Från metadata till uppmärkning av komplexa dokument: Ett ramverk för semantisk dokumentproduktion

Vinnova projekt P24478-1 A

Det finns stora mängder information tillgängligt och sökbart på elektroniska format. Tyvärr är det svårt att hitta rätt information när sökfrågan är mer komplex än den enkla matchning av ord som dagens sökmotorer erbjuder. Detta gäller speciellt om man söker efter dokument som innehåller mycket kunskap och samtidigt har en komplex struktur.

Det finns en stor mängd elektroniska dokument som oftast innehåller betydligt mer information än vanliga webbsidor. Dessvärre kan inte tekniker för semantisk uppmärkning som Ontology Web Language (OWL) användas för dessa dokumentformat. Exempelvis är dokument i MS Word och PDF formaten ofta betydligt mer genomtänkta, bearbetade och innehållsrika än webbsidor, men dessa format kan inte idag kodas med semantisk information, och det är därför inte möjligt att göra avancerade sökningar bland dessa typer av dokument.

Målet med detta projekt är att utveckla en infrastruktur som stöder systematiskt författande och uppmärkning av komplexa elektroniska dokument, speciellt i väletablerade format som MS Word och PDF. Dessa dokument skall, förutom text, figurer och tabeller, även innehålla semantisk information som beskriver innehållet och gör det möjligt för nya typer av sökmotorer att hitta relevant information.

Inom projektet arbetar vi med att ta fram ett ramverk för att hantera semantisk information som stödjer hela kedjan från fakta- och kunskapsinsamling över bearbetning och textförfattande till dokumentpublikation. Hittills har projektet tagit fram ett verktyg för att semantisk uppmärkning av PDF-dokument. Detta verktyg är baserat på Protégé, som är ett ledande verktyg för att definiera och redigera ontologier för bl a den semantiska webben, och Adobe Acrobat Professional, den mest avancerade programvaran för hantering av PDF-dokument. Vårt verktyg gör det möjligt att dels märka PDF-dokument med semantisk information lagrat i OWL formatet på liknande sätt som webbsidor, och dels att koppla text, grafik och områden i dokument till ontologier som beskriver dokumentets terminologi och semantiska innehåll. Denna uppmärkning gör det möjligt att göra avancerade sökningar i stora mänger PDF-dokument.

Projektet utgår från applikationsområdet statistikrapportering, som utnyttjar data och metadata systematiskt för att producera dokument. Statistikrapporter är dokument som innehåller ett stort antal tabeller och diagram, och det är en stor fördel om rapporterna kan göras sökbara genom semantisk märkning av innehållet. Projektet genomförs i samarbete med Statistiska centralbyrån.

Kontaktpersoner:
Prof Henrik Eriksson, Linköpings universitet, tel 013-28 26 73, e-mail her@ida.liu.se
Tekn dr Magnus Bång, Linköpings universitet, tel 013-28 44 43, e-mail magba@ida.liu.se