Open Data and Information for a Changing Planet

DAILY REPORTS

Time & Location: **October 30, 2012** @ Academia Sinica, Taipei, Taiwan
Scientific Domain: Social Science (sessions: D2, D4, HL 2.1, HL 2.2, E2, F1 and F2)
Report prepared by: Joseph Chen and Andrea Huang

What do we mean by open access to data? It's highly unlikely that we will find a global consensus for the meaning of the term, especially when the term concerns about openness within diverse scientific domains. After some empirical practices of data sharing and an overall introduction on open knowledge environments (OKE) in the previous day, the challenge is hence to be initiated for an open discussion in Day 3 of CODATA2012 high level session, and together with a parallel session to discuss the ethical integrity of data science, from the perspective of humanity.

Following by two specific sessions on data publication and citation (part I & II) and data journal as a fourth paradigm, researchers, scientists and participants thus have an opportunity to enter dialogues for mutual understanding. Projects and policies are introduced in two sessions: from a regional/national level perspective, and from a mass collaboration/citizen science perspective.

The Meaning and the Ethics: The open data is to be free, liable, accessible and permanent, and to be opened intelligently. In this context, governments are suggested to make clarification and legislation for open data. On the other hand, the international cooperation could work beyond copyright. The principle for the full and open exchange of data is suggested to be non-discriminatory for public interest, and to be used from use prospective. The proposal for more related business models will also help data sharing more sustainable. A more deep clarification of the open data meaning can be referred to the keynote speech of the same day.

Policies and laws are outside regulations, the discussion in the high level session suggest that we should look inside, the Ethics of Data Sciences. Ethical standards and confidential issues should be taken into mind when it comes to paper publishing. In this aspect, we are dealing with "character" instead of "data". If we want to prevent professors from being "paper publishers", students from being "copy and paste editors" or "paper translators", we shall review the relation between school ranking and grant mechanism or sponsorship and faculty structure. Sponsors, peer-review, and media can all be the supporting mechanism.

Data Publication, Citation, and the Fourth Paradigm: In the context of big data and data-intensive science, the Fourth Paradigm takes the view on our capacity to store, validate, analyze, visualize, and curate the information. Interdisciplinary efforts to the open access have been demonstrated in several cases of e-Journals, such as *Earth System Science Data Journal* (ESSD), *Institute for Scientific Information Journals* (ISI), *Directory of Open Access Journals* (DOAJ), or the *Data Science Journal*. Several proposals like observation data shall not be limited on visible items; changes need to be

made for the current mostly used PDF format; and semantic links between article and data are suggested.

However, one study indicates that 75 % of research data is never made openly available. The biggest obstacle for data publication is argued for the reason that scientists are lack of incentives to share their data after their journal submission. Thus, the prerequisite for open accessed data to gain more recognition is to focus on quality assurance and quality control as the fundamental incentives to users. Different perspectives and approaches like standardization, metadata format, or peer review (i.e. combine reviews between different data archives and journals), have been proposed as follow:

(1.) peer review of data paper as the main component in the data publication workflow is illustrated from a data journal editor's point of view (ESSD) ;

(2.) Digital Object Identifier (DOI) and metadata format are general approaches to increase the discoverability and visibility of research data, as well as to add valuable context to the data publishing and citation.

For instance, the ARM Data Citation Service in the Climate Change Science Institute provides services of DOI to link data and documents. Projects such as the Article of the Future incorporated with interactive widgets (e.g. 3D viewing of chemical structure), and supplemental materials (which include multimedia, table, figures, text, data sets, and even computer algorism and code) are cases which have also been addressed with such technical issues. In addition, the UK-based Digital Curation Centre (DCC) further works on how to cite datasets and links between metadata and publications.

(3.) Interactive design. More integration of text and data, viewers and seamless links to interactive datasets is top of the concern in an ideal of Data Publication Pyramid. The other project introduces the Open AIRE, which is a publication repository of OA journals. In particular, the ability to link from a publication to a citable database, or to other research materials shall enable users to interact different information objects.

However, citing data is not equal to the way we cite literature. The differences include:

(1) the dataset may only exist one copy in one location but a published work of literature generally has multiple copies in different locations;

(2) the attribution stacking problem in which the appropriate proportion of the intellectual credit becomes a challenge if the dataset is constituted by multiple sources;

(3) a citation to a literature object generally has sufficient functions in the metadata to address *Identification*, *Disambiguation* and *Equivalence* (hardcopy vs. paperback) but a citation to a data object is a subset of the metadata.

Thus, the argument for the metadata of data citation is made to be about the practice, not the format.

Projects and Policies: Discussions have been made both for regional/national level and mass collaboration level. European Data Infrastructure (EUDAT), National Research Council of Canada (NRC) and some projects in Taiwan and Hon Kong are such regional/national cases.

Based on a collaborative data infrastructure, the EUDAT brings together 25 partners from 13 countries to create a pan-European e-infrastructure. The NRC in Canada is on its national and organizational transformation process, and which brings it into an increased focus on industry and economy needs. For example, the launch of DataCite Canada, collaborating with a UK-based international DOI foundation, is a service for research community to add value through data registration.

Furthermore in Asia, a Hong Kong project is presented through the showcase of digital cultural content, while two other types of data projects in Taiwan are presented with statistical data and geospatial data. In general, most researchers in these projects are aware of the importance of enhancing data semantics through using semantic web technologies and Linked Open Data principles. Meanwhile, the demand of data transparency and openness are called to be freed from governments' constraints.

Among the fundamental lessons of mass collaboration projects are the importance of user participation, data management and data licensing. From the perspective of copyright history, the role of user normally been treated passively in the proprietary regime. This has been improved through the implementation of mass collaboration projects. A Roadkill project in Taiwan using social media (e.g. Facebook) demonstrates how user participation could assist endemic species research.

Data citation, tracking and licensing are main concerns for the Earth-Base project; the implementation challenges of the System for Earth Sample Registration (SESAR) project are the diversity of user requirements for sample registration procedures, naming protocols, metadata, and access policies.

In contrast, Open Street Map (OSM) adopts relatively loose policy principles, such as defining minimal basic rules, staying practically to collaborate with dynamic open source projects, avoiding heavy standards, and not being architecture astronaut. The final argument of this session is that the key to data management is the data policy choices made by mass collaboration projects; and at what level should "policy" be considered depends on issues such as international, institutional, jurisdiction, standards, projects, and individual contributor.

何謂開放資料呢？由於開放的定義牽涉多門科學領域，因此要定義出一個普世認定的意義似乎是不可能的。因此，基於前日會議中已經概略介紹資料分享的實例、以及開放式知識環境（OKE），COADATA 2012 會議第三天則在高階論壇中開啟更廣泛的討論，同時也在平行的場次，從人文的角度出發，針對科學倫理進行討論。

藉由參與資料出版與引用二個議程、以及討論資料期刊為第四典範的議程，研究者、科學家、以及參與者因而有機會做更深入的對話。另外，政策與計畫則分別由二個不同層級的觀點進行個案觀察，分別為區域性/國家的層級、以及大眾協同合作/公民科學層級。

意義與倫理: 開放資料的主旨是在使大眾能免費自由存取交換、可靠的、及永久的資料、同時必須能「明智的開放/ intelligent openness」。在此脈絡下，高階論壇中建議政府有責任立法闡明開放資料，但另一方面也期望能在著作權外與國際合作的關係間取得平衡。對大眾全面的資料開放，必須是無差別待遇、且能基於使用者觀點為出發的原則。而為達永續發展的目標，則需要更多的開放資料企業模型的發展。另外一個更完整的開放資料的意義，可參考同一天的專題演講內容。

政策或法規均係外在規範，高階論壇中的討論建議我們應該反躬自省，檢視資料科學的道德。研究人員在資料發表時需時時謹記道德標準和機敏資料的保密。在此觀點下，我們處理的其實是「人格」而非「資料」。若我們不希望看到教授變成「研究報告出版者」、不希望看到學生變成「剪貼編輯師」或是「報告翻譯師」，那就必須重新審視「大學排名—補助／資助—師資」的架構。與此同時，（大學）贊助者、同儕審閱和媒體都可以是支持資料科學道德的機制。

在巨量資料與資料密及科學的脈絡下，第四典範就我們對於處理資料的能力，認為基本能力包括資料儲存、分析、視覺化、與規畫展示。在此會議中，各學科間致力於資料開放的努力成果，則由一些電子期刊的案例分享來說明，其中包括地球科學資料期刊(ESSD)、 科學資訊機構期刊 (ISI)、開放取用指南期刊 (DOAJ)、或是資料科學期刊等。

然而研究指出，大約 75%的研究資料從未公開大眾取用。其中最大的障礙是科學家在期刊出版後，失去使資料一起出版的動機。因此，學者建議能驅使開放取用資料受到使用者認可的基本前提是，強調資料的品質確保性、以及資料品質的控管。而其他不同的觀點與方法建議包括標準化、後設資料格式、或是同儕審查 (結合不同資料庫與期刊的審查資料)，分別敘述如下:

(1.) 以資料文件的同儕審查，作為資料出版流程中的主要部分 (ESSD 編輯提出的觀點)

(2.) 物件辨識碼 (DOI)以及後設資料格式，是大多數研究所提出的方法，目的是能增加研究資料的能見度、以及被尋獲的能力。同時也能提高資料出版與引用的脈絡。

例如，氣候變遷科學研究資料中心旗下的 ARM 資料庫介紹他們所提供的連結資料與文件的 DOI 服務。或是在「未來的文章」計畫中結合互動機制 (如將化學結構 3D 視覺化)、補充資料(包括多媒體、圖表、資料集、甚至是電腦程式演算法與原始碼等)。除此之外，英國的數位策展中心(DCC)也進一步展示他們如何引用資料集、以及連結後設資料與出版物。

(3.) 互動式設計。例如在資料出版金字塔模型中，即強調整合文字與資料、緊密連結視覺化介面與互動式資料集。另外一個開放資料期刊的資料庫計畫 Open AIRE 則強調在出版文章內文中的連結，以及與引用資料庫的連結、或是連結到其他研究材料，使用者因而能與不同的資訊物件產生互動。

然而，引用資料與我們引用文獻有所不同。其差異點包括:

(1) 資料集可能只存放在一處，但一般文獻則在許多地點均有複製的文件存放;

(2) 若資料集是由多方來源所組合成，適當的智慧權分配比例將成為屬性堆疊的問題;

(3) 一個文獻物件的引用，一般說來在後設資料中以具備足夠的功能處理識別(Identification)、釐清(Disambiguation)、與對等性(Equivalence/精裝版與平裝版)，但一個資料物件的引用，卻是引用後設資料中的子集。

因此，論點著重在資料引用的後設資料是關於實作，並非格式。

計畫與政策: 在此議題下的討論分為兩方面，分別為區域性/國家的層級、以及大眾協同合作/公民科學層級。泛歐洲倡議資料庫（EUDAT）、加拿大國家研究協會 (NRC)、以及亞洲的台灣與香港的案例分享，均包括在區域性/國家層級的案例中。

泛歐洲倡議資料庫（EUDAT）是建立 在一個資料協作的基礎建設架構中，此計畫包括來自 13 個國家的 25 個組織成員。而加拿大國家研究協會 (NRC)，目前正面臨組織再造的過程，也因此轉而強調工業與經濟需求的取向。例如與英國的 DOI 組織合作的 DataCite Canada，積極推動透過資料的註冊服務，因而期望能增加研究社群研究成果的價值。

另外在亞洲的香港則介紹以數位化文化內容的經驗、在台灣的案例分享則以政府的統計資料、以及空間資訊資料為說明。一般而言，這些研究者的計畫經驗，都使他們瞭解到透過語意網技術來增加資料的語意、以及開放式連結資料(LOD)原則的重要性。同時，要求政府放寬資料的限制，在資料透明化與開放的需求上都提出呼籲。

在群眾協同合作計畫案例中，作為政策考量的重點是使用者參與、資料管理、以及資料授權。就著作權歷史的角度觀察，使用者的角色在所有權領域往往是被動消極的，但藉由群眾協同合作計畫的實行，使用者角色已漸漸被提升。另外，在台灣的一個利用社群工具(Facebook and Google Map)的路殺研究計畫中，也顯示出使用者的參與，能協助地域性物種的研究。

在 Earth-Base 地球科學資料計畫中，特別重視資料的引用、追蹤與授權; 而地球樣本註冊系統計畫(SESAR)則強調在樣本註冊過程、命名的通信與資料傳輸規則、後設資料、與資料取用政策。將因使用者需求的多樣化而面臨挑戰。

相對的，開放街道計畫 (OSM)則採取政策寬鬆原則，例如定義低限度的基本規則、與開放原碼計畫機集合作、避免採用繁重的標準、並且不要過度強調架構。此議程的最後論點是對於群眾協同合作計畫而言，資料管理的關鍵在於他們對於資料使用策略所做出的選擇; 而資料策略的選擇，決定於國際的、機構的、司法裁判範圍、計畫、以及個人貢獻等因素的考量。