WEB ARŞİVLEME YAKLAŞIMLARI VE ÖRNEKLERLE WEB ARŞİVLERİ

Esin Sultan OĞUZ*

GİRİŞ

İlk kez 1992 yılında kullanılan WWW (World Wide Web) kısaltması "Dünyayı Saran" Ağ anlamına gelmektedir. Geleneksel bilgi kaynaklarıyla (kitap, dergi vb.) karşılaştırıldığında çok daha yüksek büyüme oranlarına sahip olan WWW kapsamında, yazılı ve görsel içerikli olmak üzere birçok farklı türde materyal yer almaktadır. WWW ortamında yer alan materyalleri erişilebilenler ve erişilemeyenler olmak üzere iki gruba ayırmak mümkündür. "Yüzeysel web" adı verilen, erişimi serbest olan web kaynaklarını barındıran ve "derin web" adı verilen, erişimi kısıtlı web kaynaklarına sahip olan iki boyutuyla web, büyüklüğü ile araştırmacıların merak konusu olmuştur. Yapılan araştırmalar çoğunlukla web büyüklüğüne yönelik kesin bir yargıda bulunmanın oldukça zor olduğunu dile getirmekte ve çoğunlukla Web'in erişilebilen kısmı olarak tanımladığımız yüzeysel web üzerinde yoğunlaşmaktadır. Bu doğrultuda yapılmış araştırmalardan biri arama motorları tarafından erişilebilen web büyüklüğünü her bir arama motoru için değerlendirirken, en çok web belgesine erişen ilk iki arama motorunun 8 milyar sayfa ile Google ve 5 milyar sayfa ile MSN olduğunu belirlemiştir (Sullivan, 2004). Yüzeysel web'de erişilebilen belge miktarına yönelik yapılmış bir diğer araştırma ise 2005 yılı itibariyle web büyüklüğünün 11,5 milyar sayfa olduğunu belirtmektedir (Gulli ve Signorini, 2005). Yüzeysel web'in büyüklüğüne yönelik yapılan tahminler derin web'in büyüklüğünün ne olabileceği sorusunu akla getirmektedir. Erişilemez ya da erişimi kısıtlı olarak tanımladığımız derin web'in büyüklüğüne yönelik ortaya konan tahminler ise çok daha çarpıcıdır. Derin web'in büyüklüğü, 11,5 milyar olarak tahmin edilen yüzeysel web'den 550 kat fazladır (Hellerstein, 2004).

Web'deki bilginin miktarı beraberinde kalıcılığı ile ilgili sorunları da getirmektedir. Lyman (2002) bir web sitesinin ortalama ömrünün 44 gün olduğunu belirtmektedir. Bununla birlikte web

* *Araştırma Görevlisi, Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü.Hacettepe Üniversitesi Beytepe Ankara (esino@hacettepe.edu.tr)*

sitelerinin %44'üne ertesi gün de erişilemediği yapılan araştırmalarla ortaya konmuştur. Lyman ve Varian (2003) tarafından elde edilen veriler ışığında, 2001 yılında üretilmiş yeni bilgilerin %93 oranında anadan doğma sayısal olduğunu da göz önünde bulundurursak, bu bilgilerin gelecek nesillere aktarılması için gerekli uzun süreli korumaya yönelik atılacak adımların söz konusu bilgiler için hayati önem taşıdığını söyleyebiliriz.

WEB ARŞİVLEMENİN SORUNLARI

Web sitelerinin uzun süreli koruma amacıyla toplanmalarında yaşanan sıkıntılara neden olarak gösterilebilecek birçok unsur yer almaktadır. Bunlar içinde web sitelerinin büyüklüğü ve dinamik yapısından kaynaklanan sıkıntıların yarattığı teknik sorunlar ile bir takım yasal ve örgütsel sorunlar yer almaktadır.

Web'e yönelik en genel sorun olarak webin büyük ve sürekli olarak gelişen hacmi gösterilmektedir. Hiçbir arşivleme girişimi, tek başına web'i tamamen toplama işlevini gerçekleştirecek kapasiteye sahip değildir. Web büyüklüğüne yönelik yapılan tahminlerde dikkate alınması gereken en önemli nokta, bu ölçümlerin durağan, arama motorlarının ve kullanıcıların kolayca erişimine açık web sitelerine yönelik olduğudur. Web'in hareketli yapısı ve sürekli değişken doğası web sayfalarının teknik açıdan olabildiğince hızlı ve kapsamlı bilgi toplama için yazılımlar geliştirmek yönünde zorlamaktadır.

Web arşivleri aracılığıyla sağlanan bilgiler için en önemli konulardan birisi de bu bilgilerin telif haklarına ve güvenilirliğine ilişkin konulardır. Yapılan araştırmalar birçok ülkenin mevcut yasalarının web arşivlemeyi desteklemediğini göstermektedir.

Web kaynaklarının bağımsızlığı, bir başka deyişle merkezi bir yapının dışında gelişiyor olması, onunla ilgili konularda sorumluluğun hiç bir kurum ya da kuruluşça üstlenilememesi anlamına gelmektedir. Web arşivleme için de aynı durum söz konusudur. Web arşivlemede standartlar ve politikalar geliştirmekle yükümlü bir kuruluş bulunmamaktadır. Web sitelerinin içeriği ve dağıtımı ile ilgili kararlar, genellikle site sahipleri tarafından verilmektedir.

WEB ARŞİVLEME YAKLAŞIMLARI VE ÖRNEKLER

Elektronik bilgilerin sayısındaki bu artışın yarattığı eğilim, bilgilerin gelecek nesillere aktarılması amacıyla web sayfalarının arşivlenmesi yönünde olmuştur. Yasal, teknik ve örgütsel taraflarıyla oldukça kapsamlı olarak nitelendirebileceğimiz web arşivleme girişimi için uygulanacak temel stratejiler bulunmaktadır. Ulusal ve uluslararası ölçekli olmak üzere yürütülen web arşivleme çalışmaları bu stratejiler

temel alınarak yürütülmektedir. Web arşivleme yaklaşımları olarak adlandırılan bu stratejiler aynı zamanda arşive yönelik uzun vadeli politikaların oluşturulmasında da belirleyici bir unsurdur.

Web kaynaklarının diğer elektronik kaynaklardan ayrılan özellikleri içinde en belirgini kısa süreli oluşlarıdır. Bu özellik web arşivleme çalışmaları için oldukça önemli bir dezavantaj olarak görülmektedir. Bu nedenle Web arşivleme amacıyla geliştirilen yaklaşımların hepsinde, belirlenen hedefe yönelik, en iyi teknikle en çok Web sitesinin arşivlenmesi amacı yatmaktadır. Bu amaçla geliştirilen arşivleme yaklaşımları arasında harmanlama (harvesting), seçimli (selective) yaklaşım, tematik yaklaşım, derleme (deposit) yaklaşımı ve karma (combined) yaklaşım bulunmaktadır.

Harmanlama (harvesting) terimi, İnternet arama motorlarının örümcekleri (crawler) aracılığı ile Web sayfalarını yakalayarak dizinlemelerini ifade etmek veya Web sayfaları üzerinden elektronik posta adreslerinin toplanarak reklâm vb. nedenlerle yığın iletilerin (spam mail) gönderilmesi için kullanılmaktadır (Harvesting, 2005). Web sitelerinin arşivlenmesi sırasında da aynı işlemler gerçekleşir. Arşiv için geliştirilmiş örümcek kendisine verilen komutlar doğrultusunda toplama işlemini yapar. Örneğin, İsveç web arşivleme girişimi olan Kulturarw3 örümceği İsveç alan adlı web sitelerini toplamak üzere programlanmıştır. Harmanlama yaklaşımını kullanan arşivleme girişimleri içinde İnternet Arşivi (The Internet Archive) başta gelmektedir. Kâr amacı gütmeyen bir oluşum olan İnternet Arşivi herhangi bir dil ya da içerik sınırlaması olmaksızın web arşivleme çalışmalarını dünya çapında gerçekleştirmektedir. İlk sayısal koruma çalışması olma özelliğini taşıyan İnternet Arşivi'nin koleksiyon büyüklüğü 55 milyar sayfadır ve her geçen ay 20 terabyte büyümektedir. Sayısal ortamda yer alan web kaynaklarının arşivlenmesinin yanı sıra sayısallaştırma çalışmaları da yapan İnternet Arşivi, Amerikan Kongre Kütüphanesi ve İskenderiye Kütüphanesi ile ortaklaşa gerçekleştirdiği projelerle önemli çalışmalara imza atmıştır. Kütüphane kaynaklı diğer web arşivleme çalışmalarından çeşitli özellikleriyle ayrılan İnternet Arşivi, aynı zamanda Uluslararası Internet Koruma Konsorsiyumu üyesidir (About IA, 2006).

Web arşivleme yaklaşımlarından bir diğeri "seçimli" yaklaşımdır. Bu yaklaşımda web siteleri, belirlenmiş seçim ölçütleri doğrultusunda doğrudan kütüphanecinin katkısı ile toplanır. Böylece arşivdeki her bir parçanın kalitesi onaylanmış olduğundan web arşivi için ulusal bibliyografyanın bir parçası olma şansı doğar. Aynı zamanda belirli bir konuda arşive dâhil edilmesi planlanan web siteleri

tam olarak toplanır (Phillips, 2005). Avustralya web arşivleme girişimi olan PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) web arşivleme için seçimli yaklaşımı kullanan ülkelerin başında gelmektedir. 1996 yılında başlamış olan çalışmaları kapsamında elektronik dergiler, web siteleri ve hükümet yayınları yer almaktadır. PANDORA'da seçilmiş web kaynakları konu başlıkları altında listelenmektedir (Koerbin, 2004).

Web arşivleme için "tematik" yaklaşım bir ölçüde seçimli yaklaşımla benzer özellikler taşımaktadır. Tematik yaklaşımda seçimli yaklaşımda olduğu gibi arşivleme işlemleri seçime dayalı olarak gerçekleşir. Belirli bir konuda oluşturulmuş çekirdek koleksiyon üzerindeki bağlantıların izlenmesi ve bu web sayfalarının arşivlenmesi temeline dayalı tematik yaklaşımda arşivin sürekliliği, bir bakıma çekirdek koleksiyonda yer alan kaynakların sürekliliğine bağlıdır. Amerikan Kongre Kütüphanesi'nin web arşivleme girişimi olan MINERVA arşivinde tematik arşivlere de yer vermektedir. Bunlardan bir tanesi 11 Eylül Web Arşivi, diğeri ise 2002 Seçimleri Web Arşividir. Her iki arşivde de web siteleri İnternet Arşivi'nin örümcekleri ile toplanmaktadır (Schneider, Foot, Kimpton ve Jones, 2003, ss. 1-2).

Web arşivlemede derleme yaklaşımı giderek yaygınlaşmaya başlamıştır. Özellikle web arşivleme girişimlerinin ülkelerin milli kütüphaneleri aracılığıyla yürütülmeye başlaması da bunun altını çizen bir özelliktir. Web arşivlemenin yasal bir hale gelmesi ve kanunlarla desteklenmesi anlamına gelen derleme yaklaşımı geleneksel derleme yaklaşımına benzer. Web sitesi sahibi ya da elektronik bilgi üreticileri web sitesindeki bilgilerin, görüntülerin birer kopyasını kütüphaneye göndererek arşivlenmesine katkıda bulunur. Bu anlayışla Alman Milli Kütüphanesi bünyesinde yürütülen web arşivleme girişimi, The Archive Server olarak adlandırılmaktadır. Kapsamındaki web kaynakları arasında, yayıncıların ve üreticilerin çevrimiçi yayınları (Eylül 2001-), elektronik tezler ve doktora sonrası çalışmalar (1998-), Alman iş yasasının yasal elektronik kopyaları (2000) ve Nazi rejimi süresince yayımlanmış Yahudi dergileri (2005-) yer almaktadır (Day, 2003).

Web arşivleme çalışması yapan kimi ülkeler, hali hazırda kullanılmakta olan Web arşivleme yaklaşımlarından hiç birinin Web ortamındaki sayısal kültürün korunmasında yeterli olamayacağını düşünmektedirler. Bu nedenle, yeni yaklaşımlar geliştirmek yerine, bazı ülkeler tarafından mevcut olanların bir arada kullanılması fikri benimsenmiştir. Literatürde karma yaklaşımlar olarak anılan bu yaklaşım Web arşivlerinde mümkün olabilecek en kapsamlı koleksiyonu oluşturmak amacıyla harmanlama, seçimli, tematik ve

derleme yaklaşımlarının en az ikisinin bir arada kullanılması şeklinde tanımlanmaktadır. Web arşivleme çalışmalarında bu yaklaşımı benimseyen ülkeler içinde Fransa Milli Kütüphanesi başta gelmektedir. Harmanlama ve seçimli yaklaşımı bir arada kullanan Fransa Milli Kütüphanesi, harmanlamanın derleme yaklaşımını desteklediği gerekçesiyle yararlı olduğunu; seçimli yaklaşımın ise derin web olarak adlandırılan alandan web sitesi toplarken kullanıldığını belirtmektedir (Day, 2003).

ULUSLARARASI WEB ARŞİVLEME PROJELERİ

Uluslararası ölçekli web arşivleme çalışmaları, birkaç ülkenin ve/veya özel/tüzel kuruluşların bir araya gelmesiyle oluşturulan işbirliği aracılığı ile yürütülmektedir. Hali hazırda yürütülmekte olan uluslararası projeler arasında The International Internet Preservation Consortium, Networked European Deposit Library Project'in yanı sıra The Nordic Web Archive ve UK Web Archiving Consortium bulunmaktadır. The International Internet Preservation Consortium, Uluslararası İnternet Koruma Konsorsiyumu (IIPC), 24 Temmuz 2003 tarihinde içinde Avustralya, Kanada, Danimarka, Finlandiya, Fransa, İzlanda, İtalya, Norveç ve İsveç gibi ülkelerin milli kütüphanelerinin yanı sıra İngiliz Milli Kütüphanesi'nin, Amerikan Kongre Kütüphanesi'nin ve kâr amacı gütmeyen bir oluşum olan IA'nın bir araya geldiği, 12 katılımcı kuruluşla oluşturulmuş bir konsorsiyumdur (Halgrimsson, 2005). IIPC, üç aşamada ele aldığı hedeflerini Internet ortamında yer alan dünya çapındaki bilgilerin arşivlenmesi, korunması ve zaman içinde erişilmesini sağlamak; uluslararası arşivler oluşturma hedefine yönelik ortak araçlar, teknikler ve standartlar geliştirmek ve milli kütüphanelerin Web'de arşivleme yapmalarını özendirici çalışmalarda bulunmak şeklinde sıralamıştır (About the Consortium, 2006).

1 Ocak 1998 tarihinde düzenlenen Avrupa Milli Kütüphaneler Konferansı ile hayata geçen Networked European Deposit Library Project, Avrupa Derleme Kütüphanesi Ağı Projesi (NEDLIB) projesi, aynı yıl Avrupa Komisyonu Telematik Uygulamalar programınca tahsis edilen bir fonla desteklenmiş ve 2000 yılında tamamlanmıştır. Hollanda Milli Kütüphanesi başkanlığında yönetilen NEDLIB, Fransa, Norveç, Finlandiya, Almanya, Portekiz, İsviçre ve İtalya'nın dahil olduğu bir üye profiline sahiptir. Sekiz ülkenin milli kütüphanelerinin yanı sıra, NEDLIB projesinin katılımcıları arasında, bir milli arşiv, iki bilgi iletişim teknolojileri organizasyonu ve Kluwer Academic, Elsevier Science ve Springer-Verlag gibi yayımcılar da bulunmaktadır (Factsheet, 2001). NEDLIB projesinin başlatılmasındaki temel amaç, Avrupa milli kütüphanelerinin sayısal yayınların yönetiminde ve

yaşatımında ortaklaşa hareket etmeleridir. NEDLIB projesi ile sayısal yayınların derlenmesine yönelik sistem geliştirmeyi sağlayacak ortak bir teknik altyapının oluşturulması ve temel araçların geliştirilmesi hedeflenmektedir. Bu noktada milli kütüphaneler aynı zamanda derleme kütüphanesi olmalarından dolayı önemlidir. Projenin çalışma alanı, milli kütüphanelerin sayısal belgelere yönelik dermelerini geliştirirken karşılaştıkları belli başlı teknik ve yasal konuları kapsamaktadır (Werf-Davelaar, 1999).

Kuzey Ülkeleri Web Arşivi (The Nordic Web Archive), Danimarka, Finlandiya, İzlanda, Norveç, ve İsveç milli kütüphanelerinin aralarında bulunduğu bir işbirliği projesidir. Bu proje ile ülkeler web arşivleme konusundaki deneyimlerini paylaşmayı amaçlamaktadırlar. Birleşik Kraliyet Web Arşivi Konsorsiyumu (UK Web Archiving Consortium), bilimsel, kültürel ve akademik kaynakların gelecek nesillere aktarılmasını sağlamak amacıyla oluşturulmuştur. Hak sahiplerinin izinleri ile seçimli web arşivleme çalışmalarını yürüten bu konsorsiyum, İngiliz Milli Kütüphanesi öncülüğünde yürütülmektedir. The Internet Archive (İnternet Arşivi) diğer örneklerden farklı olarak kâr amacı gütmeyen ticari bir ortaklığın sonucunda ortaya çıkmıştır. Adı geçen uluslararası çalışmaların ortak noktasını ülkelerin ulusal çapta yürüttükleri web arşivleme çalışmalarını uluslararası platformda geliştirmek olarak tanımlamak mümkündür. The Internet Archive, İnternet Arşivi (IA) bir yönüyle diğer uluslararası çalışmalardan ayrılmaktadır. IA'nın hareket noktasında web'i dünya çapında arşivlemek hedefi yatmaktadır.

SONUÇ

Elektronik ortamda üretilen bilginin uzun süreli koruma hedefi, bilginin üretim aşamasından başlayan ve çeşitli aşamalar içeren bir süreçtir. Başarılı bir web arşivleme çalışması, elektronik bilginin yaşam döngüsündeki tüm aşamaların başarı ile yerine getirilmesine bağlıdır. Arşivleme aşamasına geçildiği andan itibaren karşılaşılan teknik, yasal ve örgütsel engellerin aşılmasına yardımcı politikaların ve standartların varlığı önem kazanmaktadır. Ulusal düzeyde yürütülen çalışmalarda milli kütüphaneler aynı zamanda derleme kütüphaneleri olmaları dolayısıyla önemli bir misyon taşımaktadırlar. Bu nedenle konuya dikkat çekecek, gerekli yasal ve örgütsel girişimlerin başlatılması için kamuoyu oluşturması gereken kuruluşların başında milli kütüphaneler gelmektedir. Bununla birlikte uluslararası oluşumlarda yer alarak işbirliği aracılığıyla çözüm arayışları içinde yer almakta fayda vardır.

Kaynakça

About IA [Internet Arşivi Hakkında]. (2006). 25 Mart 2006 tarihinde Internet Archive web sitesinden erişildi: http://www.archive.org/about/about.php

About the Consortium [Konsorsiyum hakkında]. (2006, Ocak). 28 Şubat 2006 tarihinde IIPC web sitesinden erişildi: http://netpreserve.org/about/index.php

Day, M. (2003). *Preserving the fabric of our lives: A survey of Web preservation initiatives*. paper presented at the ECDL 2003: 7th European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norway. 11 Mart 2006 tarihinde http://www.ukoln.ac.uk/metadata/presentations/ecdl2003-day/day-paper.pdf adresinden erişildi.

Factsheet. [Gerçek bilgiler]. (2001, 11 Mart). 1 Nisan 2006 tarihinde NEDLIB web sitesinden erişildi: http://nedlib.kb.nl/index.html

Gulli, A. ve Signorini, A. (2005). The indexable web is more then 11,5 billion pages. 5 Nisan 2006 tarihinde http://www.cs.uiowa.edu/~asignori/web-size/ adresinden erişildi.

Halgrimsson, T. (2005). Special presentation: The International Internet Preservation Consortium (IIPC). 1 Nisan 2006 tarihinde http://consorcio.bn.br/cdnl/2005/HTML/Presentation%20Thorsteinn%20Halgrimsson.htm adresinden erişildi.

Harvesting [Harmanlama]. (2005). 15 Şubat 2006 tarihinde CORDIS web sitesinden erişildi http://lu.com/odlis/odlis_h.cfm

Hellerstein, J. (2004). *Data on the deep Web: Queries, trawls, policies and countermeasures*. 30 Mart 2006 tarihinde http://www.citrisuc.org/research/projects/data_on_the_deep_web_queries_trawls_policies_and_countermeasures adresinden erişildi.

Koerbin, P. (2004, September). *The Pandora Digital Archiving System (PANDAS) Managing Web archiving in Australia: A case study*. Paper presented at the 4th. International Web Archiving Workshop, Bath, UK. 13 Nisan 2006 tarihinde http://www.nla.gov.au/nla/staffpaper/2004/koerbin2.html adresinden erişildi.

Lyman, P. (2002). Archiving the world wide web. *Building a national strategy for preservation: Issues in digital media archiving* içinde. Washington D.C.: Council on Library and Information

Resources. 5 Nisan 2006 tarihinde http://www.clir.org/pubs/reports/pub106/web.html adresinden erişildi.

Lyman, P. (2002). Archiving the world wide web. *Building a national strategy for preservation: Issues in digital media archiving* içinde. Washington D.C.: Council onlibrary and information resources. 5 Nisan 2006 tarihinde http://www.clir.org/pubs/reports/pub106/web.html adresinden erişildi

Lyman, P. ve Varian, H. (2003). How much information? 12 Nisan 2006 tarihinde http://www.sims.berkeley.edu:8000/research/projects/how-much-info-2003/ adresinden erişildi.

Phillips, M. (2005, June). Selective archiving of Web resources: A study of acquisition Costs at the National Library of Australia. *RGL Digi News*. 22 Şubat 2006 tarihinde http://www.rlg.org/en/page.php?Page_ID=20666#article0 adresinden erişildi.

Schneider, S. M., Foot, K., Kimpton, M. ve Jones, G. (2003, Agustos). *Building thematic web collections: Challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive.* 7. European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norweç'de sunulan bildiri. 11 Mart 2006 tarihinde http://bibnum.bnf.fr/ECDL/2003/proceedings.php?f=schneider adresinden erişildi.

Sullivan, D. (2004). *Search engine size wars V Erupts*. 12 Nisan 2006 tarihinde http://blog.searchenginewatch.com/blog/041111-084221 adresinden erişildi.

Werf-Davelaar, T. (1999). Long term preservation of electronic publications. *D-Lib Magazine*, 5 (9). 1 Nisan 2006 tarihinde http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html adresinden erişildi.