Problem preklapanja hijerarhijskih struktura u postupku označavanja teksta

Jokić, Andrea

Sveučilište J.J. Strossmayera u Osijeku

Filozofski fakultet

Diplomski studij informatologije

Andrea Jokić

Problem preklapanja hijerarhijskih struktura u postupku označavanja teksta

Diplomski rad

Mentor: doc.dr.sc. Boris Bosančić

Osijek, 2014.

Sadržaj

Sažetak

Rad razmatra teorijske postavke označavanja teksta kao područja digitalne humanistike. Dan je prikaz razvoja označiteljske teorije u literaturi poznate kao „OHCO teorija“ u kojoj se tekst promatra kao hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu. OHCO teorija imala je ambiciju postati općom teorijom teksta, ali zbog nepredvidivih svojstava samog teksta u tome ipak nije uspjela. U radu se ukazuje na nedostatke ove teze i problem preklapanja hijerarhijskih struktura u tekstu koji se za navedenu teoriju pokazao pogubnim. Iako se modifikacijom teorije - najprije OHCO-2, potom i OHCO-3, pokušao riješiti problem preklapanja hijerarhijskih struktura u tekstu, ta ideja nikad u potpunosti nije realizirana budući da bi nakon svake revizije na površinu isplivale nove anomalije teorije. Kao moguća rješenja ovog problema, u tehničkom smislu, u okviru TEI zajednice razvijeni su mehanizmi temeljeni na XML-u kojima se pokušalo doskočiti ovom problemu, a njihov opis, kao i praktični primjeri, također su detaljno izneseni i obrađeni u ovom radu.

Ključne riječi: označiteljska teorija, označavanje teksta, OHCO teorija, XML, TEI

1. UVOD

Nedugo nakon pojave računala, kao posljedica mogućnosti korištenja i proučavanja humanističkih tekstova u elektroničkom okruženju nastaje digitalna humanistika – znanstvena grana u okviru humanističkih znanosti koja se bavi ulogom računala u znanstveno-istraživačkom radu. Poseban dio digitalne humanistike odnosi se na postupak označavanja teksta (engl. *text encoding*) koji se pored velike praktične važnosti i uporabe pokazao i kao uzbudljivo i teorijski produktivno područje analize i istraživanja. Ontološko pitanje *što je tekst?* bit će često spominjano u teorijskim raspravama o označavanju teksta, a kao pokušaji odgovora nastat će brojne teorije. Ovaj rad bavit će se jednom od često zastupanih teorija u kojoj se tekst promatra kao hijerarhija objekata sadržaja razvrstanih prema točno određenom redoslijedu. Svrha je rada iznijeti pregled označiteljske teorije koja je pretendirala postati općom teorijom teksta, ali koja zbog nepredvidivih svojstava samog teksta u tome ipak na kraju nije uspjela. Odatle proizlaze i ciljevi rada, ukazivanje na razloge 'pada' spomenute označiteljske teorije u kojoj se nastojala zastupati ideja o tekstu kao hijerarhiji objekata sadržaja razvrstanih prema točnom redoslijedu.

U skladu s tim, problematika rada izložit će se u tri velika poglavlja. Nakon uvodnog razmatranja i upoznavanja sa sadržajem rada, u drugom poglavlju govorit će se općenito o označavanju teksta, izložit će se razvoj postupka označavanja teksta, navesti vrste označavanja teksta, predstaviti proceduralni (LaTex) i opisni označiteljski jezici (SGML, XML) kao jezici koji drže primat na označiteljskoj pozornici. Na koncu poglavlja predstavit će se i Inicijativa za označavanje teksta (engl. *Text Encoding Initiative – TEI*) kao vodeći standard u označavanju teksta temeljen na XML-u. Poslije općenitog izlaganja o označavanju teksta u trećem poglavlju prikazat će se razvoj označiteljske teorije poznate pod akronimom OHCO (engl. *Ordered Hierarchy of Content Objects)* u kojoj se polazi od aksioma da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu. Ova teorija imat će važnu ulogu u promicanju opisnih označiteljskih jezika - posebice danas najzastupljenijeg XML-a te će poslužiti kao odgovarajući okvir za stvaranje TEI-a. Na kraju poglavlja zaključit će se zbog čega je OHCO teorija – koja je pretendirala postati općom teorijom teksta – ipak opovrgnuta. U četvrtom poglavlju predstavit će se modeli razvijeni u okviru TEI zajednice kojima se u tehničkom smislu pokušalo doskočiti problemu preklapanja hijerarhijskih struktura. Neki od mehanizama koji se nastoje uspješno nositi s problemom preklapanja hijerarhijskih struktura u tekstu su uporaba

milestone elemenata (tzv. 'praznih' elemenata), višestruko označavanje istog teksta, postupak fragmentiranja sadržaja i *stand-off* označavanje. Svi navedeni mehanizmi bit će teorijski i praktično predstavljeni. Također, pobrojat će se i neki od ne-XML pristupa rješavanja ovog problema. A na koncu, ukazat će se na neka nova promišljanja o tekstu i drugačije perspektive viđenja problema (platonizam, pluralizam, antirealizam) koji su nastali kao posljedica nemogućnosti da se u kontekstu OHCO teorije formulira prihvatljiva definicija teksta.

2. OZNAČAVANJE TEKSTA

2.1. Uvodna razmatranja o postupku označavanju teksta

Pojam „označavanja teksta“[1] može se dovesti u vezu s nekom vrstom obilježavanja, ukazivanja pa i umetanja dodatnih informacija u tekst. Svaki autor prilikom pisanja teksta ujedno ga označuje koristeći interpunkcijske znakove; ostavlja prazan prostor između riječi označavajući njihove granice, stavlja zareze kako bi označio pojedine fraze, rečenične dijelove i sl. Pod tekstom se s označiteljskog stajališta ne ubrajaju samo slova i riječi nego i svi drugi znakovi, simboli i sl. poput interpunkcijskih znakova ili simbola u obliku ilustracija kao specifičnih ne-tekstualnih dijelova knjige. Svi ti znakovi i simboli u određenom smislu ulaze u domenu onoga što se može označiti te predstavljaju svojstva teksta koja se mogu tretirati kao njegovi elementi i/ili obilježja.

Računalno označavanje teksta (engl. *computer text encoding*) može biti definirano vrlo jednostavno kao „reprezentacija tekstualnih informacija na računalu“.[2] Proces računalnog označavanja teksta sastoji se od pohrane lingvističkog sadržaja (najčešće alfanumeričkih znakova i interpunkcije) i *markupa* (računalnih oznaka, odnosno označiteljskog koda kojim se definira kako će sadržaj biti oblikovan).[3] M. Sperberg-McQueen opisuje sheme za označavanje teksta (engl. *text encoding schemes*) kao sheme koje pružaju mehanizme za reprezentaciju teksta i njegove logičke i fizičke strukturu te osiguravaju pomoćne informacije za analizu ili interpretaciju teksta.[4] Označavanje teksta predstavlja način stvaranja apstraktnih modela

[1] U terminologiji vezanoj uz postupak označavanja teksta termini „tekst“ i „sadržaj“ svojevrsni su sinonim te su međusobno zamjenjivi. Sukladno tome, sintagma „označavanje teksta“ može biti zamijenjena sintagmom „označavanje sadržaja“ bez da se ne izgubi ništa od prvotnog značenja.

[2] Usp. Renear, A. Out of praxis: three (meta)theories of textuality. New York, Oxford University Press, 1997. URL: http://books.google.hr/books?id=4dZ2XSr8Q2cC&pg=PA107&lpg=PA107&dq=Renear++Out+of+praxis+:+three+(meta)theories+of+textuality&source=bl&ots=15c0WWkRPg&sig=w5S45h87zoc39Scw8AYljG08C3A&hl=hr&sa=X&ei=uAuRUqDnEYba4QShnIDgAQ&ved=0CC0Q6AEwAA#v=onepage&q=Renear%20%20Out%20of%20praxis%20%3A%20three%20(meta)theories%20of%20textuality&f=true (2013-12-01) Str. 108.

[3] Usp. Renear, A. Out of praxis. Str. 108 – 109.

[4] McQueen, M.S., Text Encoding and Enrichment. Oxford: Oxford University Press, 1991. Citirano prema: Renear, A. H. Text encoding. A companion to digital humanities. Susan Schreibman, Raymond George Siemens and John M. Unsworth. Wiley-Blackwell, 2004. URL:

računalno čitljivog teksta i logičkih odnosa unutar teksta. Ti modeli teksta koriste se da bi se definiralo kako tekst treba bit strukturiran kad je napisan (pomoću SGML-a, XML-a), kako ga prikazati u web pregledniku (s HTML-om) ili da se ukaže na važne dijelove tekstualnih podataka (pomoću TEI-a).

Sam termin *markup* započinje se koristi u izdavaštvu, a nastao je od engleskog izraza *mark up* koji se na hrvatski može prevesti kao „označavanje". Pri uređivanju rukopisa za objavljivanje, rukopisi su se označavali (engl. *marks up*) dodavanjem anotacija ili simbolima kako bi se odredio njegov izgled u tiskanom obliku.[5] Iako su označiteljske oznake (engl. *markup codes*) u početku indicirale na postupak formatiranja teksta, kasnije se počinju primarno upotrebljavati za identificiranje elemenata teksta (npr. naslova, poglavlja, rečenica itd.), a tek onda, indirektno, za formatiranje i obradu.[6]

Pojavom i rasprostranjenijom primjenom računala proces formatiranja i ispisa tekstova automatiziran je te se termin *markup* širi i na elektronički kontekst odnosno njime se pokrivaju sve vrste oznaka kojima se formatira i oblikuje *elektronički* tekst.[7] Postupak označavanja teksta danas je aktivnost koja se veže uz elektroničko okruženje i stoga kada se u ovom radu govori o označavanju teksta ono se uvijek odnosi na označavanje elektroničkog teksta, odnosno sadržaja. Prema TEI smjernicama za označavanje teksta, *markup* se definira kao bilo koja eksplicitna interpretacija teksta.[8] Računalno označavanje teksta u principu je kao i stvaranje manuskripta iz *scriptura continua,*[9] to je proces u kojem se ističe ono što je vjerojatno ili se podrazumijeva te se usmjerava čitatelja kako bi se sadržaj teksta trebao interpretirati.

Softverski dizajneri i znanstvenici 1970-ih dolaze do zaključka kako je dizajniranje sustava koje će omogućiti učinkovitu i funkcionalnu obradu teksta najbolje da se temelji na stavu kako postoje određene karakteristike teksta koje su temeljne i važne, a sve one u obradi teksta trebaju

---

http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-5&toc.depth=1&toc.id=ss1-3-5&brand=default (2013-12-01)

[5] Usp. A Gentle Introduction to XML. // TEI P5: Guidelines for Electronic Text Encoding and Interchange. A TEI Consortium eds,2013. URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html (2013-12-01)

[6] Usp. Renear, A. Out of praxis. Nav.dj. Str. 109.

[7] Kada govorim o elektroničkom tekstu potrebno je razlikovati dva temeljna načina elektroničkog prikaza teksta: „digitalni faksimil" odnosno digitalizirane slike izvornika i „strojno čitljivi tekst" (omogućuje pretraživanje).

[8] Usp. Renear, A. Out of praxis. Nav.dj. Str 109.

[9] Obilježje rukopisa iz rane klasike, riječi su pisane kontinuirano bez razmaka ili interpunkcija.

biti zastupljene neizravno - kroz identifikaciju i obradu tih značajki.[10] Ove značajke nazvane su „objekti sadržaja“ (engl.*content objects*), a takav pristup obradi teksta „sadržajna obrada teksta“ (engl. *content-based text processing*).[11]

U obradi teksta 1980-ih godina počinju dominirati *interaktivni sustavi*. Osnovna karakteristika interaktivnih sustava bila je da uređivački softver (korišten za kreiranje i modificiranje teksta) odmah i automatski oblikuje tekst onako kako će on izgledati u tiskanom izdanju. Iako ovi sustavi također uključuju postojanje označiteljskog sustava, računalne oznake u interaktivnim sustavima su obično skrivene od korisnika. U interaktivnim sustavima korisnik indirektno i nesvjesno (odabirom u izborniku, klikom na određenu funkciju) stvara oznake koji se prikriveno umeću u dokument; ove oznake softver odmah potom obrađuje i tekst se istovremeno formatira u uređenu stranicu sa željenim izgledom (proredom, tipom fonta, veličinom i sl.). Iako korisnici interaktivnih sustava često zaključuju da zapravo, budući da nisu unijeli nikakav označiteljski kôd ili zato što isti ne vide, ne postoje nikakvi kodovi već oni, na neki način, jednostavno imaju direktan i transparentan pristup samom tekstu. Ipak, bilo da je vidljiv ili nevidljiv, uvijek postoji kompleksan sustav označavanja koji reprezentira značajke teksta i podržava prikaz i uređivanje istog, a svi koji obrađuju neki tekst na računalu neizbježno su uključeni u njegovo označavanje, bili oni toga svjesni ili ne.[12]

2.2. Vrste označavanja teksta

J. H. Coombs, A. H. Renear i S. J. DeRosea otkrivaju da autori redovito pri pisanju svojih tekstova koriste dvije vrste označavanja; interpunkcijsko i prezentacijsko označavanje. Ove, kako su ih autori nazvali, tradicionalne vrste označavanja teksta (engl. *traditional types of markup*) koje pojašnjavaju napisane izraze i njihovu uporabu, jednostavno je nemoguće izbjeći

[10] Usp. Biggs,M., Huitfeldt, C.: Philosophy and Electronic Publishing. Theory and Metatheory in the Development of Text Encoding. Edited Discussion. URL: http://www.philo.at/mii/mii/node8.html#SECTION00230000000000000000 (2013-12-01)
[11] Od 1950-ih do 1970-ih obrada teksta temeljila se na tzv. formatu temeljenu računalnu obradu (engl. *format-based text processing*) koja se zatim napušta i prelazi se na sadržaju temeljenu računalnu obradu.
[12] Usp. Renear, A. Out of praxis. Nav. dj., Str. 108 – 110.

budući da ih sustav pisanja zahtjeva. Ovakvo označavanje iako nije sastavni dio teksta govori nešto o njemu – ne čita se izravno nego se interpretira.[13]

Pojavom računalnih tekstualnih sustava javljaju se nove vrste označavanja i nove vrste obrade teksta. Kad je tekst pohranjen u elektroničkim datotekama on je obilježen s posebnim tipovima elektroničkih oznaka namijenjenih za obradu putem računalnih aplikacija.[14] Neki sustavi koriste označavanje kako bi ukazali na procedure koje bi određeni program trebao slijediti, drugi se koriste označavanjem da bi definirali komponente teksta[15], neki pak koriste označavanje kojim se upućuje na vanjski izvor informacija, dok se drugi služe označavanjem koje omogućuje deklaraciju elemenata[16] koji će se dalje koristiti u postupku označavanja. U skladu s tim, J. H. Coombs, A. H. Renear i S. H. DeRosea razlikuju sljedećih šest vrsta označavanja teksta:

a) Interpunkcijsko označavanje

Interpunkcijsko označavanje sastoji se od korištenja zatvorenog skupa oznaka za pružanje prvenstveno sintaktičkih informacija o pisanom tekstu. Prethodno je već spomenuto da autori pri kreiranju teksta obavljaju interpunkcijsko označavanje primjerice, kada rečenice završavaju točkama (.), upitnicima (?), uskličnicima (!) ili ih odjeljuju zarezom (,), dvotočkom (:), točka-zarezom (;) i sl.). Kao glavni nedostatak interpunkcijskog označavanja navodi se stilistička nedosljednost; svaki autor njeguje svoj stil s obzirom na upotrebu interpunkcijskih oznaka pa je ovaj sustav označavanja podložan znatnim stilskim varijacijama.[17] Još jedan nedostatak je što interpunkcijski znakovi mogu biti dvosmisleni.[18]

b) Prezentacijsko označavanje

[13] Usp. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Markup systems and the future of scholarly text processing // Communications of the ACM. 30, 11(1987). Str. 934. URL: http://cpe.njit.edu/dlnotes/CIS/CIS732_447/Cis732_6R.pdf (2013-12-01) Str. 935.
[14] Usp. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Nav.dj. Str. 935.
[15] Komponente teksta označavaju bilo koje strukturno svojstvo ili obilježje koje se može konkretno označiti poput naslova, odlomka, retka i sl.
[16] Elementi su oznake koje se koriste u postupku označavanja komponenti teksta.
[17] Usp. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Nav. dj. Str. 935.
[18] Primjerice točka (.) obično stoji kao oznaka kraja rečenice, ali se upotrebljava i da označi skraćenicu (prof., npr.).

Prezentacijsko označavanje odnosi se na prikaz teksta i okrenuto je prema krajnjem korisniku. Kako bi prezentacija teksta bila jasnija, tekst se obilježava na dodatne načine tako što se dodaju vertikalne i horizontalne praznine oko teksta, numeričke liste, brojevi stranica teksta, brojevi paragrafa te mnoštvo *ad hoc* simbola u samom tekstu ali i izvan njega (npr. inicijal koji se javlja na početku poglavlja glagoljičkih tekstova). Tipičan primjer oznaka prezentacijskog označavanja su tipografske funkcije poput podebljanja (*bold*), zakrivljavanja (*italic*) i podcrtavanja (*underline*) teksta.[19]

c) Proceduralno označavanje

U računalnim sustavima za obradu teksta prezentacijsko označavanje zamijenjeno je proceduralnim. Proceduralno označavanje sastoji se od naredbi ugrađenih u tekst koje govore tekstualnom procesoru na koji način će formatirati određenu komponentu teksta za prikaz.[20] Nedostataka ovakvog označavanja jest da je ono nefleksibilno (npr. ako se poželi promijeniti izgled dokumenta potrebno je ponoviti postupak označavanja kako bi promijene bile uočljive, što zahtjeva dodatne napore). Primjer naredbe proceduralnog označavanja izgleda ovako:

```
.skip 4;
```

Naredba iz primjera govori da se preskoče četiri linije. Delimiteri točka (.) i točka-zarez (;) osigurat će da računalo razlikuje računalne oznake, tj. označiteljski kod od teksta koji se označuje.

d) Opisno označavanje

Opisno ili deskriptivno označavanje teksta definira što određena komponenta teksta jest odnosno kojoj klasi unaprijed definiranih svojstava teksta pripada. Nerijetko se opisno označavanje miješa s proceduralnim, stoga valja ukazati na njihovu primarnu distinkciju: proceduralno označavanje sastoji se od naredbi ili instrukcija o načinu formatiranja te iste komponente teksta dok se opisnim označavanjem opisuje ili identificira određena komponenta teksta.[21] Primjer opisnog označavanja teksta izgledao bi ovako:

[19] Isto.
[20] Isto. Str. 936.
[21] Usp. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Nav. dj. Str. 937.

```
<head>Označiteljska teorija<head>
<p><q>Markup reflektira teoriju teksta</q></p>
```

Gdje je <head> oznaka naslova, <p> oznaka odlomka i <q> oznaka citata u odlomku.[22]

e) Referentno označavanje

Referentno označavanje upućuje na vanjski izvor informacija koji može biti pohranjen u zasebnoj datoteci ili na drugom računalu. Dio teksta označen referentnim označavanjem zamjenjuje se vanjskim izvorom informacija.[23] Referentno označavanje najčešće se povezuje s opisnim označavanjem teksta i svoju pravu primjenu doživjelo je u SGML-u. Naknadno će se u sklopu HTML dokumenta pojaviti poveznice kao predstavnice referentnog načina označavanja u mrežnom okruženju. Primjer poveznice:

```
<a href="http://www.tei-c.org/index.xml">Text Encoding Initiative</a>
```

f) Metaoznačavanje

Metaoznačavanje uključuje deklaraciju elemenata koji će se koristiti u postupku označavanja teksta.[24] Ono predstavlja temeljni mehanizam vezan uz razvoj SGML-a. U tom smislu SGML treba promatrati kao metajezik koji služi generiranju elemenata opisnih označiteljskih jezika. Deklaracije elemenata smještenih u zaglavlju dokumenta predstavljaju primjere metaoznačavanja, primjerice:

```
<!ELEMENT title (#PCDATA)>
```

[22] O opisnom označavanju bit će više riječ u nastavku rada u dijelu o opisnim označiteljskim jezicima.
[23] Isto.
[24] Isto.

2.3. Vrste označiteljskih jezika

2.3.1. Općenito o vrstama označiteljskih jezika

Opisno i proceduralno označavanje teksta danas drže primat na 'označiteljskoj pozornici'. Ta je reputacija za posljedicu imala pojavu i razvoj pripadnih označiteljskih jezika te u tom smislu razlikujemo proceduralne i opisne označiteljske jezike. U nastavku rada u kontekstu proceduralnih označiteljskih jezika detaljnije će biti predstavljen LaTeX kao najpoznatiji jezik ove označiteljske skupine. U kontekstu opisnih označiteljski jezika najprije će biti riječi o SGML-u kao jeziku koji je ugazio put razvoju danas najrasprostranjenijeg označiteljskog jezika na svijetu – XML-a. Na koncu, iznijet će se temeljne prednosti opisnih označiteljskih jezika u odnosu na druge vrste označiteljskih jezika.

2.3.2. Proceduralni označiteljski jezici

Najpoznatiji jezik proceduralne označiteljske skupine svakako je LaTeX. Ovaj proceduralni jezik razvio je L. Lamport 1985. utemeljivši ga na tada postojećem TeX-u D. E. Knutha. LaTeX predstavlja sustav za pripremu dokumenata (eng. *document preparation system*) te se najčešće koristi za tehničke i znanstvene dokumente.[25] Isto tako, osigurava i mehanizme za slaganje teksta koji onda omogućuju njegov prikaz u unaprijed definiranoj formi. LaTeX je prikladan za slaganje članaka u časopisima, tehnička izvješća, kao i za prikaz kompleksnih matematičkih formula, automatsko generiranje bibliografija, kazala itd.[26]

Primjer sintakse LaTex dokumenta:

```
\documentclass{article}
\title{O LaTeX-u}
\author{Leslie Lamport}
\date{September 1985}
\begin{document}
```

[25] Usp. Lamport , L. LaTeX: A Document Preparation System. User's Guide and Reference Manual. Addison-Wesley Pub. Co. 1994. Str. 1-26. URL: http://www.stat.pitt.edu/stoffer/freetex/latex%20basics.pdf (2013-12-01) Str.2.
[26] Usp. LaTeX: a document preparation system. URL: http://www.latex-project.org/ (2013-12-01)

```
  \maketitle
  \LaTeX{} sustav za pripremu dokumenata
  \TeX{}, koristi se za tehničke dokumente.
\end{document}
```

2.3.3. Opisni označiteljski jezici

2.3.3.1. Općenito o opsinim označiteljskim jezicima

Opisni označiteljski jezici kao učinkoviti alati za obradu teksta predstavljaju nezaobilazno rješenje u označiteljskoj praksi. Iako su opisni označiteljski jezici najveću zastupljenost doživjeli u kontekstu mrežnog okruženja – kroz uporabu HTML-a za kreiranje mrežnih stranica, svoju primjenu opisni jezici pronašli su i u okrilju Inicijative za označavanje teksta (engl. *Text Encoding Initiative – TEI*) danas konzorcija, koji stoji iza najpoznatijeg standarda za označavanje i razmjenu humanističkih tekstova u digitalnom okruženju. O TEI-u će biti više riječi u nastavku rada. Slijedi prikaz dva najvažnija predstavnika opisnih jezika, najprije SGML-a, a potom njegove mnogo poznatije inačice XML-a.

2.3.3.2. SGML

Razvoj opisnih označiteljskih jezika započeo je sredinom 1960-ih u grafičko-nakladničkoj industriji gdje su u svrhu pojednostavljenja pripreme i proizvodnje knjiga razvijena dva opisna označiteljska jezika – GML (engl. *Graphic Communication Association)* i *GenCode*. [27] Krajem 1970-ih dolazi do njihovog ujedinjenja s ciljem razvoja prvog standardiziranog označiteljskog jezika SGML-a (engl. *Standard Generalized Markup Language*). Nastali standard objavljen je 1986. kao međunarodna norma ISO 8879.[28] SGML predstavlja mehanizam za opisivanje hijerarhijskih strukturiranih dokumenata koji se sastoje od različitih objekata sadržaja.[29] On ne propisuje određeni skup objekata sadržaja koji treba biti zastupljen u dokumentu nego samo

[27] Ups. Goldfarb, C. F. The SGML history niche: the roots of SGML - a personal recollection - a personal recollection, 1996. URL: http://www.sgmlsource.com/history/roots.htm (2013-12-18)
[28] International organization for standardization: information processing – text and office systems – Standard Generalized Markup Language (SGML). URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387 (2013-12-01)
[29] Pri čemu je sintagma 'objekti sadržaja' istovjetna izrazu 'komponente teksta'.

način na koji pojedini objekti sadržaja i njihovi međusobni odnosi trebaju biti označeni.[30] Temeljni koncept na kojem se temelji SGML je 'tip dokumenta' (engl. *document type*). Tip dokumenta predstavlja klasu[31] dokumenata koji dijele određeni skup objekata sadržaja, a mogu se odnositi jedni prema drugima u nekoliko definiranih kombinacija.[32] U SGML-u ne postoji nekakav unaprijed definirani tip dokumenta koji bi trebali slijediti svi ostali tipovi dokumenta, nego se 'definicijom tipa dokumenta' (engl. *Document Type Definition* – DTD) definira opisni označiteljski jezik temeljen na SGML-u. DTD određuje sintaksu i rječnik elemenata novog označiteljskog jezika.[33]

Nerijetko se SGML shvaća kao jezik koji se sastoji od vlastitih elemenata za komponente teksta poput odlomka, poglavlja, sažetka i sl. A. H. Renear upozorava na često pogrešno razumijevanje SGML-a kao jezika za označavanje teksta te ističe kako bi se on ipak trebao tretirati kao metajezik koji služi za kreiranje opisnih označiteljskih jezika odnosno strojno razumljivih definicija elemenata pojedinih opisnih označiteljskih jezika. SGML nije u pravom smislu označiteljski jezik koji se sastoji od vlastitih elemenata, već metajezik koji služi kreiranju opisnih označiteljskih jezika.[34] Osim što je metajezik SGML je i svojevrsna metagramatika, gramatika koja definira druge gramatike tj. odnose između elemenata pojedinih opisnih označiteljskih jezika. Primarni nedostatak SGML-a bio je u njegovoj opsežnosti. Kreatori SGML-a pokušali su pokriti svaku moguću primjenu jezika te je nastali produkt bio opsežan i složen za korištenje; u konačnici, 'skup' pri upotrebi. Kao odgovor na nedostatke SGML-a nastaje *eXtensible Markup Language* (XML).

2.3.3.3. XML

XML predstavlja označiteljski jezik koji je namijenjen opisivanju i davanju značenja podacima.[35] Službenu preporuku XML-a objavljuje World Wide Web Consortium (W3C) 1998. Kao i u slučaju SGML-a i XML valja promatrati kao metajezik, on nije bio predviđen da bude

[30] Usp. Renear, A. Nav. dj.
[31] Klasa u ovom smislu predstavlja skup ili grupu članova koji imaju ista svojstva.
[32] Primjeri tipa dokumenta su roman, pjesma, esej, katalog itd.
[33] Usp. M.Sperberg-McQueen, Burnard, L. AGuidelines for Electronic Text Encoding and Interchange. (TEI P3) Gentle Introduction to SGML. URL: http://www-sul.stanford.edu/tools/tutorials/html2.0/gentle.html (2013-12-01)
[34] Usp. Renear, A. Nav. dj.
[35] XML se često uspoređuje s HTML-om. Iako su sa stajališta sintakse vrlo slični, ono što ih razlikuje je svrha: XML je namijenjen opisivanju i davanju značenja podacima, dok se HTML bavi prikazom.

označiteljski jezik nego označiteljski standard koji preko vlastitih pravila propisuje kako kreirati pojedini označiteljski jezik.

Osnovne prednosti XML-a u odnosu na druge označiteljske jezike ogledaju se u tome da XML stavlja naglasak na opisno, a rjeđe na proceduralno označavanje teksta te je neovisan o bilo kojem hardverskom ili softverskom sustavu.

a) Namjena XML-a

Namjena XML-a ogleda se u pohrani, razmjeni odnosno prijenosu i strukturiranju podataka. Pod pohranom se prvenstveno misli na čuvanje i zaštitu podataka. Mnoga programska rješenja danas posjeduju sigurnosne kopije baza podataka i u XML formatu. Namjena XML-a koja se ogleda u pohrani podataka upućuje na još jedan važan princip u razvoju označiteljskih jezika – odvajanje sadržaja od njegovog prikaza. Što se tiče razmjene, tj. prijenosa podataka XML datoteka razumljiva je svakom programskom rješenju i bazi podataka čime je olakšana migracija podataka iz jednog računalnog sustava u drugi. Posebnu namjenu XML-a predstavlja mogućnost strukturiranja dokumenta (naslov, poglavlje, potpoglavlje itd.).

Struktura XML dokumenta definira se *XML Schemom*, koja predstavlja na XML-u temeljenu alternativu DTD-u. *XML Schemom* definiraju se elementi i atributi koji se mogu pojaviti u XML dokumentu, kao i podelementi (engl. *child elements*) te njihov broj i redoslijed.[36]

Svojstva XML-a

Obilježja odnosno svojstva XML-a su sljedeća:

- XML je zbog svoje dvojake prirode „metajezika“ i „običnog“ označiteljskog jezika proširiv, ne sastoji se od definiranog skupa elemenata već se navedeni skup neprestano može mijenjati i nadopunjavati novim elementima;
- XML dokument mora biti dobro oformljen (engl. *well-formed*) prema strogim sintaktičkim pravilima koja propisuje XML specifikacija u obliku preporuke W3 konzorcija;

[36] Njome se također definiraju tipovi podataka za sadržaj elemenata i vrijednost atributa te početne i fiksne vrijednosti za sadržaje elemenata i vrijednosti atributa.

- XML dokument može biti formalno valjan ili ne valjan (engl. *valid*) u odnosu na propisanu strukturu dokumenta;
- XML je mnogo zanimljiviji sa stajališta značenja podataka a ne njihova prikaza.[37]

b) Sintaksa XML-a

Tehnički termin koji se koristi u XML-u, a pomoću kojeg se označavaju komponente teksta naziva se 'element'. XML elementi nisu unaprijed definirani te svatko može kreirati svoj vlastiti skup elemenata.[38] Svaki element sastoji se od svoje početne oznake, pripadajuće komponente teksta i završne oznake, što se može uočiti u sljedećem primjeru:

početna oznaka komponenta teksta završna oznaka

Sintaktička pravila XML-a su stroga, jednostavna i logična:

- XML elementi moraju imati početnu i završnu oznaku;
- XML elementi su osjetljivi na mala i velika slova;
- XML elementi moraju biti propisno ugniježđeni;
- XML dokument mora i može imati samo jedan korijenski element (engl. *root element*) u kojem su ugniježđeni svi drugi elementi;
- Vrijednost atributa XML elemenata mora biti navedena u navodnim znakovima.[39]

Struktura XML dokumenta zbog mogućnosti međusobnog ugnježđivanja elemenata jednih u druge naziva se i XML stablo (engl. *XML tree*). Odnosi koji vladaju između elemenata u takvoj strukturi podsjećaju na one koji vladaju u strukturi obiteljskog stabla. Za opisivanje odnosa između elemenata koriste se elementi: roditelj-element (engl. *parent element*), dijete-element

[37] Usp. A Gentle Introduction to XML. Nav.dj.
[38] Pri davanju naziva XML elementima potrebno se pridržavati sljedećih pravila: nazivi mogu sadržavati slova, brojeve i druge znakove, ali ne mogu započeti s brojem ili znakom zareza (,), ne smiju započeti sa slovima XML (ili XML, Xml i sl.) te ne smiju sadržavati prazne prostore.
[39] Usp. W3schools: XML Syntax Rules. URL: http://www.w3schools.com/xml/xml_syntax.asp (2013-12-01)

(engl. *child element*) i "sestrinski" ili „bratski" elementi (engl. *siblings elements*). Roditelj element ima djecu. Djeca na istoj razini nazivaju se potomci (braća ili sestre).

```
<root>
  <child>
      <subchild>.....</subchild>
   </child>
 </root>
```

Roditelj-element je element koji sadrži promatrani element, dijete-element je element koji je ugniježđen u promatrani element, a sestrinski ili bratski elementi su elementi sadržani u istom roditelj-elementu.[40] (U Prilogu 1 prikazan je grafički prikaz XML sintakse).

2.3.4. Prednosti opisnih označiteljskih jezika

Prednosti opisnih označiteljskih jezika u odnosu na druge označiteljske jezike brojne su i raznovrsne. Neke od njih su:

- pojednostavljeno kreiranje teksta;
- omogućeno je strukturno-orijentirano uređivanje teksta;
- olakšano je premještanje i brisanje teksta;
- moguće je više alternativnih prikaza istog teksta;
- formatiranje je na globalnoj razini;
- automatski se generiraju dodatne komponente teksta (indeksi, sadržaji i sl.)
- pojačana je podrška za vanjske uređaje (printere);
- osigurana je interoperabilnost sadržaja (prijenos iz jednog označiteljskog sustava u drugi);
- omogućeno je pretraživanje (sadržaj strukturiran prema poljima);
- podrška analitičkim metodama analize teksta (stilometrija, metoda analize sadržaja i dr.).[41]

Pojedini autori idu toliko daleko da tvrde kako opisni označiteljski jezici ne predstavljaju samo najbolji pristup u području računalne obrade teksta nego i najbolji uopće mogući pristup u tom

[40] Usp. Isto.
[41] Usp. Isto.

području.[42] Uloga opisnih označiteljskih jezika najviše se ogleda u podršci razvoju TEI-a, označiteljskog standarda baziranog na XML opisnom označiteljskom jeziku o čemu će biti riječ u idućem potpoglavlju.

2.4. Inicijativa za označavanje teksta

Inicijativa za označavanje teksta (engl. *Text Encoding Initiative - TEI*) osnovana je 1987. godine na Sveučilištu Vassar sa svrhom razvoja opisnog standarda za označavanje teksta baziranog na SGML-u. TEI predstavlja zajednički pothvat istraživačke zajednice koja se bavi označavanjem teksta, a koji za zadaću ima formulirati smjernice za označavanje i razmjenu strojno čitljivih tekstova namijenjenih za književna, lingvistička, povijesna ili druga proučavanja i istraživanja. SGML je bio traženi alat koji je označiteljska zajednica u godinama koje su prethodile željno očekivala kako bi omogućila standardizaciju označavanja teksta.

U osnivanju TEI-a sudjelovale su tri poznate organizacije: Udruženje za računala i humanistiku (engl. *Association for Computers and the Humanities – ACH*)[43], Udruženje za literarno i jezično računarstvo (engl. *The Association for Literary and Linguistic Computing – ALLC*)[44] i Udruženje za računalnu lingvistiku (engl. *Association for Computational Linguistics – ACL*)[45]. Cilj TEI-a bio je dvojak: osigurati uspješnu razmjenu humanističkih tekstova u istraživanjima te sugerirati principe označavanja teksta u istom formatu.[46] Kao glavni produkt djelovanja TEI-a nastale su TEI smjernice (engl. *TEI Guidelines*) koje su opisivale TEI aplikaciju sastavljenu od propisnog seta elemenata i atributa za označavanje teksta. Principi koji su zastupani pri izradi TEI smjernica u literaturi su poznati kao „Poughkeepsie principi“[47] i njima se eksplicitno definiraju načela kojih se treba pridržavati pri izradi smjernica. Prva inačica (P1) TEI smjernica realizirana je u lipnju 1990. Specifikacija P1 inačice bila je temeljena na SGML-u. U periodu između 1990.

[42] Usp. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Nav. dj. Str. 946.
[43] Association for computers and the humanities. URL: . http://www.ach.org/ (2013-12-01)
[44] Association for literary and linguistic computing. URL: http://www.allc.org/ (2013-12-01)
[45] Association for Computational Linguistics. URL: http://www.aclweb.org/ (2013-12-01)
[46] Usp. The Poughkeepsie Principles Closing Statement of Vassar Conference The Preparation of Text Encoding Guidelines. Poughkeepsie, New York: 1987. URL: http://www.tei-c.org/Vault/ED/edp01.htm#b2b1b3b3b3 (2013-12-01)
[47] Design Principles for Text Encoding Guidelines: The Poughkeepsie Principles. URL: http://projects.oucs.ox.ac.uk/teiweb/Vault/ED/edp01.xml?ID=body.1_div1.2 (2013-12-01)

i 1993. petnaest različitih radnih skupina radilo je na reviziji TEI smjernica. P1 inačicu standarda su 1992. na kratko zamijenili novom inačicom P2. Konačno, u svibnju 1994. objavljena je prva službena verzija TEI smjernica za označavanje teksta u P3[48] inačici, a koja je također bila temeljena na SGML-u. TEI smjernice postigle su veliki uspjeh i danas ih gotovo svaki projekt vezan uz označavanje humanističkih tekstova koristi. U lipnju 2002. godine objavljena je i P4 inačica smjernica. I dok su prve tri inačice TEI smjernica bile bazirane na sintaktičkim pravilima SGML-a, P4 inačica omogućavala je izbor između SGML-a i XML-a, da bi se P5 inačica, objavljena 2007., na kraju temeljila isključivo na XML-u i koristila njegovu sintaksu za označavanje teksta što je u skladu s mrežnim okruženjem u kojem se većina elektroničkih tekstova danas nalazi.[49]

U ovom poglavlju dan je uvid u teorijske postavke područja označavanja teksta. Obrađene su definicije osnovnih pojmova označavanja teksta, taksonomija postupka označavanja i označiteljskih jezika kao i modeli teksta. U sljedećem poglavlju izložit će se označiteljska teorija.

3. OZNAČITELJSKA TEORIJA

3.1. Općenito o označiteljskoj teoriji

U literaturi o obradi i označavanju teksta često je zastupana teza da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu (engl. *Ordered Hierarchy of Content Objects* - OHCO).[50] Tekst promatran na ovaj način u suštini se sastoji od objekata kao što su poglavlja, odjeljci, rečenice itd. ugniježđenih u određenu strukturu. OHCO teorija važnu je ulogu imala u promicanju i zastupanju opisnih označiteljskih jezika (najprije SGML-a, a zatim XML-a[51]) u postupcima označavanja tekstova. Osim toga, poslužila je i kao odgovarajući teorijski okvir za

[48] Glavni urednici P3 inačice TEI smjernica su dva ugledna stručnjaka TEI zajednice i dobra poznavatelja SGML-a i označiteljske teorije, L.Burnard i M.Sperberg-McQueen (koji će se 90-ih odazvati na poziv W3 konzorcija i T.Berners-Leea za sudjelovanje u razvoju XML-a kao novog označiteljskog standarda u mrežnom okruženju.
[49] Usp. Isto
[50] Usp. Combs, J. H., Renear A.H., DeRose, S. Nav.dj. Usp. DeRose, S.; Durand, D.; Mylonas, E. What is text, really? // Journal of Computing in Higher Education. 1, 2(1990) URL: http://www.hki.uni-koeln.de/sites/all/files/courses/3226/Renear_ea-1997.pdf (2013-12-01) Str. 6-9.
[51] OHCO koncept teksta podrazumijeva da je svaka struktura pravilno ugniježđena unutar one više razine. XML je označiteljski jezik koji zastupa strukturu stabla te je u tom smislu dobro prilagođen za reprezentaciju teksta kao hijerarhije razvrstanih objekata sadržaja. Međutim, do problema dolazi kada strukture nisu u obliku stabla. O svemu tome bit će riječ više u nastavku ovog poglavlja.

nastanak TEI-a. No, i prije nego je poslužila za nastanak označiteljske inicijative ova teza o tekstu bila je implicitno zastupljena u ranijim teoretiziranjima o razvoju računalne obrade teksta i izdavačkim softverima.[52]

Iako pomalo čudi, OHCO teorija doživjela je iznimno malo dorada, proširenja i pojašnjenja tijekom godina u kojima je služila kao teorijski okvir u istraživanjima o obradi teksta i standardizaciji označavanja. TEI zajednica prisvojila je ovaj pogled na tekst bez da su ga posebno propitali, objasnili ili obranili.[53] Tek u praktičnoj primjeni, odnosno procesu označavanja teksta, uočeni su problemi koji su OHCO teoriju doveli u pitanje i uvjetovali njenu reviziju.

Kao jedan od središnjih nedostatka teorije navodi se problem preklapanja hijerarhijskih struktura (engl. *problem of overlapping hierarchies*) kojeg prvi primjećuje i izlaže D. Barnard sa suradnicima u svom radu iz 1988.[54] Ovaj problem obično se smatra problemom tehničke naravi i njegovo rješenje može se relativno lako pronaći s obzirom na praktične ustupke različitih tehnika označavanja koje označiteljski jezici, konkretno XML, omogućuje. Autori A. H. Renear, E. Mylorans i D. Durand smatraju da zabrinutost koja vlada oko problema preklapanja ne treba biti vezana uz tehnička pitanja, već uz puno fundamentalniji problem a to je razumijevanje onoga što zapravo radimo kada neki tekst označavamo. Pritom, ukoliko se želi shvatiti što se točno zbiva pri procesu označavanja teksta, stav da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu više ne zadovoljava. Štoviše, budući da pretpostavka o hijerarhiji objekata teksta počiva prvenstveno na praktičnim prednostima za razne programe (aplikacije), a ne na načelima analize, navedena grupa autora smatra kako zapravo nema odgovarajuće terminologije za opisivanje problema koji se javlja.[55]

Tekst koji nije označen je strogo govoreći „teorijski slobodan“, ali bez označavanja nema ni strojno čitljivog testa. Trebalo bi biti uobičajeno da je strojno čitljivi tekst „subjektivan“ i „interpretativan“, a ne izričito „subjektivan“ ili „interpretativan“. OHCO teorija, u tom smislu, podržava M. Sperberg-McQueena i njegov aksiom o *markup-u* da “markup reflektira teoriju

[52] Goldfarb, C.F. Nav.dj. Str. 69.
[53] Usp.Renar, A. Mylonas, E., Durand,D. Nav. dj.
[54] Barnard, D., Hayter R., Karababa M., Logan G., and McFadden, J. (1988), 'SGML-Based Markup for Literary Texts: Two Problems and Some Solutions', Computers and the Humanities 22: 265-276 URL: http://link.springer.com/article/10.1007%2FBF00118602#page-1 (2013-12-01)
[55] Usp.Renar, A. Mylonas, E., Durand,D. Nav. dj.

teksta“ (engl. *markup reflects theory of the text)*.[56] Zapravo, OHCO teorija predstavlja prošireno razmišljanje o ovom aksiomu.

U nastavku rada izložit će se priroda postavke da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu te njezina veza s obradom i postupkom označavanja teksta. Isto tako, ukazat će se na problem preklapanja hijerarhijskih struktura teksta na koji su zajednice za označavanje teksta naišle u praktičnom radu te predstaviti dvije, progresivno slabije, inačice ove teze koje su formulirane kao odgovor na nastali problem. No prije toga, kao uvod u OHCO teoriju, predstavit će se modeli teksta koji su bili zastupljeni u programima za računalnu obradu teksta prije pojave OHCO modela na kojem je spomenuta teorija i utemeljena.

3.2. Modeli teksta

Prije pojave koncepta teksta kao hijerarhije objekata sadržaja razvrstanih prema točnom redoslijedu, prema S. DeRoseu, D. Durandu i E. Mylonasu, u programima za računalnu obradu teksta bili su zastupljeni sljedeći modeli teksta:

- tekst kao bitmapa (engl. *bitmap*) – ovakav tekst nastao je skeniranjem tiskanog tekstualnog izvornika. Bitmapa je svojevrsna slika sastavljena od pixela s kojom je gotovo nemoguće manipulirati. Naime, tekst u obliku bitmape moguće je samo gledati te je pogodan samo za najjednostavnije analize; ne može se pretraživati, formatirati, ažurirati i sl. te predstavlja najmanje prihvatljiv model teksta za obradu u računalnim programima.
- Tekst kao niz znakova (engl. *a stream of characters*) – u ovom modelu riječ je o tekstu s eksplicitno označenim pojedinačnim znakovima (npr. slova) među kojima nema hijerarhijskih odnosa. Moguće je razlučiti pojedine riječi, no nije moguće utvrditi pripadnost riječi nekom poglavlju itd. Svi problemi prethodnog modela teksta jednako se odnose i na ovaj, dakle tekst je i dalje nedostupan za neku svrsishodniju obradu.
- Tekst kao niz znakova s umetnutim instrukcijama formatiranja (engl. *formatting instructions*) – ovaj model teksta nadovezuje se na prethodni dodajući u sam tekst instrukcije za potrebe njegovog formatiranja (npr., boja teksta, veličina slova).

[56] Usp. Sperberg-McQueen, M. Text in the electronic age: textual study and text encoding, with examples from medieval texts. // Literary and Linguistic Computing. 6, 1(1991).

- Tekst kao izgled stranice (engl. *page layout*) – predstavlja model teksta koji je najrasprostranjeniji među programima za računalnu obradu teksta, prvenstveno iz razloga što podržava izgled tiskanog dokumenta. Ovaj model omogućuje utvrđivanje hijerarhijskih odnosa unutar stranice te omogućuje određivanje mjesta svakog pojedinog znaka, što je pogodno za pripremu dokumenta za tiskanje, no s druge strane ne razlikuje naslov od običnog teksta i neki od osnovnih objekata sadržaja, poput retka teksta, podložni su promjeni ali ne u ovisnosti o sadržaju već o formatiranju.
- tekst kao niz objekata sadržaja (bez hijerarhijskih veza) (engl. *a stream (not hierarchy) of content objects*) – ovaj model teksta uključuje primjenu stilova uporabom kojih je riješen nedostatak *page layout* modela te je moguće razlikovati naslov od ostatka teksta. Ipak, stilovi sami za sebe ne podržavaju hijerarhijske odnose te je u dokumentu nemoguće razlikovati pripadnost određenog teksta poglavlju itd.[57]

3.3. OHCO model teksta i OHCO teorija

3.3.1. Općenito o OHCO teoriji

Nezadovoljni prethodno spomenutim modelima teksta, etiketirajući ih kao manje sofisticirane, S. J. DeRose, D. Durand, E. Mylonas i A. H. Renear 1990-ih predlažu novi model teksta, model koji je jednostavniji i koji predstavlja funkcionalniji način za kreiranje, mijenjanje i oblikovanje teksta, a koji podržava pretraživanje, pregledavanja, analiziranje i druge vrste posebne obrade teksta te omogućuje jednostavniju razmjenu tekstova među različitim softverskim i računalnim sustavima. U literaturi je ovaj model teksta poznat pod akronimom OHCO (engl. *Ordered Hierarchy of Content Objects)*, odnosno OHCO teorija, a prema njemu tekst je definiran kao hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu.[58] Razlog za tretiranjem teksta kao hijerarhije objekata sadržaja razvrstanih po točnom redoslijedu inicijalno je bio vezan uz, prethodno navedene, praktične prednosti takvog načina promatranja teksta.

Osnovna teza u OHCO teoriji je da svaka knjiga, dokument ili tekst predstavlja *hijerarhiju objekata sadržaja razvrstanih prema točnom redoslijedu*. Konkretno, svaki tekst sastoji se od

[57] Usp. DeRose, S.; Durand, D.; Mylonas, E. Nav.dj. Str. 6-9.
[58] Usp. Isto. Str. 3.

temeljnih dijelova koji se nazivaju *objekti sadržaja* (engl. *content objects).* Tekst se tako sastoji od slijeda poglavlja; poglavlja se pak sastoje od potpoglavlja koje pak čine odlomci teksta: odlomci se sastoje od rečenica, citata, jednadžbi itd. Ovakva struktura je *hijerarhijska* (engl. *hierarchy*), objekti sadržaja raščlanjuju se i ugnježđuju jedan u drugi po određenom redoslijedu. I upravo točno određen redoslijed objekata sadržaja u strukturi teksta, odnosno njihova *razvrstanost po točno određenom redoslijedu* (engl. *ordered*), čini definiciju teksta s označiteljskog stajališta potpunom. Koliko je sam redoslijed objekata sadržaja u tekstu važan uočava se u tvrdnji da su „dva teksta jedan te isti tekst samo ako posjeduju potpuno isti redoslijed objekata sadržaja".[59]

A. Renear, E. Mylonas. i D. Durand[60] hipotezu da je tekst hijerarhija razvrstanih objekata sadržaja detaljnije propituju i pojašnjavaju kroz tri vrste argumenta: *pragmatičkim argumentima* – koji pojašnjavaju pojam hijerarhije te navode praktične prednosti promatranja teksta kao hijerarhije razvrstanih objekata sadržaja; *empirijskim argumentima* – koji definiraju objekte sadržaja i njihovu ulogu u teoriji, hipotezama, objašnjenima i opisima vezanim uz tekst; *teorijskim argumentima* – koji ukazuju na važnost redoslijeda odnosno razvrstanosti objekata sadržaja te se utvrđuje kako su *x* i *y* isti tekst ako - i jedino ako - posjeduju potpuno isti redoslijed objekata sadržaja.[61]

3.3.2. Problem preklapanja hijerarhijskih struktura teksta

Kada su enkoderi započeli s praktičnom primjenom OHCO koncepta teksta gdje su trebali svaki dokument prikazati kao jednu logičnu hijerarhijsku strukturu, naišli su na probleme. Ubrzo je uočen glavni nedostatak vezan uz OHCO teoriju teksta odnosno postojanje višestrukih logičkih i ne samo logičkih hijerarhija, a što tvori problem preklapanja hijerarhijskih struktura s kojim se opisni označiteljski jezici suočavaju u postupku označavanja teksta. Pod pojmom preklapanja (engl. *overlap*), u kontekstu OHCO teorije, misli se na preklapanje komponenti teksta koje

[59]Renear, A.; Mylonas, E.; Durand, D. Refining our notion of what text really is: the problem of overlapping hierarchies. // Research in Humanities Computing. Oxford University Press, 1996. URL: http://www.stg.brown.edu/resources/stg/monographs/ohco.html (2013-12-01)

[60] Usp. Renear, A.; Mylonas, E.; Durand, D. Nav.dj.

[61] Ako se mijenja izgled nekog teksta (prored, font), može se tvrditi kako tekst ostaje isti, no ukoliko je broj ili struktura objekata sadržaja teksta izmijenjena (broj poglavlja se poveća/smanji, redoslijed stavaka se zamijeni) onda više ne možemo govoriti o istom tekstu.

pripadaju jednoj hijerarhiji teksta s komponentama teksta druge hijerarhije.[62] Kada dođe do preklapanja objekti ne mogu biti dobro oformljeni (engl. *well-formed*) u XML smislu, odnosno oznake i sadržaj koji uključuju nije moguće zastupati u strukturi stabla kako to propisuje XML specifikacija.[63] Primjerice, poglavlje neke knjige može započeti na jednoj stranici a završiti na drugoj, rečenica može započeti u jednom redu a završiti u drugom itd. U obje situacije dolazi do preklapanja logičke i fizičke strukture teksta u kojima su kao elementi logičke strukture teksta uzeti poglavlja i rečenice, a fizičke broj stranice i broj reda.

Problem preklapanja hijerarhija u strukturiranom tekstu bio je toliko očit da su i sami autori koji su osmislili OHCO teoriju ostali iznenađeni previđanjem istog pri prvotnom formuliranju teorije. Smatra se da je uzrok previđanja usko vezan uz činjenicu da se različito gledalo na koncept teksta, odnosno da su SGML zajednica u prvoj polovini 1980-ih i TEI zajednica u drugoj polovini 1980-ih različito promatrali tekst.[64]

Tijekom početnog razvoja opisnih označiteljskih sustava i pristupa tekstu kroz objekte sadržaja, svaki dokument se promatrao kao da ima **jednu** prirodno zastupljenu logičku strukturu, a ona je bila određena prema tipu dokumenta. Smatralo se kako objekti sadržaja zastupaju tip ili kategoriju teksta kojoj dokument pripada: pravni ugovori tako imaju jednu skupinu objekata, znanstvene monografije drugu – pjesme, romani, drame, pisma, peticije, potvrde itd., imaju vlastite skupine objekata i gramatike koji definiraju sintaktičke odnose koji vladaju među njima. U svakoj reprezentaciji objekti su formirali strogu hijerarhijsku strukturu, tj. uvijek su bili 'ugniježđeni' (engl. *nested*) i nikad se nisu 'preklapali' (engl. *overlapped*). Konkretno, svaki tekst posjeduje logičku i fizičku strukturu. Objekti u logičkoj strukturi nikad se ne preklapaju jedan s drugim, a isto vrijedi i za objekte unutar fizičke strukture koji se također ne mogu međusobno preklapati. Logički objekti poput rečenice, odlomka i poglavlja ne preklapaju se jedan s drugim kao što se niti fizički objekti poput redova, stupaca i stranica međusobno ne preklapaju. Ipak, do preklapanja dolazi kada se susretnu objekti logičke i fizičke strukture; primjerice, rečenica započinje u jednom redu a završava u drugom i sl. U tablici 1 prikazani su mogući odnosi između struktura i situacije u kojima dolazi do preklapanja.

[62] Usp. Renear, A.; Mylonas, E.; Durand, D. Nav.dj.
[63] Usp. Schmidt, D., Colomb, R. A Data Structure for Representing Multi-version Texts Online. Str. 1-20. URL: http://itee.uq.edu.au/~schmidt/_articles/elsevier.pdf (2013-12-01) Str.2.
[64] Usp. Renear, A.; Mylonas, E.; Durand, D. Nav.dj.

Tablica 1. Mogući odnosi između različitih struktura:[65]

Nema preklapanja **<a>Povijestoznačiteljsketeorije** **Povijestoznačiteljske<a>teorije**
Elementi dijele jednu završnu/početnu točku **<a>Povijest označiteljsketeorije** **Povijest označiteljske<a>teorije**
'Klasično' preklapanje **<a>Povijestoznačiteljsketeorije** **Povijest<a>označiteljsketeorije**
Elementi dijele završnu točku **<a>Povijestoznačiteljske teorije** **Povijest<a>označiteljske teorije**
Jedan element sadržan je u drugom element **<a>Povijestoznačiteljsketeorije** **Povijest<a>označiteljsketeorije**
Elementi dijele početnu točku **<a>Povijest označiteljsketeorije** **<a>Povijest označiteljsketeorije**
Elementi dijele početnu i završnu točku **Povijest<a>označiteljsketeorije** **Povijest<a>označiteljsketeorije**

[65] Izrađeno na temelju primjera iz rada Durusau, P., O'Donnell. Cuncurrent Markup for XML Documents. URL: http://www.durusau.net/publications/Concurrent_markup.pdf (2013-12-01)

U SGML zajednici smatralo se da tekst posjeduje samo jednu strukturu, a ona je određena prema tipu dokumenta (u Tablici 2 nalaze se primjeri tipova dokumenta s odgovarajućim objektima sadržaja).

Tablica 2. Primjeri **tipova dokumenta** i objekata sadržaja:

Knjiga	Tijelo, poglavlje, odjeljak, paragraf, izvadak, fusnota ...
Članak	Naslov, autor, afilijacija, sažetak, poglavlje, potpoglavlje, odlomak, izvadak ...
Pismo	Adresa pošiljatelja, adresa primatelja, pozdrav, tijelo, pošiljateljevi inicijali ...
Pjesma	Naslov, strofa, linija ...

U prvotnoj OHCO teoriji polazi se od pogleda da tip dokumenta indicira tip strukture koja je zastupljena. To je stav, odnosno pretpostavka, na osnovu koje je definirana prvotna verzija OHCO teorije. Tek će kasnije TEI istraživači i praktičari utvrditi da je struktura teksta zapravo višeslojna odnosno da se tekst sastoji od više hijerarhija koje se najčešće preklapaju.[66] Iz razloga što u najčešćem broju slučajeva ne postoji jedna, jedinstvena hijerarhija objekata sadržaja, više se ne može tvrditi da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu. Nakon što se klasa logičkih elemenata u određenom tekstu proširila na različite perspektive preklapanje je postalo neizbježno te se stoga više ne može govoriti o tekstu kao jedinstvenoj hijerarhiji objekata sadržaja razvrstanih prema točno određenom redoslijedu. U tom smislu, prvotna OHCO teorija je opovrgnuta, a kao pokušaj odgovora na nastali problem nastaje revidirana inačica OHCO teorije – OHCO-2.

3.4. Modifikacija OHCO teorije – OHCO-2

3.4.1. Općenito o OHCO-2

Kada su istraživači književne i lingvističke zajednice započeli s praktičnom primjenom SGML-a, ubrzo su naišli na probleme. Iako se prvotno pretpostavljalo kako je tekst moguće zastupati kroz jednu hijerarhijsku strukturu, naknadno se uspostavilo da postoje brojne hijerarhijske strukture i

[66] Usp. Renear, A.; Mylonas, E.; Durand, D. Nav.dj.

da se one najčešće preklapaju.[67] U starom pogledu na tekst objekti sadržaja grupirani su u skupine odnosno obitelji (engl. *families*) koje su određene po tipu ili kategoriji elemenata teksta. U novom pogledu obitelji su određene analitičkim i metodološkim perspektivama teksta. Koncept „analitičkih perspektiva" (engl. *analytical perspective*) definiran je kao „prirodna obitelj metodologije, teorije i analitičke prakse".[68] (U Tablici 3 nalaze se primjeri perspektiva i elemenata od kojih se sastoje).

Tablica 3. Primjeri **perspektiva** i elementi od kojih se sastoje:

Dramska:	čin, prizor, didaskalije, govor …
Prozodijska:	pjesma, stih, strofa, katren, dvostih, linija, pola linije, stopa …
Retorička:	uvod, naracija, argument, zaključak …
Diskurzivna:	otvaranje, provjera, izmjena tema, završetak …
Aksiomatička:	osnovno, aksiomi, definicije, teoremi, dokazi, protuprimjeri, klauzule ...
Sintaktička:	rečenice, imenski izrazi, glagolske fraze, determinatori, pridjevi, imenice, glagoli ...

Istraživači iz područja označavanja teksta ustanovili su da iako se objekti iz različitih analitičkih perspektiva međusobno preklapaju, parovi objekata iz jedne analitičke perspektive nikad se ne preklapaju međusobno. Primjerice, prozodijski objekti (strofa, distih, stih, pola stiha, itd.) ne preklapaju se međusobno, kao niti lingvistički objekti (rečenice, fraze, riječi); i svaka perspektiva – prozodijska, lingvistička itd – ima egzaktnu hijerarhiju te se može tvrditi da su objekti koji su određeni analitičkim perspektivama, bez iznimke, organizirani u hijerarhije. Sukladno tome, analitičke perspektive odražavaju višestruke hijerarhije u samom tekstu te OHCO teorija u svojoj drugoj inačici OHCO-2 glasi: „analitičke perspektive su ono što određuje hijerarhije objekata sadržaja u tekstu". OHCO-2 zastupa stav kako nema jednoznačne definicije teksta te da je on suviše kompliciran i da posjeduje mnogo različitih aspekata. Svojevrsna nadopuna ove teze u inačici OHCO-2.1 glasi „ukoliko se dva označena objekta sadržaja (*x* i *y*) preklapaju to znači da ne pripadaju istoj analitičkoj perspektivi".[69]

[67] Renear, A.; Mylonas, E.; Durand, D. Nav.dj.
[68] Isto.
[69] Isto. Spomenuti autori u jednom djelu izlaganja OHCO-2 teorije postavljaju zanimljivo filozofsko pitanje: je li tekst po svojoj prirodi doista hijerarhijski nastrojen ili je analitička perspektiva ta koja nas tjera da svijet gledamo kroz hijerarhijske strukture. U nastavku promišljaju je li možda hijerarhijska podjela samo češće korištena a time i

3.4.2. Problem analitičkih perspektiva

Međutim, u daljnjoj analizi označiteljske teorije došlo se do uvida da postoje i takve analitičke perspektive koje se sastoje od objekata sadržaja koji se međusobno mogu preklapati, što narušava OHCO-2 teoriju. Sam pojam perspektive širi je od pojma hijerarhije tako da je moguće postojanje perspektiva poput, primjerice, književnih studija koje analiziraju tekst književnog djela preko njegovih različitih komponenti i svojstava poput duljine rečenica, teme djela, broja stranica, metričkih linija teksta itd., a koji se međusobno mogu preklapati. Spomenuta analiziranja navela su A. H. Renear, E. Mylonass i D. Durands da započnu s promišljanjem nove inačice teorije.

3.5. Modifikacija OHCO-2 teorije – OHCO-3

3.5.1. Općenito o OHCO-3

U svojoj trećoj inačici (OHCO-3) uvodi se koncept „pod-perspektiva“ (engl. *sub-perspective*) te se OHCO teorija u sklopu nove inačice definira na sljedeći način: „*x* je pod-perspektiva od *y* onda i samo onda ako je *x* perspektiva i *y* perspektiva a teorija i praksa perspektive *x* uključena je u teoriju i praksu perspektive *y*, ali ne i obrnuto“.[70] Pod-perspektiva predstavlja jedinstven i koherentan dio analitičke perspektive. Na primjer, povijest književnosti, kritika književnosti ili kritika teksta mogu se smatrati „dijelovima od“, „područjima od“ ili „pod-područjima“ književnih studija. Svaki od njih opet ima svoja pod-područja. Tako na primjer u područje književne kritike spada recenzija. Taj pojam dijela ili potpodručja ili discipline je ono što se u kontekstu OHCO-3 teze podrazumijeva pod pod-perspektivom.

Treća verzija OHCO teze koja dopušta perspektive u kojima se objekti preklapaju, ali i dalje zagovara i naglašava značajnu ulogu hijerarhije u našem razumijevanju onoga što je tekst, glasi ovako: „za svaki različiti par objekata *x* i *y* koji se preklapaju sa strukturom neke perspektive P(1) postoje različite perspektive P(2) i P(3), tako da su P(2) i P(3) pod-perspektive P(1), a *x* je objekt u P(2) ali ne i u P(3), dok je *y* objekt u P(3), ali ne i u P(2).“[71] Sažeto: objekti sadržaja

uobičajenija ali da tekst zapravo ne posjeduje nikakvu strukturu već je ona posljedica ljudskog rasuđivanja. Usp. Renear, A., Mylonas, E., Durand, D. Nav.dj.

[70] Usp. Renear, A.; Mylonas, E.; Durand, D. Nav.dj.

[71] Isto.

mogu se preklapati u okviru jedne perspektive, ali ako to čine onda su pripadnici različitih pod-perspektiva promatrane perspektive.[72]

3.5.2. Problem pod-perspektiva

Ipak, ubrzo je uočeno da postoje objekti sadržaja koji se međusobno preklapaju ali se ne mogu svrstati čak ni u različite pod-perspektive, kako predviđa OHCO-3 teorija. Primjer za takav slučaj može se pronaći u narativnim objektima koji sadrže „priče u priči“ (tzv. Šeherezada problem). „Priča u priči“ nikako se ne može svesti na pod-perspektivu bilo koje perspektive koja promatra tekst jer predstavlja cjelinu za sebe.

Još neke od situacija u kojima dolazi do preklapanja na ovoj razini:

- referentne strukture poput hipertekstualnih linkova;
- poetski objekti poput metafora i aluzija;
- diskurzivni objekti, kao što je recimo tema;
- konkretno napisani objekti poput pjesme u akrostihu;
- lingvistički objekti poput proizvoljnih kolokacija.[73]

Kao opći zaključak koji nude autori rada može se rezimirati sljedeće:

- Perspektive – analitičke, teorijske i metodološke prakse – jednako su važne kao i tip u identifikaciji objekata sadržaja;
- Perspektive često određuju hijerarhije objekata;
- Nehijerarhijske perspektive mogu često biti raščlanjene u hijerarhijske pod-perspektive.

Ali isto tako ističu da:

- perspektive ne određuju baš uvijek hijerarhijske strukture u tekstu;
- nehijerarhijske perspektive se ne mogu uvijek razlomiti u hijerarhijske pod-perspektive.[74]

Tekst predstavlja svojevrstan sustav perspektiva koje se raščlanjuju na pod-perspektive, koje se pak dalje mogu raščlanjivati na pod-pod-perspektive i taj se proces razgrađivanja nastavlja u nedogled.

[72] Isto.
[73] Isto.
[74] Isto.

3.6. Povijesne faze razvoja označiteljske teorije

Na kraju izlaganja OHCO teorije može se rezimirati da je neku sveobuhvatnu teoriju teksta na ovoj razini razvoja postupka označavanja teksta i označiteljskog jezika ipak teško postići, usprkos najavama i očekivanjima članova zajednice digitalne humanistike. Problem preklapanja hijerarhijskih struktura u tekstu predstavljao je najveći praktički problem s kojim su se teoretičari označiteljske teorije susreli. Iako se modifikacijom OHCO teorije pokušao riješiti ovaj problem ta ideja nikad u potpunosti nije realizirana budući da bi nakon svake revizije isplivali novi nedostaci teorije. Početna verzija OHCO teorije ubrzo se suočila s problemom preklapanja logičkih i fizičkih elemenata u tekstu što je u suprotnosti s idejom o tekstu koji posjeduje jednu hijerarhijsku strukturu kako se početnom verzijom teorije tvrdilo. U tom smislu prvotna OHCO teorija je opovrgnuta. Kao odgovor nastaje OHCO-2 teorija kojom se uvodi koncept analitičkih perspektiva koje odražavaju višestruke hijerarhije u samom tekstu. Ipak, u daljnjoj analizi označiteljske teorije došlo se do uvida da postoje i takve analitičke perspektive koje se sastoje od objekata sadržaja koji se međusobno mogu preklapati, što narušava OHCO-2 teoriju. U novoj, OHCO-3 inačici predlaže se razgrađivanje perspektiva na pod-perspektive. Međutim, ubrzo je uočeno da postoje takvi objekti sadržaja koji se međusobno preklapaju, ali se ne mogu svrstati čak ni u različite pod-perspektive, kako predviđa OHCO-3 teorija. Pokazalo se da tekst predstavlja svojevrstan sustav perspektiva koje se raščlanjuju u pod-perspektive, a koje se pak dalje mogu raščlanjivati na pod-pod-perspektive i taj se proces razgrađivanja nastavlja dok se ne stigne do tzv. 'atomskih perspektiva' (engl. *atomic perspectives*).[75] Stoga, sve dokle god perspektive sadrže preklapajuće objekte to se može tumačiti kao indikator da nisu svedene na razinu atoma i da se mogu dalje razgrađivati. Ovim postaje očito kako problem preklapanja odlazi u beskonačnost – on je nerješiv. Tekst je prekompliciran, posjeduje mnoge aspekte i ne može ga se ustaviti. U tom smislu nemoguće je govoriti o definiciji teksta koja će biti apsolutna. I konačno, tekst se pokazao kao suviše heterogena cjelina da bi u cijelosti bio obuhvaćen bilo kojom teorijom. Problem preklapanja hijerarhijskih struktura u tekstu u teorijskom smislu predstavlja nerješivu prepreku, prepreku zbog koje je OHCO označiteljska teorija na kraju i opovrgnuta.[76] Ipak, u tehničkom smislu ovaj problem nije nerješiv s obzirom na praktične postupke različitih tehnika označavanja koje označiteljski jezici, a poglavito XML,

[75] Usp. Renear, A.; Mylonas, E.; Durand, D.Nav. dj.
[76] U ovom smislu opovrgnutost označiteljske teorije odnosi se na njezinu nepotpunost.

omogućuju. U idućem poglavlju stoga će biti izloženi modeli razvijeni u okviru TEI zajednice sa svrhom rješavanja problema preklapanja hijerarhijskih struktura u tekstu.

Kako u okviru OHCO teorije nije formulirana prihvatljiva definicija teksta, istraživači iz područja digitalne humanistike i označavanja teksta nastavili su svoje napore u promišljanju prirode i svojstava teksta razvijajuću nove teorije – teorije s drugačijom perspektivom viđenja problema, a koje su korisne ne samo za praktičare koji kreiraju, upravljaju ili koriste tekst nego i za sve one koji žele razumjeti tekst/tekstualnost kroz teorijsku perspektivu. A. H. Renear kategorizira ih u tri povijesne faze: *platonizam, pluralizam* i *antirealizam*.[77]

Platonizam predstavlja prvu povijesnu fazu razvoja označiteljske teorije u kojoj je razvijena i OHCO teorija. Glavna pretpostavka platonizma tekst promatra kao hijerarhijsku strukturu objekata sadržaja razvrstanih prema točnom redoslijedu, te je iscrpno izložena u trećem poglavlju.

Iduća povijesna faza, *pluralizam,* razvija se 1990-ih. U ovoj povijesnoj fazi prepoznata je kritička uloga koju metodologija, teorija i analiza imaju u kontekstu označavanja teksta. Pluralizam smatra da je struktura koja je odabrana i korištena pri označavanju teksta rezultat ljudskog interesa, trenutne istraživačke prakse i specifične prosudbe enkodera. U pluralizmu dolazi do disperzije mišljenja, ali ipak kakvog-takvog stava o prirodi teksta: smatra se da je tekst kompliciran i posjeduje mnogo različitih aspekta, da je označavanje povijesno uvjetovana ljudska aktivnost i da nema razloga za tvrdnju kako tekst posjeduje objektivnu strukturu neovisnu o našim teorijama o njoj.[78] Ono što pronađemo o tekstu u dijelovima ovisi o načinu na koji smo to istraživali, a odgovor na pitanje što je tekst ovisi o „kontekstu, metodama i namjeri istraživanja“.[79] Tekst nije objektivno postojeći entitet koji samo treba biti definiran i prezentiran, već je to entitet koji treba biti konstruiran.[80] *Pluralistički realizam,* kao 'podfaza' pluralizma,

[77] Usp. Renear, A.Nav.dj. Str. 117.
[78] Usp. Isto. Str. 122.
[79] Huitfeldt, C. Multi-Dimensional Texts in a One-Dimensional Medium. Str.142-161 URL: http://wab.uib.no/ojs/agora-wab/article/view/1860 (2013-12-01) Str. 143.
[80] Usp. Pichler, A. Advantages of a Machine-Readable Version of. Wittgenstein's Nachla B. Str. 770-776 URL: http://wab.uib.no/alois/pichler-kirchb95b.pdf (2013-12-01) Str. 774.

dozvoljava pak mnoštvo različitih perspektiva o tekstu, ali pretpostavlja da tekst ima strukturu neovisnu o interesima, kontekstu, teorijama i praksi.[81]

U trećoj fazi teoretiziranja o tekstu – *antirealizmu* – odbacuju se prethodni pogledi i tekst se promatra kao „produkt naših teorija i analitičkih alata koje koristimo kada transkribiramo, uređujemo, analiziramo ili označavamo tekst".[82] Dvije su osnovne tvrdnje o tekstu promatranom kroz prizmu antirealizma:

a) naše razumijevanje teksta (prikaz, označavanje, analiza, transkripcija i sl.) je interpretacijsko: „nema činjenice o tekstu koja je objektivna u smislu da ne bude interpretacijska."[83] S tim da to ne znači da su sve činjenice o tekstovima posve subjektivne na način da „postoje neke stvari s kojima se svi čitatelji slažu".[84]
b) postoje mnoge različite metodološke perspektive o tekstu. „Tekst može imati mnogo različitih vrsta struktura (fizičku, logičku, narativnu, gramatičku itd.)". [85]

Pesimistički stav u pogledu pokušaja definicije teksta, koji karakterizira ovu povijesnu fazu, u skladu je sa suvremenim raspoloženjem poststrukturalističke i postmodernističke teorije koje su na ovaj način iskazale svoj utjecaj i u okviru razmišljanja o samom tekstu.

4. PRAKTIČNA RJEŠENJA PROBLEMA PREKLAPANJA HIJERARHIJSKIH STRUKTURA U TEKSTU

4.1. Problematika preklapanja hijerarhijskih struktura

Problematika postojanja višestrukih hijerarhijskih struktura uočena je na samom početku razvoja OHCO teorije i upravo je ona u daljnjem razvoju istraživanja dovela do otkrivanja nepotpunosti i manjkavosti same teorije te opovrgavanja prve OHCO teze u kojoj se tekst promatra kao hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu. Preklapanja hijerarhijskih struktura u tekstu odnosi se na to da elementi jedne, postupkom označavanja teksta, označene

[81] Usp. Renear, A. Nav.dj. Str. 122.
[82] Usp. Isto. Str 122.
[83] Huitfeldt, C. Nav.dj. Str. 149.
[84] Isto.
[85] Isto. Prethodno predstavljen antirealizam je *ontologijski* antirealizam i on promatra samu prirodu teksta, općenitiprikaz onog što tekst jeste. Uz njega postoji i *semantičk*i antirealizam koji promatra prirodu našeg znanja o tekstu i našu reprezentaciju teksta.

perspektive nisu 'dobro oformljeni' s obzirom na elemente druge označene perspektive.[86] Pod 'dobro oformljenim' misli se da su elementi pravilno ugniježđeni po XML sintaktičkim pravilima, koja zahtijevaju da se zastupa struktura teksta u obliku stabla. Posebna pažnja ovom problemu posvećuje se u okviru TEI zajednice gdje je razvijeno nekoliko mehanizama koji se nastoje uspješno nositi s njim, a koji su temeljeni na XML-u. Isto tako, u okviru TEI zajednice osnovana je i *Overlapping Markup*[87] interesna skupina koja preko *mailing* liste potiče komunikaciju među TEI stručnjacima o daljnjoj diskusiji o ovom problemu i njegovom mogućem rješenju u postupku označavanja teksta.

U nastavku će se najprije definirati situacije u kojima dolazi do preklapanja hijerarhijskih struktura u tekstu, zatim će se predstaviti modeli i praktični primjeri rješenja ovog problema temeljeni na XML-u, a koje preporučuje TEI zajednica. Na koncu, navest će se i neki od ne-XML pristupa koji su razvijeni u svrhu rješavanja istog problema.

4.2. Definiranje konfliktnih situacija

U 20. poglavlju TEI smjernica posvećenom nehijerarhijskim strukturama[88], raspravlja se o problemima koji se javljaju prilikom korištenja XML-a za kodiranje neugniježđenih komponenti i svojstava odnosno značajki teksta, a koji ne dopuštaju da ih se prikazuje u strogo hijerarhijskom obliku. U smjernicama su navedeni neki od najčešćih konflikata, a to su:

- konflikt između fizičke strukture dokumenta (svezak, stranice, stupac, red) i njegove retoričke ili lingvističke strukture (poglavlja, odlomci, rečenice);
- konflikt između metričke strukture stiha (npr. grupiranje stihova u strofe) i njegove retoričke ili lingvističke strukture (fraze, rečenice, a za dramu činovi, scene, govori);
- konflikt između metričke, retoričke ili jezične strukture i reprezentacije izravnog govora, posebno ako je citirani govor prekinut drugim elementima (npr. „Što“ upitala je „treba učiniti?“) ili prelazi metričke, retoričke ili jezične granice;

[86] Na početku ovog rada, u izlaganju o XML-u kao jedno od njegovih temeljnih svojstava navedeno je kako XML dokument mora biti dobro oformljen (engl. *well-formed*) prema strogim sintaktičkim pravilima koja propisuje XML specifikacija. Pogledati poglavlje 2.3.2.2.XML.
[87] Usp.Overlapping Markup SIG. Text Encodin Initiative. URL: http://www.tei-c.org/Activities/SIG/Overlap/ (2013-12-01)
[88] Usp. A Gentle Introduction to XML. Nav.dj.

- konflikt između različitih analitičkih pregleda ili opisa dokumenta, primjerice između oznake namijenjene za označavanje informacija o izgledu neke riječi u rukopisu i oznake namijenjene za označavanje morfologije ili izgovora te riječi.[89]

4.3. Rješenja problema preklapanja hijerarhijskih struktura u tekstu – XML pristupi

U TEI smjernicama predloženo je nekoliko metoda, temeljenih na XML-u, za rješenje problema preklapanja hijerarhijskih struktura u tekstu:

- metoda redundantnog (višestrukog) označavanja istog teksta;
- korištenje praznih elementa za obilježavanje granica neugniježđene strukture teksta;
- podjele jednog logičkog neugniježđenog elementa u segmente koji se potom 'gnijezde' ispravno u hijerarhijsku strukturu;
- *stand-off* označavanje odnosno obilježavanje teksta upućivanjem na njega, a ne umetanjem XML oznaka.[90]

U nastavku slijedi prikaz spomenutih metoda, koje će biti objašnjene na praktičnim primjerima te će se iznijeti njihove prednosti i nedostaci. Rješenja problema bit će ilustrirana pomoću izvatka iz pjesme Tina Ujevića „Čin sputanih ruku“:

> U ovom glasu zavija, kao da orguljaju grobovi
> i ječi nešto sumorno, strahovito i muklo.
> U ovom glasu kao da preklinju robovi:
> "O rasti, srce svijeta, crveno, samo da ne bi puklo!"
>
> U labirintu panike, duž acetilena,
> nešto svirepo jednoliko, s provalom nadahnuća,
> pretače se u gvožđe; paralelno, desna je uposlena,
> mozak brojeve misli, a srce kovanja vruća.[91]

[89] Usp. Isto.
[90] Usp. Isto.
[91] Ujević, T. Izabrane pjesme. Čin sputanih ruku. Str. 38. URL: http://ponude.biz/knjige/t/Tin%20Ujevic%20-%20Izabrane%20pjesme.pdf (2013-12-01)

Tekst je moguće analizirati kroz različite poglede odnosno perspektive. Prema tzv. *metričkoj perspektivi* tekst se označava prema svojim metričkim značajkama poput linija, strofa ili pjevanja, ali i prozodijskim značajkama kao što su naglasci, struktura sloga, aliteracija i rima. Kroz *gramatičku perspektivu* opisuju se lingvističke i retoričke značajke teksta poput fonema, morfema, riječi, fraza, klauzula[92], rečenica i sl. *Dijaloška perspektiva* usredotočena je na pripovijedanje odnosno razlikovanje pripovjedača i njegovog sugovornika te identificiranje pojedinih segmenata kao izravnih citata.
U primjerima će se prikazati relativno jednostavni konflikti; kao primjer *metričke perspektive* označit će se samo grupa redaka i redak, za zastupanje *gramatičke perspektive* teksta označit će se rečenice, a za zastupanje *dijaloške perspektive* samo će se razlikovati i naznačiti direktan citat drugog naratora (sugovornika).

4.3.1. Višestruko označavanje istog teksta

Koncepcijski, najjednostavniji način odvajanja dvaju (ili više) hijerarhijskih odnosa koji su u proturječju postiže se na način da se označe dva (ili više) puta s tim da se svaki put zauzme jedna određena perspektiva.

Tako, primjerice, *metrička perspektiva* na tekst može biti zastupljena korištenjem `<l>` elementa za označavanje svake pojedine metričke linije pjesme:

```
<lg>
<l>U ovom glasu zavija, kao da orguljaju grobovi </l>
<l>i ječi nešto sumorno, strahovito i muklo. </l>
<l>U ovom glasu kao da preklinju robovi: </l>
<l>"O rasti, srce svijeta, crveno, samo da ne bi puklo!" </l>
<lg>
```

Gramatički perspektiva pjesme bit će omogućena oznakama koje će označiti strukturu rečenica, poput elementa `<seg>`:

```
<p>
<seg>U ovom glasu zavija, kao da orguljaju grobovi i ječi nešto sumorno,
strahovito i muklo. </seg>
<seg>U ovom glasu kao da preklinju robovi: </seg>
<seg>"O rasti, srce svijeta, crveno, samo da ne bi puklo!" </seg>
</p>
```

[92] Klauzula - u antičkoj retorici i poetici, dočetak, završetak (stiha, kolona itd.).

Dijaloška perspektiva zastupljena je označavanjem direktnog citata drugog naratora, uporabom elementa <said>:

```
<ab>U ovom glasu zavija, kao da orguljaju grobovi i ječi nešto sumorno,
strahovito i muklo. U ovom glasu kao da preklinju robovi:
<said>"O rasti, srce svijeta, crveno, samo da ne bi puklo!" </said>
</ab>
```

Prednost metode višestrukog označavanja istog teksta je što je svaki način gledanja na informaciju eksplicitno zastupljen u podacima te su pojedinačni prikazi vrlo jednostavni za obradu. Nedostatak se odnosi na potrebu za održavanjem više kopija identičnog tekstualnog sadržaja kao i činjenica da ne postoji jasna naznaka da su različite perspektive pogleda na tekst, koje se nalaze u zasebnim datotekama, međusobno povezane.[93]

4.3.2. Obilježavanje granica praznim elementima

Druga metoda za prilagodbu nehijerarhijskih objekata u XML dokumentu uključuje označavanje početka i kraja neugniježđenog teksta. U ovom dijelu potrebno je istaknuti kako se u TEI smjernicama razlikuje privilegirana hijerarhijska struktura i alternativna hijerarhijska struktura. Privilegirana hijerarhija normalno se označava, no da bi se izbjeglo njeno preklapanje s drugim, alternativnim hijerarhijama nužno je da te alternativne hijerarhije koriste prazne elemente – i to isključivo kako bi pomoću njih ukazale na granice vlastitih komponenti sadržaja, ne i da bi ih u sebe ugnijezdile, jer bi se time, u XML smislu, preklopile s privilegiranom hijerarhijom.

Identificiranjem početka i kraja komponenti alternativnih hijerarhijskih struktura sprječava se da one ostanu zanemarene pri budućoj računalnoj obradi.

Nedostatak ove metode je što takvi pojedinačni XML elementi ne predstavljaju ugniježđeni tekst, tako da je znatno otežana računalna obrada ovih dokumenata.[94]

Za neke uobičajene strukturalne značajke, TEI uvodi prazne, tzv. *milestone*[95] elemente poput `<pb/>`, `<lb/>`, `<cb/>`, `<gb/>` koji se mogu koristiti kako bi se obilježio početak tekstualne značajke. Primjerice, korištenjem `<lb/>` elementa koji označava prijelaz u novi red

[93] Usp. A Gentle Introduction to XML. Nav.dj.
[94] Usp. Isto.
[95] *Milestone* elementi nazivaju se još i granični elementi (engl. *boundary elements*), odnosno prazni elementi.

(engl. *line breaks*) moguće je naznačiti fizički raspored linija pjesme i njenu gramatičku podjelu na rečenice:

```
<p>
<seg> <lb n="1"/>U U labirintu panike, duž acetilena, <lb n="2"/> nešto
svirepo jednoliko, s provalom nadahnuća, <lb n="3"/> pretače se u gvožđe;
</seg>
<seg> <lb n="4"/> mozak brojeve misli, a srce kovanja vruća. </seg>
</p>
```

Fizički raspored linija može biti označen i generičkim elementom `<anchor>`. Atributi se u tom slučaju koriste kako bi se naznačio tip tekstualne značajke koja se delimitira te na kojem mjestu se dodijeljena instanca značajke teksta otvara i zatvara.

```
<l><anchor subtype ="sentenceStart" type ="delimiter" />
U ovom glasu zavija, kao da orguljaju grobovi </l>
<l>i ječi nešto sumorno, strahovito i muklo.
<anchor subtype ="sentenceEnd" type ="delimiter" /></l>
```

4.3.3. Fragmentacija i rekonstruiranje virtualnih elemenata

Treća metoda uključuje fragmentiranje nečega što bi se moglo smatrati jedinim logičkim (ali ne ugniježđenim) elementom u više manjih strukturiranih elemenata koji se uklapaju u dominantnu hijerarhiju, ali mogu biti rekonstruirani virtualno. Elementi se razbijaju u što veći broj fragmenata koji se nazivaju parcijalni elementi (engl. *partial elements*) i tako sve dok se problem preklapanja ne riješi. [96]

U sljedećem primjeru atributom n označena su dva dijela iste rečenice koju prekida nova linija stiha <l>:

```
<l><seg n="1">U ovom glasu zavija, kao da orguljaju grobovi</seg></l>
<l><seg n="2">i ječi nešto sumorno, strahovito i muklo.</seg></l>
```

Tehnika fragmentacije često se nadopunjuje tehnikom virtualnog pridruživanja (engl. *virtual joins*). Virtualno pridruživanje može se koristi za kombiniranje objekata u tekstu u novu hijerarhiju. U idućem primjeru, odnos između dijelova razdvojene rečenice naznačen je eksplicitno korištenjem atributa @next i @prev:

[96] Usp. Marinelli, P. Vitali, F. Zacchiroli, S. Towards the uni cation of formats for overlapping markup. Str. 1 -30. URL: http://upsilon.cc/~zack/research/publications/nrhm-overlapping-conversions.pdf (2013-12-01) Str. 9.

```
<l><seg xml:id="s1" next="s2">U ovom glasu zavija, kao da orguljaju
grobovi</seg></l>
<l><seg prev="s1" xml:id="s2">i ječi nešto sumorno, strahovito i
muklo.</seg> </l>
```

Glavna prednost metoda fragmentacije i virtualnog pridruživanja je da dozvoljavaju izravno upravljanje svim hijerarhijama u tekstu: kako privilegiranim hijerarhijama koje su izravno zastupljene tako i alternativnim koje su pridružene. Glavni nedostaci su (kao i kod većine navedenih metoda) da privilegiranje jedne hijerarhije nad ostalim zahtjeva posebnu obradu za rekonstruiranje elemenata drugih hijerarhija.[97]

4.3.4. Stand-off označavanje

Većinu označavanja karakterizira umetanje elemenata u tekst. Alternativni je pristup razdvajanje teksta i elemenata koji se koriste za opisivanje tog teksta. Takav pristup poznat je kao *stand-off* označavanje. On uspostavlja novu hijerarhiju izgradnjom novog stabla, čija čvorišta, u vidu XML elementa koji se nalazi unutar tog stabla, ne sadrže tekstualni sadržaj već poveznice na drugu razinu: čvor u drugom XML dokumentu ili dijelu teksta.

Ova metoda može se provoditi na nekoliko načina. *Prvi* se odnosi na sadržaj na koji će se primijeniti obilježavanje. Ponekad poveznica upućuje na sadržaj eksplicitno, kao u sljedećem primjeru gdje se koristi atribut `@xml:id` s vrijednosti "`w`" kako bi ih se pridružilo i povezalo s imenskim prostorom `<xi:include>`.

```
<l>
<w xml:id="w001">U</w>
<w xml:id="w002">labirintu</w>
<w xml:id="w003">panike</w>,
<w xml:id="w004">duž</w>
<w xml:id="w005">acetilena</w>,
</l>

<l>
<w xml:id="w006">nešto</w>
<w xml:id="w007">svirepo</w>
<w xml:id="w008">jednoliko</w>,
<w xml:id="w009">s</w>
<w xml:id="w010">provalom</w>
```

[97] Usp. A Gentle Introduction to XML. Nav.dj.

```
<w xml:id="w011">nadahnuća</w>
</l>

<l>
<w xml:id="w012">pretače</w>
<w xml:id="w013">se</w>
<w xml:id="w014">u</w>
<w xml:id="w015">gvožđe</w>; <w xml:id="w016">paralelno</w>,
<w xml:id="w017">desna</w>
<w xml:id="w018">je</w>
<w xml:id="w019">uposlena</w>,
</l>

<l>
<w xml:id="w020">mozak</w>
<w xml:id="w021">brojeve</w>
<w xml:id="w022">misli</w>,
<w xml:id="w023">a</w>
<w xml:id="w024">srce</w>
<w xml:id="w025">kovanja</w>
<w xml:id="w026">vruća</w>
</l>

<!-- drugdje u istom dokumentu -->

<p xmlns:xi="http://www.w3.org/2001/XInclude">
<seg>
<xi:include xpointer="range(element(w001),element(w015))"/>
</seg>
<seg>
<xi:includ xpointer="range(element(w016),element(w026))"/>
</seg>
</p>
```

Dio kôda koji upotrebljava `<xi:include>` za izgradnju druge hijerarhije može biti pohranjen u drugom dokumentu i u tom slučaju vrijednost za `@href` od `<xi:include>` treba biti URL od dokumenta koji sadrži osnovni dio kôda, u ovom slučaju `<w>` elemente.[98]

Drugi način provedbe ove metode odnosi se na broj dokumenata koji mogu poslužiti kao poveznice. Najčešće se za jednu oznaku upotrebljava poveznica koja će se onda koristiti i za sve druge oznake. Također, pomoću linkova moguće je povezati više razina.

[98] *XInclude* i *XPointer* predstavljaju XML tehnologije. *XInclude* je generički mehanizam za spajanje XML dokumenata. Usp. XML Inclusions (XInclude) Version 1.0 (Second Edition). URL: http://www.w3.org/TR/xinclude/ (2013-12-19) *XPointer* dopušta poveznicama da upute na mnogo specifičnije dijelove u XML dokumentu poput primjerice na tipove elemenata, vrijednosti atributa i sl. Usp. XML Pointer Language (XPointer) Version 1.0. URL: http://www.w3.org/TR/2001/CR-xptr-20010911/ (2013-12-19)

Brojne su prednosti *stand-off* označavanja, a neke od njih su:

- moguće je označavati tekst čak i ako je izvorni dokument samo za čitanje (engl. *read-only*);
- metodu je moguće primjenjivati i kada je osnovni tekst označen jednostavno, tj. nema XML oznaka;
- označeni dokumenti mogu se distribuirati bez da se distribuira i izvorni tekst;
- diskontinuirani dijelovi teksta mogu se kombinirati u zasebnoj oznaci;
- neovisni enkoderi mogu kreirati nezavisne oznake;
- označene datoteke mogu sadržavati različite razine informacija.[99]

No, postoji i nekoliko nedostataka ove metode. Jedan od njih je da novo obilježene razine zahtijevaju zasebna tumačenja, a razine – iako zasebne – ovise jedna o drugoj. Štoviše, iako su uključeni svi podaci o višestrukim hijerarhijama, njima je teško pristupiti pomoću generičke metode.[100]

Kao rezime o XML pristupima može se zaključiti sljedeće; tekstovi označeni u XML formatu imaju brojne prednosti budući da se svaka postojeća XML tehnologija i alat može koristiti kako bi se obradio tekst. Međutim, struktura koju zahtijeva stroga XML sintaktička pravila je stablasta i da bi se označile različite strukture teksta preduvjet je da je tekst u toj formi. Ukoliko nije, njega se na neki način 'prisiljava' da formira stablastu strukturu. TEI zajednica ponudila je nekoliko rješenja, a oni su prethodno izneseni. Svaka od metoda ima brojne prednosti ali isto tako i nedostatke te je potrebno još jednom istaknuti kako trenutno nema jedinstvenog rješenja i da svaka situacija zahtjeva jedinstveni pristup rješenju ovog problema.

[99] Usp. A Gentle Introduction to XML. Nav.dj.
[100] Usp. Isto.

4.4. Pristupi rješenju problema preklapanja hijerarhijskih struktura u tekstu bez uporabe XML-a

Kako bi se prevladao problem hijerarhijskih struktura razvijene su i druge metode koje problem prevladavaju van okvira standardnog XML-a. Pojedine su utemeljene na nestandardnoj XML sintaksi dok se druge uopće ne temelje na XML sintaksi. Neka od ovih rješenja su:

- Struktura podataka GODDAG[101] (engl. *General Ordered-Descendant Directed Acyclic Graph*) i označiteljski jezik TexMECS[102] čiji su glavni zagovornici M. Sperberg-McQueen i C. Huitfeldt.
- Označiteljski jezik LMNL[103] (engl. *Layered Markup and Annotation Language*) čiju primjenu zastupaju J. Tennison i W. Piez.
- MuLax sintaksa i SGML-ova funkcija CONCUR-a[104] čiju uporabu zagovaraju M. Hilbert i A. Witt;
- Koncept JITTs (engl. *Just-in-Time-Trees*)[105] i drugi.

Opis navedenih rješenja preklapanja hijerarhijskih struktura u tekstu bez uporabe XML-a zbog izrazite kompleksnosti prevazilazi granice ovog rada.

[101] Usp. Sperberg-McQueen, M., Huitfeldt. C. GODDAG: A Data Structure for Overlapping Hierarchies. URL: http://www2.iath.virginia.edu/ach-allc.99/proceedings/sperberg-mcqueen.html (2013-12-01)
[102] Usp. Texmacs. TeXmacs style files. URL: http://www.texmacs.org/tmweb/manual/webman-style.en.html(2013-12-01)
[103] Usp. Tennison, J., Piez, W. The Layered Markup and Annotation Language (LMNL). URL: http://xml.coverpages.org/LMNL-Abstract.html (2013-12-01)
[104] Usp. Hilbert, M., Witt, A. Making CONCUR work. A Conference of IDEAlliance: Extreme Markup Languages 2005. Montréal, Québec August 1-5, 2005. Str. 1 – 20. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.634&rep=rep1&type=pdf (2013-12-01)
[105] Usp. Durusau, P. JITTs (Just-in-Time-Trees). 2004. URL: http://www.durusau.net/publications/NY_xml_sig.pdf (2013-12-01)

5. ZAKLJUČAK

Pojavom računalnih tekstualnih sustava javljaju se nove vrste označavanja i nove vrste obrade teksta. Način na koji je tekst reprezentiran na računalu utječe na mogućnosti korištenja tog istog teksta. Kad je tekst pohranjen u elektroničkim datotekama on je obilježen posebnim tipovima elektroničkih oznaka namijenjenih za obradu putem računalnih aplikacija. Prvotni modeli teksta koji su bili zastupljene u programima za računalnu obradu, i u kojima je tekst promatran kao bitmapa, niz znakova, kao *page layout* i dr., bili su vrlo restriktivni; tekst je bio pogodan samo za najjednostavnije analize, a mogućnosti formatiranja, ažuriranja i sl. vrste obrade bile su vrlo ograničene. Kao odgovor na te 'manje sofisticirane' modele 1990-ih predložen je model teksta, poznat kao OHCO, koji je bio jednostavniji i predstavljao je funkcionalniji način za kreiranje, mijenjanje i oblikovanje teksta, s tim da je podržavao pretraživanje, pregledavanje, analiziranje i druge vrste posebne obrade teksta te omogućio jednostavniju razmjenu tekstova među različitim softverskim i računalnim sustavima. U izvornoj verziji OHCO modela tekst je promatran kao hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu. No, OHCO teorija ubrzo se suočila s problemom preklapanja logičkih i fizičkih elemenata u tekstu što je bilo u suprotnosti s idejom o tekstu koji posjeduje jednu hijerarhijsku strukturu i u tom smislu prvotna OHCO teorija je opovrgnuta. Kao odgovor na ovaj problem nastala je OHCO-2 teorija, u kojoj je uveden koncept analitičkih perspektiva kojima se odražavaju višestruke hijerarhije u samom tekstu. Ipak, kasnije se došlo do spoznaje da postoje i takve analitičke perspektive koje se sastoje od objekata sadržaja koji se međusobno mogu preklapati, što narušava OHCO-2 teoriju. U novoj, OHCO-3 inačici predloženo je razgrađivanje perspektiva na pod-perspektive. No, ubrzo je uočeno da postoje objekti sadržaja koji se međusobno preklapaju, ali se ne mogu svrstati u različite pod-perspektive, kako predviđa OHCO-3 teorija. Pokazalo se da tekst predstavlja svojevrstan sustav perspektiva koje se raščlanjuju u pod-perspektive, a koje se pak dalje mogu raščlanjivati na pod-pod-perspektive i tako u nedogled. Modifikacijom prvotne OHCO teorije, najprije u OHCO-2, zatim i u OHCO-3, problem preklapanja pokušao se nadoknaditi, no pokazalo se da izmjene nikad u potpunosti nisu funkcionirale – preklapanje se i dalje pojavljivalo i pobijalo osnovnu ideju o tekstu kao hijerarhiji razvrstanih objekata sadržaja.

Kako u okviru OHCO teorije nije formulirana prihvatljiva definicija teksta, istraživači iz područja digitalne humanistike i označavanja teksta nastavili su svoje napore u promišljanju prirode i svojstava teksta te su razvili nove poglede s drugačijom perspektivom promatranja teksta: platonizam, pluralizam i antirealizam.

Problemu preklapanja hijerarhijskih struktura, u tehničkom smislu pokušalo se doskočiti razvojem metoda temeljenih na XML-u koje su omogućile rješenje problema neugniježđenih elemenata odnosno onih elemenata koji ne dopuštaju da ih se prikaže u strogo hijerarhijskom obliku. Neke od tih metoda su uporaba *milestone* elemenata (tzv. 'praznih' elemenata), višestruko označavanje istog teksta, postupak fragmentiranja sadržaja i *stand-off* označavanje, a sve su one opisane i praktično predstavljene u radu. Bez obzira na to što su i u tehničkom smislu ostali neki problemi i ograničenja – predložene metode omogućile su, više ili manje uspješno, razrješenje problema preklapanja hijerarhijskih struktura u tekstu.

Na koncu valja zaključiti kako problemi vezani uz označiteljsku teoriju ne trebaju biti primarno vezani uz tehnička pitanja i rješenja, nego uz puno fundamentalniji problem – razumijevanje onoga što zapravo radimo kada neki tekst označavamo. Na kraju se i pokazalo da ukoliko se želi shvatiti što se točno zbiva pri procesu označavanja teksta stav da je tekst hijerarhija objekata sadržaja razvrstanih prema točnom redoslijedu više ne zadovoljava. OHCO, kao teorija kojom se isključivo zastupala struktura, ali ne i semantika teksta, jednostavno nije bila zadovoljavajuća za razvoj opće teorije teksta. U ovom radu prikazan je razvoj OHCO teorije koja je imala ambiciju postati općom teorija teksta, no u tome, kako se pokazalo, nije uspjela. Radom se htjelo ukazati na uzroke i razloge zbog kojih je spomenuta teorija opovrgnuta, ali koja bez sumnje nosi zasluge za promišljanje o tekstu i generalno označavanju teksta. Isto tako, htjelo se ukazati na značajnu ulogu koju je teorija bez obzira na propuste odigrala u promicanju i zastupanju opisnih označiteljskih jezika (SGML-a, XML-a) u postupcima označavanja tekstova te poslužila kao odgovarajući okvir za razvoj TEI-a – vodećeg standarda za označavanje teksta.

LITERATURA

1. A Gentle Introduction to XML. // TEI P5: Guidelines for Electronic Text Encoding and Interchange. A TEI Consortium eds, 2013. URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html (2013-12-01)
2. Association for Computational Linguistics. URL: http://www.aclweb.org/ (2013-12-01)
3. Association for computers and the humanities. URL: . http://www.ach.org/ (2013-12-01)
4. Association for literary and linguistic computing. URL: http://www.allc.org/ (2013-12-01)
5. Barnard, D., Hayter R., Karababa M., Logan G., and McFadden, J. (1988), 'SGML-Based Markup for Literary Texts: Two Problems and Some Solutions', Computers and the Humanities 22: Str. 265-276 URL: http://link.springer.com/article/10.1007%2FBF00118602#page-1 (2013-12-01)
6. Biggs,M., Huitfeldt, C.: Philosophy and Electronic Publishing. Theory and Metatheory in the Development of Text Encoding. Edited Discussion. URL: http://www.philo.at/mii/mii/node8.html#SECTION00230000000000000000 (2013-12-01)
7. Coombs, J. H.; Renear, A. H.; DeRose, S. J. Markup systems and the future of scholarly text processing . // Communications of the ACM. 30, 11(1987), Str. 933-947. URL: http://cpe.njit.edu/dlnotes/CIS/CIS732_447/Cis732_6R.pdf (2013-12-01)
8. DeRose, S.; Durand, D.; Mylonas, E. What is text, really? // Journal of Computing in Higher Education. 1, 2(1990) Str.1-24 URL: http://www.hki.uni-koeln.de/sites/all/files/courses/3226/Renear_ea-1997.pdf (2013-12-01)
9. Design Principles for Text Encoding Guidelines: The Poughkeepsie Principles. URL: http://projects.oucs.ox.ac.uk/teiweb/Vault/ED/edp01.xml?ID=body.1_div1.2 (2013-12-01)

10. Durusau, P. JITTs (Just-in-Time-Trees). 2004. URL: http://www.durusau.net/publications/NY_xml_sig.pdf (2013-12-01)
11. Durusau, P., O'Donnell. Cuncurrent Markup for XML Documents. URL: http://www.durusau.net/publications/Concurrent_markup.pdf (2013-12-01)
12. Goldfarb, C. F. A generalized approach to document markup. // ACM SIGPLAN-SIGOA Portland, SAD, 1981. Str. 68-73. URL: http://www.fdi.ucm.es/profesor/jlsierra/e-learning/segunda-sesion/goldfarb.pdf (2013-12-01)
13. Goldfarb, C. F. The SGML history niche: the roots of SGML - a personal recollection - a personal recollection, 1996. URL: http://www.sgmlsource.com/history/roots.htm (2013-12-18)
14. Hilbert, M., Witt, A. Making CONCUR work. //A Conference of IDEAlliance: Extreme Markup Languages 2005. Montréal, Québec August 1-5, 2005. Str. 1-20. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.634&rep=rep1&type=pdf (2013-12-01)
15. Huitfeldt, C. Multi-Dimensional Texts in a One-Dimensional Medium.Str.142-161. URL: http://wab.uib.no/ojs/agora-wab/article/view/1860 (2013-12-01)
16. International organization for standardization: information processing – text and office systems – Standard Generalized Markup Language (SGML). URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387 (2013-12-01)
17. Kovačić, I.G. Jama. URL: http://ponude.biz/knjige/i/Ivan%20Goran%20Kovacic%20-%20Jama.pdf (2013-12-19)
18. Lamport , L. LaTeX: A Document Preparation System. User's Guide and Reference Manual. Addison-Wesley Pub. Co. 1994. Str. 2-25 URL: http://www.stat.pitt.edu/stoffer/freetex/latex%20basics.pdf (2013-12-01)
19. LaTeX: a document preparation system. URL: http://www.latex-project.org/ (2013-12-01)
20. M.Sperberg-McQueen, Burnard, L. AGuidelines for Electronic Text Encoding and Interchange. (TEI P3) Gentle Introduction to SGML. URL: http://www-sul.stanford.edu/tools/tutorials/html2.0/gentle.html (2013-12-01)

21. Marinelli, P. Vitali, F. Zacchiroli, S. Towards the uni cation of formats for overlapping markup. Str. 1 -30. URL: http://upsilon.cc/~zack/research/publications/nrhm-overlapping-conversions.pdf (2013-12-01)
22. Overlapping Markup SIG. Text Encodin Initiative. URL: http://www.tei-c.org/Activities/SIG/Overlap/ (2013-12-01)
23. Pichler, A. Advantages of a Machine-Readable Version of. Wittgenstein's Nachla B. Str. 770-776. URL: http://wab.uib.no/alois/pichler-kirchb95b.pdf (2013-12-01)
24. Renear, A. H. Text encoding. A companion to digital humanities. Susan Schreibman, Raymond George Siemens and John M. Unsworth. Wiley-Blackwell, 2004. URL: http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-5&toc.depth=1&toc.id=ss1-3-5&brand=default (2013-12-01)
25. Renear, A. Out of praxis: three (meta)theories of textuality. New York, Oxford University Press, 1997. Str. 108-126. URL: http://books.google.hr/books?id=4dZ2XSr8Q2cC&pg=PA107&lpg=PA107&dq=Renear++Out+of+praxis+:+three+(meta)theories+of+textuality&source=bl&ots=15c0WWkRPg&sig=w5S45h87zoc39Scw8AYljG08C3A&hl=hr&sa=X&ei=uAuRUqDnEYba4QShnIDgAQ&ved=0CC0Q6AEwAA#v=onepage&q=Renear%20%20Out%20of%20praxis%20%3A%20three%20(meta)theories%20of%20textuality&f=true (2013-12-01)
26. Renear, A.; Mylonas, E.; Durand, D. Refining our notion of what text really is: the problem of overlapping hierarchies. // Research in Humanities Computing. Oxford University Press, 1996. URL: http://www.stg.brown.edu/resources/stg/monographs/ohco.html (2013-12-01)
27. Schmidt, D., Colomb, R. A Data Structure for Representing Multi-version Texts Online. Str.1-20 URL: http://itee.uq.edu.au/~schmidt/_articles/elsevier.pdf (2013-12-01)
28. Sperberg-McQueen, M. Text in the electronic age: textual study and text encoding, with examples from medieval texts. // Literary and Linguistic Computing. 6, 1(1991)
29. Sperberg-McQueen, M., Huitfeldt. C. GODDAG: A Data Structure for Overlapping Hierarchies. URL: http://www2.iath.virginia.edu/ach-allc.99/proceedings/sperberg-mcqueen.html (2013-12-01)

30. Tennison, J., Piez, W. The Layered Markup and Annotation Language (LMNL). URL: http://xml.coverpages.org/LMNL-Abstract.html (2013-12-01)
31. Texmacs. TeXmacs style files. URL: http://www.texmacs.org/tmweb/manual/webman-style.en.html (2013-12-01)
32. The Poughkeepsie Principles Closing Statement of Vassar Conference The Preparation of Text Encoding Guidelines. Poughkeepsie, New York: 1987. URL: http://www.tei-c.org/Vault/ED/edp01.htm#b2b1b3b3b3 (2013-12-01)
33. Ujević, T. Izabrane pjesme. URL: http://ponude.biz/knjige/t/Tin%20Ujevic%20-%20Izabrane%20pjesme.pdf (2013-12-01)
34. W3schools: XML Syntax Rules. URL: http://www.w3schools.com/xml/xml_syntax.asp (2013-12-01)
35. W3schools: XML Inclusions (XInclude) Version 1.0 (Second Edition). URL: http://www.w3.org/TR/xinclude/ (2013-12-19)
36. W3schools: XML Pointer Language (XPointer) Version 1.0. URL: http://www.w3.org/TR/2001/CR-xptr-20010911/ (2013-12-19)

PRILOG

Prilog 1. Primjer XML sintakse i shematski prikaz iste (u strukturi stabla)

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="KNJIŽEVNOST">
     <title lang="hr">Lanzarote i drugi tekstovi</title>
     <author>Michel Houellebecq</author>
     <year>2002</year>
     <price>195.00</price>
  </book>
  <book category="FILOZOFIJA">
     <title lang="hr">Brevijar poraženih</title>
     <author>Emile Michel Cioran</author>
     <year>2009</year>
     <price>100.00</price>
</book>
</bookstore>
```