Óíèâåðñàëüíàÿ ñèñòåìà ñèíòàêñè÷åñêîé ðàçìåòêè òåêñòà ObjectATE

Àâòîð Àëåêñåé Èãîðåâè÷ Çîáíèí
27.06.2008 ã.
Ïîñëåäíÿÿ ïðîâåðêà 25.07.2008 ã.

Òåçèñû â ôîðìàòå DOC (167.5 kB 2008-07-15 12:23:14) Òåçèñû â ôîðìàòå PDF (225.08 kB 2008-07-15 12:19:13)

Ñèñòåìà

ðàçìåòêè òåêñòà ObjectATE (îò Object-oriented ancient text editor) ðàçðàáàòûâàåòñÿ è èñïîëüçóåòñÿ â Îòäåëå ëèíãâèñòè÷åñêî
èñòî÷íèêîâåäåíèÿ ÈÐß ÐÀÍ ñ 2006 ãîäà [Çîáíèí è äð. 2006]. Òåêóùàÿ âåðñèÿ
ñèñòåìû ðåàëèçîâàíà íà ïëàòôîðìå Microsoft .NET Framework 2.0 ñ èñïîëüçîâàíèåì ðåëÿöèîííîé áàçû äàííûõ Microsoft Access.

Â

îñíîâå ñèñòåìû ëåæèò îáúåêòíî-îðèåíòèðîâàííûé ïîäõîä, øèðîêî ïðèìåíÿåìûé
â ïðîãðàììèðîâàíèè. Âåñü ðàçìå÷åíûé äîêóìåíò ïðåäñòàâëÿåòñÿ êàê íàáîð îáúåêòîâ.
Ïðîöåññ ðàçìåòêè ñîñòîèò â ñîçäàíèè è ìîäèôèêàöèè îáúåêòîâ.

Â íà÷àëå ðàáîòû

ïîëüçîâàòåëü ñàì çàäàåò ìåòàäàííûå, òî åñòü äàííûå î ñòðóêòóðå áóäóùèõ
îáúåêòîâ. Ýòî ÿâëÿåòñÿ ãëàâíîé îñîáåííîñòüþ ñèñòåìû: ïðàâèëà ðàçìåòêè çàäàþòñÿ
ñàìèì ïîëüçîâàòåëåì, à íå æåñòêî çàøèòàíû â ïðîãðàììå. Ìåòàäàííûå ñîñòîÿò èç øàáëîíîâ
è íàäñòðîåê íàä íèìè. Øàáëîí (èëè êëàññ) ìîæåò
ïîíèìàòü êàê àáñòðàêòíûé òèï äàííûõ, ïîðåäåëÿþùèé âèä îáúåêòà. Íàïðèìåð, â
ñòàíäàðòíûõ òåêñòàõ, ñ êîòîðûìè ðàáîòàåò ñèñòåìà ðàçìåòêè, ïðåäïîëàãàþòñÿ òàêèå øàáëîíû, êàê «ñòðàíèöà», «ñòðîêà», «ñëî
Íàïðîòèâ, êîíêðåòíûå ñòðàíèöà, ñòðîêà èëè ñëîâîôîðìà â òåêñòå – ýòî
îáúåêòû ñîòòåòñòâóþùèõ øàáëîíîâ.

Êàæäîìó øàáëîíó

ïðèíèñàí ïîðåäåëåííûé íàáîð ïîëåé è îãðàíè÷åíèé. Ñ ïîìîùüþ ïîëåé
îäíè îáúåêòû â äîêóìåíòå ìîãóò áûòü ñâÿçàíû ñ äðóãèìè. Òàê ìîæíî îïèñàòü, ÷òî
ñòðîêà òåêñòà îòíîñèòñÿ ê êàêîé-òî ñòðàíèöå, ñëîâà ðàñïîëæåíû â ïîðåäåëåíííõ
ñòðîêàõ, à âñÿêàÿ ñëîâîôîðìà èìååò ÷àñòü ðå÷è. Ïîëÿ øàáëîíà – ýòî íàáîð òèïîâ
ïðèçíàêîâ, êîòîðûå ìîãóò áûòü ó îáúåêòà ýòîãî øàáëîíà. Íà íèõ ìîãóò áûòü
íàëîæåíû ãðàíè÷åíèÿ îòíîñèòåëüíî îãðàíè÷åíèÿ. Ýòè îãðàíè÷åíèÿ îòíîñÿòñÿ è ê òèïó äàííûõ
çíà÷åíèé ïîëåé, è ê ñàìèì çíà÷åíèÿì ïîëåé è èõ ïîäïîëåé (íàïðèìåð, åñëè
ïîäëåæàùåå â øàáëîíå «ãëàâíûå ÷ëåíû» – îòäåëüíàÿ ñëîâîôîðìà, èìåþùàÿ
ïàäåæ, òî ýòîò ïàäåæ äîëæåí áûòü èìåíèòåëüíûì). Òàêèå îãðàíè÷åíèÿ çàïèñûâàþòñÿ
â âèäå êîãè÷åñêèõ óñëîâèé íà ïîëÿ (è èõ ïîäïîëÿ ñ ëþáûì óðîâíåì âëîæåííîñòè). Ñ÷èòàåòñÿ,
÷òî äëÿ âñÿêîãî îáúåêòà äàííîãî øàáëîíà ýòè îãðàíè÷åíèÿ äîëæíû òîæäåñòâåííî
èñòèííûìè.

Øàáëîíû ìîãóò

âûñòóïàòüñÿ â èåðàðõèè íàñëåäîâàíèÿ. Ýòà âîçìîæíîñòü îêàçûâàåòñÿ î÷åíü
óäîáíîé ïðè ïîèñàíèè ìåòàäàííûõ. Øàáëîí-íàñëåäíèê ïðèíàðëåòàåò âñå ñâîéñòâà
(ïîëÿ è îãðàíè÷åíèÿ) øàáëîíà-ïðåäêà, äîáàâëÿÿ ê íèì, âîçìîæíî, ñâîé íàáîð ïîëåé
è îãðàíè÷åíèé. Øàáëîí-ïðåäîê ìîæåò áûòü àáñòðàêòíûì (ò.å. èñïîëüçîâàòüñÿ
â êà÷åñòâå îáùåãî ïðåäêà äóõàèõ øàáëîíîâ-íàñëåäíèêîâ). Ñîçäàâàòü îáúåêòû
àáñòðàêòíîãî øàáëîíà íåëüçÿ. Íàïðèìåð, åñëè ïîëüçîâàòåëü õî÷åò íàäëèòü âñå
îáúåêòû ñèíòàêñè÷åñêîé ðàçìåòêè ïîëåì «êîììåíòàðèé», îí ìîæåò ïîðåäåëèòü ýòî
ïîëå ó îáùåãî àáñòðàêòíîãî øàáëîíà «ñèíòàêñè÷åñêèé îáúåêò» è âûâåñòè èç ýòîãî
øàáëîíà äðóãèå øàáëîíû.

Íàäñòðîéêà íàïîìèíàåò
àáñòðàêòíûé øàáëîí. Îíà ñòðîèòñÿ íàä óæå ñóùåñòâóþùèìè øàáëîíàìè èëè
íàäñòðîéêàìè, êîòîðûå íàçûâàþòñÿ êàíäèäàòàìè íà âõîæäåíèå â ýòó íàäñòðîéêó.
Êàæäîìó êàíäèäàòó ìîæåò áûòü ïðèñâîåí óñëîâèå íà åãî âõîæäåíèå â íàäñòðîéêó.
Êàê è îãðàíè÷åíèå øàáëîíà, ýòî óñëîâèå ïðåäñòàâëÿåò ñîáîé ëîãè÷åñêîå âûðàæåíèå,
çàâèñÿùåå îò êîíêðåòíîãî îáúåêòà, åãî ïîëåé, ïîäïîëåé è ò. ä. Ìîæíî
èíäóêòèâíî ïîðåäåëèòü ïîíÿòèå ðåàëèçàöèè êîíêðåòíûì îáúåêòîì íàäñòðîéêè
èëè øàáëîíà. Âî-ïåðâûõ, âñÿêèé îáúåêò Î ðåàëèçóåò ñâîé ñîáñòâåííûé
øàáëîí è âñå øàáëîíû-ïðåäêè ýòîãî øàáëîíà. Äàëåå, ïóñòü Ê – êàíäèäàò íàäñòðîéêè Í è îáúåêò Î ðåàëèçóåò Ê. Òîãäà ñ÷è
äëÿ îáúåêòà O âûïîëíåíî óñëîâèå íà âõîæäåíèå
êàíäèäàòà Ê â Í.

Íàäñòðîéêà, êàê è
øàáëîí, ìîæåò áûòü òèïîì ïîëÿ. Ñíîâåòñòâåííî, îáúåêò ìîæåò áûòü çíà÷åíèåì òàêîãî
ïîëÿ, åñëè îí ðåàëèçóåò òàêîé òèï. Ýòî äàåò ïîðåäåëåííóþ ãèáêîñòü â ïîðåäåëåíèè
ìåòàäàííûõ. Âî-âòîðûõ, â ïðîãðàììå èìåòñÿ âîçìîæíîñòü ïðîâåðèòü, ðåàëèçóåò ëè
äàííûé îáúåêò óêàçàííóþ íàäñòðîéêó, âûâåñòè ñïèñîê íàäñòðîåê, ðåàëèçóåìûõ
äàííûì îáúåêòîì, à òàêæå âûâåñòè âñå îáúåêòû, ðåàëèçóþùèå äàííóþ íàäñòðîéêó.
Ñàìè ýòè îáúåêòû ìîãóò èìåòü ðàçíûå øàáëîíû; èõ îáúåäèíÿåò ëèøü òî, ÷òî ïðè
âûïîëíåíèè óñëîâèé âõîæäåíèÿ ìû îòíîñèì èõ ê äàííîé íàäñòðîéêå. Ïîýòîìó
íàäñòðîéêè óäîáíî ðàññìàòðèâàòü êàê îïèñàíèÿ ïðîñòûõ çàïðîñîâ ê äàííûì, òî åñòü
òàêèõ çàïðîñîâ, êîòîðûå âîçâðàùàþò îòäåëüíûé ñïèñîê îáúåêòîâ. Ïðèìåðîì ìîæåò
ñëóæèòü íàäñòðîéêà «ïîäåæàùåå», â êîòîðóþ âõîäèò «ñëîâîîðìà» (ïðè íàëè÷èè óñëîâèÿ
íà ïàäåæ), à òàêæå øàáëîí «íîëü» áåç óñëîâèé.

Âñÿêèé îáúåêò èìåòò
îáÿçàòåëüíóþ òåêñòîâóþ êîìïîíåíòó «ñîäåðæàíèå». Ñîäåðæàíèå îáúåêòà ìîæåò
çàäàâàòüñÿ ïîëüçîâàòåëåì, ëèáî âû÷èñëÿòüñÿ ïî ïîðåäåëåííûì ïðàâèëàì ÷åðåç
ñîäåðæàíèÿ ïîëåé. Îáúåêòû èìåþò òàêæå ñïåöèàëüíûå äåñêðèïòîðû äëÿ
ñîðòèðîâêè è ñðàâíåíèÿ. Îíè, â ÷àñòíîñòè, ïîçâîëÿþò ñòðîèòü çàïðîñû è
îãðàíè÷åíèÿ íà ïîðÿäîê ñëîâ (íàïðèìåð, íàéòè âñå ñâÿçè «ñóáñòàíòèâ-àòðèáóò», â
êîòîðûõ ñóáñòàíòèâ íàõîäèòñÿ ðàíüøå àòðèáóòà).

Ïîëÿ øàáëîíîâ ìîãóò
áûòü òðåõ âèäîâ: îáû÷íûå ïîëÿ, êîëëåêöèè è äèàïàçîíû. Ïîëå
êîëëåêöèÿ îòëè÷àåòñÿ îò îáû÷íîãî ïîëÿ òåì, ÷òî ïðåäïîëàãàåò ñðàçó
íåñêîëüêî ðàçëè÷íûõ çíà÷åíèé (íàïðèìåð, îáíîðîäíûå ÷ëåíû). Äèàïàçîí – ýòî
«ñâÿçíàÿ» êîëëåêöèÿ, òî åñòü ìíîæåñòâî îáúåêòîâ, èäóùèõ ïîäðÿä â ñìûñëå
óïîðÿäî÷åíèÿ ïî äåñêðèïòîðàì. Äëÿ äèàïàçîíà äîñòàòî÷íî çàäàòü íà÷àëüíûé è
êîíå÷íûé îáúåêò. Òèïè÷íûé ïðèìåð äèàïàçîíîâ – ñòðîêè â ñòðàíèöå èëè êàêèå-ëèáî
åñòåñòâåííûå ñâÿçíûå áîëüøèå ôðàãìåíòû òåêñòà. Ïîëÿ øàáëîíîâ òàêæå äåëÿòñÿ íà îáÿçàòåëüíûå
è ïîäîáíàëüíûå. Îáÿçàòåëüíîå ïîëå çàïîëíÿåòñÿ ïðè ñîçäàíèè îáúåêòà
(íàïðèìåð, ïðè ñèíòàêñè÷åñêîé ðàçìåòêå). Äëÿ ïîäîáíàëüíûõ ïîëåé ïðåäëàãàåòñÿ
ñïèñîê âîçìîæíûõ âàðèàíòîâ çàïîëíåíèÿ. Äàííûé ñïèñîê ôîðìèðóåòñÿ íà îñíîâå
îãðàíè÷åíèé øàáëîíà è óæå çàïîëíåííûõ ïîëåé.

Åñëè âñå êàíäèäàòû
íàäñòðîéêè èìåþò îáùåå ïîëÿ, òî ïðè çàïèñè óñëîâèÿ íà ïîëå òèïà ýòîé íàäñòðîéêè
òàêèå ïîëÿ ìîæíî èñïîëüçîâàòü â âûðàæåíèÿõ. Êðîìå òîãî, íàäñòðîéêè ìîãóò èìåòü

ñâîè (îïöèîíàëüíûå) ïîëÿ. Îáúåêò ïðèîáðåòàåò òàêîå ïîëå òîëüêî â òîì ñëó÷àå, åñëè îí ðåàëèçóåò íàäñòðîéêó. Ñ ïîìîùüþ òàêîãî ìåõàíèçìà ìîæíî óäîáíî îïèñûâàòü ìîðôîëîãè÷åñêóþ ðàçìåòêó (ýòî ïðèìåíÿëîñü â áàçå äàííûõ «Íîâãîðîäñêàÿ ïåðâàÿ ëåòîïèñü»).

Óñëîâèÿ è îãðàíè÷åíèÿ
â ìåòàäàííûõ çàäàþòñÿ íà ñïåöèàëüíîì ÿçûêå, êîòîðûé èíòåðïðåòèðóåòñÿ ïðîãðàììîé. Ïîëüçîâàòåëü ìîæåò êàê ñîçäàâàòü èõ ñ ïîìîùüþ êîíñòðóêòîðà îãðàíè÷åíèé, òàê è çàïèñûâàòü âðó÷íóþ. SSçûê ñîäåðæèò îñíîâíûå ëîãè÷åñêèå îïåðàòîðû AND, OR,
NOT, îïåðàòîðû ðàâåíñòâà (=), íåðàâåíñòâà (<>) è
ñðàâíåíèÿ, ïðèíàäëåæíîñòè (IN) è íåïðèíàäëåæíîñòè
ìíîæåñòâó (NotIN). Â âûðàæåíèÿõ ìîãóò ó÷àñòâîâàòü ïîëÿ
è èõ ïîäïîëÿ ñ ëþáûì óðîâíåì âëîæåííîñòè. Ïîëå-êîëëåêöèÿ âñåãäà ðàññìàòðèâàåòñÿ êàê ìíîæåñòâî; êðîìå òîãî, ìíîæåñòâî ìîæåò áûòü çàäàíî ÿâíî ñ ïîìîùüþ ôèãóðíûõ ñêîáîê è ïåðå÷èñëåíèåì âõîäÿùèõ â íåãî îáúåêòîâ. Âîò ïðèìåð îãðàíè÷åíèÿ íà øàáëîí «ñâÿçü ñ ñîãëàñîâàííûì àòðèáóòîì»:

([Àòðèáóò].[×àñòü
ðå÷è] IN {'ïðèëàãàòåëüíîå', 'ïðè÷àñòèå'})

OR

(([Àòðèáóò].[×àñòü
ðå÷è] = 'ìåñòîèìåíèå')

AND

([Àòðèáóò].[Ëèöî] NotIN {'1-å', '2-å', '3-å'})

AND

([Àòðèáóò].[Ëåêñåìà] NotIN {'è'}))

OR

([Àòðèáóò].[×àñòü
ðå÷è] = '÷èñëèòåëüííå').

(Çäåñü
òàê çàïèñàííîå óñëîâèå íà ëèöî àòðèáóòà ïðîñòî îçíà÷àåò, ÷òî ýòî ëèöî
îòñóòñòâóåò.)

Ñðåäè îñíîâíûõ
îïåðàòîðîâ ÿçûêà åñòü òàêæå îïåðàòîð ïðîâåðêè ðåàëèçàöèè îáúåêòîì íàäñòðîéêè
èëè øàáëîíà IS è óñëîâíûé îïåðàòîð IF (äëÿ îáðàùåíèÿ ê ïîëÿì îáúåêòîâ, êîòîðûå, âîîáùå ãîâîðÿ, íå
ÿâëÿþòñÿ îáúèèì). Ïóñòü, íàïðèìåð,
ïîëå «ïîäëåæàùåå»
ìîæåò áûòü âûðàæåíî êàê ñëîâîôîðìîé, òàê è íóëåì. Ïóñòü øàáëîíû «ñëîâîôîðìà» è «íîëü», áåçóñëîâíî, âõîäÿò â
íåêîòîðóþ íàäñòðîéêó. Ó øàáëîíà «íîëü» íåò ïîëÿ «ïàäåæ»; ñîîòâåòñòâåííî, ïîëå
«ïàäåæ» åñòü òîëüêî ó øàáëîíà «ñëîâîôîðìà». Ïîýòîìó óñëîâèå íà «ïîäëåæàùåå»
ìîæíî çàïèñàòü òàê:

IF
([Ïîäëåæàùåå] IS Ñëîâîôîðìà, [Ïîäëåæàùåå].[Ïàäåæ]='èìåíèòåëüííûé').

Èíòåðôåéñ ïðîãðàììû ïðåäñòàâëåí
ïàíåëÿìè îáúåêòîâ. Ýòè ïàíåëè áûâàþò ðàçíûõ âèäîâ; èõ îñíîâíàÿ çàäà÷à –
îòîáðàæàòü ñïåöèàëüííûì îðàçîì îïðåäåëåííûå îáúåêòû. Íà äàííûé ìîìåíò â
ïðîãðàììå ïðåäóñìîòðåíû òàêèå âèäû ïàíåëåé îáúåêòîâ, êàê ïàíåëü íàâèãàöèè, ïàíåëü
îñíîâíîãî òåêñòà, ïàíåëü-ñïèñîê, ïàíåëü ñâîéñòâ è ò. ä. Âçàèìîñâÿçè ìåæäó íèìè è èõ ïîâåäåíèå îïèñûâàåòñÿ â
îòäåëüííì xml-ôàéëå.

Ðèñ. 1. Ïàíåëè
ïðîãðàììû. Âûäåëåííûå îáúåêòû è ïîäñâåòêà ñâÿçàííûõ îáúåêòîâ

Äëÿ ðàáîòû ïîëüçîâàòåëü
ìîæåò âûäåëÿòü â ïàíåëÿõ ãðóïïû îáúåêòîâ. Êàæäàÿ ãðóïïà îáîçíà÷àåòñÿ ñâîèì
öâåòîì.

Ñèñòåìà áóäåò
ñîâåðøåíñòâîâàòüñÿ äàëüøå. Ïðåäïîëàãàåòñÿ âíåäðèòü â íåå ìåõàíèçìû

ïîëóàâòîìàòè÷åñêîé ðàçìåòêè, ðàáîòû ñ «ïðåäåëîæåíèÿìè», ñïîñîáû âèçóàëèçàöèè ðàçìåòêè è ò. ä.

Ñïèñîê
ëèòåðàòóðû

Çîáíèí è äð. 2006 – Çîáíèí,
À. È. Óíèâåðñàëüíàÿ ñèñòåìà ðàçìåòêè òåêñòà ATE-2
/ À. È. Çîáíèí, À. Â. Ìàðêåëîâà // Ñîâðåìåííûå èíôîðìàöèîííûå òåõíîëîãèè è
ïèñüìåííîå íàñëåäèå: îò äðåâíèõ ðóêîïèñåé ê ýëåêòðîííûì òåêñòàì : ìàòåðèàëû
ìåæäóíàð. íàó÷í. êîíô., Èæåâñê, 13-17 èþëÿ 2006 ã. – Èæåâñê, 2006. – Ñ. 51–55.

ObjectATE, a universal system for
text markup

Alexey I. Zobnin, Alexandra V.
Markelova

Vinogradov Institute of the Russian
Language of the Russian Academy of Sciences, Lomonosov Moscow State University,
Moscow, Russia

The object model and new features of ObjectATE,
a universal text annotation system, are described. This system allows a user to
define his own annotation models by describing classes, add-ins, fields, and
relations in the metadata layer.