НАУЧНИ РАД

УДК 004.439:004.55XML
УДК 519.76

СМЕРНИЦЕ ИНИЦИЈАТИВЕ ЗА КОДИРАЊЕ ТЕКСТА И ЊИХОВА ЛОКАЛИЗАЦИЈА

Томаж Ерјавец*
Институт Јожеф Штефан, Одсек за технологије знања
Љубљана, Словенија

Превод са словеначког: Неда Бабић

Апстракт: Рад представља "Смернице иницијативе за кодирање текста", формалну спецификацију и пратећу документацију за речник XML елемената који је намењен за обележавању текстова у научне сврхе. TEI је у широкој употреби за кодирање најразноврснијих типова текстова, поготово сложених текстова у дигиталним библиотекама. Рад описује историју и организација Иницијативе за кодирање текста (TEI), структуру Смерница и даје неколико примера употребе. У раду се разматра и питање локализације: док TEI елементи могу да се користе за описивање текста на било ком језику, сâме Смерница су написане на енглеском језику. Међутим, TEI нуди више могућности за превод делова Смерница о којима се расправља на крају рада и предлаже нова, једноставна алтернатива пуном преводу.

Кључне речи: XML, TEI, стандардизација кодирања текста, локализација.

*tomaz.erjavec@ijs.si

1. Увод

Дуго времена су текстови у дигиталној форми били посматрани само као дигитални еквивалент штампаног примерка и били су тесно повезани са софтвером помоћу кога су настали или са киме су требали бити представљени. Такви текстови нису били погодни за машинску обраду и брзо су застаревали. Да би се решили ови проблеми створен је ISO стандард SGML који је пружао средства за репрезентацију текста која не зависе ни од платформе ни од излаза. SGML је допуштао корисницима да дефинишу сопствене скупове етикета погодних за кодирање произвољних типова текстова. Најпознатија примена SGML-а је, без сумње, HTML (језик за обележавање хипертекста) који се користи за кодирање докумената ради приказивања у веб прелистачима. Али ни SGML ни HTML нису били погодни као широка основа за кодирање произвољних текстова у различите сврхе: SGML је био превише сложен док је HTML дефинисао само мали речник елемената који су углавном намењени визуелном приказу.

Ово су били разлози због којих је W3C (World Wide Web Consortium) створио нови стандард, проширивши језик за обележавање XML.[1] XML је подскуп SGML-а и као такав, за разлику од HTML-а, спецификује средство за дефинисање произвољних скупова етикета. Како је XML у исто време много једноставнији од SGML-а писање софтвера који ће га обрађивати је постало много лакше. Откако је настао, XML је постао веома успешан као формат за размену података у дигиталном облику са мноштвом софтвера и повезаних стандара који подржавају његов развој, валидацију и трансформације.

Од ових стандарда, требало би поменути XSLT,[2] језик за трансформацију XML докумената; скриптови у XSLT-у трансформишу документе у другачије структуриране документе. Како се XML пре свега користи као формат за складиштење и размену података, XSLT омогућава његову трансформацију у формате који се могу директно применити, на пример у HTML за веб прелистаче.

XML је, као и SGML, мета-језик: он не дефинише одређен скуп етикета већ само обезбеђује средства за њихово дефинисање, а који ће се скуп етикета дефинисати зависи од типа података који се обрађују и предвиђеног начина употребе, иако постоји могућност да истраживачи дефинишу своје скупове етикета, или тачније, XML шеме које спецификују етикете (XML елементе и атрибуте) и валидне међусобне везе етикета, углавном је боље користити раније дефинисане, стандардне шеме јер су оне пажљиво дизајниране, имају добру документацију и одржавање, гарантују повезивање са другим апликацијама, а може их обрађивати и постојећи, њима намењен софтвер.

До сада већ постоји велики број стандардних XML шема за најразличитије области примене, као што су апликације за мобилне телефоне, за представљање математичких формула, за музичке нотације, и за техничку документацију.

Али шта је са текстовима из подручја хуманистичких наука, тј. са оним текстовима који су предмет истраживања у хуманистичким наукама, независно од тога којој врсти текста припадају, на ком језику су написани, из ког периода потичу или које методе се користе у њиховој анализи? До данас, постоји само један стандард (или боље речено, препорука) који покрива ову широку област, а то је TEI - Иницијатива за кодирање текста.

2. Иницијатива за кодирање текста

Иницијатива за кодирање текста (TEI) је развила општи стандард који представља изражајно средство за представљање текстова у дигиталном облику и који одговара истражи-

[1] http://www.w3.org/XML/

[2] http://en.wikipedia.org/wiki/XSL_Transformations

вачким потребама научника у хуманистичким наукама. TEI се може окарактерисати као:[3]

– слободно доступан скуп смерница за кодирање текстова у хуманистичким наукама уз употребу XML-а

– међународни конзорцијум који постоји да би подржао развој ових смерница

– заједницу пројеката и појединаца који користе TEI Смернице

Као и многи покушаји стандардизације и TEI се суочава са изазовима који су својствени оваквом типу пројеката. Како могу тако бројне области и заједнице у оквиру хуманистичких наука да пронађу заједничко упориште у једном језику за кодирање? Како можемо да се договоримо колико детаљно или погодно је потребно описати наш текстуални материјал? Како да помиримо предности које доносе доследност и договор са потребом за специјализацијом у појединачним случајевима? Како се носити са нечим што је стварно јединствено и неочекивано?

За разлику од многих покушаја стандардизације TEI ова и слична питања решава тако што отворено прихвата варијације и расправу у оквиру својих техничких могућности. TEI Смернице су смишљене тако да буду модуларне и прилагодљиве, па специфични пројекти могу да изаберу делове TEI који су битни за њих а остало да занемаре. Исто тако, уколико је неопходно, могу се креирати додаци TEI језика који ће обрадити аспекте текста којима се TEI још није бавио. Пошто је сам TEI сложен, процес прилагођавања није сасвим тривијалан, али је замишљен тако да буде што је могуће једноставнији.

2.1. За шта се TEI користи?

TEI Смернице се претежно употребљавају за стварање дигиталних библиотека које омогућавају приступ великој количини текстуалног материјала при чему се акценат ставља на редак и осетљив материјал који једино дигитално може да постане широко доступан. Кодирање текстова у оваквим колекцијама наглашава својства која ће бити од користи за претрагу и проналажење, као што су библиографски подаци, предметне кључне речи и други мета-подаци који корисницима помажу да пронађу материјал који их интересује.

TEI Смернице користе и специјализовани истраживачки пројекти за представљање мањих колекција текстова које су тематски усмерене. Оне се често односе на неки одређени жанр, или на аутора, период или земљу (или представљају њихову комбинацију). Пројекти овог типа често користе детаљније обележавање да би представили особена својства текста која су релевантна за одређену колекцију или су важна за аудиторијум научника којем су намењени. На пример колекција која се ограничава на дела неког аутора чији радови представљају важан извор информација о познатим савременицима може да обележава сва помињања имена и радова, уз могућност укључивања веза које ће упућивати на детаљније индексе биографских или критичких информација. Слично томе, електронско издање одређеног аутора или дела може да обухвати и представљање различитих читања, ревизија аутора, уредничких исправки и сличних уредничких информација. Намена неких колекција ове врсте је да подрже веома специфичне истраживачке циљеве, као што је лингвистичка анализа, или треба да послуже као основа за речник и у овим случајевима обележавање може да буде веома уско специјализовано.

Све употребе које су овде описане представљају неку врсту издавања: циљ је да се створи дигитална колекција коју ужа или шира публика може да користити online. Коришћење може да буде ограничено на малу заједницу корисника или на преплатнике, а може и да буде отворено за широку публику. Ређе, TEI

[3] Овај одељак прати одличан увод у TEI који је доступан на http://www.wwp.brown.edu/encoding/seminars/tei.html

користе појединци да би створили дигиталну презентацију текстуалног материјала да би подржали сопствена истраживања, у облицима који се могу објавити али и не морају. Док циљ и сврху великих колекција могу да условљавају корисници или начин финансирања, у случају радова појединаца ограничења су лична и професионална: обележени материјал може да служи као лични алат за истраживање а може се и развити у нешто што одговара дигиталној монографији која представља ауторову анализу скупа текстова. Употреба TEI Смерница у овим случајевима може бити онолико детаљна колико далеко је аутор спреман да иде: једино ограничење представљају време, енергија, маштовитост и исплативост.

2.2. Учење TEI

TEI веб станица[4] је добар извор општих информација о TEI и место где се могу пронаћи TEI Смернице. Али TEI веб станица је само један од многих извора корисних информација о томе како треба користити TEI и како разумети његов значај. Иако не постоји један извор који може дати потпуну слику, има све више литературе која се бави улогом обележавањем текста у научном раду. Библиографија Универзитета Браун[5] даје корисне изворе али је можда најбољи начин за стицање основних знања о TEI похађање неке радионице. Сваке године се одржава много таквих радионица, углавном у Сједињеним Америчким Државама и Великој Британији и оне се рекламирају на TEI веб станици.

За оне којима је потребно детаљније разумевање самог процеса обележавања, радионице су добар почетак али оне нису довољне. Учење TEI је као и учење језика: TEI има прилично обиман речник и сложен распон употребе. Нека уводна радионица даје добар увид у могућности TEI језика и упознаје полазнике са основним терминима. Али то мора бити праћено како практичним радом тако и детаљним истраживањем коришћења самог језика. Од велике помоћи је постојање конкретног пројекта у виду скупа докумената од интереса на којем ће се радити. Да би се упознао са TEI Смерницама у току корисник може уз читање Смерница, да ради још неколико ствари. TEI одржава листу слања TEI-L, где ће скоро сва питања, чак и она која постављају почетници, добити одговоре. Листа се и архивира, па се одговори на питања која се најчешће постављају као и она која се ређе постављају могу наћи у архиви. Опсег различитих ставова и приступа такође може да допринесе бољем схватању како све различите врсте пројеката користе TEI. Добар начин за учење TEI је и изучавање рада на неким конкретним пројектима обележавању текстова. Многи пројекти имају документацију, а неки чак и изузетно добру документацију која образлаже разлоге за доношење неких одлука при обележавању и даје критеријуме по којима се препознају и обележавају одређена својства текста. Многи ће радо поделити са другима узорке својих обележених текстова који могу бити веома корисни за стицање увида у комплетну слику обележавања.

3. Историјат TEI[6]

TEI је покренут 1987. године на састанку у колеџу Васар, који је окупио различите групе научника из много различитих дисциплина и представнике водећих професионалних друштава, библиотека, архива и пројеката из многих земаља Европе, Северне Америке и Азије. Резултат почетне фазе њиховог рада је издавање 1990. године првог нацрта Смерница, по-

[4] http://www.tei-c.org/

[5] http://www.wwp.brown.edu/encoding/seminars/readings.html

[6] За детаљнији приказ историје TEI видети http://www.tei-c.org/About/history.xml на који се овај одељак ослања

знатог као P1. Одмах затим је започета друга фаза рада, а резултати су објављивани од 1990. до 1993. године. Затим, након следећег циклуса ревизија, проширивања и додатавања објављења је 1994. године прва званична верзија Смерница „P3". Како се све више научника упознавало са Смерницама, њихови коментари, исправке и захтеви за проширењима су стизали са свих страна света. На крају је било скоро 200 научника из различитих дисциплина, професија и земаља који су чинили језгро групе која се бавила развојем TEI Смерница.

TEI Конзорцијум је установљен 2000. године. То је организација с међународним чланством која сада одржава, наставља да развија и промовише TEI. Циљ оснивања TEI Конзорцијума је био да се створи стални дом за TEI као демократски установљену, академски и економски независну, непрофитну организацију која се сама одржава.

Након оснивања TEI Конзорцијума, приоритет је био издавање XML верзије TEI Смерница као унапређене верзије P3 која је још била заснована на SGML-у како би се корисницима омогућио рад са XML алатима који су почели да се појављују. Верзија P4 Смерница је објављена 2002. године. Суштински, то је била XML верзија P3, у којој нису битно промењена ограничења израженим у шемама осим оних које су биле неопходне због самог преласка на XML и исправка уочених грешака у тексту P3 Смерница. Па ипак, имајући у виду да је P3 до тада већ био у сталној употреби од 1994. године, било је јасно да је потребна темељна ревизија његовог садржаја, па је одмах започет рад на верзији P5. Она је била замишљено као генерална поправка која је укључивала и јавни позив корисницима за својствима и нови развој у скупу пресудних области као што су кодирање карактера, графика, описа рукописа и језика на коме су сâме TEI Смернице написане. Верзија P5 Смерница је издата крајем 2007. године.

Утицај TEI на дигитално истраживање је био огроман. Данас је TEI међународно признат као веома важан алат, како за дуготрајно чување електронских података тако и као средство које подржава ефикасну употребу таквих података у многим предметним областима. То је шема кодирања која представља прави избор при производњи критичких и научних издања литерарних текстова, за израду научних референтних радова и велике лингвистичке корпусе као и за управљање и производњу подробнијих мета-података који су придружени разним врстама електронских текстова и колекцијама културног наслеђа.

Многобројне организације су прихватиле TEI препоруке, а међу њима су и US National Endowment (Државна задужбина за друштвене науке САД-а), the UK's Arts and Humanities Research Board (Истраживачки одбор за уметност и културу Велике Британије), the Modern Language Association (Друштво за савремене језике), the European Union's Expert Advisory Group for Language Engineering Standards (Експертска саветничка група за стандарде језичког инжењерства Европске уније) као и многе друге агенције широм света која финансирају или промовишу пројекте везане за дигиталне библиотеке и електронске текстове. Конгресна библиотека је препознајући значај TEI у општем порасту заједнице дигиталних библиотека издала Смернице за најбољу примену TEI препорука за мета-податке у пракси, да би се постигла усклађеност са другим стандардима.

Успех TEI је доста допринео да се обезбеди да се наше културно наслеђе уведе у нови, умрежени свет у настајању и да постане доступно студентима, научницима и широј јавности.

4. TEI Смернице

TEI Смернице за кодирање и размену електронских текстова дефинишу и документују језик за обележавање који служи да предста-

ви структурна и концептуална својства текста, као и она својства која се односе на његово приказивање. Главни, али не и једини акценат је на кодирању докумената који припадају хуманистичким и друштвеним наукама, а посебно на представљању примарних извора за истраживање и анализу. Ове смернице су представљене у виду модуларне и прошириве XML шеме коју прати детаљна документација и објављене су под лиценцом отвореног кода.

TEI Смернице су доступне у неколико формата: штампана и повезана копија могу да се купе, док су онлине верзије у форматима HTML и PDF бесплатно доступне са званичног TEI сајта. Са станице TEI SourceForge могу се, такође преузети шеме, изворне XML датотеке Смерница, документација и сл. као запаковани пакети. Детаљнија упутства за приступ и употребу овог материјала могу се наћи на матичној страници TEI.

Саме TEI Смернице су писане у XML-у и следе парадигму писменог програмирања Доналда Кнута, која подразумева да исти документ садржи формалну спецификацију, као и пратећу документацију у облику везаног текста.

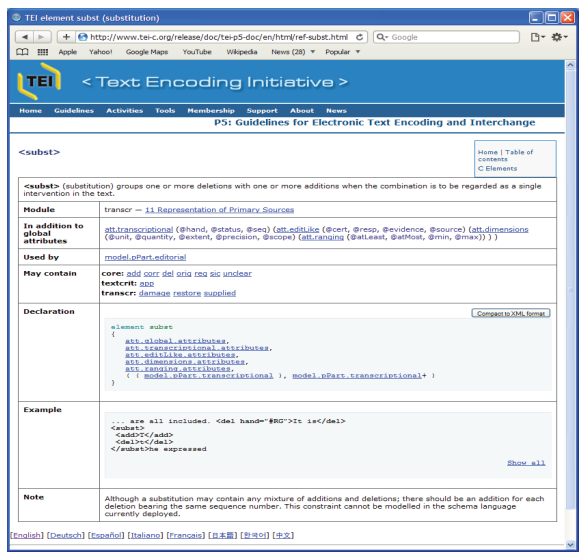
TEI Смернице дефинишу неколико стотина елемената и атрибута за обележавање докумената било које врсте. Свака дефиниција се састоји од следећих компоненти:

– текстуални опис,

– формална декларација изражена преко XML речника који је дефинисан у Смерницама и комбинован са елементима преузетим из ISO језика за RELAX NG шеме[7], и

– примери употребе.

Примера ради, слика 1. представља HTML изглед дефиниције TEI елемента <subst>. Она показује да елемент припада модулу за транскрипцију примарних извора (Transcription of Primary Sources) и даје хипервезу до пуног текста поглавља у којем је објашњен овај модул (о модулима се говори у следећем одељку). Текстуални опис објашњава да елемент "групише једно или више избацивања из текста са једним или више додавања када ту комбинацију треба посматрати као једну интервенцију у тексту". Дефиниција даље објашњава које атрибуте елемент користи, који TEI модел користи овај елемент, које елементе може да садржи, каква је формална дефиниција шеме тог елемента, и на крају даје пример употребе и напомене.



Слика 1. Дефиниција TEI елемента <subst>

4.1. TEI модули

Свако поглавље Смерница, осим што представља групу повезаних елемената, дефинише и одговарајући скуп декларација које се зову модули. Све дефиниције су окупљене у референтном одељку у додатаку Смерница. Формалне декларације за дато поглавље су окупљене у оквиру одговарајућег модула. Ради једноставности, сваки елемент је додељен једном модулу, обично да би се користио у неком специфичном пољу примене или како би подржао одређену врсту употребе. Према томе, модул је просто погодан начин за груписање већег броја повезаних декларација о елементима. У најједноставнијем случају TEI шема се добија комбиновањем мањег броја модула.

[7] http://www.relaxng.org/

У табели 1 наведени су сви модули које дефинишу TEI Смернице P5.

Име модула	*Опис модула*	*Поглавље смерница у коме се модул дефинише*
analysis	Анализа и интерпретација	17 Једноставни аналитички механизми
certainty	Извесност и неизвесност	21 Извесност, прецизност и одговорност
core	Заједничко језгро	3 Елементи доступни у свим TEI документима
corpus	Мета-подаци за језичке корпусе	15 Језички корпуси
dictionaries	Штампани речници	9 Речници
drama	Текстови који се изводе	7 Текстови који се изводе
figures	Табеле, формуле, слике	14 Табеле, формуле слике
gaiji	Документације о карактерима и глифовима	5 Представљање нестандардних карактера и глифова
header	Заједнички мета-подаци	2 TEI заглавље
iso-fs	Структуре својстава	18 Структуре својстава
linking	Повезивање, сегментација и поравнавање	16 Повезивање, сегментација и поравнавање
msdescription	Опис рукописа	10 Опис рукописа
namesdates	Имена, датуми људи и места	13 Имена, датуми људи и места
nets	Графови, мреже и дрвета	19 Графови, мреже и дрвета
spoken	Транскрибовани говор	8 Транскрипција говора
tagdocs	Елементи за документацију	22 Елементи за документацију
tei	Инфраструктура TEI	1 Инфраструктура TEI
textcrit	Критичка издања текста	12 Критички апарат
textstructure	Претпостављена структура текста	4 Претпостављена структура текста
transcr	Трансцрипција примарних извора	11 Представљање примарних извора
verse	Стихови	6 Стихови

Табела 1. TEI модули

4.2. Конструисање TEI XML шеме

Да би се утврдило да је неки XML документ валидан, а не само добро формиран, његова структура се мора проверити помоћу шеме. За валидан TEI документ ова шема мора да буде у сагласности са TEI шемом.

Спецификација за шему која је у сагласности са TEI је и сама TEI документ који користи елементе модула "Елементи за документацију". Такав документ се неформално назива "ODD" документом. Овај назив потиче од циља који је првобитно био постављен при дизајнирању система: "Један документ ради све" (One Document Does it all). Програме за обраду ODD докумената одржава TEI, а и câме Смернице су написане као такав документ.

Шему ODD чини избор TEI модула, појединих елемената се могу и забранити или им се може променити име, а могу се укључити и додатне декларације које модификују декларације елемената и атрибута које садржи сваки модул . Исти систем се може користити и за спецификовање шема које проширују TEI експлицитним додавањем нових елемената или упућивањем на друге XML речнике.

ODD документ може да обрађује ODD процесор који ће из овог скупа декларација генерисати одговарајућу XML шему, користећи неки од стандардних језика за шеме:

– језик XML DTD (део самог XML-а),

– језик ISO RELAX NG,

– језик W3C Schema,[8]

– или, у принципу, било који довољно изражајан језик за шеме.

Ове излазне шеме потом може користити било који XML процесор, на пример, валидатор или уређивач да би проверио или на други начин обрадио документе.

На званичној TEI станици доступан је ODD процесор који се назива Roma за који је

[8] http://www.w3.org/XML/Schema

обезбеђен веб интерфејс који помаже при креирању прилагођене TEI шеме.

5. Три примера TEI текстова

Док је у претходним одељцима TEI представљен у најширим цртама у овом одељку ћемо дати неке конкретне примере из нашег сопственог рада на словеначким текстовима.

Сви текстови о којима ћемо говорити су писани као XML документи који су у сагласности са TEI: за неке старије коришћен је TEI P4, док је за оне недавно завршене коришћен TEI P5. Механизам испоручивања се разликује од случаја до случаја, што зависи од конкретног сценарија за употребу, али у свим случајевима се заснива на XSLT стилским листовима који користе изворни TEI и трансформишу га у формат који омогућава његово коришћење, а то је обично HTML који је погодан за прегледање у веб прелистачу.

За следећа три примера о којима ћемо говорити ћемо прво укратко представити текстове или, боље речено, пројекате у оквиру којих су прикупљени, затим ћемо илустровати процес кодирања на малом примеру и на крају ћемо показати како су изворни XML текстови конвертовани за потребе конкретне употребе ресурса.

5.1. Библиотека eZISS

Дигитална библиотека eZISS[9] настала је у сарадњи Научно-истраживачког центра Словеначке академије наука и уметности и Института Јожеф Стефан. Садржи изабране, обично старије словеначке текстове са интегрисаним репродукцијама, преписима и научним коментарима, а у неким случајевима и аудио-визуелним записима. Ова издања су намењена за дуготрајну јавну употребу и бесплатно се могу прегледати на вебу, али се могу и преузимати као изворни TEI документи или извedени HTML документи. Пошто су објављени под лиценцом Creative Commons[10] ова издања се могу дистрибуирати и трећим лицима.

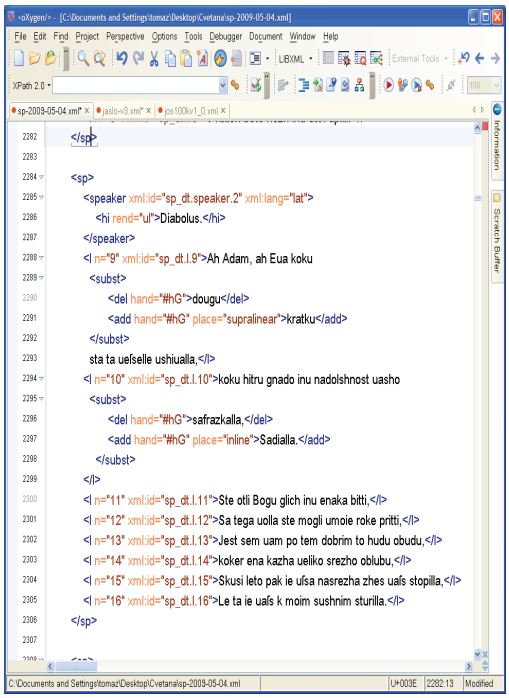
eZISS издања користе TEI шеме које се састоје од модула за опис рукописа (Manuscript Descriptions), за транскрипцију примарних извора (Transcription of Primary Sources) и за критичка издања текста (Textual Criticism) као и других модула, на пример, модула за текстове за извођење (Performance Texts).

Обележавање различитих издања није униформно јер су структуре текстова веома различите тако да су потребни различити елементи за њихов опис. Слика 2 која приказује кратак извод из eZISS “Processio Locopolis”, најстаријег текста намењеног извођењу на словеначком језику који је написан почетком 18. века укратко илуструје која врста обележавања се користи у нашим издањима. Извод приказује један говор (TEI елемент <sp>) из игроказа који изговара (<speaker>) Ђаво. Обратите пажњу да атрибут xml:id придружује јединствени идентификатор говорнику у овом говору, а да атрибут xml:lang означава језик на коме говорник прича као латински, где је “lat” трословни код за језик по ISO 639. Такође обратите пажњу да је означено да ”Diabolus” треба да буде истакнут док је захтевани изглед истицања подвлачење. Говор потом садржи осам редака (<l>), а сваком од њих је додељен текући број ретка у драми (атрибут n) и јединствени идентификатор.

Прва два ретка садрже и замену (<subst>, на слици 1 је дата дефиниција овог елемента) с којом је преписивач прецртао део текста и додао неки други текст (пун опис преписивача је дат у TEI заглављу, а вредност атрибута “hand” се односи на тај опис). Атрибут “place” означава место где је додатак направљен: у првом ретку унутар ретка, а у следећем изнад.

[9] http://nl.ijs.si/e-zrc/

[10] http://creativecommons.org



Слика 2. Текст једног говора из eZISS издања игроказа “Processio Locopolis”.

Да би се могла прегледати, издања су трансформисана у HTML документа помоћу XSLT стилских листова. Свако старије издање (P4) је имало сопствени стилски лист за трансформацију у HTML, док издања писана у TEI P5 имају придружене много једноставније стилске листове који само конвертују сложени изворни XML документ у поједностављени XML документ који је и даље у сагласности са TEI. Овај XML документ се потом трансформише у HTML помоћу стандардних TEI XSLT стилских листова који су доступни на званичној TEI станици.

5.2. Речник jaSlo

Јапанско-словеначки онлине речник “jaSlo”[11] развијен је у сарадњи Филозофског факултета Универзитета у Љубљани и Института Јожеф Стефан. То је речник за учење језика намењен словеначким студентима јапанског језика и тренутно садржи око 10.000 одредница.

Шема за речник jaСло користи TEI P4 модул за речнике. На слици 3 може се видети почетак једне одреднице ; као што се може видети, одредница прво садржи водећу реч на јапанском (<form type=”hw”>). Она се даље дели на запис у три различита писма: ромаји (транслитерација у латинично писмо), хирагана (фонетско јапанско писмо) и кањи (кинески идеограми). Граматичка група (<gramGrp>) даје врсту речи (глагол, класа 5) и његову подкатегоризацију (непрелазни). Затим следе три флективна облика глагола са суфиксима који одређују његово флективно понашање. Превод даје три преводна еквивалента на словеначком језику. Затим следи један пример на јапанском преведен на словеначки. Лексичке одреднице садрже и додатне информације: ниво тежине речи, упућивање на годину и поглавље у уџбенику у коме се реч први пут појавила, дефиниције (углавном за властите именице), етимологију, унакрсне упутнице, итд.

Слика 3. Почетак једне речничке одреднице речника jaSlo.

[11] http://nl.ijs.si/jaslo/

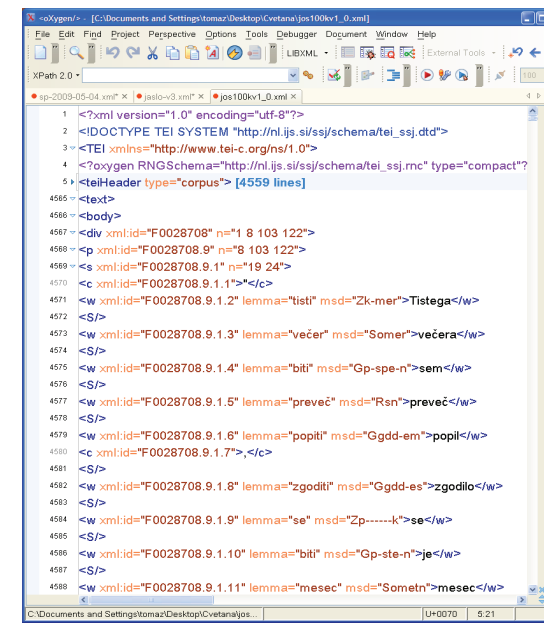
Речник је доступан онлине преко веб сучеља за претрагу. Сучеље је имплементирано као Perl CGI сервис који према критеријумима за претрагу претражује TEI кодирани речник и враћа релевантне одреднице и трансформише их из XML записа у HTML ради приказивања.

5.3. Корпус JOS

Пројекат JOS[12] развија лингвистички анотиране корпусе и придружене изворе словеначког чија је сврха да за словеначки језик олакшају развој технологија заснованих на људском језику. Тренутни резултати укључују морфосинтаксичке спецификације, два корпуса која су ручно анотирана на нивоу речи и два веб сервиса. Ови извори су доступни под лиценцом Creative Commons.

JOS корпуси садрже узорке пасуса из словеначког референтног корпуса FidaPLUS[13] који је анотиран недвосмисленим морфосинтаксичким описима и лемама. jos100k корпус садржи 100.000 речи и он је темељно ручно проверен, док jos1M садржи милион речи чије су анотације само делимично ручно проверене. Корпуси се могу користити као скупови података за обучавање тагера врста речи и лематизера за словеначки језик.

Корпуси су кодирани у TEI P5 са шемом која користи модуле за мета-податке за језичке корпусе (Metadata for Language Corpora), лингвистичку анализу (Linguistic Analysis) и структуре својстава (Feature Structures) са још неким JOS проширењима. Слика 4 приказује први део једне анотиране реченице из jos100k корпуса. Реченица садржи речи (<w>), интерпункцијске знакове (<c>) и размаке (<S>). Свака реч је анотирана морфосинтаксичким описом (msd) и лемом. Морфосинтаксички описи су компактне ниске које се директно разлажу у структуре својстава које су дефинисане у TEI заглављу. Тако се, на пример, морфосинтаксички опис "Zk-mer" разлаже у "zaimek vrsta=kazalni spol=moški število=ednina sklon=rodilnik" или, на енглеском језику "Pronoun Type=demonstrative Gender=masculine Number=singular Case=genitive", а на српском "zamenica vrsta=pokazna rod=muški broj=jednina padež=genitiv".



Слика 4. Почетак анотиране реченице "Tistega večera sem preveč popil, zgodilo se je…" из jos100k корпуса.

Корпуси су доступни у изворном XML формату, али и као табуларне датотеке које су произведене из изврних помоћу XSLT скриптова. Формат табуларних датотека је погоднији за обучавање тагера и лематизера. Као што је већ поменуто, JOS пројекат нуди и веб сервис за морфо-синтаксичко тагирање и лематизирање предатих текстова на словеначком језику, при чему су ова два алата обучена на JOS корпусима.

6. Међународно коришћење

TEI Смернице су широко прихватили пројекати и институције у многим земљама Европе, Северне Америке и Азије и оне се користе за кодирање текстова на десетинама језика.

[12] http://nl.ijs.si/jos/
[13] http://www.fidaplus.net/

TEI заједница је интернационална и вишејезична и она сваке године покрива све веће географско подручје. Међутим, сложено кодирање текстова у чему је TEI изванредан, захтева добро разумевање доступних елемената (којих има преко 500) и они који не владају енглеским језиком су у много неповољнијој ситуацији при учењу и употреби Смерница од енглеских говорника.

Зато TEI ради на стварању функционалне архитектуре за:

1) локализацију TEI извора;

2) локализацију стилских листова за испоруку референтног дела TEI Смерница;

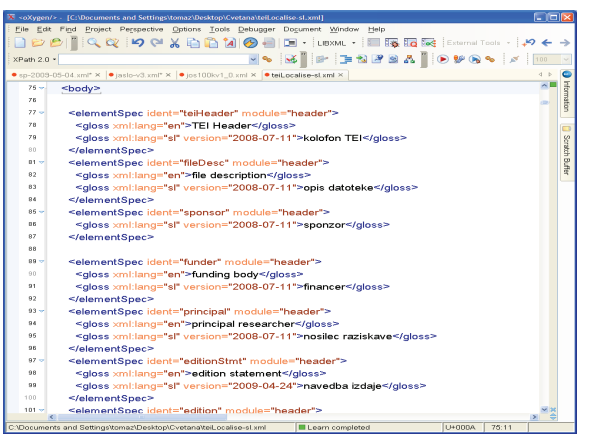
3) превод референтне документације TEI.

Овај рад је веома напредовао за шест великих језика, захваљујући донацији Удружења за књижевно и лингвистичко рачунарство ALLC (The Association for Literary and Linguistic Computing). Међутим, мало је вероватно да ће Смернице бити ускоро преведене на језике који имају мали број говорника као што је словеначки. Комплетно штампано издање има преко 1300 страница, што укључује референтни одељак од скоро 500 страница у коме се дају детаљни описи и дефиниције свих елемената.

Прецизно XML кодирање текстова такође представља специјализовану област рада тако да би број потенцијалних читаоца Смерница преведених на словеначки вероватно остао мали. Упркос томе што би било нереално очекивати комплетан превод Смерница или њиховог референтног дêла на сваки језик, постоје, ипак, неке области у којима се са малим напорима већ постижу корисни резултати.

За словеначки језик превели смо језичке ниске у TEI XSLT стилским листовима који производе HTML излаз, а који, такође, извршавају и задатке као што је прављење садржаја, подела великих TEI докумената у више HTML датотека итд, на пример “Table of Contents” или “Next” превели смо у “Kazalo” и “Naslednji” у словеначком (“Садржај” и “Следећи” на српском).

Такође смо превели називе свих TEI елемената на словеначки; ово не значи да су имена елемената у XML-у преведена већ да су уз дефиниције елемената приписане глосе на словеначком језику, као што је приказано на слици 5. Као што се може видети, структура спецификације једног елемената је овде прилично једноставна: елемент “teiHeader” припада модулу “header” и има глосу на енглеском “TEI Header” са словеначким преводом “kolofon TEI” (“TEI заглавље” на српском). Треба напоменути да, у ствари, само мали број TEI елемената има глосу на енглеском језику, јер већина елемената, као нпр. “sponsor”, имају назив који је истовремено и глоса.



Слика 5. Коришћење елемената за документацију за локализацију глоса TEI елемента на словеначки.

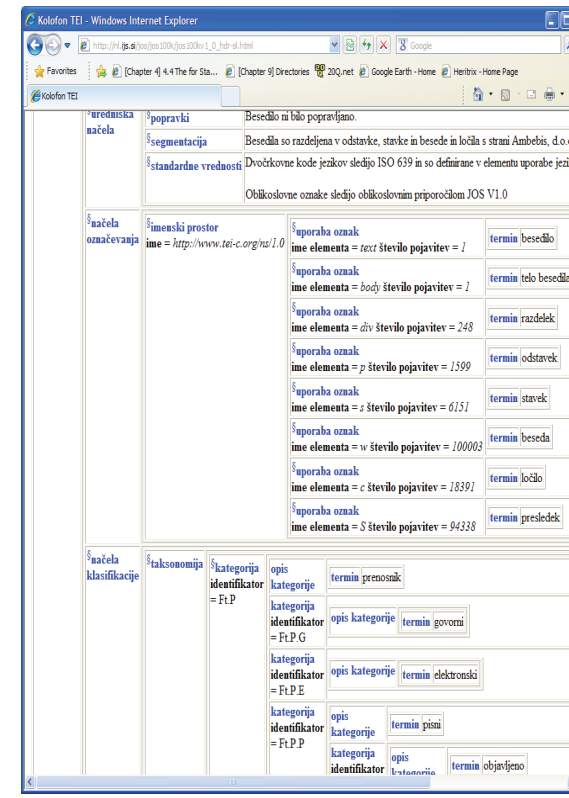
Датотека која садржи овакве скраћене спецификације свих TEI елемената, заједно са њиховим преводом на словеначки је доступна на вебу.[14] Надамо се да ће у будућности бити додати и други језици.

И док таква објашњења могу бити интересантна као додатак самим Смерницама, ми их користимо на другачији начин у нашим издањима. Написали смо XSLT стилски лист који конвертује TEI заглавље у једноставан HTML документ замењујући имена елемената њихо-

[14] http://nl.ijs.si/tei/localise/teiLocalise-sl.xml

вим глосама, на енглеском или на словеначком, или било ком другом језику који би био додат спецификационој датотеци. Ово омогућава превод мета-података документа; како типично TEI заглавље садржи и информације о употреби етикета у документу, у њему се налазе и преводи назива свих елемената који су употребљени у документу.

Слика 6 представља део TEI заглавља за jos100k корпус, у коме се види и део заглавља о употреби етикета. Заглавље jos100k је написано и на енглеском и на словеначком језику (разликовање језика се врши помоћу атрибута xml:lang) тако да XSLT стилски лист умеће у HTML садржај заглавља на одговарајућем језику, као и глосе елемената тог језика. Такве HTML датотеке се потом користе за представљање сваког издања, било на енглеском било на словеначком језику.



Слика 6. TEI заглавље корпуса jos100k трансформисано у HTML са енглеским и словеначким глосама имена елемената и садржајем заглавља.

7. Закључци

У раду је представљена Иницијатива за кодирање текста (TEI), међународни напор да се развије заједнички стандард за репрезентацију дигиталних текстова на начин који нуди мноштво могућности и истовремено одговара истраживачким потребама научника у хуманистичким наукама. Разматрали смо организацију TEI, њену историју, Смернице и конструисање шема, а дали смо и три примера из наше праксе у којима смо користили TEI шеме кодирања. Рад смо закључили дискусијом о међународном коришћењу TEI Смерница.

Представили смо "лаган" приступ локализацији који је погодан за мале језике у коме су глосе TEI елемената преведене да би се потом

могле користити, на пример, за HTML презентацију TEI заглавља. XML датотека са свим елементима, њиховим глосама на енглеском и одговарајућим преводима на словеначки језик, је слободно доступна на вебу и могу јој се додати преводи и на друге језике.

Као што се и може видети из овог рада, TEI Смернице су обиман и прилично сложен документ тако да је сасвим оправдано питање да ли вреди утрошити потребно време и труд да би се препоруке савладале и користиле и није ли можда боље да се за одређене пројекте развије сопствени начин кодирања који би савршено одговарао потребама конкретног пројекта. На основу наших искустава не можемо дати коначан одговор. За једноставније текстове које ће користити одређени појединац или који ће се користити у оквиру једне институције, одговор би могао бити негативан. Међутим, ако је потребно много времена за анотирање текстова, што их чини драгоценим и вредним за дугорочно чување, и чиме се ствара могућност да се они користе на разне начине, а посебно ако текстови треба да буду широко доступни у свом изворном облику, онда је TEI прави одговор: обележавање текстова се може формално проверити, по дефиницији је оно добро документовано а пројекти могу рачунати и на све више софтвера који је придружен TEI.

Референце

Овде дајемо листу најважнијих URL адреса које су у тексту поменуте. У овом раду не дајемо класичне библиографске референце до којих се, међутим, може доћи ако се прате доле наведене адресе. Свим станицама које су повезане с овим адресама је приступљено 10. 09. 2009.

XML – Проширив језик за обележавање (Extensible Markup Language):
http://www.w3.org/XML/

ISO Relax NG – Шема језик за проверу валидности XML докумената (Schema language for validating XML documents):
http://www.relaxng.org/

XSLT – XSL трансформације (XSL Transformations):
http://en.wikipedia.org/wiki/XSL_Transformations

TEI Consortium Web Site: Веб станица TEI конзорцијума:
http://www.tei-c.org/

eZISS – Дигитална библиотека словеначких књижевних текстова:
http://nl.ijs.si/e-zrc/

jaSlo – Јапанско-словеначки on-line речник за учење језика:
http://nl.ijs.si/jaslo/

JOS – Језички корпус словеначког за истраживања у области технологија заснованих на људским језицима:
http://nl.ijs.si/jos/

TEI element name translations into Slovene: Превод на словеначки имена TEI елемената:
http://nl.ijs.si/tei/localise/teiLocalise-sl.xml