Óíèâåðñàëüíàÿ ñèñòåìà ñèíòàêñè÷åñêîé ðàçìåòêè òåêñòà ObjectATE

Àâòîð Äëåêñåé Èãîðåâè÷ Çîáíèí
27.06.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 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÷åñêîé èñòîð÷åñêîé 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÷åííûé 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 ñèſòåìû: ïðàâèëà ðàçìåòêè çàäàþòñÿ
ñàìèì ïîëüçîâàòåëåì, à íå æåſòêî çàïèſàíû â ïðîãðàììå. Ìåòàäàííûå ñîſòîÿò èç øàáëîíîâ
è íàſñòðîåê íàä íèì. Øàáëîí (èëè êëàſſ) ìîæíî
ïîſèſàòü êàê àſſòðàêòíóé òèï äàííûõ, îïðåäåëÿþùèé âèä îáúåêòà. Íàïðèìåð, â
ſòàíäàðòíûõ òåſòàõ, ſ êîòîðûì ſèſòåìà ðàçìåòêè, ïðåäïîëàãàþſÿ òàêèå øàáëîíû, êàê «ſòðàíèöà», «ſòðîêà», «ſëî
Íàïðîòèâ, êîíêðåòíûå ſòðàíèöà, ſòðîêà èëè ſëîâîôîðìà â òåêſòå – ýòî
îáúåêòû ſîſòâåòſòâóþùèſ øàáëîſîâ.

Êàæäîó øàáëîſó

ïðèſàſí îſðåäåëåſíûé íàáîð ſſëåé è îſðàſíè÷åſèé. S ſſſſùüþ ſſëåé
îſíè îáúåêòû â äſêóſåſòå ſſáóò áûòü ſſâÿçàſû ſ äðóſàèſè. Òàê ſſæſſ ſſèſàòü, ÷òſ
ſòðſſêà òåſſòà ſòſſſſèſòſÿ ê êàſſé-òſ ſòðſàſèöå, ſſſſâà ðàſſſſëſæſåſû â ſſðåäåëſſſſûſ
ſòðſſêàſ, à âſÿſêàÿ ſſſſâſſôſðſà èſåå ÷àſòü ðå÷è. Sſſÿ øàáſſſſà – ýòſ íàáſð òſſſâ
ſðſçſàſſâ, ſſòſðſûå ſſãóò áûòü ó ſáúåſêà ýòſſâſ øàáſſſſà. Íà ſſſ ſſãóò áûòü
ſàſſſæſåſû åſòåſſòâſſſûå ſſãðàſſ÷åſſÿ. Ýòſ ſſãðàſſ÷åſſÿ ſòſſſñÿòſſÿ è ê òſſſó äàſſſûſ
çſà÷åſſé ſſſåé, è ê ſàſſſ çſà÷åſſÿſ ſſſåé è ſſ ſſäſſſåé (ſàſðſſſåð, åſſſ
ſſäſſåæàſåå â øàáſſſå «ãſàâſûå ÷ſåſû» – ſòäåſſſàÿ ſſſſâſſôſðſà, ſſåþſàÿ
ſäàåæ, òſ ýòſò ſäàåæ äſſſæåſ áûòü ſſåſſòåſſſſûſû). Òàſſſå ſſãðàſſ÷åſſÿ çàſſſſûâàþòſÿ
â âſäå ſſſãſ÷åſſſſ óſſſſâſſé ſà ſſſÿ (ſ ſſ ſſäſſſſÿ ſ ſþáſû óðſâſſå ſſſſæåſſſſòſ). S÷ſòåòſÿ,
÷òſ äſÿ âſÿſêſãſ ſáúåſòà äàſſſſãſ øàáſſſſà ýòſ ſãðàſſ÷åſſÿ äſſæſſû áûòü òſæäåſòâåſſſ
ſſòſſſſûſſ.

Øàáſſſû ſſãóò

âûſòðàſâàòüſÿ â ſåðàðõſſ íàſſëåäſâàſſÿ. Ýòà âſçſſæſſſòü ſſàçûâàåòſÿ ſ÷åſü
óäſáſſé ſðſ ſſſſàſſſ ſåòàäàſſſûſ. Øàáſſſ-ſàſſſåäſſſſ ſðſſſáðåòåò âſå ſâſſſòâà
(ſſſÿ è ſãðàſſ÷åſſÿ) øàáſſſſà-ſðåäſà, äſáàâſÿÿ ê ſſſ, âſçſſæſſ, ſâſſ ſàáſð ſſſåé
è ſãðàſſ÷åſſé. Øàáſſſ-ſðåäſê ſſæåò áûòü àáſòðàſòſûſ (ò.å. ſſſſſüçſâàòüſÿ
â ſà÷åſòåå ſáſåãſ ſðåäſà äðóãſò øàáſſſſſâ-ſàſſëåäſſſâ). Sſſçäàâàòü ſáúåſòû
àáſòðàſòſſãſ øàáſſſà ſåëüçÿ. Íàſðſàſð, åſſë ſſëüçſâàòåëü õſ÷åò ſàäåëſòü âſå
ſáúåſòû ſſſòàſſ÷åſſſ ðàçſåòſ ſſëåſ «ſſſàſíòàðſé», ſſ ſſæåò ſſðåäåëſòü ýòſ
ſſëå ó ſáſåãſ àáſòðàſòſſãſ øàáſſſà «ſſſòàſſ÷åſſſ ſáúåſò» è âſâåſòſ èç ýòſãſ
øàáſſſà äðóãſå øàáſſſû.

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 ïëåé, ïîäïîëé 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. Ñîîòâåòſòâóþùèå îáúåêòû ìîæåò áûòü çíà÷åíèåì òàêîãî
ïîëÿ, åñëè îí ðåàëèçóåò òàêîé òèï. Ýòî äàåò ïîðåäåëåíèþ ãèáêîñòü â ïîðåäåëåíèè
ìåòàäàííûõ. Âî-âòîðûõ, â ïðîãðàììå èìåþòſÿ âîçìîæíîſòü ïðîâåðèòü, ðåàëèçóåò ëè
äàííûé îáúåêò óêàçàííóþ íàäſòðîéêó, âûâåſòè ſïèſîê íàäſòðîâê, ðåàëèçóåìûõ
äàííûì îáúåêòîì, à òàêæå âûâåſòè âſå îáúåêòû, ðåàëèçóþùèå äàííóþ íàäſòðîéêó.
Ñàìè ýòè îáúåêòû ìîãóò èìåòü ðàçíûå øàáëîíû; èõ îáúäàëíÿåò ëèøü òî, ÷òî ïðè
âûïîëíåíèè óñëîâèé âõîæäåíèÿ ìû îòíîſìè èõ ê äàííîé íàäſòðîéêå. Ïîýòîìó
íàäſòðîéêè óäîáíî ðàſſſìàòðèâàòü êàê ſïèſîſûé çàïðîſſâ ê äàííûì, òî åſòü
òàêèõ çàïðîſſâ, êîòîðûå âſçâðàùàþò ſòàåëüſûé ſſèſſîſ îáúåêòîâ. Ïðèìåðſſ ſſæåò
ſëóæèòü íàäſòðîéêà «ſſîäëåæàùåå», â êîòîðóþ âõîäèò «ſëîâſîſſîðſà» (ſðè íàëè÷èè óſëîâèÿ
íà ſàäåæ), à òàêæå øàáëîſ «ſſſëü» áåç óſëîâèé.

Âſÿêèé îáúåêò èſååò
îáÿçàòåëüſóþ òåêſòſâóþ êſſſſſåſòó «ſſſäåðæàſèå». Ñſſäåðæàſèå îáúåêòà ſſæåò
çàäàâàòüſÿ ſſëüçſſâàòåëå, ëèáſ âû÷èſëÿòüſÿ ſſ ſſðåäåëåſſſûſ ſðàâèëàſ ÷åðåç
ſſſäåðæàſèÿ ſſëåé. Îáúåêòû èſåþò òàêæå ſſſåöèàëüſſûå äåſſêðèſòſðû äëÿ
ſſſðòèðſâêè è ſðàâſſåſèÿ. Îſè, â ÷àſòſſſòè, ſſçâſſëÿþò ſòðſèòü çàſðſſſû è
îãðàſè÷åſèÿ ſà ſſðÿäſê ſëſâ (ſàſðèſåð, ſàéòè âſå ſâÿçè «ſóáſòàſòèâ-àòðèáóò», â
êſòſðûõ ſóáſòàſòèâ ſàõſäèòſÿ ðàſüøå àòðèáóòà).

Sſëÿ øàáëſſſâ ſſãóò
áûòü òðåõ âèäſâ: ſâû÷ſûå ſſëÿ, êſëëåêöèè è äèàſàçſſû. Sſëå
êſëëåêöèÿ ſòëë÷àåòſÿ ſò ſâû÷ſſãſ ſſëÿ òåſ, ÷òſ ſðåäſſſëàãàåò ſðàçſ
ſåſêſëüêſ ðàçëë÷ſûõ çſà÷åſèé (ſàſðèſåð, ſàſſſðſäſûå ÷ëåſû). Äèàſàçſſ – ýòſ
«ſâÿçſàÿ» êſëëåêöèÿ, òſ åſòü ſſſæåſòâſ ſâúåêòſâ, èáóùèõ ſſäðÿä â ſſſûſëå
óſſðÿäſ÷åſèÿ ſſ äåſêðèſòſðàſ. Äëÿ äèàſàçſſà äſſòàòſ÷ſſ çàäàòü ſà÷àëüſûé è
êſſå÷ſûé ſâúåêò. Òèſè÷ſûé ſſðèſåð äèàſàçſſſâ – ſòðſſêè â ſòðàſèöå èëè êàêèå-ëèáſ
åſòåſſòâåſſſûå ſâÿçſûå áſëüøèå ôðàſìåſòû òåſſòà. Sſëÿ øàáëſſſâ òàêæå äåëÿòſÿ ſà ſâÿçàòåëüſûå
è ſſöèſſàëüſûå. Îáÿçàòåëüſſ̃å ſſëå çàſſſëſÿåòſÿ ſðè ſſçäàſèè ſâúåêòà
(ſàſðèſåð, ſðè ſèſòàêſè÷åſêſé ðàçſåôå). Äëÿ ſſöèſſàëüſûõ ſſëåé ſðåäëàãàåòſÿ
ſſèſſſê âſçſſæſûõ âàðèàſòſâ çàſſëſåſèÿ. Äàſſûé ſſèſſſê ôſðſèðóåòſÿ ſà ſſſſâå
ſâðàſè÷åſèé øàáëſſà è óæå çàſſëſåſſûõ ſſëåé.

Åſëè âſå êàſäëäàòô
íàäſòðîéêè èſåþò ſâúèå ſſëÿ, òſ ſðè çàſèſè óſëſâèÿ íà ſſëå òèſà ýòſé íàäſòðſéêè
òàêèå ſſëÿ ſſæſſ èſſſſëüçſâàòü â âûðàæåſèÿõ. Êðſſå òſãſ, íàäſòðſéêè ſſãóò èſåòü

ñâîè (îïöèîíàëüíûå) ïîëÿ. Îáúåêò ïðèíàäëåæàò 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] = '÷èñëèòåëüíîå').

(Çäåñü
òàê çàïèñàííûå 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 îáùèìè). Ïóñòü, íàïðèìåð,
ïîëå «ïîäëåæàùåå»
ìîæåò áûòü âûðàæåíî êàê ñëîâîôîðìé, òàê è íóëåì. Ïóñòü øàáëîíû «ñëîâîôîðìà» è «íîëü», áåçóñëîâíî, âõîäÿò â
íåêîòîðóþ íàäñòðîéêó. Ó øàáëîíà «íîëü» íåò ïîëÿ «ïàäåæ»; ñîîòâåòſòâåííî, ïîëå
«ïàäåæ» åñòü òîëüêî ó øàáëîíà «ñëîâîôîðìà». Ïîýòîìó óñëîâèå íà «ïîäëåæàùåå»
ìîæíî çàïèñàòü òàê:

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

Èíòåðôåéñ ïðîãðàììû ïðåäñòàâëåí
ïàíåëüíè îáúåêòîâ. Ýòè ïàíåëè áûâàþò ðàçíûõ âèäîâ; èõ îñíîâíàÿ çàäà÷à –
îòîáðàæàòü ñïåöèàëüíûì îáðàçîì ïîðåäëåííûå îáúåêòû. Íà äàííûé ìîìåíò â
ïðîãðàììå ïðåäóñìîòðåíû òàêèå âèäû ïàíåëåé îáúåêòîâ, êàê ïàíåëü íàâèãàöèè, ïàíåëü
îñíîâíîãî òåêſòà, ïàíåëü-ñïèñîê, ïàíåëü ñâîéſòâ è ò. ä. Âçàèìîñâÿçè ìåæäó íèìè è èõ ïîâåäåíèå îïèſûâàåſòſ â
îòäåëüíîì xml-ôàéëå.

Ðèſ. 1. Ïàíåëè
ïðîãðàììû. Âûäåëåííûå îáúåêòû è ïîäſâåòêà ſâÿçàííûõ îáúåêòîâ

Äëÿ ðàáſòû ïîëüçſâàòåëü
ìîæåò âûäåëÿòü â ïàíåëÿõ ãðóïïû îáúåêòîâ. Êàæäàÿ ãðóïïà îſíçíà÷åſòſ ſâſèì
öâåſòſ.

Ñèſòåìà áóäåſ
ſſâåðøåſſñòſſâàòüſſÿ äàëüøå. Ïðåäïîëàãàåſòſ âſåäðèòü â íåå ìåõàíèçìû

ïîëóàâòîìàòè÷åñêîé ðàçìåòêè, ðàáîòû ñ «ïðåäèêàòíûìè», 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 ïèñüìåííûå 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.