《当代语言学》第 15 卷 2013 年第 2 期 142 – 161 页，北京

汉字知识的形式表达*

周亚民　台北大学　　黄居仁　香港理工大学

提要　汉字的知识本体和形式表达的研究不仅有助于计算机处理汉语，更能够突显汉字的特色和丰富知识内涵。本文旨在说明如何在计算机建立汉字知识，以及如何用形式语言表达汉字知识。与过去的汉字数据库不同的是，本研究以语意网的形式语言描述汉字的知识，希望能够对这方面的研究有所启发。汉字知识的形式表达内容包括：字形外在结构和演变的描述、意符与声符的描述、字形内在结构的描述、字义与衍生词的描述、异体字关系的描述、字音演变的描述、时间的描述，其中，意符和字义皆与 IEEE 建议上层共享知识本体(SUMO)对应，作为汉字知识的上层知识。本研究采用的形式语言是 OWL-DL，有助于汉字知识与其他知识本体分享知识。

关键词　汉字的形式表达　汉字知识本体　建议上层共享知识本体

1. 前言

计算机知识库核心包括内容和形式，内容是知识的领域和范围，而形式则是将内容转换成可以被计算机处理的表达方式。由于形式表达方式有助于知识整合与分享，知识库的建立逐渐发展为知识本体的建立。

知识本体有两个重要的特性：概念化和外显描述。概念化是一个过程，即将某个领域中具体和不具体的对象或实体形成概念，并描述概念与概念之间的关系，其实这就是建立模型(model)。知识本体必须使用形式语言表达概念和关系，常用的形式语言可以分为不采用语意网为基础和采用语意网为基础的语言，主要的差异是后者以 RDF(resource description framework)作为核心语言。不采用语意网为基础的包括 CycL(Lenat 1995)、KIF(Genesereth and Fikes 1992)、Ontolinqa(Farquhar，et al. 1997)。这些语言发展得较早，如果用这些语言表达知识本体，在网络环境的分享能力不如采用语意网为基础的语言。以语意网为基础的知识本体语言包括 RDF、RDFS、DAML + OIL、OWL(Gomez-Perez and Corcho 2002；Fensel 2000)。

对计算机而言，语言的处理是非常重要的问题，语言知识本体和形式表达

* 本文为台湾国科会补助专题研究计划部分成果，计划名称：国科会多年期计划 – 建立汉字知识本体第二阶段：中文计算机次常用字(NSC 96-2411-H-228-002-MY2)；承蒙《当代语言学》审查人与编辑的宝贵建议，修订了许多疏漏，特此一并致谢。

是协助解决问题的关键资源。由于语言的基本单位是词，大部分的语言知识本体都是以词为内容(Huang，et al. 2004；Miller 1995；Niles and Pease 2003)。如果要处理不同的语言，需有不同的语言知识本体做为资源。若要处理汉语，除了需要建立以词为内容的知识本体，还必需建立以汉字为内容的知识本体，因为汉字在语言中的角色不同于拼音文字。赵元任(1992 [1975]：246-8)说 "汉语是不计词的……字是中心主题，词则在许多不同的意义上都是辅助性的副题"，更有汉语语言学家因此提出字本位的理论，认为字才是汉语的中心(徐通锵 2000 [1997]；潘文国 2002；鲁川 2001)。

汉字的知识本体和形式表达是相当重要的研究，不仅有助于计算机处理汉语，在语言知识本体的研究领域，更能够突显汉字在众多书写系统之中，是极富有特色和丰富知识内涵的文字。因此，我们在 2005 年提出了汉字知识本体(Hantology，见周亚民、黄居仁 2006 和 Chou and Huang 2010)。本研究的目的是针对如何在计算机建立汉字知识，以及如何用形式语言表达汉字知识。本研究与过去的汉字数据库不同，以语意网的形式语言描述汉字的知识，希望能够对这方面的研究有所启发。

2. 研究方法与步骤

语料库与电子词典是建立语言知识本体的重要资源，但是，本研究缺乏符合需要的语料库和电子字典，必须依赖传统的字书。由于缺少适用的电子字典，在建立汉字知识本体的研究上投入了很长的时间。

本研究最重要的参考字书是《说文解字》(许慎 2004 [121]，以下简称《说文》)和《汉语大字典》(徐中舒 1992)。采用《说文》有几个原因：(1)本研究以《说文》部首做为汉字的基本意符；(2)《说文》的释义以本义为主，能够帮助我们找到部首与从属字的关系；(3)《说文》不仅释义也释形，部首的作用表现在字形，缺乏字形的解释，不容易找出部首的表意概念；(4)《说文》的收字以小篆为主，保留了较多造字的字形特征，较楷体字更容易分析部首在从属字形的功能。汉字知识本体的建立分为三个阶段进行，第一阶段是《说文》意符概念的建立，第二阶段是字形和异体字关系的建立，第三阶段建立字义和衍生词汇。

2.1 《说文》意符概念的建立

我们以《说文》的部首作为基本意符，找出意符的概念，这样能够保留《说文》部首的知识结构，有了最初的知识结构，再继续建立其他的知识也是比较可行的方法。《说文》部首作为意符所表达的概念，表现在部首与从属字之间的关系，这些关系必须从部首的字形和本义开始，才能找出从属字使用部首表达的是什么意义，因此汉字知识本体概念层的建立分为两个阶段：第一阶段分析部首的本义，第二阶段找出部首所统领之从属字的概念。

2.1.1 分析部首的本义

《说文》的释义皆以本义为主，虽然有些字因为甲骨文或其他考证资料的发现，可知许慎的解释并非造字的初义，但以《说文》释义探求汉字本义，仍然是文字学主要的方法。不过许慎对字义的解说都很简短，留下很大的诠释空间，不同的文字学家有不同的解释，如果根据的考证资料不同，差异就更多了。为了找出对《说文》较好的解释，本研究参考季旭升(2002)、李孝定(1992)、《汉语大字典》和蔡信发(2002)等的观点，取多数共同的解释作为部首的本义。

2.1.2 找出部首统领从属字的概念

经由部首本义的分析后，进一步的工作是找出从属字利用部首表达的意义是什么。许慎对部首的释义，并不一定是从属字运用部首所要表达的意义。例如《说文》解释“羊”为“吉祥”，实际上，该释义并不是“羊”的本义，“羊”的本义即是分类上属哺乳动物的羊，因为“羊”之甲骨文和金文皆像羊头(李孝定1992)，如果直接采用许慎的解释，将“吉祥”作为“羊”统领从属字的概念，这么做并不恰当。作为表意的概念，必须先找出部首较为可信的解释，还要分析从属字的部首表达的概念，两相比较后，才能决定部首作为意符的概念。

2.1.3 意符概念与建议上层共享知识本体的对应

分析《说文》的540部首与从属字所得到的意符概念，本研究将之对应到IEEE SUMO[①]。因为SUMO已经将当代知识的基本概念加以组织，如果对应到SUMO，也就是将意符概念的知识结构建立起来，藉由SUMO还可以找出特定概念的相关意符。

我们允许意符概念的分类对应到两个以上的SUMO概念，另外由于SUMO为阶层式的知识结构，越接近终端节点分类越细，越能反映细微的差异，与之相反，上层的节点则较无法区别细微的差异，所以，在分类时尽可能放到终端节点，如果无法找到适当的节点，则向上寻找直到找到适当的节点。为了区分SUMO概念与意符概念的差异，进行分类时我们考虑两种不同的分类情形:

第一，意符概念与SUMO概念为同义。例如“骨”的本义依《说文》的解释为“骨骼”，可以对应SUMO概念的实体/物质/物体/自体连结物/物质/混合物/体物质/组织/骨骼。这些意符在《说文》的解释不仅是本义，也是作为从属字的意符所表达的概念。另外一种情形是《说文》对意符的解释，不能找到同义的SUMO概念，但是意符所表达的概念可以对应同义的SUMO概念。

① SUMO(Suggested Upper Merged Ontology，建议上层共享知识本体)由IEEE标准上层知识本体工作小组所建立，作为不同知识本体的上层知识，可协助分享和交换知识(参见Niles and Pease 2001)。

第二，SUMO 节点的概念为广义。由于 SUMO 为各个领域知识的共同上层结构，用来涵盖各种领域的知识本体，SUMO 节点大部分皆为其他概念的广义概念。

2.2 字形结构知识的建立

2.2.1 六书知识的建立

六书是汉字字形结构的基石，也是剖析字形最传统的方法。依六书解析字形结构有时不一定能够得到一致的结果，因为同一个汉字可能会被不同的文字学家归类到不同的类别，本研究主要以《说文》的解释和体例作为分类的主要依据，若有不清楚的地方，请参考季旭升(2002)的研究。

2.2.2 描述字形的意符和声符

汉字的字形结构功能可以分为意符与声符，意符表示字形与意义有关连，而声符则表示字形与语音有关连。裘锡圭(1995〔1988〕)认为形符与意符是不同的。我们并不区分形符与意符，只要其功能与意义连结即是意符，因为形符与意符都通过形式表达概念，故形符就是意符，意符也是形符。

确认字形中哪个部分是意符与声符，最好依赖字书，自从《说文》开始在字书中加入解形，后来的字书大部分都偏重释义和声韵，即使近代的《康熙字典》以及《汉语大字典》，虽然加入了解形的部分，但是多数限于引用《说文》。判断形符与声符，不宜在楷体字中找，因为很多楷体字由于字形的变化，已经看不出形符与声符，必须从小篆着手，而《说文》小篆有助于找出正确形符与声符。

2.2.3 构字式

六书、意符和声符是汉字的内在结构，其分析需要依赖文字学的理论，事实上，也可考虑字形本身的部件结构，尤其是无法确认字形中的意符或声符时，可直接描述字形的结构。中研院信息科学所文献处理实验室已经将《汉语大字典》的 54678 字逐一表达了构字式(庄德明 2003；庄德明、谢清俊 2005)。本文采用该研究的构字式表示字形结构，而将重心放在构字式以外的汉字知识表达。

2.3 异体字关系的建立

本文采用裘锡圭(1995〔1988〕)对异体字的定义，即音义全部或部分相同而字形不同，而甲骨文、金文、籀文、小篆、隶书、楷书等不同阶段的文字不是异体字，因为其改变是全面性的，不是发生在某一个字形的改变。裘锡圭还将异体字关系分为全同与部分异体，例如 “十”与“拾”即为部分异体字，因为只有在数字时以互相使用，如果是“路不拾遗”，则不可写“路不十遗”。由于两个字形是否真的音义都完全相同，并不是这么容易证明，除非能穷举所有的使用情形，否则很难说两个字形是完全相同，因此，本研究以通用异体而非全同异体

表示两个字形大多数音义相同。

过去的异体字数据库，多数并未区分通用异体和部分异体，因为部分异体的知识建立十分复杂，而本文的特色之一，就是区别通用异体和部分异体。建立异体字关系的知识，本文分为下列几个面向:

第一，建立字书对异体字关系描述的知识

异体字关系的知识，基本的证据来自字书，但不同字书对异体字的描述体例不同，有些比较明确地表达异体字的关系，有些只能知道异体字的存在，但是并未说明是什么关系。本文判断异体字，需找到比较明确的关系，由于不同的字书描述不一定相同，所以，尽可能同时将多本字书的考证加入，得到较完整的轮廓。

第二，异体字的时间面向

此面向描述异体字的时间关系，彼此又有哪些共同的意义，主要根据除字书外，还需要例证。字书对于汉字的使用情形，与例证所提供的信息不同，前者只能确定若某字为字书所收，则此字应出现在该字书成书之前，但是此字于字书成书时是否仍在使用，则必需根据例证。本文以《汉语大字典》为主要字书，因为它不仅是字典而且有许多的例证资料，更包括近百年来的许多考古发现，对于建立异体字的时间关系非常重要。

第三，异体字字义面向

分析异体字之间的字义关系，必须要先确认本义，然而探求汉字的本义并不容易，但是只有掌握本义才能找出词义引申的脉络，以及汉字彼此之间的关系，从而进一步确认异体字的关系。由于确认汉字的本义有相当的困难，端看所发现的证据以及对这些证据所作的诠释，本文以《说文》的释义为主。

第四，异体字字音面向

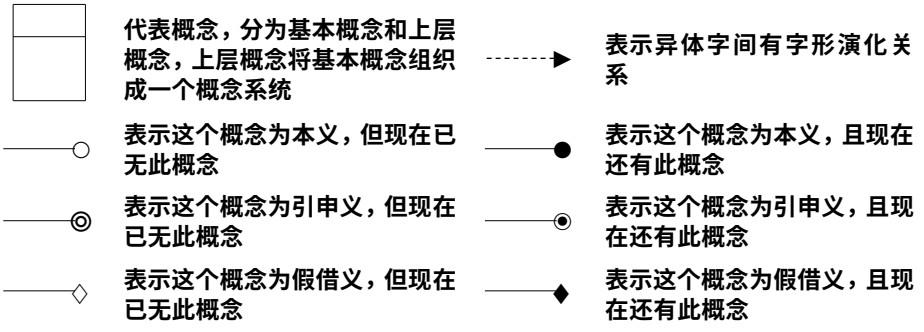
确定异体字关系，必需考虑字音，但不能以现代音作为判断的基础，而必须参考中古音和古音。由于古音的时间久远，考证困难，其中，韵部有较一致的研究成果，因此只先描述韵部，并根据《汉语大字典》的古韵为分部系统。中古音最有代表性的字书是《广韵》和《集韵》，因为同韵或同声纽可以使用不同的反切下字或上字，确认异体字不能只根据反切上字与下字，必须参考声纽和韵调。

2.4 异体字关系描述模型

由于异体字关系十分复杂，因此，本文开发两种不同分析角度的模型工具: GCD(glyph to concept diagram) 与 CGD(concept to glyph diagram) ，前者字形对应意义，反映异体字共同意义的变化，而后者则是意义对应字形，反映特定的意义曾经使用过哪些字形。

2.4.1 CGD

符号说明:



请看图 1 和图 2。

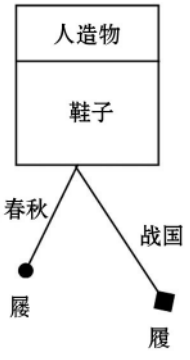


图 1　屣、履

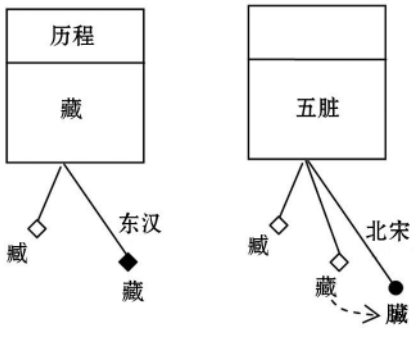


图 2　藏、脏

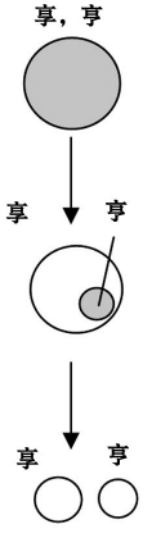
图 1 中，屣、履(有本字假借　“屣”的本义为“鞋子”，春秋时已使用“屣”表示“鞋子”的概念，到了战国时期，又假借“履”表示鞋子，为有本字的假借，这两个字到现在都还有此概念。“屣”的位置高于“履”表示使用“屣”表示“鞋子”的概念，较使用“履”早，而非造字的先后。

图 2 中，“收藏”的“藏”，先假借“臧”表示，到东汉时又借用“藏”，至今仍使用“藏”表示此概念，为无本字假借。“内脏”的概念原来也是假借“臧”，后又假借“藏”，为分担“藏”的词义，新增形符“月”成为“脏”，表示此概念，为本字后造字，“臧”和“藏”现在已无此概念。

2.4.2 GCD

GCD 描述特定时间的字形所具备的意义，主要的描述方法是利用集合。图 3 中的 GCD 集合图，函数 S 将字形对应到概念集合，可以更详细地了解不同时间“亨”与“享”的意义。周朝时“亨”的意义有: 通达、顺利、享受、祭祀、煮; “享”的意义有: 通达、顺利、享受、祭祀、煮，即“亨”与“享”在周朝的共同意义也是通达、顺利、享受、祭祀、煮。GCD 的圆形表式字义的集合，在周朝“亨”与“享”的字义集合是相同的。而唐朝时“享”与“亨”的意义都减少了，

“亨”的意义有通达、顺利，“享”则有通达、顺利、享受、祭祀的意义，两字只有在“通达、顺利”的意义时为异体字。现在“享”与“亨”则没有共同的意义，因此，在当代“亨”与“享”的字义集合没有交集。



S(亨|周朝) = {通达，顺利，享受，祭祀，煮}
S(享|周朝) = {通达，顺利，享受，祭祀，煮}
S(亨|周朝) ∩ S(享|周朝) = S(亨|周朝) = S(享|周朝) = {通达，顺利，享受，祭祀，煮}

S(亨|唐朝) = {通达，顺利}
S(享|唐朝) = {通达，顺利，享受，祭祀}
S(亨|唐朝) ∩ S(享|唐朝) = {通达，顺利}

S(亨|当代) = {通达，顺利}
S(享|当代) = {享受}
S(亨|当代) ∩ S(享|当代) = ψ

图3 “亨、享”的 GCD 集合图

2.5 字义与衍生词汇关系的建立

汉字由于使用时间长，本义经过引申或假借后产生丰富的字义。常用字由于使用频繁，累积的义项则更多。为减少研究的复杂度，我们先建立汉字的当代字义，表达当代用字的知识结构。由于大型字典所收义项虽多，但是不易区分哪些是现代使用的字义，而给小学生使用的国语小字典，多为现代常用义，比较符合本研究的需求。现代字义建立之后，逐个义项皆与 SUMO 对应，可与其他知识本体衔接。在汉字知识本体架构中，包括字义的衍生词汇，因此必须要知道每个字义产生哪些衍生词汇，大部分字典并没有作这样的区分，而国语小字典则将每个字义的生成词分别列出，因此，我们选择它作为字义与衍生词汇关系的基础。

3. 研究结果

3.1 说文意符的知识结构

本文将《说文》540 部首对应到 SUMO，发现大部分的意符表达的概念都是具体的，与汉字的特性相符合，也与其他书写系统的发展相似。文字学家认为甲骨文已是相当成熟的文字，而小篆较甲骨文更成熟，主要证据是看形声字的比例，这是由文字的结构出发。我们认为应该还可以从另一个角度来看文字是否成熟，即若某书写系统发展成熟，应该能够描述所需传达的基本概念，换句话说，如果大部分基本概念可以被记录，则该书写系统已经成熟。我们从这个

角度分析意符所呈现的知识结构，可以发现意符所表达的概念涵盖了动物、植物、基本物质、心理、疾病、宗教、饮食、交通、生死、数字、时间、交易、衣物、颜色、形状、量度单位、天体、地理、人造物、人际互动、音乐、人和身体部位、器官、感官、性别等。540 个意符就能表达这么广泛的概念，而《说文》部首很多都能找到对应的甲骨文，所以，从这个角度也可以证明甲骨文非常成熟，已经能够表达很多概念而不只是单纯的图画。

这些出现在 SUMO 结构中的意符概念是非常基本的核心概念，因为其他汉字都是由这些概念而扩展。例如由人的概念(SUMO 概念为“人类”) 扩展到描述人的行为和德性; 从手的概念(SUMO 概念为“躯体部件”) 扩展到描述手的动作。呈现在 SUMO 的意符概念，其构字的能力并非完全相同，因为有些意符只有几个从属字甚至没有从属字，有些则有数以百计的从属字，意符构字能力最高的是有关植物的概念，其次是躯体部件，第三是有关人类的概念。这些意符使用频次较高，可推论应是生活上有实际的需要所致。

我们还发现，意符与从属字所形成的知识结构是一个具有高度衍生力的知识体系。这个知识体系符合 Pustejovsky(1995) 提出的衍生词汇理论，此理论解释了词汇衍生的环境-经验结构(qualia structure) ，此结构分为四个面向: 物质(formal) 、组成(constitutive) 、功用(telic) 、产生(agentive) 。由研究结果来看，部分的意符概念结构具有与衍生词汇相似的衍生能力。

3.2 汉字知识的形式表达

汉字知识的形式表达，我们采用的是 OWL(web ontology language) -DL。这种表达方式有助于将来在网络的应用以及与其他知识本体分享知识。我们将汉字知识的形式表达分为字形描述、六书、意符与声符描述、字义和构词、字音、异体字和时间的描述等，分别加以说明。OWL-DL 的语法和描述知识的方式，请参考 W3C 的技术报告②。

3.2.1 字形的描述

字形描述包括字形外在空间结构、内在结构(六书) 、部首、Unicode 的字形、是否有异体字，与字形相关的主要概念类别是 Glyph(字形) ，所有的字形都以 Glyph 类别描述。Glyph 类别以 hasGlyphExpression relation 描述字形的构字式，此关系在汉字知识定义为 InverseFunctional，InverseFunctionalProperty 在 OWL 描述的是下列的规则:

If P(y,x) and P(z,x) then y = z

② OWL - DL 的语法，以及如何描述概念和关系，请参考 W3C 的技术报告 http: //www. w3. org/TR/owl-guide/ [accessed 5, Aug. 2012]

由于 hasGlyphExpression 是 inversefunctionalproperty，那么如果两个字形个体有相同的构字式，这两个字形个体是同一个字，即:

if hasGlyphExpression(G_i , GE_k) and hasGlyphExpression (G_j , GE_k) then $G_i = G_j$

由于不同编码系统的字形不完全相同，汉字知识以 Unicode 作为编码系统字形的来源。Unicode 收有汉字字形的异写字，汉字知识将 hasGlyphInUnicode 关系定义为 Functional，OWL 的推理规则可以根据这个定义推理两个不同 Unicode 的字形为异写字，其推理规则为: If P(x,y) and P(x,z) then y = z。故如果一个字在 Unicode 中有一个以上的字码，则这几个 Unicode 表示的是同一个字。

字形描述包括这个字形是否有楷书、小篆、金文、甲骨文，分别以 Kaishu、LesserSeal、Bronze、OracleBone 类别表达。两个字形如果有字形演变关系，则以 hasAncientGlyph 和 isAncientGlyphOf 建立字形之间的古今演变关系，彼此为反关系(inverse relation) ，且都具有递移关系。

字形描述很重要的是指出字形结构哪一个部分是意符和声符，分别利用 hasSemanticSymbol 关系和 hasPhoneticSymbol 关系描述。hasSemanticSymbol 关系建立 Glyph 与 SemanticSymbol 的关系，由于一个字形可以有很多不同的意符，因此，并没有限制其意符的个数。

描述字形结构哪个部分是声符则是利用 hasPhoneticSymbol 建立 Glyph 与 PhoneticSymbol(声符) 关系，PhoneticSymbol 也有 hasGlyphExpression 的关系描述，因此如果 SemanticSymbol 的构字式与某个 PhoneticSymbol 的构字式相同，由于 hasGlyphExpression 关系为 InverseFunctional，即可知道意符也是声符。

汉字知识的形式表达在字形描述的部分还有造字方法和部首，字形的造字方法是透过 principleOfFormation 建立 Glyph 与 LiuShu(六书) 的关系来描述该字形造字的方法; 而 hasRadical 关系则描述字形的部首，每个汉字字形以 hasVariant 建立与其他异体字的关系。

3.2.2 意符与声符的描述

字形的意符以类别 SemanticSymbol 表达，意符的字形描述方式也是利用 hasGlyphExpression 与 hasGlyphInUnicode 两个关系。每个 Glyph 的 SemanticSymbol 由 hasSemanticSymbol 建立关系，每个意符所衍生的字形，则由 hasDerivedGlyph 关系建立。意符所表达的概念透过 hasConcept 与 conceptOfSUMO 两种方式描述，hasConcept 直接描述意符的概念，所建立的是 SemanticSymbol 与 XML Schema String Data Type 的关系，意符概念与 SUMO 的对应则是由 conceptOfSUMO 关系描述。因为意符表达的概念在 SUMO 不一定能找到同义的概念，只能

找到上位概念，因此，为了精确表达意符概念，除了以 conceptOfSUMO 关系描述意符概念，另外再由 hasConcept 描述意符概念。

由于每个意符表达的概念对应不同，无法直接描述特定的 SUMO 概念与意符概念的关系，因此汉字知识的形式描述，将意符对应到 SUMO 的最上层概念 Entity。由于所有 SUMO 概念都是 Entity 的 subclass，都继承 Entity，因此每个意符所对应的 SUMO 概念，可以分别对应到各别的 SUMO 概念。conceptOfSUMO 是很重要的关系，它可以将汉字知识对意符的描述与 SUMO 之间建立连结，分享 SUMO 的知识。

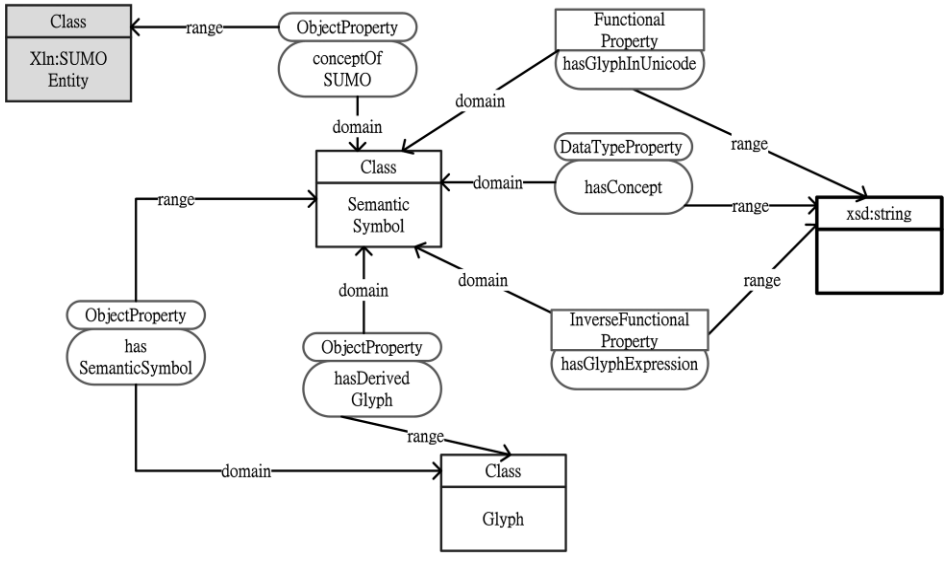


图 4　汉字知识意符描述的 OWL 语意模型

字形的声符以 PhoneticSymbol 概念类别表达，其字形结构也是以 hasGlyphExpression 描述，并用 hasGlyphInUnicode 描述声符在 Unicode 中的具体字形。每个字形的声符由 hasPronunciation 建立 Glyph 与 PhoneticSymbol 的关系，每个声符所表示的声韵调，则是透过 hasPronunciation 建立 PhoneticSymbol 和 Pronunciation 的关系。由于声符多是可以独用的字形，字音随着时间改变发生变化，Pronunciation 概念类别有 Ancient Pronunciation，Middle Pronunciation 和 Modern Pronunciation 三个 subclass 表达声符的古音、中古音和现代音。由于声符可能为多音字，因此 hasPronunication 可以建立一个 PhoneticSymbol 与多个 Pronunciation 的关系。

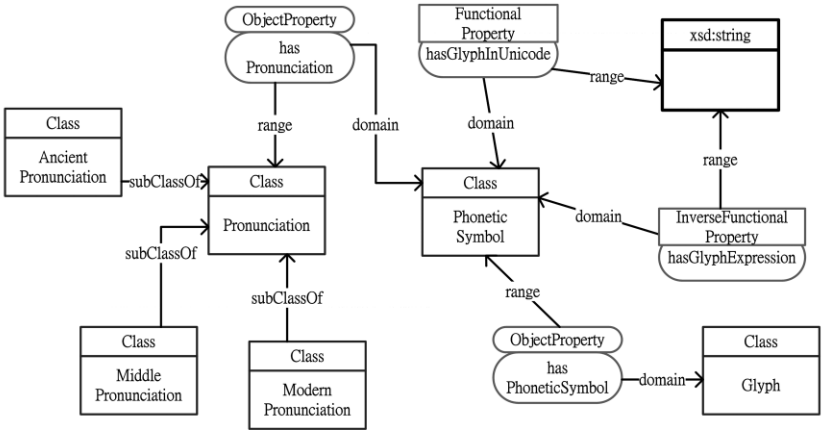


图 5　汉字知识声符描述的 OWL 语意模型

3.2.3 六书的描述

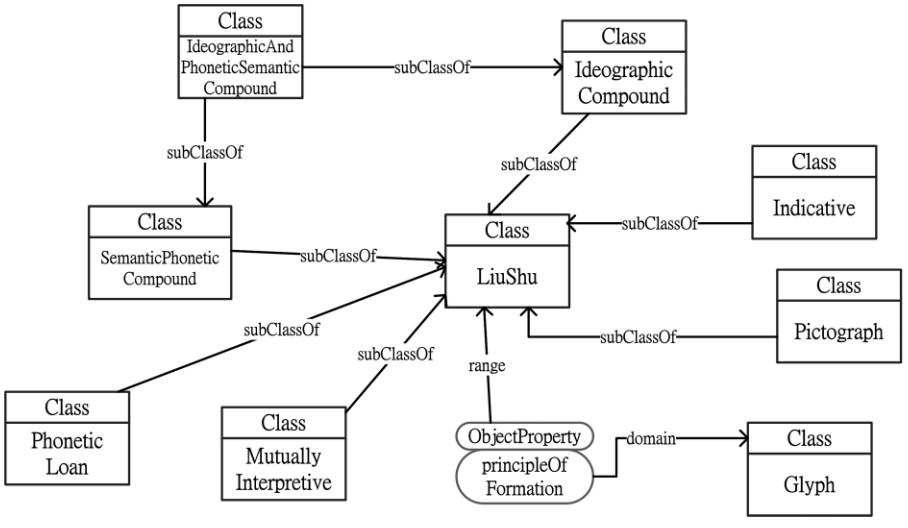


图 6　汉字知识六书描述的 OWL 语意模型

六书在汉字知识的形式表达中以 Pictograph、Indicative、Ideographic Compound、SemanticPhonetic Compound、Mutually Interpretive、Phonetic Loan **六个概念类别依序表示象形、指事、会意、形声、转注和假借，这六个概念类别都是** LiuShu **概念类别的** subclass**，由于《说文》有很多的字都是会意兼形声，因此，又增加一个** Ideographic and PhoneticSemantic Compound **概念类别，多重继承** Ideographic Compound **与** SemanticPhonetic Compound **概念类别，表达会意兼形声同时具备会意与形声字的特性。每个字形的造字方法，则是由** principleOfFormation **建立字形与六书的关系。**

3.2.4 字义与衍生词汇的描述

字义的描述需要的概念类别和关系很多，其中最主要的概念类别是 Sense，并且以 hasSense 建立 Glyph 与 Sense 的关系。由于不是所有的汉字都有字义，因此在形式描述中没有限制必需要有字义，有的汉字有多义，古汉语尤甚，所以也没有限制最多只能有一个字义。字义透过 meaning 关系和 conceptOfSUMO 关系表达，其中 conceptOfSUMO 将 Sense 与 SUMO 概念建立对应。由于每个字义对应的 SUMO 概念并不相同，因此，OWL 模型将 Sense 对应到 SUMO 的最上层 Entity 概念，而实际上每个字义再分别对应真正的 SUMO 概念。

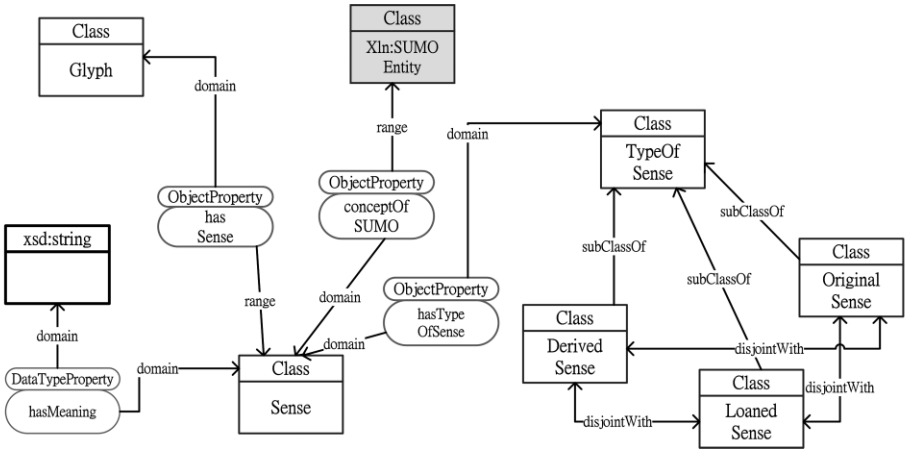


图 7　汉字知识字义描述的 OWL 语意模型(1)

每个字义使用的上下文关系和时间关系，由 Linguistic Context 概念类别表达，如果有多个上下文关系，则 hasLinguisticContext 可以建立多个 LinguistContext 关系。上下文由 Text 概念类别描述，上下文是实际的短语或句子用例，利用 hasCitation 建立 Linguistic Context 与 Text 之间的关系，并由 source 描述短语或句子用例的出处，时间则由 TimeInterval 表达，以 timecoverage 建立 Linguistic Context 的时间关系，以描述字义的使用时间。如果该字义收在字书中，则由 RecordOfZiShu 表达字书，并由 source 描述字书的名称，timecoverage 描述字书的成书时间。

不同字义的衍生词汇主要由 Word 概念类别描述，并由 hasSense 建立 Sense 与 Word 的关系。Word 以 OWL 的 SameAs 关系描述它与 SUMO 的 Word 是相同的概念。Word 有 Simple Word、Derived Word 和 Compound Word 三个子类别，分别描述单纯词、派生词和复合词，由于不同时间的衍生词汇不同，因此以 timecoverage 建立 Word 与 TimeInterval 的关系来描述衍生词汇的时间关系。对于没有词素义或词义的字形，它可以用连绵词等单纯词，以 isPartOfSimpleWord 建立 SimpleWord 与 Glyph 的关系。Compound Word 以 OWL 的 SameAs 关系描述它与

GOLD 的 OrthCompound 是相同的概念，SimpleWord 也以 SameAs 关系描述 SimpleWord 与 GOLD 的 SimpleOrthWord 是相同的概念。

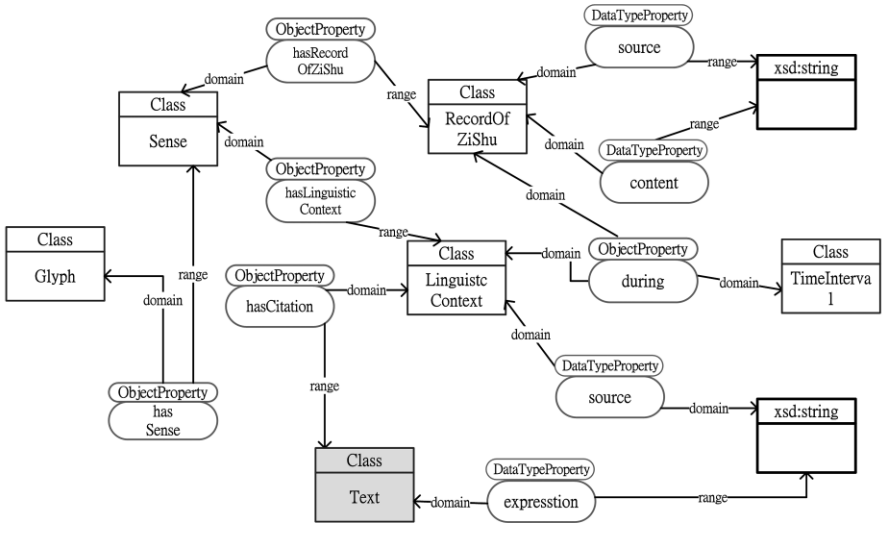


图 8　汉字知识字义描述的 OWL 语意模型(2)

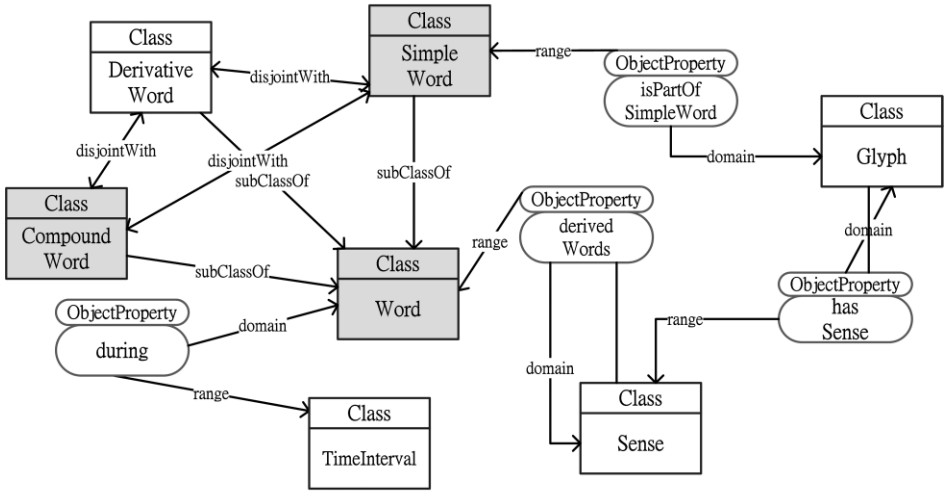


图 9　汉字知识衍生词汇描述的 OWL 语意模型

3.2.5 异体字关系的描述

异体字主要由 Variant 概念类别、Variant Relation 概念类别和 TypeOfVariant Relation 概念类别描述，由于异体字关系复杂，因此汉字知识以 Variant Relation 概念类别表示异体字关系。

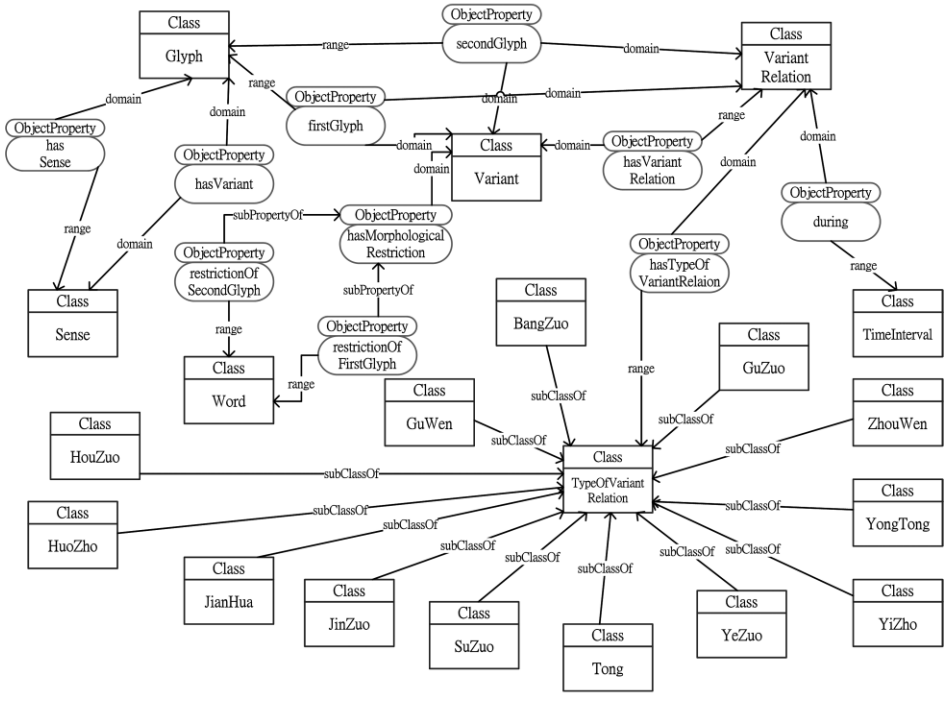


图 10　汉字知识异体字描述的 OWL 语意模型

通用异体字关系由 hasVariant **直接建立** Glyph **与** Variant **的关系，若为部分异体字则建立** Sense **与** Variant **的关系。** Variant **概念类别描述两个字形存在着异体字关系，但是并不描述异体字关系的类别，而是由** Variant Realation **描述。如果有多个异体字，可以有多个** Variant **分别描述不同的异体字。** Variant Relation **的异体字关系分为** GuZuo、BenZuo、GuWen、HouZuo、HuoZuo、JinZuo、SuZuo、Tong、YeZuo、YiZuo、YongTong、ZhouWen、JianHua **等概念类别，依序表示古作、本作、古文、后作、或作、今作、俗作、同、也作、亦作、用同、籀文和简化关系。如果异体字有构词的限制，则由** hasMorphololgicalRestriction **建立异体字与词汇的关系。**

3.2.6 字音的描述

字音由多个概念类别与关系描述，其中最主要的是 Pronunciation **概念类别。由于汉字音随义转，音与义的关系利用** hasPronunciation **建立，对于声符也是以** hasPronunciation **描述声符。**

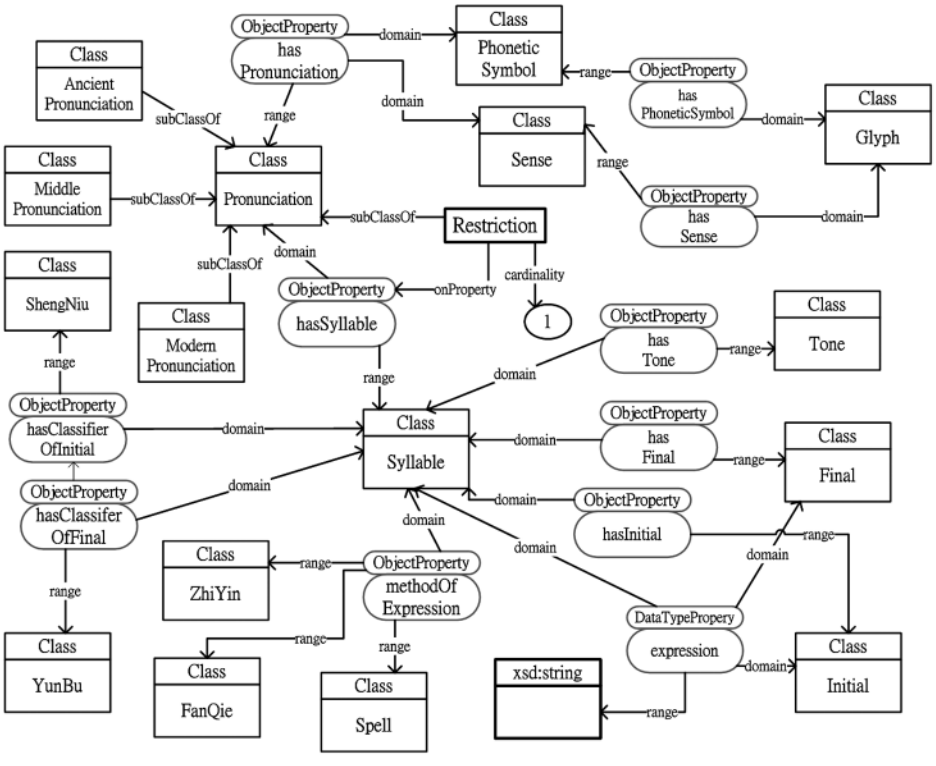


图 11　汉字知识字音描述的 OWL 语意模型

汉字为单音节，以 Syllable **概念类别表达音节，每个音节由声母、韵母和声调构成，分别以** Initial、Final **和** Tone **概念类别表达，一个音节可以没有声母，但是必定有韵母和调。上古音和中古音大多使用声纽和韵部表示，分别以** ShengNiu **和** YunBu **表达声纽和韵部，而不同的文字学家或声韵学家所使用的声纽和韵部不同，因此，**ShengNiu **和** YunBu **皆以** methodOfClassification **描述所使用的方法。由于表音的方式不同，汉字知识利用** JhihYin class、FanQie class、Spell class **分别表示直音、反切和拼音法。**

由于汉字为单音节，因此以 Restriction **关系对** hasSyllable **加以限制只能由一个** Syllabl **构成** Pronunciation。Pronunciation **有** Ancient Pronunciation，Middle Pronunciation，Modern **等三个子类别，分别描述古音、中古音和现代音，由于这三个** class **都是** Pronunciation class **的** subclass，**全部都继承** pronunciation class **的** hasSyllable property，**因此** pronunciation **可以描述古音、中古音和现代音不同时期的声韵调。**

3.2.7 **时间的描述**

汉字知识中对于时间的描述都是透过 TimeInterval **概念类别，目前我们对于时间的划分以朝代为单位，由夏到现代共分为** 45 **个时段，每个时段采用建立**

TimeInterval 的 Instance 方式(即 OWL 的 Individual)。TimeInterval 有 begin，end，dynasty 三个关系，分别描述时段的起始公元年、结束公元年、朝代名称。在汉字知识中需要建立时间关系的包括 Variant Relation、Word、RecordOfZiShu、LinguistcContext，这些概念类别都使用 timecoverage 关系描述时间关系。

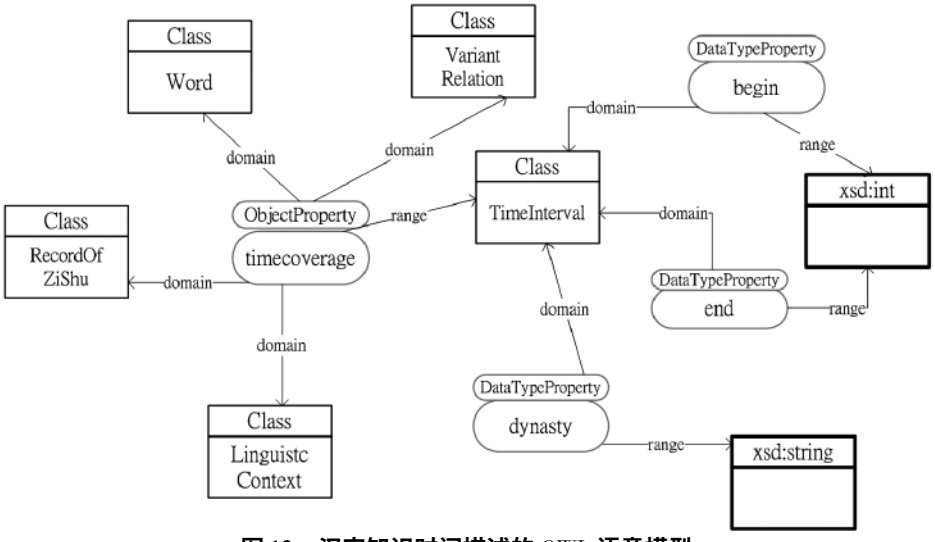


图 12　汉字知识时间描述的 OWL 语意模型

描述汉字知识的形式语言模型见图 13(插页)。

Hantology 的 OWL 形式语言描述(部分)

```
<? xml version = "1.0" ?  encoding = "UTF-8" ? >
<rdf: RDF  xmlns = http: //www. ntu. edu. tw/2005/01/Hantology. owl#
  xml: base = "http: //www. ntu. edu. tw/2005/01/Hantology. owl" >
xmlns: gold = "http: //www. emeld. org/gold. owl#"
        xmlns: j. 0 = "http: //protege. stanford. edu/plugins/owl/protege#"
        xmlns: rdf = "http: //www. w3. org/1999/02/22-rdf-syntax-ns#"
        xmlns: sumo = "http: //reliant. teknowledge. com/DAML/SUMO. owl#"
        xmlns: rdfs = "http: //www. w3. org/2000/01/rdf-schema#"
        xmlns: owl = "http: //www. w3. org/2002/07/owl#"
        xmlns: goldlinguistics = "http: //www. owl-ontologies. com/unnamed. owl#"
  <owl: Ontology rdf: about = "" >
    <owl: imports rdf: resource = "http: //protege. stanford. edu/plugins/owl/protege" / >
    <owl: imports rdf: resource = "http: //reliant. teknowledge. com/DAML/SUMO. owl" / >
    <owl: imports rdf: resource = "http: //www. owl-ontologies. com/unnamed. owl" / >
    <owl: versionInfo > 2005-01-01，edited by Ya-Min Chou < /owl: versionInfo >
  < /owl: Ontology >
<! - - Glyph - - >
```

```
< owl: Class rdf: ID = " Glyph" >
    < rdfs: label > 字形 < /rdfs: label >
< /owl: Class >
< owl: Class rdf: ID = " Pictograph" >
    < rdfs: subClassOf rdf: resource = " #LiuShu" / >
    < rdfs: label > 象形 < /rdfs: label >
< /owl: Class >
  < owl: Class rdf: ID = " Indicative" >
    < rdfs: subClassOf rdf: resource = " #LiuShu" / >
    < rdfs: label > 指事 < /rdfs: label >
< /owl: Class >
< owl: Class rdf: ID = " IdeographicCompound" >
    < rdfs: label > 会意 < /rdfs: label >
    < rdfs: subClassOf >
      < owl: Class rdf: ID = " LiuShu" / >
    < /rdfs: subClassOf >
  < /owl: Class >  < owl: Class rdf: ID = " SemanticPhoneticCompound" >
    < rdfs: label > 形声 < /rdfs: label >
```

4. 结论

对于汉字知识本体的研究，本文提出了开创性的想法，根据这些想法也具体地建立了汉字知识本体。研究进行过程中，最大的困难是必须要先建立知识工程、书写系统、文字学、声韵学、语言学等不同领域的知识，愿意参与的研究人员很少，本研究累积了近十年的投入，才有目前的结果。与过去的研究比较，本研究有下列的贡献:

第一，提出表达汉字书写系统的模型和形式

汉字的知识的重要性，可以藉由 WordNet 对自然语言处理的影响更清楚地被了解，因为词汇的知识是语言处理的基本知识，计算机处理自然语言的根本就在于语言形式与语意的关系，而词汇是语意的基本单位，但是对于汉语而言，除了应重视词汇知识，更应重视汉字知识。

汉字有丰富的知识，如果要表达汉字的知识，如何将汉字的知识抽象化、概念化？如何建立模型？如何表达？表达的形式是什么？本研究回答了这些问题，我们设计了一个汉字知识模型和表达形式，这个模型能够表达汉字不同断代的字形结构、意符、声符、本义、引申义、假借义、衍生词汇、古今字、正俗字、通假字、简化字之间互相交错的关系，还可以做为汉字知识的平台，可以利用这个架构将汉字知识不断输入计算机。

第二，《说文》意符的知识架构

本研究是一次将汉字意符知识与上层知识本体整合的研究，不仅表达《说文》意符的概念，更将这些概念之间的关系建立，呈现《说文》意符的知识架构，突破了传统文字学对汉字的分类限制，建立了意符的知识架构，并发现意符具有与衍生词汇类似的概念衍生能力。本研究与过去《说文》部首的概念分类研究比较，有下列优点：(1) 分类的互斥性较佳，好的分类架构应该满足互斥性与完整性，如果类别定义没有重叠或模棱两可即满足互斥性，完整性则要求任何待分类的对象都可以分到其中一类，不会找不到适当的类别加以归类。(2) 分类较为详细，SUMO 有 900 多个类别，可以作更细微的区别。过去的《说文》部首概念分类，其类别大部分只有十多类，许多部首的概念都被放在同一类，因而无法区分不同的《说文》部首概念。但是，如果用 SUMO 加以分类，将会归类到不同的类别，可以将它们区别开来以增加鉴别率。(3) 知识的分享力较佳。如果要增加知识的分享力，必须有一个大家接受的知识描述架构，所有的知识都利用这个架构为基础。对于计算机而言，更需要共同的知识结构才能让不同的计算机共享资源。目前 IEEE SUMO 即扮演这个角色，许多领域知识都开始与它连接，使不同的领域知识得以共享。我们以 SUMO 作为知识描述的基础，可以得到较高的知识分享力。

第三，提出异体字关系的表达架构

异体字的关系非常复杂，不同的文字学家从不同的角度看异体字，有的从古今字、有的从正俗字、有的从字的分化与合并研究、有的从字形结构，但是这些角度彼此交织而非互斥，而且并不是从同一个特性来看异体字关系。如何找出一个架构能够有系统地表达复杂的异体字关系，因为，它有字形、声韵、字义、时间、构词的交互作用。我们比较了本研究与其他异体字关系表达，证明汉字知识本体在异体字关系表达的优越性，能够以一个架构掌握异体字的关系，对于异体字关系的表达是很大的进展。

第四，提出汉字与时间关系的表达方法

语言会随着时间的推移而发生变化，共时与历时研究都是语言学的重要研究方法。历时的语言研究需要大量的语料，而这些语料如果没有描述语言与时间的关系，就难以比较语言的变化。本研究表达了字形、声韵、字义、异体字关系、衍生词汇的变化，而且彼此之间的关系都被描述，不是被切分为独立的语言学单位然后描述其变化。藉由汉字知识本体可以描述特定断代某个汉字的字形，该字形当时有什么字义，不同的字义在当时其声韵为何，不同字义当时有什么衍生词汇，而不是分别描述字形、声韵、字义、异体字关系、衍生词汇的变化。每一个汉字都应该要视为一个整体，分析和表达都要描述形音义以及与异体字和衍生词汇的关系，才能将汉字的知识结构表达清楚。

引用文献

Chou，Y. M（周亚民）. and Huang，C. R（黄居仁）. 2010. Hantology：Conceptual system discovery based on orthographic convention. In C. Huang，N. Calzolari，A. Gangemi，A. Lenci，A. Oltramari，and L. Prevot，eds.，*Ontology and the Lexicon：A Natural Language Processing Perspective*. Cambridge：Cambridge University Press. Pp. 122 – 43.

Farquhar，A.，R. Fikes，and J. Rice. 1997. The ontolinqua server：A tool for collaborative ontology construction. *International Journal of Human-Computer Studies*. 46，6：707 – 28.

Fensel，D. 2000. The semantic web and its languages. *IEEE Intelligent Systems* 15，6：67 – 73.

Genesereth，M. R. and R. E. Fikes. 1992. *Knowledge Interchange Format*. Version 3.0 Reference Manual，Knowledge Systems Laboratory，Stanford University，Stanford，CA.

Gomez-Perez，A. and O. Corcho. 2002. Ontology languages for the semantic web. *IEEE Intelligent Systems* 17，1：54 – 60.

Huang，C. R.，R. Y. Chang（张如莹），and S. B. Lee（李祥宾）. 2004. Sinica BOW（Bilingual Ontological WordNet）：Integration of bilingual WordNet and SUMO. 4th International Conference on Language Resources and Evaluation（LRE2004），Lisbon，Portugal.

Lenat，D. B. 1995. CYC：A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38，11：33 – 8.

Miller，G. A. 1995. WordNet：A lexical database for English. *Communications of the ACM* 38，11：39 – 41.

Niles，I. and A. Pease. 2001. Towards a standard upper ontology. In C. Welty and B. Smith，eds.，*Proceedings of the* 2*nd International Conference on Formal Ontology in Information Systems*（*FOIS*-2001）. Ogunquit，ME.

———. 2003. Linking lexicons and ontologies：Mapping WordNet to the SUMO ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*. Las Vegas，NV.

Pustejovsky，J. 1995. *The Generative Lexicon*. Cambridge，MA：The MIT Press.

蔡信发，2002，《说文部首类释》。台北：台湾学生书局。

黄居仁、陈圣怡、周亚民，2010，语言的知识与知识的语言：由说文解字出发的知识本体研究。见《研究之乐：王士元先生七十五寿辰学术论文集》。上海：上海教育出版社。106 – 22 页。

季旭升，2002，《说文新证》。台北：艺文印书馆。

李孝定，1992，《读说文记》。台北：中研院历史语言研究所。

鲁　川，2001，《汉语语法的意合网络》。北京：商务印书馆。

潘文国，2002，《字本位与汉语研究》。上海：华东师范大学出版社。

裘锡圭，1995［1988］，《文字学概要》。台北：万卷楼图书有限公司。

徐通锵，2000［1997］，《语言论——语义型语言的结构原理与研究方法》。长春：东北师范大学出版社。

徐中舒 主编，1992，《汉语大字典》。台北：建宏出版社。

许　慎 著，徐铉校订，2004［121］，《说文解字》。北京：中华书局。

赵元任，1992［1975］，汉语词的概念及其节奏。《中国现代语言学的开拓与发展：赵元任语言学论文选》。北京：清华大学出版社。

周亚民、黄居仁，2006，汉语文字和词汇知识在计算机的表达——历史变迁的观点。见何大安编《语言暨语言学》专刊《山高水长：丁邦新先生七秩寿庆论文集》。台北：中研院。595－611 页。
庄德明，2003，汉字智能型编码系统的现况与进展。汉字智能型编码与应用研讨会，中研院，台北。
庄德明、谢清俊，2005，汉字构形数据库的建置与应用。汉字与全球化国学术研讨会，台北。

第一作者 简介

周亚民，男，博士，台北大学中文系助理教授。研究兴趣：汉字知识本体、数字典藏、语料库。代表作 “中日汉字知识库：汉字传播与扩散观点”和“Hantology：Conceptual system discovery based on orthographic convention”。电子邮件：milesymchou@ gmail. com

CHOU Yamin，male，Ph. D. ，is an assistant professor at the Department of Chinese Literature，Taipei University. His research interests include Hanzi ontology，digital archive，corpus. His major publications are：“The knowledgebase of Chinese and Japanese characters：Based on dissemination and spread perspective”and“Hantology：Conceptual System Discovery Based on Orthographic Convention”. E-mail：milesymchou@ gmail. com

作者通讯地址：23741 台湾新北市三峡区大学路 151 号人文大楼 7 楼 台北大学中文系

全国汉语方言学会第十七届学术年会通知

全国汉语方言学会第十七届学术年会暨汉语方言国际学术讨论会将于 2013 年 12 月 13 日至 15 日在中国广州暨南大学举行，现向全国汉语方言学会全体会员征集论文，并邀请海内外研究汉语方言的专家学者参会。

有意与会者请先提交电子版论文(或提要)，邮件地址为：hyfyguangzhou@ 163. com；ncdagz17@ yahoo. cn。纸质论文请寄：广州暨南大学汉语方言研究中心，邮编 510632。或：北京建国门内大街五号 610 室，全国汉语方言学会秘书处，邮编 100732，并在信封上标明“十七届学术年会论文”字样。接收论文的时间为 2013 年 6 月 1 日至 9 月 30 日。经评审合格后即发参会邀请函和会务通知。

本届学术年会将举办“全国汉语方言学会首届青年学者论文比赛”和“全国汉语方言学会首届国际音标记音比赛”，均设有奖金，有关规定详见《方言》2013 年第 1 期及中国社会科学院语言研究所方言研究室网页：http：//ling. cass. cn/fangyanweb/index. htm。

从本届学术年会起，每届学术年会都将根据学术热点和学科发展情况组织专场讨论。本届学术年会将设“新技术与方言调查研究”、“语言信息资源与国家安全”、“方言岛与濒危汉语方言”三个专场。参加专场讨论会的名单采用报名(需经评审)和特别邀请两种方式确定。欢迎方言学科国家重大招标项目申请专场讨论。

(全国汉语方言学会秘书处)

《当代语言学》第 15 卷 2013 年第 2 期 250 – 252 页，北京

Abstracts of Articles

LU Qin，CHEN Yirong，and LI Sujian，The construction of ontology: Top-down approach vs bottom-up approach

Ontology construction aims to build conceptual knowledge in such a way that the relations among major concepts can be explicitly identified and presented in a machine operable way so as to assist in intelligent processing of computer applications. An upper-level ontology includes general concepts that are used broadly across different domains whereas ontologies acquired by computing through algorithms automatically are more likely to be domain specific. This paper first introduces domain specific core ontology (mid-level ontology) and application domain ontology (lower-level ontology). Then, it presents a top-down approach to build a core ontology for Chinese in the IT domain based on the English upper level ontology SUMO and other English-Chinese resources available. The paper also introduces a bottom-up approach to build domain specific ontology using corpus based approach.

Keywords: automatic ontology construction, upper-level ontology, mid-level ontology (domain core ontology), application domain ontology/lower level ontology, hypernym relations

CHOU Yamin and HUANG Churen，The formal representation for Chinese characters

The formal representation of Chinese characters using ontology is an important research area, and advantageous to process Chinese language. This paper aims to describe the methodology of constructing the ontology of Chinese characters and its formal representation. The formal representation proposed herein includes the external structure and derivation of Chinese characters, semantic and phonetic symbols, internal structure, sense and derived words, the relations of variants, and the pronunciations. The semantic symbols and senses of characters are connected with IEEE Suggested Upper Merged Ontology (SUMO). This study uses the OWL (Web Ontology Language)-DL to describe the knowledge of Chinese characters and share with other ontology.

Keywords: formal representation, Chinese characters ontology, SUMO

HSIEH Shukai，Lexical semantic relations in Chinese: A preliminary study on the classification, logical validation and evaluation method

In recent years, construction of lexical knowledge resources like WordNet has become one of the common interests among lexical semantics and ontological knowledge engineering. The labeling of different semantic relations in lexical resources not only constitutes the base but also has great

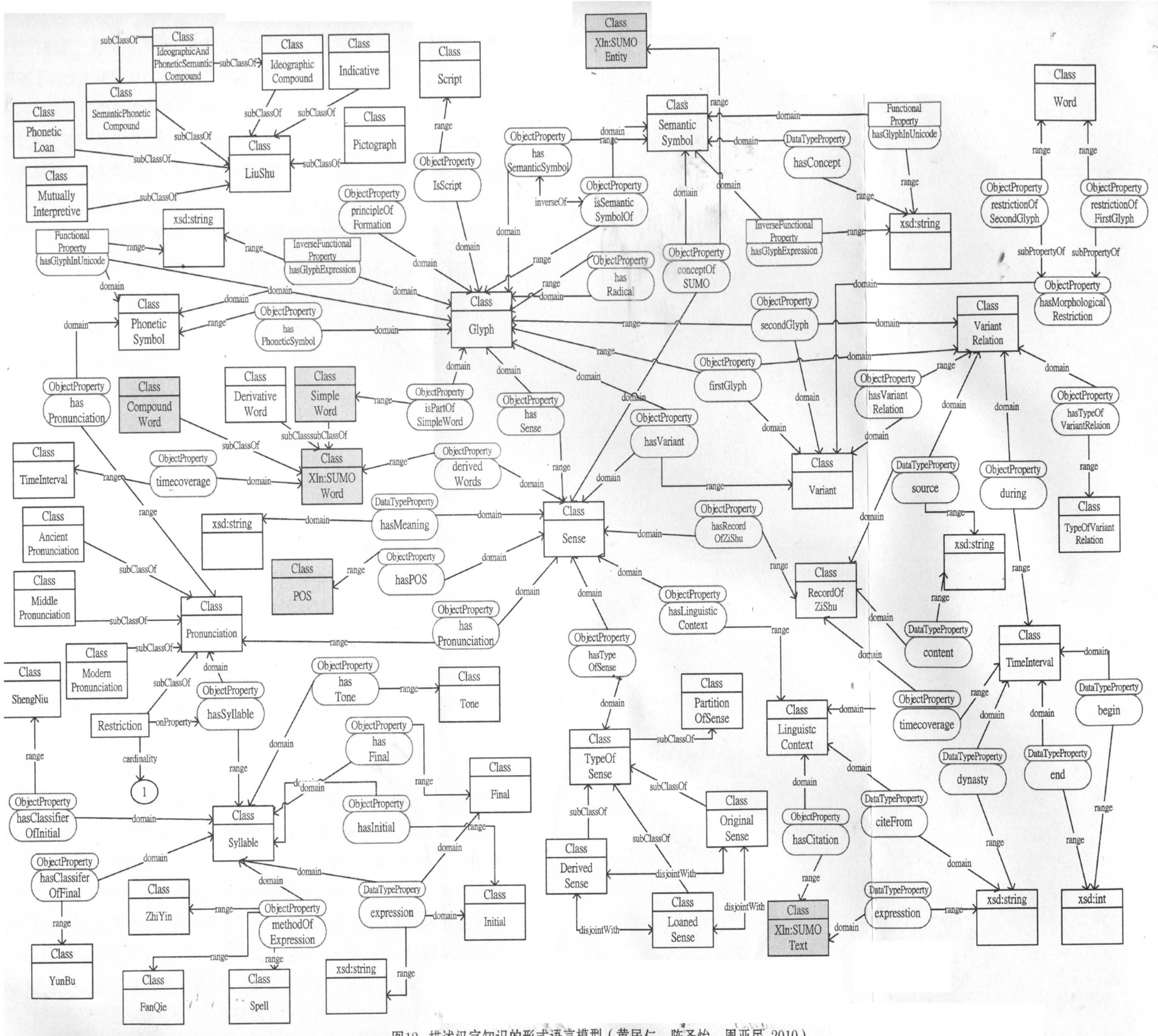


图13 描述汉字知识的形式语言模型（黄居仁、陈圣怡、周亚民 2010）