码库思体验报告

Weiwen Li

码库思体验报告

从 2019 年起，我有幸在 EnamelFC 项目中通过码库思平台处理清宫造办处档案。该项目的重要预期成果即是从巨量的清代造办处档案以及奏疏等文献中筛选有关珐琅、玻璃器物、中欧技术交流等方面的主题，提取其中涉及宫廷器物从订制到生产组织、工艺、行政管理等诸多方面的信息，并制作成互动式数据库。**码库思平台在数据库的制作过程中承担文献初处理的任务，即标注文本信息，为下一步可视化数据互动展示做准备**。

这是一项复合性极强的工作。它不仅挑战操作者的逻辑与心理，同时也挑战了码库思的性能边界。在着手进行这一工作之后不久，我便及时意识到了这一点。**这尤其要感谢我的学姐王华艳老师**，她作为 EnamelFC 项目中初涉码库思操作的第一人，自 2018 年起，以无法想象的心智付出，积累了巨量的操作经验。我和她的工作交接约在 2019 年夏季完成。**如果没有她的经验分享，我对码库思的摸索恐怕将要持续很长的时间**。

请允许我在此文中使用第一人称。当我开始写作这篇报告的时候，我的码库思实践仍在继续。我写作此文，并非恬于夸耀自己的苦劳，亦非以为自己已经探索出了某种学术性的方法与规律，而仅仅是希望能将操作中诸多零散的思考和稍纵即逝的感悟记录下来。

1. 码库思的牛刀小试

码库思本质上是一种语言学工具，它有三个主要性能：关键词标注识别、外链百科、信息格局可视化。其中关键词标注识别是码库思的核心功能，它的原理是对文本中反复出现的字句进行定义并自动识别。定义的内容呈现为一系列“标记”，可以根据使用者的具体需求进行设定。**周详地设定这些标记，并使其所覆盖的范围相互接合而不冲突，是码库思交给其使用者的隐藏任务**，如果这一步做得不到位，码库思将可能把文本处理转化成一场噩梦。

当我接手这一工作时，项目成员已经讨论确定了一组标记。我与王华艳老师交接工作时的第一步，就是将以下标记复制到了我的码库思账号中。事实证明，这组标记的设定是基本合理、经受得住考验的。但就如同音有音名和唱名的区别，标记也不能仅仅从字面上去理解。我们必须为它们设定详尽的语义学定义。

按钮名称	标记名称	在操作中的实际定义
價格	price	-
注釋	notes	-
其它主題	thematicindex	-
作品名	textTitle	-
職業	profession	-
物品	objects	不附带任何限定成分的

		器物名（器型信息），也可为器物的组成部分
物質	material	材质或质地
技術	technic	技术名，或直接作用于物体或材质的工艺性行动
機構	institutions	宫廷中的所有作坊、司局、部处，也包括本为地名但在文本中作为生产单位出现的情况
顏色	color	材质颜色或器物表面所敷施的颜色
行動	action	机构、人员之间发生的行政性行动，或作用于器物整体的间接动作，如“配、包裹、塞垫”，而不包括直接作用于物体或材质的工艺性行动（与技术相区别）
姓名	fullName	
别名	partialName	
時間	timePeriod	日期或时间段
地名	placeName	不能单独构成生产单位的各级地名
官名	officialTitle	

表格 1：作者实操开始时的既定标记设定，与作者在实操中为其厘定的进一步定义

值得肯定的是码库思平台对此有所准备，它为每个标记设定了“按钮名称”和“标记名称”两个属性。**“按钮名称”相当于唱名，它仅仅是显示在操作界面上的字样，而其完整定义则应在“标记名称”中体现**。从上表中可见，我所面对的这一组标记在当时并未被赋予明确定义，其“标记名称”基本为“按钮名称”的对应英文翻译。但即便如此，“标记名称”仍在某种程度上解释了“按钮名称”，如“作品名”标记的实际定义为“textTitle”，这就指出了标记字样中的“作品”实际上被限定于文学作品。又如标记“时间”的实际定义为“timePeriod”，这就规定了这一标记所适用的范围主要是时间段或持续时间，而不是时间点。这一定义使我在接下来的操作中没有手动标记任何节令名称，而实际上它们本在项目的关切范围内。

意识到标记的实际定义必须要比按钮上所显示的字样提供更明确的语义指导，我为这组标记中字面含义较为模糊的几个设定了更为详细的定义并交与团队参考，即上表右栏所示。当然这些定义仅仅是用来规范我的个人操作，而没有取代项目已经设定的“标记名称”。值得注意的是，**团队设定上述一组标记，是在处理康熙、雍正时期奏疏类档案时完成的，其标记必然带有此一批档案的信息格局**。而我所负责处理的档案则是乾隆六十年造办处行文档案，其信息格局难免与前不同，造成若干标记出现定义偏转，并影响其可操作性。这其中最为典型的可能是“价

格”标记。其定义仅以“price”一词代表，而实际上我所处理的档案中“价格”可能涵盖了工料、工食、赔补、赃罚等多种性质的经济数据，绝非“price”一词可以涵盖。这一标记与实际信息赋值的差异使我在接下来的操作中完全放弃了这一标记。

码库思的第二个功能，外链百科，在清宫造办处档案这个特殊的文本面前也遇到了新问题。当导入一段文献时，码库思可以自动进行初处理，利用其自带的百科数据库标注其中的历史人物、时间、官职、地名等信息。然而**造办处档案中所涉及的工匠、生产单元基层人员等大量姓名信息不见于正史，历史人物自动识别在此不仅无法体现效率，更会造成大量错误对应，将藉藉无名的匠役自动识别为正史中的重名者**，反而增加手动删改的工作量。在短暂的试验之后，我停用了历史人物识别功能。这当然有其代价，因为少部分文献中提到的正史人物也就因此而无法与外链百科中的资料建立联系。为此，团队曾经提出过一种替代方案，即利用码库思的导出标记功能，手动为人物列表赋值。但这是后话了。

如果说造办处档案的特殊性使得码库思的历史人物识别功能难以施展，其官职名和地名识别功能则有内在的不足。从实践中看，这两个自动识别项都**过于灵敏**。官职名识别往往将大量在后世已不用或者被文学化的上古官职名称加以标记，如“大官”、“供奉”、“后殿”、“三式”等字眼均被标注为官职，与文献实际语境有所脱离。地名识别也往往将大量少见、高古、低行政级别地名字眼加以标记，如“白玉”、“云龙”、“灵仙”等。但与历史人物识别功能相比，官职名和地名识别的总体效率仍然更高，所以这两项功能在操作中均得以保留。

码库思的另一项高阶功能是关键词标记，它可以让操作者通过编写简单的程序语言，有针对性地提取文献中的固定句式信息。这对于一些**句式极为稳定的技术性文献**有显著的展示信息格局功效。码库思平台在其使用场景举例中收录了多个此类操作经验，例如通过识别引号，直接提取现代句读版古文献中的引语（https://dh.chinese-empires.eu/forum/topic/16/tagging-quotations-with-markus）；又如提取古代地方志中的城垣施工信息（https://dh.chinese-empires.eu/forum/topic/32/exploring-data-on-the-construction-of-city-walls-with-regular-expressions）等。在操作初期，我曾经在两个方向上探索过这一功能的实践。首先是**试图在造办处各类行文中找到其可能的基本语句架构**，例如“某日，某作坊进某物，某太监持进”；“某日，某作坊奉某太监传旨，所进某物着修改/修理/毁为材料”；“某日，某作坊为造某物，领取某材料，某库某人发放讫”等。但文献所展示的多样性以及具体器物、行动、工艺上的多样性很快突破了这些构架分类，使得为每一种句式建立程序语言变得无利可图。理想状态下清晰可辨的信息格局仅能在档案所描述的事件极为简略的情况下才能出现，而此时依靠手动设定的标记完全可以展示这些格局，亦不需要特别的句式模板。

另一项探索则是**希望找到造办处档案中对于单体器物的描述范式**。在实操之初，我以为这种范式是很好定义的，它由“材质（M）”、“颜色（C）”、“技术（T）”三个关键标记或简单或复杂的叠加组成，最终总于标记“物品（O）”，即器型名词。在这种理想结构下呈现的器物名信息格局举例如下：

磁蓋碗 M+O
青綠寶月瓶 C+O
青花白地盤 [C+C]+O
磁青紙摺子 C+M+O
青綠流金雙耳花插 C+T+O
紅雕漆盒 C+T+M+O
銅胎法瑯蓮子壺 M+M+O
彩漆楠木三層盒 C+M+M+O
金胎洋紅法瑯盃盤 M+C+M+O
金胎綠色掐絲法瑯豆 M+C+T+M+O
銅鍍金雲福紫檀木座 [M+T+M]+M+O
銅胎黃地五彩法瑯螺螄盤 M+[C+C]+M+O
呆白攪紅玻璃雙耳盃 C+T+C+M+O
亮白玻璃長方法瑯片銅鍍金邊襯黃綾玻璃吊鏡 C+M+M+[M+T+M]+C+M+M+O
… …

然而这一理想化的格局亦难以被以程序语言总结，不仅仅因为其各种变化至于无穷，更是因为**项目的兴趣点要求将器物的一些组成部分或者附属物也当作有独立地位的物品来看待**，例如罩、盖、胎、座、表穰、表针等。在某些情况下，项目的主要兴趣甚至会首要集中在一件器物的局部构件上，例如刀匕类器物的珐琅靶、木质器物的珐琅片镶嵌等。而当我们把这些构成部分也标记为“物品”而与器物整体平级并列，整个器物命名规则将变得几乎无法描述，每一个器物都将可能是独一无二的，即便是同类器物，也可能因其不同构成部分在具体语境中的重要程度而呈现为不同表述。


於本年四月十一日 七品首領薩木哈將做得鑲法瑯片銅鍍金帶頭一件、鑲法瑯片金墻盒二件、銀墻盒二件、玳瑁盒一件交太監胡世傑、毛團呈進訖。
於本年四月十五日司庫劉山久、七品首領薩木哈將做得鑲法瑯片鑲金帶頭一件交太監毛團呈進訖。
於本年四月十六日司庫劉山久、七品首領薩木哈將做得鑲法瑯片銀盒三件、玳瑁盒一件交太監毛團呈進訖。
於本年四月十七日司庫劉山久、七品首領薩木哈將做得鑲法瑯銀盒三件交太監毛團呈進訖。
於本年四月十八日司庫劉山久、七品首領薩木哈將做得鑲法瑯片玳瑁盒一件交太監毛團呈進訖。
於本年四月二十日司庫劉山久、七品首領薩木哈將做得鑲法瑯片金盒一件交太監毛團呈進訖。
於本年四月二十二日司庫劉山久、七品首領薩木哈將做得鑲法瑯片玳瑁盒一件交太監毛團呈進訖。
於本年四月二十五日司庫劉山久、七品首領薩木哈將做得鑲法瑯片銀盒二件、玳瑁盒一件交太監毛團、胡世傑呈進訖。
於本年四月二十六日司庫劉山久、七品首領薩木哈將做得鑲法瑯片玳瑁盒二件交太監毛團、胡世傑呈進訖。

图 1：固定句式反复出现的理想状态信息格局

有趣的是，码库思的另一项自带功能，即时间信息识别，恰恰是建立在其程序语言性能之上的。这一功能可以识别如“年号-年份-月-日”结构的时间信息。但即便是最基本的时间表述，也可能在具体语境中出现变化，而使这一识别失效。一些特别的数字组合，例如在“日”后紧跟一个数字，也将使整个时间表述无法被

识别，例如“乾隆二年八月十二日**七**品首领…”、“乾隆二年八月十二日**九**江关监督…”等。这是码库思平台尚可优化之处，倒也说明了编写完善的程序语言以应对多变的文献表达绝非易事。

码库思实操开始后不久，我与王华艳老师赴柏林拜访了码库思制作团队的 Ho 先生。在他的指导下，我们基本掌握了在导入文献后复制标记、批量导入标记的操作。这对于处理巨量的造办处档案非常关键，不仅可以免于在每个文档中重新手动定义标记，更可以形成一组标记清单，并随着操作的推进而不断扩充。标记清单中附带有各个标记出现的文档名，这样又可以较快地定位若干标记，并在必要时进行删改。这些操作很快构成了之后文献处理的标准流程：
一，通过复制粘贴导入已经经过校对的文献
二，利用码库思自带的时间、地名、官职名自动识别系统进行自动标记
三，将已经定义的标记项（表 1）通过“管理标记”复制到文档处理界面
四，将已经收集的标记清单通过“批量导入”套用在文档中
五，手动操作，根据具体语境，调整、删改自动套用的标记，形成成品，

至此，除了基于程序语言的“关键词识别”外，码库思平台的各个基本功能都在 EnamelFC 项目中得到了较为充分的发挥。

2. 沉浮于语境汪洋

码库思不是自动套用标记的过程。码库思是一场考验逻辑与食指肌肉的手动操作。码库思操作终最精彩、最摄人心魄、最让人抓狂的环节，就是在手动操作中**逐一判断语境并决定标记的去留**。《Tagging quotations with MARKUS》的作者指出，“While the identification is not always spot on – you still need to use your brain – the identification errors can be easily spotted and manually corrected.”这句冷静的话放在 EnamelFC 项目对码库思的应用中也是有效的，只不过它背后涌动的逻辑与心智暗流，只有真正动手，在码库思的平台上度过一天又一天的实操者才能感受。

当操作者把已经积累起的标记清单套用在新导入的文档中时，文档将会呈现出一种因标记交叠横陈而造成的花花绿绿状。

图 2：批量导入标记清单后文档操作界面所形成的效果

在我最初的认知中，这种标记的交叠，尤其是因多重交叠而形成的方括号 [ ]，主要是源自一些字词的**语义多价**。语义多价是码库思难以处理的一种情况，即文献中的一个词同时符合两个标签的条件。**在全手动操作中，当一个词被标记赋值后，它将无法再被另一个标记赋值。而在直接批量套用一组标记清单时，一个词则可以承载多个标记，但操作者无法选择保留一个标记而删除另一个**。他必须全部将其删除，并重新标记。出于一种当时尚在萌发阶段的**逻辑洁癖**，我认定这种方括号是必须清理的。负责数据库下一步加工的信息工程师 Philippe Pons 先生则告诉我说，这种**标记堆叠**未必要清理干净，它们对于码库思操作者来说可能比较碍目，但是对于他的下一步操作来说，是有一定意义的，因为这种堆叠可能显示了某种信息层级，例如“亮绿”首先是一种“绿”，而“香几”则肯定是一种“几”。

但我很快发现，事情没有这么简单，**方括号[ ]和需要被清理的标记之间没有必然联系**。有若干标记重叠的确如 Pons 先生所说，是可以保留的，但也存在着大量没有构成重叠而必须清理的标记。这里面蕴藏的无数可能，是任何一位码库思操作者的语言学天堂和噩梦。当我操作码库思几个月后，我开始留意各种必须清理的标记与标记重叠的语义场，并试图做出总结。由于我的工作仍在继续，这场总结可能远远没有完成。但我还是将它尽可能完整地展示在本文中。

一，专有名词犯各类标记词

这是标记讹错最典型的原因之一，也是我向别人介绍我的码库思实践体验时最常举的例子。有的词既可以是颜色也可以是姓氏，例如“白”。有的词既可以是材

质也可以是姓氏，例如“金”。当我看到批量套用的标记将某位白世秀的姓氏全部标记为颜色的时候，我的码库思体验就开始了。而我所处理的档案中最看不得我手闲的历史人物大概是一位**白玉凤**，他名字的三个字分别是颜色、材质和物品。幸而他出场不多。有的词则既可以是器物也可以是限定成分，例如“圆明园”中的“圆”同时也是物品（器型）。这样的标记重叠显然没有任何意义，只可能被全部清除。在柏林，我们与 Ho 先生讨论过在码库思平台中添加筛选功能的可能，即“检索标注所有 白 而排除 白世秀 的组合”。Ho 先生表示，这在技术上是可以实现的。如果有一天这一功能真的得以出现在码库思平台上，那么这可以被看作是 EnamelFC 项目的小小贡献。

二，名词犯量词

这类讹错就需要更仔细地观察。**一些器物名同时也是惯用量词**，例如“扇”可做“门”的量词；“座”、“尊”可做相当一部分大型器物和造像的量词；“对（对联）”可做两两出现的事物的量词。**这些量词会被识别为器物，因而需要清理**。还有些情况是名词量词同形，如“法琅片二片”，其中第一个“片”为名词而得保留标记（物品），而第二个“片”则为量词而需要清理。

三，有物量词和无物量词

任何量词都不在项目的关注范围内。但**有些量词有实际对应的物品**，例如“徽墨成匣”等。在实操中，此类有物量词的处理倾向于展示物品的存储形态，而非展示数量。“成盒”中的“盒”可能被当作物品标记，而“十盒”中的“盒”则被看作纯粹的量词。

四，题材犯材质词

“竹”、“锦”、“纸”、“角”等**符合材质（物质）标记的词并不一定就是材质**，它们还可以指代对这些材质的模仿、其它相关题材、或仅仅是同形多义，例如“竹式”、“银镀金竹子”非竹；“红花锦地”、“合锦”非锦；“纸槌瓶”无纸；“入角”、“八角”并非动物角等情况。

五，名词犯材质词

有些物品名词也可能指代材质。此类词汇总量不大，但有一定迷惑性。常见的如“蜡”作为物品与材质“蜡”同形；“银”作为经济概念与材质“银”同形；“牙子”中的“牙”与材质“牙”同形。有时**名词中的限定成分也可能与材质词冲犯**，最典型的例子有如“象牙牙签”，其中第一个“牙”为材质词，适用“物质”标记，而第二个“牙”则是“牙签”的名词作定语限定成分，不适用“物质”标记。此外例如“镇纸”中并无纸，蜡扦中也不包含蜡，这样的情况都需要操作者在花花绿绿的汪洋中甄别。

六，限定成分犯名词

有些器物题材同时也可以是物品名，例如“如意”、“花”、“龙”；而有些物品名的限定成分本身也构成名词，“墨床”既无墨也不是床。这些情况构成了一类较为可观的需要清理的情境，。其中有一些需要仔细检查，例如“钉”是物品，而“乳钉昭文带”中的“钉”则仅仅是玉器上的点状凸起。有时材质等标记的限定成分也可能与作为物品的名词相冲犯，例如“象牙”中的“象”字，与作为陈设

的“象（动物）”相冲突。

七，实指颜色与非实指颜色

这是一种在实操中不多见但很有趣的语言学现象。颜色词在文献中一般是较为明确的，但**并非所有颜色词都是实指颜色**。其中最为典型的是“紫檀木”。当标记清单被批量套用至一个文档时，所有被标记为物质的“紫檀木”的“紫”字都会嵌套一个颜色标记。但紫檀木的紫色，绝非“紫”在日常生活经验中所指代的可见光波长范围，而近乎一种深沉的黑红色。这里的“紫”便不应被单独标记为颜色。而与之相对，“紫漆”中的“紫”则是真正有效的颜色信息。有些类似情况还要更微妙，例如“黄杨木”中的“黄”。黄杨木的颜色确实可称为黄，但黄杨木并非唯一的黄色调木材。故而这里的“黄”亦不必作为颜色信息看待。这样的情况还有如“红铜”、“黄飞金”等。另有**一些颜色词是材质词借用**，例如“翡翠”、“棕”，在文献的绝大部分均作颜色词而非材质，因而需要格外的小心判断。此外，有**一些颜色词的并列出现**也需要谨慎对待。例如，“红绿”用来形容瓷器，是两种颜色，需要分开标记；而“青绿”用来形容青铜器，则是一种颜色，需要二字整体标记。

八，名词多价

有一些名词，在某些语境下可能构成物品，而在另一些语境下则不构成物品，或至少不构成项目兴趣范围内的物品，例如“笔”是物品，而“御笔”中的“笔”则仅仅指皇帝翰墨而非书写工具。“字”作为书写结果并非物品，但当其被制作成各种材质的浮雕状字样并嵌安在匾额上的时候，它就成为了物品，如“铜字”、“云母字”等。这类情况还有一种特别的情况，出现在“样”字上。“样”可以是一种物品，即某器物的设计模型或图纸；它也可以构成其它语义，例如“花样”、“样式”、“一样”、“照样”等。**在实操速读中，对“样”字的标记可能是最费思量的任务之一**，因为我不仅需要判断它在语境中是否指代器物模型或图纸，还需要判断其是否作为模型或图纸的代词，例如在“照样准做”中，如果前文确实提到了一件“样”，那么这里的“照样”中的“样”就确有实指，因而应该被看作物品，而如果前文仅仅是以某个实际器物或者“样式”作参考，并无实际的“样”，那么“照样”中的“样”就非实指一物，从而不该被认为是物品。我必须承认的是，我不能保证在操作中我的每个判断都是正确的。

除了这些需要注意的纯粹语言学场景之外，码库思平台本身的操作受限于行文逻辑，而无法进行语义赋值的缺陷，又可能造成部分彻底无法处理的情况。例如“金星玻璃”是一种特殊材质，“蓝金星玻璃”则是由颜色、物质两个信息并列而形成的意群。但在实际文献中，这一意群还可能呈现为“金星蓝玻璃”的异体字样。在语义赋值的角度看，“金星蓝玻璃”和“蓝金星玻璃”是等价的，但在实际码库思操作中，前者因为行文字眼截断而无法被标记体现。

如上所述，在批量套用标记后所产生的泛滥式花花绿绿中，有至少一半都是错误的或者没有意义的。而将花花绿绿的局面加工为逻辑自洽的信息格局的过程就因此被我们称作**“打扫”**。这是本项目中码库思实操的最大工作量集中点。在操作初期，我采取了保留部分标记重叠的做法。**有些标记泛滥或重叠是逻辑谬误，必须去掉**，例如“牌楼”中的“牌”被标记为物品、人名“福海”被标注为地名等。

而**有一些标记重叠则似乎可以容忍**，例如机构“铸炉处”的“炉”被单独标记为物品；颜色“天青”中的“青”被嵌套式识别为颜色。这种**“逻辑宽容”**的一个非常有代表性的例子是文献中经常出现的出处注释“瓷器档案”字样。当标记清单被批量套用时，“瓷器档案”中的“瓷”会被标记为物质，而“案”则会被标记为物品。在这里，前者在逻辑上是可以接受的，保留之也似乎无伤大雅，而后者则构成语境讹错，是必须清除的。

我曾经探索过这种“逻辑宽容”的适当程度，这一方面是为了展示 Pons 先生所说的“有意义的重叠”，另一方面也是为了**减少个人工作量**。例如当我回看我的初期操作成果时，我发现我在有一段时间中曾经展现了一种**野兽派般的逻辑粗粝**，例如我甚至曾经允许“催长”和“副催长”；“玉作”和“金玉作”这两组标记相互重叠，这也就意味着未来数据库的使用者在检索“催长”和“玉作”时，将同时看到所有“副催长”和“金玉作”的信息。至于上文中的“瓷器档案”，我曾出于“逻辑宽容”而保留所有“瓷”字上的物质标记，那么当未来某个瓷器研究领域的使用者检索“瓷”字时，他会看到成百上千个检索结果是“瓷器档案”这个文献注释。他会因此而抓狂吗？当我意识到这可能是一个问题时，已经是许多个月之后的事情了。

现在我可以说，围绕着 EnamelFC 项目的档案阅读者有三种角色。档案的拣选录入者和录入校对者是档案的**“精读者”**（他们是故宫博物院的杨老师和北京社科院的关老师）。未来数据库建立后的使用检索用户则可称作**“点读者”**。而夹在两者之间，码库思操作者则可称作**“速读者”**。他们一方面没有时间对档案全文进行精读，但其操作又必须要求其掌握各个标记所处的具体语境，以应对上述场景，判断其去留。这些小小的语言学乐趣仅仅是一个开始，当他们构成各种更为复杂的局面时，操作者还要面对更多挑战。

3. 反复摇摆的平衡木

至此我们还没有讲到码库思实操中的另一个关键因素，即预判未来数据库成品**使用者的需求**。他们可能需要检索哪些信息，又希望这些信息以何种形态呈现呢？我逐渐意识到，考虑这些问题不仅仅是急人之所急，也是在给自己铺路。因为这将直接影响到码库思的操作策略。

EnamelFC 项目所展望的数据库，是一种极为详尽、包纳一切信息点的产品。但有些信息点是绝难通过码库思操作涵盖的，例如“风格”。器物的风格是器物的重要属性，但这些属性所可能呈现的语言表述千变万化，其中大部分无法被完善地标记，一方面是厘定这些风格是数据库使用者见仁见智的主观判断，另一方面，器物描述中的风格信息载体往往与其它信息相互杂糅，例如“白地青花”、“西洋珐琅烟壶”等，与颜色、材质等相互交叠。在我和王华艳老师共同工作期间，项目组中曾讨论过一种方案，即将每个器物名作为一个完整单元标注，而不单独标注展示其中的颜色、材质等信息。这样做固然可以避免大范围应用标记而造成的讹错，但码库思的标签识别优势也将消失殆尽。因为每个器物名都是独一无二的，**如果标记清单中的关键词独特性过强，将导致文献处理变成一场无穷无尽的手**

动圈点，而这样的操作其实完全可以在 word 文档中进行。

在实操过程中，我感受到两个**极端思路**的存在。第一种极端思路即如上所说，是**标记尽可能复杂的意群（复合策略）**。由于越复杂的意群独特性越强，这将使得语境判断的工作量下降，文献信息格局趋于明晰。但这也将带来手动操作工作量的大增。另一个极端思路则是**标记尽可能小的语素**，也就是单个文字，而抛弃掉任何可能的组合或者限定成分（**极简策略**）。由于单个文字标记相互重叠的几率较低，这样可以避免相当一部分需要手动清理的情况，而直接形成一种马赛克般的极简信息格局。

在实操的前期，我的操作策略是从第一种策略向第二种策略滑动。这一趋势的背景是随着标记清单的逐渐扩充，一些少见的简化、缩略式用语开始出现，例如“珐琅”在某些语境下会被简称为“珐”或者“琅”，而一旦这类缩略用语被标记进入清单，在批量套用时就将导致每个“珐琅”字样上被嵌套两个缩略式单字标记，而形成一组三标记重叠，如“[珐][琅]”状，从而需要手动清理。这样的清理并不能算特别费时，但的确曾让我在一段时间中认为标记应该趋向于单字，从而避免这种**长短嵌套**。如今调整完善那段时间的作品时，我还能找到当时这种操作策略的痕迹，例如各类屏风我当时均仅标记“屏”字，而没有将插屏、挂屏、屏风、桌屏等各类屏风分别标记。这样做也并不能真的避免什么混淆，更多是一种逻辑倾向。更有甚者，我还将所有作为物品存在的“狮子”均仅标注“狮”字，大部分“X 子”类物品都只标注前一个字；“海螺”则仅标记“螺”字，以避免与作为器物的“海”发生标记重叠。

但到了后期，我开始发现无论是极简策略还是符合策略，对于未来的数据库使用者来说不一定就是友好的，因为我**无法预判使用者的需求**。在未来的数据库成品中，使用者对于信息的检索将在很大程度上基于我所厘定的标记。如果我仅标记“屏”而不标记“插屏”、“挂屏”、“屏风”，可能导致部分希望检索某种特定品类的使用者的体验下降。当然，如果我把每一种屏都单独标记，也有可能造成希望快速统计文献中“屏”字数量的操作者多付出劳动。目前我还并不清楚未来的使用者能否跳过我所设定的标记来检索单字。**实际上，在极简策略与复合策略的背后所闪现的，是未来的使用者更多以词的逻辑来看待造办处档案，还是以物的逻辑来看待造办处档案。**在词的逻辑上，当然以单字标记最为理想，此时使用者甚至可以统计某一年的档案中提到了多少次某种颜色，而忽略其背后的物品。在物的逻辑上，则以尽可能完整展现一个物品的名称最为理想。项目组曾经多次设想未来的数据库使用图景，例如能否以检索某种器型来统计某年间一共**出现**了多少件此类器物。这当然很快被论证是不可能的，因为我们至多可以统计这种器型被**提及**了多少次，至于它们是不是真的存在过或者被制造出来，则需要更为详细的语境判断。例如“座”这一常见附属物，其大量出现都是在“某物无座”的语境下，而这还仅仅是最简单的否定式语境。换句话说，把文献做成数据库也无法免除未来的使用者通读文献的必要性。**数据库的制作者无法预判未来使用者的需求**，他所能做的仅仅是把握着码库思，在两种策略之间徘徊，而不过于偏向任何一侧。

所以到了现阶段，也就是我开始系统性地检查、完善我的早期工作成果的阶段，

我对当时的一些极简逻辑做了纠偏。这其中有一些是纠正错误，例如我在工作初期曾将“全带”、“傍带”、“带钩”等均仅标记“带”字，导致带和带饰不分。此外，如“多宝格”、“博古格”等带限定成分的物品名，也基本予以恢复，与简单的“格”相区别。原本被分开当作地名和物质标记的“西洋珐琅”和“广珐琅”也逐渐被合并而为一个词。一些在最初被容忍的标记重叠也被清理，比如“花囊”作为珐琅器物，清理了原本嵌套于其上的“囊”。这一工作还将持续，我**试图在两个极端之间找到合适的平衡点**。但也许终究是找不到的，毕竟我无法预测未来使用者的需求。

这样的策略调整当然也就意味着，我工作初期的成果所基于的逻辑，与后期所基于的逻辑之间有相当差异，乃至间杂以讹错。当我把阶段性的成果导出并交给团队后，团队向我反馈说，一些关键标记在各段文档中时有时无。究其原因，是我在操作中经常调整标记清单，对一些比较容易造成讹误或者标记重叠的标记进行清理更新。例如文献中常有以数字为名的历史人物，如“六十”、“六十五”等。这些数字姓名标记如果被批量应用在数字信息较多的文档段落中，将会造成大量错标；而有时其较为隐蔽，如一段文档仅造成一次错标，则又往往难于察觉，逃过清理。所以每隔一段时间，当这些标记词开始较少出现时，我就会对标记清单进行更新，设立一套**“近期标记”**。在这一过程中，可能有些标记词由于在近期工作中没有出现而被漏失，导致新处理的文档丢失部分关键标记词。归根结底，码库思操作就是与错误为伍，与错误做朋友，把对它们的修正放到更长远的时间上去考虑。否则不仅进退失据，操作者也会自身难保。

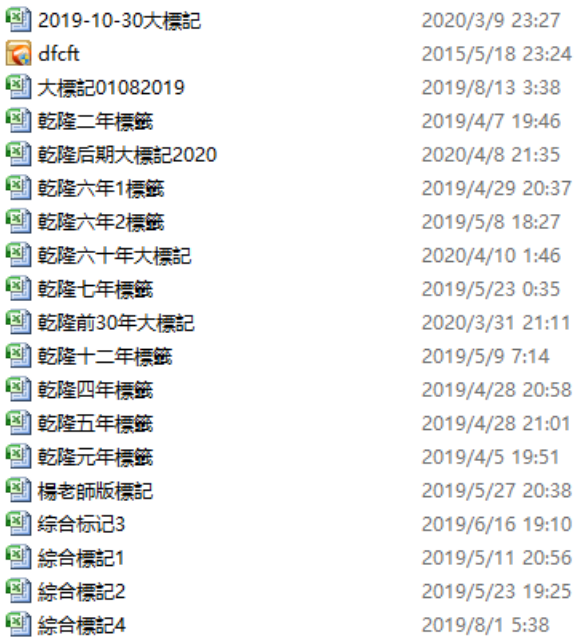

图 3：实操中应用的部分阶段性标记清单

4. 逻辑洁癖陷阱

码库思的操作者往往会陷入一种特殊的精神状态。如果其应用场景较为简单可能

尚好，而如果其应用场景如 EnamelFC 一般复杂，则不易幸免。**码库思制作团队的 Ho 先生指出，码库思平台还从未遇到如此复杂的应用场景**。故而在操作者心境这一问题上，本项目成员应该是有足够发言权的。

我暂时将这种心境称作**“逻辑洁癖陷阱”**。需要说明的是，这里所说的“逻辑洁癖”与上文中提到的“逻辑宽容”并非对立概念。“逻辑洁癖”是一种深层次的心智状态，是超乎操作本身而存在的无法遏制的体验；而允许部分在语义上不构成讹误的标记重叠或标记无意义的“逻辑宽容”则是一种主动选择的实操战术。“逻辑洁癖”的症状是，**在处理一种语境局面时，以参考类似语境局面的处理方式为第一准则，并极力保持所有同类语境局面的处理策略相同**。

这其中的例子很多，我仅举一例：在工作初期，我将“汉玉”作为一个整体处理，因为“汉玉”直到二十一世纪还是常用词。从直觉来看，“汉玉”的“汉”虽是题材定语，但附带了一种近乎材质的信息。但这一直觉处理随着实操的推进而出现动摇，尤其是在“商金银”、“周铜”等“朝代-材质”式意群出现时，我的二十一世纪直觉无法处理它们，只能选择放任不管，因为这些说法对清代人士而言可能是稀松平常的，而在当下已经基本不用。我由此返回到“汉玉”的情况，发现我不能确认这个词的准确定义：汉玉到底是一种玉，还是一种主题？或者真的是汉代遗物？如果是赝品汉代遗物或仿汉玉呢？如果是第一种情况，那么汉玉确实构成材质，而“商金银”、“周铜”也就应该被认为是某种（在化学意义上）特殊的金银和铜；如果是第二种情况，那么“汉”字就不该被标注，因为我们没有设立题材标记；如果是第三种情况，那么“汉”便应该作为时间概念存在；而如果是第四种情况，那么“汉”则又回归为题材标记。

遗憾的是，我并非玉器领域的专业人士。最初将“汉玉”做整体标记仅仅出于直觉。而**“逻辑洁癖陷阱”的第一症状就是语言直觉在反复的语义分析之下趋于崩塌**。在操作码库思半年之后，我决定将“汉玉”的“汉”字彻底淘汰，从此不作标记，以与“商金银”、“周铜”等的处理方式保持一致。这一决定随后又促使我重新审视“汉白玉”的语义模式，并最终认定，虽然“汉白玉”是一个整体常用词，但文献中的“汉白玉”未必是今日建筑概念上的汉白玉，而其“汉”字更不具有任何实际时间信息（因为“汉白玉”传统上又作“旱白玉”），故而“汉白玉”应仅标注“白”字为颜色，“玉”字为材质，而“汉”字则应舍弃不作标记。而为了与之保持一致，文献中偶尔出现的同义词“旱白玉”的“旱”字也应舍弃不作标注。整个逻辑的最终结果就是，“汉白玉=白玉”；“汉玉=玉”。

读者们，您觉得我的逻辑中有漏洞吗？或许有，或许如果是您在操作码库思，您可以提出更好的解决方案。但是我必须告诉您的是，您已经几乎无法说服我改变这一做法，我在我构建的逻辑高塔之上遗世独立。当我把这些思考与王华艳老师分享时，她评价道：“哈哈，你又入了“逻辑”的坑，走我老路了[…]这个是我们抠了很久逻辑，或者专门做研究的人，最后会得出的结论。这个是汉语的模糊性吧[…] 我们这些具体做的人就会去扣逻辑，但是一般人评价就靠直觉，所以一开始很难说服他们（2019/10/30）”。我很清楚的是，王华艳老师从零做起的时候，她所遇到过的逻辑陷阱要远远多过我，她在这其中的困苦挣扎也远比我更卓绝。她对此的发言权是远高于我的。我很希望她也能写下她的码库思探索史，这对码

库思的设计团队将有更重要的价值。

5. 眼、手、心

我在操作中，将一年的档案制作成一个文件。某些量特别大的年份则制作成两个文件。“清理”每一个文件的过程对于强迫症来说都是极大的满足，因为这从逻辑到视觉都是一个**“由乱而治”**的过程。由于码库思平台对于每个标记都可以全文检索并决定去留，故而每一种由语境造成的自动套用式逻辑讹误在理论上只需一次操作就可全文修改完成。所以每制作一个文件，必然是先慢后快，处理完第一个段落，整篇文档就已经略脱繁乱；处理完三分之一，则整篇文档的色彩已经显得有序。

当然，统一处理某标记词在一个文档中所出现的全部情况，则又是一种挑战。检索并决定文档中所有“白”字上颜色标记的去留，需要操作者首先决定是采取**“见误则删”**的战术，还是反其道而行之，采取**“见例则留”**的战术。战术的选择取决于检索出的标记词结果中有多大比例是符合语境的。假使在检索出的 200 个被标注为颜色的“白”字中，有 10 个是“白世秀”的姓氏，那么操作者便可采取“见误则删”战术，快速下拉清单，看到“白世秀”的“白”便删去颜色标记。这样只需点击 10 次，处理便完成了。而如果这 200 个被标注为颜色的“白”字中有 190 个都是“白世秀”的姓氏，那么操作者则可以采取“见例则留”战术，仅将那 10 个“白”在语境中确实作为颜色的例子锁定住，而一键删去所有其他检索结果，这样也仅需点击 10 次。但问题是，操作者往往需要处理这条检索结果清单到一半的时候才能知道到底哪种情况为多。在这个例子中，如果战术选择失误而又没有及时更改，操作者则需要点击 190 次鼠标才能完成清理。遗憾的是，这样的情况是经常出现的。

我小的时候玩过这样的游戏：班里的捣蛋鬼连续让我回答好几十个同样的问题，我都回答“是”。等我精神疲累了，他突然问我**“你是不是笨蛋”**。我在惯性之下便回答**“是”**。码库思就是一个这样的玩家，当我要处理一条长长的标记词检索结果清单而又不得不连续点击上百次删除时，它就仿佛是在等待我的疲累。然后，突然，百伪之中忽来一真。我的食指此时早已形成条件反射，来不及反应，便将那符合语境的标记也删去了。读者们，此时您觉得我会怎么办？找到那个错处并更改过来？不，**我会忽略这个错误，就让它错着**。如果之后我在操作或修订中有缘发现了这个错，我那时再把它改过来。而如果这个错误就此逃过了我的心智，那么我便由它去了。

我相信这跟 EnamelFC 项目的成员们想的不一样。他们把任务委托给我的时候恐怕不能想象我会这么做。但我还是要坦陈，我差不多就是这么处理码库思版的“我是笨蛋”游戏的。因为如果我试图在每一次被它玩弄时都忙不迭地去找回那个错误，补上那个被误删的标记，**那我早已疯了**。

操作码库思让我学会与错误共处。**在我所能探索出的操作策略在效率上的最优解中，永远都有错误的一席之地**。事实上，当我打开任何一个已经处理过、修订过

的文件，并上下随意浏览个三十秒之后，我都会发现错误，可能不止一个。我会改正它们。而当我下一次再打开同一个文件，又上下随意浏览个三十秒之后，我还会发现新的错误。直到某次，当我再打开它，浏览了一分钟都还没有发现错误的时候，我就知道它差不多做好了。我恐怕没有机会与未来的数据库使用者们对话。所以我要在这里说，**如果你们发现数据库有错误，不要惊讶**。发现一个错误就改正一个错误，这就是码库思的宇宙。

故宫博物院杨老师除了尽心拣选、录入各个主题的档案，他对数据库建设和码库思的实操也有深刻的见地。当我和王华艳老师有一次向他展示码库思的制作原理时，杨老师直言不讳地指出，**如果未来数据库的使用者知道数据库是这样由两个外行人在一个语言学信息处理平台上点点鼠标，对着满屏幕的错误做出来的，那么他绝对不会使用这个数据库**。尽管我们之后又多次聆听杨老师直率而富有哲思的指导，但在那一瞬间，我的内心第一次充满了对杨老师的感激。我们需要有人指出这一点，这并没有让我们感受到任何的挫折和失落，我很希望与包括码库思制作团队在内的大家分享这种珍贵的**“程序不信任”**，这就如同了解了汉堡包制作流程的老食客所表达的不信任对快餐店而言非常宝贵一样，对于文献加工流程的创作者也是非常宝贵的。

在码库思的操作界面上工作了一整天之后，我往往会有一种隐忧。我担忧我在码库思上所处理的档案被制作成数据库成品之后，会被未来的使用者认为是无用之物。**它对于“物的逻辑”的处理必然是不全面的；而它对于语境的“词的逻辑”的处理则又阻止了潜在的数据库使用者以纯粹语言学的方式使用这个数据库**，例如统计出这些档案中一共有多少个“样”字或“白”字，而并不在意它们的语义赋值。一位**全知的宫廷器物学家**使用者将会发现，“什么人做的这数据库，里头有错误”；而一位**全知的语言学家**使用者则会发现，“啊哟，做这数据库的人居然还自作聪明地进行了语义甄别，可惜他的头脑不过尔尔”。

当然最终我总能说服自己，这两种使用者都不可能真的存在。**对于任何一种信息逻辑的追求，都不可能被推演到极致，否则数据库便不必存在**。一个全知的宫廷器物学家和一位全知的语言学家大概不会信任任何数据库，他们可能早已通读世界上所有档案，或者早已开始属于自己的码库思之旅了。这或许才是码库思的使命。

6. 尾声

我的档案处理工作还在继续。有一次在微信上与 EnameIFC 项目的负责人赵老师估算已完成以及尚待完成的档案总量时，她无意间问我，“你处理 Markus 快吗（2020-03-29）”？我看着屏幕上她的问题，不禁乐出了声。这是一个关键性的问题，但却无法由我来回答。也许之后 EnameIFC 项目还会雇佣其它人来进行码库思实操，到那时可能这个问题才能得到解答。而码库思也将会因此而在另一个人的头脑中爆炸，诞生出另一个混杂着词与物的宇宙，而他的那个可能与我头脑中的截然不同，甚至连“快”和“慢”的定义都会不同。

Philippe Pons 先生曾指出，码库思的操作是学术性的（scientifique）（2020-03-11）。我很感谢他对我和王华艳老师工作性质的肯定。到目前为止，我的码库思探险充

满了乐趣与纠结，当然还有成就感。码库思的操作者是孤独的，他的乐趣与痛苦绝难与旁人诉说。我不敢说我就是那个最好的码库思用户，但是我的确从中得到了某种心智和知识上的锻炼，而且我知道**我的工作成果有着合格的质量，至少在我所能控制的范围内**。我想为此感谢 EnamelFC 项目，包括直接参与数据库制作流程的团队成员，尤其是把我哄赚进来的王华艳老师，还有其它所有的学术组成员。你们委托给我的是一件趣事。

李纬文 2020-04-16