ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/351196230

Kültürel Miras Bilgisi Ötesindeki Sayısal Kütüphaneler

Preprint · January 2020

DOI: 10.13140/RG.2.2.32548.19842

CITATIONS
0

READS
5

2 authors, including:



Özhan Sağlık
Istanbul University
28 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Surveying The Evidentiary Value of Archived Electronic Records In The Terms of Trustworthiness View project

All content following this page was uploaded by Özhan Sağlık on 30 April 2021.

The user has requested enhancement of the downloaded file.

41. Bölüm
Kültürel Miras Bilgisi Ötesindeki Sayısal Kütüphaneler

Piedad Garrido Picazo
Zaragoza Üniversitesi, İspanya
Jesus Tramullas Saz
Zaragoza Üniversitesi, İspanya
Manuel Coll Villalta
Zaragoza Üniversitesi, İspanya

Çeviren
Özhan Sağlık - *Uludağ Üniversitesi*

ÖZET

Bu bölümde, kültürel mirasın erişimi ve yayılmasının aracı olarak sayısal kütüphaneler tanıtılmaktadır. Pek çok bilgi teknolojisinin kullanımıyla müteşekkil olan sayısal kütüphanelerin tarihi dokümanların analizi problemine sağlam temelli bir yaklaşım getirdiği savunulmaktadır. Bununla birlikte, program ve otomatik bir yorumlayıcı kullanımıyla dokümanter ve teknolojik varsayımları anlamanın sadece kültürel miras bilgisinin etiketlenmesinde kullanılabilecek daha iyi bir şema konusunda bilgilendirmekle kalmayacağı, çok etmenli sistemler, örüntü eşleme, bilgi yönetimi ve içerik ilişkisine dayalı bilgi görselleştirmesi gibi diğer ilgili alanlarda içerik çalışmaları sırasında karşılaşılan problemleri çözerek kullanıcıların ihtiyaç duyduğu kültürel miras bilgisini içerecek çok yönlü bir sayısal kütüphane prototipinin meydana gelmesine katkıda bulunması temenni edilmektedir.

GİRİŞ

Bu araştırma projesi Bölgesel Bilim, Teknoloji ve Üniversite Bakanlığı tarafından desteklenmiştir. Araştırma ekibi, kullanıcıların fazla sayıdaki muhtemel bilgi kullanımları nedeniyle yeterli gelmeyen bilgiye erişim ve sunumun klasik modellerinin sınırlılığındaki problemleri ortaya çıkmıştır. Bu problemler, 1990'lı yılların ikinci yarısında ciddi bir biçimde ele alınmış; Berners-Lee ve Fischetti'nin (1997) semantik web olarak bilinen fikirlerine kaynaklık etmiştir.

Sadece kültürel miras içeren bilgi ve dokümanları dijitalleştirmeye dayalı yaklaşımların, yavaş yavaş gerçekleştirilen proje ve edinilen tecrübelerle farklı tür kullanıcıların bilgi ihtiyaçlarını karşılamak noktasında yeterli gelmediği anlaşılmıştır. Kültürel miras hakkındaki oldukça geniş çoklu ortam bilgisinin kullanıcıların ihtiyaç duyduğu kapsamlı referans (gösterge) bilgisiyle birlikte oluşturulmadığı ifade edilmektedir (Crane, 2002).

Bu çalışmada bölümler şu şekilde yapılandırılmıştır: Etiketleme dillerini kullanarak sayısal kültürel miras kütüphanelerine yapılan katkıların arka planı birinci bölümde incelenmiş, ikinci bölümde tarihi metinleri çözümleyerek kontrol eden dijital ortamdaki tarihi dokümantasyonun yönetilmesi problemi açıklanmıştır. Bununla birlikte, çok etmenli sistemler geliştirmede örüntü işleme teknolojileriyle birleştirilmiş bilgiye erişim süreçlerinin geliştirilmesi, bilgi yönetimi ve görselleş-

tirmede otomatik yorumlayıcı geliştirmek için takip edilen yöntemler belirtilmiştir. Son bölümde, sonuçlar ifade edilmiştir.

ARKAPLAN

Sayısal kütüphaneler, genellikler içerdikleri dermenin çeşitliliğiyle tanımlanır (Crane, 2002). Sayısal kütüphanelerdeki mevcut kaynaklar, heterojenlik durumları göz önüne alındığında, bilgiye erişim ve karşılıklı çalışabilirlik gibi sorunların çözümünde faydalanılmasından beri farklı içerikleri tanımlamada kullanılan üstveri şemalarının oldukça önemli hale geldiğini göstermektedir. Baldonado, Chang, Gravano ve Paepcke (1997) Dublin Core (Tolosana-Calasanz, Nogueras-lso, Bejar, MuroMedrano, & Zarazaga-Soria, 2006) ve MARC şeması (Chand ler, Fol ey, & Hafez, 2000) kullanarak üstveri temelli bir karşılıklı çalışabilirlik modeli geliştirmişlerdir.

Sağladığı kolaylıklardan dolayı, kültürel miras içeren sayısal kütüphanelerde Temel Dublin Core Üstveri Şeması daha çok kullanılmaktadır. Bunun yanı sıra, bazı araştırmacılar Dublin Core kullanırken karşılaşılan problemleri ifade etmiştir. Foulonneau, Cole, Habing, and Shreeves (2005) Dublin Core Üstverileri'nin Open Access Initiatives (Açık Erişim Girişimi [OAI]) kapsamındaki arşivlerde bulunan yayınlar için yeteri kadar kapsamlı gelmediği sorununa dikkat çekmiştir. Çeşitli arşivlerdeki kültürel mirası tanımlayan üstverilerle ilgili karşılaştırmalı bir çalışma yapan Hutt ve Riley (2005) aynı sorunu ifade etmiştir. Halbert, Haczmarek, and Hagedorn (2003) bu bağlamda Dublin Core kullanımında karşılaşılan problemleri açıklamıştır.

Sahadaki çalışmalar incelendiğinde, üstveri şemalarının içermediği semantik ilişkinin çeşitli araçlarla kurularak Dublin Core'nin kullanımının geliştirilmesi gerektiği anlaşılmaktadır (De Gendt, lsaac, Van Der Meijt, & Schlobach, 2006). İncelenen çalışmalar neticesinde, iki adet başlama noktası ifade edilebilir:

1. Dublin Core'nin diğer kapsamlı şemalar olmadan kültürel miras içeren sayısal kütüphanelerdeki kullanımı oldukça basit kalmaktadır.
2. Malzeme ve derme türüne ilişkin tanımlayıcı üstveri şeması ile thesaurus (kavramlar dizini), ontoloji gibi tanımlama aracı şemalarını bünyesinde barındıran daha kapsamlı bir modelin geliştirilmesi gerekli görülmektedir.

TARİHSEL BİLGİ ANALİZİ

İncelenen Web bilgi servisi Sayısal Aragon Ansiklopedisidir (http://www.enciclopedia-aragonesa.com/). Kültüre miras bilgisini muhafaza etmek çok kapsamlı ve çeşitlilik arz eden bir alan olduğundan, bu bilginin bir kısmı olan tarihi metinlerin nasıl yönetileceği bu bölümde açıklanacaktır.

Günümüzde, sayısal kütüphanelerde tarihi metinlerin analizi için iki bakış açısı mevcuttur. Birincisinde, bu tür tarihi metinlerdeki mevcut bilgilerin analizinin insan çabasıyla tamamlanması mümkündür fakat tam olarak koordine edilmiş okuyucu grubuna ihtiyaç duyulmaktadır. Analiz zamanı, okuma hızı analizi ve okuyucu sayına göre değişecektir. İkincisi ise, ilgili bilgileri ayıklayabilen insan ve harici otomatik sistem tarafından okunaklı hale getirilebilecek bir bilgisayar aracılığıyla yapılandırarak otomatik bir analiz yapmaktır.

Tarihi metinlerin otomatik analizi yapılırken karşılaşılan en ilginç özellikler, metinlerin şu iki

hususu karşılayıp karşılamamasına göre şekillenmektedir:

1. Olaylar grubu, varlıklar grubuyla ilişkili olarak ortaya çıkarlar (somut bir yer ve zamanda gerçekleşirler).
2. Ayrıca, büyük sayıdaki varlıkların iç kaynaklı bilgileri diğer varlıklarla kurdukları ilişki ağlarında bulunabilmektedir. Semantik ilişkiler, onların rolünü yeteri kadar yansıtamayabilir.

Bu özelliklerden, analiz sürecinden sonra elde edilen bilgilerin her düğümün başka bir düğümle ilişkilendirildiği varlıklar olarak kurgulandığı ve insan etkileşimli semantik bilgilerin düğümler arasındaki ilişkiyi ve belirli bir zamanda gerçekleşmiş vakaların tarih ve mekânını içeren kemerler şeklinde etiketlendirildiği bir ağ yapısına sahip olması gerektiği anlaşılmaktadır.

Sayısal kütüphane hizmetlerine yönelik diğer yönler arasında, sayısal bilgi ürünü, sesli arama motoru geliştirmek için çaba göstermektedir. Böylece, bilgiye erişimin sadece konuya dayalı kategorilere ve/veya mevcut abecesel dizinlere bağlı olmaması hedeflenmektedir (Brocks, Thiel, Stein & Dirsch-Weigand, 2001). Bu nedenle, los Reyes de Aragon (Aragon Kralı ve Kraliçesi) ile beraber çalışabilecek ses setleri seçilmiştir.

Ses setleri analiz edilince, bu etiketin kültürel miras bilgisine ait parçanın yönetim, saklama ve erişimine yeteri kadar sağlıklı çözüm sunamadığı, fakat düzenleme yapabilmek için metin yapısına katkı sağladığı görülmüştür. Bu meseleyi inceleyen mevcut girişimler değerlendirilmiş ve sadece müze koleksiyonlarıyla ilgili heterojen bilimsel dokümantasyonun değişimi ve benimsenmesiyle ilişkili oldukları görülmüştür. Bu nedenle, özel bir amaca matuf olacak bir ontoloji geliştirilmesine karar verilmiştir (ISO 21127, 2016).

COLLATE Projesinin aksine, bu projede ontoloji kullanımı, ISO 13250'de konu haritaları paradigması etrafında şekillendirilen ve tanımlanan düzen ışığında koleksiyondaki malzemelerle kurulan semantik ilişkinin kısıtlılığına bağlı değildir (ISO 13250, 2003). Bu standart, bilgilerin nasıl organize edildiği hakkında belirgin bir fikir oluşturmak ve bu konudaki ses setlerinin artmasıyla sadece otomatik tercümanın hata yapmamasını önlemekle görevli olmayan fakat hanedan ve soy araştırmaları gibi diğer ansiklopedik bilgi sektörlerinde de değerlendirilebilecek ilişki modellerinin nasıl belirleneceği konusunda kullanılmaktadır (Giunchiglia, Yatskevich & Shvaiko 2007).

Bilgiyi belleğe almak için hibrid XTM/DITA şeması geliştirilmiştir. Bununla birlikte, XTM varlıklar arasındaki ilişkinin nasıl yorumlanacağı ve hafızaya alınacağını şu üstverilerle belirlemektedir: "konu", "dosya yolu (baseName)", "dosya yolu dizisi (baseNameString)", "nesne tipi sorgusu (instanceOf)", "konu sorgusu (topicRef)", "üye", "ilişki", "olay", ve bir vakayı tanımlamak için yeni bir etiket olarak "tarih". Bunun yanı sıra, her varlığın içyapısı şu temel DITA etiketleriyle tanımlanabilir. "konu id", "kısa açıklama", "içerik (conbody)". Sonrasında, bu şemayı kullanarak ses setlerini otomatik nitelendiren bir yorumlayıcı geliştirilmiştir (Maier & Delcambre, 2000).

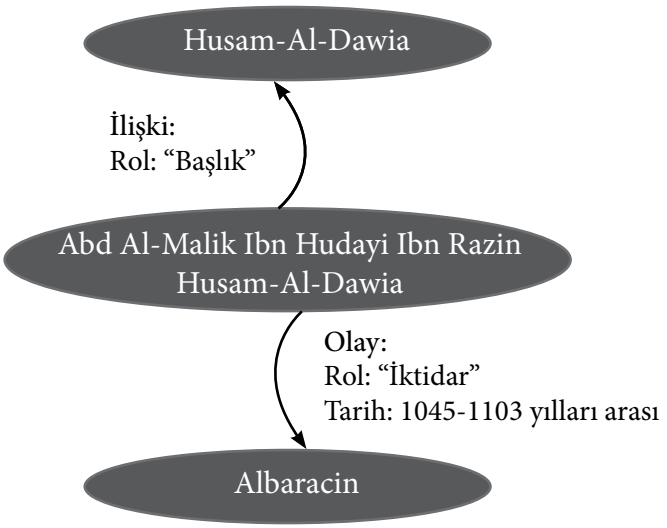
Bilginin gösteriminde kullanılan bu iki format arasındaki farklar belirgin olup, birleştirilmiş DITA ve XML üstveri şemalarının yararları sayesinde proje katkıları kısmında belirtildiği üzere bazı problemler çözülmüştür. Bu, sahada oldukça yenilikçi bir yaklaşım olarak kabul edilmektedir. Kültürel mirasın teknik dokümantasyonu için bir üstveri şeması olarak DITA (2007) ve nitelendirilmiş bilgi kaynakları sistemlerinin üretimi için hazırlanan XTM (Park ve Hunting, 2002) semantik un-

surlar ile bilgilendirici tanımlayıcılar kullanılarak diğer kaynaklarla bağlantı kurmak için özel olarak tasarlanmıştır.

Şekil 1: XML etiketli örnek bir ses

```
<voz subcategoriaId="38">
  <vozId>
    98
  </vozId>
  <nombre>
    Abd al-Malik ibn Hudayl ibn Razin
  </nombre>
  <descripcion>
    Segundo soberano de la taifa de Albarracín, entre 1045 y 1103, con el título de Husam al-Dawla (Sable del Estado).
  </descripcion>
</voz>
```

Şekil 2: Konu Haritası Yapısı



OTOMATİK TERCÜMAN

Tarihsel bilgi analizinin amacı dokümanın kapsadığı varlıklar hakkında ilgili tüm bilginin tanımlandığı bir ilişki grubu açığa çıkarmaktır. Bu yüzden tercüman algoritması metni okuyabilmeli ve bir şekilde varlıklar arasındaki ilişkiyi ya da ilgili bir varlığı tespit edebilmeli ve onları bir ya da çeşitli şekillerde saklayabilmelidir.

Sorunun kaynağı, üstveri, rol, veri veya mekân ve ilişkili varlıkları ortaya çıkarmak için kullanılan yöntemlerde olduğu gibi somut bilgilerin bir ilişki ya da olaya dönüşüp dönüşmediğini belirlemek için kullanılan tespit yönteminde bulunmaktadır. Bu nedenle iki yaklaşım benimsenmektedir. İlki temel seviye algoritması olup, analiz, metin üzerinde gerçekleştirilir. Önce, metinden cümleler çıkarılır, sonrasında kelimeler incelenir. Her kelime, rollerin bir ilişkiyi (soy, hanedanlık, miras vb. gibi) ya da varlığı ifade etmesi bakımından analiz edilir. Eğer bir varlık ifade ediliyorsa, roller diğer cümleler kapsamında sorgulanır; sorgulama başarılı sonuçlar verirse, konu başlıkları arasındaki ilişki tesis edilmiş olur. Varlıkları tespit etmek için ise mevcut köşeli parantezler (olayların tarihini belirtmek için genellikle kullanılan şekil) üzerinde temel bir inceleme yapılır.

Yukarıda açıklanan algoritmada, inceleme doğrudan metindeki kelimeler üzerine olduğu için

ilk seviye olarak tanımlanmakta ve ilişkiler ile varlıkların oluşumu daha önce belirlenen kelimeler kapsamında gerçekleştirilir. Bu kelimeler, metinde var olan bilgiler hakkında ipuçları sunmaktadır. Buradaki inceleme yöntemiyle gerçek hayat arasında da bir benzerlik kurulabilir. Bu benzerliğe, kapsamlı bir metinde bazı bilgiler bulmak istediğimizde metindeki bilgiler hakkında bize ipuçları verecek anahtar kelimeler aracılığıyla bir dizi okuma yapmak örnek verilebilir.

Bu tür analizlerin sonuçları, otomatik tarihsel bilgi analiziyle yakınlık gösterdiği sürece olumlu kabul edilmektedir. Mesela, 200 ifadenin geçtiği bir metinde insan etkileşimine kıyasla algoritma yaklaşık %70'lik bir doğruluk oranıyla yüz ilişki ve 10 varlık tesis etmiş olabilir. Fakat bu oran artırılabilir. Gerçekleştirilen analizler sözlüğe bağlı olup, tanımlanacak ifadenin karmaşıklığına göre başarısı azalabilir. Bu nedenle sözlükten bağımsız olarak daha üst bir seviyede inceleme yapmak gerekli görünmektedir. Bu yeni yaklaşım, karmaşık ifadeleri azaltarak daha basit hale getirir ve kolay bir tercüman aracılığıyla analiz edilmesini sağlar.

Böylece ikinci bir algoritma geliştirilmiştir. Bu algoritmada tanımlayıcı bir metin ve morfolojik analizler Freeling uygulaması aracılığıyla gerçekleştirilmektedir. Burada, kelimeler metinsel formlarıyla analiz edilmez, varlıklar ve ilişkileri tesis etmek ve aranan bilgiyi ortaya çıkarmak için cümleler morfolojik yapıları bağlamında değerlendirilir.

İlk olarak morfolojik yapı düzenlenir, İspanyolca'da dikkate alınmayacak kelimelerin listesi oluşturulur. Sonrasında, Freeling uygulamasından kaynaklı hataları (cins ve özel adları ayırmak, bazı fiilleri cins ada çevirmek ve herkesçe bilinen kısaltmaların karşılıklarını algılatmak gibi) düzeltmek için kelime kelime küçük analizler yapılır. Çift giriş sistemi uygulanır ve takip eden analiz safhalarında hata yapmamak için kelimeler simultane olarak çeşitli biçimlerde kataloglanır (İspanyolca'da fiil olarak "era" veya cins ad olarak "epoch" gibi).

Sonraki süreçlerde, söz dizimlerini daha önceden tanımlanmış örüntü serileriyle karşılaştırarak cümle cümle analizler yapılır. Bu önceden tanımlanan işaretler, algoritmaların hedeflediği sonucu yani bize aradığımız bilgiyi veren ifadelerin dil bilimsel biçimlerini tanımlar. Aslında, algoritma varlıkları açıklayan metinlere ait cümlelere karşılık daha önceden programlanmış örüntüleri kıyaslamak için bir kural motoru kullanır. Bu doğrulama, örüntüleri çeşitli biçimlerde kolaylaştırmak için her cümlede yapılır ve doğrulama sırasında cümlelerin ön ve son eklerinin analiz yeterliliğini olumsuz etkilememesine önem verilir. Çoğu İspanyolca cümlede, sonradan düzeltme yapma imkânı olmasına rağmen cümlelerin başındaki ve sonundaki ifade formüllerini dikkate alarak muhtemel söz dizimi örüntülerini elemek oldukça zordur. Bu nedenle, gerekli olan önceden tanımlanmış örüntüler, yorumlanan varlıkla ilgili olaylar veya varlıklar arasındaki ilişkileri kurarak aradığımız bilgiyi bildiren genel ifade formüllerini tanımlayan morfolojik kombinasyonlar olacaktır.

Sonrasında, aşağıdaki örnekle açıklayacağımız algoritmanın işleme alınması için tanımlanması gerekmektedir. Elimizde şöyle bir cümle olsun: "A comienzos de su reinado, se casô con Magdalena de Folcaquier" (Saltanatının başında, Magdalena Folcaquier ile evlendi). İlk aşamada, aradığımız bilgi çerçevesinde birbiriyle ilişkili olmayan kelimeler ayıklanacaktır. İkinci aşamada aşağıdaki yapı bağlamında morfolojik analiz gerçekleştirilir. "**A** (Bulunma hâli; SP, edat), **comienzos** (başta; NC, cins ad), **de** (Vasıta hali; SP, edat), **su** (İyelik hâli; DP, iyelik eki), **reinado** (saltanat; NC, cins ad), **se casô** (evlendi, VM, birinci şahıs fiil çekimi), **con** (Yönelme hâli; SP, edat), **Magdalena Folcaquier** (Magdalena Folcaquier; NP, özel isim)". Bu morfolojik yapı, aradığımız bilgiyi bulmak amacıyla

daha önceden tanımlanan örüntülerin algoritma serisiyle eşleştirilecektir. Tanımlanan örüntülerden herhangi biri " VM, SP ya da NP" olarak belirlendiyse "casô" kelimesini içeren cümleler bu örüntülerle eşleşecektir. Bu eşleşmeyi başarıyla gerçekleştirmek için başka varlıklar üzerindeki fiilleri de tanımlayan oldukça fazla cümleyi analiz etmek gerekmektedir.

Eğer yeterli sayıda örüntü geliştirildiyse, tanımlanacak ifadelerin oldukça büyük sayıdaki dil bilimsel biçimlerini tekleştirmek mümkün görünmektedir. Fakat artık bilgiler, metnin bazen çok fazla bilgi içerip, artık ya da doğru olmayan bilgilerin yanlış bir şekilde ayıklanması gibi bazı problemler teşkil etmektedir. Bu problemleri engellemek için algoritmalar özet örüntüler kullanır. Bu özet örüntüler, yanlış bilgi sunmamak için uygun örüntüleri kullanarak artık cümlelere uygulanacak söz dizimi kurallarını tanımlar ve yukarıda açıklanan artık bilgilerin mahiyetinden dolayı bu cümleleri ayrıştırır.

Uygulanacak bu örüntü kuralları için şöyle bir gelişigüzel örnek verilebilir: "Hijo (NC) de (SP) Pedro II (NP), rey (NC) de (SP) Aragôn (NP) y (CC) Juana (NP), condesa (NC) de (SP) Urdiel (NP) (II. Peter'ın oğlu, Aragôn Kralı ve Juana, Urdiel Kontesi)." Örnekte görebildiğimiz gibi, incelenen varlıkta, "Pedro I" (Peter II) "rey de Aragôn" (Aragôn Kralı) ve "Juana" "condesa de Urdiel" (Urdiel Kontesi) olmaktadır. Burada bu bilgiler birbiriyle örtüşmediği için artık bilgiler ayrıştırılmıştır. Eğer bu tür artık bilgileri bulup, eleyebilecek özet örüntüler programlanırsa, daha sonraki analizlerde kullanılacak örüntüler daha kolay olabilecektir.

Dikkat çekici bir özet örüntü şu örnekle açıklanabilir: Son üç kelimeyi silen bir işlem süreci kuralını gündeme getiren " NP NC SP NP" örüntü düzenini ele alalım. Örneğimizdeki cümlede "Pedro II" kelimesi için bu kural uygulanacaktır. Sonuçta elde edilecek cümle şu şekilde olacaktır: " Hijo de Pedro II y Juana, condesa de Urdiel" (Peter II'nin oğlu ve Joan, Urdiel Kontesi). Örneğimize bu kural "Juana" kelimesinden itibaren işleme alınacak ve özetlenen cümle nihai olarak şöyle olacaktır: "Hijo (NC) de (SP) Pedro II (NP) y (CC) Juana (NP)" (Peter II'nin oğlu ve Juana). Daha önceden tanımlanan örüntüler bu açıklamadaki gibi programlanırsa, aranan bilgi ortaya çıkarılabilir. Bu örnekten hareketle, "Pedro II" (Peter II) ve "Juana" (Joan) varlıkları söz konusu olduğunda aranacak varlığın "hijo" (oğul) olduğu anlaşılmaktadır. Buna rağmen, özet örüntülerini kullanmanın örneklerde görüldüğü üzere aranan bilgiyle ilgili olan bilgileri ayrıştırma riski de vardır.

Olayların tespit edilmesinde üç tür örüntü kullanılmaktadır. Bunlar, başlangıç-bitiş tespiti yapan örüntüler, parantez analizi ve ortak olay örüntüleri. İlki tarihi belgelerdeki ortak özelliklere dayalıdır. Tarihi bir varlık söz konusu olduğunda, başlangıç ve bitiş tarihleri ile mekânlar her zaman ilk cümlede parantez içerisinde verilir. Bu türden bilgileri ortaya dökebilmek ve bu türden ifadelere ait kombinasyonları bulabilmek için bir dizi örüntü geliştirilir. Parantez analizi, çok bilinen bir bibliyografik formülü kullanır; cümlede parantezler arasındaki tarihler tanımlanacak olayın geçtiği zamanı gösterir. Parantez analizi için kullanılan örüntüler bu tür ifade araçlarını bularak bunları ortak olay örüntülerinin keşfedebileceği kelime dizileri haline getirir.

Bir örnekle açıklayacak olursak, "batalla de Roma (315-340)" (Roma Savaşı) cümlesi tespit edilir ve analiz edilecek olan şu cümleye dönüştürülür: "batalla de Roma entre 315 y 340" (315 ve 340 yılları arasında gerçekleşen Roma Savaşı). Bu, çok kullanılan ortak olay örüntülerine bir örnek olup, bir önceki cümlede olduğu gibi olayları tanımlayan ifadeleri içerir ve aranan bilgiyi tespit ederek ortaya çıkarır. Sonuç olarak tüm bu örüntü grupları incelendiğinde, sonraki analizleri kolaylaştırmak

için parantez arasına alınan tüm cümle ayrıştırılmıştır.

Cümlede bazı örüntüler bulunduğu için örüntü kuralları uygulandığında, ilişkili olduğu kural benimsenir ve ilişkili olduğu prosedür işleme alınır. İşleme alınan prosedür, sentez, özet veya parantez örüntülerinden önce geliyorsa cümleyi düzeltebilir veya ilişkiler ya da olaylar meydana getirir. Son olarak uygulanacak kurallar kelimelerin ifade ettikleri prosedürü işaret eder. Veri, mekân ve ilişki kuralı ait oldukları olay ya da varlığın mevcudiyetine göre uygulanırlar. Bu verilerle prosedür, cümle ve kelimeleri açığa çıkararak ilişkileri ve/veya olayları meydana getirir.

Analizin ikinci safhası bitince, ilişki ve olaylara ait bir liste oluşur. Bu liste, üçüncü safhada ilişkilerin geçersiz tarih içerip içermediği ya da anlamlı olup olmadığına bakılarak ayrıştırılması marifetiyle doğrulanır. Bu safhada, referansların çözümlenmesi de gerçekleştirilir.

Bu aramayı yapmanın şu problemi çözmek gibi bir amacı vardır: Tarihi belgelerin analizinde referans gösterilen varlıklar, oluşturulan listelerde bir konu olarak görünebilir. Bir varlığın bir başlığa karşılık gelmesi de mümkündür fakat bu durum alternatif bir biçimde adlandırılır. Aslında, algoritma, varlığı ifade eden ilişki parçasını varlığın gerçek adını kullanarak değiştirmelidir. Mesela, "Alfonso X El Sabio" [Alfonseo X Bilge] örneğinde konu "Alfonso X" veya "Alfonso El Sabio" olarak ortaya çıkabilir. Burada konu her ne kadar geçerli referanslara sahip olsa da uygun adla değiştirilmelidir. Algoritma, büyük harfle başlayan kelimeleri ayıklayarak referansları kontrol eder. Sonrasında, başlığına bakarak aynı işlem konu konu gerçekleştirilir ve konular birbirleriyle kıyaslanır.

Bu durumda her zaman 2'ye eşit veya ondan büyük bir numaraya sahip olan ve pek çok sayıda kelimeyi muhteva eden geçerli referanslar söz konusu olacaktır. Böylece, ilişkili metin referans gösterilen uygun konuyla değiştirilecektir. Farklılıklar geliştirildiğinde, çapraz aramalar aracılığıyla ilişki listelerinde tek veya çift yönlülük hesaplanacaktır.

Sonuç olarak ilişkiler ve olayların ilgili rolleri hesaplanır. Bu hesaplamayı çözümlemek için kullanılan iki yaklaşım vardır. Birincisi, diğer rolleri devre dışı bırakarak önekli rollerle birleştirilen önekli ilişkilerden oluşmaktadır. Bu durumda, algoritma daha önceden tanımlanan bir tablo aracılığıyla ilişkileri saptar. Fakat bu yöntem çok kullanışlı görülmemektedir. İkinci yaklaşımda ise, son algoritmada kullanılan rollerin sayımı benimsenmektedir. Daha sık karşılaşılan rollere daha fazla ilişki atanmakta; sık karşılanmayanlara ise daha az ilişki tayin edilmektedir. Bu yöntem, daha önceden tanımlanmış tablolar ve mevcut bilgilerden bağımsız olduğu için daha kullanışlı görülmektedir. Bir ilişki atandığında, filtreleme aracılığıyla uygun olmayan tarihler ve az kullanılan roller ayıklanmaktadır.

Aranan bilgiye ilişkin konu başlıklarını oluşturmak için ikinci bir analiz yapılması gerekli görülmedikçe özetleme işlemi sırasında bilgi ayrıştırma işlemi gerçekleştirilmemektedir. Fakat dokümanların kontekstine bağlı olarak bu durumun aksi uygulanabilir. Bu durum şöyle bir örnekle açıklanabilir: " Hijo (NC) de (SP) Pedro II (NP), rey (NC) de (SP) Aragôn (NP) y (CC) Juana (NP), condesa (NC) de (SP) Urdiel (NP)" (Peter II'nin oğlu, Aragôn Kralı ve Joan, Urdiel Kontesi). Burada "rey (NC) de (SP) Aragôn (NP)" (Aragôn Kralı) ve "condesa (NC) de (SP) Urdiel (NP)" (Urdiel Kontesi) ifadelerinin ayrıştırıldığını görürüz. Konunun konteksti analiz edildiğinde bu bilginin farklı bir yapıya sahip olduğu anlaşılır fakat genel bir bakış açısıyla incelendiğinde herhangi bir farklılık görülemeyebilir. Genel olarak aranan bilgi, dokümanda bulunduğu sürece onu geçici olarak saklamak ilginç bir yaklaşım olacaktır. Bu yaklaşımı uygulamak için algoritma muhtemel durumlar içerisinde

ilişkiler oluşturarak kurtarma yöntemini kullanacaktır. Oluşturulan ilişkiler, aralarında bağ kurdukları konular bilinmediği sürece öksüz kalacaktır.

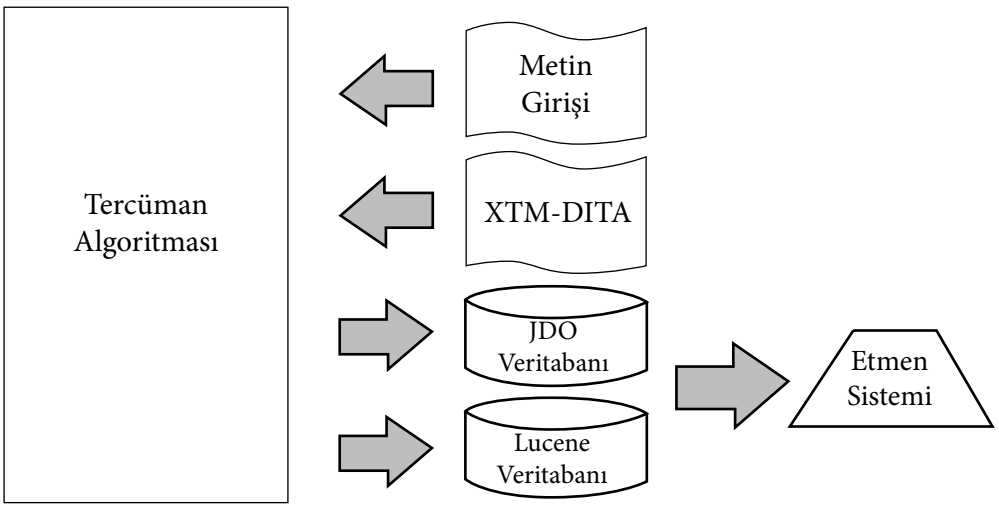
Referans analizine benzer bir şekilde buradaki algoritma da her ilişkinin referans gösterildiği konuyu arayıp, ilgili konuyu bulduğu zaman ilişki ve konu arasında bağ oluşturacaktır. Bir önceki örnekten hareket edersek, konu başlığı olarak "Pedro II" (Peter II) adı aranacak ve "rey (rol) ile de Aragôn (referans gösterilen varlık) arasında bir bağ kurulacaktır. Sonrasında "Juana" (Juana) konu başlığını bulacak ve "condesa (rol) de Urdiel (referans gösterilen varlık)" (Urdiel Kontesi) arasında bir bağ oluşturacaktır. Aslında referans gösterilen konu bulunmamış, algoritma mevcut ilişkilerle yeni bir konu meydana getirmiştir. Fakat gerekli olan yapının dâhilinde olmadığı için göz ardı edilmiştir. Yaptığımız uygulamalar ikinci yaklaşımda benimsenen algoritmanın üç binden fazla ilişki oluşturduğunu ve iki bin ifadeden bin olay meydana çıkardığını göstermektedir. Örneklem geliştikçe, sonuçların da katlanarak artacağı anlaşılmaktadır. Oluşturulan tarihlerin doğruluğu %90 civarında olup, %20'si ilk denemede doğrulanmıştır.

BİLGİ YÖNETİMİ

Olaylar ve ilişkilerin listesine sahip olduğumuzda, bunları korumanın üç yolu vardır. Bunlar şöyle belirtilebilir:

1. Konuyla ilgili mevcut olan her bilgiyle birlikte ilgili olay ve ilişkileri muhteva eden XTM/DITA hibrid formatında XML dokümanı oluşturmak. Referans gösterilen varlıklar mevcut olduğu ve dokümanın kendisinde konuları muhteva ettiği sürece kurulan ilişkiler karşılıklı etkileşime açık olmalıdır. Ayrıca, mevcut algoritma sadece orijinal dokümanda geçerli olsa bile diğer dokümanlardaki varlıklara yapılan referansları incelemek ve bu referansları birleştirmek mümkün hale getirilmelidir.
2. Java veri nesneleri (Java data object [JDO]) saklamada Java persistent Objects (Java Kalıcı Nesneler [JPOX]) teknolojisini kullanmaktadır. Aşağıdaki gibi nesneye dayalı bir veritabanı elde etmek ilişkisel veritabanı yönetim sistemine sahip olunması anlamına gelecektir. Böylece, nesneye dayalı veritabanı kullanımını benimseyen programcılar veya ilişkisel bir veritabanına ihtiyaç duyan kullanıcılar açısından oldukça faydalı olacaktır. Böylece, tüm asli bilgiler veritabanında saklanmakta ve keşif araçları, Web sistemleri ve uygulamaları gibi aracılar marifetiyle sorgulama yapılabilmektedir.
3. Üçüncü ve son yöntem, Lucene Analyzer (Tam metin arama motoru [çevirmen notu]) aracılığıyla konuların dizinlenmesidir. Bu dizin bize aradığımız konuyu bulmamızı sağlayacaktır.

Şekil 3: Tercüman algoritmasının girdi ve çıktıları



ARAMA ETMENLERİ SİSTEMİ

Java Etmen Geliştirme (Java Agent Development [JADE]) platformu aracılığıyla çok etmenli bir sistem geliştirilmiştir. Dış etmenler tarafından gelen istekleri toplamak ve arama sonucunda elde edilen konularla bunları cevaplamakla görevli olan özel bir etmen mevcuttur. Bu etmen, geliştirilen nesneye dayalı veritabanı ve kullanılan Lucene arasında arama sorgusundan hareketle hangi konunun veritabanından ayıklanacağının belirlenmesi hususundaki ilişkiyi doğrular. Etmen, ayrıntılı arama yapabilmek için çeşitli ihtimalleri dikkate alır ve pek çok arama etmeni eş zamanlı çalışır fakat hiçbiri veriyi düzeltmez.

Ayrıca, sistemin esnekliğini test etmek için aşağıdaki süreçleri takip ederek başka bir arama etmeni geliştirilir. Karakter dizinine bakarak etmen Google'ı tarar ve orijinal etmene buradan elde ettiği sonuçları gönderir. Başka bir deyişle, geliştirilen arama etmeni prototipi, farklı ortamlarda da arama yapabilmek için versiyonlar oluşturmamızı sağlar.

Bununla birlikte, Grafiksel Kullanıcı Arayüzüne (Graphical User Interface [GUI]) uyumlu olarak birleştirilmiş arayüz prototipinden hareketle iki farklı etmen türü geliştirilmiştir. Oluşturulan versiyonlar, farklı arayüz problemlerini çözmektedir. Biri konsol kullanıcı arayüzü etmeni, diğeri de grafiksel kullanıcı arayüzü etmenidir. Her ikisi de herhangi bir etmenle aynı tespit ve iletişim sistemine sahiptir. Çoklu etmen sistemi kullanılarak bir sorgudan hareketle geliştirilen 3 boyutlu konu ağacını bir etmen aracılığıyla görmek mümkündür.

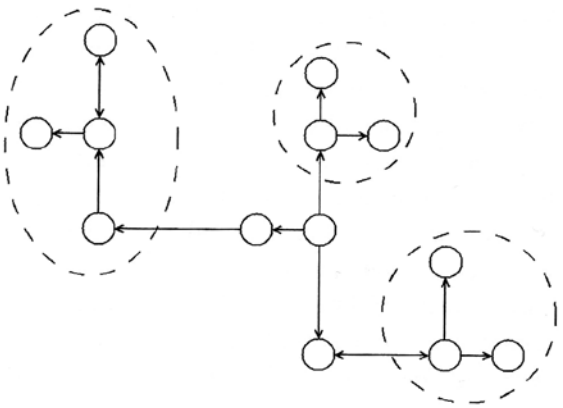
BİLGİNİN GÖRSELLEŞTİRİLMESİ

Tüm işlemlerden önce, sunulan tarihsel bilgi hazırlanır, işlenir, farklı depolama cihazlarında saklanır ve mümkün olduğunda grafiksel görselleştirmesi oluşturulur. XTMD/DITA formatı bilgi sunan her konunun arasındaki ilişkiyi kurmak için etiketlenen düğümlerin birbirine bağlandığı örgü düğümlerde olduğu gibi bilgiyi sunmamızı sağlayan özelliklere sahiptir.

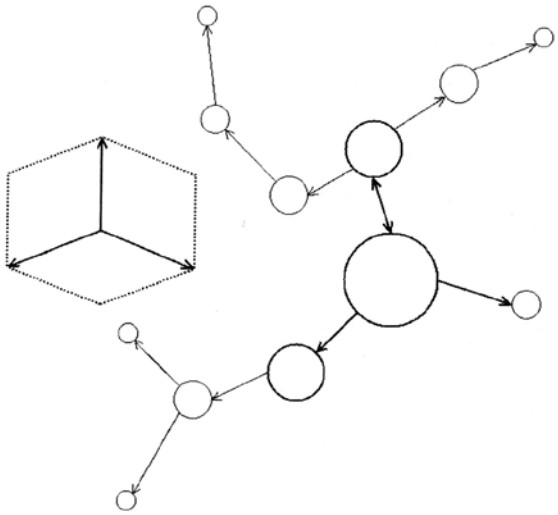
Düğümler arasındaki uzaklık, düğümlerin konular arasında ilişkiyi sağlamadaki rolleriyle ters orantılıdır. (Örnek: Aile ile ilişkilendirilen iki konu, bölgesel ilişki ağına göre kendi içerisinde birbirine daha yakındır). Algoritma, sayısal uzaklığın 100 ile çarpılıp, analizlerde elde edilen maksimum ilişkiye karşılık gelen her rolün ilişki düzeyini kıyaslamak için elde edilen yüzdeden çıkarılmasına dayalıdır. Bu sayede, hızlı bir bakış açısıyla incelenen alanlardaki konuların ve onların ilişki ağlarının birbiri içerisinde ne kadar ilgili olduğu örgü yoğunluğu aracılığıyla incelenebilir.

2 boyutlu gösterim, hem boylamasına hem de enlemesine birbiriyle ilişkili olan konu örgülerinin uzatılmasına dayalıdır (Şekil 4). 3 boyutlu gösterim ise benzer bir yöntemi benimsemektedir. Her konu örgüsü ilişkili konular tarafından çevrelenir; enlemesine ve boylamasına uzatılarak büyütülür. Tüm uzaklık hesaplamaları birbirine eşit ya da benzer olup, tek koşul düğüm örgülerinin 3 boyutlu konumlandırılmasıdır.

Şekil 4. 2 boyutlu gösterim örneği



Şekil 5. 3 boyutlu gösterim örneği



SONUÇ

Hem metodoloji geliştirme hem de sayısal bilgi işleme modelinde semantik Web'teki gereksinimleri karşılayan ve kültürel mirasla ilişkili bilgi ve dokümanlara özgü özellikleri benimsemek; sayısal kütüphaneler bağlamında ortaya çıkan bilgi ile tanımlama için birleştirilmiş bilgiyi kullanmak; konu kategorileri, abecesel dizinler ve esnekliğe indirgenmiş aramalar gibi araştırma kontekstiyle ilgili temel problemleri çözümlemek için muhtemel öneriler getiren tamamen yeni çalışma biçimlerinin özelliklerindendir.

Diğer yanda, tarihsel doküman analizi problemi, başlangıçta sınırlı bir ilk seviye sistemiyle, sonrasında ise ikinci bir seviye sistemiyle oldukça esnek ama belirli bir örüntüye bağlı kalınarak çözülmüştür.

Bu analiz problemi, programlama temelinden dilbilimsel temelli bir yaklaşıma dönüştürülmüş; böylece kültürel miras bilgisinin tanımlanmasında kullanılacak ifade biçimleri belirlenmiştir. Mevcut algoritmada işlenen örüntüler sabit olmayan ve değişkenli bir dil olan İspanyolca ile gerçek-

leştirilmiştir. Açıklanan örüntüleri geliştirmek için daha sabit yapıya sahip diller mevcut olsa da, otomatik doküman analizini başarıya götüren özellik dil değildir.

Metin analizini gerçekleştiren başka teknolojiler bulunsa da, sentezleme, aranan bilgiyi ortaya çıkarma, kataloglama ve anlaşılması güç tarihsel metinlerden bir konu başlığı oluşturma gibi süreçleri gerçekleştirebilen çok az teknoloji mevcuttur. Bahsedilen tüm bu süreçler, metni anlamaya yönelik okumalar aracılığıyla elle gerçekleştirilebilir. Günümüzde, bu teknolojinin kullanılmasını bilgisayarlar yorumlayamayabilir fakat en azından bilgiyi değerlendirerek önemli olanı sunabilir.

Başka bir bakış açısıyla da bu hususu gözlemlemek oldukça dikkat çekicidir. Ayrık dilbilimsel anlatım figürlerine dayalı uygun doküman yazımları, bu figürleri saptamak için düzenlenmiş örüntülerle daha önceden programlanmış analiz algoritmalarından yola çıkarak çok yönlü bir analiz gerçekleştirmeyi mümkün kılar. Fakat bu algoritmanın çözmesi gereken bazı sorunları bulunmaktadır. Bunlar, yazım, imlâ, dil bilgisi ve biçimsel değişkenlerdir. Bu sorunlar, algoritmanın etkinliğini her zaman düşürebilir.

Bununla birlikte, etkili bir XTM temelli bilgi etiketlemesi başarılmış ve bazı sayısal kütüphanelerde hem nesne hem de koleksiyon düzeyinde uyumluluk gösteren diğer üstveri dilleriyle birleştirilmiştir. Bunlara DITA, FOAF, SKOS, MODS ve RDF örnek verilebilir. (Park & Hunting, 2002). XTM'nin XML'e dayalı olduğundaki gibi bilgiye erişim süreçlerinin geliştirilmesi otomatik olarak gerçekleştirilecek ve sayısal bilginin korunması sürecine de yardımcı olacaktır (Premis & RLG/OCLC, 2005). Ayrıca, geliştirilen prototip, kişisel aracılara ve yapay zeka ile hesaplamalara dayalı bilgiye erişim sistemlerini de bünyesine katmakta (Crestani & Passi, 2000) ve sadece GEA kullanıcıları tarafından değil GEA çevrimiçi dışındaki kullanıcılar tarafından da talep edilen bilgiyi etkinleştirmektedir (Herrera & Passi, 2006). Böylece, gerekli olduğu için büyük bir ilgi gören ve zOAC gibi projelerde de kullanılan Açık Erişim Girişimi'nin işlevselliği veya harmanlama yaklaşımına katkıda bulunulmaktadır.

Son olarak 2 ve 3 boyutlu bilgi görseli oluşturmak gibi önemli bir amaç gerçekleştirilmiş, teknolojik cihazlardaki sürekli değişime rağmen sayısal bilginin en iyi halinin sunulmasına gayret edilmiştir.

KAYNAKÇA

Baldonado, M., Chang, C-C.K., Gravano, L., & Paepcke, A. (1997). Metadata for digital libraries: Architecture and design rationale. *Proceedings of the second ACM international conference on Digital libraries* içinde (pp. 47- 56). Philadelphia. New York: ACM.

Berners-Lee, T., & Fischetti, M. (1997). *Weaving the Web*. San Francisco: Physica-Verlag.

Brocks, H., Thie, U., Stein, A.,& Dirsch-Weigand, A. (2001). Customizable retrieval functions based on user tasks in the cultural heritage domain. *Proceedings of the 5th European Conference on Digital libraries*, LNCS 2163 içinde (pp. 37- 48). Heidelberg: Springer.

Chandler, A., Foley, D., & Hafez, A.M. (2000). Mapping and converting essential federal geographic data committee (FGDC) metadata into MARC21 and Dublin Core. Towards an alternative to the FGDC clearinghouse. *D-Lib Magazine*, 6(1).

Crane, G. (2002). Cultural heritage digital libraries: Needs and cornponents. *Proceedings of the 6th European Conference on Digital libraries*. LNCS 2458 içinde (pp. 626637). Heidelberg: Springer.

Crane, G., & Wulfman, C. (2003). Towards a cultural heritage digital library. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries* içinde (pp. 75- 86). Houston. New York: ACM.

Crestani, F., & Passi, G. (2000). Soft computing in information retrieval: Techniques and applica*tions*. Heidelberg: Physica-Verlag.

De Gendt, V., lsaac, A., Van Der Meijt, L., & Schlobach, S. (2006). Semantic Web techniques for multiple views on heterogeneous collections: A case study. *Proceedings of the 10th European Conference on Digital libraries*, LNCS 4172 içinde (pp.426- 437). Heidelberg: Springer.

DITA. (2007). *Darwin information typing architecture OASIS standard*. 28 Ağustos 2008 tarihinde http://dita.xrnl.org adresinden erişildi.

Gelb, J. (2007) Optimizing your content development using topic maps. *Proceedings of DITA North American Conferences*. içinde San Jose, CA.

Foulonneau, M., Cole, T.W., Habing, T.G., & Shreeves, S.L. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. *Proceedings of the 5th ACMIJEEE-CS-joını conference on Digital libraries* içinde (pp. 32- 41). Denver. New York: ACM.

Giunchiglia, F., Yatskevich, M., & Shvaiko , P. (2007). Semantic rnatching: Algorithms and implementation . *Journal on Data Semantics*.

Halbert, M., Kaczmarek, J., & Hagedorn, K. (2003). Findings from the Mellon metadata harvesting initiative. *Proceedings of the 7th European Conference on Digital libraries*, LNCS 2769 içinde (pp. 58- 69). Heidelberg: Springer.

Herrera, E., & Passi, G. (2006). Special issue on soft approaches to information retrieval and information access on the Web. *Journal of the American Society for Information Science and Technology*, ACM Digital library, 57, 511 - 514.

Hutt, A., & Riley, J. (2005). Semantics and syntax of Dublin Core usage in open archives initiative data providers of cultural heritage rnaterials. *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries* içinde (pp. 262- 270). Denver. New York: ACM.

ISO 13250. (2003). lnformation technology - SGML applications - Topic maps.

ISO 21127. (2006). *A reference ontology for the interchange of cultural heritage information*.

Maier, D., Delcambre, L. (1999). *Superimposed information for the lnternet*. ACMSIGMOD Workshop on The Web and Databases. Philadelphia.

Park, J., Hunting, S. (2002). *XML topic maps: Creating and using topic maps for the Web*. Addisson-Wesley.

Premis, W.G ., & Rlg/Oclc, W.G. (2005). *Preservation metadata: implementation strategies.*

Tolosana-Calasanz, R., Nogueras- lso, J., Bejar, R. , Muro-Medrano, P.R., &Zarazaga-Soria, F.J.

(2006). Semantic interoperability based on Dublin Core hierarchical one-to-one mappings. I*nternational Journal of Metadata, Semantics and Ontologies*, 1(3), 183.

ANAHTAR KELİMELER

Dublin Core (DC): Internet'teki kaynakların tanımlamaları için bir üstveri setidir. ISO 15836-2003 Standardı ışığında kaynak tanımlaması için kullanılan 15 tanımlama unsurunu içerir.

Freeling: Dil Analizörleri için Açık Kaynak Paketi (nlp.lsi.upc.edu/freeling)

JADE: Java Agents Development Framework (Java Etmen/Ajan Geliştirme Çerçevesi) (http://jade.tilab.com/)

JDO, JPOX: Java Persistent Objects (JPOX - Java Kalıcı Nesneler) JDO ve Java Veri Nesnelerinin özelliklerinin ücretsiz olarak tam bir şekilde uygulanmasıdır. API, bu kalıcı nesnelerin arayüz odaklı standart bir Java modelidir.

Lucene: Lucene ücretsiz ve açık kaynak kodlu bilgiye erişim kütüphanesidir. (https://lucene.apache.org/)

Machine Readable Cataloging Record (MARC): Makineyle Okuanbilir Kataloglama: Bir bibliyografik kaydın otomatikleştirilmiş biçiminin normalleştirilmiş hali.

Open Archive Cataloguer (zOAC): Açık Arşiv Kataloğu: Bir Web uygulaması olan Zope sunucusunda geliştirilen otomatik üstveri harmanlaması ile bibliyografik kayıtların biraraya toplanması için Açık Veri Girişimi Üstveri Harmanlama Protokolü'nü uygular.

Open Archives Initiative (OAI): Açık Arşivler Girişimi: İçeriğin verimli bir şekilde yayınlanmasını sağlamak için karşılıklı çalışabilirlik standartları geliştirmek ve uygulama girişimi (http://www.openarchives.org).

View publication stats