Journal of Geographic Information System, 2009, 1, 12-16
Published Online August 2009 in SciRes (http://www.SciRP.org/journal/jgis/).

Metadata Extended Model Based On Geological Domain Ontology

Ying HUANG[1,2], Mingqiang GUO[1,2], Xiangang LUO[1,2], Zhong XIE[1,2]
[1]*Faculty of Information Engineering, China University of Geosciences, Wuhan China*
[2]*GIS Software and Application Project Research Center of the Educational Department*

Abstract: The current metadata modeling techniques can not meet the needs of knowledge conception expression, knowledge organization, and metadata semantic consistency in geological domain. This paper introduces ontology and integrates this theory to geological domain metadata modeling. It adopts the first order logic equivalent algorithm and defines the metadata extended model as a quaternion group which is consists of geological term set, geological term definition set, attribute definition set and instance set. It also provides the formal description of each set. Finally the five steps for building geological domain metadata extended model are given. The result presents that this model not only provides the content standards for geological domain knowledge representation and knowledge organization, but also provides the basis for geological domain multi-source data and historical data integration and application in semantic consistency.

Keywords: ontology; geological domain ontology; metadata; metadata extended model

基於地質領域本體的元資料擴展模型

黃　穎[1,2]，郭明強[1,2]，羅顯剛[1,2]，謝　忠[1,2]
1. 中國地質大學資訊工程學院，湖北 武漢
2. 教育部地理資訊系統軟體及其應用工程研究中心，湖北 武漢

摘　要：針對目前地質領域現有元資料建模技術無法滿足地質領域知識概念表示、知識組織體系及元資料語義一致性等問題,借鑒本體思想,將本體理論引入到地質領域元資料建模中,採用一階邏輯等值演算方法,將元資料擴展模型形式化定義為一個四元組,並給出各元組間相互關係的形式化描述.最後,提出了建立地質領域元資料擴展模型的五步法來構建知識工程.將本體理論引入到地質領域元資料建模中來,可以把元資料模型中實體、屬性和聯繫隱含的語義顯式的表達出來,不僅為地質領域知識概念的表示和知識組織體系提供了內容標準,也為地質領域專用元資料標準的建立、地質領域多源資料的集成以及在語義一致性上解決歷史資料整合與應用等方面提供了基礎.

關鍵字：本體；地質領域本體；元數據；元資料擴展模型

1. 前　言

元資料是關於資料的資料[1]。經過幾十年工作,我國花費了大量人力、物力、財力獲取了各種地下、地面的地質空間資訊, 但目前這些資訊零散地分佈在各個部門、各個地區和各個單位,形成了“資訊孤島”的現狀[2]。同時,針對地質領域元資料,一直沒有一個系統有效的定義,從而無法形成統一的語義描述,資訊在不同的地域,以及不同電腦系統之間無法進行標準化的交流,影響了地質資料的共用.產生這些問題的主要原因是現有元資料建模技術已經無法滿足地質領域知識概念表示和知識組織體系方面的有關問題,也無法保證地質領域元資料語義上的一致性.因此,本文將本

基金專案:十一五國家 863 計畫課題“面向網路的三維空間資訊服務技術研究與軟體發展”(No. 2009AA12Z211)資助

體理論引入到地質領域元資料建模研究中,建立地質領域元資料描述框架,提供元資料建模規範,允許不同用戶在統一的元資料模型下自定義元資料,形式化描述元資料語義,從而使地質領域不同分類的元資料之間相互理解成為可能.

2. 本體及地質領域本體

在哲學上,一般認為本體論就是存在的理論.在資訊科學領域,本體論最具有代表性的定義是:共用概念模型的明確的形式化規範說明[3].地質領域本體兼具哲學本體和資訊本體的雙重含義.其目標是建立地質領域主題的、層次清晰的規範說明,形成公認的形式化的知識表示體系和地質領域主題知識的內容資訊組織,為地質領域資料模型的基本內容標準、地質領域多源資料整合的編碼問題、在語義一致性上解決歷史資料庫的整合與應用等方面提供基礎.

可以用五個基本建模元語[4]來表示地質領域本體:GO={C,R,F,A,I},其中:

C 表示地質類或概念的集合,比如地質領域本體中地層領域包括地層研究和地層劃分,而地層劃分又分為地層分類、國際年代地層單位及部分中國年代地層單位;

R 表示地質領域中概念之間的交互作用,比如地質領域本體中礦產地質領域與礦產分類屬於從屬關係,而金屬礦產與非金屬礦產屬於並列關係,具體的金屬礦鈾與能源礦產又是一種相關的關係;

F 表示函數:是一類特殊的關係,是關係的特定表達方式.函數中規定的映射關係,可以使得推理從一個概念指向另一個概念. 如 instance-of(x,y)就是一個函數，表示 x 是 y 的一個實例.

A 表示本體公理:通常都是一階謂詞邏輯的運算式,是無需再進行證明的邏輯永真式.如概念乙屬於概念甲的範疇.

I 表示實例:實例代表元素,從語義上講它表示的就是物件,也稱個體.類是實例的類,實例是類的實例,實例是本體中最小物件,它具有原子性,即不可再分性.如果某個實例還可以再進行劃分,那麼它就是類,而不是實例.比如地層-地層劃分-國際年代地層單位及部分中國年代地層單位-年代地層單位的等級-宇(宙).

3. 元資料擴展模型設計原理

定義 1 基於地質領域本體的元資料擴展模型可以用一個四元組來表示,記作 G=<GT,GT^D,GA^D,GEX>,其中 GT 表示地質術語集,GT^D 表示地質術語定義集,GA^D 表示屬性定義集,GEX 表示地質實例集.模型圖如圖 1 所示。

基於地質領域本體的元資料擴展模型旨在研究地質領域元資料不同物件及其語義關係.其要素為地質術語集、地質術語定義集、屬性定義集與實例集.它們之間存在一些重要的語義關係:

・地質術語集與地質術語定義集:包含關係(part-of)、繼承關係(kind-of)、等價關係(equal-of)、非交關係(disjointness-of);

・地質術語集與實例集:實例關係(instance-of);

・地質術語集與屬性定義集:區分關係(difference-of);

・實例集與屬性定義集:抽象關係(abstract-of)。

語義形式化描述詳見 3.3.

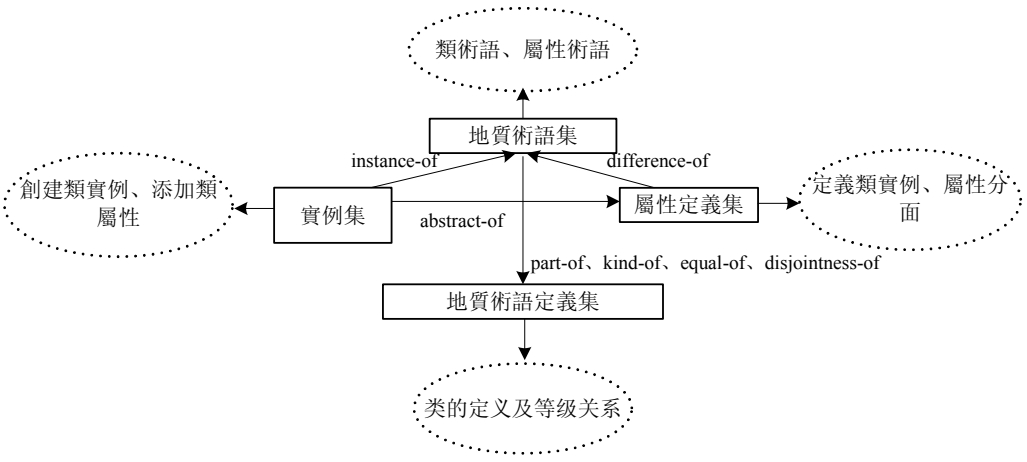

圖 1. 基於地質領域本體的元資料擴展模型

2.1 地質術語集

地質術語集 GT:由類術語與屬性術語構成.在本模型中,類術語對應著現實世界中的地質實體,包括類的定義和類的等級體系.屬性術語對應於實體間的各種關係以及實體自身的特徵.

2.2 地質術語定義集

地質術語定義集 GT^D:用來定義 GT 中的術語所屬類及類的等級體系.

完善一個等級體系有幾種可行的方法:

1)自頂向下法:由地質領域中最大的類開始,而後再將這些類細化.

2)自底向上法:由底層最小類的定義開始,它們是這個等級體系的細枝末節,然後將這些細化的類組織在更加綜合的類之下.

3)綜合法:綜合以上兩種方法.首先定義大量重要的類,然後分別將它們進行恰當地歸納和演繹.最後將它們與一些中級類關聯起來.

無論選擇哪種方法,都要從"類"的定義開始.

2.3 屬性定義集

屬性術語根據其取值的類型可分為兩種:一種是地質類屬性,表示兩個地質類之間的關係；另一種是屬性分面,表示類的特徵.一個屬性可能由多個"分面"組成.一個屬性的"分面",就是屬性取值的類型、容許的取值、取值個數和有關屬性取值的其他特徵.屬性定義集 GA^D 的目的就是將屬性術語分配給類術語.

一旦定義好了一些地質類,就必須開始描繪類的內在特徵.通常,有幾種屬性類型:

1)"內在"屬性:如金屬鐵的顏色、密度等；

2)"外在"屬性:如鐵的產地；

3)組成部分:如組成鐵的化學元素為 Fe；

4)與其他個體的關係.此處的"關係"是指某個類中的實例與其他類之間的關係.任何一個類的所有下位類都會繼承其上位類的屬性.如,鐵礦是黑色金屬礦產的下位類,繼承了其上位類黑色金屬礦產的屬性.

2.4 實例集

類是為了描述物件而形成的概念,先有物件後有類,而一旦形成類之後,便把物件歸屬在相應的類下,此時稱物件為類的實例.實例集定義類術語的下屬實例,包括:確定一個類,創建該類的一個實例,以及添加這個類的屬性值.

3. 元資料擴展模型形式化描述

3.1 元資料擴展模型關係綜述

包含關係(part-of)是地質術語集與地質術語定義集之間部分與整體的關係.

等價關係(equal-of)即兩個地質術語完全相同.

繼承關係(kind-of)即子類與父類的關係.

非交關係(non-intersection-of)即在同一劃分層次上的不同術語集間的互不相交關係.

抽象關係(abstract-of)是實例集與屬性定義集之間的關係,是從實例集中抽取一個或多個實例基本屬性的過程,即基本屬性與實例集之間的一對多關係.

實例關係(instance of)指地質術語集與實例集之間的關係,即類和其實例之間的多對多映射關係.一個類可實例化為多個物件,一個物件也可為多個物件的實例化.

區分關係(difference-of)是從屬性定義集到術語集的一種映射,即屬性與類之間的關係.兩類之間總存在差別,體現於其內涵,即基本屬性上.

基於地質領域本體五元語,對一階謂詞邏輯進行了擴展,增加了相應的定義、關係、函數,對元資料擴展模型語義進行解釋.

3.2 元資料擴展模型語法

經典一階謂詞邏輯語言(BittnerT etc.,2004)L 包括:1)個體變數:i,j；2)個體常量:m,n；3)5 個邏輯符號(Masolo C,Borgo S,2005): $\neg \rightarrow \exists \forall \Rightarrow$ ；4)3 個輔助符號:,)(.為描述該擴展模型,一階語言 L 可擴展為:1)兩個二元謂詞原語:=(等於)、≠ (不等於)；2)三個函數定義:f_{ins},f_{abs},f_{diff}.

3.3 元資料擴展模型形式化

給定 G=<GT,GT^D, GA^D, GEX >,語義解釋為一個二元組:I=<Δ^I,$\cdot^I$>,其中 $\Delta \neq \varnothing$ 為 G 的論域,GT,GT^D, GA^D, GEX 是對論域的一種劃分,I 為解釋函數,包括 f_{ins},f_{abs},f_{diff},用於約束實例、抽象、區分等關係,有以下語義形式化描述:

3.3.1 地質術語集與地質術語定義集

定義 2 GT_1,$GT_2 \in$ GT,ga_1,$ga_2 \subset GA^D$ 為 GT_1,GT_2 的屬性定義集,記為 {ga_{11},ga_{12},ga_{13},. . . ,ga_{1n}} 與 {ga_{21},ga_{22}, ga_{23},. . . ,ga_{2n}},R 是 GT 上的二元關係,包括等價、包含、繼承、非交等,則有:

1)術語具有非空二元關係:R≠ ∅；

2)$ga_{1i} \neq \varnothing$,$ga_{2i} \neq \varnothing$ $(1 \leq i \leq n)$, 若 $\{ga_{11},ga_{12},ga_{13},\ldots,ga_{1n}\}=\{ga_{21},ga_{22},ga_{23}\ldots,ga_{2n}\}$, 則 $GT_1 \equiv GT_2$,稱為 GT_1RGT_2 為等價關係;

3)設 $O=f_{ins}(GT_1)$,即 O 為 GT_1 的一個實例,若 O 在 GT_2 的所有屬性定義集 ga2 上均有真值,且值域是相關的,反之不成立,則 GT_1 語義包含 GT_2,記為 $GT_2 \subseteq GT_1$ 或 $GT_2 \subset GT_1$,GT_1RGT_2 為包含關係;

4)繼承關係: $\forall\ ga_{1i} \in ga_1, \exists\ ga_{1i} = ga_{2j} \in ga_2 (1 \leq i, j \leq n)$,則 $GT_2R\ GT_1$ 為繼承關係;

5) 若 $GT_1 \subseteq \neg\ GT_2, GT_2 \subseteq \neg\ GT_1$, 則 $GT_1 \cap GT_2 = \varnothing$,GT_1RGT_2 為非交關係;

3.3.2 實例集與屬性定義集

定義 3 GT 為術語定義集,GA^D 為屬性定義集,GEX 為實例集,f_{ins}, f_{abs} 為一元函數,則實例集與屬性定義集的抽象關係為:

1)若某一地質類術語存在,則可抽象出其屬性定義集,即:若 $\exists\ GT_i \in GT$, 則 $f_{abs}(GT_i)=\{ga_{11},ga_{12},ga_{13},\ldots,ga_{1n}\}$, 且 $\{ga_{11}, ga_{12}, ga_{13},\ldots,ga_{1n}\} \subset GA^D$,其中 $1 \leq i \leq n$;

2)基本屬性可抽取於一個或多個實例,即:若 $\{ga_{11},ga_{12},ga_{13},\ldots,ga_{1n}\} \subset GA^D$,$f_{abs}(GT_i)=\{ga_{11},ga_{12},ga_{13},\ldots,ga_{1n}\}$,則 $f_{ins}(GT_i) \in GEX$,其中 $1 \leq i \leq n$;

3)同一術語集的實例,其基本屬性相同,即: $\forall\ GT_i, GT_j \in GT$, 若 $GT_i=GT_j$, 則 $f_{abs}(f_{ins}(GT_i))=f_{abs}(f_{ins}(GT_j))$,其中 $1 \leq i, j \leq n$;

3.3.3 地質術語集與實例集

定義 4 GT 為術語定義集,GEX 為實例定義集,f_{ins},f_{abs} 為一元函數,則地質術語集與實例集的實例關係為:

1)實例集是實例的集合,凡類術語都可被實例化,即: $\forall\ GT_i \in GT, \exists\ f_{ins}(GT_i) \neq \varnothing$, 且 $f_{ins}(GT_i)=GEX_i \in GEX$,其中 $1 \leq i \leq n$;

2)凡實例都有其歸屬的類術語存在,即: $\forall\ GEX_i \in GEX, \exists\ GEX_i \in f_{ins}(GT_i)$且 $GT_i \in GT$,其中 $1 \leq i \leq n$;

3)兄弟類的實例不相交,即:若 $f_{ins}(GT_1)=GEX_1, f_{ins}(GT_1)=GEX_2, GT_1 \in GT, GEX_1, GEX_2 \in GEX$,且 $GEX_1 \neq GEX_2$,則 $f_{ins}(GT_1) \cap f_{ins}(GT_1)=\varnothing$;

4)若實例相同,則其對應的類術語具有相等或包含關係,即: $\forall\ GT_1, GT_2 \in GT$,若 $f_{ins}(GT_1)= f_{ins}(GT_2)$,則 $GT_1 \subseteq GT_2$ 或 $GT_2 \subset GT_1$;

5)若類術語相同,則其實例具有相同的基本屬性,見定義 3③.

3.3.4 地質術語集與屬性定義集

定義 5 GT 為術語定義集,GA^D 為屬性定義集,則地質術語集與屬性定義集的區分關係為:

1)屬性定義集相同則對應的地質術語相同,見定義 2②;

2)不同屬性定義集用於區分不同的地質術語,即 :$\{ga_1,ga_2,ga_3,\ldots,ga_m\} \subset GA^D$, $\{ga_1,ga_2,ga_3,\ldots,ga_n\} \subset GA^D, m \neq n$,則 $f_{equ}(GT_1) \neq f_{equ}(GT_2)$.

4. 五步法創建地質領域本體的元資料擴展模型

在遵循上述元資料擴展模型的設計原理與形式化描述的基礎上,通過抽象總結出一套創建地質領域元資料擴展模型知識工程的五步法.

第一步,列舉地質領域本體中的重要術語.

盡可能多的列舉一份地質領域術語清單,清單中的術語是元資料擴展模型想要陳述的或要向用戶解釋的所有概念.此時暫不考慮概念間會有屬性及表達上的重複.

第二步,地質領域本體術語提取

上一步驟中已經產生了地質領域中大量術語,但卻是一張毫無組織結構的辭彙表,這時需要對其中的每一個術語的重要性要進行評估,選出關鍵性術語,摒棄那些不必要或者超出地質領域範圍的術語,盡可能準確而精簡的表達出地質領域的知識,從而形成一個領域知識的框架體系,為建立地質領域本體元資料模型做必要準備.

第三步,建立元資料擴展模型框架,定義四元組關係

為了描述地質領域本體元資料擴展框架,本文提出四元組模型.該模型不僅定義了地質領域概念集,還定義了地質領域概念間的關係.

1)定義類及類間關係.創建的術語中,很大一部分屬於類術語,而對類術語的術語定義有以下 3 種方法:自頂向下法、自底向上法及綜合法(參見 2.2 節).這 3 種但不論採用哪種方法,都需要從類術語的定義開始.類間關係表示該類術語與術語定義集間包含、繼承、等價及非交關係,如子類與父類所描述的概念是包含或繼承關係.

2)定義類的屬性.僅有類間的關係根本不足以描述地質領域元資料擴展模型.因此,一旦定義好了類就要描述這個類的內部結構.已經從步驟 2 的術語列表中選擇好類,絕大多數剩下的術語可能是這些類的屬性.通常,有

幾種物件屬性的類型能夠成為一個本體中的屬性: “內在”屬性、“外在”屬性、組成部分、個體間關係(參見2.3 節).除了最初確定的一些屬性之外還要描述地質類術語與屬性、實例與屬性定義集間的關係.

3)定義屬性值.屬性術語根據其取值的類型可分為兩種:一種是地質類屬性；另一種是屬性分面(參見2.3 節).

4)創建實例集(參見 2.4 節)

第四步,對元資料擴展模型編碼形式化.

選用合適的本體描述語言對本模型進行編碼、形式化.目前大多數已經建立的本體模型都是基於一階謂詞邏輯或描述邏輯[4,7].本模型採用一階謂詞邏輯.元資料擴展模型的形式化描述可以提供比自然語言更嚴格的格式,增強機器的可讀性,進行自動翻譯以及交換,便於本模型自動進行邏輯推理.(參見 3.3 節)

第五步,模型的檢驗評價.

元資料擴展模型形式化以後,是否滿足了我們剛開始提出的需求,是否滿足模型的建立準則,模型中的術語是否被清晰的定義了,概念及其關係是否完整等問題都需要我們在模型建立過後進行核對總和評估.由於是在建立客戶本體過程中進行檢驗評價,鑒於文章篇幅的原因,在這裏不具體介紹.

5 結論

將本體理論引入到地質領域元資料建模中來,可以把元資料模型中實體、屬性和聯繫隱含的語義顯式的表達出來,不僅為地質領域知識概念的表示和知識組織體系提供了內容標準,也為地質領域專用元資料標準的建立、地質領域多源資料的集成以及在語義一致性上解決歷史資料整合與應用等方面提供了基礎.

今後還要在以下幾個方面深入研究:

1)擴展各元組間二元或多元的關係；

2)研究地質領域本體元資料分類粒度問題；

3)研究地質領域本體元資料分類標引規則.

參考文獻 (References)

[1] V. Kashyap and A. Sheth. Information brokering across heterogeneous digital data—A metadata based approach. Kluwer Academic Publishers, 2000.

[2] C. L. Li, F. D. Li, and X. G. Luo. The construction and implementation of national geological spatial information grid node computing pool. 2006, 5(5): 2.

[3] L. Su, Q. W. Zhu, and Y. J. Chen. Conceptual modeling of spatial database based on geographic ontology. Computer Engineering, 2007, 33(12): 87.

[4] H. W. Wang, J. C. Wu, and F. A. Jiang. Study on ontology model based on description logics: System engineering. 2003, 21(3): 101−107.

[5] T. Bittner, M. Donnelly, and B. Smith. Individuals, universals, collections: On the foundational relations of ontology. Proceedings of the Third Conference on Formal Ontology in Information Systems, FOIS, Turin, 2004, 37−48.

[6] C. Masolo and S. Borgo. Qualities in formal ontology P. Hitzler, C. Lutz, G. Stumme. The proceedings of the workshop on foundational aspects of ontologies. Koblenz Germany, 2005, ISTC-CNR: 2−16.

[7] H. W. Wang, J. C. Wu, and F. Jiang, Extended ontology model and ontology checking based on description logics. Journal of Shanghai Jiaotong University (Science), 2004, 1, 195−198.

