Open Data and Information for a Changing Planet
DAILY REPORTS

Scientific Domain: **Bio-Med Science (sessions: A3 and C3)**
Time & Location: **October 29, 2012** @ Academia Sinica, Taipei, Taiwan
Daily Report by: Ryan Guan and Andrea Huang

The following report summarises the discussions and outcomes of two sessions in CODATA 2012 conference. We try to recap several aspects of empirical experiences in sharing public health data, and of several technical issues discussed in microorganisms and bioinformatics.

Three case studies in the Project Tycho, which is a collaboration between University of Pittsburgh and Taiwan CDC, are presented to discuss their opportunities and challenges in opening data and information. In the sharing of public health data, three positive aspects are recognised for decision makers, research communities, as well as general public:

(1) supporting decision makers for policy making and practical actions;
(2) assisting research communities for data reuse and scientific analysis;
(3) opening the access for public to diverse health data.

On the other hand, researchers have also identified several barriers within their domains. For example, motivations and incentives; ethical conflicts, privacy and legal concerns (i.e. open access licence); economic and political difficulties, or technical issues are those obstacles remain unresolved. Among which, the technical problems have been further discussed in topics of data format, standardization or data rescue in surveillance database. In Tycho project, for instance, names of disease are hard to reach consensus between different source providers. In addition, the name formats of patients between different databases are not only difficult in standardization; they are further challenged with ethical issues, in which they are under the circumstances of lacking international ethical codes for data-sharing guidance.

Even if the Expert Finding System of Indian Council of Medical Research has tried to provide an information naming system for solving the naming problem, the system still cannot apply to PubMed (its major data source), simply because the disambiguation of author names remains a challenge to work on. Nevertheless, how to make public health data linkable is the key to make data accessible to the public. Project Tycho has therefore proposed two methods for our guidance:

(1) using *geospatial patterns* of population, transportation, migration, and climate to correlate disease dynamics

(2) linking health data to the *Semantic Web* by converting all Tycho data to RDF; linking to existing ontologies such as Infectious Disease Ontology, GeoNames, US Census, as well as US National Climatic Data Center (NCDC – NOAA)

Most of the discussion about information platform for microorganisms and bioinformatics focused on the ways how information communication technologies have assisted biomedical research to integrate existing databases into knowledge-based systems.

Started by databases of World Data Centre of Microorganisms (WDCM), the Global Catalogue of Microorganisms (GCM) is proposed to construct a data management system and a global catalogue to facilitate organizing, investigating and sharing data resources of its member collections. At the same time, quality control issues were also suggested, such as applying standards like OECD Best Practice Guidelines; providing a template of catalogue, metadata table, automatic quality checker for checking data of organism type with its species information; or checking the sequence information and the nomenclature information for microorganisms.

In Taiwan, the Bioresource Collection and Research Center (BCRC) of Food Industry Research and Development Institute (FIRDI) collaborated with the Global Biodiversity Information Facility (GBIF) has dedicated to establish data exchange services, and to make bioresource data freely and universally available on the web. Approaches adopted like using the platform workflow to manage bioresource; using barcode technologies to convert paper-based knowledge to digital data; or using the microbial common language (MCL) structure, are such examples for BCRC platforms to advance their knowledge-based system to achieve network linking visions.

In a further move to the concept of Open Knowledge Environment (OKE), which has also been set in sessions of (B5 and C5), the emphasis of this talk is on many successful OKE examples in microbiology. Suggestions of some principles are made for the bio-med community to reach an OKE vision:

(1) maximizing public good,
(2) avoiding monopolies,
(3) keeping low costs,
(4) remaining the freedom of inquiry,
(5) maintaining essential characteristics of communities

此份報告總結 CODATA 2012 國際會議中兩個場次的討論及結果。並試從各種觀點討論公共衛生資料公開的實務經驗、微生物及生物資訊的技術性問題

Tycho 計畫(匹茲堡大學、台灣疾病管制局合作)提出三項個案研究以討論開放資料與資訊的的機會與挑戰。就開放公共衛生資料來說，有三項對施政者、研究社群、及大眾的好處：

(1) 輔助施政者及各式決策者、
(2) 促進研究社群資料再利用與科學分析、
(3) 普及多元的公衛資料。

另一方面，研究者也遇到一些該領域中的阻礙。例如，動機與需求、道德衝突、隱私權及法律考量(如開放取得授權)、政治經濟難題、或技術性問題等還未解決。其中，格式、標準化、或監測資料庫的資料救援等技術性問題被進一步討論。例如，Tycho 計畫中，不同資料來源的疾病命名分歧。另外，不同資料庫中病患姓名的格式不只難以標準化，還有道德上的爭議。主因是資料開放的國際道德法規不健全。

雖然印度醫學研究委員會的專家搜尋系統針對此問題曾提出資訊命名系統，但它無法套用在生醫文獻資料庫(它主要資料來源)上，因為庫中作者姓名格式分歧。不過，如何連結公共衛生資料仍是使資料開放給大眾存取的關鍵。因此 Tycho 計畫提出兩種指導方針：

(1) 使用人口、運輸、遷徙、及氣候的地理空間模式去推估流行病學。
(2) 將 Tycho 資料轉換成 RDF 格式，以連結衛生資料與語意網。連結至既存的知識本體，如傳染病本體、基因本體、美國人口普查、美國國家氣候數據中心(NCDC – NOAA)。

大部分微生物與生物資訊平台的討論聚焦在於，資訊傳播科技如何幫助生醫研究去整合現有資料庫至知識基礎的系統。

始於國際微生物資料中心(WDCM)的資料庫，全球微生物名錄(GCM)提議建立資料管理系統及全球名錄，以促進其同盟收藏資料的整理、研究、及開放。同時也有品管問題，例如達到 OECD 最佳施行準則；提供名錄模板、後設資料表格、生物類型及其物種資訊的自動品管系統；或檢查基因序列及微生物命名法資訊。

在台灣，財團法人食品工業發展研究所(FIRDI)的生物資源保存及研究中心(BCRC)與全球生物多樣性資訊機構(GBIF)合作致力於建立資料交換服務、及於網路上自由普遍地開放生物資源資料。BCRC 能實行方法，例如使用平台工作流程去管理生物資源；使用條碼技術以轉換紙本至數位資料；或使用微生物資源共用語言(MCL)結構，來改進知識基礎的系統以達成網路連結的願景。

進入至開放式知識環境(OKE)的概念，同於 B5 和 C5 場次中提及，此演講著重於許多微生物學中 OKE 成功的例子，並提供生物醫學社區一些在 OKE 願景上的建議：

(1) 公共利益最大化，
(2) 避免市場壟斷，
(3) 壓低成本，
(4) 保持詢問的自由，
(5) 維持社區核心特質。