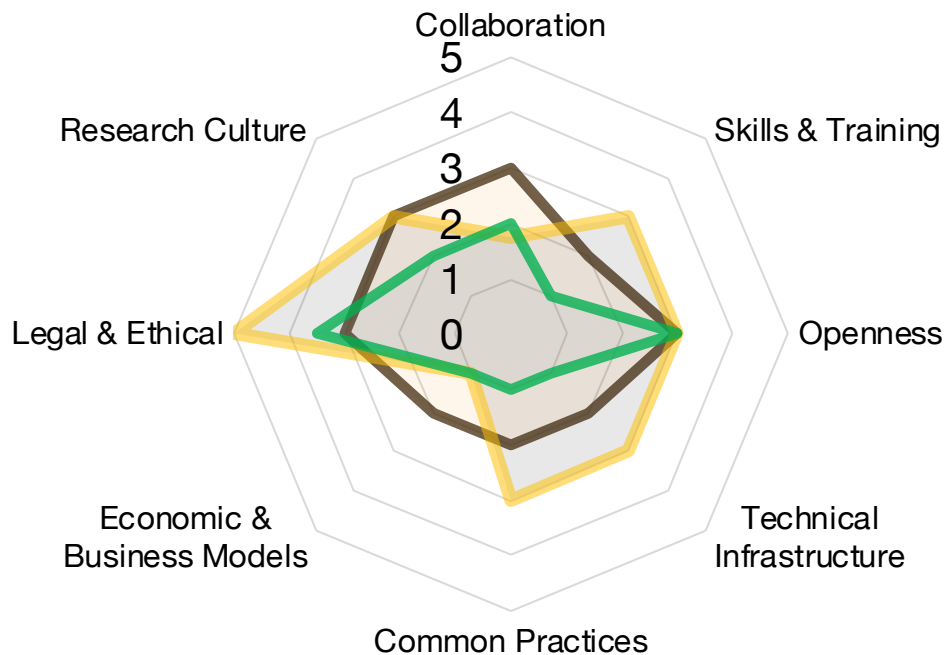
Capability summary of social sciences

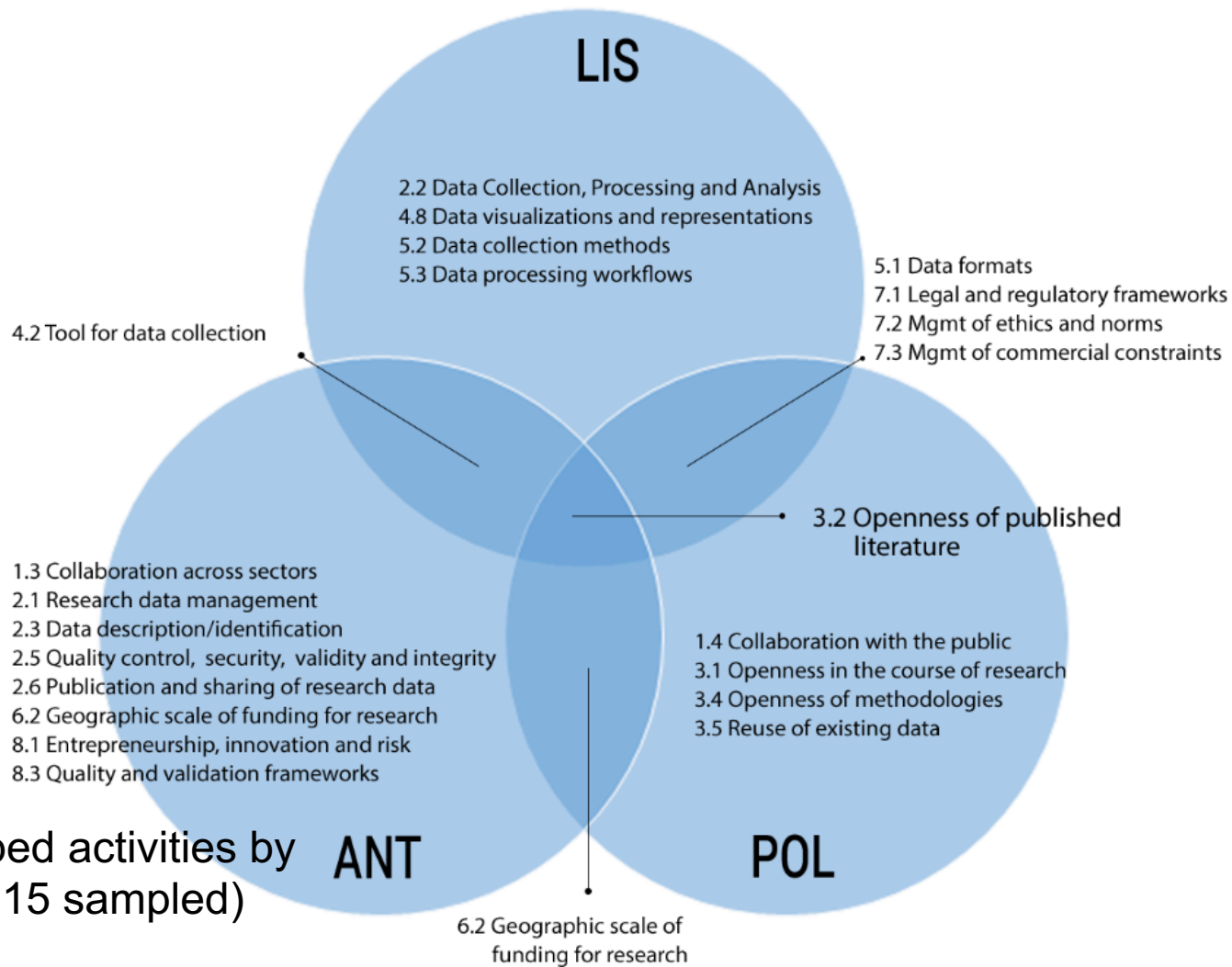


LIS participants gave better scores on Legal & Ethical, in general assigned fairly high scores on Skill & Training, Openness, Technical Infrastructure, and Common Practices than other two disciplines.

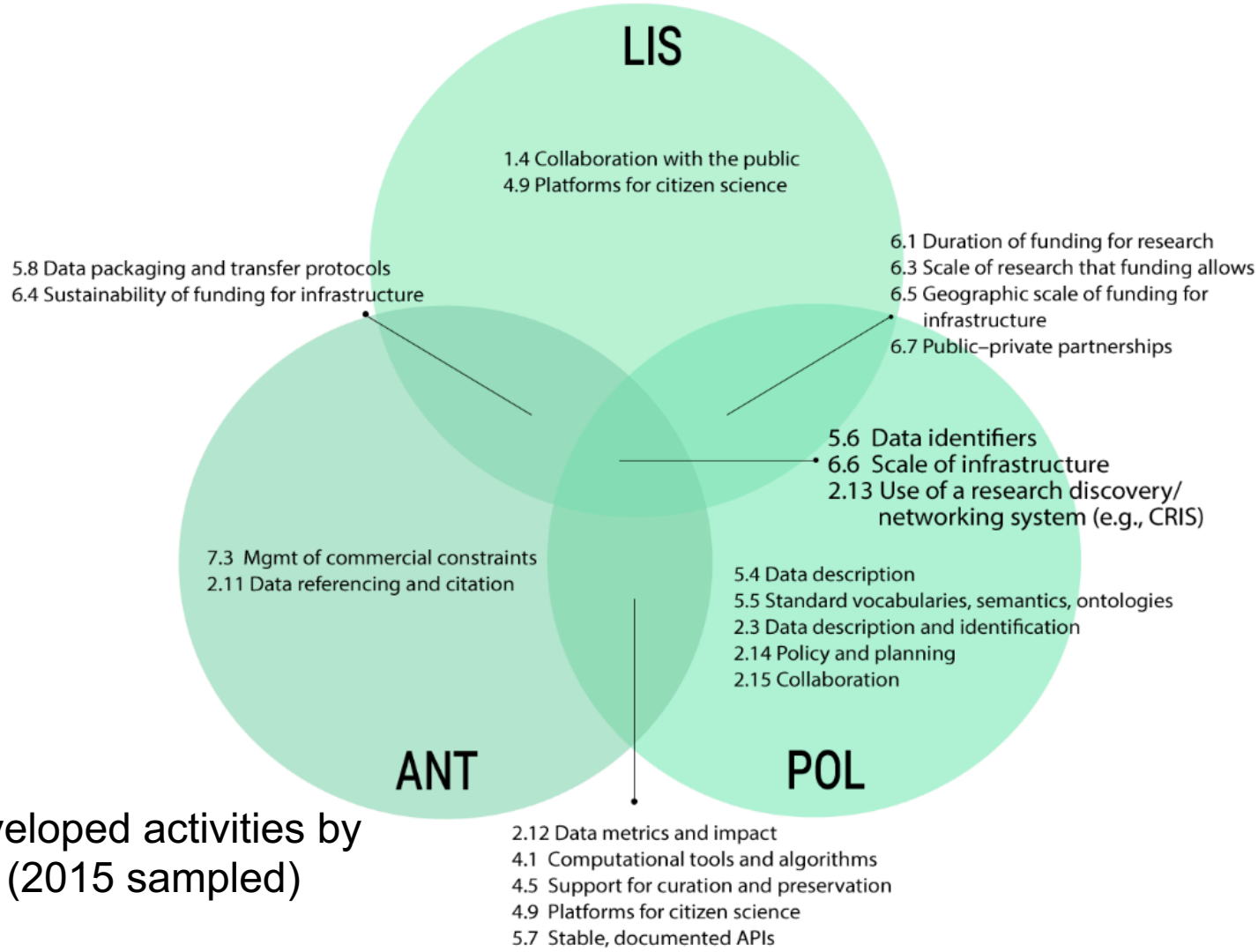
- 1: Nominal
- 2: Pockets of Activity
- 3: Moderate
- 4: Widespread
- 5: Complete Engagement.

■ Anthropology ■ Library & Info Science ■ Political Science





Most developed activities by discipline (2015 sampled)



Least developed activities by discipline (2015 sampled)

社會科學領域作為「資料集約」型科學，發展最齊全的部分

法規、倫理相關規範 (7.1, 7.2, 7.3)
資料收集、分析等相關技能 (2.2, 4.2, 4.3)

另外一方面，發展未臻的部分：

- 資料相關的描述、標準、DOI與詮釋資料 (5.4, 5.6)
- (大規模) 系統性的資料獲取協定 (4.1, 4.5, 5.7, 5.8)
- (大規模) 且充足的研究資金 (6.1, 6.3, 6.4-6.7)

Part 3 & Part 4 討論：

在收集有關校內讀者的研究模式和資料特徵之後：

- Profiling can be very time- and labor-consuming (can be up to 2-4 hours)
- The disciplines' similarity and deviation in data practices and capabilities were also worth discovering.

有助評估發展 RDS (research data service) 之其他管道：

- 多找尋領域相關文獻
近年來 The Electronic Library ISSN: 0264-0473 與 International Journal of Digital Curation 有不少與學術圖書館提供 RDS 相關的研究文章
- IDCC (International Digital Curation Conference) 國際數位策管會議
- 多與其他圖書館館員交流、分享經驗

2020 INTERNATIONAL DIGITAL CURATION CONFERENCE

#idcc20

2020 Theme | Collective Curation:
the many hands that make data work

Organised by



17 – 20 February 2020

 **Dublin, Ireland**

附錄 | 理解研究者們的資料實踐：
各領域學者如何與資料互動？
如何創造適合機構使用者的研究產品策管環境？

國立台灣大學圖書資訊學系 DxLab (Data X Lab)
Lab Director: 鄭瑋 助理教授

各領域的學者如何與資料互動？ 2017/10- present

主題資料	項目問題意識	研究方法	研究對象與樣本
生態資料 Ecological Data	公民科學計畫資料再用探索	焦點團體訪談法	6位研究人員 (南投特生中心)
人文資料 Data in Humanities	我們與次級資料的距離：華語文學者對「次級資料」之定義探析	深度訪談法+資料側寫 (data profiling)	8位研究人員 (北區大學院校)
健康資料 Health Data	生醫新創團隊對健康資料的取用與再用	深度訪談法	14位醫藥新創從業人員或醫師 (北區)

生態資料 | 公民科學計畫資料再用探索

公民科學 (citizen science) 係指

「由科學家設計計畫，利用群眾的力量，在短時間內蒐集大量的資料，以提供學術界做研究，或提供決策者參考。公民科學家從中獲得知識、自我實現、滿足感，並提高科學素養。」

動機：

公民科學在美國是圖書館的重要業務

科學傳播與增進公民科學素養與公共圖書館之任務相符

生態資料 | 公民科學計畫資料再用探索

焦點團體法

研究對象：行政院農業委員會特有生物研究保育中心（特生中心）的六位研究員（分別代表不同 project）

核心研究問題：

特生中心學者與公民之間，如何合作並且與研究數據互動？

生態資料 | 公民科學計畫資料再用探索

早期成果：

公民科學家所蒐集的資料類型包含照片、非結構文字（如生物分類或其他控制編碼）、結構化文字（如敘述）、地理座標等，最特別的是可能會攜回實體樣本（已處理過的標本或野生動物大體）。

共同的困難：公民的投入可能會帶來資料品質與資料重複的情形（如重複登錄同一件動物大體），並且目前除了人工控管以外，無法用系統或其他自動化方式檢查。另一困難為品質控管，同樣目前是以人工處理。

目前在國內公共圖書館與代表性公民科學計畫的連結還相當初步

Research data lifecycle and different stakeholders' data related activities

	公民科學家	特生中心	第三方組織
Data created/collected	✓	✓	
Data processed	✓	✓	
Data analyzed		✓	✓
Data published		✓	
Data archived		✓	✓
Data reused		✓	✓

公民科學家在資料蒐集及整理階段涉入強度最高（合乎研究團隊期待）

第三方組織則在資料分析、儲存及再利用上扮演重要角色（真沒想到，就順理成章變成碩論了？）

第三方單位舉例

中研院、暨南大學、工研院
各地鳥會、民間劇團

INPUT
技術、公民科學家

Input的單位會再
利用資料嗎？



OUTPUT
將分析後的結果提供相關單位做後續
推廣生態教育

單純利用資料的單位
會有投入的一天嗎？

中研院、義守大學、高工局、防檢局、ebird
(康乃爾大學)、民宿業者、雲林縣政府……

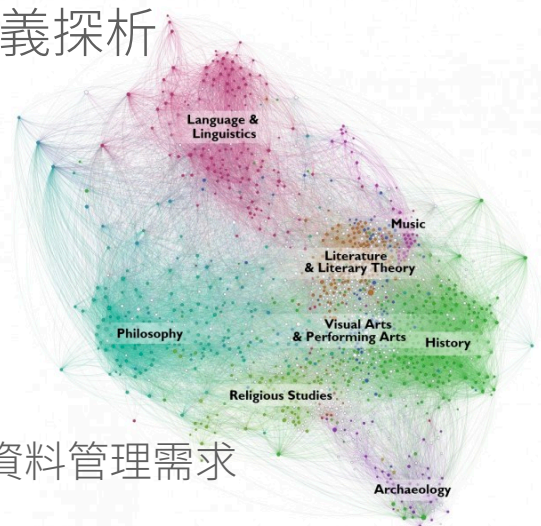
人文資料（華語文） | 華語文學者對「次級資料」之定義探析

你的資料不是我的資料、我的資料不是你的資料、
你的資料是我的資料、我的資料是你的資料、…

重要性：唯有瞭解「意義」與「情境」，才能覺察出人文學者的資料管理需求

深度訪談法與資料側寫法 (Data curation profile, DCP)

研究對象：八位華語文領域研究人員（含二位教授、一位助理教授與五位碩博士生）。



人文資料（華語文） | 華語文學者對「次級資料」之定義探析

Findings:

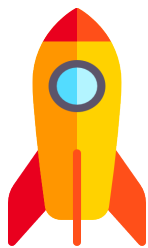
- 華語文學者在與研究材料互動時，對於原始資料 (raw data) 與次級資料 (secondary data) 之間的定義與分野模糊（訪員和受訪者越討論越模糊）
- 學者多半不願提供未完整經消化、精化拋光（polished）後的資料，亦不願將自己的錯誤攤下陽光下，也不願分享再供他人二次使用。
→與「原始資料」的定義背離
- 華語文學者對於自身的資料有高度的擁有權與保護權意識，但對於應如何保護智慧財產權的相關做法尚未達到共識。

健康資料 | 生醫新創團隊對健康資料的取用與再用

深度訪談法

研究對象：14位醫藥生技新創產業的從業人員與醫師

研究目標：探討民間新創公司對於醫療資料的需求，以及近用管道。



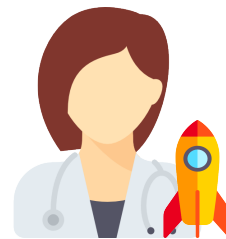
A BMI startups (N=5)

Rationale: having the fewest resources in the medical field



B Physicians (N=4)

Rationale: playing an important role in the data access and collecting process



C BMI startups + Physicians (N=5)

Rationale: After facing data-accessing difficulties, medical personnel may have different opinions

Case Study of a BMI Startup Accelerator Program: H.Spectrum

BMI Startup Accelerator



A platform that we could reach different stakeholders in BMI startups



Have an Industry-Cooperation Research Project with National Taiwan University

Hospitals



Reached out to some group B's physicians in local hospitals

Findings |

新創團隊取得健康資料的管道：

- 社群媒體，如Facebook或LINE，
- 現實世界的個人網絡(人脈)，甚至是團隊成員的病人。
- 一些新創團隊有向國家級研究院（即中央研究院）和民間的保險公司尋求幫助，但發現此類組織無權提供任何資訊（個資法）。

健康資料 | 生醫新創團隊對健康資料的取用與再用

新創團隊取得和蒐集臨床資料的困境（目前國內）：

- 幾乎沒有資源與時間獲得IRB 與TFDA 批准
- 醫護人員的血汗情形（工作都做不完了，沒有時間參與新創研發）
- 醫護人員向患者推廣新創產品也存在著風險；
- 取得和蒐集資料的成本非常高（如以前取用健保資料庫需要數十萬至百萬的規費，而現在也已經停止，釋憲中）
- 無法取得和蒐集資料（目前健保資料庫自 2017 年底始停止申請服務，民間團體「台灣人權促進會」申請釋憲中）

創造適合機構使用者的研究產品策管環境 (NTUData) 2018/05- present

日期	項目問題意識	研究方法	研究對象與樣本
2018/07	研究資料策管系統： 機構典藏資料之使用 需求初探	使用者研究+資料側寫 (data profiling)	9位校內研究人員(資工 / 資管 / 圖 資 / 數位人文)
2019/09 (預)			12-18 位生醫藥領域學者 (北區)
2020/01 (預)	參與式設計：	參與式設計工作坊	12-18 位基礎科學學者 (北區)
2020/09 (預)	研究資料策管系統		12-18 位應用科學學者 (全國)
2021/01 (預)			12-18 位社會科學學者 (全國)

研究資料策管系統：機構典藏資料之使用需求初探

研究驅動：在於建立一個學術資料機構典藏之前，先探知資訊學群學者的資料儲存管理之需求。

研究對象：9位台灣大學教師與博士生，包含圖書資訊、資訊管理、資訊工程、數位人文等

Findings |

即便同屬資訊學群，個體研究人員的團隊所產製出來的資料在數量與形式上都有很大的差異。

警示：設計系統原型之時，時刻注意以偏概全的研究限制（不能假設同個領域大家的 practices 就會很相同，根據經驗跨領域、選用相似研究方法的學者更相同），進而疏忽領域內可能的細微變異。

我的計畫

我的資料

資料使用分析

設定

我的計畫 > 新增計畫

取消

1. 計畫內容

標題

促進式資訊分享研究

副標題

以促進式資訊分享減少醫療照護之雙邊資訊不對稱

簡介

本報告包含三個互相關聯的研究。這三個研究有一個共同的主題：醫療照護產業中的資訊議題。三者各自從醫療照護提供者的供應商端、顧客端與內部營運端切入。

領域

資訊管理學

增加

關鍵字

健康照護管理

營收分享

資訊不對稱

委託單位

科技部

計畫執行期間

年/月/日



-

年/月/日



計畫書與核定清單

可上傳計畫相關重要文件（非研究資料）

新增檔案

選擇檔案

執行單位

國立臺灣大學資訊管理學系暨研究所

 我的計畫 我的資料 資料使用分析 設定我的資料 > 促進式資訊分享研究 > **新增資料**

取消

1. 上傳資料 

檔案一

檔案

 重新上傳

血栓溶解劑施打率影響因子紀錄.xlsx

格式

大小

Excel

5 MB

檔案簡述

簡單的描述這份檔案，例如：「欄位說明」

 增加檔案 下一步

NTUData

| 支持「科研資料」特色的機構典藏

1. 連動校務帳號
2. 為校內PI 打造
3. 支援研究歷程中產生之資料的管理與策展
4. 暫以「計畫」為單位
5. 支援博碩士論文專案

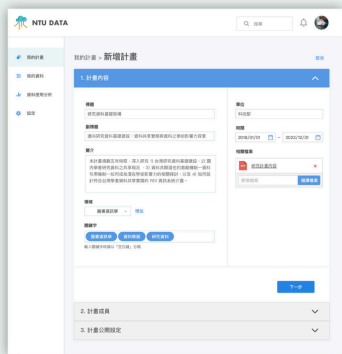
The screenshot displays the NTUData user interface. At the top, the NTU DATA logo is on the left, and a search bar with the text '搜尋' (Search) is on the right. Below the header, there are two main panels. The top panel shows the '我的計畫 > 新增計畫' (My Plans > Add Plan) workflow, with a '取消' (Cancel) button. The bottom panel shows the '我的資料 > 促進式資訊分享研究 > 新增資料' (My Data > Promoted Information Sharing Research > Add Data) workflow, also with a '取消' (Cancel) button. The '新增資料' section is titled '1. 上傳資料' (1. Upload Data) and includes a '檔案一' (File 1) section. Under '檔案一', there is a '檔案' (File) section with a '重新上傳' (Re-upload) button and a file named '血檢溶解劑施打率影響因子紀錄.xlsx' (Blood test solvent injection rate impact factor record.xlsx). Below the file name, there is a table with columns '格式' (Format) and '大小' (Size). The table shows 'Excel' and '5 MB'. There is also a '檔案簡述' (File Description) section with a text input field and a placeholder text: '簡單的描述這份檔案，例如：「欄位說明」' (Simply describe this file, for example: 'Field description'). At the bottom right of the '新增資料' section, there is a '增加檔案' (Add file) button and a '下一步' (Next step) button.



User
Principal Investigator (PI)
/ other members of the
research team

use NTU account
to **log in to**
NTUData

Create a Research Project



On the dashboard of NTUData, the PI can create a project in a very intuitive way. The activities include project description filling, team member adding (so that those member have permission to access the project and its data).

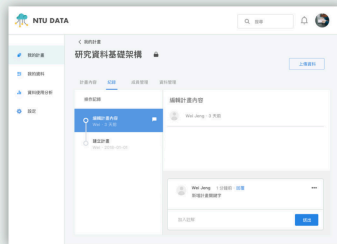
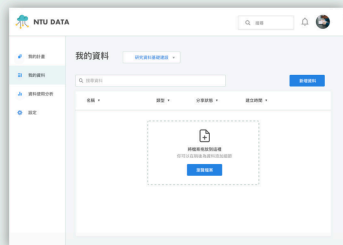


View Projects

In the project detail page, user can view or edit the description of the project. It would be convenient for the research team to record and communicate.

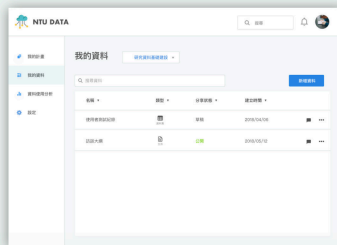
User can select a certain project to add research data in. These include metadata filling, file uploading, embargo setting, and data use permission setting.

Update Research Data



Manage Projects

Manage Dataset



View Data Use Metrics



User can view data use metrics (who views, downloads, or shares their published data) in a visualized manner. In addition, these metrics can be exported as a excel format to help user better track the usage.