对非专利文献数据深加工文献筛选和标引方法的进一步研究

专利检索咨询中心 邓小敏

摘要:针对目前使用的规则和实际加工中存在的不相适应的情况,建议扩大化学物质加工范围,并为目前不能标引但具有标引价值的几类化学物质提供标引方法。

关键词:数据加工 文献筛选 化学结构 混合物 大分子

在非专利文献加工中，目前使用的《中国非专利文献数据深加工标引规则》（以下简称标引规则）是2008年针对含可专利技术的非专利文献制定的，其涉及生物方法、制剂方法、化学方法、联合方法、新治疗应用、提取方法、物理方法和分析方法等八项可专利技术主题。每个主题的文献，有一部分需要标引化合物结构，另一部分不需要标引化学结构或无化学结构，其技术主题通过关键词、摘要和其他标引项目体现。标引规则要求对文献中确定结构的化合物进行“化合物信息”（CN号、职能符、化合物名称以及化合物结构等）标引[1]。2010年以来，非专利文献数据专项

加工项目要求根据是否有确定结构的化合物来筛选文献，再根据标引规则进行筛选，只标引同义词、方剂信息和化合物结构信息，其它项目暂不标引。集中进行化合物化学信息特别是化学结构标引并建立化学物质数据库将会为检索医药化学领域非专利文献提供一条有效途径，然而在实际加工过程中逐渐发现，标引规则“对结构确定的化合物进行标引”这一规定限制了大量涉及非化合物类化学物质的文献和技术信息的标引。本文试图对标引规则提出符合实际的解释，并对目前不能标引但具有标引价值的几类化学物质提出标引方法的建议。

化合物仅为化学物质的一部分，医药化学领域非专利文献中还有大量涉及非化合物类化学物质（如混合物、单质、阴阳离子及结构不确定的物质等）的技术方案，虽然这类物质不属于严格定义上的“化合物”，但也能够用化学结构来表征。根据是否有确定结构的化合物来筛选和标引文献，必然会导致大量涉及这类非化合物类化学物质技术方案的文献和技术信息漏标，无论是从构建数据库的完整性方面还是从检索的全面性方面来讲，都有必要对这部分文献和技术信息进行标引。从文献筛选的角度来讲，有必要将筛选含有“化合物”的文献扩展到筛选含有“能够标引化学结构的化学物质”的文献；从标引的角度来讲，有必要将对“化合物”进行标引扩大对“能够标引化学结构的化学物质”进行标引。下文总结了几类不属于标引规则规定加工范围之内的化学物质，以案例形式阐述了标引这几类化学物质的重要性，并提供了具体的标引方法。

一、混合物

混合物是由多种物质（化合物或单质）混合而成，按照笔者提供的标引方法可分为两类：一类为各组成成分和比例均确定；另一类为组成成分或比例不确定。

（一）组分和比例均确定的混合物，多见于药用组合物。很多数据库如美国化学文摘、各种药典都将这些药用组合物作为一个整体进行收录，给予其单独的CAS号和药物登记号，这类混合物的标引方法可以参照化合物的标引，如案例1。

案例1：丝裂霉素联合优福定治疗晚期胃癌24例疗效观察[2]

文献采用丝裂霉素（MMC）和优福定（UFT）联合给药方案治疗晚期胃癌。优福定为尿嘧啶和呋喃氟尿嘧啶的混合物，两者的当量比为4:1。优福定和丝裂霉素在该联合用药方案中的地位是等同的，但按照规则，只需要标引丝裂霉素而不标引优福定，这

会导致重要信息优福定漏标。因此建议将优福定作为一个整体进行标引，具体标引方式如下：

表 1 优福定的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	74578-38-4	M;K;T	优福定	CAweb

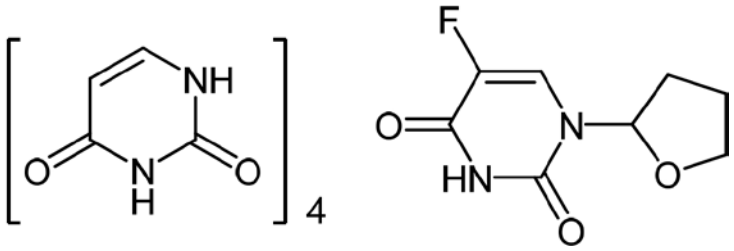


图 1 优福定的结构

这类药用组合物常见的还有：奥格门汀（阿莫西林和克拉维酸）、他唑西林（哌拉西林和他唑巴坦）、复方新诺明（磺胺甲噁唑和甲氧苄啶）、替门汀（替卡西林和克拉维酸）、异烟丁醇（乙胺丁醇和甲磺酸异烟肼）、天门冬氨酸钾镁（天门冬氨酸钾和天门冬氨酸镁）等。

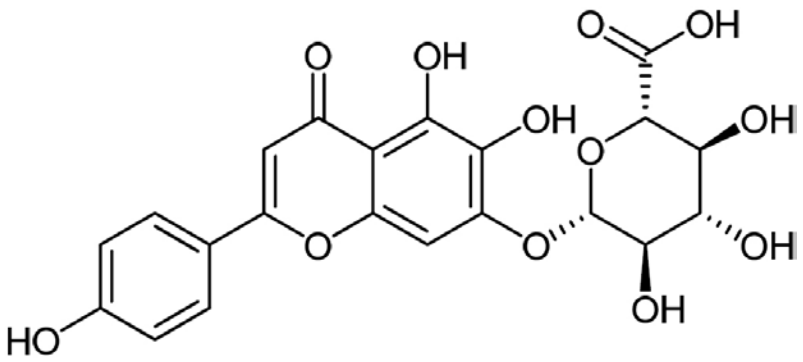
（二）对于组成成分或比例不确定的混合物的标引，需要视情况而定。一般来说，大部分难以绘制化学结构，但如果组成中某一种成分的含量很高或者该成分在所组成的混合物中具有代表性，比如是主要的活性成分，建议标引该具体成分的化学结构来代表该混合物，并在结构 GIF 图中标注该具体成分的名称，如案例 2 和案例 3。

案例 2：灯盏花素缓释微丸制备工艺与处方优化的研究 [3]

文献利用挤出滚圆法制备灯盏花素骨架型缓释微丸，考察了其制备工艺和最优处方，并探讨其释药机制。灯盏花素为该药物制剂中的活性成分，有必要进行标引。文中指出“灯盏花素是灯盏细辛中提取的黄酮类成分，主要为灯盏花甲素和灯盏花乙素，其中灯盏花乙素含量占 95% 以上”。因此灯盏花素的主要成分为灯盏花乙素 (scutellarin)，可以用灯盏花乙素的化学结构来代表灯盏花素的结构，具体标引方式如下：

表 2 灯盏花素的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	116122-36-2	K;T	灯盏花素	CBmed



scutellarin

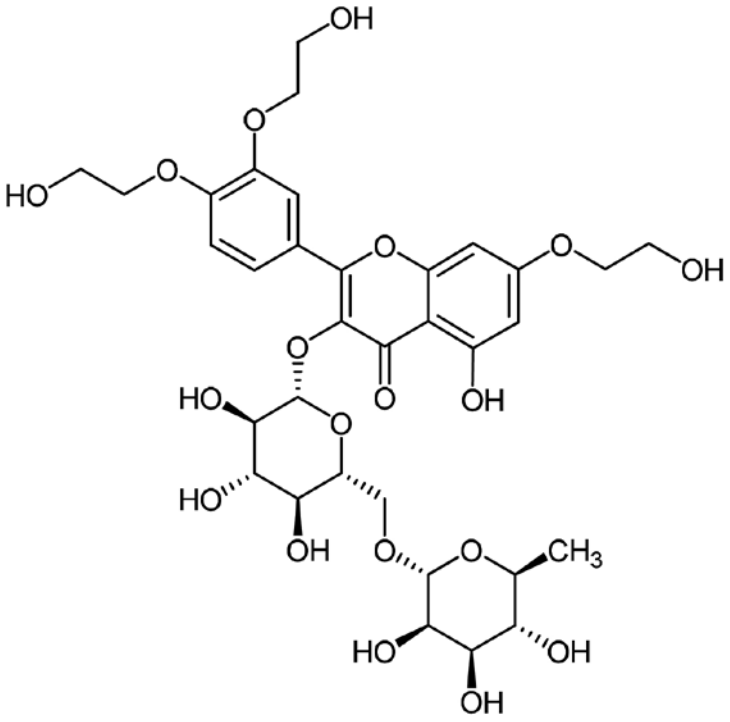
图 2 灯盏花素的结构

案例 3：用环氧乙烷法制备维脑路通[4]

文献采用环氧乙烷法制备维脑路通。维脑路通为多种羟乙基芦丁的混合物，其中 7,3′,4′－三羟乙基芦丁为主要有效成分，可以用 7,3′,4′－三羟乙基芦丁的化学结构来代表维脑路通的结构，具体标引方式如下：

表 3 维脑路通的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN		K;T;P	维脑路通	CBmed



7,3′,4′—Tris（hydroxyethyl）rutin

图 3 维脑路通的结构

二、单质、阴阳离子、原子团等

单质、阴阳离子、原子团等虽然不是化合物,但化学结构均是确定的,其具体标引方法可以参照化合物的标引，如案例 4 和案例 5。

案例 4：复方硫磺洗剂的制备及质量控制 [5]

文献以硫酸锌、硫磺、樟脑为主药，甘油、甲基纤维素为辅料，制备了复方硫磺洗剂。其中单质硫磺为主药之一，具体标引方式如下：

表 4 硫磺的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	7704-34-9	K;T;M	硫磺	CBmed

S

图 4 硫磺的结构

常见的药用单质还有 I_2、C、Au 等。

案例 5：双频超声强化提取黄柏中小檗碱的研究 [6]

文献采用双频超声强化法从黄柏提取中小檗碱，为热敏性药物的提取提供了新的强化方法。小檗碱为提取分离的主题产品，其为一价阳离子，文献没有指出与哪种阴离子形成了盐，因此标引小檗碱，具体标引方式如下：

表 5 小檗碱的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	2086-83-1	K;T;P	小檗碱	CBmed

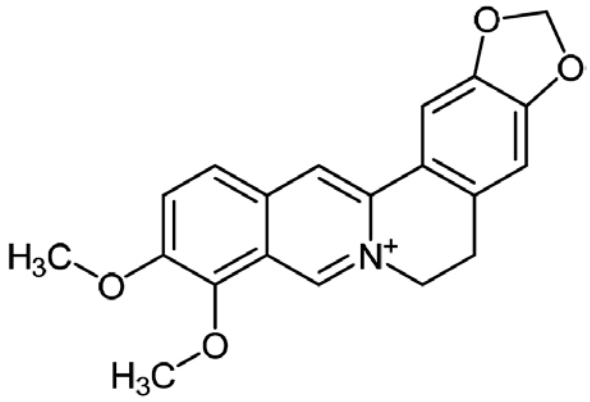


图 5 小檗碱的结构

医药领域经常涉及到的离子还有新斯的明、乙酰胆碱、血根碱、筒箭毒碱等。

三、结构不确定的化学物质

在非专利文献中，结构不确定的化合物主要涉及到以下几类：一类化合物的总称；化合物取代基位置或链的长度不确定；化合物中各组成部分比例或连接方式不确定。结构不确定的化学物质和混合物有部分交集，同混合物一样，这类化合物大部分也无法绘制化学结构，但以下几种特殊的类型还是可以用化学结构来表征的。

（一）一类具有相同母核的化合物的总称，比如香豆素类化合物、博来霉素、头孢菌素（先锋霉素）、卷曲霉素、乙酰螺旋霉素、新霉素、多粘菌素 B、螺旋霉素、麦白霉素及柱晶白霉素等。这类化合物具有相似的结构，即具有共同的母核，仅取代基不同。在标引这类物质时，可以将总称所包含的化合物作为一个整体进行标引，化学结构可以采取母体加取代基的形式表示，例如案例 6。

案例 6：金银花水煎液抗绿脓杆菌生物膜作用及其与庆大霉素的协同作用 [7]

文献研究结果表明金银花水煎液对绿脓杆菌生物膜具有较强的抑制作用，且与庆大霉素具有明显的协同作用。庆大霉素为一类具有相同氨基糖苷母核的抗生素，包括庆大霉素 C1(Gentamicin C1)、庆大霉素 C2(Gentamicin C2)、庆大霉素 C1a(Gentamicin C1a) 及庆大霉素 C2a(Gentamicin C2a) 等。原文并未指明具体是哪一种庆大霉素，因此将庆大霉素作为一个整体进行标引，具体标引方式如下：

表 6 庆大霉素的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	1403-66-3	K;T;M	庆大霉素	CBmed

（二）化合物中各组成部分的比例不确定，比如盐的酸部分和碱部分的比例不确定，此时绘制化学结构时可以将盐的各部分还原成酸和碱或者金属原子，用变量 x、y、z 等表示各部分不确定的比例关系。例如案例 7。

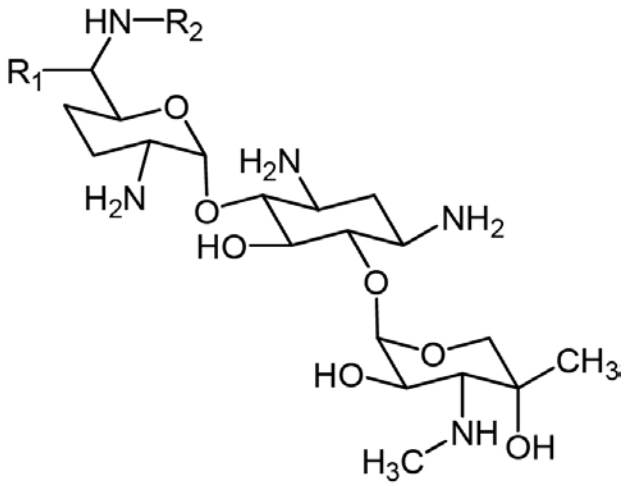
案例 7：盐酸川芎嗪口服定时释药微丸的制备 [8]

文献采用双层包衣法制备了盐酸川芎嗪定时释药微丸。盐酸川芎嗪为该药物制剂中的活性成分，有必要进行标引。盐酸川芎嗪为盐酸和川芎嗪形成的盐，在 CAweb 中检索得到盐酸川芎嗪中盐酸与川芎嗪的比例是不确定的，

可以采用以下方式对其进行标引：

表 7 盐酸川芎嗪的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	76494-51-4	K;T	盐酸川芎嗪	CAweb



Gentamicin C_1 $R_1=R_2=CH_3$
Gentamicin C_{1a} $R_1=R_2=H$
Gentamicin C_2 $R_1=CH_3$, R2=H

图 6 庆大霉素的结构

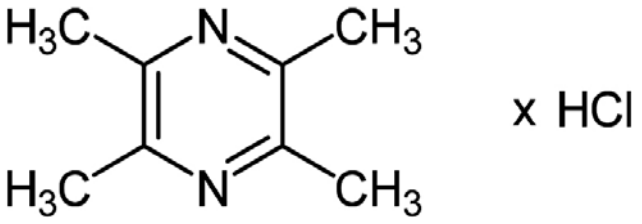


图 7 盐酸川芎嗪的结构

此类结构不确定的化学物质常见的还有:硫酸依替米星、硫酸壳糖胺、盐酸氯环利嗪、盐酸万古霉素等。

（三）化合物中各组成部分的连接方式不能确定，这种情况在络合物和无法判断反应位置的有机酸盐或有机碱盐中比较常见。在绘制化学结构时可以忽略各组成部分的连接方式，仅需绘制各组成部分及比例，比如案例 8 和案例 9。

案例 8：硫磺酸锌的合成研究 [9]

文献采用液相合成路线和醇沉技术，利用超声波合成牛磺酸锌。牛磺酸锌为化学合成方法制备的主题产

品，其为一种配合物，中心原子为正二价锌，配体为硫磺酸根，两者比例为 1:2，但并没有检索到有关锌与硫磺酸配位方式的信息，此时可以采用以下方式对其进行标引：

表 8 牛磺酸锌的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN		K; T; P	牛磺酸锌	CBmed

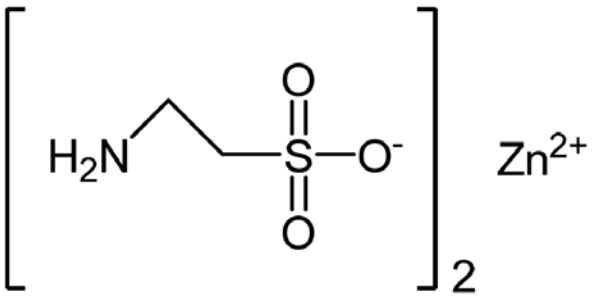


图 8 牛磺酸锌的结构

案例 9:三磷酸腺苷二钠的回收 [10]

文献提供了一种回收三磷酸腺苷二钠的方法。三磷酸腺苷二钠为主题产品，应该标引。1 分子三磷酸腺苷中有 3 个成盐位点，不能确定是其中的哪两个参与反应形成了三磷酸腺苷二钠，因此在绘制化学结构时将三磷酸腺苷二钠还原为三磷酸腺苷和钠原子进行绘制。

表 9 三磷酸腺苷二钠的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	987-65-5	K; T; P	三磷酸腺苷二钠	CBmed

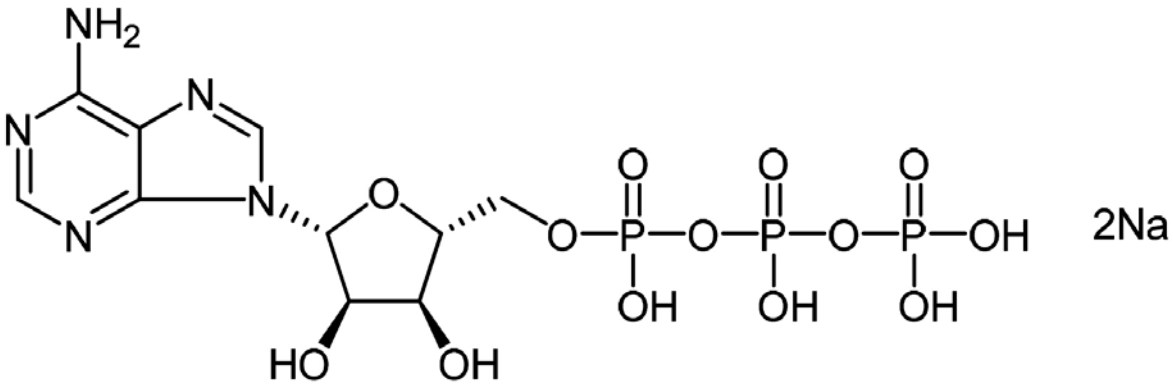


图 9 三磷酸腺苷二钠的结构

四、大分子化合物

大分子化合物主要为有机物，例如蛋白质、多糖、脂类以及其他聚合物等。由于分子量大，结构复杂，结构绘制时存在一定难度，甚至大部分大分子化合物没有确定的结构，因此大多数大分子化合物标引化学结构的意义不大，但聚合物的单体和聚合方式均是已知的大分子，可以进行结构标引，如案例 10。

案例 10：透明质酸的发酵及提取工艺[11]

文献提供了一种透明质酸的微生物发酵及分离纯化方法。透明质酸为通过发酵制备的主题产品，有标引其化学信息的价值。透明质酸是以 D-葡萄糖醛酸和 N- 乙酰氨基葡萄糖单体为结构单元，通过 β -1,4 糖苷键反复交替连接而形成的链状高分子酸性粘多糖，可采用以下方式对其进行标引：

表 10 透明质酸的结构信息

No.	CN	RN	Role	Name	StructureReference
1	CN	9004-61-9	K;T;P;E	透明质酸	CBmed

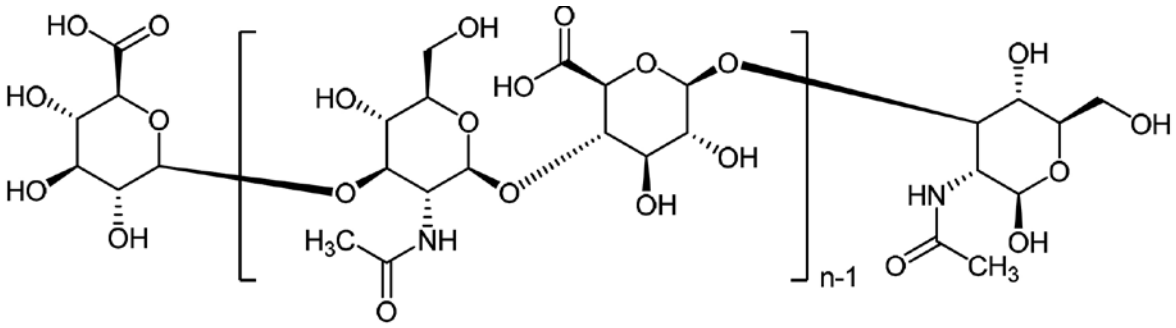


图 10 透明质酸的结构

这类大分子化合物常见的还有右旋糖酐、右旋糖酐铁、肝素、玻璃酸钠等。

以上可以标引化学结构的化学物质，在标引结构的 MOL 图时，可参考化合物的 MOL 图标引方式。对于如案例 10 的聚合物，MOL 图仅标引聚合物的单体结构。

小结

仅要求对文献主要技术主题中结构确定的化合物进行化合物信息标引，是基于同时标引关键词和重新撰写摘要的基础上制定的标引规则。对于实际工作中遇到大量不在目前标引范围之内的其他化学药物或含有在医药领域常用的混合物而言，标引规则

有很大的局限性，需要对“结构确定的化合物”的定义进行拓展，有必要将不属于化合物但能用化学结构进行表征的化学物质也包括进来。标引这些化学物质是对目前数据加工范围的拓展和加工内容的必要补充。笔者对这些化学物质进行总结分类，并提供了具体的标引方法，可为全面标引化学结构和建立完整的化学物质数据库提供参考。

（专利检索咨询中心　杨晓春　审校）

参考文献

1. 中国非专利文献数据深加工标引规则.2010年
2. 彭大为．“丝裂霉素联合优福定治疗晚期胃癌24例疗效观察”.《临床荟萃》1991年第6卷第6期第267-268页。
3. 陈大为、张彦青、邹艳霜、李淑斌、赵秀丽．“灯盏花素缓释微丸制备工艺与处方优化的研究”．《中草药》2003年第11卷第34期第990-993页。
4. 屠世忠．“用环氧乙烷法制备维脑路通”．《华西药学杂志》1986年第3卷第1期第66页。
5. 张流明．“复方硫磺洗剂的制备及质量控制”．《海峡药学》2006年第1卷第18期第35-36页。
6. 曾荣华、陆海勤、丘泰球．“双频超声强化提取黄柏中小檗碱的研究”.《天然产物研究与开发》2005年第6卷第17期第769-772页。
7. 袁秀丽、吕嘉枥、程慧娟．“金银花水煎液抗绿脓杆菌生物膜作用及其与庆大霉素的协同作用”．《西北药学杂志》2010年第3卷第25期第201-203页。
8. 胡海洋、宋华先、陈大为、刘丹．“盐酸川芎嗪口服定时释药微丸的制备”．《沈阳药科大学学报》2008年第5卷第25期第342-346页。
9. 孟洁、赵红坤、李娟、任保增、雒廷亮、刘国际．“硫磺酸锌的合成研究”．《浙江化工》2004年第6卷第35期第1-3页。
10. 李建淮、张天民．“三磷酸腺苷二钠的回收”．《中国生化药物杂志》1983年第2期第18-19页。
11. 汪江波、黄金明、张晶．“透明质酸的发酵及提取工艺”．《华中农业大学学报》2010年第5卷第29期第648-653页。