Apliko de XML al fakaj laboroj pri Esperanto

Dum la lastaj jaroj okazis impona disvolviĝo de la uzo de XML-a "teĥnologio" en ĉiuj kampoj de komputiko kaj, precipe, en TTT-servoj. Vulgarigistoj eĉ ne hezitas diri, ke XML estas la Esperanto de komputiko. Ĉiuokaze, tiu ekscitiĝo naskis plurajn XML-lingvojn, por plej diversaj uzoj, kaj senpagajn ilojn, kiuj signifagrade simpligas la laborojn de aŭtomata pritraktado de tekstoj kaj datumbazoj. La plej populara apliko de XML en Esperantujo estas la vortara projekto REVO (Reta Vortaro), sed preter ĉi tiu ekzemplo XML estas rekomendinda solvo por ĉiu leksikografia laboro kaj ĉiu terminologia datumbazo.

Ces dernières années ont vu un décollage impressionnant de l'utilisation de la « technologie » XML dans tous les domaines de l'informatique et, plus particulièrement, dans les services Web. Des vulgarisateurs n'hésitent pas à dire que XML est l'espéranto de l'informatique. Quoi qu'il en soit, cette effervescence a produit un grand nombre de langages XML, destinés à des usages très variés, et des outils gratuits qui simplifient grandement tous les travaux de traitement automatique de textes et bases de données. L'application phare du XML dans le monde espéranto est le projet de dictionnaire REVO (Reta Vortaro) mais, au-delà de cet exemple, XML est une solution de choix pour tout travail lexicographique et toute base de données terminologique.

De kelkaj jaroj XML (legu: ikso-mo-lo) fariĝis moda temo en komputika mondo. La prikomputila gazetaro ĝin celebras kiel "la universalan lingvon" aŭ eĉ "la Esperanton de komputiko". Por scii, pri kio temas, kaj kiel la afero povus vere rilati kun Esperanto, ni unue provu elvolvi la anglan akronimon al ĝia signifo, nome *eXtensible Mark-up Language*, laŭvorte tradukebla per "etendebla mark-lingvo", kio ne multe pli klaras. Pri "lingvo" certe temas, sed formala, de la tipo uzebla en dialogo kun komputiloj, ne homa. Cetere, temas pri "marklingvo", t.e. formala lingvo uzebla por evidentigi la *strukturon* de dokumento: ni ĉiuj scias, ke teksto, frazoj, vortoj ktp havas logikan strukturon, kaj tio momente sufiĉu, sed precizigo pri tiu nocio baldaŭ venos. Fine restas por klarigi "etendebla". Nu, malantaŭ ĉi tiu senkulpa adjektivo kaŝiĝas multo, unue ke fakte XML ne estas marklingvo, sed normo difinanta tutan familion da marklingvoj laŭ la sama modelo, kaj due, ke laŭ la donita modelo vi tute rajtas konstrui la marklingvon, kiu plej bone taŭgas por evidentigi la specifan strukturon de *viaj* dokumentoj. Laste ni aldonu, ke kiam oni nun parolas pri XML, oni ne nur parolas pri la normo pri marklingvoj, sed ankaŭ pri la iloj ebligantaj operacii super la tiele markitaj dokumentoj, kaj ankaŭ pri la uzoj de tiaj dokumentoj kadre de TTT (Tut-Tera Teksaĵo).

Strukturo

Kio do estas tiu strukturo, kiun oni povus deziri evidentigi, kaj kial entute ĝin evidentigi? Nu, imagu, ke vi skribas leteron. Vi supozeble havas kutiman strukturon por tio: unue vi skribos la daton kaj la lokon. Poste venos alvoko de la tipo "Kara Oĉjo", "Estimata samideano" aŭ simila. Poste vi supozeble referencos la antaŭe ricevitan korespondaĵon, respondos al kelkaj demandoj el ĝi, kaj poste informos pri viaj travivaĵoj. Fine de la letero venas adiaŭa formulo, subskribo kaj eventuala postskribaĵo. Homa leganto facile povas kompreni la strukturon de via letero, unue pro ĝia kutimeco, sed ankaŭ ĉar vi uzis tipografiajn rimedojn por ĝin evidentigi: la loko kaj la dato kuŝas supre de la paĝo, la alvoko konsistigas izolitan linion tuj poste ktp.

Ni nun transiru al pli faka ekzemplo, nome Esperanta vortaro de la tipo PIV. Se preteratenti la antaŭparolon, postparolon, eldonejan reklamon ktp, ni povas diri, ke vortaro esence konsistas el listo da artikoloj. Ĉiu artikolo rilatas al unu radiko kaj komenciĝas per kapvorto, kiu mem havas tre komplikan strukturon, kiel ni poste montros. Sekve prezentiĝas la diversaj sencoj de la kapvorto, kaj por ĉiu senco la teksto ampleksas diversajn elementojn, devigajn aŭ ne, kiel vinjetoj, difino, ekzemploj, rimarkoj kaj referencoj al aliaj vortoj. Ankoraŭ poste venas la derivaĵoj kaj kunmetaĵoj de la koncerna radiko, kaj por ĉiu el ili ripetiĝas la sama strukturo, kiel por la artikola kapvorto. Denove tipografiaj rimedoj servas por substreki la strukturon kaj helpi homan leganton analizi la tekston en la ĝusta kadro.

Imagu nun, ke vi estas terminologo-fizikisto, deziranta trovi ĉiujn vortojn temantajn pri fiziko en la *Nova Plena Ilustrita Vortaro*. Via unika rimedo por tion atingi estas preni la vortaron en la manojn, foliumi ĉiujn paĝojn, serĉi la fizikan vinjeton per viaj okuloj kaj, kiam vi trafas ĝin, retropaŝi al la koncerna derivaĵo (supozu, ke vi trovis "**~igi**"), kaj poste al la kapvorto de la koncerna artikolo (ekz-e "**nebul/o**") por malkovri la radikon. Nun vi rekunigis la elementojn de la koncerna fizika termino (en nia ekzemplo "**nebuligi**") kaj vi skribas ĝin sur apuda papero. Iom longe kaj tede, ĉu ne? Supozeble komputilo povus helpi, do vi afable petas la saman taskon al la respondeculo de S.A.T., kiu sur la disko de sia komputilo havas la tutan dosieron de la *Nova PIV*, ekzemple en formato de *Microsoft Word*, aŭ de simila teksto-prilaborilo. Certe por li la trovo de la vinjetoj estos pli rapida, ol per permana foliumado, sed la procedo por rekonstrui la koncernajn terminojn estos same neaŭtomata, ol ĝi estas por vi, ĉar neeblas instrukcii al la programo *Microsoft Word*, ke ĝi malkovru la formon de la koncerna derivaĵo kaj poste anstataŭigu la tildon per la koncerna radiko. Kompreneble, oni teorie povus krei programon kapablan fari ĉi tion, sed ĝi devus esti kapabla kompreni la tipografian formaton de la tekstoprilaborilo kaj reprodukti la homan rezonadon rilate tipografion... Malfacila tasko, por tre specifa programo.

Memkomprenebla konkludo estas, ke se oni volas vere aŭtomatigi tiun ĉi laboron, necesas evidentigi la strukturon de la dokumento (la vortaro) tiamaniere, ke komputila programo povu ĝin ekspluati. Tipografiaj rimedoj estas nesufiĉe formalaj, nesufiĉe fidindaj kaj tro hom-orientitaj por tio: la plej bona rimedo estas enkonduki en la tekston markojn en

formo de konvenciaj ŝlosilvortoj, kiuj eksplicite indiku “tiu ĉi segmento de la teksto estas kapvorto, difino, fak-indiko, derivaĵo ktp”. Ni sekve montros ekzemplon de marklingvo plenumantan ĉi tiun celon.

Skizo de marklingvo por vortaristoj

El la supra diskuto sekvas, ke la strukturaj elementoj aspektas kiel skatoloj: la elemento “vortaro” entenas multajn elementojn “artikolo”, respondantajn al ĉiu radiko; kiuj artikoloj mem entenas elementon “kapvorto” kaj unu aŭ plurajn elementojn “derivaĵo”; kiuj derivaĵoj mem entenas... kaj tiel plu. Tamen teksto estas unudimensia strukturo kaj por prezenti tiajn skatolojn, sufiĉas uzi du limmontrilojn, nome la komencan kaj la finan *markojn*. Komenca marko konsistas el iu vorto inter angulaj krampoj, kiel `<vortaro>`, `<artikolo>`, `<kapvorto>`, `<derivaĵo>`, `<senco>`... La vorto povas esti en ajna lingvo (bulgara, ĉina, franca, japana, korea, portugala, rusa..., eĉ Esperanto) prezentebla per la vaste uzata signokodo Unikodo, kaj ĝi laŭeble rilatu kun la struktura rolo de la koncerna elemento. Ni poste vidos, ke post ĉi tiu vorto povas aperi *atributoj*. La fina marko havas similan formon, krom ke tuj post la malferma krampo aperas oblikva streko, do: `</vortaro>`, `</artikolo>`, `</kapvorto>`, `</derivaĵo>`, `</senco>`... Ni resumu: la teksto konsistas el diversaj “skatoloj”, nomataj *elementoj*, komenciĝantaj per *komenca marko* kaj finiĝantaj per *fina marko*. Ĉio inter la du respondaj markoj estas la *enhavo* de la koncerna elemento. Se estas nenio inter la du markoj, oni diras, ke la elemento estas malplena, kaj oni rajtas uzi la mallongigon `<elemento/>` anstataŭ `<elemento></elemento>`.

Fig. 1 montras tre simpligitan modelon de nia vortaro: aperas nur la unua kaj lasta artikoloj, kaj en la unua aperas la elemento “kapvorto”.

La tabelo de Fig. 2 prezentas, kiamaniere la interna strukturo de la artikola kapvorto estas prezentita en PIV per tre pedantaj tipografiaj rimedoj: ekzemple, la grado de oficialeco estas prezentita jen per komenca steleto (por Fundamentaj radikoj), jen per nombra supera indico (por nefundamentaj, sed oficialaj radikoj), kiu indico povas anstataŭi, sekvi aŭ antaŭiri la oblikvan strekon, jen per fina supera indico Z (por neoficialaj, sed Zamenhofaj radikoj), jen per neniu signo (por radikoj nek oficialaj, nek Zamenhofaj).

```
<vortaro>
   <artikolo>
      <kapvorto>
         ...
      </kapvorto>
      ...
   </artikolo>
   ...
   <artikolo>
      ...
   </artikolo>
</vortaro>
```

Fig. 1: Vortara modelo

Kapvorto	Radiko	Finaĵo	Oficialeco	Diversaj
*fenestr/o	fenestr	o	Fundamenta	
eskap^{8}i	eskap	i	8-a OA	
esoter/a^{Z}	esoter	a	Zamenhofa	
biosfer/o	biosfer	o	neoficiala	
*antaŭ	antaŭ		Fundamenta	memstara
*-ad/	ad		Fundamenta	sufikso
-aĉ/^{1}	aĉ		1-a OA	sufikso
*bo/	bo		Fundamenta	prefikso
mis^{4}/	mis		4-a OA	prefikso

Fig. 2: Strukturo de la kapvorto

Krom oficialecon, la kapvorto provizas ankaŭ la formon de la radiko[1], la t.n. gramatikan karakteron de la radiko kaj diversajn aliajn gramatikajn informojn, kiujn eblas kombini kun la gramatika karaktero. Ni do povas prezenti la strukturon de la artikola kapvorto per tri elementoj: “radiko”, “oficialeco”, “karaktero”. Jen ekzemploj en Fig. 3 kaj Fig. 4.

```
<kapvorto>
  <radiko>antaŭ</radiko>
    <oficialeco>FU</oficialeco>
  <karaktero>memstara</karaktero>
</kapvorto>
```

Fig. 3

```
<kapvorto>
  <radiko>biosfer</radiko>
  <oficialeco>neniu</oficialeco>
  <karaktero>o</karaktero>
</kapvorto>
```

Fig. 4

1. En ĉi tiu teksto ni uzas la vorton “radiko” pli amplekse, ol oni kutime faras.

Ĉi-okaze ni vidas por la unua fojo, ke elementoj povas enhavi ne nur aliajn elementojn, sed ankaŭ puran tekston. Eĉ miksaĵo de elementoj kun pura teksto estas permesita enhavo.

Plia instruo de la ekzemplo estas la graveco kontroli la enhavon de iuj elementoj: ja por oficialeco kaj karaktero ekzistas nur kelkaj akcepteblaj valoroj. Tial la komputilo estas taŭga ilo por kontroli, ke la enhavo kongruas kun la akcepteblaj valoroj, ke ni, ekz-e, ne erare tajpis "c" anstataŭ "o" por la elemento karaktero, aŭ ke ni ne forgesis entajpi la elementon "oficialeco", aŭ ke ni ne erare enmetis ĝin en elementon "radiko". Por tio sufiĉas doni al kontrola programo la formalan modelon de nia marklingvo, kiun modelon fakuloj nomas la *dokument-tip-difino* (DTD), aŭ la *ŝemo* de la lingvo[2].

```
<artikolo>
  <kapvorto>
    <radiko>antaŭ</radiko>
    <oficialeco>FU</oficialeco>
    <karaktero>memstara</karaktero>
  </kapvorto>
```

Fig. 5

```
<artikolo>
  <radiko>antaŭ</radiko>
  <oficialeco>FU</oficialeco>
  <karaktero>memstara</karaktero>
  ...
</artikolo>
```

Fig. 6

Denove ni ne forgesu, ke la marklingvon difinas *ni*, kaj ke ni do rajtas adapti ĝin al la logiko de nia problemo. Ekz-e, nun kiam estas difinita la enhavo de elemento "kapvorto", ni povas veni al la konkludo, ke servas por nenio apartigi ĉi tiujn elementojn en "kapvorto" kaj ke same bonus havi ilin en rekta dependeco de elemento "artikolo", t.e. havi la strukturon de Fig. 6 anstataŭ de Fig. 5.

Eblas eĉ decidi, ke la tri elementoj "radiko", "oficialeco" kaj "karaktero", kiuj mem ne povas enhavi aliajn elementojn, estu ŝovitaj en la komencan markon de elemento "artikolo" en formo de *atributoj*[3], kiel montras Fig. 7.

```
<artikolo radiko="antaŭ" oficialeco="FU" karaktero="memstara">
...
</artikolo>
```

Fig. 7

Pere de la vortara ekzemplo ni povis familiariĝi kun la plej gravaj reguloj de XML, sufiĉaj por la cetero de ĉi tiu prelego. Kompreneble eblas lerni pli multe, legante nacilingvan vulgarigan libron, da kiuj ekzistas granda nombro.

Operacioj super XML-a dokumento

Ekzistas du tipoj da kutimaj operacioj efektivigeblaj super XML-a dokumento. Unue kontrolo, t.e. aŭtomata pruvo, ke la dokumento kongruas kun la difino de la marklingvo, en kiu ĝi estas skribita. Por tio oni uzu specialan ilon[4], kiu helpe de la DTD aŭ de la ŝemo de la koncerna lingvo kapablas pritaksi la konformecon de la dokumento.

---

2. Ŝemoj estas pli nova teĥniko, pli potenca en la senco, ke ĝi ebligas pli fajne difini la marklingvon. Aldone, la lingvo uzata por redakti ŝemojn, kiu nomiĝas XML-Ŝemo, estas XML-a marklingvo.

3. Ĉiu atributo prezentiĝas kiel esprimo `identigilo="valoro"`, en kiu "valoro" estas pura teksto. La DTD-oj ne kapablas difini la akcepteblajn valorojn de atributo, kio iom bremsis ilian uzon, sed la ŝemoj kapablas.

4. Por la kutimaj operacioj ekzistas pluraj senpagaj iloj. Por homo ne tro komputillerta povas esti malfacile trovi la ĝustan ilaron, pro kio ni konsilas peti helpon de pli spertaj homoj, ekz-e de la redaktantoj de REVO.

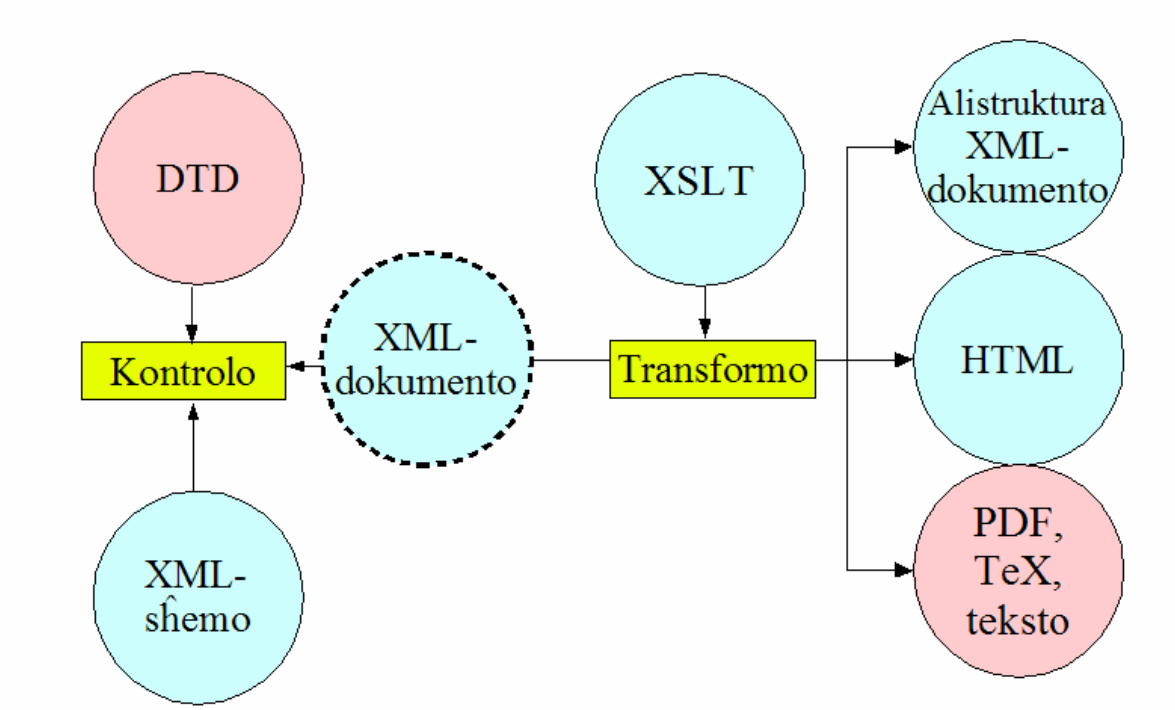

Fig. 8: Operacioj super XML-dokumento

La dua tipo de operacio estas transformo, t.e. aŭtomata transigo de la dokumento de unu marklingvo al alia, aŭ eĉ al simpla teksto. Por kio ĝi servas? Nu, ekzemple, por ekstrakti de nia vortaro la liston de ĉiuj substantivaj radikoj, aŭ de ĉiuj terminoj kun difinita fako, aŭ de ĉiuj homonimaj radikoj, aŭ por prezenti al homa leganto nian dokumenton uzante specifajn tipografiajn konvenciojn por ĉiu aparta elemento de ĝi ktp. Transformo povas servi ankaŭ por kunfandi du dokumentojn, ekzemple por aldoni al nia vortaro tradukojn ĉerpotajn el alia XML-dokumento. Ĉiujn ĉi tipojn de transformo la aŭtoro uzis por produkti la manuskripton de sia *Matematika Vortaro*: unue necesis ĉerpi la matematikan materialon de REVO kaj traduki ĝin al specifa marklingvo, due necesis aldoni la ĉeĥajn kaj hungarajn tradukojn (provizitajn de eksteraj kunlaborantoj) kaj fine necesis produkti la malnetan manuskripton en HTML-formato.

Por efektivigi la transformojn oni uzas specialan ilon[4], kiu aplikas al la dokumento iun "transform-programon", nomatan XSLT-skripto. Pro nehazarda koincido XSLT estas samtempe programlingvo kaj XML-a marklingvo[5].

Popularaj marklingvoj

Plej konata estas HTML (legu: ho-to-mo-lo), la marklingvo servanta por redakti TTT-paĝojn[6]. Transformo de XML-dokumento al TTT-paĝo estas aparte interesa por homoj, kiuj deziras publikigi siajn vortarojn en la reto. Ni jam aludis pri XSLT kaj XML-Ŝhemo. Malpli konataj, sed tre potencaj estas SVG (*Scalable Vector Graphics*) kaj SMIL (*Synchronized Multimedia Integration Language*): SVG ebligas desegni komplikajn ebenajn figurojn kaj SMIL ebligas difini multmediajn sinsekvojn[7].

En la Esperanta mondo menciindas la marklingvo de REVO, pri kiu ni parolos pli poste. Tute nekonata, sed potenciale utila, estas la marklingvo ellaborita de la aŭtoro por strukturi datumbazon de kapvortoj de la nova kaj malnova PIV-oj[8].

---

5. Malmulte utilas precizigi, ke la akronimo XSLT signifas *XML Style Sheet - Transform* (XML-a stilfolio - Transformo), ja pri stilfolioj apenaŭ temas. Pli gravas noti, ke la kapablo uzi XML-on por redakti *kaj* alte strukturitajn datumojn, *kaj* programojn eventuale operaciantajn super tiuj datumoj, estas verŝajne unu el la fontoj de la mito pri la universaleco de XML.

6. Ni precizigu, ke HTML naskiĝis antaŭ XML kaj ne tute obeas ĝiajn regulojn. Ekzistas tamen pli rigora versio, nomata XHTML, kiu estas vera XML-a marklingvo.

7. SMIL estas legebla per la senpaga legilo *RealPlayer*. SVG-dokumentoj estas videblaj per iuj TTT-legiloj (*Internet Explorer* kun taŭga aldonaĵo). La aŭtoro uzis SVG por la ilustraĵoj de sia *Matematika Vortaro*, kiuokaze montriĝis tre oportuna la kapablo de SVG-dokumento enhavi alilingvan skripton por kalkuli la punktojn de kurbo.

8. Por pli da detaloj vidu la retpaĝojn:
`http://perso.wanadoo.fr/kursoj/piv1/piv1.htm` kaj `http://perso.wanadoo.fr/kursoj/piv2/-piv2.htm`. La deveno de la koncernaj datumbazoj estas klarigita kaj ekzemplaj uzoj de la XML-a versio estas montritaj.

REVO (Reta Vortaro)

Iniciatite kaj prizorgate de Wolfram Diestel, jam de pluraj jaroj funkcias ĉi tiu elstara Esperanta vortaro, uzata kaj prilaborata per interretaj rimedoj[9]. Ĝi estas difina vortaro, kun ekzemploj, plurlingvaj tradukoj, bildoj, sed ankaŭ meta-enhavo (font- kaj fak-indikoj, referencoj, rimarkoj...) kaj ĝi disponigas plurtipajn indeksojn: Esperantan indekson de vortoj, sed ankaŭ nacilingvajn, fakajn kaj tezaŭrajn indeksojn.

De teĥnika vidpunkto, REVO baziĝas sur tri pilieroj: unue la diskutlisto[10], en kiu redaktantoj kaj uzantoj interŝanĝas opiniojn; due ĝia TTT-ejo[11], en kiu eblas konsulti la vortaron, elŝuti la XML-dosieron de ĉiu artikolo kaj efektivigi simplan redaktolaboron pere de formularo; kaj trie la redakta servilo, kiu retpoŝte akceptas la modifitajn artikolojn kaj pritraktas ilin ĉiutage.

Bedaŭrinde la limigita amplekso de ĉi tiu prelego ne ebligas paroli pri la REVO-a marklingvo. Sufiĉas scii, ke ĝi baziĝas sur vortara strukturo simila al tiu, kiun ni ĉi-supre prezentis. Cetere la marklingvo estas prezentita en la koncerna TTT-ejo, kune kun diversaj detalaj manlibroj.

REVO estas uzebla diverscele, kaj tio donas al ĝi grandan valoron: iuj redaktantoj kontentiĝas per aldono de tradukoj al sia nacia lingvo; iuj faras pure leksikografian laboron; iuj priserĉas tezaŭrojn aŭ la reton por trovi interesajn uz-ekzemplojn kaj enkondukas ilin en koncernajn artikolojn; iuj pritraktas difinitan fakon, revizias ĉiuj ĉi-temajn artikolojn kaj enkondukas terminojn ĉerpitajn el aŭtoritataj fontoj...

Konkludo

XML montriĝas aparte valora por krei vortarojn, terminarojn, lekskikonojn, sed ankaŭ por ekspluati XML-ajn leksikografiajn datumbazojn. Al XML estas asociitaj multaj valoraj senpagaj iloj, kiel: TTT-legiloj, sintaksaj kontroliloj kaj transformiloj. La projekto REVO, bazita sur XML, montriĝis ideala sistemo por leksikografia kunlaborado kaj samtempe ĝi estas la ĝis nun plej bona tereno por evoluigi ĉiun seriozan terminologian projekton.

---

9. Rimarkindas, ke, krom en la reto, eblas ĝin eldoni ankaŭ paperforme aŭ por poŝkomputiloj. La paperforma eldono restas ĝis nun teoria ebleco, sed jam pretas XSLT-skripto, kiu kapablas transformi la artikolojn al TeX-formato.

10. `http://www.yahoogroups.com/group/revuloj`.

11. `http://purl.org/net/voko/revo`.