謝清俊研究員

現職：

中央研究院資訊科學研究所研究員（已退休）

學歷：

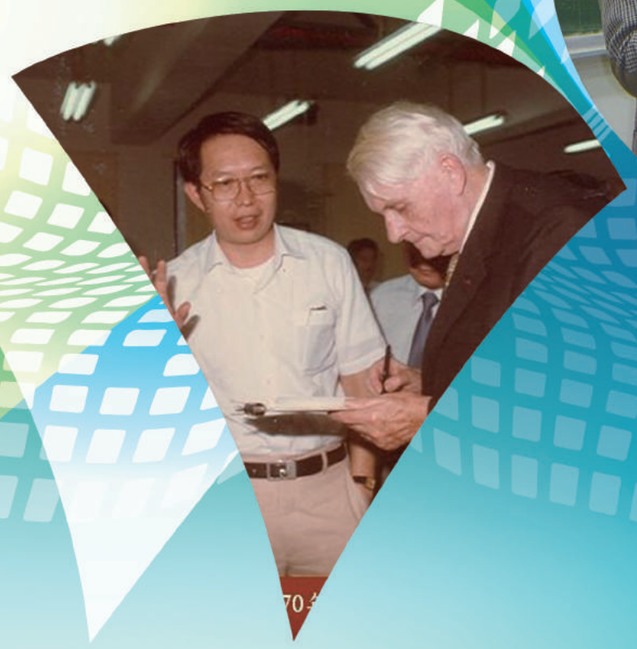
國立臺灣大學電機工程學系學士
國立交通大學電子研究所碩士 / 博士

經歷：

國立交通大學 控制系 / 控制與計算機系 / 計算機系 / 計算機工程系（創系）/
計算機工程研究所（創所）教授

國立臺灣工業技術學院（現為臺灣科技大學）電子系 / 六個研究所系主任 / 所長

中央研究院 資訊科學研究所研究員 / 兼任研究員 / 計算中心主任 /
語言學研究所兼任研究員



人文資訊學的開創、發展與數位典藏的實踐

人文資訊學先驅

深耕古籍典藏數位化　跨領域資料庫典範推手

受尊為本國人文資訊學的創始者，謝清俊以資訊科學本科出發，從基礎的整理中文字形、中文資訊編碼，創設古籍全文資料庫，推動國家文物典藏的數位工程，再到佛學義理的系統處理，謝研究員及其團隊，把人文學者理解的豐富文化資產和見解，彙整建構在電腦，形成有系統的知識本體，幫助人類文明走向嶄新境界。

「謝清俊研究員缺乏一長串的國際一流學術期刊論文，沒有傲人的國內外學術桂冠。但三、四十年來，臺灣人文資訊學的發展有今天領先世界的成果，缺乏謝研究員，這篇歷史是寫不成的。」

用這段文句，總結投身學術研究一輩子的學者，有些非典型，卻畫龍點睛，句句到位。這是出自史學研究者杜正勝院士，推薦謝研究員參加行政院傑出科技貢獻獎選拔所寫的文。

2017 年歲末，頒獎典禮前，主辦單位專刊編輯小組來到桃園八德拜訪謝清俊研究員。聽他娓娓道來諸多往事，讓我們一窺在學術研究生涯的哪個時點、怎樣的時空機緣，促成他從電腦科技的本位，跨接文史領域，終而發展出一套人文資訊理論學說，並廣泛應用於古籍文獻、佛經典藏的數位工程。

生於 1941 年，謝清俊老家湖南。幼時隨著父親的軍旅抗日、剿共而遷徙，1949 年底隨父遷臺。臺灣大學電機工程系畢業後，取得交通大學電子研究所計算機科學組碩士，於 32 歲通過交大電子研究所計算機科學組的國家博士學位。

謝清俊是位資訊工程學者，擅長開發跨領域的資訊系統，所學涵蓋資訊科學、社會資訊學與人文資訊學等。細數他的研究成果，主要涵蓋：中文資訊處理；中文圖書館自動化，中文資訊交換碼 (CCCII) 的設計與應用；二十五史全文資料庫的研究與開發，建構「漢字構形資料庫」解決「缺字問題」；參加世界標準組織 ISO9541 中日韓字型標準的訂定；並將 ISO8876 SGML 引進臺灣等。

博士論文時期 激盪人文科技火花

謝清俊 40 年來致力中文資訊交換、數位化與中文知識表達的建置與開發工作，是我國人文數位化工程的奠基者。

回首往昔，謝清俊娓娓說起一段往事，是他與史學家毛漢光在哈佛大學時期的「腦力激盪」。當時，謝清俊到麻省理工學院準備他的博士論文，毛先生時任哈佛大學客座教授，兩人住在同間民房，閒暇時持續談論起文史和電腦的關係，二者間的借力可能性。

這段腦力火花，直到謝清俊寫完論文，回到交通大學，才有初始的應用機會。1976 年，謝清俊獲得交通大學正教授資格，自此不再有為升等而發表論文的煩惱。

「當時回來之後（1971 年下半年），我跟交大圖書館館長林樹教授合作了一個中文電腦相關的專案。要知道，從清末到民國七十年間，中文字最初的統計都是傳教士的努力成果。那一次專案，我們算是首度用電腦技術把中文字的排序做了加權。」

然而，那時讓謝清俊有機會參與推動中文電腦計畫，是在蔣經國擔任行政院院長時期。當時，他受邀前去行政院就電腦的議題，向中央部會官員演講做說明，接著，答應李國鼎的邀約，替國家做相關的政策規劃。因為這個委任，必要避嫌，謝清俊也自覺與產業界保持距離，專注投入研究。

基礎功：催生中文字形產生系統

2001 年 7 月，謝清俊自中研院退休前夕，資訊所跟他進行了一次深度訪談，就回顧他參與中文資訊處理的基礎工作，以及中文字形碼在國際交換標準競賽中險勝的歷程。

所謂處理中文的輸入與輸出，就是要把中文寫的東西，用科學、數學的方式表達出來。一旦轉換成數學語言，電腦就能處理中文。然而，中文與外國語文的結構不同，讓中文字形交換碼變成問題。

外國人不懂中文字的結構，當他們開始處理中文編碼時，把中文字跟英文字母直接對應。這麼做問題來了。有些中文字無法處理，產生缺字現象。這種不精準甚至可視為某種文化歧視。

於是，謝清俊想承續先人在文字學的努力，找出中文構字的法則，用現代科學、數學的語言表達出來。這項工程相當浩大。最初，在交大執行時，受限電腦處理容量，不得不忽略一些細節。轉任到中研院後，有了運算資源，才重新整理一遍，補足過去忽略的細節，以忠實呈現中文字學的結構。

在 2001 年，謝清俊帶領的研發團隊整理出 1,200 個字根，是中文字形最基本的結構，字根與字根間透過一些規則，結合成文字。另外，該系統考慮到使用者的方便運用，把字根擴展約 4,000 個元件，任何人使用時只需做一個層次的分析，只要考慮一個組合運算，就能直觀地知道一個字是如何組成。

這個系統也用來處理缺字，是中文交換碼的下一代。現在的交換碼都是封閉集合，此系統卻是一套產生系統，因此可以處理的字數無限，是現行其他系統所不及。透過這套系統可做中文字的交換，附加在任何編碼系統都沒有問題，還能利用構字系統表達缺字。此系統還可延伸處理日本、韓國、越南的漢字，甚至中國各個朝代的文字，真正突破時空限制。

國際戰得勝：中文資訊交換碼列標準

謝清俊另一項重要貢獻是參與「中文資訊交換碼 (Chinese Character Code for Information Interchange, CCCII)」的制訂與推廣，並規劃與建立中央研究院圖書館自動化系統。

1979 年 11 月，美國為了處理東亞文字，想訂定一套標準碼，由國會圖書館委託史丹佛大學，由 John Haeger 負責的研究圖書館組織召集會議。謝清俊由國科會指派參加與會，卻發現當時全世界只有一套日本的 JIS 交換碼可處理漢字，「如果我們沒有一套中文資訊交換碼的話，可能日本的交換碼就會變成全部漢字的標準。我覺得這是個非常嚴肅的事。」

那時，謝清俊才開始收集資料，教育部也剛好發表一份 4,808 個常用字的字集，可以開始做中文資訊交換碼的工作，然而一切並未真正開動。但美國在軍事與其他用途必須用計算機處理中、日、韓文等東方語文的資料，非常急於訂立一套標準。打算在隔年三月，於華盛頓召開亞洲研究學會年會時，決定採用哪套編碼系統。

為了避免中文系統採用日本漢字的編碼方式，謝清俊在會議現場果斷地宣稱：「臺灣正在做中文編碼的工作。」一回國，就趕緊跟電機工程學會會長李國鼎報告，取得贊同，在業界募資三百萬元，成立「國字整理小組」。

「當時我對李先生說，外國人在設計 ISO 646(字元集資訊交換碼) 時，我們沒有參與，所以電腦沒有辦法處理中文。」謝清俊直言：「現在這個中文資訊交換碼的標準如果不制訂的話，我們會對不起以後的子孫。」

1979 年底，國字整理小組花了三個月的時間，做出 4,808 個字的字集編碼。當時投入共事的張仲陶教授過年也沒回家，在臺灣技術學院的計算中心全力跑資料。1980 年 3 月，謝清俊飛去華府，參加亞洲研究學會年會，「我上飛機時帶了幾本字集編碼，裝訂的膠都還沒有乾，拿在手上還是軟的。」

這番眾心協力的成果，在該場年會打了漂亮的勝仗。謝清俊回憶當時，面對美國圖書館界和國家標準局的多位編碼專家、語言專家，輪番考問了一上午，簡直比考博士論文還辛苦。但隨後午餐時分，他們向謝清俊恭喜：美國決定採用臺灣的系統，不用日本系統了。日本派了七、八個代表去，最後失望而歸。

中文資訊交換碼從 1979 年發展出來，經過 20 多年後，一直維護到 1998 年 (張仲陶教授過世後)，就不再做維護了。因為謝清俊認為階段任務已完成，而且缺字系統問世後，也能取而代之。

籌建中研院計算中心 擔綱文史數位化推手

1983 年，謝清俊離開臺灣技術學院 (現臺灣科技大學)，來到中研院資訊科學研究所。轉職的主要目的，是想做古籍資料庫。當年他在哈佛結識的毛漢光在歷史語言所，兩人說好要合作開動當年熱切討論的古籍數位化。

然而，轉任中研院後約三個月，謝清俊就被吳大猷院長找去做中研院成立計算中心的籌備工作。

當時中研院的計算環境很差，除了資訊研究所有些電腦外，就只有植物所有一台迷你電腦。吳院長曾語重心長地說，他雖然不懂電腦，也不需要用電腦做研究。但知道以後研究不用電腦的話，中研院的學術地位在世界上將遭遇極大的挑戰，為了年輕的研究者，他瞭解電腦對中研院的發展很重要。

謝清俊當時的想法是要盡全力幫忙文史研究者，「我認為，電腦直接給自然科學和生命科學的研究者就好了，他們自己會用，但文史方面的研究者必須要有人照顧他們。」

接了計算中心主任後，謝清俊也沒耽誤他要做古籍數位化的初衷，反而利用這個機會，把想做的二十五史的資料做出來。從那時一直到他卸任之前，中研院計算中心的資源幾乎有七成都在支持文史數位化。

「一直到現在，我仍覺得這個策略沒有錯。」謝清俊說：「因為電腦買了，對自然科學、生命科學的人來說，他們很容易就會使用，但是文史的部分，非有人帶不可。」

古籍全文資料庫：二十五史打先鋒

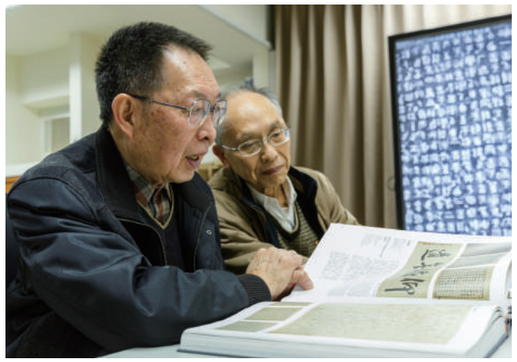
中研院 1985 年開始做二十五史全文資料庫，國外約在 1984 年底開始有一些全文資料庫出現。相比臺灣的資訊科學技術約落於國外五到十年的差距來說，中研院全文資料庫的起步早，相當有前瞻性。

然而，製作古籍全文資料庫的過程，遭遇了很多文史與資訊科學不同思考角度帶來的衝突。

例如，訂定二十五史全文資料庫的規格時，謝清俊的組員拿了一些國外的論文，指稱全文資料庫的檔案結構都是一頁文稿一個檔案，為什麼他堅持要一個段落做一個單位？為什麼一定要堅持保留二十五史原書的段落、行數跟字數？

「我跟他們講道理講不通，他們從電腦的技術來看，認為我在找麻煩，但是那是錯的。」謝清俊表示：「因為結構分成好多種，版面結構是一種，文章內容結構是另一種。我們必須知道哪些資訊在做全文資料庫時，必須保留下來。這個點到今天，都是一個好問題。」

謝清俊堅持之下制訂而出的規格，其實是非常領先全球的。例如，SGML(Standard Generalized Markup Language，通用標示語言語法)，於 1986 年正式發表，臺灣在 1985 年就訂出自己的標示語言，一直到今天，中研院計算中心還在使用。後來資料庫移到網路時，才將這套系統對應到 HTML 格式。



合作典範起效應
建立數位人文基礎

雖然在發展人文資訊學說的初始階段，謝清俊面臨了資訊科學與文史研究兩方人士，眼界只落於專業本位的問題。因為都只從自身角度出發，對於如何開展文獻與數位化的未知可能性，普遍欠缺想像力。

但謝、毛兩人在古籍數位計畫的合作，對後續兩個領域研究者的合作共識，也起了正向效果，被評為資訊學與人文學合作的好範例。

2010 年出版的《數位人文要義：尋找類型與軌跡》論文集就提到：「中研院《漢籍全文資料庫》的成功，很大的原因在於前身計劃《史籍自動化－食貨志輸入電腦》；在執行過程中，謝、毛兩位教授彼此的尊重和互動，讓數位與人文的結合成為可能，之後不管計畫如何擴充，這樣的合作模式提供了堅實、穩定的基礎。」

該論文集並提醒，「人文資訊學改變了文史哲學研究的習慣。數位資源有助於研究者推動跨領域的團隊研究。而大量數位資料資源的出現，或將有助於人文界開展議題，深化研究質量。然而，人文研究學者必須加強史料來源的掌握與研讀能力，特別是掌握與之結合的關鍵資料群，把資料還原到原有時空脈絡中，充分瞭解史料的內涵與當時社會關切的事物，以顯現具有時代意義的議題。」

全文資料庫百花齊放
資料精準好用是王道

事實上，做中文電腦絕非本國學者想做而已。可想見，兩岸不同的簡、繁體字，誰主導誰勝出的競爭不會少。非僅如此，還有第三方想角逐，就是日本的漢字體。三方各有企圖心，也都有自身的難題待克服。

中國大陸當時的問題，在於欠缺統一。很多人跳進來做，但每套文史資料庫之間欠缺通用的使用方法。拿基本的查詢來說，查紅樓夢有一套方式，查其他史書又是另一套；用現代說法就使用者介面 (UI) 的體驗很差。連簡單的查詢都欠缺一致性，更別說其他資料庫結構各吹各的調，資料根本無法互通應用。

謝清俊觀察，資訊理工者由技術角度出發，沒考慮使用者的用法，這是造成資料庫製作出來，難用而乏人問津的主因。

有個在梁實秋文集出現的故事，謝清俊拿來打了比方：「據說湖南的筷子比其他地方都長。伸手夾菜很方便，但要送進自己嘴巴就有點費事。其實長筷子是用來夾菜給同桌其他人，和樂融融吃頓好飯。組一個跨人文和資訊領域的團隊，就得在每個組員發揮各自專長之際，還真正讓大家的努力產生綜效。」

史哲典藏或文獻製作的資料庫，最不樂見的下場就是乏人問津。謝清俊為了確保資料庫製作出來，叫好叫座，花了一番心力。

早年，掃描技術還不成熟，二十五史全文都是人工打字，校對達五次，謝清俊要求錯字率的門檻是萬分之五。因為費心費工求精準，才成就了一套漂亮的二十五史資料庫；因為資料正確又好用，創造了好口碑，學人做研究引用無數。

提出跨行業資訊界說
嶄新傳播模式興起

當人類文明的紀錄和傳承，從紙本移轉到網路，全球強國積極推動數位圖書館、數位博物館或數位典藏等計畫，以強化其文化的功能與影響、競爭力。臺灣自然不能落於其後。

2001 年，謝清俊從資訊所退休，旋即被楊國樞院士借重，出任《國家數位典藏計畫》辦公室主任。二年期間，謝清俊傾畢生研究功力，大力向參與計畫的文物典藏機構，落實他倡導的「人文資訊學」。這個階段延續了謝清俊的長期努力，幫助人文學科研究者瞭解資訊科技在網路時代帶來的影響與意義，如何理解資訊做為媒介形式，並如何回應之。

要探討資訊科技對人文社會科學的影響，首要之務是對人文、社會、科學的範疇建立「資訊的界說」，要釐清資訊的概念、界說（定義），並據以說明資訊、傳播（溝通）和文化之間的關係。

謝清俊綜合了東方傳統哲學的系統觀、東西方對「存有」問題的看法，以及佛學的觀點等，發展出解決問題的基本方法論和研究方法，以探討資訊的生成與現起；並據此推演出對資訊的觀點，稱為「資訊的緣起觀」。透過這套緣起觀，界定一個跨行業、通用的資訊界說與嶄新的傳播模式，解決了 60 年來懸而未決的資訊界說問題。

謝清俊提出，資訊是所知表現在媒介的形式，所知是資訊的內容。在應用時，我們用的是所知，而非資訊。把資訊界定為「形式」有其正當性。一是資訊可被偵知、測知，並非抽象，是一種憑藉物質或物理現象呈現的形式。其次，電腦是只能直接處理數位形式的機器，它本身就是一套制式系統，凡是數位化資訊都能處理。因此，此界定完全符合電腦處理資訊的特質。

溝通是文明的肇始處，沒有溝通就沒有文明。謝清俊認為，尋取資料、閱讀資料、觀賞藝術品等，都是溝通的行為；推而廣之，一般的學術研究、知識

處理，都與溝通息息相關。沒有溝通，就沒有研究，也沒有知識的發現與積累。因此，借用傳播學的溝通性質，以界定資訊，不僅自然且有其必要。

他對資訊的定義借用了認知學、美學、傳播學等學科的精華，並集結資訊在運用時，對人文、社會和科技方面的考量。這樣的組合也符合科技濟世 (Technology Practice) 的要求。

謝清俊依據「資訊即所知表現在媒介上的形式」的界定，推導出資訊的四項基本性質，彼此還有交互影響而孳生出的新性質。這些導出的資訊性質，宛如基因，無時無地影響著學術研究的進行。因為它已提供嶄新的溝通與知識處理的行為，而溝通與知識處理正是所有學術研究的基礎。

國家數位典藏成果：歷史地圖和語言典藏

而根據這套人文資訊學說推展的《國家典藏數位化計畫》，是一個整合人文和科技的國家型計畫，也是少數以人文導向的科技計畫。參與的機構有故宮博物院、國家圖書館、國立歷史博物館、臺灣省文獻委員會、自然科學博物館、臺灣大學、和中央研究院等十餘個單位。計畫的目標除了典藏重要文物之外，普及精緻文化典藏的應用，亦是任務重點。

該計畫於 2002 年主要的數位化產出是：歷史地圖和語言典藏檔案，這是所有數位典藏共同參照和相融合的基礎。此外還有原住民、近代史料、動物、植物、礦物、考古、拓片、金石銅玉瓷陶等珍藏、古舊照片、書法、繪畫、善本書、清代宮庭檔案、以及臺灣早期的報紙雜誌等各方面的數位典藏。

這些產出一方面以公共資訊系統的方式，免費提供國人使用，如中小學教育、知識普及和社區文化發展等應用；另一方面公開在數位典藏市場問世，以平價提供國人精緻的文物數位典藏，以利於運用在學術研究、商業、產業、出版等加值應用。

佛經電子化另闢觸角 貢獻獎表彰肯定

謝清俊畢生不遺餘力倡導、推動古籍文獻電子化的重要性。文獻電子化最大的好處，就是讓大家充分共享。他把文獻比喻為資源或財富，「大家各自擁有的電子文獻拿出來共享的話，每個人都會越來越富有，國家的整體力量也會隨之增強。」

文獻要共享，必須能在網路流通，內容也要突破時空限制，任何人隨時隨地都能解讀文獻的內容。為此，就必須遵從一些文獻的標準來呈現文獻，而這些標準就是 ISO 8879 SGML 與它相關文件制定的事。

除了文史典籍外，謝清俊也把文獻電子化觸角延伸到佛學經典，曾帶領團隊發展出科文處理系統，是一套放諸四海通用的具體成果。他與佛學的接觸，來

自某認知心理學家的提點。他直言，即使看了幾年的西洋哲學書，對於「中文知識如何在電腦表達」仍有疑惑，一直都在尋找那把真正解答的鑰匙。

直到就教於當時一位加州大學的電腦科學和認知科學客座教授。對方提點：「你找錯方向了，要看佛經才對。因為西方哲理重視邏輯數理，而佛經於邏輯之外，還有人文的價值體系，是表達知識的寶庫。」

簡要而言，佛學的論辦基礎是因明學，不同於西方邏輯學偏重形式的演繹，不干涉思想內容，講求形式的結構和規則。因明學著重於論題的真實性及其原因，研究方法是立論、論證，是內容求真之學。

找到中文知識在電腦的表達關鍵後，謝清俊帶領的團隊就做過心經電子化。心經目前依據周止菴的蒐集，共有 15 個不同版本，有梵文譯漢文、梵文音譯，也有梵文譯藏文，再由藏文譯為漢文。

「若只把不同版本的佛經儲存在電腦，而不表明各版本之間的對應關係，是無意義的。」謝清俊點出問題：「因為電腦在這種情況下，很難做內容的比對，或任何進一步的處理。」

因此，開發團隊採用字串運作的指令，開發一套科文對照結構的系統，可以把某個版本改變為另一個版本，藉此觀察二個版本之間的關係。當然，這樣的系統可廣泛應用於處理古籍多版本。只要古籍有較詳盡的目錄，或在原目錄內，可找出主題、眉批與評語，就可構成一種類似科文的結構，適用於這套分析系統。

2003 年卸任國家數位典藏辦公室主任之後，謝清俊仍持續治學做研究，先後在法鼓佛學院和多所大學暨研究所授課或擔任講座，推廣他畢生鑽研的人文資訊學說，嘉惠無數學子。目前安居在桃園綠地環繞的鄉間社區，過著清幽的退休生活。此次傑出科技貢獻獎的表彰肯定，也在他畢生跨領域的學術成就列表中，增添一筆澹然的光彩。