اولین همایش ملی جویشگر بومی
۲۷ و ۲۸ بهمن ماه سال ۱۳۹۴
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
وزارت ارتباطات و فناوری اطلاعات
ITRC

آرشیو ملی وب: پایلوت، پیش‌نیازها و استانداردها

محمدمهدی اثنی‌عشری[۱]، طاهره میرسعیدقاضی[۲]، محمد آزادنیا[۳]

[۱] گروه سکوهای فناوری اطلاعات، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران
esnaashari@itrc.ac.ir

[۲] گروه مهندسی دانش، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران
a_ghazi@itrc.ac.ir

[۳] گروه سکوهای فناوری اطلاعات، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران
azadnia@itrc.ac.ir

چکیده

امروزه سامانه‌های آرشیو وب متعددی در سطح جهان، به جمع‌آوری و آرشیو نمودن صفحات وب‌سایت‌های مختلف مبادرت می‌ورزند. این کار، با توجه به ماهیت فرّار وب و تغییرات مداوم در آن، در کنار جذابیت تاریخی خود، امکان دسترسی به اطلاعاتی که در گذشته وجود داشته‌اند، اما امروزه به هر دلیل دیگر در دسترس نیستند را فراهم می‌سازد. به علاوه، امکان انجام تحلیل‌های مختلف آماری روی داده‌های جمع‌آوری شده نیز وجود خواهد داشت. در ایران، چندین سال است که بحث ایجاد یک سامانه آرشیو ملی وب و حفظ و نگاه‌داری تولیدات فکری و ذهنی فارسی‌زبانان در سطح اینترنت مورد توجه قرار گرفته است، اما هنوز اقدامی جدی در زمینه راه‌اندازی چنین سامانه‌ای صورت نپذیرفته است. در مقاله حاضر، سعی شده است که با بررسی پیش‌نیازها و استانداردهای قابل استفاده در یک سامانه آرشیو وب، و با راه‌اندازی یک پایلوت از سامانه آرشیو وب، پیشنهاداتی را برای حصول به یک سامانه واقعی آرشیو وب ارائه گردد.

کلمات کلیدی

آرشیو وب، خزشگر، نمایه‌گذار، جستجوگر، استانداردهای فراداده‌ای، شناسگر دائمی اشیاء رقمی.

۱- مقدمه

نهادهائی مانند کتابخانه‌ها و آرشیوهای ملی برحسب وظایف خود ملزم به گردآوری انتشارات و تولیدات فکری کشور خود هستند [۱]. برای سال‌های متمادی، تنها بستر انتشارات و تولیدات فکری، آثار مکتوب (چاپی و خطی) بود. اما با ظهور کامپیوتر و گسترش روزافزون استفاده از آن در طی چند دهه گذشته، بستر دیجیتالی نیز جای خود را، به عنوان یکی از بسترهای قابل استفاده برای انتشار مطالب و تولیدات فکری، باز نموده است. این مسأله خصوصاً در کنار سادگی استفاده از این بستر و نیز مبحث توسعه پایدار، که استفاده بی‌رویه از منابع طبیعی نظیر درختان را غیرقابل قبول دانسته و در پی استفاده از روش‌هائی است که تهدیدات کمتری برای این منابع ایجد نمایند، باعث شده است که جایگاه بستر دیجیتال به شکل چشم‌گیری ارتقاء یافته و حجم عظیمی از تولیدات فکری بشر تنها در این بستر ارائه گردند. دقیقاً به همین علت، گردآوری، حفاظت و سازماندهی آثار غیرمکتوب نیز به عنوان جزئی از وظایف کتابخانه‌ها و آرشیوهای ملی محسوب می‌گردد.

با توجه به گسترش روزافزون اینترنت، بخش عظیمی از محتوای دیجیتالی مذکور در محیط وب منتشر می‌شود. اما متأسفانه محیط وب کاملاً ناپایدار است و مستندات موجود در آن در هر لحظه ممکن است به دلایل گوناگونی (از جمله پائین آمدن وب‌سایت حاوی مستندات، برداشته شدن مستندات توسط تولید کننده آنها، تغییر آدرس URI مستندات) از بین بروند. به همین دلیل، لازم است که انباره‌ای سازمان‌یافته از منابع موجود در وب ایجاد شود تا امکان نگاه‌داری طولانی مدت از این مستندات فراهم گردد. در این انباره، نسخه‌های متعددی از هر وب‌سایت نگهداری می‌گردند که هر یک نسخه‌ای از آن وب‌سایت در مقطع زمانی مشخص خواهد بود. بدین ترتیب، امکان دسترسی به اسناد و مستنداتی که در گذشته بر روی یک وب‌سایت قرار گرفته و اکنون وجود ندارند، فراهم خواهد گردید. این انباره منابع وب را به صورت یکجا در اختیار محققان فعلی و نسل‌های آینده قرار می‌دهد و از این لحاظ منبع غنی از داده محسوب می‌گردد. به این انباره آرشیو وب گفته می‌شود.

مقاله حاضر در پی بیان چرایی نیاز به در اختیار داشتن یک آرشیو ملی از وب و شناسایی تکنیک‌ها، پیش‌نیازها و استانداردهایی است که برای راه‌اندازی چنین آرشیوی مورد نیاز می‌باشد. در این راستا، و به منظور تسلط هرچه بیشتر بر جزئیات فنی و تکنیکی مورد نیاز، پایلوت کوچکی نیز ایجاد گردیده که در آن، وب‌سایت‌های مرکز تحقیقات مخابرات ایران، کتابخانه ملی و خبرگزاری تابناک برای برهه زمانی کوتاهی آرشیو شده‌اند. نتایج حاصل از ایجاد این پایلوت و نیز پیشنهاداتی در زمینه راه‌اندازی آرشیو ملی وب در این مقاله ارائه گردیده‌اند.

ادامه این مقاله به صورت زیر سازماندهی شده است. ابتدا در بخش دوم به بیان تبیین ضرورت ایجاد یک سامانه آرشیو ملی وب پرداخته خواهد شد. بخش سوم به معرفی مختصری از ساختار

کلی یک سامانه آرشیو وب اختصاص خواهد داشت. ابزارهای مورد نیاز جهت راه‌اندازی یک سامانه آرشیو وب در بخش چهارم بیان خواهد گردید. در بخش پنجم، استانداردهای فراداده‌ای مورد نیاز در یک سامانه آرشیو وب معرفی شده‌اند. توصیف مختصری از پایلوت راه‌اندازی شده در بخش ششم ارائه خواهد شد. در نهایت، بخش هفتم حاوی جمع‌بندی، نتیجه‌گیری و ارائه پیشنهاد برای فعالیت‌های آتی در این خصوص خواهد بود.

۲- چرا آرشیو ملی وب لازم است؟

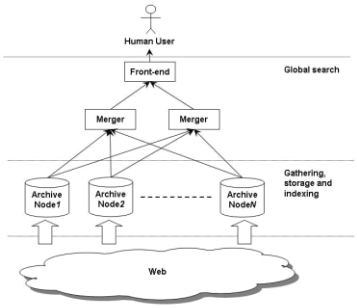
اولین و مهم‌ترین آرشیو وب موجود در سطح دنیا، انباره نگاه‌داری شده توسط سازمان غیرانتفاعی Internet Archive است که از سایت www.archive.org قابل دسترسی است. این انباره از سال ۱۹۹۶ به آرشیو کردن کل محتوای موجود در سطح اینترنت پرداخته و تا کنون بیش از ۲ پتابایت حجم داده را (حجم داده در حالت فشرده) آرشیو نموده است. بنابراین، شاید این طور به نظر برسد که با وجود انبارههائی از این دست در سطح دنیا، چه نیازی به ایجاد یک انباره موازی بدین شکل در سطح کشور وجود دارد. دلایل متعددی را در این زمینه می‌توان برشمرد که در زیر به برخی از آنها اشاره می‌شود:

- **سیاست‌گذاری‌های آرشیو:** آرشیوهای وب موجود در سطح دنیا بر اساس اهداف و سیاست‌گذاری‌های مختلفی عمل می‌نمایند. یک وب‌سایت یا یک صفحه وب مشخص ممکن است از نظر یک آرشیو کننده کم اهمیت باشد، اما از نظر آرشیو کننده دیگری مهم محسوب گردد. با توجه به حجم بالای مطالب موجود در وب، آرشیو کنندگان امکان آرشیو کردن کلیه محتوای موجود در هر وب‌سایت را ندارند. آنها مجبورند صفحات مهم‌تر را مشخص کرده و آرشیو نمایند. همچنین، تکرار آرشیو از یک وب‌سایت یا صفحه وب مشخص نیز تابع سیاست‌های آرشیو کننده خواهد بود. با توجه به آنکه هیچ یک از آرشیوهای وب موجود در سطح دنیا برای سیاست‌گذاری‌های خود منطبق بر ویژگی‌های ملی و فرهنگی ما تعهدی ندارند، آرشیوهای تهیه شده توسط آنها جوابگوی کلیه نیازمندی‌های ملی و فرهنگی ما نخواهد بود.
- **عدم دسترسی به کل آرشیو:** با توجه به آنکه آرشیوهای موجود در سطح دنیا توسط مؤسساتی نگاه‌داری می‌شوند که ارتباط تجاری خاصی با کشور ما ندارند، امکان دسترسی به کل آرشیو آنها وجود ندارد. استفاده از این آرشیوها تنها در نقش یک کاربر امکان‌پذیر است و بدیهی است که برای یک کاربر معمولی، محدودیت‌های زیادی برای دسترسی به کل محتوا وجود دارد.
- **احتمال جلوگیری از دسترسی به آرشیو:** به دلیل آنکه آرشیوهای موجود در سطح دنیا در اختیار ما نیستند، این احتمال همواره وجود دارد که بنا به هر دلیلی، نظیر تحریم‌های یک‌جانبه بر علیه نظام مقدس جمهوری اسلامی ایران، از بین رفتن آرشیو به دلایل مختلف، قطعی اینترنت و ...، امکان استفاده از آنها برای ما از دست برود.
- **عدم وجود کنترل بر آرشیو:** آرشیوهائی که توسط مؤسسات و سازمان‌هائی خارج از ایران نگاه‌داری می‌شوند کاملاً از کنترل ما خارج بوده و امکان هیچ‌گونه نظارتی بر نحوه عملکرد آنها وجود ندارد. همین عدم امکان نظارت باعث شده است که بسیاری از صفحات وب‌سایت Internet Archive در حال حاضر در ایران فیلتر گردد، زیرا نظارتی بر محتوای نگاه‌داری شده در این آرشیو وجود نداشته و متأسفانه محتویات غیراخلاقی زیادی در آن قرار دارد.

با توجه به جمیع موارد فوق، به نظر می‌رسد که در حال حاضر، ایجاد یک آرشیو ملی از وب نیاز اساسی و فوری برای کشور می‌باشد. از سوی دیگر، آرشیو وب یک منبع غنی محتوایی است که با در اختیار داشتن آن، می‌توان به انجام تحلیل‌های کلان فرهنگی، سیاسی، اجتماعی و اقتصادی مبادرت ورزید. بدیهی است که هرچه زمان ایجاد این آرشیو بیشتر به تعویق بیفتد، بخش بیشتری از تولیدات فکری کشور، که در حال حاضر در اینترنت قابل دسترسی هستند، از سطح اینترنت خارج شده و امکان دسترسی به آنها در آینده برای هیچ‌کس وجود نخواهد داشت.

۳- ساختار کلی یک سامانه آرشیو وب

یک سامانه آرشیو وب به صورت کلی ساختاری مشابه ساختار ارائه شده در شکل ۱ دارد. مطابق با این شکل، در پائین‌ترین لایه، تعدادی گره خزشگر قرار دارند که به جستجو در سطح اینترنت و خزش اطلاعات پرداخته و این اطلاعات را ذخیره‌سازی می‌نمایند. با توجه به گستردگی وب، وجود تنها یک گره خزشگر نمی‌تواند جوابگوی فعالیت جمع‌آوری اطلاعات از سطح وب باشد. به همین دلیل، عموماً در سامانه‌های آرشیو وب، چندین گره به صورت موازی و در یک ساختار توزیع شده به خزش در وب و جمع‌آوری اطلاعات از سطح وب می‌پردازند.



شکل ۱: ساختار کلی یک سامانه آرشیو وب

به منظور آنکه بتوان اطلاعات ذخیره شده را در اختیار کاربر قرار داد، نمایه‌سازی بر روی اطلاعات صورت می‌پذیرد. کاربر از طریق یک واسط کاربری (Front-end) به اطلاعات ذخیره شده و نمایه‌سازی شده دسترسی پیدا می‌نماید. با توجه به آنکه اطلاعات جمع‌آوری شده بر روی گره‌های مختلفی قرار گرفته‌اند، برای ارائه به کاربر، لازم است که این اطلاعات با یکدیگر ترکیب شده(توسط Mergerها) و سپس به کاربر ارائه گردند. بدیهی است که کاربر برای دریافت اطلاعات خود نیازمند جستجو بر روی مجموعه عظیمی از اطلاعات است که توسط خزشگرها جمع‌آوری شده‌اند. در آرشیوهای مختلف سیاست‌های متفاوتی برای جستجو مد نظر قرار گرفته‌اند که از آن جمله می‌توان به موارد زیر اشاره نمود:

- **جستجوی مبتنی بر URL:** ساده‌ترین نوع جستجو بر روی آرشیو وب جستجوی مبتنی بر URL است. در این روش، کاربر یک URL مشخص را به سامانه آرشیو وب ارائه می‌کند و سامانه، لیستی از زمان‌هائی را که آن URL در آرشیو خزش شده است در اختیار کاربر قرار می‌دهد. سپس کاربر با انتخاب زمان مد نظر خود، می‌تواند نسخه آرشیو شده از آن URL را در آن زمان مشاهده نماید.
- **جستجوی کلیدواژه:** روش جستجوی معمول در جویشگرهاست.
- **مرور موضوعی:** فهرستی از موضوعات لیست شده‌اند و کاربر می‌تواند با انتخاب یک موضوع، به صفحات مرتبط با آن موضوع دست یابد.
- **مرور الفبائی:** دسترسی به صفحات بر اساس حروف الفبا صورت می‌گیرد.

۴- ابزارهای مورد نیاز

دو زیربخش عمده یک سامانه آرشیو وب، یکی خزشگر و دیگری نمایه‌گذار و جستجوگر هستند. در ادامه، در خصوص هریک از این دو بخش و ابزارهای قابل استفاده برای پیاده‌سازی هریک از آنها صحبت خواهیم نمود.

۴-۱- خزشگر

سامانه‌های خزش متعددی در سطح دنیا توسط شرکت‌ها و سازمان‌های مختلف با اهداف گوناگونی توسعه داده شده‌اند. این سامانه‌ها عموماً با دریافت تعدادی لینک اولیه، دریافت صفحات وب را آغاز نموده و با دریافت هر صفحه، لیست لینک‌های موجود در آن صفحه را به صف دریافت خود می‌افزایند. این پروسه تا زمانی که لینک جدیدی در صف دریافت وجود نداشته باشد ادامه می‌یابد. با توجه به گستردگی وب و تعداد بالای صفحات موجود در آن، اغلب این سامانه‌ها امکانات متعددی برای محدودسازی خزش دارند. به عنوان نمونه‌هائی از این محدودیت‌ها می‌توان به محدودسازی خزش به دامنه‌های خاص، محدود کردن عمق دنبال‌کردن لینک در هر سایت، حذف لینک‌های دارای الگوی خاص، و ... اشاره نمود.

هدف بسیاری از سامانه‌های خزش موجود، خزش وب به منظور ارائه اطلاعات بروز شده در قالب یک جویشگر است. اما تفاوتی که وجود دارد آن است که در یک سامانه آرشیو وب، نسخه‌های متعددی از هر URI نگاه‌داری می‌شوند که هر نسخه وضعیت آن URI را در یک مقطع زمانی مشخص نشان می‌دهد. بر خلاف جویشگر، که بروز بودن اطلاعات در آن اهمیت بالائی دارد، در یک سامانه آرشیو وب، کامل بودن نسخه‌ها از حیث دارا بودن اکثر تغییرات هر URIاهیمت ویژه‌ای دارد.

بهترین مرجع برای دریافت مشورت، اطلاعات و انتخاب سامانه‌های مرتبط با آرشیو وب کنسرسیوم بین‌المللی نگاه‌داری اینترنت (IIPC) است که وب‌سایت آن http://netpreserve.org است. این کنسرسیوم مجموعه مؤسسات، سازمان‌ها، کتابخانه‌ها و شرکت‌هائی را در خود جای داده است که هریک با هدف‌گذاری خاص خود، به آرشیو تمام یا بخشی از اینترنت می‌پردازند. با توجه به آنکه اعضای این کنسرسیوم عموماً جزء بهترین‌های

فعالیت در این حوزه هستند، ابزارها و سامانه‌های پیشنهاد شده توسط آن معتبرترین پیشنهادهای ممکن خواهند بود. لیست ابزارهای معرفی شده توسط این کنسرسیوم به منظور انجام خزش در یک سامانه آرشیو وب به قرار زیر هستند:

- Heritrix: ابزاری است که توسط Internet Archive توسعه داده شده است. می‌توان گفت که کامل‌ترین ابزار موجود به منظور انجام خزش در یک سامانه آرشیو وب همین ابزار است. اغلب اعضای کنسرسیوم IIPC به منظور خزش و تولید آرشیو خود از این ابزار و یا ابزارهائی که مبتنی بر آن توسعه داده شده‌اند (و در ادامه معرفی خواهند شد) استفاده کرده‌اند.
- NetarchiveSuite: ابزار NetarchiveSuite را می‌توان عملاً یک واسط کاربری مناسب برای Heritrix دانست که اجازه استفاده ساده از Heritrix را می‌دهد.
- CINCH: ابزاری است که با دریافت لیستی از URLها، فایل‌های مشخص شده توسط این URLها را دریافت می‌نماید.
- HTTrack: این نرم‌افزار برای تولید یک کپی از یک وب‌سایت می‌تواند مورد استفاده قرار گیرد.
- WARCreate: افزونه‌ای بر مرورگر Chrome است که همانند HTTrack می‌تواند یک کپی از یک وب‌سایت را تولید نماید. کپی ایجاد شده توسط WARCreate در قالب فرمت Web ARChive (WARC) ذخیره می‌شود. این فرمت، فرمت استاندارد ذخیره‌سازی آرشیوهای وب است که Heritrix نیز از آن استفاده می‌کند.
- WebSite-Watcher: ابزاری است که با دریافت یکURL ، صفحه مرتبط با آن URLرا به همراه تمامی اجزاء آن (تصاویر، اسکریپت‌ها و ...) دانلود نموده و ذخیره می‌نماید.

از میان ابزارهای فوق، ابزار Heritrix برای ایجاد یک سامانه آرشیو پایلوت مورد استفاده قرار گرفت که در بخش ششم در مورد آن صحبت خواهد شد.

۴-۲- نمایه‌گذار و جستجوگر

لیست ابزارهای معرفی شده توسط کنسرسیوم IIPC جهت نمایه‌گذاری شامل ابزارهای زیر است:

- Wayback Machine: ابزاری است که توسط Internet Archive توسعه داده شده و مورد استفاده قرار گرفته است. اغلب اعضای کنسرسیوم IIPC به منظور نمایه‌گذاری و جستجو بر روی آرشیو خود از این ابزار و یا ابزارهائی که مبتنی بر آن توسعه داده شده‌اند (و در ادامه معرفی خواهند شد) استفاده کرده‌اند.
- Warrick: این ابزار توسط Frank McCown در دانشگاه Old Dominon توسعه داده شده و هدف از آن فراهم آوردن امکان بازیابی یک سایت به صورت کامل از روی آرشیو وب آن در Internet Archive است.
- WCT: ابزار Web Curator Tool که اختصاراً WCT نامیده می‌شود، ابزار قدرتمندی است که Heritrix و Wayback Machine را با یکدیگر ترکیب نموده، واسط کاربری مناسبی برای آنها ایجاد نموده تا کاربران معمولی، خصوصاً کتابدارانی که باید آرشیوی از وب تهیه نمایند، راحت‌تر بتوانند از آنها استفاده نمایند و در نهایت، امکان افزودن فراداده دوبلین کور (در ادامه معرفی خواهد شد) به آرشیو وب ایجاد شده را فراهم نموده است.
- JWAT: ابزاری به زبان جاوا با عنوانJava Web Archive Toolkit است که از آن می‌توان برای خواندن فایل‌های WARC استفاده نمود.
- WAT: ابزار Web Archive Transformation Format توانائی استخراج فراداده از فایل‌های WARC و تبدیل آنها به فرمتی که قابل نمایه‌گذاری توسط ابزارهای دیگری نظیر Hadoop باشد را فراهم می‌نماید.
- WarcManager: ابزاری است که امکان مرور و بررسی محتویات فایل‌های WARC را فراهم می‌کند.
- WARC Tools: این ابزار قابلیت خواندن محتویات فایل‌های WARC و در صورت لزوم اِعمال تغییرات در آنها را فراهم می‌آورد. همچنین، توانائی تبدیل فایل‌های ARC به WARCرا دارد.
- NutchWax: امکان نمایه‌گذاری و جستجو بر روی فایل‌های آرشیو از طریق موتور جستجوی Nutch را فراهم می‌آورد.
- WERA: این ابزار قدرتمند کاملاً مشابهWayback Machine ارائه شده توسط Internet Archiveعمل می‌کند و تنها تفاوت آن با Wayback Machine این است که امکان جستجوی مبتنی بر کلمه را نیز ارائه می‌نماید.

از میان ابزارهای فوق، ابزار Wayback Machine برای ایجاد یک سامانه آرشیو پایلوت مورد استفاده قرار گرفت که در بخش ششم در مورد آن صحبت خواهد شد.

۵- استانداردهای فراداده‌ای

لازمهٔ دسترسی به منابع گردآوری شده در انبارهٔ یک آرشیو وب سازماندهی مناسب است. سازماندهی منابع وبی علیرغم تفاوتهای عمدهٔ این دسته از منابع با انتشارات چاپی، در کلیت خود تابع اصول و مفاهیمی‌مشترک با منابع غیر وبی است. این سازماندهی نیازمند استفاده از فراداده‌هایی است که بتوانند منابع را به گونه‌ای مناسب و استاندارد توصیف کنند [۲]. فراداده‌ها به اشکال مختلف می‌توانند در سازماندهی آرشیو وب نقش ایفا کنند [۳]:

- فراداده می‌تواند بازیابی را با ایجاد بستری برای توصیفگرهای خاص، بهبود بخشد. به عنوان مثال کلمه «سبز» در فیلد مؤلف، نام یک شخص را نشان می‌دهد، در حالی که «سبز» در عنوان یک مدرک، ممکن است واژه‌ای جهت بازیابی موضوعی باشد. برچسب‌های فراداده‌ای مناسب در اطراف عناصر داده‌ای متفاوت، این امکان را به جویشگرها می‌دهند که اطلاعات را در یک مسیر قابل شناسایی‌تر جستجو کنند.
- فراداده‌ها راهی را برای نگاه‌داری پیشینه اسناد وب و تغییرات اِعمال شده در آنها فراهم می‌سازند. زمانی که شیء رقمی، برای آخرین بار بررسی یا روزآمد شده باشد، این نظامها مسئولیت ایجاد آن را بر عهده داشته و یا شرایط دسترسی به آن را ایجاد می‌کنند.
- فراداده می‌تواند به تصمیم‌گیری درباره اعتبار داده کمک کند. فراداده شاهد و گواهی برای منشأ و منبع ارائه داده است و این زیربنای حاکمیت، شفافیت و مسئولیت را تشکیل می‌دهد. این مسئله به طور فزاینده‌ای برای بسیاری از سازمان‌هایی که بیشتر به پیشینه‌های الکترونیکی متکی هستند تا به فایل های کاغذی، دارای اهمیت است. این امر از آن جهت که مشخص می‌کند، مدرک الکترونیکی به لحاظ امنیتی محفوظ بوده، و پیشینه نیز کامل است و در آن دخل و تصرفی نشده، ضروری است.
- فراداده کلیدی برای میانکنش‌پذیری است و می‌تواند فرایند انتقال داده‌ها از یک سامانه به سامانه دیگر را تسهیل نماید.

طبق بررسی‌های انجام شده در [۲]، معروف‌ترین فراداده‌هایی که در آرشیوهای وب موجود در دنیا مورد استفاده قرار گرفته‌اند عبارتند از:

- فراداده‌های توصیفی: XML MARC، Doublin Core، MODS، MADS، EAD، RDF و METS.
- فراداده‌های مدیریتی: TextMD، MIX، TEI و فراداده حفاظتی PREMIS.
- فراداده‌های میانکنش‌پذیری (پروتکل‌های مبادله اطلاعات): OAI-PHM، SRU/W و Z39.5.

نکته دیگری که باید در سازماندهی یک آرشیو وب مورد توجه قرار گیرد، استفاده از آدرس‌هایی به عنوان شناسگر یک سند است که برای مدت زمان طولانی و نامحدود تداوم داشته باشند. روال آدرس‌دهی فعلی در اینترنت که مبتنی بر URI است، این قابلیت را ندارد. زیرا با جابجایی یک سند در داخل پوشه‌های وب‌سایت، آدرس قبلی از درجه اعتبار ساقط خواهد گردید. شناسگر باید طوری تعریف شود که تا زمانی که سازمان مربوطه پابرجاست، وجود داشته باشد و محدود به مکانی خاص یا فرآیندی خاص نباشد. به هنگام جابجایی اسناد از یک مکان به مکان دیگر یا از یک محمل ذخیره به محملی دیگر همچنان کاربر بتواند آنها را بیابد [۴]. بدین منظور، تا کنون نظام‌های متعدد شناسگر اشیاء رقمی تعریف شده‌اند که برخی از معروف‌ترین آنها عبارتند از: DOI، PURL، URN، ARK و XRI [۵].

۶- پایلوت

به منظور تسلط هرچه بیشتر به جزئیات فنی و تکنیکی مورد نیاز برای راه‌اندازی یک سامانه آرشیو ملی وب، پایلوتی ایجاد گردید که در آن، صفحات سه وب‌سایت ۱) مرکز تحقیقات مخابرات ایران، ۲) کتابخانه ملی و ۳) خبرگزاری تابناک برای بازه زمانی محدودی آرشیو شدند. به منظور ایجاد آرشیو، از خزشگر Heritrix استفاده شد. جدول ۱ نتایج حاصل را ارائه می‌کند. لازم به ذکر است که به طور متوسط، پهنای باندی در حدود ۷۶۰ کیلو بیت بر ثانیه برای خزش‌ها مورد استفاده قرار گرفته است. علت بالا بودن نسبی پهنای باند مصرفی، قرار گرفتن سامانه آرشیو کننده و وب‌سایت مرکز تحقیقات مخابرات ایران روی یک شبکه محلی و لذا استفاده از پهنای باند شبکه محلی برای آرشیو این وب‌سایت بوده است. به منظور فراهم‌سازی امکان دسترسی به آرشیوهای ایجاد شده، از ابزار Wayback Machine، که تنها امکان جستجو از طریق URL را فراهم می‌سازد، استفاده گردید.

جدول ۱: نتایج حاصل از آرشیوهای پایلوت

ردیف	پارامترهای خزش	وب‌سایت‌های آرشیو شده	زمان تکمیل خزش	تعداد لینک‌های خزش شده	حجم آرشیو نهائی (به صورت فشرده)
۱	پارامترهای پیش فرض Heritrix	www.itrc.ac.ir	۶ شبانه‌روز	۶۲۷,۱۰۰	۱۰GB
۲	عدم رعایت مکانیزم مؤدب بودن	www.itrc.ac.ir	۴۸ ساعت	۶۲۷,۱۰۰	۱۰GB
۳	عدم رعایت مکانیزم مؤدب بودن، حذف لینک‌های بی‌اهمیت	www.itrc.ac.ir	۱ ساعت و ۴۰ دقیقه	۲۵,۶۲۱	۱GB
۴	عدم رعایت مکانیزم مؤدب بودن، حذف لینک‌های بی‌اهمیت	www.itrc.ac.ir www.nlai.ir	۲۴ ساعت	۴۸,۰۹۷	۳/۲GB
۵	عدم رعایت مکانیزم مؤدب بودن، حذف لینک‌های بی‌اهمیت	www.itrc.ac.ir www.nlai.ir www.tabnak.ir	بیش از ۴ روز	۴۶۸,۸۸۹	۴۰GB
۶	عدم رعایت مکانیزم مؤدب بودن، حذف لینک‌های بی‌اهمیت، خزش افزایشی	www.itrc.ac.ir www.nlai.ir	۲۴ ساعت	۴۸,۰۹۷	۵۴۷MB
۷	عدم رعایت مکانیزم مؤدب بودن، حذف لینک‌های بی‌اهمیت، خزش افزایشی	www.itrc.ac.ir www.nlai.ir www.tabnak.ir	بیش از ۴ روز	۴۶۸,۸۸۹	۲۴GB

۷- جمع‌بندی، نتیجه گیری و ارائه پیشنهاد

با توجه به جمیع مواردی که در بخش‌های قبل بیان گردید، تهیه یک سامانه آرشیو ملی وب، یکی از حیاتی‌ترین اقداماتی است که باید در اسرع وقت بدان پرداخته شود تا بتوان تولیدات فکری بومی را که در سطح وب منتشر می‌شوند، آرشیو نموده و برای استفاده نسل حاضر و آتی نگاه‌داری نمود. در این بخش پیشنهاداتی در خصوص ابزارهای قابل استفاده، حداقل پیکربندی سخت‌افزاری و پهنای باند مورد نیاز، و در نهایت، استانداردهای فراداده‌ای قابل استفاده به منظور ایجاد و راه‌اندازی آرشیو ملی وب ارائه خواهد گردید.

۷-۱- ابزارهای قابل استفاده

به منظور راه‌اندازی یک آرشیو وب، دو دسته از ابزارها مورد نیاز هستند که عبارتند از: ابزارهای خزش وب و ابزارهای نمایه‌گذاری و فراهم آوردن امکان جستجو. در زمینه ابزارهای خزش وب، Heritrix گزینه مطلوبی است، اما با توجه به آنکه این ابزار واسط کاربری مناسبی ندارد، ابزار NetarchiveSuite در این زمینه پیشنهاد می‌گردد که در بدنه اجرائی خود از Heritirx استفاده کرده و به همین دلیل، تمام ویژگی‌ها و توانمندی‌های Heritrix را دارد، و به علاوه، با ارائه یک واسط کاربری ساده و مناسب، نقطه ضعف اصلی Heritirx را نیز مرتفع ساخته است. در زمینه ابزارهای نمایه‌گذاری و جستجو، استفاده از ابزار WERA پیشنهاد می‌شود که تمامی قابلیت‌های Wayback Machine را دارد و به علاوه، امکان جستجوی مبتنی بر کلمه را نیز فراهم می‌سازد. در کنار این دو پیشنهاد، همچنین می‌توان از ابزار Web Curator Tool استفاده نمود که Heritrix و Wayback Machine را در کنار یکدیگر قرار داده و به علاوه، امکان افزودن فراداده Doublin Core را به صورت اتوماتیک نیز فراهم نموده است.

۷-۲- نیازمندی‌های سخت‌افزاری

بر اساس تجربیات به دست آمده از پایلوت ایجاد شده، زیربخش تهیه خزش سربار قابل توجهی را بر سخت‌افزار تحمیل نمی‌کند، اما طبیعتاً نیازمند فضای دیسک سخت بالایی است. در این زمینه، با توجه به اینکه حجم آرشیو تهیه شده توسط مؤسسه Internet Archive از سال ۱۹۹۶ تا انتهای سال ۲۰۱۲ برابر با ۱۰ پتابایت (۱۰,۰۰۰ ترابایت) بوده است و سالانه نیز در حدود ۱۰۰ ترابایت افزایش می‌یابد، در نظر گرفتن فضائی در حدود لااقل ۵۰۰ ترابایت برای آرشیو ملی وب منطقی به نظر می‌رسد. بر عکس، زیر بخش نمایه‌گذاری و جستجو از لحاظ حافظه RAM و پردازنده سربار نسبتاً بالایی را اِعمال می‌نمایند که در طراحی سخت‌افزاری باید لحاظ گردد.

۷-۳- استانداردهای فراداده‌ای پیشنهادی

مطالعات و پژوهشهای صورت گرفته در سطح بین المللی نشان می‌دهند که هیچ یک از استانداردهای فراداده‌ای بر اساس کارکردی خاص به تنهایی قادر به پاسخگویی نیازهای بومی و محلی بافتی خاص (جغرافیایی، موضوعی، زبانی و ...) نیستند. لذا، لازم است که ترکیبی از این استانداردها در آرشیو ملی وب مورد استفاده قرار گیرند. در این زمینه، با توجه به بررسی‌های صورت پذیرفته روی آرشیوهای ملی در سراسر دنیا و مطالعاتی که در [۲] و [۳] انجام شده است، پیشنهادهای زیر قابل ارائه هستند:

- فراداده‌های توصیفی: به کارگیری استاندارد METS، فراداده توصیفی آرشیوی EAD، فراداده توصیفی شی‌گرای (MODS) و فراداده توصیفی مستندات (MADS) پیشنهاد می‌شوند.
- فراداده‌های مدیریتی: به کارگیری فراداده‌های PREMIS، TEI و MIX پیشنهاد می‌شود.
- فراداده‌های میانکنش‌پذیری: استفاده از استانداردهای پروتکل طرح آرشیو باز (OAI-PMH) (برای پشتیبانی از مدل تعامل مجتمع) و نیز پروتکل (SRU/W) پیشنهاد می‌شوند.

همچنین، در زمینه نظام شناسگر اشیاء رقمی، با توجه به رایگان بودن مدل ARC و نیز وجود ساختار سلسله‌مراتبی آدرس‌دهی در آن، که امکان مدیریت توزیع‌شده آدرسی‌دهی را فراهم می‌سازد، استفاده از این نظام پیشنهاد می‌گردد.

مراجع

۱ سازمان اسناد و کتابخانه جمهوری اسلامی ایران، "قانون اساسنامه سازمان اسناد و کتابخانه جمهوری اسلامی ایران"، تهران: سازمان اسناد و کتابخانه جمهوری اسلامی ایران، ۱۳۸۶.

۲ سازمان اسناد و کتابخانهٔ ملی ج ا ا، شورای عالی پژوهش، گروه پژوهشهای توسعه ای فناوری اطلاعات، گزارش نهایی، طرح پژوهشی آرشیو وب ایران (فاز اول): امکان سنجی ایجاد آرشیو وب در سازمان اسناد و کتابخانه ملی ج ا ا.

۳ دکتر میترا صمیعی، "بررسی و تحلیل فرمتها و استانداردهای مختلف در حوزه ذخیره سازی محتوای الکترونیکی متنی"، گزارش فنی: مستخرج از پروژه: طراحی و پیاده سازی و استقرار چارچوبی برای ارائه خدمات به اشتراک گذاری محتوای الکترونیکی میان سازمان‌ها، مبتنی بر استانداردهای موجود، ۹۰/۱۱/۰۸.

۴ فاطمه، نبوی، " کتابخانه دیجیتالی: مبانی نظری، محتوا، ساختار، سازماندهی، استانداردها و هزینه ها"، صص۵، ۱۳۸۴.

۵ سعیده اکبری داریان ، " آرشیو وب ایران و شناسگر دائمی‌اشیای رقمی"، مقالات پوستری، مجموعه مقالات نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب، سازمان اسناد و کتابخانه ملی جمهوری اسلامی‌ایران، تهران، دوم اسفند ۱۳۹۱.