

《山海经》知识库的知识图谱构建与知识元标引实践

□ 曹菁菁 国家图书馆古籍馆

《山海经》是中国古代经典，是保留了大量上古传说的一部地理书。这本书原本分为《山经》《海经》。秦末战火之后，典籍散佚。汉代朝廷向民间征集书籍，《山海经》从民间被发现，但图已经亡佚。武帝之后，朝廷整理的《山海经》包括《山经》《海内经》《海外经》《大荒经》。学者们对于《山海经》的成书过程众说纷纭。一般认为，今存本中《大荒海内经》五篇成书最早，《海内四经》《海外四经》及《山经》成书于战国至秦汉之时。

《山海经》作为古代经典，其所记载的史料可以上溯到殷商时期。它本是图与文并存的一种典籍。在漫长的传承中，许多地理信息随着图表的散佚而缺失了。同时，经过历朝历代的修改，《山海经》的文本也出现了一定程度的变化。在建设本知识库时，我们使用了国家图书馆藏的 92 种版本，书目基本数据如下：

序号	索书号	题名	版本
1	08272	山海经十八卷	宋淳熙七年（1180）池阳郡斋刻本
2	02607—第304 函	山海经十八卷	明正统十年（1445）刻道藏本
3	08273	山海经传十八卷	明成化元年（1465）吴宽抄本
4	11354	山海经传十八卷	明成化四年（1468）北京国子监刻本
5	09000	山海经水经合刻	明嘉靖十三年（1534）黄省曾刻本
6	03526	山海经传十八卷	明嘉靖十五年（1536）潘侃前山书屋刻本
7	00215	山海经释义十八卷	明嘉靖（1522—1566）刻本
8	12087	山海经释义十八卷	明嘉靖（1522—1566）刻本
9	11355	山海经传十八卷	明万历十三年（1585）吴琯刻合刻山海经水经本
10	A00918	山海经释义十八卷图一卷	明万历二十五年（1597）蒋一葵尧山堂刻本
11	12274	山海经传十八卷	明万历吴琯刻古今逸史本
12	07878	山海经释义十八卷图七十五回	明万历刻本
13	08275	山海经图赞二卷	明万历刻秘册汇函本
14	04954	山海经释义十八卷	明万历刻清修本
15	17006: 6	山海经释义十八卷	明万历刻清乾隆十年（1745）重修本
16	A00503	山海经传十八卷	明抄本

17	13810	山海经传十八卷	明刻本
18	11125	山海经十八卷	明刻本
19	t4495	山海经传十八卷	明周如山大业堂刻本
20	17689: 4	新刻山海经传十八卷	明胡文焕刻格致丛书本
21	A02850: 16	山海经补注一卷	明刻杨升庵杂著十四种本
22	t4496	新刻山海经图二卷	明胡文焕刻格致丛书本
23	06505	山海经图赞一卷	明天启刻藁古介书前集本
24	T01026: 9	山海经十八卷	明刻增定古今逸史本
25	19102: 5	山海经十八卷	清康熙五十三年至五十四年（1714—1715）项綢群玉书堂刻本
26	A02850: 16	山海经传十八卷	清康熙五十三年（1714）项綢群玉书堂刻本
27	A02848: 43	山海经广注十八卷读山海经语一卷杂述一卷引用书籍一卷图五卷	清康熙刻本
28	A05381: 2	山海经十八卷	清康熙刻秘书廿一种本
29	15911	山海经十八卷	清乾隆黄晟槐荫草堂刻本
30	7085: 7	山海经补注一卷	清乾隆刻艺海珠尘本
31	38296: 5	山海经新校正十八卷古今本篇目考一卷	清乾隆四十六年（1781）毕沅灵岩山馆经训堂刻本
32	地 701/314.5	山海经图赞一卷	清乾隆刻艺海珠尘本
33	38562: 9—10	山海经图赞补逸一卷	清乾隆四十九年至嘉庆元年（1784—1796）卢氏抱经堂刻群书拾补本
34	153801: 13—14	山海经广注十八卷读山海经语一卷杂述一卷图五卷	清乾隆五十一年（1786）刻本
35	56352: 38	山海经新校正十八卷篇目考一卷	清乾隆四十八年（1783）镇洋毕沅灵岩山馆刻经训堂丛书本
36	13089: 51—52	山海经补注一卷	清乾隆间绵州李调元刻函海本
37	55469: 62—64	山海经补注一卷	清乾隆间绵州李调元刻清末广汉钟登甲乐道斋重修函海本
38	39196: 1	山海经新校正十八卷篇目考一卷图五卷	清乾隆刻本
39	XD532: 7	郭璞山海经图赞一卷	清嘉庆刻汉唐地理书钞本
40	41888	山海经十八卷	清嘉庆十九年（1814）文星堂刻本
41	地	山海经笺疏十八卷图赞一	清嘉庆至光绪间刻郝氏遗书本

	701.1/744.1	卷订讹一卷叙录一卷	
42	地 701.1/744.2	山海经补注一卷	清嘉庆南汇吴省兰听彝堂刻艺海珠尘 本
43	37618: 29	山海经图赞一卷补遗一卷	清嘉庆南汇吴省兰听彝堂刻艺海珠尘 本
44	40551: 57	山海经图赞	清嘉庆南汇吴省兰听彝堂刻艺海珠尘 清道光三十年金山钱氏漱石轩重印本
45	42957: 37	山海经补注	清嘉庆南汇吴省兰听彝堂刻艺海珠尘 清道光三十年金山钱氏漱石轩重印本
46	41848: 86	山海经图赞一卷补遗一卷	清嘉庆南汇吴省兰听彝堂刻艺海珠尘 清文萃堂后印本
47	地 701.1/824.1	山海经赞佚文	清嘉庆道光间秀水姚东升佚书拾存稿 本
48	地 701/314.8	山海经十八卷	清道光二十六年（1846）刻秘书廿八 种本
49	地 701.1/846 部一	山海经四卷	清咸丰五年（1855）海清楼刻本
50	40653: 1	读山海经一卷	清同治十年（1871）德清俞氏刻光绪 七年（1881）增刻春在堂全书本
51	地 701.1/846.1	读山海经一卷	清同治十年（1871）德清俞氏刻清光 绪二十三年（1897）增修春在堂全书 本
52	13359: 81— 83	读山海经一卷	清同治十年（1871）德清俞氏刻光绪 二十五年（1899）增修春在堂全书本
53	地 701.1/846.3	读山海经一卷	清同治十年（1871）德清俞氏刻光绪 二十八年（1902）增修春在堂全书本
54	41653: 40— 41	山海经十八卷图赞一卷补 注一卷	清光绪元年（1875）崇文书局刻子书 百家本民国四年（1915）石印本
55	58133: 38— 39	山海经补注一卷	清光绪八年（1882）总纂升庵合集本
56	13361: 38— 40	山海经十八卷	清光绪三年（1877）浙江书局刻光绪 二十七年重修二十二子本
57	57962: 39— 40	山海经十八卷	清光绪十三年（1887）大同书局经训 堂丛书石印本
58	55930: 46— 48	山海经笺疏十八卷图五卷 图赞一卷订讹一卷叙录一 卷	清光绪二十一年（1895）上海书局石 印本

59	地 701.1/846.4	山海经图赞二卷	清光绪间湘潭叶德辉刻光绪二十一年（1895）民国八年（1919）重修民国印观古堂汇刻书本
60	地 701.1/865.1	山海经新校正十八卷篇目考一卷图五卷	清光绪二十三年（1897）上海图书集成局铅印子书二十二种本
61	38987: 13—16	山海经存九卷首一卷	清光绪二十三年（1897）汇印汪双池先生丛书本
62	地 701.1/865.2	山海经笺疏十八卷图五卷图赞一卷订讹一卷叙录一卷	清光绪二十五年（1899）上海江左书林影印本
63	地 701.1/865.4	山海经图赞二卷	清光绪间湘潭叶德辉刻光绪三十四年（1908）重修观古堂汇刻书本
64	地 701.1/865.5	山海经图赞二卷	清光绪间湘潭叶德辉刻民国八年（1919）重修观古堂汇刻书本
65	地 701.1/865.3	山海经笺疏十八卷图赞一卷订讹一卷叙录一卷	清光绪刻本
66	141835: 11—14	山海经图赞二卷	清光绪间湘潭叶德辉刻观古堂汇刻书本
67	37489: 1058—1061	山海经十八卷	清宣统三年（1912）上海图书集成局铅印子书二十八种本
68	地 701/314.6	山海经笺疏十八卷图赞一卷订讹一卷	清阮氏琅环仙馆本
69	8759: 109—112	山海经四集十八卷杂述一卷图五卷	清刻本
70	40170: 41	山海经补注一卷	清刻本，听彝堂藏版，据清嘉庆间南江吴省兰听彝堂刻版后印
71	39611: 12	山海经	清刻经余必读续编本
72	37474: 54	山海经腴词	清刻小嫫嬛山馆汇刊类书本
73	10199: 54	山海经十八卷	民国三年（1914）育文书局石印子书十六种本
74	37306: 49	山海经十八卷	民国五年（1916）上海书局子书三十二种本
75	8094: 1	山海经笺疏十八卷图赞一卷订讹一卷	民国五年至六年（1916—1917）潮阳郑国勋龙溪精舍刻朱印龙溪精舍丛书本
76	10603: 144	山海经十八卷篇目考一卷	民国八年（1919）上海扫叶山房石印三十六子全书本

77	42957: 53	山海经十八卷	民国九年（1920）上海五凤楼刻子书四十八种本
78	38537: 46	山海经十八卷	民国十年（1921）育文书局石印子书三十六种本
79	38537: 33	山海经十八卷图赞一卷补注一卷	民国十年（1921）上海扫叶山房石印本
80	8830: 55	山海经为诗经旧传考	民国十年（1921）四川存古书局刻六译馆丛书本
81	144243: 5	山海经十八卷篇目考一卷	民国十二年（1923）上海扫叶山房石印三十六子全书本
82	38163: 2	山海经十八卷	民国十二至十五年（1923—1926）上海商务印书馆正统道藏影印本
83	7793: 4	山海经新校正十八卷篇目考一卷	民国十三年（1924）上海扫叶山房石印本
84	39150: 2	山海经十八卷图赞一卷补注一卷	民国十四年（1925）扫叶山房子书百家上海石印本
85	40581: 2	山海经图赞一卷	民国二十四年（1935）上海大东书局影印指海本
86	10253: 102	山海经图赞二卷	民国二十四年（1935）长沙叶氏观古堂汇印郎园全书本
87	7086: 15—17	山海经十八卷	民国二十六年（1937）上海商务印书馆古今逸史影印本
88	20350: 675—676	山海经类编	民国抄本
89	55395: 64—65	山海经笺疏十八卷图赞一卷订讹一卷	民国间上海中华书局铅印四部备要本
90	55659: 64	山海经地理今释六卷	上海古籍书店一九六三年重刻求恕斋丛书本
91	10230: 70—75	山海经图赞一卷	一九八七年扬州江苏广陵古籍刻印社影印群书拾补初编本
92	8830: 37	山海经笺疏十八卷图赞一卷订讹一卷	二十世纪九十年代中国书店影印潮阳郑国勋龙溪精舍刻龙溪精舍丛书本

上述 92 种版本主要分为三种类型：

《山海经》的十八卷本：十八卷本《山海经》以汉代刘向、刘歆父子整理本为基础，东晋学者郭璞对《山海经》进行了注释，含有注、赞、图。《隋书·艺文志》和《旧唐书·艺文志》有《山海经》二十三卷，今已经不存。今本郭璞注本《山海经》为十八卷本，图画散佚。这一版本是目前最为流行的《山海经》版本。是知识库中标引的主要对象。

《山海经图赞》：此书相传为郭璞所著，部分赞的内容存于郭璞注本《山海经》中，单行二卷本《山海经图赞》有人质疑为后人托名之著。在知识库中，由于有图有文，我们将其当做一种独立的文献类型进行标引。

《山海经》其他相关著作：这些著作与《山海经》相关，但文本只是与《山海经》的内容有关系，比如引用了部分经文或词句，但文献内容主要为作者的创作。

本次建设主要需要将上述三种类型的文献数据进行深层标引,并建立一个便于读者理解的知识图谱。

一、知识元标引流程

知识元是指语义完整、不可再分的最小知识单元。知识元逻辑关系抽取是在定义逻辑关系类型的基础上,判断两条知识元间是否存在以及存在何种关系,进而构建描绘领域整体态势的事理逻辑图谱可应用于知识体系架构、知识智能推荐、情报监测与预警等场景。^①对知识元的界定、抽取、处理、标引、描述是知识图谱构建的基础,也是对文本数据进行深层标引的基本工作。

（一）选取知识元

《山海经》充满了大量的专名词，有学者统计约有 2142 个，^②涉及古代生活的各个方面。本次标引的主要对象即为专名词。机器在全文识别的文本基础上，必须对所有数据进行知识元的抽取。首先，确定知识元的类型，并提供知识元的样例大约 1000 个左右，然后训练机器进行知识元+词性的识别。具有古籍训读初步实践的机器很快会对文本进行全面分析，并切出初步的名词性知识元。得到机器提供的知识元总表后，需要人工进行审核和干预，对其中的知识元进行标准化界定并训练机器记住这些知识元所出现的具体位置，以保证它在其他文本中快速定位到相似性较高的知识元并进行判断和抽取。

[illegible]

图 1 机器初步抽取知识元的结果 (局部)

^①周京艳、刘如、李佳娱等：《情报事理图谱的概念界定与价值分析》，《情报杂志》2018年第5期，第31—36、42页。

^②贾雯鹤：《论山海经专名研究的理论与方法》，《天府新论》2007年第1期，第138—144页。

（二）知识元分类

知识元作为文本中最小的知识单位，在《山海经》中高频出现，并承载了信息动态关联的作用。我们对每一个知识元都需要进行足够细致的标引。对于知识元本身，需要进行聚合管理，否则难以呈现知识图谱的可视化。对于专名来说，初步工作在于有效地对知识元进行分类标引。经过对于专名的研究并参考大量研究文献之后，本知识库采用了对知识元进行三级类目管理，第一级类目是大类目，第二级类目是子类，第三级类目是标准控制类目。一级二级分类模板如下：

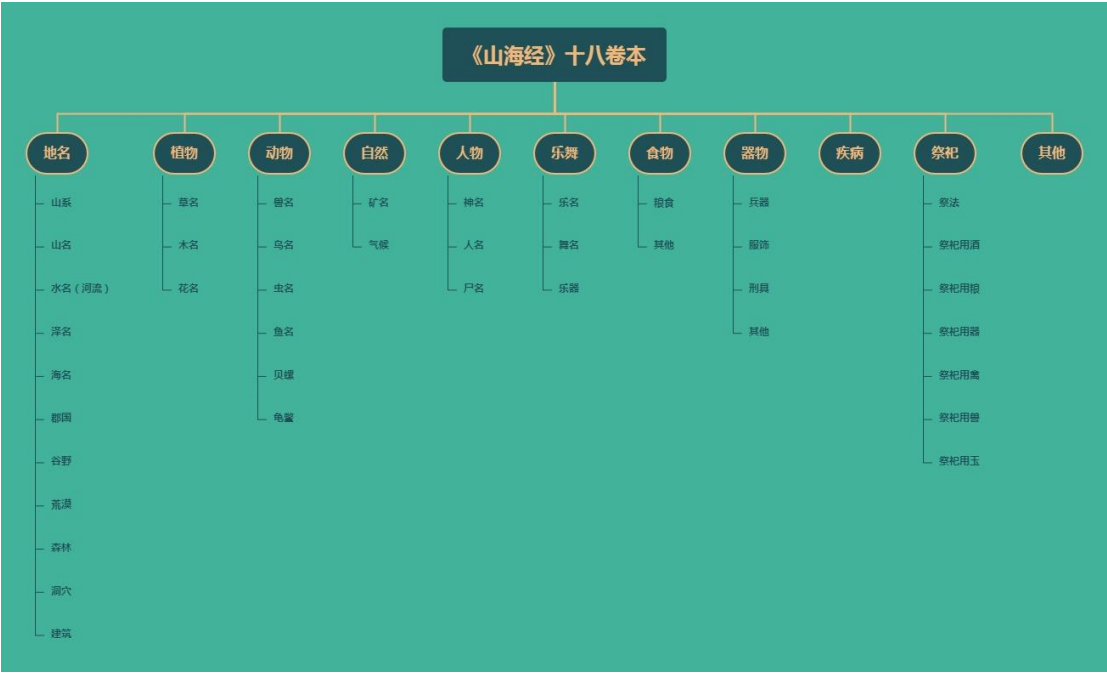


图 2 十八卷本《山海经》专名词的二级分类模版

从知识元分布的情况来看，《山海经》出现了大量的地名和丰富的动植物名，祭祀用语和与之相关的人物名、器物名。这种知识元的分布情况大致符合《山海经》作为早期中国山川地理与宗教信仰典籍的文献性质。

由于二级分类对于知识元的分类聚合十分重要，所以在标引知识元的过程当中，它占据两个独立的重要字段。

（三）标引知识元

知识元虽然独立被抽取出来，但是它们在文献中的定位、内涵、外延、逻辑关系等必须加以细粒度标引，方能支持后续知识图谱的建设。针对《山海经》的文献特征，知识元在多个维度上需要进行细粒度标引（表 1）：

表 1：术语标引著录主要字段表：

序号	字段名称	著录内容
1	加工编号	
2	术语名称	
3	一级类目	
4	二级类目	

5	出处书名	
6	出处书目加工编号	
7	出处版本	
8	出处章节编号	
9	地理层级	参见地理方位标引。
10	描述	此字段可重复，每个字段容纳一项元素描述，如“有花”“黑理”。
11	能否利用	是为“1”，否为“0”。
12	利用方法	是为“1”时需具备。一项标引一种利用方法。
13	功效	一项标引一种功效。
14	关联项 ID	关联术语的加工编号。
15	关系类型说明	相关术语的关系类型说明。

唯一标识号。每一个出现的专名词都被赋予唯一的标识号。每一个知识元因为其所出现的文本位置不同，都具有独一无二的标引意义。

术语坐标：这是机器自动生成的、标注专名在原书页双层 PDF 中的固定位置，以便实现原图对照显示及原文位置跳转。

一级类目和二级类目。这是知识元进行分类和聚合的重要字段。

出处书名、书名唯一标识号、出处章节编号。这是系联知识元与原书书目信息及原文位置的重要字段。

地理层级。《山海经》每一个知识元都在特定的地理位置上出现，故而地理位置是最基本的标引点。有一个知识元出现在多个地点的情况，将它们标引清楚，才能得到每一个知识元的地理分布情况。地理层级与地理标引相互关联，可以系联同一个地点出现的所有知识元，也可以聚合同一个知识元出现的所有地点。

描述。描述部分抽取知识元的上下文关键字词，是逻辑分析时使用的重要字段。

能否利用与利用方法。是专名词知识元使用场景的描述字段。

关联项与关系类型说明。这两个字段是人工干预字段，主要作用在于描述机器不能识别的专名项之间的联系。比如描述从属位置关系，“虡勺”这个知识元下有“南流注于虡勺”的上下文，那么关联项是南流的唯一识别号，关系类别是“注入”。又比如界定同一个事物，唯一识别号为 GJSH20210001001SY0119 的“诸毗”和唯一识别号为 GJSH20210001001SY0089 的“诸毗”是同一个地点，故而关联项字段填写 GJSH20210001001SY0089，关系类型填写“相等”。

加工编号	术语名称	术语坐标	文件名称	一级类目	二级类目	三级类目
term_id	term_name	term_coordin	file_name	firsts_category	seconds_category	thirds_category
1 GJSH20210001001SY0001	南山經之首	["fileName": "0017.pdf]		地名	山系	
2 GJSH20210001001SY0002	離山	["fileName": "0017.pdf]		地名	山系	
3 GJSH20210001001SY0003	招搖之山	["fileName": "0017.pdf]		地名	山名	
4 GJSH20210001001SY0004	西海	["fileName": "0017.pdf]		地名	海名	
5 GJSH20210001001SY0005	桂	["fileName": "0017.pdf]		植物	木名	桂树
6 GJSH20210001001SY0006	金	["fileName": "0017.pdf]		自然	矿名	金属
7 GJSH20210001001SY0007	玉	["fileName": "0017.pdf]		自然	矿名	玉
8 GJSH20210001001SY0008	韭	["fileName": "0017.pdf]		植物	草名	韭

出处书名	出书书目加工编号	出处版本	出处章节	地理层级
ccsm	ccsmjgbh	ccbb	cczjbh	dlcj
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	GJSH20210001001DM0001
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	GJSH20210001001DM0001
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	
山海經十八卷	GJSH20210001001	宋淳熙七	GJSH20210001001PZ0005	

地名	描述	是否利用	利用方式	功效	关联项ID	关系类型说明	状态
dm	descriptio	nfly	lyff	gx	glx_id	gxlxsm	status
南山經之	["descl":	0			/		2
離山	["descl":	0			/		2
招搖之山	["descl":	0			/		2
	{}	0			/		2
南山經-離	["descl":	0			/		2
南山經-離	["descl":	0			/		2
南山經-離	["descl":	0			/		2
	{}	0			/		2
南山經-離	["descl":	1	食	不飢	/		2

图 3 标引字段一览表

（四）提取语义与构建语义关联

在本期建设中，《山海经》文献中的标引重点是专名词。专名词的语义分析则重点在语义聚合方面。

在文献中，有一些专名虽然同形，但并非指向同一事物，这种专名有唯一识别号进行区分。但有些专名同形亦指向同一语义，需要人工校验并标引。比如，文献中有 63 处“河”，都指向“黄河”。这 63 处黄河虽然都有唯一识别号，但均在“关联项”字段中注明了第一处指向黄河意义的专名“河”的唯一标识号，并将“关系类型说明”字段标引为“相等”。同义词聚合情况主要出现在地名的标引中。由于河流流域大、山系绵延长，在多个地点中作为参照物出现是很常见的现象。对这类专名词的语义标引有利于对于山川示意图的勾勒。

而在动植物专名的近义聚合方面，提取语义并进行语义关联同样重要。比如，豹、猛豹、玄豹、雌豹，都指向“豹”这类野兽，而且均多次出现。分析文献上下文语义时，豹、猛豹、玄豹、雌豹为同类近义词，但是有具体的指向事物。所以近义词的语义聚合，通过“三级类目”字段进行系联。而同形专名词多次出现，上下文分析之后，又为指向为同一语义的专名词，则依旧在“关联项”字段和“关系类型说明”字段进行标引。

962	GJSH20210001001SY0962	河	0064.pdf	地名	水			GJSH20210001001SY048
1029	GJSH20210001001SY1029	河	0068.pdf	地名	水			GJSH20210001001SY048
1039	GJSH20210001001SY1039	河	0069.pdf	地名	水			GJSH20210001001SY048
1057	GJSH20210001001SY1057	河	0069.pdf	地名	水			GJSH20210001001SY048
1065	GJSH20210001001SY1065	河	0069.pdf	地名	水			GJSH20210001001SY048
1083	GJSH20210001001SY1083	河	0070.pdf	地名	水			GJSH20210001001SY048
1152	GJSH20210001001SY1152	河	0073.pdf	地名	水			GJSH20210001001SY048
1220	GJSH20210001001SY1220	河	0076.pdf	地名	水			GJSH20210001001SY048
1225	GJSH20210001001SY1225	河	0076.pdf	地名	水			GJSH20210001001SY048
1347	GJSH20210001001SY1347	河	0081.pdf	地名	水			GJSH20210001001SY048
1388	GJSH20210001001SY1388	河	0083.pdf	地名	水			GJSH20210001001SY048
1404	GJSH20210001001SY1404	河	0084.pdf	地名	水			GJSH20210001001SY048
1432	GJSH20210001001SY1432	河	0085.pdf	地名	水			GJSH20210001001SY048
1439	GJSH20210001001SY1439	河	0085.pdf	地名	水			GJSH20210001001SY048
1446	GJSH20210001001SY1446	河	0086.pdf	地名	水			GJSH20210001001SY048
1474	GJSH20210001001SY1474	河	0087.pdf	地名	水			GJSH20210001001SY048
1477	GJSH20210001001SY1477	河	0087.pdf	地名	水			GJSH20210001001SY048
1484	GJSH20210001001SY1484	河	0087.pdf	地名	水			GJSH20210001001SY048
1504	GJSH20210001001SY1504	河	0088.pdf	地名	水			GJSH20210001001SY048
1530	GJSH20210001001SY1530	河	0089.pdf	地名	水			GJSH20210001001SY048
1535	GJSH20210001001SY1535	河	0089.pdf	地名	水			GJSH20210001001SY048
1594	GJSH20210001001SY1594	河	0092.pdf	地名	水			GJSH20210001001SY048
1598	GJSH20210001001SY1598	河	0092.pdf	地名	水			GJSH20210001001SY048
1601	GJSH20210001001SY1601	河	0092.pdf	地名	水			GJSH20210001001SY048
1605	GJSH20210001001SY1605	河	0092.pdf	地名	水			GJSH20210001001SY048
1611	GJSH20210001001SY1611	河	0092.pdf	地名	水			GJSH20210001001SY048
1621	GJSH20210001001SY1621	河	0093.pdf	地名	水			GJSH20210001001SY048
1626	GJSH20210001001SY1626	河	0093.pdf	地名	水			GJSH20210001001SY048
1973	GJSH20210001002SY0310	河	0019.pdf	地名	水			GJSH20210001001SY048
1990	GJSH20210001002SY0327	河	0019.pdf	地名	水			GJSH20210001001SY048
2143	GJSH20210001002SY0480	河	0026.pdf	地名	水			GJSH20210001001SY048
2153	GJSH20210001002SY0490	河	0027.pdf	地名	水			GJSH20210001001SY048
2161	GJSH20210001002SY0498	河	0027.pdf	地名	水			GJSH20210001001SY048
2167	GJSH20210001002SY0504	河	0027.pdf	地名	水			GJSH20210001001SY048
2300	GJSH20210001002SY0637	河	0032.pdf	地名	水			GJSH20210001001SY048
2315	GJSH20210001002SY0652	河	0033.pdf	地名	水			GJSH20210001001SY048
2433	GJSH20210001002SY0770	河	0038.pdf	地名	水			GJSH20210001001SY048
2441	GJSH20210001002SY0778	河	0038.pdf	地名	水			GJSH20210001001SY048
2444	GJSH20210001002SY0781	河	0038.pdf	地名	水			GJSH20210001001SY048
2457	GJSH20210001002SY0794	河	0038.pdf	地名	水			GJSH20210001001SY048
2471	GJSH20210001002SY0808	河	0039.pdf	地名	水			GJSH20210001001SY048
2477	GJSH20210001002SY0814	河	0039.pdf	地名	水			GJSH20210001001SY048
2583	GJSH20210001002SY0920	河	0044.pdf	地名	水			GJSH20210001001SY048
2609	GJSH20210001002SY0946	河	0045.pdf	地名	水			GJSH20210001001SY048
4122	GJSH20210001003SY0390	河	0030.pdf	地名	水	陽汙之山		GJSH20210001001SY048
4124	GJSH20210001003SY0392	河	0030.pdf	地名	水	凌門之山		GJSH20210001001SY048
4130	GJSH20210001003SY0398	河	0030.pdf	地名	水			GJSH20210001001SY048
4222	GJSH20210001003SY0490	河	0036.pdf	地名	水	崆峒山		GJSH20210001001SY048
4244	GJSH20210001003SY0512	河	0037.pdf	地名	水	鳥鼠同穴山		GJSH20210001001SY048
4276	GJSH20210001003SY0544	河	0038.pdf	地名	水	洛西山		GJSH20210001001SY048
4280	GJSH20210001003SY0548	河	0039.pdf	地名	水			GJSH20210001001SY048
4284	GJSH20210001003SY0552	河	0039.pdf	地名	水	井陘山		GJSH20210001001SY048
4413	GJSH20210001003SY0681	河	0045.pdf	地名	水	玄股		GJSH20210001001SY048
4945	GJSH20210001003SY1213	河	0069.pdf	地名	水	先轅大澤之		GJSH20210001001SY048

图 4 指向“黄河”的“河”的语义聚类

加工编号	术语名称	术语坐标	文件名称	一级类目	二级类目	三级类目	出处书名	地理层级	地名	关联项ID
GJSH20210001001SY0078	滑	[[{"fileName": "0021.pdf		地名	水名	山海经十八卷				/
GJSH20210001001SY0079	白玉	[[{"fileName": "0021.pdf		自然	矿名	玉	山海经十八卷		南山经·雒山·箕尾之	GJSH20210001001SY0025
GJSH20210001001SY0080	雒山	[[{"fileName": "0021.pdf		地名	山系		山海经十八卷	GJSH20210001001DM0002	雒山	GJSH20210001001SY0002
GJSH20210001001SY0081	招摇之山	[[{"fileName": "0021.pdf		地名	山名		山海经十八卷	GJSH20210001001DM0002	招摇之山	GJSH20210001001SY0003
GJSH20210001001SY0082	箕尾之山	[[{"fileName": "0021.pdf		地名	山名		山海经十八卷		箕尾之山	GJSH20210001001SY0074
GJSH20210001001SY0083	《雒山》神	[[{"fileName": "0021.pdf		人物	神名		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0084	龍	[[{"fileName": "0021.pdf		动物	兽名	龙	山海经十八卷			/
GJSH20210001001SY0085	禋毛	[[{"fileName": "0021.pdf		祭祀	祭法		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0086	瘗玉	[[{"fileName": "0021.pdf		祭祀	祭祀用玉		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0087	瘞	[[{"fileName": "0021.pdf		祭祀	祭法		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0088	糗	[[{"fileName": "0021.pdf		祭祀	祭祀用粮		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0089	糗	[[{"fileName": "0021.pdf		祭祀	祭祀用粮		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0090	壁	[[{"fileName": "0021.pdf		祭祀	祭祀用玉		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0091	稻米	[[{"fileName": "0021.pdf		祭祀	祭祀用粮		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0092	白蜜	[[{"fileName": "0021.pdf		植物	草名	茅	山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0093	席	[[{"fileName": "0021.pdf		祭祀	祭祀用器		山海经十八卷	GJSH20210001001DM0001	南山经·雒山·箕尾之	/
GJSH20210001001SY0094	南次二经	[[{"fileName": "0021.pdf		地名	山系		山海经十八卷	GJSH20210001001DM0001	南次二经	/
GJSH20210001001SY0095	柜山	[[{"fileName": "0021.pdf		地名	山名		山海经十八卷	GJSH20210001001DM0002	柜山	/

图 5 近义聚合和同义聚合的标引层次

二、地理方位标引

（一）对于地理方位的界定

《山海经》也是上古的地理文献，记载了古人认知中的世界。山经共分东、南、西、北、中五个区域，每个区域都有若干条山系，山系以“经”为名。《海外经》和《海内经》分别

指向两个区域：海外、海内，它们可能是以同心圆的位置分布的。海外区域和海内区域都按照东南西北四个方向划分为四个区域。《大荒经》所记载的大荒，是在《海内外经》记录的区域之外的地方，也分为东南西北四个区域。最后的《海内经》，记载的区域似乎位于《山经》和《海内外经》之间，目前看，有很多地名可以与中原地区周边的古代地名相对应。

本平台采用两层地图叠加的方法，底层地图是现在使用的实际地图，上层地图就是《山海经》所描绘的地图。《山海经》所描绘的地图，本期建设并未建设，但是构建专名与地图逻辑关系的标引是本期工作的重点。

（二）标引地理方位的方法

表 2 地理方位标引著录主要字段表：

序号	字段名称	必备性	著录内容
1	地名 ID 号	必备，不可重复	地理方位唯一编号
2	地理层级	必备，不可重复	地理方位层级编号
3	地名	必备，可重复	根据文献实际标引
4	自长（里）	有则必备，可重复	根据文献实际标引
5	参照物	有则必备，可重复	根据文献实际标引
6	方向	有则必备，可重复	根据文献实际标引
7	距离（里）	有则必备，可重复	根据文献实际标引
8	东至	有则必备，可重复	根据文献实际标引
9	西至	有则必备，可重复	根据文献实际标引
10	南至	有则必备，可重复	根据文献实际标引
11	北至	有则必备，可重复	根据文献实际标引
12	经纬度	有则必备，可重复	根据地名对应的今地经纬度标引

从《山经》到《海内经》，每一部分都有分区，每一个分区都有一个标记地点。所以，每一个标记地点都拥有一个唯一标识号“地理 ID 号”。

“地理层级”的一级为山系、国家、地标等。《山经》的每一个山系中都有数座山峰，为了表示它们之间的关系，我们用隶属的层级关系来表示。比如“南山经之首”是南山的第一条山系，山系的名称为“鵠山”，鵠山山系中又有九座山，各有其名。在标引时候，鵠山山系的层级为第一层，属于鵠山山系的山峰标引为二级层级。如“鵠山山系”的地理层级为“GJSH20210001001DM0001”，其所属的山峰招摇之山、堂庭之山、猿翼之山、柢阳之山、柢山、亶爰之山、基山、青丘之山、箕尾之山则依次标引为 GJSH20210001001DM0001—0001，GJSH20210001001DM0001—0002，……，GJSH20210001001DM0001—0009。由于《山海经》中详细记载了山系和山峰的自长、相距距离、走向、相连的山峰、四至等信息，所以一并标引在其相关地点之下。

地理ID号	地理层级	地名	自长（里）	参照物
GJSH20210001001DM0001	GJSH20210001001DM0001	離山	二千九百五十里	
GJSH20210001001DM0002	GJSH20210001001DM0001-0001	招搖之山		
GJSH20210001001DM0016	GJSH20210001001DM0001-0002	堂廡之山		招搖之山
GJSH20210001001DM0021	GJSH20210001001DM0001-0003	獫狁之山		堂廡之山
GJSH20210001001DM0025	GJSH20210001001DM0001-0004	柎陽之山		獫狁之山
GJSH20210001001DM0039	GJSH20210001001DM0001-0005	柎山		柎陽之山
GJSH20210001001DM0045	GJSH20210001001DM0001-0006	宣爰之山		柎山
GJSH20210001001DM0049	GJSH20210001001DM0001-0007	基山		宣爰之山
GJSH20210001001DM0055	GJSH20210001001DM0001-0008	青丘之山		基山
GJSH20210001001DM0069	GJSH20210001001DM0001-0009	箕尾之山		青丘之山
GJSH20210001001DM0086	GJSH20210001001DM0002	南次二經	七千二百里	
GJSH20210001001DM0087	GJSH20210001001DM0002-0001	柜山		
GJSH20210001001DM0101	GJSH20210001001DM0002-0002	長右之山		柜山

方向	距离（里）	东至	西至	南至	北至	经纬度
東	三百里					
東	三百八十里					
東	三百七十里					
東	三百里					
東	四百里					
東	三百里					
東	三里					
東	三百五十里	東海				
		長右	流黃		諸毗	

图 6 地理方位标引

三、图像标引

《山海经图赞》文献中含有大量图像，少数十八卷本《山海经》也有插图。对这些图像也有固定的标引方式。由于对古籍图像的标引工作尚无成熟的智能处理方式，图像标引主要依靠人工校验。人工校验时，根据古籍文献实际情况，科学合理确定著录内容是否符合古籍文献著录的规范，是否符合古籍文献的真实状况。图表人工校验的主要任务有：

人工校验项	人工校验内容	备注
图表类型	根据图表的类型进行判断	参考图表类型分类表进行人工校验
图表内容	分析图中所画事物为何种动物、植物或事物。	根据所绘内容，结合章节、文献内容进行判断，力求做到对图表内事物的一一对应。
图表链接	确认图表链接是否准确有效	数据阶段为抽检，数据导入平台后进行逐一人工校验

确认了图表之后，必须对其进行标引。对于图像来说，准确的定位和图像及图像对应文本所包含的信息都是十分重要的信息源（表 3）。

表 3 图像标引著录主要字段表：

序号	字段名称	必备性	著录内容
1	图像 ID 号	必备，不可重复	
2	图像名称	必备	

3	出处图像	有则必备	
4	出处页码	有则必备	
5	出处书名	有则必备	
6	出处书目 加工编号	有则必备	
7	出处版本	有则必备	
8	出处章节 编号	有则必备	
9	术语加工 编号	有则必备，可重复	对应的术语加工编号，可重复。一个字段对应一个术语加工编号。
10	图像描述	有则必备，可重复	此字段可重复，每个字段容纳一项元素描述。
11	图像路径	必备	

图像 ID 号是标引时给定的唯一识别号。图像名称按照原文照录。关于其出处的所有信息都由“出处图像”“出处页码”“出处书名”“出处书目加工编号”“出处版本”“出处章节编号”标引。如果图像名称与专名知识元有关，则在“术语加工编号”中填入相关专名知识元的唯一标识号。“图像描述”字段主要标引图像所蕴含的其他信息，是可以重复的。

图像ID	图像名称	出处图像	出处页码	出处书名	出处书目加工编号
GJSH20210048001TZ0012	羈耳國	{"x":"244","y":"913","w":"819","h":"1531"}	0100	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0013	毛民國	{"x":"206","y":"1338","w":"819","h":"1093"}	0101	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0014	鼎陽國	{"x":"288","y":"1150","w":"975","h":"1288"}	0102	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0015	忸人國	{"x":"50","y":"1444","w":"1356","h":"731"}	0103	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0016	小人國	{"x":"263","y":"1588","w":"1106","h":"506"}	0104	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0017	一臂民	{"x":"50","y":"1244","w":"1238","h":"1287"}	0105	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0018	三面人	{"x":"388","y":"1025","w":"687","h":"1494"}	0106	山海經四集十八卷雜	GJSH20210048001
GJSH20210048001TZ0019	釘靈國	{"x":"325","y":"1094","w":"850","h":"1369"}	0107	山海經四集十八卷雜	GJSH20210048001
GJSH20210048002TZ0001	狴狴	{"x":"35","y":"1025","w":"1240","h":"1510"}	0005	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0002	鹿蜀	{"x":"40","y":"1379","w":"1353","h":"1103"}	0006	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0003	類	{"x":"58","y":"1322","w":"1266","h":"1200"}	0007	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0004	搏訢	{"x":"145","y":"1195","w":"1219","h":"1102"}	0008	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0005	九尾狐	{"x":"56","y":"1214","w":"1242","h":"1269"}	0009	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0006	長右	{"x":"118","y":"1034","w":"1147","h":"1393"}	0010	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0007	滑羆	{"x":"124","y":"1039","w":"1202","h":"1388"}	0011	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0008	蜺	{"x":"191","y":"1399","w":"1180","h":"1107"}	0012	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0009	獬豸	{"x":"51","y":"1439","w":"1271","h":"939"}	0013	山海經四集十八卷雜	GJSH20210048002
GJSH20210048002TZ0010	蠪蛸	{"x":"57","y":"1143","w":"1303","h":"1272"}	0014	山海經四集十八卷雜	GJSH20210048002

出处章节编号	术语加工编号	图像描述	图像路径
GJSH20210048001PZ0009	GJSH20210001003SY0035	爲人耳長，行則以手攝持之，在無長國東	data/GJSH20210048001/Images/GJSH20210048001_0100_
GJSH20210048001PZ0009	GJSH20210001003SY0172	爲人身生毛，在海外東北	data/GJSH20210048001/Images/GJSH20210048001_0101_
GJSH20210048001PZ0009	GJSH20210001003SY0197	人面，長唇，黑身，有毛，見人笑亦笑，笑則唇掩其目	data/GJSH20210048001/Images/GJSH20210048001_0102_
GJSH20210048001PZ0009	GJSH20210001003SY0238	人面，魚身，無足，在建水西	data/GJSH20210048001/Images/GJSH20210048001_0103_
GJSH20210048001PZ0009	GJSH20210001003SY0629	人長九寸，在大荒東	data/GJSH20210048001/Images/GJSH20210048001_0104_
GJSH20210048001PZ0009	GJSH20210001003SY1183	一手，一足，一目，一鼻孔，在大荒之西	data/GJSH20210048001/Images/GJSH20210048001_0105_
GJSH20210048001PZ0009	GJSH20210001003SY1186	人頭，三邊各有面，無左臂，居大荒山	data/GJSH20210048001/Images/GJSH20210048001_0106_
GJSH20210048001PZ0009	GJSH20210001003SY1576	其從跡已，下有乚，馬蹄善走，在康居北	data/GJSH20210048001/Images/GJSH20210048001_0107_
GJSH20210048002PZ0002	GJSH20210001001SY0013	狀如禺而白耳，伏行人走，出招搖山	data/GJSH20210048002/Images/GJSH20210048002_0005_
GJSH20210048002PZ0002	GJSH20210001001SY0034	狀如禺而白首，其文如虎而赤尾，佩其皮宜子孫，出阻陽山	data/GJSH20210048002/Images/GJSH20210048002_0006_
GJSH20210048002PZ0002	GJSH20210001001SY0051	狀如狸而有黃，自爲牝牡，出置憂山	data/GJSH20210048002/Images/GJSH20210048002_0007_
GJSH20210048002PZ0002	GJSH20210001001SY0057	狀如羊，九尾，四耳，其目在背，出基山	data/GJSH20210048002/Images/GJSH20210048002_0008_
GJSH20210048002PZ0002	GJSH20210001001SY0064	狐身，九尾，能食人，出青邱山	data/GJSH20210048002/Images/GJSH20210048002_0009_
GJSH20210048002PZ0002	GJSH20210001001SY0098	狀如禺而四耳，見則大水，出長右山	data/GJSH20210048002/Images/GJSH20210048002_0010_
GJSH20210048002PZ0002	GJSH20210001001SY0116	狀如人而鰐鰕，音如柝木，見則其縣有蘇，出囊光山	data/GJSH20210048002/Images/GJSH20210048002_0011_
GJSH20210048002PZ0002	GJSH20210001001SY0115	狀如虎而牛尾，音如吠犬，是食人，出浮玉山	data/GJSH20210048002/Images/GJSH20210048002_0012_
GJSH20210048002PZ0002	GJSH20210001001SY0161	狀如羊，而無口，出海山	data/GJSH20210048002/Images/GJSH20210048002_0013_
GJSH20210048002PZ0002	GJSH20210001001SY0180	狀如雕而有角，是食人，出鹿耳山	data/GJSH20210048002/Images/GJSH20210048002_0014_

图 7 图像标引实例

四、知识图谱

（一）知识图谱构建

知识图谱的构建建立在知识元的基础上。知识元相对类型较多，且依附于古籍文献影像和全文数据。平台需要既可以在文献图像的维度上展示知识元，又能在全文数据的维度上展示知识元，还要能在地图维度上展示知识元。

利用字段进行跨库逻辑关联，是我们运用的主要处理方式。通过对关联字段的标准化和定向关联，形成知识元与文献图像、文献全文的相互关联，并在特定的检索条件下实现多库字段的抽取与展示。

本平台通过术语唯一识别号、术语坐标、出处章节、地理层级、图像 ID 号、图像坐标、图像路径、经纬度等字段实现知识元库与古籍图像库、古籍全文数据库之间的定位、跳转与信息链接。

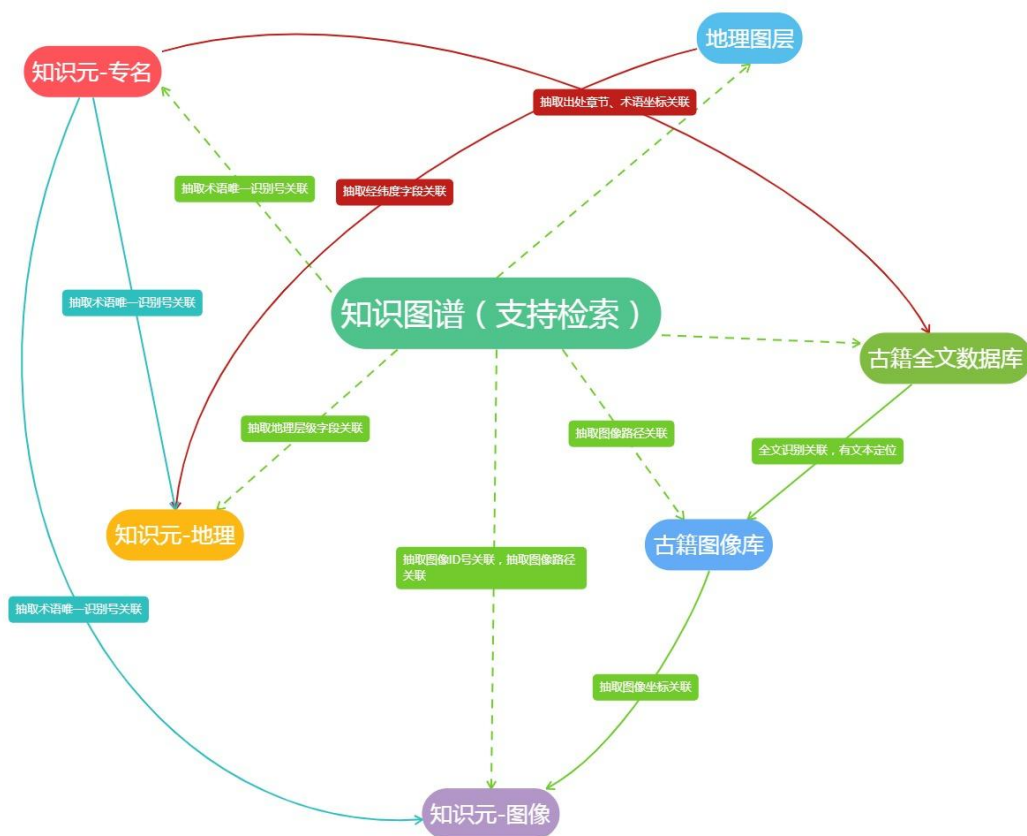


图 8 知识图谱关系图

（二）知识图谱检索的可视化

与传统检索不同，知识图谱检索提供的是一种关系网络，而非信息条目列表。在此基础上，展现检索词的相关检索结果的最好方式应该为可视化的检索结果。

为了形成简洁的可视化页面，检索结果的筛选也进行了初步的层级控制：优先展示检索词相关度最高的第一个知识元，并同时展示该知识元在上下文中出现的相关专名词，且标注它们与该知识元的关系。

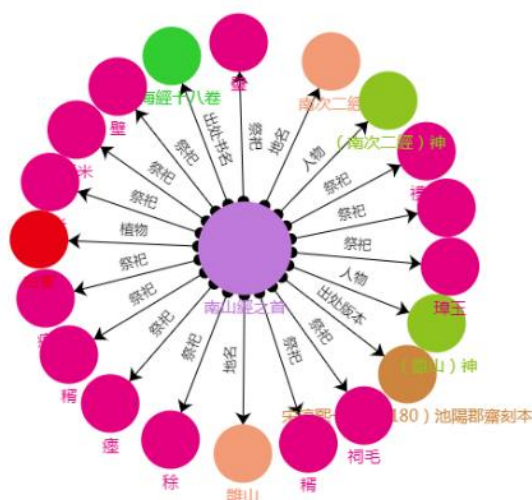


图 9 以“南山经”为检索词的第一次检索可视化页面

由于知识图谱检索可视化展示的工作仍在进一步探索当中,所以很多设想暂时并未在本平台中实现。

知识图谱检索还需要具有可探索性。所谓可探索性,即可以在该可视化页面中对部分重要相关的知识元安置超链接,使读者可以通过超链接跳转进入该知识元的独立页面。该独立页面中,应该展现该知识元的重要相关知识元以及它们之间的关系。比如,该知识元为地名,那么应该展现它所处的地理位置、其中所存在的物产,以及它的出处章节。又如,该知识元为人名,那么应该展现它所处的地理位置、人物的相关知识元、以及其特征描述。

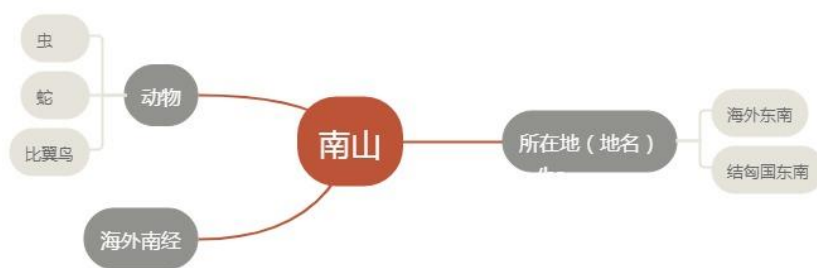


图 10 与地名知识元重要相关的知识元

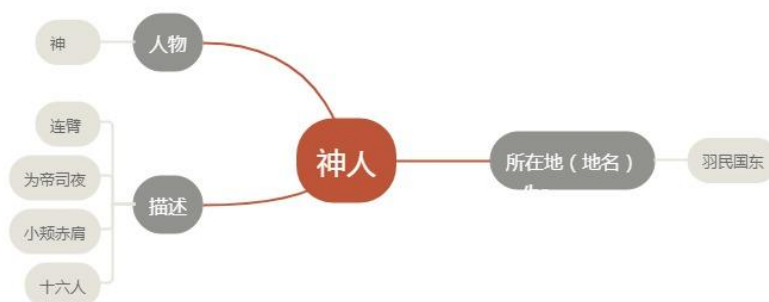


图 11 与人物重要相关的知识元

五、讨论

《山海经》知识库平台基于《山海经》古籍专题文献基础数据库，采用 SOA 体系，基于 HTML5 技术的 Web 前端设计与开发技术，开展网络应用系统建设。设计建设了多个模块，包括：文献导航与检索，古籍文本校勘比对，知识元检索与知识图谱索引，图片检索与图像识别，地理定位与标记等。同时，通过定位分层的全文识别技术、细粒度标引、知识图谱构建等技术手段，知识库平台对收录的 92 种《山海经》古籍的内容实现了文本、图像、地理信息在浏览、检索、数据分析方面的多维多向关联。

《山海经》知识库平台更注重专业人员的使用体验。知识库平台按照题名、作者、成书时代和文献分类等线索分类索引，组织分类导航树视图，便于读者定位浏览。平台提供图像索引、全文索引、图文对照索引、题名索引、作者索引、分类索引、专名索引、版本索引、成书时代索引，支持简繁通检和异体字通检，支持图像识别。这里不再一一介绍。

《山海经》的知识元标引和知识图谱的建设工作只是智慧图书馆建设工程的一个初步尝试，是在实践中不断学习的过程。如何优化知识图谱的可视化结果以及如何进一步提高知识元标引在不同场景下的使用，是我们需要继续研究的课题。这篇粗浅的小文也希望得到方家们的赐教。