DOI:10.16113/j.cnki.daxtx.2020.06.001

数字人文取径下的淡新档案重整与分析（上）

项　洁　洪一梅

（台湾大学资讯工程学系，台湾大学数位人文研究中心，台北 10617）

摘要：所谓数字人文取径，是将数字资源或数字工具的运用导入到一个人文学者完整的研究历程中，从问题意识开始，包含搜集与取得材料、整理与组织材料，到分析与观察，研究者可在此基础上进行诠释，进一步产出研究成果。本文便是从这样的方法论出发，陈述台湾大学数字人文研究中心在淡新档案的重整与系统化工作上，以数字人文研究为系统导向，融入三个重要的系统概念：学科元素的深化、研究功能的延伸、研究意识的连结；采取多重脉络的关联结构，将档案组织的层级从“案”细致化到“件”，逐件进行全文的人、时、地、物与客家相关词汇的标注，并在戴炎辉分类之外，针对客家研究的需要，建立客家事件主题架构。在这样的档案内容深化后，运用 DocuSky 的云端资料库模式，建置“淡新档案客家研究数字分析系统”（Danxin Archives System for Hakka Studies，DASH），提供系统使用者与材料间互动探索的多元功能，并以标准格式 DocuXML 让使用者完整汇出需要的文本材料及加值信息，以衔接个人的议题研究，启动研究者个人化的数字人文研究历程。

关键词：数字人文；淡新档案；客家研究

中图分类号：G270.7　　**收稿日期：**2020-07-03

作者简介：项洁，博士，特聘教授，研究方向为自动推论、数字典藏、数字人文，E-mail：jhsiang@ntu.edu.tw；**洪一梅，博士，博士后，研究方向为信息组织、数字典藏、数字人文。**

1　前言

档案是人类历史活动的记录，由机关团体、个人或家族，因组织业务功能或个人职能活动所产生或接收的资料，是事件发生的原始记录，在时间中增长并汇集，因而具备历时性价值而保存成为档案。档案的整理包含了“编排（arrangement）”与“描述（description）”，堪称档案学圭臬的《荷兰手册》中，分别就来源原则与全宗原则制订了档案编排与描述方法，其目的在于维护档案的背景脉络，并达成对资料实体与智能控制的信息组织。[1]

淡新档案典藏于台湾大学图书馆，是清乾隆五十四年（1789）[2]至光绪二十一年（1895）淡水厅、台北府及新竹县的衙门公文书，是现存清代台湾省、府、州、县厅署档案中最具规模且完整的，且为研究彼时台湾行政、司法、经济、社会、农业等极有价值之第一手资料，深具学术价值，弥足珍贵。淡新档案的整理工作可溯及日据时期台北帝国大学文政学部的典藏研究，往后于 1947—1953 年间，由台湾大学法学院戴炎辉教授开始正式整理这批档案，裱褙修补、分类编号，并且将之更名为淡新档案。戴炎辉教授以现代法学观点重整淡新档案，建立“淡新档案分类表”，先将档案分为行政、民事、刑事三大门，门之下依序再分为类、款、案、件；此举不仅将淡新档案分门别类，更使其便于利用。1986 年戴炎辉教授将淡新档案全数移交台大图书馆特藏组珍藏，并随即着手淡新档案的出版计划[3]；与此同时，台大图书馆特藏组亦逐步进行淡新档案电子化，以图书馆典藏管理的准则著录元数据（metadata），并搭配原件影像档

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

案建置数字典藏系统。[4]

数字人文中许多的概念发展与技术创新，使得档案的整理除保有原来维护档案背景脉络的数字化外，在档案整理、组织与系统化上都有了突破性的发展，尤其可以针对使用者不同的需求做不同的脉络分析与呈现。本文将从数字人文取径出发，论述台湾大学数字人文研究中心在淡新档案的重整与系统化，所采行的方法论、系统模式以及实践成效。

本文分为两部分（上下两篇）。第一部分（上篇）除了说明淡新档案的重整外，也包括连瑞枝和陈丽华两位教授针对淡新档案中的客家相关元素所做的“淡新档案客家事件主题表”。第二部分（下篇）则着重于DASH的系统设计与功能运用。

2 数字人文导向的系统方法论

近二十年来，随着数字科技的发展与变迁，“数字人文”成为人文研究环境中一个新兴且热门的术语。许多信息专家与人文学者相继投入数字人文中，跨越新领域的尝试为学术场域带来新的体验与冲击。此间，许多研究者不断探究何谓“数字人文”，纷纷由各自观点尝试定义数字人文，莫衷一是。譬如斯坦福大学在其人文中心网站上，陈述了他们对数字人文的看法：数字人文是电脑科学与人文学的结合，各式各样的电脑工具得以让人文学者进行过去所不能的大规模研究。数字人文不仅促进跨领域的协作，并开展了学科与方法论上的论述。[5]这样的界定，对实际进行研究的人文学者来说，仍然难以具体想象可能的数字人文研究。

在2018年台湾《数位典藏与数位人文》的期刊发刊词上，我们倡议应在数字典藏的丰硕成果上，开展数字人文研究的新视野。许多学研机构以大量人力全面地搜集数据并建置为数字典藏资料库，提供研究者使用。从资源提供者的角度来说，供给了丰富的研究材料，为研究者减省许多搜集整理的时间和气力；然而，对研究者而言，材料的搜集与资料库的设计并非为特定研究者量身打造，这样的帮助事实上尚未进入研究者真正的研究工作中。我们想象中的数字人文研究，应是将数字资源或工具的协作导入到一个人文学者整个的研究历程中，从问题意识开始，搜集与取得材料、整理与组织材料、检索的功能、到资料的分析与脉络观察，研究者可在此基础上进行诠释，进一步产出研究成果。将数字人文从抽象的学术定义，转为对人文研究历程，不论是方法上或工具上，数字协作的拓展与延伸。[6]我们在这里强调数字人文的本质就是跨领域合作的，而且这个合作不应是人文学者单方向的提出需求，由信息专家来满足，或是信息专家设计工具，由人文学者削足适履的去使用。而是一个对等的、共同学习、共同成长的有机体。我们用“协作”来表达这个观念。

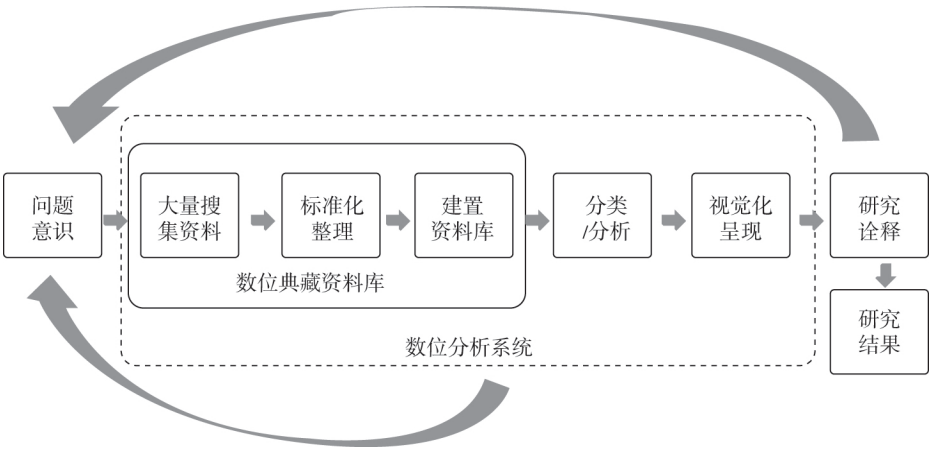

图1 人文研究历程中数字资料库的系统转型

因应数字人文时代的开展，数字系统建构的方法论也要随之改变，方能满足人文学者研究的需求。数字时代与以前最大的差别在于取得研究资源不再困难，人文研究者面对的不再是信息匮乏，而是难以驾驭的大量以数字形式存在的资料。在数字时代，高质量的数字研究素材来自各个机构（如图书馆、档案馆、博物馆、学术单位或业界）有系统建构的数字典藏库。这些典藏库建制的主要目的是通过完整收集和标准化整理将特定领域的资料有系统的数字化提供使用（见图1的“数位典藏资料库”部分）。它们在资料收集上以尽量完全为目标，功能上要考虑最大使用社群的检索需求。这种“最大公约数”的做法往往无

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

法照顾到个别研究者的特殊需要。从资料面来说，检索得到大量与特定研究议题不相关的资料对使用者而言造成太多资料上的杂音；从工具面来说，一个可以适应大多数使用者或素材的工具，对特定资料集或议题的处理上往往反而碍手碍脚。换句话说，研究者在使用一个资料库寻找资料时，往往是怀着问题意识而来，为了怕错失重要的资料，研究者宁愿从资料库中尽量收集可能相关的资料，然后依循问题意识梳理材料，通过分类与分析的手段，探索材料间的脉络，藉以获得线索乃至启发。在过去，数字典藏资料库在分类、分析与呈现的阶段几乎无能为力，研究者只能依赖传统的训练在取得材料后自力自强。因此，我们认为，数字系统除了将大量搜集来的资料做标准化整理并提供使用者取用外，应拓展系统功能，协助研究者对材料进行分类与分析，并通过视觉化图像，直观地呈现材料反映的现象，将系统可协作的范围延伸至分析与观察的阶段。这样的延伸对研究者而言不仅是崭新的体验，更能将问题意识与材料关联起来，在两者的互动中修正自己的研究，甚至带来新的研究启发。[7]但无论是元数据的描述或是内容的解析，仍是以群体对象的典型为设定，对个别学科或专门议题的研究者而言，总未能切中要害。事实上，多数的人文研究在实作面都是独特的、个别化的，而典型的元数据，对个别研究者而言，仅能协助资料检索，对于后续的分类分析与观察助益不大，也因此，当我们在进行系统建置的使用者调查时，使用者的学科背景、研究旨趣及议题倾向都是我们关注的项目。

2018年，台湾大学数字人文研究中心受台湾客家委员会客家文化发展中心（以下简称“客发中心”）的委托，建构一个以淡新档案为主针对客家研究需求的数字系统。系统建置目的是为提供客家研究者使用，协助他们利用淡新档案进行客家文化的相关研究。为了达到这个目的，我们以“数字人文导向”的概念来建置系统。这个策略分成如下几个方面：

其一，学科元素的深化：淡新档案的重整应走向学科化、议题化，在元数据的描述及内容的解析上，深化客家文化相关知识及萃取客家文化相关元素。

其二，研究功能的延伸：系统除了提供检索与浏览等基本功能外，要以《淡新档案》中元数据及标记（尤其是客家元素）为基础，提供各种后分类与人机间互动的资料分析，并且以直观的视觉化来辅助研究观察。

其三，研究意识的连结：一个以研究为目的的系统，应能连接至个别的研究议题与问题意识。因此，强调系统与使用者之间的互动操作、脉络的追踪，以及个人化服务，让研究者在使用过程中不仅可以修正原始的研究意向，亦可延伸或启发新的研究旨趣，让个人研究的衔接与推进成为可能。

3　多重脉络的关联结构

档案整理“以保持文件的历史关系和便于利用为原则”[8]，更细地说，档案组织的三大原则是立基于尊重档案来源的理念[9]：其一，尊重档案来源原则——维护档案的证据价值，并保存档案生命历程的各项记录。其二，尊重全宗原则——将看似个别存在的档案，纳入一定的社会活动范围内，成为一个不可分割的有机体，以特定的方式使其聚集，以记录其形成过程与联系。其三，尊重档案的原始顺序——强调档案来源机构为档案归档与保存所建立的原有次序，以保持文书系列案件之间的相互关系。

淡新档案在台北帝国大学（台湾大学前身）取得时，似乎已经没有按照清代地方衙门用“房”分类的全宗面貌。戴炎辉教授在1947年整理时，以“案”为单位，依据现代法律将档案整理成1163案（共19152件）[10]，分属行政、民事、刑事三大门（每门各有数类，每类又有数款）[11]。所以淡新档案在纸本的层次已经无法按照原生的脉络编排。戴炎辉教授的分类固然给予研究者许多的便利，但并不是没有检讨的空间[12]，但是对于实体档案而言，一旦分类固定了，就很难用别的方法重整。到了数字化时代，这个局限终于可以被打破。从管理上来说，档案仍是唯一的，但数字化档案可以呈现各种不同的脉络；从使用上来说，传统档案的查检依靠的是档案编号，或是典藏管理机关所提供的档案描述，这样的档案描述通常偏重于管理的目的，也较局限于档案来源机构的原编信息及详简程度，例如：编号、形制、层级关联、产生信息等档案客观信息。但数字化档案可以依据特定的使用对象与目的进行学术专业的编排、描述并建置系统，在查检上多了选择，除了档案的客观信息外，更可深入档案的内容与细节，甚至超脱原生脉络的局限。[13]也就是说，数字化档案的编排与描述可以在系统性的方法论上，不违背传统档案组织的理论原则，维系住档案的原生脉络，呈现出更多在内容细节中的关联线索，并提供更多的使用弹性与研究的可能性。

我们曾提出“多重脉络”的系统性方法论，它利用数字系统的特性，让档案同时具备原生脉络，或是典藏单位的档案编排脉络，甚至是学科观点的诠释脉

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

络，彼此间不混杂，却可以观察不同脉络间的关系，并可通过档案的内容细节去串起可能的脉络线索，让多重脉络彼此互动，提供研究者更多可延伸的观察视角。[14]我们在进行淡新档案重整与客家研究系统建置时，便是以“多重脉络”作为系统关联结构的方法论，主要的作法仍是建基在数字化档案的编排与描述上。

3.1 利用档案多元层次的特性，将淡新档案的处理层级细致化

在传统的档案管理作业中，囿于各种原因，对档案的描述往往停止在“案”的层级，不会对案中的每一“件”去做描述。但是因为数字化大大增加了利用检索取得资料的便利，所以在数字化的过程中，需要在“件”的层次处理，才能够充分利用数字化所能提供的检索、后分类和多重脉络的便利。

淡新档案经重新整编后，共有1162案，计19246件[15]。为了研究的分析需求，我们对每一“件”做了以下的处理：原本在台大图书馆的努力下，淡新档案已经被缮打成全文，也建了元数据，如标题、摘要、日期、相关人员与关键词等，在本计划中，我们对每一件做了更深入的元数据描述，如将原属“案”层次的元数据（如案号、“成案时间”与戴炎辉分类）或地理信息录入“件”层次的元数据中。此外在元数据中，也对客家相关主题做了处理。此外也针对每一件档案的全文内容，进行七种词汇的标记，分别为人名、地名、时间、客家身份、客家宗教文化、客家社会组织和客家特色名词。

3.2 以“件”为单元，描述详细且兼具学科专业的元数据

每一件档案的元数据描述，除包含编号、形制、层级关联、产生信息等档案客观信息外，也纳入研究需要的相关学科专业信息。每个描述栏位皆需有明确的定义及描述规范。以下特别说明与档案脉络相关的栏位：

3.2.1 戴炎辉分类（门/类/款）

淡新档案历经长时期的辗转留存与整理，其于清领时期的原生脉络早已无法完整考证，现存淡新档案的编排是由戴炎辉教授整件为案，以近代法律学的概念建立门、类、款三层分类架构，被叫作戴炎辉分类。[16]这个分类架构是传统的树状编排方式，每一件档案只会归于一案，而每一案也仅会有一组分类（门/类/款）。戴炎辉分类为使用者提供一个明确的切入观点，然而以近代法学观念诠释清代台湾衙门的运作自有其局限性[17]，但即便如此，戴炎辉分类仍是与淡新档案实体保存体系最为一致的编排架构，况且历经出版与资料库建置，研究者也早已习惯依此分类进行查找，因此，在新建系统中，每一件淡新档案的元数据除了有所属案号外，也录有戴炎辉分类。系统以树状结构呈现分类，通过这个结构，可以看出档案的戴炎辉分类脉络。

淡新档案中案的命名也充分反映了戴炎辉分类的架构。每一案由五个数目字组成：“第xyznn案”。其中x对应“门”的层级，y对应“类”的层级，z对应“款”的层级，而两位数的nn则代表在xyz结构中的第几案。譬如第16305案就是行政门（1）中军事类（6）中城工款（3）的第05案。

3.2.2 客家事件（L1/L2/L3）

既然建构新系统的目的是为了提供客家研究者使用并进行与淡新档案相关的客家文化研究，当然需要在淡新档案与客家之间的关系做比较深入的分析。淡新档案与客家文化关系非常密切。从时空信息来看，淡新档案涉及的清代台湾地方治理范围，正是台湾北部客家族群的主要生活空间，亦即现今的桃竹苗（桃园、新竹、苗栗）地区。从生活内涵来看，淡新档案潜藏很多客家的故事及客家先民生活的记录。从客家研究的学科基础来看，淡新档案涉及的客家生活面向有政治（如：乡保、董事）、经济（如：义仓、垦户、号记、土地纠纷）、社会（如：族群）、文化（如：民俗、饮食）、宗教信仰（如：尝、祀）、语文符号（特色词，如：伯公），正攸关客家研究的相关学科，包括政治学、经济学、社会学、人类学、历史学与语言学等。[18]

因此，若以客家研究的观点，淡新档案中的客家相关元素，尤其是人、事、时、地、物，成为研究者期待发掘并掌握的档案内涵。其中，我们认为“事”在档案内文中并无相应词汇或文句，使用者难以通过内文的检索寻得，再则，“事件”的名称往往是后人诠释后给定的（例如林爽文事件），而在事件发生时所产生的文件中，当然不会有后见之明的事件名称。要诠释“客家事件”必得由客家研究的专家学者针对档案的内容细细耙梳。参与系统建置的客家研究学者认为，这不仅是辨识“客家事件”的问题，事件所被赋予的名称必须是领域内大家的共识，这为“客家事件”的描述带来相当的困难度。几经评估及试行作业后，决定描述为“客家事件”的主题分类，不给定事件名称，而是将档案内涵中可能涉及的客家事件主题予以分类。如此一来，避免了给定名称的狭隘性，多重主题的做法也突破了单一线性编排的局限。每一件

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

档案可以涉及多个客家事件主题，同一案的档案也可以有不同的关联。客家事件主题也采三层的树状结构，每层的主题词至多以十个字描述，避免过于冗长。使用者通过“客家事件主题分类”来理解淡新档案，是本系统提供的另一专业视角的档案脉络。依主题分类动态聚合检得的档案，在未阅览档案内容之前，已呈现些许的脉络线索，并可通过其他元数据栏位的交互筛选，探索档案中的种种蛛丝马迹。

经过仔细地件件耙梳，我们发现淡新档案的 19246 件中有 6832 件与客家主题相关，这些文件分属于 371 案[19]。我们在这些文件的元数据的“客家事件主题”中著录它们所属的主题，而在“案”层级的客家主题则是那“案”中所有“件”的客家主题的联集。客家主题分类和戴炎辉分类最大的不同在于后者中每一案有一个唯一的分类，然而客家主题是对于案件/文件的形容，所以每案（或件）可以同时拥有数个主题。我们所设计的客家主题在第一层有七个主题，分别为亲属、宗教、族群关系、社会秩序、产业活动、社会组织与行政事务。在下节中我们提供了完整的“客家事件主题”表与每个主题下的案/件数。

以第 22510 案为例做说明。第 22510 案中共有 22 件，在戴炎辉分类里属于“民事/田房/争界”，如果看每一件的客家事件主题，则丰富得多。以下列举本案中的 3 件：

案号	第 22510 案	件号	22510_005
戴炎辉分类	民事/田房/争界	客家事件主题	族群关系/客原关系 社会组织/拓垦组织
		件号	22510_006
		客家事件主题	社会秩序/土地/侵占 产业活动/茶业 族群关系/客家内部关系 社会组织/拓垦组织
		件号	22510_007
		客家事件主题	社会秩序/土地/侵占 族群关系/客家内部关系 族群关系/客原关系

此三件同属一案，皆归于戴炎辉分类中的“民事/田房/争界”；第 22510_005 件，有二组客家事件主题分类词条：“族群关系/客原关系”“社会组织/拓垦组织”；第 22510_006 件则有四组：“社会秩序/土地/侵占”“产业活动/茶业”“族群关系/客家内部关系”“社会组织/拓垦组织”；第 22510_007 件，则有三组：“社会秩序/土地/侵占”“族群关系/客家内部关系”“族群关系/客原关系”，三件内容各自涉及多重的客家事件主题，当我们在“件”的层级做细致的描述，即可发掘出件与件之间的主题关联，也可区别件与件之间的主题差异。从主题的整体来看，大约可以猜到本案与茶园的土地纠纷有关。

3.3　客家事件主题表

以下所列的三层架构的淡新档案客家事件主题表系由本计划委托新竹交大连瑞枝教授与新竹清大陈丽华教授设计并整理。与客家相关的案与件的元数据标记亦由两位所带领的团队所完成。

淡新档案客家事件主题表

L1	L2	L3	L1/L2/L3	案数	件数
亲属	[－]	[－]	亲属/[－]/[－]	1	75
亲属	女性	[－]	亲属/女性/[－]	8	105
亲属	婚姻	[－]	亲属/婚姻/[－]	9	123
亲属	婚姻	招赘	亲属/婚姻/招赘	1	2
亲属	婚姻	赘婚	亲属/婚姻/赘婚	4	30

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

续前表

L1	L2	L3	L1/L2/L3	案数	件数
亲属	财产	[－]	亲属/财产/[－]	19	406
亲属	财产	土地	亲属/财产/土地	1	12
宗教	尝会	[－]	宗教/尝会/[－]	12	52
宗教	尝会	妈祖祀	宗教/尝会/妈祖祀	1	1
宗教	尝会	福德祀	宗教/尝会/福德祀	2	2
宗教	尝会	义民祀	宗教/尝会/义民祀	1	1
宗教	庙宇	[－]	宗教/庙宇/[－]	4	14
宗教	庙宇	三官庙	宗教/庙宇/三官庙	1	5
宗教	庙宇	天后宫	宗教/庙宇/天后宫	1	2
宗教	庙宇	有应公	宗教/庙宇/有应公	1	1
宗教	庙宇	福德庙	宗教/庙宇/福德庙	1	14
宗教	庙宇	福德祠	宗教/庙宇/福德祠	1	5
宗教	庙宇	义民宫	宗教/庙宇/义民宫	1	1
宗教	庙宇	义民爷	宗教/庙宇/义民爷	1	5
宗教	庙宇	褒忠亭	宗教/庙宇/褒忠亭	1	1
宗教	庙宇	广泽尊王	宗教/庙宇/广泽尊王	1	3
宗教	神明会	[－]	宗教/神明会/[－]	7	15
宗教	神明会	义民爷	宗教/神明会/义民爷	1	2
族群关系	[－]	[－]	族群关系/[－]/[－]	1	75
族群关系	女性	[－]	族群关系/女性/[－]	21	222
族群关系	客原关系	[－]	族群关系/客原关系/[－]	56	850
族群关系	客原关系	馘首	族群关系/客原关系/馘首	16	102
族群关系	客闽关系	[－]	族群关系/客闽关系/[－]	104	1934
族群关系	客闽关系	女性	族群关系/客闽关系/女性	3	51
族群关系	客闽原关系	[－]	族群关系/客闽原关系/ [－]	40	704
族群关系	客闽原关系	土地	族群关系/客闽原关系/土地	1	11
族群关系	客闽原关系	客语	族群关系/客闽原关系/客语	1	1
族群关系	客闽原关系	馘首	族群关系/客闽原关系/馘首	7	33
族群关系	客闽原关系	蔡牵事件	族群关系/客闽原关系/蔡牵事件	1	1
族群关系	客闽原关系	林爽文事件	族群关系/客闽原关系/林爽文事件	2	2
族群关系	客闽原关系	戴潮春事件	族群关系/客闽原关系/戴潮春事件	3	3
族群关系	客家内部关系	[－]	族群关系/客家内部关系/[－]	112	2102
族群关系	客家内部关系	女性	族群关系/客家内部关系/女性	1	11
族群关系	客家内部关系	吴阿来事件	族群关系/客家内部关系/吴阿来事件	1	2

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

续前表

L1	L2	L3	L1/L2/L3	案数	件数
社会秩序	[－]	[－]	社会秩序/[－]/[－]	2	83
社会秩序	人身	[－]	社会秩序/人身/[－]	4	54
社会秩序	人身	伤害	社会秩序/人身/伤害	54	1166
社会秩序	人身	勒索	社会秩序/人身/勒索	18	175
社会秩序	人身	人口贩卖	社会秩序/人身/人口贩卖	10	115
社会秩序	人身	男女关系	社会秩序/人身/男女关系	11	132
社会秩序	土地	[－]	社会秩序/土地/[－]	18	476
社会秩序	土地	侵占	社会秩序/土地/侵占	50	1197
社会秩序	财物	[－]	社会秩序/财物/[－]	37	838
社会秩序	财物	侵占	社会秩序/财物/侵占	26	408
社会秩序	财物	强盗	社会秩序/财物/强盗	51	1032
社会秩序	财物	窃盗	社会秩序/财物/窃盗	22	211
社会秩序	财物	赌博	社会秩序/财物/赌博	7	95
社会秩序	人口贩卖	[－]	社会秩序/人口贩卖/[－]	1	1
产业活动	林业	[－]	产业活动/林业/[－]	10	69
产业活动	樟脑	[－]	产业活动/樟脑/ [－]	7	49
产业活动	糖业	[－]	产业活动/糖业/[－]	11	76
产业活动	茶业	[－]	产业活动/茶业/[－]	7	264
社会组织	尝会	[－]	社会组织/尝会/[－]	13	120
社会组织	拓垦组织	[－]	社会组织/拓垦组织/[－]	87	1697
社会组织	拓垦组织	金广福	社会组织/拓垦组织/金广福	5	45
社会组织	拓垦组织	隘垦合约	社会组织/拓垦组织/隘垦合约	4	5
社会组织	水利组织	[－]	社会组织/水利组织/[－]	11	100
社会组织	义仓社仓	[－]	社会组织/义仓社仓/[－]	1	17
行政事务	人口	[－]	行政事务/人口/[－]	2	10
行政事务	女性	[－]	行政事务/女性/[－]	2	19
行政事务	文书	[－]	行政事务/文书/[－]	7	18
行政事务	水利	[－]	行政事务/水利/[－]	5	110
行政事务	矿业	[－]	行政事务/矿业/[－]	3	98
行政事务	米谷	[－]	行政事务/米谷/[－]	2	11
行政事务	米谷	稳定米价	行政事务/米谷/稳定米价	4	169
行政事务	隘务	[－]	行政事务/隘务/[－]	48	868
行政事务	隘务	裁隘	行政事务/隘务/裁隘	2	9
行政事务	隘务	隘租	行政事务/隘务/隘租	3	43
行政事务	隘务	隘粮	行政事务/隘务/隘粮	13	189
行政事务	盐务	[－]	行政事务/盐务/[－]	10	205

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

续前表

L1	L2	L3	L1/L2/L3	案数	件数
行政事务	公共事务	[一]	行政事务/公共事务/[一]	6	34
行政事务	公共事务	堤防	行政事务/公共事务/堤防	3	22
行政事务	公共事务	军需	行政事务/公共事务/军需	2	27
行政事务	公共事务	铁路	行政事务/公共事务/铁路	2	45
行政事务	公共事务	台北建城	行政事务/公共事务/台北建城	6	144
行政事务	土地税务	[一]	行政事务/土地税务/[一]	70	1624
行政事务	土地税务	官租	行政事务/土地税务/官租	5	26
行政事务	土地税务	屯租	行政事务/土地税务/屯租	13	142
行政事务	土地税务	水患	行政事务/土地税务/水患	13	196
行政事务	土地税务	社租	行政事务/土地税务/社租	5	62
行政事务	土地税务	隘租	行政事务/土地税务/隘租	1	44
行政事务	土地税务	土地清丈	行政事务/土地税务/土地清丈	1	31
行政事务	土地税务	土地清查	行政事务/土地税务/土地清查	18	222
行政事务	地方示禁	[一]	行政事务/地方示禁/[一]	14	74
行政事务	地方菁英	[一]	行政事务/地方菁英/[一]	46	789
行政事务	地方菁英	土目	行政事务/地方菁英/土目	1	1
行政事务	地方菁英	垦户	行政事务/地方菁英/垦户	6	44
行政事务	地方菁英	通事	行政事务/地方菁英/通事	6	37
行政事务	地方菁英	隘首	行政事务/地方菁英/隘首	8	141
行政事务	地方菁英	港口争夺	行政事务/地方菁英/港口争夺	1	1
行政事务	村落组织	[一]	行政事务/村落组织/[一]	25	341
行政事务	村落组织	联庄保甲	行政事务/村落组织/联庄保甲	4	8
行政事务	海上事务	[一]	行政事务/海上事务/[一]	1	8
行政事务	海上事务	船只救护	行政事务/海上事务/船只救护	1	14
行政事务	军事治安	[一]	行政事务/军事治安/[一]	10	78

4 在档案的电子全文内容中萃取并标记重要研究元素

提到数字化时，绝大多数人想到的是影像扫描与建置元数据，近年来将文字资料全文化也变成重要的趋势。但对于某些领域，如历史研究，单有全文是不够的。历史书写往往会提到时间、人物、地点，这些信息如果能够萃取出来并标准化，对于建立与解读文件间的关联性有莫大的助益。在台湾大学以前建好的包含淡新档案的资料库中（http：//thdl. ntu. edu. tw 和 http：//dtrap. lib. ntu. edu. tw），已经对每一件建立了“人名注解”“主要关系人及职称/身分”“关键词”“印记种类”等元数据，而这些也仍然保留在新系统的元数据中。在新系统中，我们进一步提取淡新档案中的重要标签词汇，并直接标记于档案全文里，并可通过标记的规则纳入权威控制与权威参照的功能。我们强调两者的不同在于元数据是对文件的形容，而标记是对个别文字的动作，当然比较繁琐，但用处也比较大。

为什么需要提取标签词汇，进行标记的目的又是什么呢？对使用者来说，要了解档案的内容，最直接的方式是阅读内文，但大量的档案需逐件逐字的阅读，除了旷日费时外，更容易忽视掉一些线索，譬如某个案件中的两个人名可能在另外一个案件中也共同出现。因为淡新档案中出现的人大多是市井小民，忽

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

略人名的关联十分容易发生。通过文内对人名词汇的标记，系统可以轻易地通过人名的共现来提醒使用者的关注。除了人名的标记外，地名、时间、物品也可以提供很多档案间的线索。譬如哪一个胥吏会特别专注某种类型的案件，或特定类型的抢劫案件的地理分布，等等。通过这些标签词汇，深埋在档案内文中的历史情境与现象将逐一浮现。过往只能通过编排才能追踪的档案脉络，藉由这些标签词汇的关联性，延伸出更多的探索路径。对系统来说，所有的档案内容文字就是一篇篇的组合符号，没有任何文意或语意，提供给使用者的方式，仅是一种文字的比对结果；但事实上，使用者想要的不仅是文字比对的结果，更期望此结果贴近他心中构思的脉络，例如：使用者查询“太平”，他真正想查的是地名“太平”。因此，系统必须在提供给使用者之前先行以使用者的观点为这些档案内容赋予意义。提取标签词汇并标记于档案内容中，成为预先标识、意义化档案内容的系统性方法。

除了典型的人、时、地、物的标签词汇外，我们更以客家文化研究的观点，提取“客家身分”“客家宗教文化”“客家社会组织”“客家特色词”等学科专业元素[20]。在全部19246件的淡新档案中，我们总共撷取了25953个人名词汇、13450个时间词汇、4971个地名词汇、3040个客家身份词汇、155个客家特色词、107个客家社会组织词汇、48个客家宗教与文化词汇，总标记数为478311。以第17330_007件为例（如图2），人名标记有34个，地名标记有38个，时间标记有2个，客家身份标记有10个，客家社会组织标记有7个，客家特色词有25个。当我们为每一件都进行详尽的标签词汇标记，每个标记都将串起档案间的词汇关联，成为使用者可用以追踪内容的细致线索。

ntul-od-th17330_007.txt

一七三三〇-七 典史為勘理詳覆事（新竹縣典史傅若金為詳覆勘定民番界址並姜胡氏樂捐買地歸還化番請獎各緣由申請新竹縣知縣方祖蔭案核轉詳）詳文（註一）光緒十二年十二月十四日到署典史為勘理詳覆事。本年十一月廿三日，奉憲臺札開：「轉奉爵撫憲劉 批據新竹墾戶金廣福、眾佃戶徐慶等稟請派勇撑要紮守，並番民各照舊界管業等情，奉批：『現據總土目朱打馬即謝寶順等以該墾戶金廣福曾經分界，將大南坑、柏色林一帶地面立石歸伊等經理給有墾批，乃復連年私佔，眾番等無有寸土耕種，稟乞追還業界等情，核與該墾戶等所稟互異。仰新竹縣即派典史傅若金，會同都司鄭有勤并墾戶黃南球，切實清查。如係民佔番地，縱不照數歸還，亦須按年照給番租，憑官定數，以後劃清界限，不得絲毫再佔。如該墾戶等恃強不遵，即由傅若金等稟請嚴辦。該墾戶等內佔番地，視為自有之產，外抗官糧，久執無賦之業，形同化外，鄭有勤奉委撫番，自係一秉至公，豈能縱容該墾戶等欺佔化番產業？本爵部院視民番如一體，該墾戶等自此以後須謹守法律，毋謂暗無天日，任意欺凌化番致蹈法網。並由縣傳諭沿山各墾戶知之。此稟仍繳』等因，批發下縣。奉此，並據監生陳汝湘等暨總土目朱打馬即謝寶順、土目夏矮績等稟，奉爵撫憲批飭前因。除分飭外，合行札飭。為此，札仰該典史即便遵照辦理」等因，並粘原抄壹紙。奉此，卑職遵即函會訂期履勘。適鄭都戎赴五指山一帶分布防勇，至本月初六日回防，卑職當邀墾戶黃南球前往會同，並傳集墾佃徐慶等、番目朱打馬等，於初七、八、九等日引導入山，循途履勘化番朱打馬等地段，東達內山，南抵黃龍章墾地，西聯民墾十四份、新藤坪，北至民墾十份坪等處為界。其西邊交界之所，均於峰頭立有界石，上刊「朱打馬眾番界址」等字。山下概循往來大路，民番剖分，路內向係番目夏矮績管理，外即新藤坪、十四份，現為民人周國山等結庄墾耕。當勘該墾民所闢園土內，有侵越路內番地者四處，數尚無多，合計不滿一甲。山高土瘠，收穫甚微，姑念耕種有年，且與眾番尚相親睦，免予歸還，即遵帥批酌給番租由官定數之諭，著周國山等年納番租四石，歸番目夏矮績領收。北邊亦經立有界石，甫入番界，即該民番等稟爭之九芎坪，再入七、八里許，即該番等稟爭之大坪、長坪，惟大坪已為金廣福紮隘，其間並有墾佃八家開種田土，細詢其故，緣大坪為後山生番出草必經之路，化番朱打馬等丁單莫敵，因請金廣福隘丁移入該處以壯聲援，厥後守禦漸堅，生番漸與通好，應用犒番牛、酒無處取資，遂將大坪、長坪、九芎坪等處賣歸金廣福招墾為業，藉得地價，以充犒賞之需，此番地之所由變歸民墾也。卑職擬仍遵照帥批酌抽番租，無如眾番交相索地，嘖嘖不休，若令退地還番，非特金廣福地價無歸，各墾佃生業頓失，倀倀何之。正疑難間，適候選縣丞姜紹基之母姜胡氏諭遣其子到地稟稱：「現蒙爵帥勤勞招撫保衛民生，吾儕踐土食毛，應深感激。盍請捐備原價，向金廣福買出大坪、長坪、九芎坪等處番地歸還該番，庶各化番從此衣食有餘，漸興禮義」等語。據此，卑職等經即質之各番甚為欣感，且各番以大坪、長坪附近番社便于耕種，願將離社較遠之九芎坪一處讓歸姜姓，以效報投之意。似此民番交讓，長治可期，遂如所請，當將姜胡氏捐價買出之長坪、大坪兩處番地歸還該番，所有大坪已墾田土，著該原墾佃民向番另議年納番租，其餘未墾處所或由該番自行墾耕，抑再招佃，均聽其便，擬俟轉詳奉准後，請予掣單該番目等收執，以杜附近墾民復相覬覦，致啟爭端。其九芎坪一處，即令姜姓管闢。雖係各番自願讓送，究在番界之內，亦應完給番租，以免混淆而符原界，但地面甚窄，將來墾田無多，已著姜紹基年完租錢八百文，付番目朱打馬領回分給，該民番等均尚悅服，願遵安業。惟大坪、長坪現雖多係荒埔，而地面平闊，水道暢通，詢之鄉耆，簽云：將來墾闢，年可收谷八、九百石，約可值價二千餘元，姜胡氏慨然捐價備買還番，固由帥恩所感，究亦義舉可嘉。查例載凡義舉至千兩以上者准予請獎，茲姜胡氏本係青年守節，尤復好義樂捐，可否仰懇憲臺俯援千兩以上義舉准予請獎例，詳請題旌，抑懇帥憲賞給匾額，以勵節義之處，出自恩施。合將卑職遵札會同勘理番界緣由，是否有當，理合備文，詳乞憲臺察核，俯賜轉詳，實為公便。須至詳者。右 詳欽加五品銜特授埔裏撫民分府署理新竹縣正堂方光緒拾貳年拾貳月拾參日（註二）署典史傅若金註一：滿漢文鈐記一枚，漢文曰：「臺北府新竹縣典史鈐記」，註二：滿漢文鈐記一枚，漢文曰：「臺北府 新竹縣典史鈐記」。附註：私記一枚。

图2　第17330_007件电子全文影像含标记

客家文化研究观点的元数据描述和客家元素的标签词汇使得我们可以更有效地将客家研究的概念在案件中交互串联，为使用者提供案件与词汇交互间多重脉络的研究线索。

注释与参考文献

［1］［9］薛理桂，王丽蕉.档案编排与描述：理论与实务［M］.台北：文华图书馆管理资讯股份有限公司，2010：35-41.

［2］戴炎辉将淡新档案的起始年定为嘉庆十七年（1812）（戴炎辉，1953），Buxbaum指出1812年为戴炎辉分类中最早的刑事案件，如果要连其他（行政、民事）也算在内，最早的案件应为乾隆五十四年（1789）（吴密察.清末台湾之“淡新档案”及其整理［J］.中国社会经济史研究，2017（2）：78-88.）。台湾大学图书馆特藏组网站上将淡新档案起始年设为乾隆四十一年（1776），其来源是一篇研讨会的论文（台湾文献资料合作发展研讨会议，“国立中央”图书馆台湾分馆.“国立”台湾大学图书馆特藏介绍［M］.台北：“国立”台湾大学图书馆特藏组，1991.），但这个时间如何而来则不可考。今从现存档案的起始成案年。

［3］台大图书馆淡新档案的出版计划于2010年结束，将内容完整出版为36册。

［4］［10］“国立”台湾大学图书馆.淡新档案简介［EB/OL］.［2019-05-30］.http://speccoll.lib.ntu.edu.tw/node/56＃2.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[5]Stanford University. Digital Humanities[EB/OL]. [2019-05-30]. https://shc.stanford.edu/digital-humanities.

[6]项洁.发刊词：从数位典藏到数位人文[J].数位典藏与数位人文，2018(1)：i-v.

[7]项洁，洪一梅.新世纪数位与人文的研究取径——"国立台湾大学数位人文研究中心"简介[J].汉学研究通讯，2019(150)：32-38.

[8][13][14] 项洁，翁稷安.多重脉络——数位档案之问题与挑战[M].项洁.数位人文要义：寻找类型与轨迹，台北：台湾大学出版中心，2012：25-60.

[11][16]戴炎辉.清代淡新档案整理序说[J].台北文物，1953(2).

[12][17]吴密察.清末台湾之"淡新档案"及其整理[J].中国社会经济史研究，2017(2)：78-88.

[15]在整理淡新档案的过程中，台大图书馆特藏组的同人发现有三件没有被戴炎辉归类到的文件。在本系统中我们将每件当成一案，并以0开头处理(戴炎辉的三门的代号分别是1，2，3)。这三件收录在台大图书馆所刊《淡新档案》第三十六册的最后。

[18]张维安.思索台湾客家研究[M].台北：远流出版事业股份有限公司，2015：18-25.

[19]淡新档案绝大多数的案件都发生在客民的生活区，所以严格来说几乎所有的案均与客家相关。在这里我们强调的是客家事件主题，所以如果客家因素没有在案件主题中被凸显，则没有被算在内。

[20]这项工作主要由新竹交通大学的连瑞枝教授与新竹清华大学的陈丽华教授与她们的团队负责，在此致谢。

Reorganizing and Analyzing Danxin Archives via Digital Humanities

HSIANG Jieh[1], HONG Yimei[2]

(Department of Computer Science and Information Engineering, Taiwan University;
Research Center for Digital Humanities, Taiwan University, Taibei 10617, China)

Abstract: Digital humanities (DH) is a research paradigm that employs digital resources and information technology to humanities research. From initiating research problem and collecting and organizing material, to analyzing and observing (visualizing) contexts, DH attempts to participate in every step of the research process. Based on this methodology, this paper describes how we built a context discovery system for Danxin Archives. We incorporated three important factors: in-depth analysis of research elements embedded in the texts, expansion of functionalities for context discovery, and the connection between the material at hand and research awareness. We organized and refined the archives from the "case" level to "document" level, and annotated the text by labeling the persons, locations, objects, as well as Hakka elements. In addition to the original classification of Professor Dai Xianhui, we also designed a new classification system according to Hakka affairs. Using DocuSky, a DH platform developed by the Research Center for DH at the Taiwan University, we built DASH, Danxin Archives System for Hakka Studies. DASH allows the user to do full-text query, analyze the resulting set using the two classification systems, observe and interact with the contexts provided by the system, and export the outcome in DocuXML format.

Keywords: Digital Humanities; Danxin Archives; Hakka studies

【责任编辑：梁继红】

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net