Open Data and Information for a Changing Planet | DAILY REPORTS

Time & Location: **October 30, 2012** @ Academia Sinica, Taipei, Taiwan
Scientific Domain: Bio-Med Science (sessions: D3, E3, F3 and F5)
Daily Report by: Ryan Guan and Andrea Huang

Given that most discussion on information platform for microorganisms and bioinformatics (part I) has focused on how to integrate existing databases into knowledge-based systems, we might wonder some questions. Whether biomedical science matters to the problem of the planet under pressure, or what existing technologies has been provided to improve data integration and analysis while the diversity of bio-medical data is complex? .

A look at photosynthesis capabilities in bacterial species tackling CO2 emission makes a database of oxygenic photosynthetic gene clusters important. The extensive diversity of fimbriae and lactamase identify the significance of structure of the task of genes encoding in database management; the use of open databases (HAGR, BiGRID) is proved positively to assist the research of mus musculus aging genes' topological analysis and prediction; and the semantic enhancing to the virology modelling study is benefited by applying data visualization techniques. These five talks yield some helpful insights into the relation of biomedical science to the global change problem with illustrations of diversities in bio-med data characters, and cover some issues of the scientific topic, *Semantic Web and Linked Data*, that might offers technical solutions to the data integration and application.

Of course, if we recall the Tycho project in the previous day, open access platform and semantic web approaches are underlined in the conference discussion. Similar to Tycho project, the case of INDICATOR platform is designed to be a public health platform with open access principles. Nevertheless, while Tycho project focuses on disease data linkage, INDICATOR tries to build a platform for multi-sources monitoring and modelling. The integration of different domains is further to be proposed to combine human, animal and environmental data sources as a whole picture. Aside from the discussion of technical issues in INDICATOR, the platform of bioinformatics (part II) has also moved its concerns to big data, which is presented with several following issues:

(1) challenges of data-intensive science that deal with approaches to process big data simultaneously;
(2) ideas for genome information society which takes the view of regarding big genome data as a basis for understanding human evolution (from the perspectives of Biomedical Genomics and Environmental Metagenomics);
(3) semantic interoperability of big data for biomedical knowledge management, as well as approaches like de-identification, data encryption, or access control differentiations are suggested for the privacy issue.

Moreover, the double insight of big data is followed by two sessions of information infrastructure design for biological and biodiversity science, namely infrastructures for biodiversity information and infrastructures for bio-economy.

Considering the infrastructure efforts for biodiversity data, the Chinese Academy of Sciences has followed the global trend in developing the e-Science infrastructure (*cyber-infrastructure*) with the recent progress in biodiversity informatics, open source techniques, as well as the emerging framework of user participation as citizen science. For instance, the Biodiversity Heritage Library provides an open access repository for biodiversity publication, which allows online group discussion; biological field observation data of the Chinese Field Herbarium biodiversity information system is collected and built by online community collaboration.

In addition, infrastructure designs such as specimen catalogues, service-oriented architecture, or visualization layers with geospatial system design are incorporated in several platform introduction in the national project: the National Specimen Information Infrastructure (NSII). In particular, NSII categorize its services into three categories: Data as a Service (DAAS), Software as a Service (SAAS), and Knowledge as a Service (KAAS).

Apart from big data, technologies, or legal and policy concerns, the bio-economy is a relatively new issue, especially to the CODATA community and to the bio-med science. It discusses the value of biological science collections that can transfer scientific knowledge into innovative and sustainable products; the biotechnological processes of data storages and knowledge managements; as well as financial constrains of organizations or sectors who take part in biological science collections.

For example, in the panel discussion, participants recognize the decline of specimen data collection causing a knowledge gap. And reasons for this phenomenon include the lack of funding, the decreasing trend of field works, and the uncertainty of related policies and legal systems (e.g. the conflict of animal rights in Japan, or the ownership of specimen should be remained in the original country or to the one who collect them).

By contrast, technical issues are more inspired while comparing with above complex social and policy issues. Avoiding building a single huge database for the big biodiversity data, and adopting advance biotechnological processes of data storages and knowledge-based managements are evidently shown in several case studies:

(1.) the university museum collection network is collaborated by Japan, China and Vietnam;

(2.) the Integrated Digitized Biocollections (iDigBio) in U.S. is established by the community of the Network Integrated Biocollections Alliance with the implementation of a scalable and open assessable cloud-based infrastructure;

(3.) the concept of heterotopia with its spatio-temporal dimension is realized and put into data services in the case of National Digital Archive Program of Taiwan;

(4.) the Biodiversity Research Museum of Academia Sinica, Taiwan adopts open data and information principles and constructs its database infrastructure as a three-tier web architecture (Presentation tier, Logic tier and Data Service tier).

To sum up, while cases of information platform and big data give us the best continuous of technical discussions from the previous day, bio-economy provides an echo to the challenges and opportunities in open data sharing. The domain of bio-medical science in CODATA 2012 conference has six sessions in two days. (Two more e-Health sessions focusing on specific technological issues are in B3 and E4, summarized in the computer science daily reports) The information platform, big data, and infrastructure design are three main components, which highlight the domain. The key to link these fundamentals may rely on open access, semantic web technology, as well as the universal consensus-international collaboration.

既然在大部分微生物與生物資訊平台(場次一)的討論裡，我們已著重如何整合現有的資料庫至知識系統，接下來我們或許會思考生物醫學科學是否能協助解決壓力下星球的問題、或是如何在複雜多樣的生醫資料中，尋找現有的科技以幫助資料整合及分析。

當我們檢視細菌光合作用是具備對抗二氧化碳排放的能力時，生氧光合作用基因簇資料庫則顯現出其重要性。在資料庫管理中，大規模的菌毛和內醯胺的多樣性確認了資料架構在建構基因編碼時的重要性；而開放資料庫(HAGR, BiGRID)則協助研究者進行小家鼠老化基因的拓樸性質分析及預測；同時，資料視覺化技術亦協助增進病毒生物學模型研究的語意研究。簡而言之，本五場次對生醫科學及全球變化的關係提出了深刻的見解、展示了生醫資料的多重角色、以及含蓋了被視為可能解決資料整合及應用等技術問題的語意網路與資料連結 (此會議的一項新的科學主題)。

當然，在前一天的 Tycho 計畫中，已強調開放存取平台與語意網的方法。同樣的，指標平台(INDICATOR)的目的是在開放存取原則下，幫助公共衛生的發展。然而，Tycho 計劃著重在疾病資料連結，相對的，指標平台則嘗試打造多源觀測與計算模型平台。該計畫的特色是，全面的跨領域整合人類、動物及環境的資料來源。另外，除了討論指標平台技術等問題，微生物與生物資訊信息平台(第二場次) 則將焦點轉移至巨量資料。要點如下：

(1) 密集資料科學的挑戰，包括同時處理巨量資料的方法；
(2) 提出基因體社群的想法: 藉由巨量的基因體資料(結合生醫基因體學、環境元基因組)來詮釋人類演化；
(3) 生醫知識管理的巨量資料在跨平台語意溝通力。另外亦包括隱私權、去識別化、資料加密、存取控制區別等知識管理議題。

此外，對於巨量資料的雙面觀察，則由資訊基礎建設的兩個面向切入。由緊接的二大議題，即生物多樣性資訊基礎建設及生物經濟資訊基礎建設，進行經驗分享與討論。

在對生物多樣性資訊基礎建設的案例中，中國科學院已跟隨 E 化科學基礎建設(資訊基礎建設)的全球潮流。近期進展包括生物多樣性資訊、開放原始碼技術、以及所謂包括使用者參與的公民科學的科學領域新架構。例如，生物多樣性遺產圖書館提供生物多樣性文獻的開放存取庫(提供線上群組討論功能)；中國自然標本館的生物多樣性資訊系統的生物田野調查資料，由線上社群合作收集與建立。

另外，在國家標本資源共享平台(NSII)的許多平台簡介之中，資訊基礎建設也嘗試設計整合樣本名錄、服務導向設施、或視覺化地理空間層級。而 NSII 也因此將服務導向設施的架構分成三類：資料服務(DAAS)、軟體服務(SAAS)、以及知識服務(KAAS)。

除了巨量資料、科技、或法律及政策的考量，生物經濟則相對是一個新的議題。對 CODATA 社群和生醫科學界更是如此。次議題論及生物科學樣本在轉換科學知識至創新與永續的產品價值、資料存取及知識管理的生物科技進展、以及生物科學樣本收集的機構或部門的經濟限制等課題。

例如，在專題討論中，與會者發現目前學界對樣本資料收集的減少趨勢，已經造成知識斷層的現象。其原因包括經費補助不足、田野調查工作趨向沒落、相關政策與法律系統的不確定性(如日本的動物權益抬頭、及樣本收集者與樣本原屬國家所屬權爭議)等因素。

相對於以上複雜的社會與政策議題，技術性議題則在生物經濟資訊基礎建設的討論中，更具啟發性。其中，避免為龐大的生物多樣性資料建造單一巨大的資料庫、或是採納資料儲存與知識導向管理的進階生物科技程序，已經具備許多個案研究，如下所述：

(1.) 日本、中國與緬甸合作的大學博物館樣本網絡；
(2.) 生物樣本網路整合聯盟社群，成立美國整合數位生物典藏(iDigBio)，該計畫採用具有可擴張性及可公開存取的雲端設施；
(3.) 台灣數位典藏國家型科技計畫以異質地域的時空維度概念，應用在資料服務；
(4.) 台灣中央研究院生物多樣性研究博物館採用開放資料與資訊原則，並建構三層式網路架構(使用者介面層、邏輯層、和資料服務層)的資料庫設施。

總結 CODATA 2012 會議這兩天在生物醫學科學領域，總共有六個場次 (另外兩場有關 E 化健康管理議題，著重在更深入的技術層面探討，因此收錄在計算機科學領域的報告中)。資訊平台與巨量資料的個案沿續了兩天技術性議題，而生物經濟議題的討論，則對於本會議的主要課題: 資料開放存取的挑戰與機會，進行回應。其中三大中心主題橫跨資訊平台、巨量資料、與資訊基礎建設。而串聯此三大基礎的關鍵則是開放存取、語意網技術、以及跨國合作的普世共識。