Èäåîëîãèÿ è òåõíèêà ðàçìåòêè â Òåõíîëîãèè ñìåøàííîãî íàáîðà

Àâòîð Àëåêñàíäð Âèêòîðîâè÷ Êîâàëåíêî
27.06.2008 ã.
Ïîñëåäíåå îáíîâëåíèå 16.07.2008 ã.

Òåçèñû â ôîðìàòå RTF (57.69 kB 2008-07-13 21:38:16) Òåçèñû â ôîðìàòå PDF (122.14 kB 2008-07-13 21:37:33)

Òåõíîëîãèÿ ñìåøàííîãî íàáîðà (ÒÑÍ) åñòü ñïîñîá ðàáîòû ñ òåêñòàìè, ñîäåðæàùèìè ôðàãìåíòû ðàçëè÷íûõ ïèñüìåííûõ ñèñòåì. Ñ òåõíè÷åñêîé òî÷êè çðåíèÿ ìîæíî ãîâîðèòü ïðîñòî î ôðàãìåíòàõ, òðåáóþùèõ ðàçíîé èíòåðïðåòàöèè, ÷òî ñðàçó äåëàåò ôîðìóëèðîâêó ïðîáëåìû áîëåå îáùåé è àêòóàëüíîé è äëÿ òåêñòîâ, îäíîðîäíûõ ïî èñïîëüçóåìîé ïèñüìåííîñòè. Äëÿ ðàáîòû ñ òàêèìè òåêñòàìè òðåáóåòñÿ ðåøåíèå òð¸õ âçàèìîñâÿçàííûõ çàäà÷: 1) ñîçäàíèå åäèíîé ïëàò-ôîðìû äëÿ ïðåäñòàâëåíèÿ â îäíîì ôàéëå ðàçíîãî ðîäà ôðàãìåíòîâ; 2) íàõîæäåíèå ñïîñîáà ðàçáèåíèÿ òåêñòà íà ôðàãìåíòû; 3) ïðåäîñòàâëåíèå ïîëüçîâàòåëþ ñïîñîáà îïèñàíèÿ èíòåðïðåòàöèè ôðàãìåíòà.

Öåëü äàííîé ðàáîòû – ñîñðåäîòî÷èòüñÿ íà âòîðîñòè ðàçìåòêè è å¸ èíòåðïðåòàöèè, ðàññìîòðåâ èõ îáùåå îïèñàíèå, è ïîêàçàòü âîçìîæíîñòè äëÿ ïîñëåäíåãî ðàáîòû èññëåäîâàòåëÿ ïðèíÿòîãî â ÒÑÍ ïîäõîäà â ñðàâíåíèè ñ äðóãèìè èìåþùèìèñÿ ïîäõîäàìè ê ðàçìåòêå, ïðåæäå âñåãî XML.

1. Òåðìèíû.
Òåêñò ìû ïîíèìàåì êàê çàôèêñèðîâàííóþ ïîñëåäîâà-òåëüíîñòü çíàêîâ. Òåêñò ïîäâåðãàåòñÿ èíòåðïðåòàöèè ÷åëîâåêîì èëè àâòîìàòîì ñ êîíêðåòíîé ïðàêòè-÷åñêîé èëè èññëåäîâàòåëüñêîé öåëüþ. Êàæäîå èññëåäîâàíèå îïèðàåòñÿ íà èíôîðìàöèþ, ñâÿçàííóþ ñ êîíêðåòíûìè ìåñòàìè â òåêñòå, è âñÿêàÿ ñìåíà ñïîñîáà èíòåðïðåòàöèè íóæäàåò-ñÿ â óêàçàíèè ìåñòà òàêîé ñìåíû. Ñî-âî-êóï-íîñòü óêàçàíèé íà ìåñòà â òåêñòå, ñâÿçàííûõ ñ êîíêðåòíîé èíòåðïðåòàöèåé, ìû íàçûâàåì ðàçìåòêîé. Òàêèì îáðàçîì, ñ òåêñòîì ìîæåò áûòü ñâÿçàíî ìíîãî ðàçìåòîê, îòâå÷àþùèõ ðàçíûì öåëÿì.

Âàæíûìè ïðèìåðàìè ðàçìåòêè ÿâëÿþòñÿ ðàçáèåíèå òåêñòà íà ñîñòàâëÿ-þùèå åãî ôðàãìåíòû è îòðàæåíèå ñòðóêòóðû òåêñòà â âèäå èåðàðõèè. Ñâÿçü èåðàðõèè ñ òåêñòîì îáóñëîâëåíà âçãëÿäîì íà òåêñò (öåëüþ èíòåðïðåòàöèè), òàê ÷òî ðàçíûå ðàçìåòêè îäíîãî òåêñòà ìîãóò ïîäðàçóìåâàòü ðàçíûå èåðàðõèè. Íî â îáùåì ñëó÷àå ðàçìåòêà – ýòî íå ðàçáèåíèå è íå èåðàðõèÿ, à òîëüêî íàáîð óêàçàíèé íà ìåñòà â òåêñòå.

Ïî ñïîñîáó ñâÿçè ñ òåêñòîì ðàçìåòêè äåëÿòñÿ íà äâà êëàññà: âíåøíèå è âíóòðåííèå.

Âíåøíÿÿ ðàçìåòêà – ýòî íàáîð ññûëîê íà ýëåìåíòû òåêñòà, õðàíèìûé îòäåëüíî îò òåêñòà. Òàêàÿ ðàçìåòêà íàèëó÷øèì îáðàçîì ñîîòâåòñòâóåò íåçàâèñèìîñòè îáúåêòà (òåêñòà) îò åãî îïèñàíèÿ, îäíàêî îáåñïå÷èâàÿ íåçàâè-ñèìîñòü ðàçìåòîê. Ïðèìåðîì âíåøíåé ðàçìåòêè ìîæåò áûòü ïåðå-÷åíü ìåñò â ïîñëåäîâà-òåëüíîñòè çíàêîâ, âûðàæåííûé íîìåðàìè ñëîâ òåêñòà, áóêâ ñëîâà.

Âíóòðåííÿÿ
ðàçìåòêà –
òàêîé ñïîñîá ðàçìåòêè, ïðè êîòîðîì ìåñòî â òåêñòå îò-ìå÷àåòñÿ âíåñåíèåì ïðÿìî â ïîñëåäîâàòåëüíîñòü çíàêîâ òåêñòà ñïåöèàëüíûõ ìåòîê – ýëåìåíòîâ ÿçûêà ðàçìåòêè. Ïðè ýòîì òåêñò óæå íå ÿâëÿåòñÿ íåèçìåííûì îáúåêòîì, à èçìåíÿåòñÿ ñ äîáàâëåíèåì êàæäîé íîâîé ðàçìåòêè. Âìåñòå ñ âíóòðåííåé ðàçìåòêîé óäîáíî ðàññìàòðèâàòü è íåÿâíóþ ðàçìåòêó – âèäèìûå è íåâèäèìûå ýëåìåíòû èñõîäíîãî òåêñòà, êîòîðûå òîæå ìîãóò ñëóæèòü èññëåäîâàòåëÿì îïîðîé ïðè èíòåðïðåòàöèè.

Âíåøíÿÿ
ðàçìåòêà èìååò äâà íåóäîáñòâà: 1) äëÿ å¸ ðåäàêòèðîâàíèÿ íåîáõîäèìî ñïåöèàëüíîå ïðîãðàììíîå îáåñïå÷åíèå (ÑÏÎ), 2) ðåäàêòèðîâàíèå ñàìîãî òåêñòà âëèÿåò íà âñå èìåþùèåñÿ ðàçìåòêè, â òîì ÷èñëå ñîçäàííûå íåçàâèñèìûìè èññëåäîâàòåëÿìè è âíå ÑÏÎ. Òåì íå ìåíåå, âíåøíÿÿ ðàçìåòêà èñïîëüçóåòñÿ â ïðàêòèêå, òàê êàê å¸ âàæ-íûì äîñòîèíñ-âîì ÿâëÿþòñÿ ïîääåðæêà ìíîæåñòâà íåçàâèñè-ìûõ ðàçìåòîê è íåèç-ìåí-íîñòü (çàùèù¸ííîñòü) ñàìèõ òåêñòîâ êîðïóñà. Èçâåñòåí è ïîäõîä, ïðè êîòîðîì âíåøíÿÿ è âíóòðåííÿÿ ðàçìåòêè ïåðå-âîäÿòñÿ äðóã â äðóãà – äëÿ ðåäàêòèðîâàíèÿ âñå âíåøíèå
ðàçìåòêè âíîñÿòñÿ â òåêñò, ñòàíîâÿñü âíóòðåííèìè, à âíóò-ðåí-íèå ðàçìåòêè äëÿ õðàíåíèÿ è èíòåðïðåòàöèè ñíîâà ïåðåâîäÿòñÿ âî âíåøíþþ ôîðìó. Â êîðïóñå ìîãóò èñïîëüçîâàòüñÿ è ðàçìåòêè îáîèõ êëàññîâ, òåì áîëåå ÷òî âíåøíÿÿ ðàçìåòêà ìîæåò ñ óñïåõîì îïèðàòüñÿ íà ÿâíóþ è íåÿâíóþ âíóòðåííþþ ðàçìåòêó.

Ñâîè íåóäîáñòâà
èìåþò è âíóòðåííèå ðàçìåòêè. Âî-ïåðâûõ, â òåêñòå îíè ïðèñóòñòâóþò âñå ñðàçó, â òîì ÷èñëå íå íóæíûå êîíêðåòíîìó èññëåäîâàòåëþ, ÷òî äåëàåò îñòðûì âîïðîñ î íàãëÿäíîñòè ñèíòàêñèñà ðàçìåòêè. Âî-âòîðûõ, âñòà¸ò âîïðîñ î çàùèù¸ííîñòè èñõîäíîãî òåêñòà, êîòîðûé ìîæåò ñíèìàòüñÿ òàêæå ïîñðåä-ñòâîì èñïîëüçîâàíèÿ ñïåöèàëüíûõ ðåäàêòîðîâ ðàçìåòêè. Êðîìå òîãî, êîíêðåòíûå âèäû ñèñòåì âíóòðåííåé ðàçìåòêè íàêëàäûâàþò ñâîè îãðàíè÷åíèÿ íà âîçìîæíîñòè ðàçìåòêè. Òàê, ñðåäè ÿçûêîâ ðàçìåòêè íà îñíîâå SGML, áëàãîäàðÿ íàëè÷èþ âîçìîæíîñòè îòäåëüíî çàäàâàòü èíòåðïðå-òàöèþ ýëåìåíòîâ ðàçìåòêè, íàèáîëüøåå ðàñïðîñòðà-íåíèå ïîëó÷èë XML. Îí, îäíàêî, â ñâîåé îñíîâå ïðåäúÿâëÿåò ê ñàìîé ðàçìåòêå æ¸ñòêèå òðåáî-âàíèÿ, ïðàêòè-÷åñ-êè èñêëþ÷àþùèå èñ-ïîëüçî-âàíèå ìíîæåñòâà íåçàâèñèìûõ ðàçìåòîê. Â èòîãå äëÿ ðàçíûõ èññëåäîâàòåëü-ñêèõ è ïðàêòè÷åñêèõ çàäà÷ ðàçìåòêà äîëæíà äåëàòü-ñÿ íà ðàçíûõ ýêçåìïëÿðàõ èñõîäíîãî òåêñòà, ÷òî ñíèæàåò íàä¸æíîñòü ðàáîòû.

2. Ðàçìåòêà
â ÒÑÍ – âíóòðåííÿÿ. Å¸ ïðèíöèïû ñêëàäûâàëèñü ïîä âëèÿíèåì ñòðåìëåíèÿ: à) ê ëàêîíè÷íîñòè ðàçìåòêè (èñõîäíûé òåêñò äîëæåí îñòàâàòüñÿ êàê ìîæíî ëó÷øå ÷èòàåìûì); á) êî ìíîæåñòâåííîñòè ðàçìåòîê â îäíîì òåêñòå (ðàáî÷àÿ ôîðìà òåêñòà äîëæíà îñòàâàòüñÿ åäèíñòâåííîé äëÿ ðàçíûõ çàäà÷); â) ê íåçàâèñèìîñòè ýòèõ ðàçìåòîê (íîâàÿ âíîñèìàÿ ðàçìåòêà íå äîëæíà âëèÿòü íà èíòåðïðåòàöèþ ðàíåå âíåñ¸ííûõ); ã) ê ïðîñòîòå îïèñàíèÿ èíòåðïðåòàöèè ðàçìåòêè (îïèñàíèå äîëæíî áûòü äîñòóïíûì äëÿ íåïðîãðàììèñòîâ).

À. Ëàêîíè÷íîñòü
îáåñïå÷èâàåòñÿ ñëåäóþùèìè ñðåäñòâàìè.

Âî-ïåðâûõ,
èíòåðïðåòàöèÿ òåêñòà ïîèðà-åòñÿ íå òîëüêî íà ÿâíóþ ðàçìåòêó ñòàíäàðòíîãî
ôîðìàòà, íî è íà íåÿâíóþ, êàê è íà ëþáûå äîïîëíèòåëüíûå ñèìâîëû èëè êîìáèíàöèè
ñèìâîë-ëîâ. Íå íóæäàþòñÿ â ñïåöèàëüíûõ ìåòêàõ è îñîáûå ñëó÷àè èíòåðïðåòàöèè,
èäåíòèôèöèðóåìûå ïî ñî÷åòàíèþ ìåòêè ñ ýëåìåíòàìè èñõîäíîãî òåêſòà, òàê êàê
òàêîìó ñî÷åòàíèþ ïîæíî ïðèäàòü îñîáóþ èíòåðïðåòàöèþ òåì æå ìåõàíèçìîì, ÷òî è
ñàìîé ìåòêå.

Âî-âòîðûõ,
õîòÿ ýëåìåíòû ðàçìåòêè îòêðûâàþò ñîáñòâåííûå ôðàãìåíòû òåêſòà, ê íèì íå
ïðåäúÿâëÿåòſ ñòðîãîå òðåáîâàíèå èìåòü ïàðíûé (“çàêðûâàþùèé”) ýëåìåíò (÷òî
íå ìåøàåò èſïîëü-çî-âàòü ïàðíîſòü ïðè íåîáõîäèìîſòè, ïîäðàçóìåâàÿ å¸ ïðè
îſèſàíèè èíòåðïðåòàöèè).

Ýòè
ðåøåíèÿ äåëàþò ðàçìåòêè â òåêſòå òåîðåòè÷åſêè ìèíèìàëüíûìè: ìåòêà ſòàâèòſÿ òàì
è òîëüêî òàì, ãäå âîçíèêàåò íåîáõîäèìſòü â ſìåſå èíòåðïðå-òàöèè, íå óſìàòðèâàåìàÿ
èç äðóãèõ ìàðêåðſâ. Â ðåçóëüòàòå íåêîſ-ðûå çàäà÷è ſîãóò ðåøàòüſÿ áåç âíåſíèÿ
äîſſîëíèòåëüſíé ðàçìåòêè â óæå èſåþùèåſÿ êſðſóſà.

Á. Ìſſæåſòàåſſſſòü
ðàçìåòſê îáåſſïå÷èâàåòſÿ âſçſſæſſſòüþ ſåðåä êàæ-äûì ſſſå÷åſſſûì ſðàãſåſòſſ
âſòàâèòü ſſâûé, ſëóæåáſſûé ſðàãſåſò ſſ ſâſåé ſåò-êſé. Òàêèå ſðàãſåſòſ
ſèſſòàêſè÷åſêè íå ſòëè÷àþſſÿ ſò ſðàãſåſòſâ èſõſïäſſſãſ òåêſòà (è ſſſóò íåſòè ëþáſå
êſſëè÷åſſòâſ èſſſðſàöèè, ſſſïäåðæàòåëüſſſé äëÿ ſſïäåðæèâàåſſſſãſ äàſſſſé ðàçſåòêſé
àſſſåêòà). Ïðè ýòſſ, åſſëè ſåòêå ſðàãſåſòà íå ſſðèäàſſ íèêàêſſãſ çſà÷åſèÿ ſſðè
ſſſèſàſſèè èſſòåðſðåòàöèè, òſ âåſſü ſðàãſåſò (à íå òſëüêſ ſåòêà) èãſſſðèðóåòſſÿ.
Òàêèſ ſſáðàçſſ, äſſáàâëåſèå â òåêſò ſſâûſ ſåòſê íå âëèÿåò ſà óæå ſſóùåſſòâóþſèå èſſòåðſðåòàöèè, â ſſſèſàſèè êſòſðûſ ſſðſ ýòè ſåòêſ
íè÷åſſ íå ſêàçàſſ.

Â. Íåçàâèſèſſſòü
ðàçſåòſê ðåàëèçóåòſÿ, êðſſå óêàçàſſſé ſſſſſſſáſſſſòè ſåſàſèçſà èſòåð-ſðå-òàöèè
èſſſſðèſðſâàòü “÷óæèå” ſðàſſſåſòſ, òåſ, ÷òſ ſſ (ſåſàſèçſ) äſſſóſêàåſ è ſðſèçâſſëüſſſå
çàäàâàſèå òðàêſſâêè ſàſſſſãſ ðàçáèåſèÿ íà ſðàſſſåſ-òſ, ſſçâſëÿÿ ſ÷èòàòü
ſóùåſòâåſſſûſè äëÿ ðàçáèåſèÿ òſëüſſ íóæſſûå ſåòêſ.

Ã. Ïðſſòſòà
ſſèſàſèÿ èſòåðſðåòàöèè ðàçſåòêſ ſſåſſïå÷èâàåòſÿ òåſ, ÷òſ ſ òſ÷êſ çðåſèÿ
ðåäàê-òſ-ðà êſðſóſà ſſà ſſſſòſèò â âſſſèſſûâàſèè â ſðſſſòſſ òâêſòſ-âſſ ſâéëå
(“ſòèëåâſſ ſâéëå”) äëÿ êàæäſé ſåòêſ ſſſſſſòàâëåſèÿ åé öåſſ÷êſ ſèſàſſëſâ è ſðåſáà-çſ-âàſèé,
êſòſðûå ſëåäóåò ſðèſåſèòſ ſ ððàſàſåſòó. Ïðåſáàçſſâàſèÿ, ſ êſòſ-ðûſ èä¸ò ðå÷ü, óſàçſûâàþſſÿ
ſâſèſ èſåſåſ, çà êſòſðûſ ſſæåò ſòſÿòſ ëèáſ ſðſſãàſſſſàÿ ôóſêſëÿ, ëèáſ
ſåſðſſãàſſſſå ſðåſſáſàçſſâàſèå, çàäàâàåſſſå ſåðå÷-ſåſ ſðſſſòſûſ èëè ðåſóðſſèſſſûſ
çàſſåſ â ſſſåöèàëüſſſé òåſſòſſâſſ ôſðſå. Ñſſòàſſſ÷åſêſ èſåſà òåſ ëèë äðóãſſ
ſðåſſáſàçſſâàſèé íå ðàçëè÷àþſſÿ, òàê ÷òſ
ðåäàêòſð êſðſóſà ſſæåò ſſ-ſâſåſó ſåðåſſſðåäåëÿòſ êàêèå-òſ èç ſſſòàſſſûſ ôóſêſëé.

Ïðèìåðàìè
íåïðîãðàììíûõ ïðåîáðàçîâàíèé ìîáóò ñëóæèòü: ïåðåâîä ïðåäñòàâëåíèÿ ñèìâîëîâ â òó
èëè èíóþ øðèôòîâóþ êîäèðîâêó, óäàëåíèå óäàðåíèé, ðàññòàíîâêè ïåðåíîñîâ è/èëè
ìîðôåìíûõ ãðàíèö, ëåììàòèçàöèÿ, ñíÿòèå/ðàñïðàíîâêà òèòë è äðóãèå îïåðàöèè,
ñïåöèôè÷íûå äëÿ îðôîãðàôè÷åñêîé òðàäèöèè è ïðåäñòàâëÿþùèå ñëîæíîñòü íå ñòîëüêî
àëãîðèòì÷åñêóþ, ñêîëüêî ïðåäìåòíî-ñïåöèàëüíóþ. Íåñìîòðÿ íà òåõí÷åñêóþ
ïðîñòîòó, ñàìè ïðåîáðàçîâàíèÿ, ïîëó÷åííûå òàêèì îáðàçîì, ìîãóò áûòü äîâîëüíî
ñëîæíûìè, âî âñÿêîì ñëó÷àå çà íèì îáû÷íî ñòîèò áîëüøîé òðóä ñïåöèàëèñòà,
êîòîðûé ïðåäñòàâëÿåò ñîáîé çàêîí÷åííûé ðåçóëüòàò. Ïðåäëàãàåìàÿ ôîðìà äà¸ò âîçìîæíîñòü
ñïåöèàëèñòàì îáìåíèâàòüñÿ è òàêèè ðåçóëüòàòàìè.

3. Ïðîöåññ
èíòåðïðåòàöèè ðàçìåòêè. Ñòèëåâîé ôàéë ïðèìåíÿåòñÿ íà âòîðîì ýòàïå
èíòåðïðåòàöèè. Ïåðâûé è òðåòèé ýòàï çäàâþ óñëîâèÿ è öåëü ðàáîòû ðàçðàáîò÷èêà
ðàçìåòêè.

Ìåòêè
ñòàíäàðòíîãî âèäà íà÷èíàþòñÿ ñïåöèàëüíûì ìàðêåðîì “íà÷àëî” (÷àñòî ýòî çíàê
“%”). Ïåðâûé ýòàï èíòåðïðåòàöèè ñòàâèò ïðåä íèì äîïîëíèòåëüíûé ìàðêåð
“êîíåö”, òàê ÷òî íà ãðàíèöàõ ôðàãìåíòîâ îêàçûâàåòñÿ êîíòàìèíàöèÿ ýòèõ ìàðêåðîâ.
Íà òðåòüåì ýòàïå óäàëÿþòñÿ âñå ó÷àñòêè òåêſòà ìåæäó ýòîé êîíòàìèíàöèåé è
áëèæàéøèì ê íåé ìàðêåðîì “êîíåö”. Ïîýòîìó çääà÷à âòîðîãî ýòàïà, êîòîðûé
îïðåäåëÿåò ñîñòàâèòåëü ñòèëåâîãî ôàéëà, – ïðîïèñàòü â í¸ì òàêå çàìåíû, â
ðåçóëüòàòå êîòîðûõ íóæíûå ôðàãìåíòû íå òîëüêî íå óäàëèëèñü áû íà òðåòüåì ýòàïå,
íî è ñîäåðæàëè áû â ñåáå ïðåäïèñàíèÿ íóæíûõ ïðîïóñêàíèé. Èñïîëüçóÿ â òàêèõ
çàìåíàõ è ñàìì ìàðêåðû “íà÷àëî” è “êîíåö”, ìîæíî ãàáêî âëèÿòü íà ñëîæèâøåñÿ â
ðåçóëüòàòå ïðåñå÷åíèÿ ðàçìåòîê ïðîâîíà÷àëüíîå “ðàçáèâíå”.

ÒÑÍ
ðåàëèçîâàíà íà ïëàòôîðìå PHP, ÷òî áûëî âûçâàíî å¸ ïåðâîíà÷àëüíûì
íàçíà÷åíèåì – ïðåäñòàâëÿòü â èíòåðíåòå â ðàçíûõ âèäàõ äðåâíåðóññêèå ïåñíîïåíèÿ.
Ñïåöèôèêàöèè, íåîáõîäèìûå äëÿ ïåðåíîñà Òåõíîëîãèè íà äðóãèå ïëàòôîðìû,
ïðåäñòàâëåíû íà ñàéòå ïðîåêòà “Ôîíä
çíàìåííûõ ïåñíîïåíèé” (URL: http://znamen.ru/tsn/tsnp-zam.php).

Ideology and technique for encoding using the mixed scripts technique

Alexander V. Kovalenin

A. P. Ershov Institute of Informatics Systems of the Siberian Branch of
the Russian Academy of Science, Novosibirsk, Russia

This paper discusses common problems in text encoding
and proposes their solution using a “mixed scripts technique” (MST), a way of
working with texts containing parts which need different interpretation. MST’s
way of text markup, together with a procedure of markup interpretation, is
given in comparison with other solutions like XML. A special form of
non-programmed user-defined transformations is shown, which can become common
for carrying out certain kinds of scientific research.