Time & Location: **October 30, 2012** @ Academia Sinica, Taipei, Taiwan
Scientific Domain: Computer & Information (sessions: D1, E1, E4, and F4)
Report prepared by: Henry Chang, Ming-Huang Wang and Andrea Huang

---

Increasing amounts of data are being produced worldwide, and the archiving, curation, and sharing of data are critical for analysis and thus making data meaningful. Day 3 of computer science sessions follows the previous issues in big data and data intensive, but moves further the theme toward some practical dimensions: data preservation and dissemination; tools in open data environment; computational infrastructure; as well as e-Health management. While technologies about *cloud computing, standardization* and *interoperability*, as well as some semantic web discussions remain key concerns, more conceptual frameworks for an open sharable, accessible and citable environment are proposed.

Computational Infrastructure: Issues such as data management in 4V (volume, velocity, variety, validation), or complexity, lineage, and user interface etc, are necessary challenges to computational infrastructure design. This session presents national information structures in China and Japan.

The Cyber infrastructure and e-Science program (from 2001-) of Chinese Academy of Sciences develops the CAS Data Cloud, which integrates various cloud services. Among theses, Geospatial Data Cloud servers for geospatial research aggregate open data resources (i.e. crawling metadata and cache data entities from open data, reorganizing raw data and push them to the cloud database). In addition, the National Science and Technology Infrastructure of China project (NSTI, 2002-) uses the Data Pyramid Model to describe data in four dimensions (raw data, basic attributes, physical description, and semantics). Both national infrastructure programs have developed and will continue to improve systems like its Data Management & Sharing and NSTI Portal, for its national scientific endeavours.

Furthermore, “Data shared, but not found” and ”Found, but not comprehensive” are the main challenges for Japan’s National Institute of Information and Communications Technology (NCIT). Search engines are thus developed to fully extract archive information by correlating spatiotemporal, ontological, and citational structures.

Tools and Methods: New tools are developed for big data mining, leaning, analysis and dissemination. A research project has been proposed for understanding current states of data analysis for mining big data, and for identifying gaps in theory and practice. Three big data technologies (two are open source software) are compared for scientists to conduct interactive big data analysis. The result shows that the commercial tool displays the best performance while the other two open source tools posse potential for prime-time operation and processing the locality of data.

In addition, the Common Data/Service Environment, based on OGC, W3C and OASIS standards, is proposed for enabling easy access to and integration of distributed multi-source geospatial data. GeoBrain and Global Agriculture Drought Monitoring and Forecasting System (GADMFS) are such cases for implementation. A new method, Data Prospecting, which combines two traditional data exploration and mining techniques, focuses on finding the right subset of data amongst all the data files based on the content stored in the file.

How tools can be integrated into the open collaborative environment for data processing and analysis, assimilation, mining, and visualization is interlinked with previous studies in Day 2. Toolkits for building collaborative environments by open-source components (Talkoot), for data mining, for interactive visualizer and image classifier, or for semantic search are some cases which can be integrated within the GLIDER framework that has been introduced in the previous day.

Another method for the e-Health management following the previous B3 talks on standard and interoperability, system framework and knowledge modelling is drawn to the discussion in the panel. And, the unification efforts between the HL7 (USA) and ISO13606/openEHR Archetypes (Europe) by the Clinical Information Modeling Initiative (CIMI) are further investigated. In addition, XML technologies in the e-Health issue are once again to be explored with some promising implementation cases in Japan's Dolphin Project and Brazil's industry's experience. At the same time, the other highlight of today's discussion is about an object model of the semantics of archetypes (the openEHR Archetypes), that is one of the design components in the aforementioned openEHR project. It is proposed for industry structure design. The character of this archetype is introduced for being capable of re-usable and composable (i.e. like a LEGO instruction sheet, which defines the configuration of LEGO bricks making up a tractor.)

Preservation and Dissemination: Pre-digital data may locate in - somewhere - paper, film, photographic plate, but are not accessible on-line so they are "lost" in the digital era context. Therefore, the issue of data rescue and preservation is to recover and to preserve those "lost" data of various kinds (including High Energy Physics data in E1). Pilot projects regarding the rescue of data at risk and forgotten databases are proposed to assist research outcomes to be sharable, citable, and discoverable. Tools for optimizing the dissemination of information easily and low costs are presented by the case of e-book. The research uses the Free and Open Source Software (FOSS) to assist the implementation of digital library applications, and develops the e-book system with three-dimensional format integrated with a variety of formats. Thus data and information are ensured to be easily accessed from any device, and to be disseminated to everywhere.

Specifically, Figshare is a data repository where scientists can make all of their research data and results available in a citable, sharable and discoverable manner. This system allows users to upload any file format to be visualisable in the browser, so that figures, datasets and media can be disseminated in a way that the current scholarly publishing model does not allow.

Meanwhile, knowing the nature of archives has been changed from preservation focus to data accessibility and user expectations, researchers start to rethink how science data archive systems can be improved for their usability. Challenges for the archive institutions include concerns of sustainability and development costs (e.g. tools that add data value are expensive to create and maintain, and most client software is not portable). Thus, how the archiving approach might be adapted and restructured is proposed in three dimensions:

(1) **Data**: What gets archived? File formats/ Reformatting/recasting archive products; on the fly product creation; and what happens to non archived data that are used to add value?
(2) **Access and Dissemination**: Tools for data access; data distribution mechanism;
(3) **Institution and Actors**: pre-mission planning/requirements; information required by user to use the data; interaction with users: forums/social media; interaction with providers.

隨著全世界的資料逐漸增加，資料的典藏，管理，與分享也相對的重要。只有資料的分析才能使資料有意義。也因此在會議第三天中，計算機科學相關議題接續之前的巨量資料與資料密集二大主題外，更將焦點放在實用案例的面向，主要包括: 資料保存與傳輸、開放資料環境中的工具、計算資訊基礎建設、以及 E 化健康管理等討論。同樣的，雲端運算、標準化與相互操作性、語意網技術等仍是重點，另外也在開放分享、易取用及易引用環境的概念架構議題上更進一步進行討論。

計算資訊基礎建設: 對於計算資訊基礎建設而言，在資料管理中所面臨的挑戰包括所謂的 4 V(數量 volume, 速度 velocity, 差異 variety, 確認 validation), 或是複雜性、譜系(lineage), 以及使用者介面等。這個議程主要分享中國與日本的資訊基礎建設。

自 2001 年起，中國科學院(CAS)的資訊基礎建設與 e 化科學計畫，發展整合 CAS 資料雲端服務。其中地裡空間資料雲端服務，主要整合開放資料資源(如後設資料的自動瀏覽、開放資料的資料抓取、原始自料的辨識、以及資料上傳雲端資料庫等工作)。此外自 2002 年展開的中國國家科學與技術基礎建設計畫，則使用資料金字塔模型，透過四個面向來描述資料 (原始資料、基礎資料屬性、外在描述、以及相關語意)。 這些計畫以研發與持續精進資料管理與分享為目標，視整合多領域的資料為重大任務。

不僅如此，"資料分享，但無法尋得" 或是 "資料尋得，但無法理解"等問題，則是日本國家情報通信研究所(NCIT)視為資訊基礎建設的重大挑戰。因此，為發展更完整提取資料庫所需的相關資訊，目前正研發的搜尋引擎，將時空、知識本體、以及可被引用等結構進行相關整合。

工具與方法: 為達成巨量資料探勘、資料學習、分析與傳輸等目的所發展的新工具，研究計畫提出針對巨量資料探勘目前學界研究的現況進行分析，同時並在理論與應用基礎上了解並分析其中的橫溝。會中包括現有資料工具的比較分析 (其中兩樣工具是運用開放源始碼的工具)。結果顯示，商業客製化工具呈現最佳效能，但開放源始碼工具則在主要時間的操作、以及計算資料位置上具有發展的潛力。

此外像地腦 (GeoBrain) 和全球農業旱災關測與監視系統 (GADMFS) 等實作案例，則是以 OGC, W3C 和 OASIS 等標準為基礎而發展的公用資料/服務環境，作為基礎架構。其特色是有利於資訊的易取用、以及分散式多源的空間資料整合。而另外一個新方法 –資料勘探 (Data Prospecting)—則結合傳統的資料蒐尋、與資料探勘技術，強調在資料庫的資料夾中，找出正確的資料子集。

如何在開放協同環境下進行整合資料處理與分析、傳輸、探勘、以及視覺化等議題，與前一日的研究討論相互呼應。開放原始碼工具像是 Talkoot 則再度被討論，而另外的工具應用則在資料探勘、互動式視覺化與影像分類、或是語意搜尋等相關工具，則以前日曾討論的 GLIDER 架構為整合環境，另以工具和方法層面和與會者進行不同面向切入的分析討論。

另外一項有關資料方法的討論，則在 E 化健康領域的專題討論中，繼續延伸日前針對標準化與相互操作性、系統架構以及知識模型等項目進行討論。臨床資訊模型推動(CIMI) 進行整合了美國的 HL7 標準、與歐洲的 ISO13606/openEHR Archetypes 標準。另外 XML 技術再次在 e-化健康領域中分享成功的實作案例。分別是日本中 Dolphin 計畫，以及巴西的工業研究案例。同時，也再次討論針對工業架構設計的開放式電子醫療紀錄原型(openEHR Archetypes)，以語意原型的物向導向模式為基礎，強調可重複被使用、以及重新組合元素的能力 (就向樂高積木的使用說明設計概念，定義了樂高積木的積木如何組成一個拖引機的結構配置說明)

保存與傳輸: 數位化時代前的資料可能存在 – 某處 -- 紙本、膠片、底片等處，但因無法線上取用，因此對數位化時代而言，他們就是所謂的資料 ”遺失”。也因此，所謂的資料救援(data rescue)以及保存等議題因應而起。其目的是在遺失資料的重新發現與保存工作上進行探究。此議題包括的資料種類多元，即使是 E1 議程中所討論的高能物理資料，亦包含在此研究議題中。相關的實驗性計畫陸續提出，針對資料具有遺失危險、以及被遺忘的資料庫等問題進行探究，同時亦包括達成相關目標如: 研究成果能有效的被分享、被引用、以及被發現。此日議程中，包括了電子書有效針對資訊傳播、與成本降低的最佳化進行研究案例分享。另外像是運用開放原始碼所完成的數位圖書館案例、或是另一項整合多種資料格式的運用的 3D 電子書案例，提供了能在任何地點與任何裝置上，取用資料格式的設計討論。

特別值得一提的是 Figshare。這是一個使用者能使他們的研究資料與成果，同時保存在一個可被引用、可被分享、以及可被發現的資料庫系統。此系統允許使用者上傳任何格式的資料，並能有效在瀏覽器上呈現。因此提供了目前學術發表模型中，尚無法完全達成的成果，也就是提供了一個能使得研究內容相關圖案、資料集、多媒體等資料與資訊能同時被發佈與分享的方法。

當學者了解到目前資料儲存歸檔的本質，已由資料的保存漸漸趨向-- 強調資料的易取用性和使用者期望等觀點時，他們提出了如何改進科學資料儲存歸檔使用性的再思考方向。對於資料儲存歸檔的機構而言，他們必須面臨永續性、發展成本 (如增加資料價值的工具，取得與維護的成本昂貴、大部分使用者端的工具不易移轉) 等挑戰。 因此研究針對資料儲存歸檔的方法的重新調整，提出了三個面向的考量建議:

(**1) 資料:** 儲存歸檔的資料涵蓋哪些範圍? 檔案格式 / 重新格式化/ 儲存歸檔重新配置，系統運作中產生的成品、用來增加資料價值但未被進行資料儲存歸檔的資料如何處置?
(2) 取用與傳輸: 資料取用的工具，以及資料散佈的機制;
(3) 機構與使用者: 任務前的規劃、資訊使用者的需求、與使用者的互動包括論壇/ 社群媒體等、與資訊提供者的互動。