KAYNAK KEŞİF YETENEĞİNİN ARTIRILMASI İÇİN INTERNET KAYNAKLARININ İÇERİKLERİNİN STANDART BİÇİMDE TANIMLANMASI

Baha OLGUN* - Hayri SEVER**

Öz

Internet kaynaklarının makinaca anlaşılabilir olmamasından dolayı, kullanıcıların bilgi ihtiyacını karşılamada sorunlar yaşanmaktadır. Kaynakların yapısal olarak gösterilmemesi ve nasıl yorumlanması gerektiğinin ilgili kaynağa özgün çözümlerle halledilmeye çalışılması, ilk göze çarpan nedenleri oluşturmaktadır. Internet kaynaklarının gerek sayısal ve gerekse de hacimsel olarak çok hızlı artışı göz önünde bulundurulduğunda, içerik sözcüklere dayalı arama makinalarına ilaveten, otomatik kaynak keşfine ve bilginin harmanlanmasına olanak veren yazılım araçlarının gerekliliği ortaya çıkmaktadır. Bu tür yazılım araçlarının başarısı ise, işlenecek kaynakların modellenmesinin standart bir biçimde yapılmasına çok yakından bağlıdır. RDF (Resource Description Framework), böyle bir çabadan doğan anlamsal bir modellemedir ve bu model üzerinde yapılan çalışmalar WWW (World Wide Web) Konsorsiyumu tarafından kontrol edilmektedir. DC (Dublin Core) üstveri elemanları, elektronik katalog bilgilerini tutmak için, RDF'in genişletilebilirlik özelliği kullanılarak tanımlanmıştır. Bu makalede, RDF/DC modeli kullanılarak Türkçe elektronik kaynakların içeriklerinin tanımlanmasını sağlayan H-DCEdit adlı editör aracı tanıtılacaktır. RDF modelinin serileştirme sözdizimi olarak SGML (Standard Generalized Markup Language) kullanılmıştır. Bu çalışmaya ek olarak, RDF/DC belgelerinin, DSSSL (Document Style Semantics and Specification Language) standardı yardımıyla farklı belge biçimlerinde yeniden biçimlenmesi de sağlanmıştır.

***Anahtar Sözcükler**: Internet kaynak keşfi, Web madenleme, RDF, DC, SGML, DSSSL, editör.*

ABSTRACT

Since Internet resources are not yet machine-understandable, there are some problems with satisfaction of information needs of users. Unstructural representation of resources and employing ad-hoc solutions for

* *Arş. Gör., Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü (baha@hacettepe.edu.tr)*
** *Doç. Dr., Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü (sever@hacettepe.edu.tr)*

the issue of how to interpret these resources constitute the reasons that an eye would easily catch at first sight. Given that a drastic and staedy increase in the number as well as in the volume of Internet resources has been observed over years, in addition to search engines based on content terms, the necessity of software agents being capable of automatically discovering and harvesting these resources has arisen. Success of this type of software agents very closely depends on standartization of modeling of resources to be processed. A semantic modeling defined as a result of such an effort is called RDF (Resource Description Framework), and the studies on this model is controlled by WWW (World-Wide Web) Consortium. The DC (Dublin Core) metadata elements have been defined using the property of extensibility of RDF to handle electronic catolog information. In this article, an authoring editor, called H-DCEdit, is introduced. This editor makes use of RDF/DC model to define contents of Turkish electronic resources. To serialize (or to code) a RDF model, SGML (Standard Generalized Markup Language) has been used. In addition to this work, by using DSSSL (Document Style Semantics and Specification Language) standard, the format information in regard to a given RDF/DC document has been separated from its content, and hence H-DCEdit provides different views of an RDF/DC document.

Keywords: *Discovery of Internet resources, RDF, DC, SGML, DSSSL, Web mining, authoring editor.*

1. GİRİŞ

Internetin günümüzde ilgiyle karşılanmasının bir nedeni, büyük miktarlarda bilgi içermesidir. Internet üzerindeki her şey makinaca-okunur özelliğe sahip olmasına rağmen, makinaca-anlaşılır özelliğe sahip değildir (Lassila 1998). Bu durum Internet üzerindeki verinin kullanıldığı işlerin otomatik olarak yapılmasını engellemektedir. Büyük miktarlardaki Internet verilerini klasik yöntemlerle işlemekse bilgi miktarının büyüklüğü açısından imkansızdır.

Web[1] tabanlı bir yazılım aracının (software agent) bilgiyi otomatik olarak tanıma, yorumlama, keşfetme yeteneği HTML (HyperText Markup Language) diline özgü bir takım karakteristik kısıtlamalarla sınırlı kalmaktadır. Örneğin, çeşitli sitelerdeki elektronik satış kataloglarını tarayarak belirli bir ürün bazında müşteriye en uygun fiyatı sunmaya çalışan bir yazılım aracının, belirli bir katalogda

1 *Web ve Internet kelimeleri, makale boyunca eş anlamlı olarak kullanılmışlardır.*

ürün adının "<l>ad:</l>" ile nitelendirildiğini öğrenmesi, başka bir katalogdaki ad nitelendirmesi konusunda ona pek bir ipucu vermez. Bu "adı" olabileceği gibi "isim:" de olabilir (Doorenbas, Etzioni ve Weld 1996). Fiyat bilgisi birimi katalogdan kataloğa değişebilir. Benzer şekilde, bir elektronik (daha özel olarak Web) katalogda ürünler arası ilişkiler hiper bağlaçlar (hyper links) aracılığı ile kurulur. Verilen bir katalog bazında, hangi ilişkinin (benzer, opsiyon, genel, özel, parça, vb.) hangi hiper bağlaç tarafından sağlandığının öğrenilmesi, -ki bu öğrenme süreci zaman açısından oldukça pahalı ve elle (manuel) gerçekleştirilen bir süreçtir- başka kataloglara da uygulanabilen bir kazanıma tekabül etmeyebilir. Ayrıca, HTML sayfasının içeriği doğal dil ile yazılmış saf metinlerden oluşur. Doğal dil ile yazılmış metinlerden, içerik ile ilgili bilgi çıkarmak ise pratik değildir (Etzioni 1996; Manber ve Bigot 1998). Bundan dolayı, metin üzerinde yapılan kelime arama yönteminden, daha farklı ve daha etkin bir yöntem arayışı başlamıştır. Buna bağlı olarak, arama işlemlerinde ihtiyaç duyulan bilgilerin oluşturulması gerekliliği de ön plana çıkmıştır. Üzerinde arama yapılan belgenin konusu, yazarı ya da yayım tarihi gibi bilgilerin anlaşılması gerekmektedir. Bu bilgilerin doğal dil ile yazılmış belgenin metin içeriğinden çıkarılması da oldukça güçtür, hatta imkansızdır. Öyleyse bu bilgiler iyi tanımlanmış olmalı ve makinaca-anlaşılır bir biçimde tasarlanmalıdır. Sorgu sonuçlarının iyileştirilebilmesi, daha akıllı sorguların yapılabilmesi, kaynakların özelliklerinin tanımlanması ve standart bir yöntemin varlığı ile sağlanabilir.

Internet üzerindeki kaynaklar, saklanma biçimlerine göre, yapısal olmayan, yarı yapısal, ya da tam yapısal verilerden oluşur. Veri biçimlerindeki çeşitliliğe paralel olarak, bu verileri işleyen yazılım araçları mimarileri de farklılık göstermektedir. Yapısal olmayan veriler için **arama makinaları** (search engines) yaygın seçenek olarak yer almasına karşın, yapısal olmayan ya da yarı yapısal veriler için **meta-arayıcılar**[2] (metasearchers) ve yapısal veriler için **arabulucular** (mediators) kullanılmaktadır (Gravano ve Papakonstantinou 1998).

2 *Bir meta-arayıcı, yapısal olmayan Internet kaynakları üzerinden kullanıcının bilgi ihtiyacını karşılamada, arama makinalarını kullanırlar. Meta-arayıcıların bu bağlamdaki ana işlevi, taban aldıkları arama makinalarının veritabanlarını kullanıcıya şeffaf hale getirmektir.*

Aşağıda, meta-arayıcılarda ve arabulucularda ortak (ya da işlevsel olarak benzer) işlemler açıklanmaktadır.

Sistem yapısında her bir dağıtık veri kaynağıyla ilişkili bir **paketleyici** (wrapper) bulunur. Veri kaynakları, her biri farklı biçimlerde verileri bulunan ve farklı sorgu girdi ve çıktılarına sahip bağımsız arama makinaları olabilir. Paketleyicinin görevi, veri kaynakları için ortak bir veri modeli görünüşü sağlamaktır. Paketleyici, aynı zamanda, bir ortak sorgu arayüzü kullanma olanağı da sağlar. Meta-arayıcı ya da arabulucudan gelen sorguyu ilişkili olduğu kaynağın özel sorgu diline çevirir. Kaynak tarafından iletilen sorgu sonuçlarını da ortak veri modeline dönüştürür. Bu sayede arayüz şeffaflığı sağlanır.

Meta-arayıcıların ve arabulucuların sistem işlevleri de üç ana işlemle ifade edilebilir: **Veri tabanı seçimi, sorgu dönüştürümü** ve **sonuçların birleştirilmesi** (Gravano ve Papakonstantinou, 1998). Veri tabanı seçiminde sorgu ile ilgili verinin bulunduğu veri tabanı belirlenir. Sorgu dönüştürümü adımında, sorgu, seçilen her bir veritabanı üzerindeki paketleyicinin anlayabileceği bir biçime dönüştürülür. İlgili veritabanı paketleyicisi sunulan sorguyu işletir ve sonuçları döndürür. Sonuçların birleştirilmesi adımında ise, döndürülen bu sorgu sonuçları belirli bir takım işlemden (tekrarlı bilgilerin ayıklanması, farklı sıralama algoritmalarınca sıralanmış belgelerin yeniden sıralanmaya tabi turulması, farklı kodlanmış aynı tür bilgilerin tek bir kodlamayla ifade edilmesi[3], vb.) geçirildikten sonra birleştirilir (Gauch ve Wang 1996).

Meta-arayıcılar, prensip olarak yapısal olmayan metinler (düz ya da serbestçe hazırlanmış metinler) ya da yarı yapısal sayılan takı (tag) ile işaretlenmiş belgeler (HTML, SGML ya da XML) üzerinde işlem yapar ve dağıtık veri kaynakları üzerinden kullanıcılara tek bir küresel sorgu arayüzü sunarlar. Dağıtık ve tam yapısal veriler, örneğin ilişkisel veri tabanları, ilgili paketleyiciler tarafından modellenir ve bu yerel veri modelleri üzerinden arabulucu (mediators) katmanı/ aracı aracılığı ile tek bir küresel sorgu arayüzü, kullanıcıya sunulur (Manber ve

3 *Kodlamadaki (Encoding) farklılık, ilgili bilginin nasıl tutulduğu ile ilgilidir. Örneğin, sıcaklık bilgisi soğuk, ılık ve sıcak ile ifade edilebileceği gibi santigrat ile de gösterilebilir.*

Bigot 1998). Arabulucular aynı zamanda etiketlerinin ve düğümlerinin özel anlamları olan ve çizge ile ifade edilen yarı yapısal belgeler üzerinde de çalışırlar.

Yukarda yapılan tartışmalar çerçevesinde, arama araçlarının yeterince kesin sonuçlar döndürebilmesi ve heterojen Internet kaynaklarının tümleştirilmesi (integration) için gereksenen model, şu veya bu şekilde üstveri (veri hakkında veri) tanımlama ve sözcük haznesi kontrolu olanaklarını sağlamalı ve çeşitli uygulama alanlarına genişletilebilir olmalıdır, denilebilinir (Singh 1998). Bu konudaki en iddialı seçenek de, RDF[4] (Resource Description Framework) modelidir. Çünkü RDF karmaşık ilişkileri tanımlama yeteneğine sahiptir, kullanımı kolaydır, model üzerinde sözcük haznesi kontrolu mümkündür ve genişletilebilir esnek bir anlayışa sahiptir.

Üstveri çeşitli uygulamalara kaynak olmaktadır. Üstveri, Web kaynaklarının içeriklerinin tanımlanmasında yani kataloglamada kullanılabilir. Arama araçları tarafından daha kesin sonuçlar elde edebilmek amacıyla, katalog girişleri olarak kullanılabilir. Üstveri, elektronik ticaret alanında bilginin kodlanmasında da kullanılabilir. Bunların yanında, Web sayfalarına erişimin filtrelenmesi amacıyla içerik etiketi olarak kullanılabilir. İçerik seçimi ile ilgili çalışmalar büyük ölçüde Web üstveri kavramına dayandırılmaktadır. Elektronik ortamda, elektronik imza olarak kullanılması da üstveri uygulama alanlarındandır. Ayrıca üstveri, kişilerin ya da Web sitelerinin gizlilik özelliklerinin de tanımlanmasında kullanılabilir. En yaygın olarak ise; üstveri, elektronik belgelerin içerik bilgilerinin ve entellektüel özelliklerinin tanımlanmasında kullanılır. Anlaşılacağı gibi, Web kaynaklarının da içinde olduğu elektronik kaynakların ve tüm elektronik ortam nesnelerinin (Web sayfaları, Web siteleri, kullanıcılar, vs.) kesin ve açık tanımlarının yapılması çok önemlidir (Lassila 1998; World... 1998).

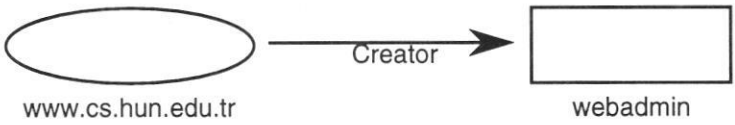
4 *RDF (Resource Description Framework), üstveri işlenmesiyle ilgili bir modeldir ve Web üzerindeki makinaca-anlaşılır bilginin uygulamalar arasında işlenebilirliğini sağlar. RDF, Web kaynaklarının otomatik işlenmesini sağlayan olanakları içerir. World Wide Web Konsorsiyumu, Internet üstveri için yeni bir standart olarak RDF modelini sunmuştur. RDF tasarımı sırasında çeşitli disiplinler üstveri gösterimi ve taşınmasıyla ilgili konularda katkıda bulunmuşlardır. HTML (Hypertext Markup Language) üstveri, PICS (Platform for Internet Content Selection), SGML (Standard Generalized Markup Language) ve XML (Extensible Markup Language) standartlarından yararlanılmıştır. Bunlar dışında, nesneye-yönelik programlama kavramlarından ve veri tabanı kavramlarından da yararlanılmıştır.*

RDF modelinin amacı, Web kaynaklarına ait üstverinin tanımlanması, ilişkilendirilmesi ve standartlaştırılmasıdır. RDF ile tanımlanacak kaynaklar URI (Uniform Resource Indicator) ile isimlendirilebilecek tüm kaynaklardır. Fakat kaynak ile ilgili herhangi bir kısıtlama getirilmemiştir. RDF uygulamadan, sözdizimden ve ortamdan bağımsız bir modeldir (World... 1999b).

RDF modeli, yönlü etiketli çizge ile gösterilebilir. RDF verisi, düğümler ve düğüme iliştirilmiş öznitelik/değer çiftleri ile ifade edilebilir. Düğümler herhangi bir Web kaynağı olabilir. Öznitelikler düğümlerin isimlendirilmiş özellikleridir ve çizgede Web kaynağını gösteren düğümden çıkan yay olarak gösterilirler. Öznitelik değerleri ise ya atomiktir (karakter dizgi) ya da başka düğümdür. Öznitelik değerleri de çizgede yayın diğer ucundaki düğüm olarak görünürler. Yani RDF modelinin temeli **düğüm**lere, **özellik**lere ve **değer**lere dayanır. Sözgelimi, A'nın bir B özelliği varsa ve B'nin değeri C ise; A ile C bir çizgenin düğümleri olurlar. B ise aralarındaki yayın etiketidir. Yayın yönü ise A'dan C'ye doğrudur.

RDF modelinin tanımlarını bir kütükte saklamak ve uygulamalar arasında kullanmak için, çizge gösterimini serileştirme dili olarak XML (Extensible Markup Language) kullanılmaktadır. RDF ile XML birbirlerini tamamlamaktadır (Lassila 1998). XML, RDF modeline serileşme sözdizimi olarak destek vermektedir. RDF ise XML dilini güçlendirmiştir ve XML, üstveri tanımlarını anlamlı hale getirmek için RDF tanımlarına ihtiyaç duymaktadır. Her ne kadar XML, RDF gösterimi için seçilmiş olsa da, diğer sözdizim biçimleri de kullanılabilir ve bu konuda kesin bir kısıtlama yoktur.

RDF Tanımının Çizge ile Gösterimi



RDF Tanımının XML İle Gösterimi

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/RDF"
    xmlns:dc="http://purl.org/DublinCore">
<rdf:Description about="http://www.cs.hun.edu.tr">
<rdf:Description>
<rdf:RDF>
```

RDF, üstveri kodlaması için önceden tanımlı bir sözlük içermez. Çeşitli amaçlarla geliştirilmiş sözlüklerden destek alır. Bir sözlük, kaynak tanımlama komiteleri tarafından tanımlanmış üstveri elemanları ve öznitelikler kümesidir. **Dublin Core**[5] **Initiative** (Dublin... 1998) uluslararası bir çalışma komitesidir ve RDF modelinin üzerine kütüphane katalogları için gerekli üstveri elemanlarını eklemek üzerine yoğunlaşmıştır. Bu çalışmanın sonucu DC (Dublin Core) üstveri elemanları kümesi olarak bilinir. DC, RDF modeli üzerine inşa edildiği için, kısaca RDF/DC olarak da adlandırılır.

Var olan üstveri elemanları kümelerinin farklılığı; Web üzerinde kullanılacak, birbirini tamamlayacak ama birbirinden de bağımsız ele alınabilecek üstveri elemanları kümelerinin tasarlanmasını gerektirmektedir. WWW Konsorsiyumu (kısaca **W3C** olarak da bilinir) tarafından geliştirilen RDF, genişletilebilme özelliği sayesinde temel yapı taşı görevi yapıp, bir çok üstveri kümesini destekleyebilmektedir. DC, bu amaç doğrultusunda yapılan bir çalışmadır ve RDF modeline sayısal kütüphaneler açısından destek vermektedir.

DC elektronik kaynaklar için basit bir içerik tanımlama modeli sunar. Anlamları üzerinden uluslararası çalışma grupları ve uzmanlar tarafından fikir birliği

5 *Dublin Core elektronik kaynak keşfinde, kullanılan üstveri elemanları kümesini ifade eder. Dublin Core girişimi, Internet'in gelişen altyapısında önemli bir yer tutmaya başlamıştır. Dublin Core birçok komite tarafından, kaynak tanımlama konusunda anlamsal bir ortak yaklaşım geliştirme açısından desteklenmektedir. Bu sayede, yapılan çalışmalar uluslararası bir geçerliliğe ulaşmıştır. Dublin Core standardı on beş eleman içerir. Elemanların genel anlamları üzerinde uluslararası fikir birliği oluşmuştur.*

oluşmuş on beş üstveri elemanından oluşur (Dublin... 1998). DC üstveri elemanları şunlardır : Başlık *(Title)*, Yaratıcı *(Creator)*, Konu *(Subject)*, Tanımlama *(Description)*, Yayımcı *(Publisher)*, Orijinal Kaynak *(Source)*, Katkı Yapan *(Contributor)*, Tarih *(Date)*, Tür *(Type)*, Biçim *(Format)*, Belirleyici *(Identifier)*, İlişki *(Relation)*, Dil *(Language)*, Kapsam *(Coverage)* ve Haklar *(Rights)*. Bu elemanların bazılarının anlamı açıkça tanımlanmıştır, bazıları ise henüz deneyseldir. Tanımları açık olan elemanların da kullanım biçimleri konusunda bir sınırlama yoktur. DC üstveri elemanlarının uluslararasılaşması için de çalışmalar sürmektedir. Böylece üstveri elemanlarının kullanımının ulusal niteliklerden bağımsız hale getirilmesi sağlanmaya çalışılmaktadır.

DC üstveri elemanlarının avantajlarının başında kullanımındaki basitlik gelir. Bunun yanında, elemanların anlamlarının da aynı şekilde yorumlanması ve arama işlemlerinin aynı sonuçları üretilmesi sağlanmaktadır. Bu özelliği, üstveri elemanları üzerindeki uluslararası anlaşma ve katkılar da desteklenmektedir. Ayrıca elemanların genişletilebilmesi olanağı sayesinde, özel uygulamalarda DC elemanlarına ek yapılabilmesi de mümkündür.

Internet artık sadece tanıtım sayfalarının sergilendiği ve kullanıcıların bu sayfalar üzerinde gezindiği sanal bir ortam olarak düşünülemez. Internet büyük miktarlarda bilginin saklandığı heterojen veri tabanları kolleksiyonu olarak düşünülmelidir (Singh 1998). Kullanıcılara, Internet bilgileri üzerinde sorgulama olanağı sağlamak da önemli bir çalışma konusu olmuştur. Klasik anlamda sorgu işlemleri, okunabilir Web kaynakları üzerinde yapılabilen işlemlerdir ve kaynak içeriği içinde kelime arama yöntemini temel almıştır. Bu yöntem her zaman doğru sonuçlar oluşturmamakta, sorgular çoğunlukla istenilen dışında sonuçlar üretmektedir. Yeterince kesin sonuçlar üretilememesinin en önemli nedenleri, kaynakların tanımlanmasıyla ilgili bir yöntemin eksikliği ve kaynak hakkında bilgiye ulaşmak için kaynağın kendisine ulaşılması zorunluğudur. Üstveri kavramı, klasik anlamda, Web kaynaklarının sorgulanması yöntemlerinin değişmesini sağlamıştır. Anahtar kelime arama yöntemi yerine, daha üst düzeyli bir kavramsal model üzerinde inceleme yapma yöntemi gelmiştir.

Üstveri üzerindeki ilk çalışmalar WWW Konsorsiyumu'nun **PICS** *(Platform for Internet Content Selection)* adlı çalışması ile başlar. PICS (World...1999c) bir sunucudan istemcilere, Web sayfalarıyla ilgili bazı değerlendirmeler iletmek için kullanılan bir yöntemdir. Bu değerlendirmeler, Web sayfası içerik bilgileridir. Örneğin, bir sayfanın yetkili bir araştırmacı tarafından yaratılıp yaratılmadığı; içeriğinde şiddet, cinsellik ya da bozuk dil yer alıp almadığı gibi bilgiler değerlendirme ölçütlerdir. PICS, sabit ölçütler kullanmak yerine, bir değerlendirme sisteminin yaratılması için standart bir yöntem sunar. Böylece kullanıcılar Web sayfalarının kendi ölçütlerine uymayanlarını filtreleme olanağı bulurlar. Özellikle aileler çocuklarının Internet kullanımında belirli ölçütler koyarak, onların kötü etkilenmelerini engelleyebilirler. PICS sınırlı bir kullanım olanağına sahiptir ve daha genel amaçlı içerik tanımlamasında kullanılamaz. Örneğin sayısal kütüphanelerin ve Internet kaynaklarının içerik tanımlayıcı bilgilerinin oluşturulması gereksinimlerini karşılayacak işlevsel yeterliliğe sahip değildir. Bu nedenle WWW Konsorsiyumu, Web için daha genel üstveri tanımlama modeli olarak RDF çalışmasını başlatmıştır. RDF kısaca Web kaynaklarının işlenmesinin otomatikleştirilmesini sağlayan olanakları ifade eder.

Benzeri bir üstveri çalışması da **KIF** *(Knowledge Interchange Format)* dilidir. Bir gösterim dili olarak KIF'in çok iyi tanımlanmış bir sözdizimi ve semantiği vardır (Singh 1998). Aynı zamanda işleç önde tarzında bir ifade dilini (prefix version of first-order predicate calculus) kullanır ve KIF'le karmaşık ifadeler tanımlamak mümkündür. KIF sadece üstveri tanımlanmasında kullanıldığından, KIF cümleleri ile ilişkili eylemi ifade etmek mümkün değildir. Bu nedenle, KIF dilinin üzerine bir iletişim katmanı görevi yapacak ve KIF cümleleri için istem türünü tanımlamaya yarayacak başkaca dil özelliklerini sağlayan KQML (Knowledge Query and Manipulation Language) kullanılır (Singh 1998). Her KQML iletisi, bir işlem türü ve ilgili KIF ifadelerini içerir. KQML ve KIF kullanılması, anlamsal açıdan kuvvetli üstveri tanımlama ve etkin işleme olanağı sunar. Ancak ifade yeteneği gelişkin olan bu katmanlı modelin uygulamaya konması kolay değildir. Bu nedenle RDF, basitliğiyle bu modele göre önemli bir avantaj sağlar. Ayrıca RDF modelinin anlaşılması ve ifade edilmesi daha kolaydır.

2. INTERNET KAYNAK İÇERİKLERİNİ TANIMLAYAN YAZILIM : H-DCEDİT

Geliştirilen yazılımın amacı, RDF modeli ve DC üstveri elemanları kullanılarak, elektronik kaynakların içeriklerinin tanımlanmasını sağlayan bir yazılımın gerçekleştirilmesidir. Yazılımda DC üstveri elemanları RDF modelinde bir sözlük olarak kullanılmıştır. RDF modeline yazılımda sade biçimde yer yerilmiştir ve temel RDF kuralları kullanılmıştır. Yazılımın ürünü ise RDF/DC modelini temel alan katalog bilgileridir. Katalog yapısı, RDF/DC sözdizimi temel alınarak SGML[6] (International... 1986) tarafından tanımlanmıştır. Yapılan SGML belge tür tanımı, bu tür tanımına uyan SGML belgelerinde, Türkçe karakterleri desteklemektedir ve RDF/DC modelinde tanımlanan belgelerin serileştirme dili olan XML'e (World... 1998) dönüştürülmelerinde bir takım kolaylıklar sağlamaktadır. Sistem işlev çizgesi, alt kesimleriyle birlikte Şekil 1'de gösterilmiştir. Burada koyu gösterilen kesimler, dışarıdan sisteme eklenen kesimlerdir.

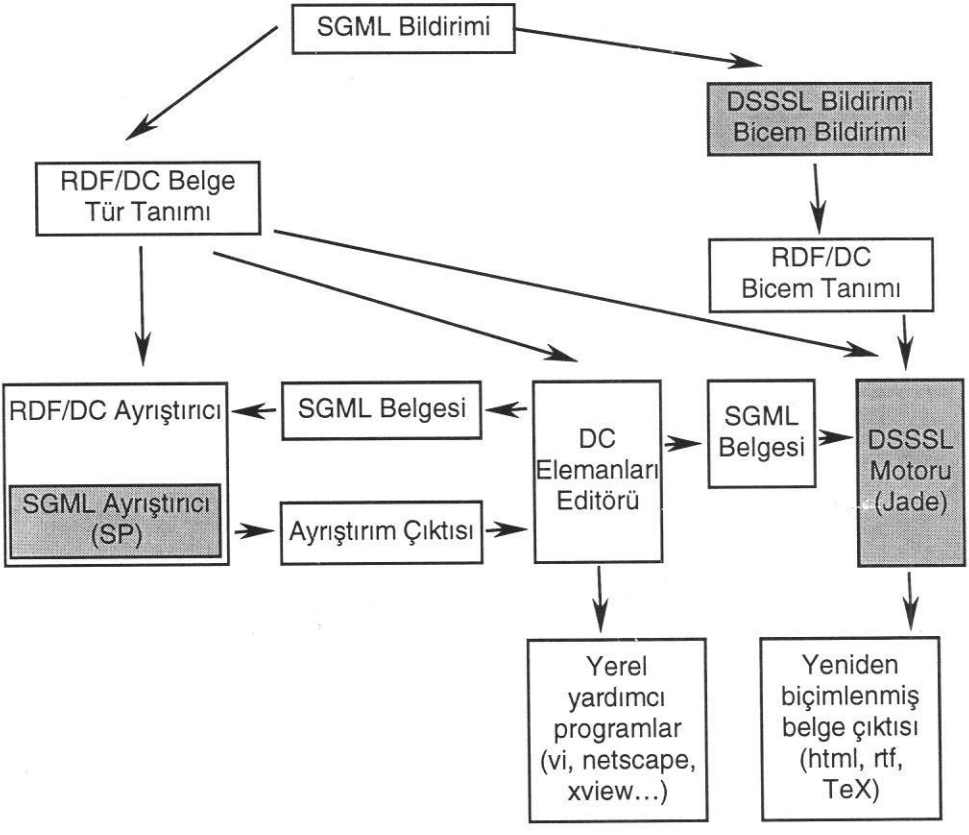
Şekil 1'de görüldüğü gibi sistem SGML temelli bir sistem olduğundan, işlevler SGML bildirimine bağlıdır. SGML bildirimine dayalı belge tür ve biçem tanımları ise, ilişkili oldukları alt kesimlerin işleyişlerini denetlemektedirler. Sistem tüm alt kesimleriyle anlamlı bir işlevi gerçekleştirmesinin yanında, ayrı olarak da ça-

---

6 *SGML (Standart Generalized Markup Language) Nedir?*
Kelime işlemcilerin yaygınlaşması, oluşturulan belgelerin elektronik ortamda bir kullanıcıdan diğerine taşınmasını önemli hale getirmiştir. Her ne kadar, belgeleri makinadan makinaya aktarma için çeşitli yöntemler mevcutsa da, bazı durumlarda bu belgelerin görünüşleri ile ilgili ayrıntıları taşınamamaktadır. Bunun nedeniyse, hazırlanan belgelerin makinaya ait özellikleri de içermesidir. Bazen içerik de aktarılamamaktadır, çünkü hazırlanan belgenin içeriği doğrudan görünüşle ilgili ayrıntılara bağımlıdır. Genel anlamda sorun, aktarılan belgelerin yeniden yapılandırılmasını sağlayacak yapı bilgisinin eksikliğidir.
*Bir metin içindeki kesimlerin rollerini belirleyen, **yapı** bilgisidir. Bir belge içindeki başlıklar, paragraflar ya da listeler yapı bilgisiyle belirlenir. Belgenin yapısı, o belgenin elemanlarının nasıl görüntülendiğiyle ilgili olmaktan çok, elemanların dizilişiyle ilgilidir. Bir belge içindeki başlık elemanının yazılı belge içindeki rolü ile, belgenin dökümündeki rolü aynıdır. Dökümdeki görünüş farklı olsa bile, bu başlık elemanının rolünü değiştirmez.*
Elektronik ortamda hazırlanmış belgelerin dünya üzerinde dağıtılmasıyla ilgili sorunların çözüldüğü varsayıldığında bile, başka bir sorun ortaya çıkacaktır. Belge yazarları belgelerini standart bir yapıya dayandırmak isterler. Fakat tek bir standartlaşmış belge yapısı kullanıcıların isteklerini karşılayamaz. Çünkü kullanıcılar bilgiyi her zaman aynı biçimde kullanmak istemeyebilirler. Sonuçta hem yazarların hem de kullanıcıların gereksinimlerini karşılayacak bir yönteme ihtiyaç duyulmaktadır. Bu yöntem, hem yazarlara esnek belge yapısı tasarlama olanağı sunmalı, hem de kullanıcılara belge içindeki bilgi üzerinde denetim olanağı sunmalıdır.
Bu tür sorunların üstesinden gelmek için, ISO (International Standards Organization) tarafından ISO 8879: Information processing-Text and office system- Standard Generalized Markup Language (SGML) standardı hazırlandı. SGML, yapısal bilgi içeren belge oluşturabilmek ve kullanıcıların istedikleri içeriği belgenin içine gömerek yaratabilmeleri için tasarlanmıştır.

lıştırılabilir. Sistemin alt kesimleri mümkün olduğu kadar birbirlerine sıkı bağlı olmayacak biçimde tasarlanarak, her bir alt kesimin yeniden kullanılabilirliği amaçlanmıştır. Ancak editör kesimi sistemde, diğer alt kesimlerden destek alan kesim olduğundan bu tanımlamanın dışındadır. Çizelge 1'de yazılım alt kesimlerinin satır sayısı bilgileri verilmiştir.

Şekil 1. Sistem modeli



Yazılım Alt Kesimi	Programlama Dili	Satır Sayısı
SP-1.2 (SGML Ayrıştırıcı)	C++	70.000
RDF/DC Ayrıştırıcı	C++	800
H-DCEdit 1.0	C(MOTIF)	5.000
JADE1.0.1 (DSSSL Motoru)	C++	52.000

Çizelge 1. Yazılım alt kesimleri satır sayıları

2.1. SGML Bildiriminin Oluşturulması

Yazılım içinde kullanılan SGML bildirimi, referans somut sözdizimi tanımına uyumlu olarak tasarlanmıştır. Bu tanıma eklemeler yapılarak SGML işleme sisteminin Türkçe içeriği de desteklemesi sağlanmıştır. Ayrıca SGML belgelerinde bazı XML uzantılarının da kullanılabilmesi için SGML bildirimine bazı ekler yapılmıştır (sistem içerisinde kullanılan SGML bildirimi için bkz. Ek 1).

SGML bildiriminde CHARSET bölümünde sekiz ikillik (bit) belge karakter kümesi tanımlanmıştır. Belge karakter kümesi iki temel karakter kümesine dayandırılmıştır : ISO 646 ve ECMA-128 (ISO Latin 5). Sekiz ikillik karakter kodlarını simgeleyecek karakter kümesinin sol tarafı için (0-127) ISO646, sağ tarafı için ise (128-255) ECMA-128 kullanılmıştır. Bu sayede SGML belgeleri için tanımlanan karakter kümesi Türkçe desteği verebilecek duruma gelmiş olur. Aşağıdaki SGML bildirimi kesimi SGML belgeleri için karakter kümesi tanımı yapmaktadır.

```
BASESET "ISO 646-1983//CHARSET International Reference Version
        (IRV)//ESC 2/5 4/0"
DESCSET          0       9       UNUSED
                 9       2       9
                 11      2       UNUSED
                 13      1       13
                 14      18      UNUSED
                 32      95      32
                 127     1       UNUSED
BASESET  "ISO Registration Number 148//CHARSET ECMA-128
         Right Part of Latin Alphabet Nr. 5//ESC 2/13 4/13"
DESCSET  128     32      UNUSED
         160     95      32
         255     1       UNUSED
```

SGML belgeleri, XML ile uyumlu olabilecek biçimde tasarlanacağından, SGML bildiriminde ilgili değişikliklerin yapılması gerekmektedir. RDF modelinde şema (World... 1998) ve sözlük kullanımında XML isim uzayları (namespaces) kullanılmaktadır. Bu kullanım RDF modelinin içinde nesneye yönelik yaklaşımları barındırmasının bir sonucudur. XML isimuzayı, eleman adının isimuzayıyla nitelenmesi biçiminde kullanılır. Bu sayede, ayrı şema tanımları içindeki elemanlar belirlenebilmekte ve anlamsal bütünlük sağlanmaktadır. "Şema : Eleman" biçiminde bir yazımda Şema isimuzayını, Eleman ise şema içindeki elemanı göstermektedir. ":"ayıracı ise isimuzayı ile elemanı ayırmak için kullanılır. Fakat SGML referans somut sözdizimi içindeki SGML isimlendirme kurallarında, eleman adlarının belirtiminde (specifacation) ":" ayıracının kullanımı standart değildir. Bu nedenle SGML bildiriminde NAMING kesimine ":" ayıracının eklenmesi gerekmekterdir.

```
NAMING  LCNMSTRT      ""
        UCNMSTRT      ""
        LCNMCHAR      "-.:"
        UCNMCHAR      "-.:"
```

Yukarıda verilen satırlardaki belirtimler, isimlendirme kurallarını açıklamaktadır. NAMING kesimindeki belirtimler, referans somut sözdizimi üzerine eklenmek istenen isimlendirme kuralları ifade eder. LCNMCHAR kesiminde küçük harf olarak kullanılacak isimlendirme karakterleri tanımlanır. UCNMCHAR kesiminde ise büyük harf olarak kullanılacak isimlendirme karakterleri tanımlanır. Her iki kesime de "- . :" karakterleri eklenerek, bu karakterlerin isimlendirme karakterleri olarak algılanması sağlanır. Bu karakterlerden ":" karakterinin eklenmesi de XML isim uzayı benzeri bir kullanımı SGML belgelerinde kullanma olanağı sağlamıştır.

2.2. RDF/DC Belge Tür Tanımı

Yazılımların kullandığı belge tür tanımı hem temel RDF modelini hem de DC elemanlarını içerdiği için "RDF/DC" olarak adlandırılacaktır. RDF/DC belge tür tanımı, SGML bildirimiyle birlikte anlamlıdır. SGML bildirimindeki XML uyumu ve Türkçe desteği RDF/DC belge türü içinde de sağlanmıştır. Başka bir deyişle, XML uyumu ve Türkçe desteği belge tür tanımında tamamlanmıştır. Aşağıda RDF/DC belge tür tanımı kesimleri ve açıklamaları verilmiştir.

```
<!-- RDF and DC elements in the same DTD -->
<!-- In addition, XML like output is supported -->

<!--        RDF Elements -->

<!ELEMENT rdf:RDF - - ( rdf:Description )* >
<!ATTLIST rdf:RDF
          xmlns:rdf CDATA "http://www.w3.org/RDF/"
          xmlns:dc CDATA "http://purl.org/DC/"
>
```

Belge tür tanımının yukarıda görülen kesiminde, oluşturulacak SGML belgesi için sıradüzensel olarak en dıştaki rdf:RDF elemanı tanımlanmıştır. Bu SGML elemanı belgenin aynı zamanda RDF modeline göre tanımlandığını da göstermektedir. Yani oluşturulacak belge içeriği <rdf:RDF> ve </rdf:RDF> takıları (veya belirleyicileri) arasında oluşacaktır. SGML belgesinin içeriği ise RDF ile tanımlanan her bir kaynak tanımlamaları olacaktır. RDF tanımlamalarında, DC üstveri elemanları özellik (property) olarak kullanılacaklardır. Aynı zamanda XML ile uyumluluk desteğinin ilk parçası olarak, RDF çerçevesi kuran rdf:RDF elemanı içine XML isimuzayı tanımları eklenmiştir.

```
<!ENTITY % property "ANY">

<!ELEMENT rdf:Description - - %property;>

<!ATTLIST rdf:Description
      ID NMTOKEN #IMPLIED
      about CDATA #IMPLIED
      aboutEach CDATA #IMPLIED
      bagID NMTOKEN #IMPLIED
>
```

Yukarıdaki her bir RDF tanımlaması için varlık, eleman ve öznitelik tanımları verilmiştir. RDF tanımlamasını içerecek olan SGML elemanı rdf:Description elemanıdır. rdf:Description elemanının içerik modeli bir parametre varlık tanımıyla gösterilmiştir. Bu tanıma göre RDF kaynak tanımları belge tür tanımı içinde yer alan her eleman (ANY) olabilir. Bu biçimdeki bir gösterim belge tür tanımına sonradan yapılacak ek bazı tanımların kolay bütünleşebilmesini sağlamak amacıyla kullanılmıştır. Öznitelik tanımındaysa, tanımlanacak kaynaklar için belirleyicinin ifade edileceği öznitelik isimleri (ID, about, aboutEach, bagID) gösterilmiştir. Geliştirdiğimiz yazılım şu anda sadece ID özniteliğini desteklemektedir. Ayrıca XML standardının kaçındığı ama SGML içinde bulunan takı kısaltmalarının eleman ta-

nımlarından çıkarılması da XML diline yaklaşma kapsamında ele alınmıştır. Örneğin rdf:Description elemanınında takı minimizasyonu ihmal edilmiştir.

```
<!--            DC Elements          -->

<!ENTITY % dccontent "(#PCDATA)">

<!ELEMENT  DC:TITLE          --   %dccontent; >
<!ELEMENT  DC:CREATOR        --   %dccontent; >
<!ELEMENT  DC:SUBJECT        --   %dccontent; >
<!ELEMENT  DC:DESCRIPTION    --   %dccontent; >
<!ELEMENT  DC:PUBLISHER      --   %dccontent; >
<!ELEMENT  DC:CONTRIBUTOR    --   %dccontent; >
<!ELEMENT  DC:DATE           --   %dccontent; >
<!ELEMENT  DC:TYPE           --   %dccontent; >
<!ELEMENT  DC:FORMAT         --   %dccontent; >
<!ELEMENT  DC:IDENTIFIER     --   %dccontent; >
<!ELEMENT  DC:SOURCE         --   %dccontent; >
<!ELEMENT  DC:LANGUAGE       --   %dccontent; >
<!ELEMENT  DC:RELATION       --   %dccontent; >
<!ELEMENT  DC:COVERAGE       --   %dccontent; >
<!ELEMENT  DC:RIGHTS         --   %dccontent; >

<!ATTLIST DC:DATE
   year       CDATA #IMPLIED
   month      CDATA #IMPLIED
   day        CDATA #IMPLIED>

<!ATTLIST DC:RELATION
    type                          CDATA #IMPLIED
    resource   CDATA #IMPLIED >

<!-- Added Elements for Robots  -->
<!ELEMENT BODY - - %dccontent; >
<!ATTLIST BODY
     location  CDATA #IMPLIED >
```

Belge tür tanımının yukarıdaki kesiminde RDF kaynak tanımının ögeleri olan DC üstveri elemanları, birer SGML elemanı olarak tanımlanmıştır. Bir varlık tanımıyla DC üstveri elemanlarının içerik modeli, işlenebilir karakter verisi (#PCDATA) olarak belirlenmiştir. DC elemanlarından "DATE" ve "RELATION" elemanlarının içerikleri doğrudan yazılmamakta, bunun yerine öznitelik tanımlarıyla eleman değerleri belirlenmektedir. On beş adet DC elemanları dışında "BODY" adlı bir eleman daha tanımlanmıştır. Bu eleman, tanımlanan kaynağın yerel sistemdeki yerini göstermesi için dahil edilmiştir ve aslında kaynak üzerinde tarama yapacak Web robotlarına yönelik tasarlanmış üstveri elemanıdır. Bu amaçla elemanın öznitelik tanımında, yerel dosya adı bilgisinin tutulması amaçlanmıştır. RDF çerçevesi içinde, on beş DC elemanı ve bir de sonradan eklenen elemanla birlikte on altı özellik kullanılmaktadır. Başka bir deyimle, SGML belgesi şeklinde oluşturulacak katalog kaydı on altı özellik ile belirlenmektedir.

```
<!-- Entities For Turkish Support -->
<!ENTITY Ccedil    CDATA  "&#199;">
<!ENTITY ccedil    CDATA  "&#231;">
<!ENTITY Ouml      CDATA  "&#214;">
<!ENTITY ouml      CDATA  "&#246;">
<!ENTITY Scedil    CDATA  "&#222;">
<!ENTITY scedil    CDATA  "&#254;">
<!ENTITY Idot      CDATA  "&#221;">
<!ENTITY iwhdot    CDATA  "&#253;">
<!ENTITY Uuml      CDATA  "&#220;">
<!ENTITY uuml      CDATA  "&#252;">
<!ENTITY Gbrewe    CDATA  "&#208;">
<!ENTITY gbrewe    CDATA  "&#240;">
```

Yukarıdaki SGML varlık tanımlarıyla, oluşturulacak SGML katalog bilgileri içeriğinde Türkçe karakterlere karşılık varlık referanslarının kullanılması amaçlanmıştır. Türkçe desteği için ECMA-128 içinde bulunan ancak ISO 8859-1 içinde bulunmayan on iki karaktere karşılık olarak, varlık tanımları yapılmıştır. Varlık tanımlarının karşılıkları, SGML bildiriminin tanımlanan belge karakter kümesinde

de bulunan kodlardır. Yani Türkçe diline özel on iki karaktere karşılık referansları getirilecektir. Çizelge 2'de on iki Türkçe karakter ve ECMA-128 içinde tanımlı sayısal karşılıkları bulunmaktadır.

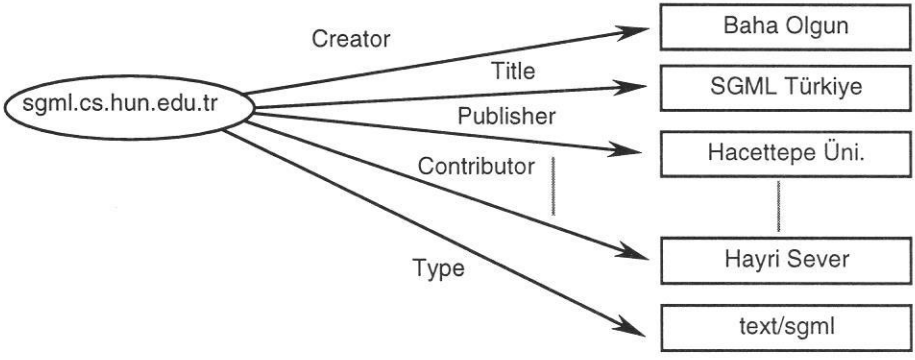
Bahsedilen XML ile uyumluluk özelliklerine rağmen oluşturulan belgeler birer SGML belgesidir. Tamamen XML diline dönüştürülmeleri için ayrıca bir ön işlemeye gerek duyulmaktadır. Ancak yapılan tanımlar bu işlemin yükünü hafifletecektir.

Karakter	Varlık	ECMA-128 sayısal kod karşılığı
Ç	Ç	199
ç	ç	231
Ö	Ö	214
ö	ö	246
Ş	Ş	222
ş	ş	254
İ	İ	221
ı	&iwhdot;	253
Ü	Ü	220
ü	ü	252
Ğ	&Gbrewe;	208
ğ	&gbrewe;	240

Çizelge 2. Türkçe karakterler için tanımlı varlıklar

2.3. Oluşturulan SGML Belgeleri

RDF/DC belge tür tanımına uyumlu SGML belgeleri, yinelenebilir rdf:Description elemanlarından oluşacaktır. Her bir rdf:Description elemanı bir katalog bilgisi girişini ifade eder. Şekil 2'de oluşturulan bir SGML belgesinin RDF tanımlarına uygun çizge gösterimi verilmiştir.



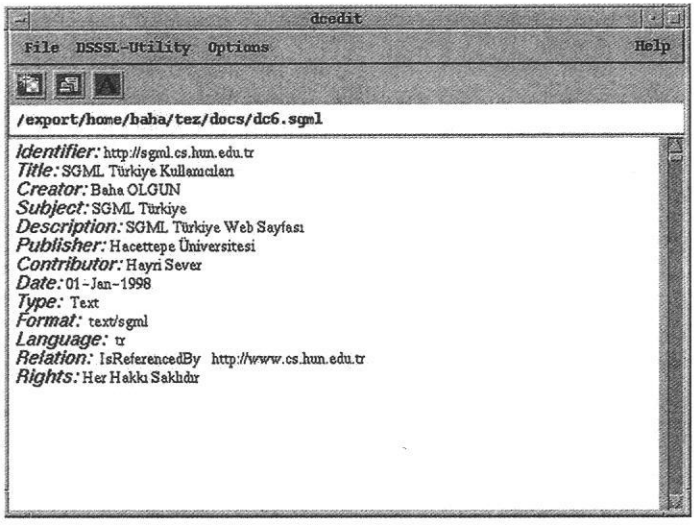
Şekil 2. SGML belgesinin çizge ile gösterimi

Şekil 2'de çizge gösterimi ile gösterilen kaynak tanımının SGML ile serileştirilmiş biçimi ise Ek 2'de gösterilmiştir.

SGML belgesinde XML isimuzayı kullanılmaktadır. DC:IDENTIFIER elemanında DC şeması içinde IDENTIFIER elemanı ifade edilmektedir. Ayrıca Türkçe içerik desteği de Türkçe'ye özel karakterlerin SGML varlıklarıyla gösterilmesi sağlanmıştır.

2.4. Hacettepe-DC Üstveri Elemanları Editörü : H-DCEdit

SGML bildirimine ve belge tür tanımına uygun SGML belgelerini oluşturmak için, UNIX ortamında X-MOTIF ile geliştirilmiş bir editör programı kullanılmaktadır. Şekil 3'de H-DCEdit programının arayüzü görülmektedir.



Şekil 3. H-DCEdit programı arayüzü

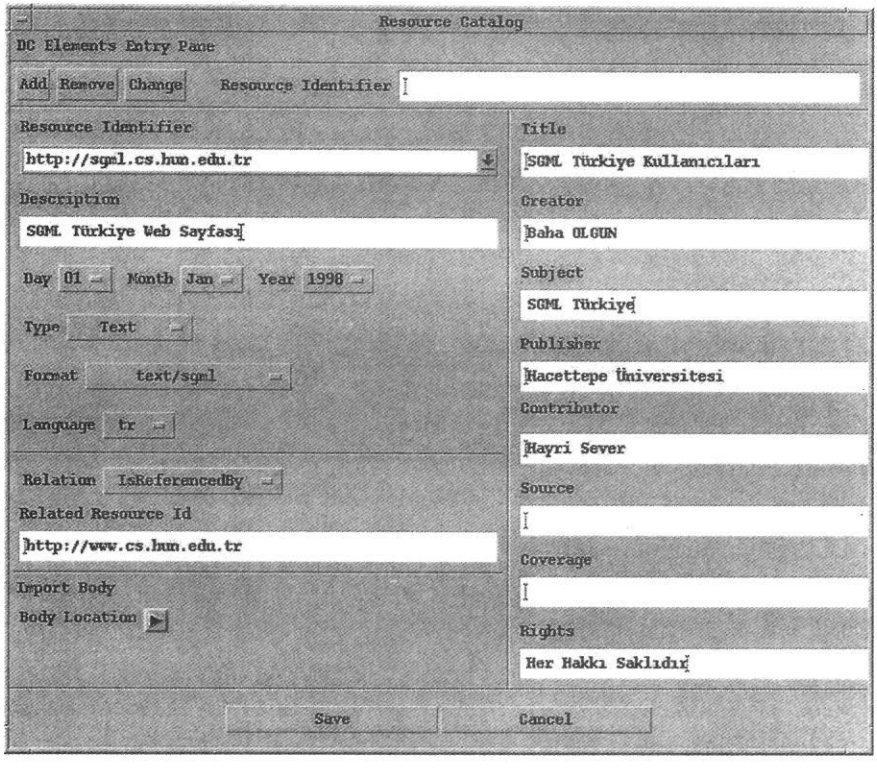
H-DCEdit programında DC elemanlarını RDF modeli içinde tutacak SGML belgeleri oluşturulur. DC elemanları girişi Şekil 4'de görülen arayüz aracılığıyla sağlanır.

DC elemanları giriş arabirimi ile istenilen sayıda kaynak tanımı yapılabilir. Ayrıca DC elemanlarına eklenen "BODY" elemanı için de giriş yapılabilir. Eğer bu eleman için değer girişi yapılmışsa yani kaynak için yerel işletim sistemi üzerinde yol adresi belirtilmişse; kaynağın görüntülenmesi için olanak sağlanmıştır. Görüntüleme işleminde yerel programlar, yardımcı olarak kullanılmıştır. İçeriği text/html biçimde tanımlanan bir kaynak için HTML destekli bir Web göstericisi kul-

lanılabilir. Aynı şekilde içeriği image/gif olan bir kaynak, gif biçimini destekleyen bir grafik gösterici program ile kullanılabilir. Kullanılacak yerel programlar, H-DCEdit programı içinde tanımlanmalıdır. Yardımcı uygulamaların tanımlanmasını sağlayan arayüz Şekil 5'de görülmektedir.

DC Elemanları Giriş Arayüzü aracılığı ile katalog bilgileri oluşturup saklandıktan sonra; ilgili katalog bilgileri, H-DCEdit programının arayüzü tarafından listelenir. Şekil 3'de, Ek 2'de verilen örnek bir SGML belgesi ile ilişkilendirilen bir liste görünümü yer almıştır.

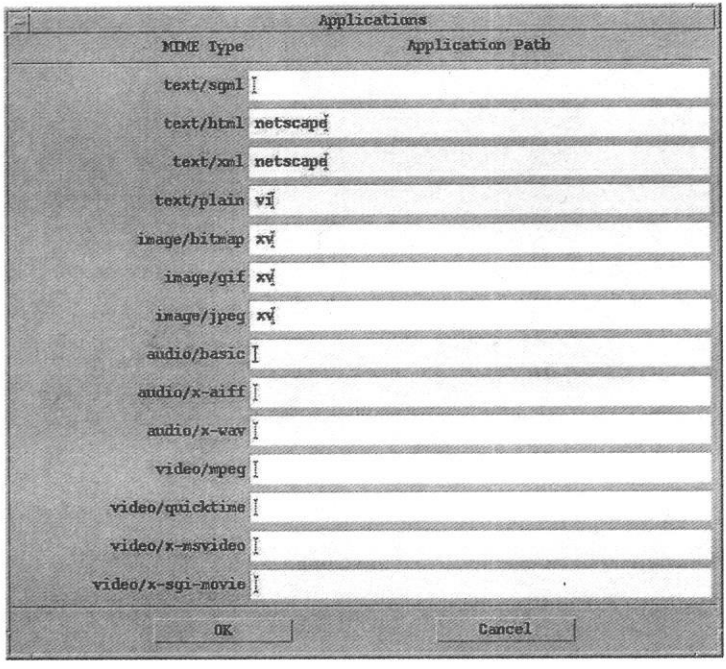
Çizim alanında listelenmiş olan katalog bilgilerinin görünüşüyle ilgili değişiklikler yapma olanağı da H-DCEdit programı tarafından sağlanmıştır. Şekil 6'da görünüş değişikliği sağlamak için kullanılan arabirim verilmiştir.



Şekil 4. DC elemanları giriş arayüzü

H-DCEdit programı oluşturduğu SGML belgelerini yükleyip yeniden işleme olanağı da verir. Bunun için SGML belgesi RDF/DC ayrıştırıcıya (parser) giriş olarak sunulur ve ayrıştırıcının çıktısından yararlanılarak belge bilgileri sisteme taşınır.

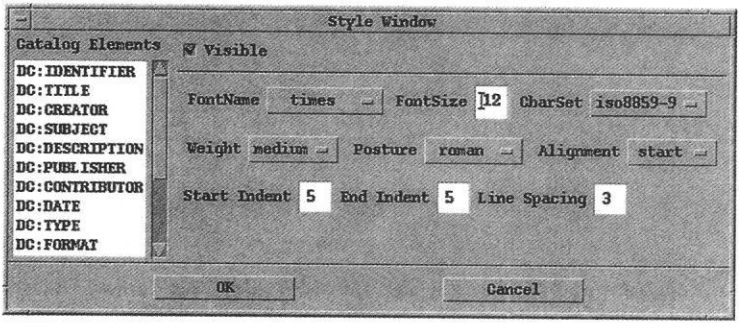
H-DCEdit programı, DSSSL[7] (International... 1996) standardının biçimleme olanağını da örneklemektedir. Programın çıktısı olan SGML belgeleri yardımcı program olarak kullanılan Jade DSSSL motoru (James'DSSSL Engine) yar-



Şekil 5. Yardımcı uygulama tanımlama arayüzü

7 *DSSSL (Document Style Semantics and Specification Language) nedir?*
Genelleşmiş işaretleme oluşturmanın amacı, biçimleme ve diğer işlem bilgilerini, belgenin kendisinden ayırmaktır. Her genelleşmiş, işaretleme şemasında, SGML işaretleme ile işleme belirtimlerini birleştirmek için bir yöntem vardır. DSSSL standardının ana amacı, SGML işaretleme ile işleme bilgilerini bir arada kullanabilmek için standart bir yöntem sağlamaktır. DSSSL, ISO/IEC 10179:1996 standart numarasıyla yayımlanmıştır. DSSSL standardı içinde, bir dönüştürüm dili, bir biçimleme dili, bir sorgu dili ve bir ifade dili tanımlanmıştır.

dımıyla HTML, RTF ya da TeX olarak yeniden biçimlenebilmektedir (Çevrimiçi, elektronik adres: http://www.jclark.com/jade http://www.jclark.com/jade, [25.12.1999]). Böylece, SGML standardının ilişkili olduğu DSSSL standardı da örneklenmiştir. Biçimleme işlemi için, belge tür tanımıyla uyumlu bir biçem tanımı yapılmıştır. Bu tanım sayesinde başka biçimlere çevirim işlemi yapılmıştır.



Şekil 6. Çizim alanı görünüş değişikliği sağlayan arayüz

2.5. RDF/DC Ayrıştırıcısı

RDF/DC ayrıştırıcı programı SGML belgesini ayrıştırır ve H-DCEdit programının anlayacağı biçimde bir ayrıştırma çıktısı oluşturur. Çıktı H-DCEdit programı tarafından sisteme yüklenerek eski bilgilerin yeniden işlenmesi sağlanır. Ayrıştırıcı program tarafından oluşturulan ara çıktı H-DCEdit programına yöneliktir.

RDF/DC programı, bir SGML ayrıştırıcısı olan SP (SGML Parser) paketi üzerine oturtulmuştur (Çevrimiçi, elektronik adres: http://www.jclark.com/sp 77 http://www.jclark.com/sp [25.12.1999]).

SP paketi SGML standardını büyük ölçüde desteklemektedir. Bu da RDF/DC programına zaman içinde istenildiği biçimde değiştirilebilecek olma esnekliğini kazandırır.

SP paketi programcılara SGML belgeleri üzerinde işlem yapma olanağını bir API (Application Program Interface) aracılığıyla sunar. RDF/DC ayrıştırıcısı da sunulan API olanağını kullanarak SP paketini bir kitaplık (library) olarak kullanmıştır. Ayrıştırma işlemi için class rdfdc oluşturulmuştur ve RDF/DC ayrıştırıcısı da class rdfdc'yi kullanan bir SGML uygulaması olarak geliştirilmiştir. rdfdcparse adlı uygulamanın, giriş olarak aldığı Ek 2'de görülen SGML belgesi için ürettiği ayrıştırma çıktısı Ek 3'te gösterilmiştir.

RDF/DC ayrıştırıcısı SGML belgesinde tanımlanan kaynak sayısı bilgisiyle başlar ve her eleman için eleman adı ve içeriği bilgilerini içerir. Ancak bazı elemanların içeriği sadece H-DCEdit programının anlayacağı biçimdedir. Görüldüğü gibi ayrıştırım çıktısı özel amaçlı bir içeriğe sahiptir ve H-DCEdit uygulaması tarafından kullanıldıktan sonra silinmektedir. Yani bu çıktıyı kullanıcının anlamasına gerek yoktur. Çünkü sadece bir ara çıktı olarak kullanılır, ihtiyaç kalmadığında ise silinir.

2.6. DSSSL Biçem Tanımı

DSSSL standardı dönüştürme, biçimleme ve sorgulama olmak üzere üç ana kesime ayrılır ve bunlar dışındaki kesimler ise bu kesimlere destek olarak kullanılır. H-DCEdit programı kullanıldığı DSSSL motorunun biçimleme olanağını kullanmıştır.

DSSSL motoru, SGML işleme sisteminin içine yerleştirilmiş biçimdedir ve hazırlanan SGML bildirimi ve belge tür tanımı, DSSSL motoru için de geçerlidir. Ayrıca, kullanılacak DSSSL dili için genel bir DSSSL belirtimine ve DSSSL ağaçlığına uygulanacak bir biçem dil tanımına ihtiyaç duyulmaktadır (Olgun 1999:57-65). Bu genel tanımlara uygun bir biçem tanımı yapılarak, SGML belgelerinin yeniden biçimlenmesi sağlanmıştır.

H-DCEdit programı tarafından oluşturulan bir SGML belgesi ve DSSSL biçem tanımının yönlendirdiği DSSSL motorunun biçimleme işlemi sonucunda HTML,

RTF, TeX belgeleri olarak yeniden oluşturulabilmektedir. RTF ve TeX belgeleri tek parça olarak oluşturulurken, HTML belgesi, biçim bilgisini taşıyan CSS (Cascading Sytle Sheets) tanımıyla birlikte üretilir (Çevrimiçi, elektronik adres: http://www.w3.org/Style/CSS http://www.w3.org/Style/CSS, [18.12.1999]). Ek 4'de Ek 2'deki SGML belgesinin karşılığı olan HTML belgesi ve Ek 5'de HTML biçim bilgisini gösteren CSS tanımı gösterilmiştir.

3. SONUÇ

Kaşgarlı Mahmut Bilgi Geri-Getirim Projesi kapsamında, meta-arayıcılar veya arabulucular gibi yazılım aracılarına girdi teşkil eden RDF/DC katalog bilgilerini üreten, H-DCEdit yazılımı gerçekleştirilmiştir. SGML ayrıştırıcısı, sp-1.2, ve DSSSL motoru, jade-1.0.1, modüllerini kullanan bu yazılım, UNIX ortamında MOTIF geliştirim aracı ile C/C++ dillerinde yazılmıştır. Yazılımda model olarak WWW Konsorsiyumunun bir çalışması olan RDF modeli temel alınmış, sözlük olarak da DC üstveri elemanları seçilmiştir. Oluşturulacak elektronik kaynaklar bilgileri ise SGML standardıyla desteklenmiştir. Böylece RDF çerçevesi içinde DC üstveri elemanlarını içeren SGML belgelerini oluşturan H-DCEdit yazılımı ortaya çıkmıştır. H-DCEdit programı ismi, "Hacettepe-Dublin Core Üstveri Elemanları Editörü" kelimelerinden türetilmiş bir kısaltmadır.

Yazılımda, ayrıca SGML standardının bir uzantısı olan DSSSL standardı da örneklenmiştir. Buna ek olarak; SGML belgelerinin, günümüzde HTML yerine geçecek olan XML standardına da kolay dönüşebilmesine destek verilmiştir. Oluşturulan SGML belgelerin Türkçe içeriği desteklemesi de sağlanmıştır. Kaynak kod ve belgeler http://www.cs.hun.edu.tr/~km Web sayfasında bulunmaktadır.

Çalışmamız kapsamında geliştirilen yazılım kesimleri genişletilmeye açıktır. RDF modelinin gelişmesi sürdükçe, kullanılan model karmaşıklaştırılabilir, DC

üstveri elemanları dışında diğer sözlükler de kullanılarak bilgi tanımlama yeteneği artırılabilir. Yapılan çalışma, bir dizinleme/arama aracı ile Web sitelerinde ya da kütüphanelerde kaynak tanımlamaya ve kaynak hakkında bilgi sunmaya yardımcı olarak da kullanılabilir (bkz. KMBGS: Gerçekleştirim Bölümü, [Çevrimiçi]. Elektronik adres: http://www.cs.hun.edu.tr/~km/gerceklestirim.html. [05.10.1999]).

Ek 1. Yazılım sisteminin kullandığı SGML bildirimi

```
<!SGML "ISO 8879:1986"
CHARSET
BASESET "ISO 646-1983//CHARSET International Reference Version
    (IRV)//ESC 2/5 4/0"
DESCSET     0               9    UNUSED
            9               2    9
            11              2    UNUSED
            13              1    13
            14              18   UNUSED
            32              95   32
            127             1    UNUSED
BASESET  "ISO Registration Number 148//CHARSET ECMA-128
     Right Part of Latin Alphabet Nr. 5//ESC 2/13 4/13"
DESCSET     128             32   UNUSED
            160             95   32
            255             1    UNUSED
CAPACITY   PUBLIC "ISO 8879:1986//CAPACITY Reference//EN"
SCOPE    DOCUMENT
SYNTAX
SHUNCHAR CONTROLS 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
    18 19 20 21 22 23 24 25 26 27 28 29 30 31 127 255
BASESET "ISO 8859-1//CHARSET International Reference Version
    (IRV)//ESC 2/5 4/0"
DESCSET     0               255  0
FUNCTION    RE                   13
            RS                   10
            SPACE                32
```

```
            TAB     SEPCHAR     9
NAMING      LCNMSTRT        " "
            UCNMSTRT        " "
            LCNMCHAR        "-.:"
            UCNMCHAR        "-.:"
            NAMECASE        GENERAL     YES
                            ENTITY  NO
DELIM       GENERAL         SGMLREF
            SHORTREF        SGMLREF
NAMES       SGMLREF
QUANTITY    SGMLREF     NAMELEN   40
FEATURES
MINIMIZE  DATATAG NO OMITTAG YES RANK NO SHORTTAG YES
LINK  SIMPLE NO IMPLICIT NO EXPLICIT NO
OTHER  CONCUR YES 999999 SUBDOC YES 99999999 FORMAL YES
APPINFO NONE>
```

Ek 2. H-DCEdit programı çıktısı bir SGML belgesi

```
<!DOCTYPE RDF:RDF PUBLIC "-//Baha Olgun//DTD RDF and DC//EN">
<RDF:RDF xmlns:rdf="http://www.w3.org/RDF/"
xmlns:dc="http://purl.org/DC/">
<RDF:Description about="http://sgml.cs.hun.edu.tr">
<DC:IDENTIFIER>http://sgml.cs.hun.edu.tr</DC:IDENTIFIER>
<DC:CREATOR>Baha Olgun</DC:CREATOR>
<DC:TITLE>SGML T&uuml;rkiye Kullan&iwhdot;c&iwhdot;lar&iwhdot;
</DC:TITLE>
<DC:SUBJECT>SGML T&uuml;rkiye</DC:SUBJECT>
<DC:DESCRIPTION>SGML T&uuml;rkiye Web Sayfas&iwhdot;
</DC:DESCRIPTION>
<DC:PUBLISHER>Hacettepe &Uuml;niversitesi</DC:PUBLISHER>
<DC:CONTRIBUTOR>Hayri Sever</DC:CONTRIBUTOR>
<DC:RIGHTS>Her Hakk&iwhdot; Sakl&iwhdot;d&iwhdot;r</DC:RIGHTS>
<DC:TYPE>text</DC:TYPE>
<DC:FORMAT>text/sgml</DC:FORMAT>
<DC:LANGUAGE>tr</DC:LANGUAGE>
<DC:DATE year=1998 month=Jan day=01></DC:DATE>
```

```
<DC:RELATION resource="http://www.cs.hun.edu.tr" http://
www.cs.hun.edu.tr type=IsReferencedBy></DC:RELATION>
</RDF:Description>
</RDF:RDF>
```

Ek 3. Ayrıştırım çıktısı

```
1
RDF:DESCRIPTION
DC:TITLE
SGML Türkiye Kullanıcıları
DC:CREATOR
Baha Olgun
DC:SUBJECT
SGML Türkiye
DC:DESCRIPTION
SGML Türkiye Web Sayfası
DC:PUBLISHER
Hacettepe Üniversitesi
DC:CONTRIBUTOR
Hayri Sever
DC:TYPE
0
DC:FORMAT
0
DC:IDENTIFIER
http://sgml.cs.hun.edu.tr
DC:LANGUAGE
0
DC:RIGHTS
Her Hakkı Saklıdır
DC:RELATION
6
http://www.cs.hun.edu.tr
DC:DATE
0
0
8
```

Ek 4. SGML belgesinin karşılığı olan HTML belgesi

```
<HTML>
<LINK REL=STYLESHEET TYPE="text/css" HREF="dc6.css">
<BODY>
<DIV CLASS=DC:IDENTIFIER>
<SPAN CLASS=DC:IDENTIFIER>
Identifier: http://sgml.cs.hun.edu.tr http://sgml.cs.hun.edu.tr
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION> </SPAN>
<DIV CLASS=DC:CREATOR>
<SPAN CLASS=DC:CREATOR>
Creator: Baha Olgun
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:TITLE>
<SPAN CLASS=DC:TITLE>
Title: SGML Türkiye Kullan&#305;c&#305;lar&#305;
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:SUBJECT>
<SPAN CLASS=DC:SUBJECT>
Subject: SGML Türkiye
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:DESCRIPTION>
<SPAN CLASS=DC:DESCRIPTION>
Description: SGML Türkiye Web Sayfas&#305;
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:PUBLISHER>
<SPAN CLASS=DC:PUBLISHER>
Publisher: Hacettepe Üniversitesi
```

```
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:CONTRIBUTER>
<SPAN CLASS=DC:CONTRIBUTER>
Contributer: Hayri Sever
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:RIGHTS>
<SPAN CLASS=DC:RIGHTS>
Rights: Her Hakk&#305; Sakl&#305;d&#305;r
</SPAN>
</DIV><SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:TYPE>
<SPAN CLASS=DC:TYPE>
Type: text
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:FORMAT>
<SPAN CLASS=DC:FORMAT>
Format: text/html
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:LANGUAGE>
<SPAN CLASS=DC:LANGUAGE>
Language: tr
</SPAN>
</DIV>
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:DATE>
<SPAN CLASS=DC:DATE>
Date: 01-Jan-1998
</SPAN>
</DIV>
```

```
<SPAN CLASS=RDF:DESCRIPTION></SPAN>
<DIV CLASS=DC:RELATION>
<SPAN CLASS=DC:RELATION>
Relation: IsReferencedBy http://www.cs.hun.edu.tr http://
www.cs.hun.edu.tr
</SPAN>
</DIV>
</BODY>
</HTML>
```

Ek 5. HTML belgesinin CSS uzantısı

```
SPAN.RDF:DESCRIPTION {
    font-family: Times New Roman,serif;
    font-weight: 500;
    font-style: normal;
    font-size: 10pt;
    color: #000000;
}
SPAN.DC:RELATION, SPAN.DC:DATE, SPAN.DC:LANGUAGE,
SPAN.DC:FORMAT, SPAN.DC:TYPE, SPAN.DC:RIGHTS,
SPAN.DC:CONTRIBUTOR, SPAN.DC:PUBLISHER, SPAN.DC:DESCRIPTION,
SPAN.DC:SUBJECT, SPAN.DC:TITLE, SPAN.DC:CREATOR,
SPAN.DC:IDENTIFIER {
    font-family: Times New Roman,serif;
    font-weight: 500;
    font-style: normal;
    font-size: 14pt;
    color: #000000;
}
DIV { margin-top: 0pt; margin-bottom: 0pt; margin-left: 0pt; margin-right:
0pt }
DIV.DC:RELATION, DIV.DC:DATE, DIV.DC:LANGUAGE, DIV.DC:FORMAT,
DIV.DC:TYPE, DIV.DC:RIGHTS, DIV.DC:CONTRIBUTOR,
DIV.DC:PUBLISHER, DIV.DC:DESCRIPTION, DIV.DC:SUBJECT,
DIV.DC:TITLE, DIV.DC:CREATOR,
DIV.DC:IDENTIFIER {
    text-align: left;
    line-height: 12pt;
    text-indent: 0pt;
```

KAYNAKÇA

Doorenbas, R.B., Etzioni, O. ve Weld, D.S. (1996). "A Scalable Comparison-Shopping Agent for the World-Wide Web", Technical Report No. UW-CSE-96-01-03, Depatment of Computer Science and Engineering, University of Washington. [Kaynak Sayfa]. [Çevrimiçi].
http://www.cs.washington.edu/research/projects/softbots/www/projects.html
[1998, Aralık]. ShopBot Projesi 1998 sonlarında sonlandırılıp, ticari olarak geliştirilmeye başlanmıştır
(bkz. http://www.jango.excite.com http://www.jango.excite.com)

Dublin Core Metadata Initiative. (1998). Dublin Core Element Set, Version 1.0. [Kaynak Sayfa]. [Çevrimiçi]. http://www.purl.org/dc. [1999, Aralık 25]

Etzioni, O. (1996). "The World Wide Web: Quagmire or Gold Mine?" , **ACM Comms.**, 39(1):65-68.

Gauch, S. ve Wang, Guijun, W. (1996). "Information Fusion with ProFusion" , WebNet'96: The First World Conference of the Web Society, San Fracisco, CA.

Gravano, L. ve Papakonstantinou, Y. (1998). "Mediating and Metasearching on the Internetî, **IEEE Data Engineering**, 21(2): 28-36.

International Organization for Standardization. (1986). ISO 8879/1986: Information Processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML). Ref. No. ISO 8879:1986 (E).

————. (1996). ISO/IEC 10179:1996: Information technology -- Processing languages -- Document Style Semantics and Specification Language (DSSSL).

Lassila, O. (1998 Temmuz-Ağustos). "Web Metadata A Matter of Semantics", **IEEE Internet Computing**, 30-37.

Olgun, B. (1999). Dublin Core Üstveri Elemanları Editörü. Yüksek mühendislik tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.Ankara [Çevrimiçi].
http://www.cs.hun.edu.tr/~km/belgeler.html [1999, Aralık 25]

Manber, U. ve Bigot, P.A. (1998). "Connecting Diverse Web Search Facilities" , **IEEE Data Engineering**, 21(2): 21-27.

Singh N. (1998 Mayıs). "Unifying Heterogeneous Information Models", **ACM Comms.**, 41(5) 37-44.

World Wide Web (W3C) Consortium (1999a). W3C: Cascading Sytle Sheets (CSSs). [Kaynak Sayfa]. [Çevrimiçi]. http://www.w3.org/Style/CSS. [1999, Aralık 18]

———. (1999b). W3C: Resource Desciption Framework (RDF) Model and Syntax Specification. [Kaynak Sayfa]. Haz. Lassila O. ve Swick, R.R. [Çevrimiçi]. http://www.w3.org/TR/REC-rdf-syntax [1999, Şubat 24].

———. (1999c). W3C: PICS - The Platform for Content Selection. [Kaynak Sayfa]. [Çevrimiçi]. http://www.w3.org/PICS/. [1999, Ekim 14].

———. (1998). W3C: Resource Description Framework (RDF) Schema Specification. [Kaynak Sayfa]. Haz. Brickley, D., Guha, R.V., and Layman, A. [Çevrimiçi]. http://www.w3.org/TR/WD-rdf-schema [1999, Mart 4]

———. (1997). W3C: Extensible Markup Language (XML). [Kaynak Sayfa]. Haz. Bray, T., Paoli, J., ve Sperberg-McQueen, C.M. [Çevrimiçi]. http://www.w3.org/TR/PR-xml-971208. [1999, Aralık 25].

TEŞEKKÜR

Bu makalede söz konusu edilen çalışma, T.C. D.P.T. tarafından desteklenen 97K121330 numaralı Kaşgarlı Mahmut Bilgi Geri-Getirim Sistemi Projesi kapsamında gerçekleştirilmiştir. Proje hakkında geniş bilgi, http://www.cs.hun.edu.tr/~km çevrimiçi Ev Sayfası adresinden elde edilebilinir (05/10/99, son günlenme tarihi).

Yazarlar ayrıca, bu makalenin hakemine, titiz çalışmasından ve yol gösterici önerilerinden dolayı teşekkür etmeyi bir borç bilmektedirler.