ÈÑÏÎËÜÇÎÂÀÍÈÅ TEI È RDF Â ÈÍÔÎÐÌÀÖÈÎÍÍÎ-ÏÎÈÑÊÎÂÎÉ ÑÈÑ-ÒÅÌÅ «ÐÓÑÑÊÈÅ ËÈÒÅÐÀÒ

Àâòîð Àðòåì Âèêòîðîâè÷ Àíäðååâ
27.08.2012 ã.
Ïîñëåäíåå îáíîâëåíèå 27.08.2012 ã.

Summary. The conceptual data model of the information retrieval system «Russian Literature of XVIIIth century» challenges efficient indexing. The system is based on a set of TEI P4 conformant documents where available search criteria are derived from the mark-up. Initial relational approach proved to be very inadaequate and was replaced by a RDF-based system. Standard RDF schemata are used as much as possible but are too coarse-grained in many cases, so a custom schema was devised.

Èíôîðìàöèîííî-ïîèñêîâàÿ ñèñòåìà «Ðóññêàÿ ëèòåðàòóðà XVIII âåêà» îáåñïå÷èâàåò ïîèñê òåêñòîâ è ôðàãìåíòîâ òåêñòîâ ïî ðÿäó ïà
ïàðàìåòðû è ò.ï. [Àíäðååâ è äð., 2009] Îñíîâîé ÈÏÑ ÿâëÿåòñÿ êîðïóñ òåêñòîâ, ðàçìå÷åííûõ â ñîîòâåòñòâèè ñ ðåêîìåíäàöèÿìè TEI P
ÿâëÿåòñÿ òî, ÷òî íàáîð ïîèñêîâûõ ïàðàìåòðîâ íå çàäàí æåñòêî, à ïîðääåëÿåòñÿ èñïîëüçóåìûìè òåãàìè TEI.

Íåñìîòðÿ íà îòíîñèòåëüíî íåáîëüøîé îáúåì êîðïóñà, âîïðîñ ýôôåêòèâíîé îðãàíèçàöèè èíâåðòèðîâàííûõ èíäåêñîâ âñòàë óæå íà ðàííèõ
ñèñòåìû âûÿâèë ñëåäóþùèå ñëàáûå ìåñòà:

-
òðåáîâàíèå íîðìàëèçàöèè ìîäåëè íàêëàäûâàåò îãðàíè÷åíèÿ íà èñïîëüçóåìóþ ðàçìåòêó, ëèáî ïðèâîäèò ê ÷åðåñ÷óð ãðîìîçäêèì ñõåìà

-
äèíàìè÷åñêàÿ ïðèðîäà íàáîðà ïàðàìåòðîâ ïëîõî ñî÷åòàåòñÿ ñ íåîáõîäèìîñòüþ èìåòü æåñòêóþ ñõåìó ÁÄ

-
îáëàñòüþ äåéñòâèÿ ïàðàìåòðà ìîæåò áûòü êàê òåêñò, òàê è ïðîèçâîëüíûé åãî ôðàãìåíò, ÷òî òàêæå ïëîõî âûðàçèò â ðåëÿöèîííé

-
îáíîâëåíèå ïîèñêîâûõ èíäåêñîâ îêàçûâàåòñÿ íåòðèâèàëüíîé çàäà÷åé.

Ïîýòîìó ïîñëå ðÿäà ýêñïåðèìåíòîâ áûëî ïðèíÿòî ðåøåíèå ñëåäîâàòü ïîãõîäó, íàìå÷åííîìó â [Tummarello et al., 2005] è ïðåäñòàâëÿòü
ïîëó÷èòü RDF-ïðåäñòàâëåíèå, ïîëíîñòüþ ýêâèâàëåíòíîå TEI-ðàçìåòêå, à òîëüêî îáñíîâíûå ïîçîæèòèâíûé äîñòîãî íàõîæäåíèÿ òåê

Ìû ñòàðàëèñü ïî âîçìîæíîñòè ñëåäîâàòü ñóùåñòâóþùèì ñòàíäàðòàì è ðåêîìåíäàöèÿìè â îáëàñòè ìåòàäàííûõ, îäíàêî îá äëÿ âñåõ ïàð

Dublin
Core Terms äëÿ ïðåäñòàâëåíèÿ áèáëèîãðàôè÷åñêîãî ïîèñàíèÿ è ñòðóêòóðèðîâàíèÿ òåêñòîâ

OAF
äëÿ ïðåäñòàâëåíèÿ èìåí ñîáñòâåííûõ

Äëÿ íåêîòîðûõ ïàðàìåòðîâ, â ïåðâóþ î÷åðåäü ñòèõîâåä÷åñêèõ, ãîòîâûõ RDF-ýëåìåíòîâ, êàæåòñÿ, íå ñóùåñòâóåò âîîáùå, à äëÿ íà
ñîáñòâåííûõ (íàïðèìåð, èìåíà èñòîðè÷åñêèõ ëèö vs èìåíà ìèôîëîãè÷åñêèà). Ñ äóóãîé ñòîðîíû, îòíîøåíèå dcterms:isPartOf, åñòåñòâî
îñîáûõ ÷àñòåé òåêñòà, íàïðèìåð, ýïèãðàôîâ. Ïîýòîìó íàìè áûëà ðàçðàáîòàíà ñîáñòâåííàÿ RDF-ñõåìà, (a) ðåàëèçóþùàÿ íåäîñòàþùè
ìåòàÿçûê RDFS, ïîñêîëüêó èñïîëüçîâàíèå OWL â íàøåì ñëó÷àå ïðèâîäèëî áû òîëüêî ê óâåëè÷åíèþ íàêëàäíûõ ðàñõîäîâ.

Ïîñëå îïðåäåëåíèÿ ìîäåëè äàííûõ âñòàë âîïðîñ î âûáîðå ïëàòôîðìû ðåàëèçàöèè, êîòîðàÿ áû ïîçâîëÿëà ëåãêî îïåðèðîâàòü êàê SGML

êàê âñòðîåííûé SGML-ïàðñåð (ðåàëèçóþùèé ïîäìíîæåñòâî ïîëíîãî SGML, äîñòàòî÷íîå äëÿ îáðàáîòêè äîêóìåíòîâ TEI P4), òàê è ðàç

Áëàãîäàðíîñòè

Ðàáîòà âûïîëíåíà ïðè ôèíàíñîâîé
ïîääåðæêå ÐÔÔÈ, ãðàíò ¹–07–00493à.

Ñïèñîê
ëèòåðàòóðû

Àíäðåâà è äð.,
2009 Àíäðåâà À. Â., Áóõàðêèí Ï. Å., Ìàòâåâà Å. Ì., ÏîíîìàðåâàÌ. Â. Î ðàçðàáîòêå íîâîé òåîðåòè÷åñêîé ìîäåëè ðåïðåçåíòàöèè èñòî
ÑÏá, 2009. Ñ. 303–310

Tummarello et
al., 2005 Tummarello G., Morbidoni
C., Pierazzo E. Toward textual encoding based on RDF // Proceedings of the
9th ICCC International Conference on Electronic Publishing. Leuven-Heverlee, 2005. Pp. 57–64.