



استاندارد ملی ایران

۲۱۷۱۲

چاپ اول

۱۳۹۵



INSO

21713

1st.Edition

2017

Identical with
ISO 14873:2013

جمهوری اسلامی ایران
Islamic Republic of Iran
سازمان ملی استاندارد ایران

Iranian National Standardization Organization

اطلاعات و مستندسازی—مباحث آماری
و کیفی مربوط به آرشیوسازی وب

Information and documentation—
Statistics and quality
issues for web archiving

ICS: 01.140.20

سازمان ملی استاندارد ایران

تهران، ضلع جنوب غربی میدان ونک، خیابان ولیعصر، پلاک ۲۵۹۲

صندوق پستی: ۱۴۱۵۵-۶۱۳۹ تهران - ایران

تلفن: ۸۸۸۷۹۴۶۱-۵

دورنگار: ۸۸۸۸۷۱۰۳ و ۸۸۸۸۷۰۸۰

کرج ، شهر صنعتی، میدان استاندارد

صندوق پستی: ۳۱۵۸۵-۱۶۳ کرج - ایران

تلفن: ۰۲۶ (۳۲۸۰۶۰۳۱) - ۸

دورنگار: ۰۲۶ (۳۲۸۰۸۱۱۴)

ایمیل: standard@isiri.org.ir

وبگاه: <http://www.isiri.gov.ir>

Iranian National Standardization Organization (INSO)

No.1294 Valiasr Ave., South western corner of Vanak Sq., Tehran, Iran

P. O. Box: 14155-6139, Tehran, Iran

Tel: + 98 (21) 88879461-5

Fax: + 98 (21) 88887080, 88887103

Standard Square, Karaj, Iran

P.O. Box: 31585-163, Karaj, Iran

Tel: + 98 (26) 32806031-8

Fax: + 98 (26) 32808114

Email: standard@isiri.org.ir

Website: <http://www.isiri.gov.ir>

ب

به نام خدا

آشنایی با سازمان ملی استاندارد ایران

سازمان ملی استاندارد ایران به موجب بند یک ماده ۳ قانون اصلاح قوانین و مقررات مؤسسه استاندارد و تحقیقات صنعتی ایران، مصوب بهمن ماه ۱۳۷۱ تنها مرجع رسمی کشور است که وظیفه تعیین، تدوین و نشر استانداردهای ملی (رسمی) ایران را به عهده دارد.

تدوین استاندارد در حوزه‌های مختلف در کمیسیون‌های فنی مرکب از کارشناسان سازمان، صاحب‌نظران مراکز و مؤسسات علمی، پژوهشی، تولیدی و اقتصادی آگاه و مرتبط انجام می‌شود و کوششی همگام با مصالح ملی و با توجه به شرایط تولیدی، فناوری و تجاری است که از مشارکت آگاهانه و منصفانه صاحبان حق و نفع، شامل تولیدکنندگان، مصرفکنندگان، صادرکنندگان و واردکنندگان، مراکز علمی و تخصصی، نهادها، سازمان‌های دولتی و غیردولتی حاصل می‌شود. پیش‌نویس استانداردهای ملی ایران برای نظرخواهی به مراجع ذی‌نفع و اعضای کمیسیون‌های مربوط ارسال می‌شود و پس از دریافت نظرها و پیشنهادها در کمیته ملی مرتبط با آن رشته طرح و در صورت تصویب، به عنوان استاندارد ملی (رسمی) ایران چاپ و منتشر می‌شود.

پیش‌نویس استانداردهایی که مؤسسات و سازمان‌های علاقه‌مند و ذی‌صلاح نیز با رعایت ضوابط تعیین شده تهیه می‌کنند در کمیته ملی طرح، بررسی و در صورت تصویب، به عنوان استاندارد ملی ایران چاپ و منتشر می‌شود. بدین ترتیب، استانداردهایی ملی تلقی می‌شود که بر اساس مقررات استاندارد ملی ایران شماره ۵ تدوین و در کمیته ملی استاندارد مربوط که در سازمان ملی استاندارد ایران تشکیل می‌شود به تصویب رسیده باشد.

سازمان ملی استاندارد ایران از اعضای اصلی سازمان بین‌المللی استاندارد (ISO)^۱، کمیسیون بین‌المللی الکترونیک (IEC)^۲ و سازمان بین‌المللی اندازه‌شناسی قانونی (OIML)^۳ است و به عنوان تنها رابط^۴ کمیسیون کدکس غذایی (CAC)^۵ در کشور فعالیت می‌کند. در تدوین استانداردهای ملی ایران ضمن توجه به شرایط کلی و نیازمندی‌های خاص کشور، از آخرین پیشرفت‌های علمی، فنی و صنعتی جهان و استانداردهای بین‌المللی بهره‌گیری می‌شود.

سازمان ملی استاندارد ایران می‌تواند با رعایت موازین پیش‌بینی شده در قانون، برای حمایت از مصرفکنندگان، حفظ سلامت و ایمنی فردی و عمومی، حصول اطمینان از کیفیت محصولات و ملاحظات زیستمحیطی و اقتصادی، اجرای بعضی از استانداردهای ملی ایران را برای محصولات تولیدی داخل کشور و/یا اقلام وارداتی، با تصویب شورای عالی استاندارد، اجباری کند. سازمان می‌تواند به منظور حفظ بازارهای بین‌المللی برای محصولات کشور، اجرای استاندارد کالاهای صادراتی و درجه‌بندی آن را اجباری کند. همچنین برای اطمینان بخشیدن به استفاده کنندگان از خدمات سازمان‌ها و مؤسسات فعال در زمینه مشاوره، آموزش، بازرگانی، ممیزی و صدور گواهی سیستم‌های مدیریت کیفیت و مدیریت زیست محیطی، آزمایشگاه‌ها و مراکز واسنجی (کالیبراسیون) وسائل سنجش، سازمان ملی استاندارد این‌گونه سازمان‌ها و مؤسسات را بر اساس ضوابط نظام تأیید صلاحیت ایران ارزیابی می‌کند و در صورت احراز شرایط لازم، گواهینامه تأیید صلاحیت به آن‌ها اعطا و بر عملکرد آن‌ها نظارت می‌کند. ترویج دستگاه بین‌المللی یکاهای واسنجی وسائل سنجش، تعیین عیار فلزات گرانبهای و انجام تحقیقات کاربردی برای ارتقای سطح استانداردهای ملی ایران از دیگر وظایف این سازمان است.

1- International Organization for Standardization

2- International Electrotechnical Commission

3- International Organization for Legal Metrology (Organisation Internationale de Métrologie Legale)

4-Contact point

5- Codex Alimentarius Commission

کمیسیون فنی تدوین استاندارد

«اطلاعات و مستند سازی- مباحث آماری و کیفی مربوط به آرشیو سازی وب»

سمت و / یا محل اشتغال:

رئیس:

سرپرست دایرۀ المعارف پزشکی اسلام و ایران-
فرهنگستان علوم پزشکی
قاسملو، فرید
(کارشناسی ارشد زبان های باستانی)

دبیر:

کارشناس دفتر تدوین- پژوهشگاه استاندارد
مرادی، مهتاب
(کارشناسی مترجمی زبان انگلیسی)

اعضا: (اسمی به ترتیب حروف الفبا)

سرپرست بخش اعلام و مدخل دایرۀ المعارف
پزشکی اسلام و ایران- فرهنگستان علوم پزشکی
ابراهیم زاده، فهیمه
(کارشناسی ارشد کتابداری و اطلاع رسانی)
عضو هیئت علمی- مرکز تحقیقات مخابرات ایران
اثنی عشری اصفهانی، محمد مهدی
(دکترای هوش مصنوعی)

- مشاور دایرۀ المعارف پزشکی اسلام و ایران-
فرهنگستان علوم پزشکی
تجلى، وحید
(کارشناس ارشد زبان شناسی همگانی)

کارشناس مسئول IT- پژوهشگاه استاندارد
ملکی، امین
(کارشناسی فناوری اطلاعات)

مدیر گروه امنیت فناوری و اطلاعات- سازمان
پژوهش های علمی و صنعتی
باسری، نوید
(کارشناسی ارشد کامپیوتر)

کارشناس مسئول- نهاد کتابخانه های عمومی
کشور
پازوکی، فاطمه
(کارشناسی ارشد کتابداری و اطلاع رسانی)

سرپرست بخش آموزش- فرهنگستان علوم
پزشکی
پزشکی اردکانی، سیده فاطمه
(کارشناسی ارشد فلسفه)

عضو هیئت علمی دانشگاه شهید بهشتی
 حاجی زین العابدینی، محسن
(دکترای کتابداری و اطلاع رسانی)

کارشناس- فرهنگستان علوم پزشکی
دادرس، پری ناز
(کارشناسی ارشد زبان شناسی همگانی)

سمت و/یا محل اشتغال:

اعضا: (اسمی به ترتیب حروف الفبا)

مدیر سازماندهی و تحلیل اطلاعات- پژوهشگاه علوم و فناوری اطلاعات ایران

دهسرایی، زهرا

(کارشناسی ارشد علوم اطلاعات و دانش شناسی)

کارشناس- پژوهشگاه استاندارد

شیرازی میگون، مریم

(کارشناسی ارشد فناوری اطلاعات)

کارشناس پژوهشکده اسناد- سازمان اسناد و کتابخانه ملی ایران

ضرغامی، زهرا

(کارشناسی ارشد زبان انگلیسی)

سربرست گروه ویرایش اصطلاحی- دایرهالمعارف پژوهشکی اسلام و ایران

سعیده، قدردوست نخچی

(کارشناسی ارشد زبان شناسی همگانی)

کارشناس پژوهشکده اسناد- سازمان اسناد و کتابخانه ملی ایران

ضرغامی، زهرا

(کارشناسی ارشد زبان انگلیسی)

کارشناس دفتر تدوین- پژوهشگاه استاندارد

قشقایی، محمد مهدی

(کارشناسی مهندسی معدن)

رئیس گروه اطلاع رسانی- کتابخانه تربیت مدرس

مهرمندوست، طاهره

(کارشناسی ارشد کتابداری)

ویراستار:

کارشناس استاندارد- بازنیسته سازمان ملی استاندارد

طوماریان، سهیلا

(کارشناسی مهندسی الکترونیک)

فهرست مندرجات

عنوان	صفحة
پیش گفتار	ز
هدف و دامنه کاربرد	۱
اصطلاحات و تعاریف	۲
روش‌ها و اهداف آرشیوسازی وب	۳
روش‌های جمع آوری	۱-۳
روش‌های توصیف و دسترسی	۲-۳
روش‌های محافظت	۳-۳
بنیان قانونی برای آرشیوسازی وب	۴-۳
دلایل دیگر برای آرشیوسازی وب	۵-۳
آمارها	۴
کلیات	۱-۴
آمارهای مربوط به توسعه مجموعه	۲-۴
مشخصه‌یابی مجموعه	۳-۴
استفاده از مجموعه	۴-۴
محافظت از آرشیو و ب	۵-۴
محاسبه هزینه‌های آرشیوسازی وب	۶-۴
شاخص‌های کیفیت	۵
کلیات	۱-۵
محدودیت‌ها	۲-۵
توصیف	۳-۵
استفاده و مزایا	۶
کلیات	۱-۶
کاربردها و مخاطبین مورد نظر	۲-۶
مزایا برای گروه‌های کاربری	۳-۶
به کارگیری آمارهای پیشنهادی توسط گروه کاربران	۴-۶
فرآیند آرشیوسازی وب با شاخص‌های کارایی مربوطه	۵-۶
کتاب نامه	۸۶
	۸۴
	۸۱
	۸۰
	۷۹
	۷۹
	۷۹
	۶۷
	۶۷
	۶۶
	۶۶
	۶۳
	۵۷
	۵۱
	۴۲
	۳۲
	۳۱
	۳۱
	۳۰
	۲۷
	۲۳
	۱۹
	۱۵
	۱۵
	۱
	۲
	۱
	ز

پیش‌گفتار

استاندارد «اطلاعات و مستند سازی- مباحث آماری و کیفی مربوط به آرشیوسازی وب» که پیش‌نویس آن در کمیسیون‌های مربوط بر مبنای پذیرش استانداردهای بین‌المللی به عنوان استاندارد ملی ایران به روش اشاره شده در مورد الف، بند ۷، استاندارد ملی ایران شماره ۵ تهیه و تدوین شده، در دویست و یکمین اجلاسیه کمیته ملی استاندارد اسناد و تجهیزات اداری و آموزشی مورخ ۱۳۹۵/۱۲/۰۴ تصویب شد. اینک این استاندارد به استناد بند یک ماده ۳ قانون اصلاح قوانین و مقررات مؤسسه استاندارد و تحقیقات صنعتی ایران، مصوب بهمن ماه ۱۳۷۱، به عنوان استاندارد ملی ایران منتشر می‌شود.

استانداردهای ملی ایران بر اساس استاندارد ملی ایران شماره ۵ (استانداردهای ملی ایران- ساختار و شیوه نگارش) تدوین می‌شوند. برای حفظ همگامی و هماهنگی با تحولات و پیشرفت‌های این‌ملی و جهانی در زمینه صنایع، علوم و خدمات، استانداردهای ملی ایران در صورت لزوم تجدیدنظر خواهند شد و هر پیشنهادی که برای اصلاح یا تکمیل این استانداردها ارائه شود، در هنگام تجدیدنظر در کمیسیون فنی مربوط، مورد توجه قرار خواهد گرفت. بنابراین، باید همواره از آخرین تجدیدنظر استانداردهای ملی ایران استفاده کرد.

این استاندارد ملی بر مبنای پذیرش استاندارد بین‌المللی زیر به روش «معادل یکسان» تهیه و تدوین شده و شامل ترجمه تخصصی کامل متن آن به زبان فارسی می‌باشد و معادل یکسان استاندارد بین‌المللی مذبور است:

ISO/TR 14873: 2013, Information and documentation - Statistics and quality issues for web archiving

«اطلاعات و مستند سازی- مباحث آماری و کیفی مربوط به آرشیوسازی وب»

۱ هدف و دامنه کاربرد

هدف از تدوین این استاندارد، تعیین آمارها، اصطلاحات و معیارهای کیفی مربوط به آرشیوسازی وب است. این استاندارد همچنین نیازها و شیوه‌هایی که بسیاری از سازمان‌ها نظری کتابخانه‌ها، آرشیوها، موزه‌ها، مراکز تحقیقاتی و بنیادهای حفظ آثار با آن مواجه شده‌اند را مدنظر قرار می‌دهد. مثال‌های ذکر شده از بخش کتابخانه برگرفته شده‌اند، زیرا کتابخانه‌ها به ویژه کتابخانه ملی وظیفه جدید آرشیوسازی وب را بر عهده گرفته‌اند. توصیه می‌شود این امر به هیچ وجه اهمیت سهم سایر موسساتی غیر از کتابخانه را کم نکند و نیز از کاربرد پذیری اصول این استاندارد برای موسسات حفظ آثار و متخصصان تولید آرشیو نکاهد.

این استاندارد برای متخصصینی درنظر گرفته شده است که اغلب در گروه‌های ترکیبی شامل کتابداران، متصدیان آرشیو، مهندسان و کارکنان مدیریتی مستقیماً در آرشیوسازی وب دخیل هستند. این استاندارد همچنین برای سرمایه‌گذاران مالی موسسات آرشیوسازی وب و ذی‌نفعان برون سازمانی، کاربرد دارد. اصطلاحات به کار رفته در این استاندارد قصد دارد گستره وسیعی از مزايا و تخصص‌های مخاطبینی را منعکس نماید که بین علم رایانه، مدیریت و کتابداری توازن برقرار می‌کنند.

این استاندارد مدیریت منابع الکترونیکی و دانشگاهی و تجاری نظری مجلات، روزنامه‌ها یا کتاب‌های الکترونیکی که معمولاً به صورت مجزا و با استفاده از سامانه‌های مدیریتی متفاوتی انباشته و پردازش می‌شوند را مد نظر قرار نمی‌دهد. این منابع به عنوان منابع اینترنتی درنظر گرفته می‌شوند و در این استاندارد به عنوان یکی از شاخه‌های متمایز آرشیوسازی وب به آن‌ها، پرداخته نمی‌شود. برخی از سازمان‌ها مدارک الکترونیکی که می‌تواند از طریق وب، از طریق ذخایر الکترونیکی ناشر محور و از طریق سامانه‌های دارای مخزن^۱ ارائه شود را نیز جمع‌آوری می‌کنند. این موارد بسیار خارج از دامنه کاربرد این استاندارد است. اصول و روش‌هایی که برای جمع‌آوری این نوع از اطلاعات استفاده می‌شود در حقیقت بسیار متفاوت از اصول و روش‌های آرشیوسازی وب است. آمار و نمایه‌های کیفی مرتبط که برای یک روش مناسب است، الزاماً مناسب روش‌های دیگر نیست.

در نهایت این استاندارد اساساً بر روی اصول و روش‌های آرشیوسازی وب تمرکز دارد و روش‌های جایگزین جمع‌آوری منابع الکترونیکی را شامل نمی‌شود. در حقیقت برخی منابع اینترنتی، به ویژه منابعی که روی وب توزیع نمی‌شوند (برای مثال: خبرنامه‌هایی که به صورت پست الکترونیکی

توزیع می‌شوند)، باروش‌های آرشیوسازی وب برداشت نشده‌اند و با سایر روش‌هایی که در این استاندارد توصیف یا تحلیل نشده است، جمع‌آوری شده‌اند.

۲ اصطلاحات و تعاریف

در این استاندارد، اصطلاحات و تعاریف زیر به کار می‌روند:

۱-۲

دسترسی

access

درخواست موفق از یک کتابخانه که خدمت برخط را ارائه می‌کند.

یادآوری ۱- دسترسی، یک چرخه از فعالیت‌های کاربر است که عموماً زمانی شروع می‌شود که کاربر به کتابخانه ای متصل می‌شود که خدمت برخط را ارائه می‌کند و با یک فعالیت آشکار (ترک پایگاه اطلاعاتی از طریق خارج شدن یا خروج) و یا فعالیت ضمنی (پایان زمان به علت عدم فعالیت کاربر) پایان می‌یابد.

یادآوری ۲- دسترسی هایی که به وب‌گاه کتابخانه صورت می‌گیرد، بازدیدهای مجازی محسوب می‌شود.

یادآوری ۳- درخواست‌های مربوط به صفحه درگاه یا ورود عمومی، مستثنی هستند.

یادآوری ۴- در صورت امکان، درخواست‌های دریافت شده از موتورهای جستجو مستثنی هستند.

[منبع: تعریف ۱-۲-۲ از استاندارد ISO 2789: 2013]

۲-۲

ابزار دسترسی

access tool

نرم افزار تخصصی که برای یافتن، بازیابی و بازپخش منابع اینترنتی آرشیو شده به کار رفته است

یادآوری- ابزار دسترسی ممکن است به وسیله تعدادی از بسته‌های نرم افزاری مجزا که همراه با هم کار می‌کنند، پیاده سازی شوند.

۳-۲

فراداده اجرایی

administrative metadata

اطلاعات ضروری که مدیریت مناسب اشیاء رقمی در یک مخزن را ممکن می‌سازد

یادآوری- فراداده اجرایی را می‌توان به گروههای زیر تقسیم کرد:

- فراداده زمینه یا اصالت (اصل): فراداده‌ای است که چرخه‌ی حیات یک منبع تا یک نقطه خاص، شامل هستارها و فرآیندهای مرتبط را توصیف می‌کند؛ برای مثال پیکربندی و فایل‌های ثبت وقایع.

- فراداده فنی: فرادادهای است که ویژگی‌های فنی یکشیء رقمی، برای مثال قالب آن را توضیح می‌دهد.
- فراداده حقوقی: فرادادهای است که مالکیت و استفاده مجاز قانونی از شیء را تعریف می‌کند.

۴-۲

آرشیو

archive

آرشیو وب

Web archive

- مجموعه کامل از منابع خزش شده از وب در طول زمان، که از یک یا چند مجموعه تشکیل شده است.

۵-۲

جریان بیت

bit stream

مجموعه‌ای از رقم‌های صفر و یک که فایل رقمی را تشکیل می‌دهد.

۶-۲

بودجه (خزش)

budget (crawl)

محدودیت مربوط به یک خزش یا تک تک نقاط شروع^۱ مورد استفاده، که می‌تواند برای مثال به صورت تعداد فایل‌ها، حجمی از داده‌ها، یا زمان صرف شده برای هر خزش به شکلی که در تنظیمات خشکش تعریف شده است، بیان شود.

۷-۲

خزش انبوه

bulk crawl

برداشت انبوه

bulk harvest

خزشی که در جهت جمع‌آوری همه دامنه‌های سطح بالای منفرد یا چندگانه یا زیرمجموعه‌ها) است.

یادآوری ۱- در مقایسه با خزش‌های انتخابی، خزش‌های انبوه محدوده وسیع‌تری دارند و عموماً کمتر اجرا می‌شوند.

یادآوری ۲- خزش‌های انبوه عموماً به آرشیوهای وب با مقیاس بزرگ منجر می‌شوند که تضمین کیفیت را غیر ممکن می‌سازند. این فرایند اغلب از طریق نمونه برداری، انجام می‌شود.

1- Individual seeds

۳

۸-۲

نسخه برگرفته (نسخه‌برداری)

Capture

نمونه

instance

رونوشتی^۱ از یک منبع که در نقطه‌ای خاص از زمان، خزش شده است
یادآوری- اگر یک منبع سه بار در زمان‌های مختلف خزش شده باشد، سه نسخه برگرفته وجود خواهد داشت.

۹-۲

مجموعه

collection

مجموعه آرشیو وب

Web archive collection

منابع بهم پیوسته که به عنوان یک گروه ارائه شده است

یادآوری ۱- مجموعه می‌تواند به طور مشخص قبل از جمع‌آوری انتخاب شود (برای مثال یک رویداد یا یک عنوان) یا با نگاه به گذشته از منابع موجود در آرشیوها به دست آید.

یادآوری ۲- آرشیو وب ممکن است شامل یک یا چند مجموعه باشد.

۱۰-۲

خزش

crawl

برداشت

harvest

فرایند بررسی عمومی و رونوشت برداری از منابع با استفاده از یک خزشگر است

یادآوری- خزش‌ها می‌توانند به دو شکل انبوه یا انتخابی دسته بندی شود.

۱۱-۲

تنظیمات خزش

crawl settings

پارامترهای خزش

crawl parameters

1-Copy

تعریف این که بهتر است کدام منابع جمع‌آوری شوند و تکرار و عمقی که برای هر مجموعه از نقاط شروع

مورد نیاز است

یادآوری - تنظیمات خزش، حدگزاری^۱ خزشگر (تعداد درخواست‌های فرستاده شده به کارساز^۲ میزبان منابع در هر ثانیه یا دقیقه)، سازگاری با فایل روباتس.تی.ایکس.تی و فیلترها برای فرار از تله‌های خزشگر را نیز شامل می‌شود.

۱۲-۲

خزشگر

crawler

برداشت‌کننده

harvester

خزشگر آرشیوکننده

archiving crawler

نام منسخ شده: عنکبوت

DEPRECATED: spider

نرم افزاری که به طور متوالی، مکان‌های یکنواخت منبع (یو.آر.آل‌ها)^۳ را درخواست و منابع حاصل را برای یو.آر.آل‌های بیشتر تجزیه می‌کند^۴.

یادآوری - منابع ممکن است ذخیره شوند و یو.آر.آل‌ها بر طبق مجموعه‌های از قواعد از پیش تعریف شده کنار گذاشته شوند [به تنظیمات خزش (۱۱-۲) و محدوده خزش (۴۰-۲) مراجعه شود].

۱۳-۲

تله خزشگر

crawler trap

صفحه وب (یا مجموعه‌های آن) که موجب می‌شود خزشگر از کار بیفتند یا ارجاعات به منابع کم ارزش یا بی ارزش را بی وقهه دنبال کند.

یادآوری - تله‌های خزشگر می‌توانند عمدتاً در جایی کار گذاشته شوند که از جمع‌آوری منابع به‌وسیله خزشگرها جلوگیری کنند. این کار می‌تواند سه‌وْنیز اتفاق بیفتند، برای مثال به هنگام دنبال کردن بی وقهه روزهای تقویم به‌وسیله یک خزشگر.

1-Politeness

2-Server

3-Uniform Resource Locators(U.R.L)

به تعریف ۵۰-۲ مراجعه شود.

۴- بر اساس استاندارد ملی ایران شماره ۵، اختصارات انگلیسی استفاده شده در متن استاندارد برای اولین بار باید به صورت کامل زیر نویس انگلیسی شوند. به دلیل تکرار زیاد اختصارات انگلیسی در متن این استاندارد، به منظور سهولت، در ادامه متن این اختصارات آوانویسی شده‌اند.

۱۴-۲

ابزار متصدی

curator tool

برنامه کاربردی که روی خزشگر وب اجرا می شود و از فرایندهای برداشت پشتیبانی می کند.
یادآوری- عملکرد اصلی این ابزار، مدیریت مقاصد آرشیو و فراداده توصیفی و اجرایی مرتبط با آن است . ممکن است این عملکرد شامل مولفه هایی برای زمان بندی و کنترل کیفیت نیز باشد .

۱۵-۲

داده کاوی

data mining

فرایندهای محاسباتی است که الگوها را با تجزیه و تحلیل داده های کمی از چشم اندازها و ابعاد مختلف، دسته بندی کردن آن و خلاصه سازی روابط و تاثیرات بالقوه، استخراج می کند.

[منبع: تعریف ۳ از استاندارد ISO 16439]

۱۶-۲

وب عمیق

deep Web

نام منسخ شده: وب پنهان

DEPRECATED: hidden Web

نام منسخ شده: وب نامرئی

DEPRECATED: invisible Web

بخشی از وب که نمی تواند با موتور جستجوگر خزش داده شود یا نمایه سازی شود و به طور مشخص شامل منابعی است که به طور پویا ایجاد شده یا با اسم رمز حفاظت می شوند.

۱۷-۲

فراداده توصیفی

descriptive metadata

اطلاعاتی که محتوای فکری شیء رقمه‌ی را توصیف می کند

۱۸-۲

نام دامنه

domain name

رشته شناساگری است که محدوده استقلال اجرایی، اختیار یا کنترل بر اینترنت را تعریف می‌کند که با قوانین و رویه‌های سامانه نام دامنه (دی.ان.اس)^۱ تعریف شده است.

۱۹-۲

سامانه نام دامنه

domain name system

دی.ان.اس

DNS

سامانه نام‌گذاری جهانی است که به صورت توزیع شده و سلسله مراتبی، برای شناسایی هستارهای متصل شده به اینترنت، مورد استفاده قرار می‌گیرد.

یادآوری - دامنه‌های سطح بالا (تی.ال.دی‌ها) در سلسله مراتب، بالاترین سطح هستند.

۲۰-۲

نمونه سازی

emulation

بازتولید کارکردی و رفتاری یک سامانه منسوخ، با استفاده از نرم‌افزاری (به نام نمونه‌ساز) در سامانه‌های رایانه‌ای رایج است.

یادآوری - نمونه سازی، راهبردی کلیدی در محافظت رقمی است.

۲۱-۲

میزبان

host

بخشی از یو.آر.آل که منبع شبکه‌ای محتوا را نام‌گذاری می‌کند.

یادآوری - میزبان، عموماً نام دامنه برای مثال web.archive.org، www.archive.org، یا زیردامنه برای مثال

۲۲-۲

HTML

زبان نشانه‌گذاری ابرمنتنی

Hypertext Markup Language

زبان نشانه‌گذاری اصلی برای صفحات وب شامل عناصری که برای افزودن اطلاعات ساختاری و معنایی به متن پردازش نشده، استفاده شده‌اند.

۱- به تعریف ۴۸-۲ مراجعه شود.

۲۳-۲

اچ.تی.تی.پی

HTTP

پروتکل انتقال ابرمتن

Hypertext Transfer Protocol

پروتکل ارتباطات کارفرما/کارساز که برای انتقال اطلاعات در وب استفاده می شود

۲۴-۲

ابرپیوند

hyperlink

پیوند

Link

ساختار ارتباطی که برای پیوند اطلاعات در اینترنت استفاده می شود

۲۵-۲

داده ناخواسته

junk

هرزنامه

spam

محظوبات ناخواسته که به نظر می رسد بی ربط یا بدون ارزش بلندمدت است

یادآوری ۱- هرزنامه عمده عموما برای دستکاری نمایه های موتور جستجوگر استفاده می شود. همچنین داده ناخواسته می تواند بطور غیرعمدی به هنگام افتادن خزشگر در تله خزشگر ایجاد شود.

یادآوری ۲- موسسات جمع آوری به طور کلی سعی می کنند تا از جمع آوری داده ناخواسته و هرزنامه اجتناب کنند به طوری که بتوان از منابع برای جمع آوری منابع "خوب" استفاده کرد . با این حال، برخی موسسات نمونه کوچکی از این داده ها را به عنوان بخشی از سابقه وب، حفظ می کنند.

۲۶-۲

پیوند کاوی

link mining

پردازش و تحلیلی که بر استخراج الگوها و ساختارهای مکاشفه ای از ابرپیوندها، برای مثال رسم گراف های شبکه تمرکز می کند.

۲۷-۲

نشت وب زنده

live Web leakage

مشکل رایج در ارائه منابع آرشیو شده، زمانی اتفاق می‌افتد که پیوندها در منبع آرشیو شده به جای آن که به نسخه‌ی آرشیو شده درآرشیو وب هدایت شوند، به نسخه فعلی که روی وب‌گاه قرار دارد، هدایت می‌شوند.
یادآوری - نشت وب زنده نیز وقتی اتفاق می‌افتد که فایل‌های آغازگر صفحات وب آرشیو شده به منابع وب زنده ارجاع می‌دهند و با موقیت درخواست می‌کنند. برای مثال این اتفاق ممکن است موجب شود نقطه نظرات رسانه‌های اجتماعی بر روی وب زنده، یا ویدئوهای جریانی در صفحه وب آرشیو شده، ظاهر شود.

۲۸-۲

فایل ثبت وقایع

log file

فایلی که به‌طور خودکار به‌وسیله کارسازی که سابقه‌ای از فعالیت‌های خود را نگه می‌دارد، ایجاد می‌شود.

۲۹-۲

فراداده

metadata

داده‌هایی که زمینه، محتوا و ساختار شیء رقمی و مدیریت آن‌ها را در طول زمان توصیف می‌کند.

[منبع: تعریف ۱۲-۲ از استاندارد ISO 15489-1:2001]

یادآوری - فراداده را می‌توان به فراداده توصیفی، ساختاری و اجرایی رده بندی کرد.

۳۰-۲

گذار

migration

تبديل قالب‌های فایل قدیمی‌تر یا منسوخ به قالب فایل‌های جدیدتر یا جاری با هدف حفظ دسترسی پذیری به شیء رقمی می‌باشد

یادآوری - گذار، راهبرد کلیدی محافظت است.

[منبع: تعریف ۱۳-۳ از استاندارد ISO 15489-1:2001]

۳۱-۲

نوع ا.آ.ام.ای

MIME type^۱

نوع رسانه اینترنتی

Internet media type

نوع محتوا

content type

1- Multipurpose Internet Mail Extension

عبارت است از شناساگر دو بخشی برای قالب‌های فایل در اینترنت

یادآوری - توسعه چند منظوره رایانمۀ اینترنت (ام.آی.ام.ای) از سرآیندۀ نوع - محتوا، شامل نوع و زیرنوع، برای نشان دادن قالب منبع استفاده می کند، برای مثال، .image/jpeg

۳۲-۲

نامزدی (کاندیداتوری)

nomination

منبع کاندیداتوری که برای شمول در آرشیو وب باید درنظر گرفته شود.

۳۳-۲

صفحه

page

صفحه وب

Web page

منبع ساختار یافته، که علاوه بر هرگونه محتوای قابل خواندن برای انسان، شامل صفر یا تعدادی ارتباط با دیگر منابع است و بهوسیله یک یو.آر.آل مشخص می شود.

۳۴-۲

مجوز

permission

اجازه خزیدن یک وب‌گاه زنده و/یا اجازه نمایش عمومی محتوای آن در یک آرشیو وب می باشد

یادآوری - مجوز می تواند با یک پروانه^۱ رسمی از طرف صاحب حقوق^۲ اعلام، یا به واسطه قوانین حفظ آثار ملی در کتابخانه ملی، تفویض شود.

۳۵-۲

کاربر ثبت شده

registered user

شخص یا سازمانی که عضویت او در کتابخانه ثبت شده است تا بتواند از داخل یا خارج از کتابخانه، از مجموعه و/یا خدمات کتابخانه، استفاده کند

یادآوری ۱- کاربران می توانند با ارائه درخواست به عضویت ثبت شده در کتابخانه درآیند، یا به صورت خودکار پس از عضویت در موسسه به عضویت انها در کتابخانه ثبت شود.

یادآوری ۲- فرایند ثبت کاربران در فواصل منظم، حداقل هر سه سال یکبار، پایش می شوند، به طوری که بتوان کاربران غیرفعال را حذف کرد.

1-License
2-Right holder

[منبع: تعریف ۲-۲-۲۸ از استاندارد ISO 2789:2013]

۳۶-۲

درخواست

request

پیامی است که در قالب اج.تی.پی بهوسیله سامانه درخواست‌کننده (برای مثال یک مرورگر یا یک خوشگر) به یک کارساز کنترل از راه دور برای منبع خاص مشخص شده بهوسیله یک یو.آر.آل، فرستاده می‌شود.

۳۷-۲

پاسخ

response

جواب یک کارساز دور دست به یک درخواست اج.تی.پی برای یک منبع، شامل منبع درخواست شده، یا باز ارجاع^۱ به یک یو.آر.آل دیگر یا یک پاسخ منفی (خطا) است، که نشان می‌دهد چرا منبع درخواست شده نمی‌تواند در اختیار گذاشته شود.

۳۸-۲

کد پاسخ

response code

کد وضعیت

status code

عدد سه رقمی که به کارساز درخواست کننده، وضعیت منبع درخواست شده را نشان می‌دهد یادآوری-کدهایی که با رقم ۴ شروع می‌شود، برای مثال (4xx) نشان می‌دهد که منبع درخواست شده، موجود نیست.

۳۹-۲

روباتس.تی.ایکس.تی

robots.txt

استاندارد، جلوگیری کننده از نمایه سازی توسط روباتس ها

robots exclusion standard

پروتکل استفاده شده که از دسترسی خوشگرهای وب به تمام یا بخشی از وب‌گاه، جلوگیری می‌کند یادآوری^۱-رباتس.تی.ایکس.تی. الزام آور قانونی^۲ نیست.

یادآوری ۲- این پروتکل ممکن است برای درخواست حداقل تاخیر بین درخواستهای متوالی یا حتی فراهم ساختن پیوندی به نقشه سایت جهت تسهیل خرش کردن بهتر سایت، نیز استفاده شود.

۴۰-۲

محدوده (خرش)

scope (crawl)

مجموعه‌ای از پارامترها که گستره خرش را تعریف می‌کند، برای مثال، بیشترین تعداد گام‌ها^۱ یا بیشترین عمق مسیری که توصیه می‌شود خرشگر دنبال کند

یادآوری ۱- محدوده خرش می‌تواند به گستردگی کل دامنه سطح بالا (برای مثال...ir) یا به محدودیت یک فایل باشد.

۴۱-۲

محدوده (آرشیو وب)

scope (Web archive)

وسعت آرشیو یا مجموعه وب، که به وسیله اجبار قانونی موسسه‌ای^۲ یا خط مشی مجموعه، تعیین می‌شود

۴۲-۲

دامنه سطح دوم

second level domain

عبارت است از زیربخش‌ها در دامنه‌های سطح بالا برای رده‌های خاص سازمان‌ها یا مراکز مورد نظر(برای مثال .gov.ir برای وب‌گاه‌های دولتی، ac.ir برای وب‌گاه‌های دانشگاهی)

۴۳-۲

نقطه شروع (دانه)

seed

یو.آر.آل مورد هدف

Targeted URL

یو.آر.آل مربوط به مکان^۴ منبع خاصی که باید خرش شود، و به عنوان نقطه شروع به وسیله یک خرشگر وب مورد استفاده قرار می‌گیرد.

۴۴-۲

2-Hops
3-Institutional legal mandate
4- Location

انتخاب

selection

فرآیند تصمیم‌گیری متصدیانه، تعیین می‌کند که آیا مجموعه معنی‌داری از منابع، در محدوده آرشیو وب قرار دارد و براساس خطمشی توسعه مجموعه، ارزیابی شده است.

۴۵-۲

خرش انتخابی

selective crawl

برداشت انتخابی

selective harvest

خرشی که در جهت جمع‌آوری منابع انتخاب شده مطابق با معیارهای معین است

یادآوری ۱- در مقایسه با خوش‌های انبوه، خوش‌های انتخابی محدوده‌ی کمتری دارند و عموماً به تعداد دفعات بیشتری اجرا می‌شوند.

یادآوری ۲- خوش‌های انتخابی پیوسته، خوش‌هایی در جهت جمع‌آوری منابع منتخب مطابق با معیارهای معین هستند، از قبیل اهمیت علمی، ارتباط با موضوع یا تکرار به روزرسانی پیوسته منبع.

یادآوری ۳- خوش‌های انتخابی رویداد، خوش‌های مبتنی بر زمان در جهت جمع‌آوری منابع مرتبط با رویدادهای منحصر به‌فرد هستند که در تاریخ معین به پایان می‌رسند، از قبیل انتخابات، رویدادهای ورزشی و بلایا.

۴۶-۲

فراداده ساختاری

structural metadata

اطلاعاتی که توصیف می‌کند چگونه اشیاء مرکب جهت ایجاد واحدهای منطقی با یکدیگر ترکیب شده‌اند

۴۷-۲

مقصد آرشیو

target

مجموعه معنی‌داری از منابعی که باید آن‌گونه که به‌وسیله یک یا چند نقطه شروع و تنظیمات خوش مرتبط تعریف شده است، جمع‌آوری شوند

۴۸-۲

دامنه سطح بالا

top level domain

تی.ال.دی

TLD

بالاترین سطح از دامنه‌ها در سامانه نام دامنه (دی.ان.اس)، از جمله دامنه‌های سطح بالای کد کشور (برای مثال .fr..de) می‌باشد. می‌توان از اختصارات نام کشورها (ISO 3166) و دامنه‌های سطح بالای عمومی (برای مثال ..com..net..org و .paris).

یادآوری - این اصطلاح به معنای دامنه‌های سطح بالای کد کشور در گزارش می‌باشد، مگر اینکه به طور ویژه غیر از این ذکر شده باشد.

۴۹-۲

شناساگر یکنواخت منبع

Uniform Resource Identifier (URI)

یو.آر.آی

رشته توسعه پذیری از نویسه‌ها که برای مشخص کردن یا نام‌گذاری منبع در اینترنت، استفاده شده است

۵۰-۲

مکان‌یاب یکنواخت منبع

Uniform Resource Locator (URL)

یو.آر.ال

زیرمجموعه‌ای از شناساگر یکنواخت منبع (یو.آر.آی) که مکان یک منبع و پروتکل بازیابی آن را مشخص می‌کند.

۵۱-۲

قالب WARC

WARC format

قالب فایل که روشی را برای ترکیب چندین منبع رقمنی در یک فایل آرشیو کل، همراه با اطلاعات مربوط مشخص می‌کند

یادآوری - قالب WARC (آرشیو وب)، از سال ۲۰۰۹ در سازمان ISO استاندارد شده است (ISO 28500:2009)

۵۲-۲

وب‌گاه

website

مجموعه‌ای از صفحات وب که از جهت قانونی و/ یا مدیریتی، بهم پیوسته هستند

یادآوری - معمولاً وب‌گاه‌ها نشان‌دهنده موسسات رسمی، سازمان‌های رسمی، شرکت‌های خصوصی و صفحات شخصی هستند.

۵۳-۲

وب

Web

کاربرد اصلی نشر در اینترنت که با سه استاندارد کلیدی: یو.آر.آی، اج.تی.بی و اج.تی.ام.آل، فعال می‌شود

۳ روش‌ها و اهداف آرشیوسازی وب

خطمی‌های موسسه‌ای و امکانات فنی، شکل و محتوای آرشیوهای وب را تعیین می‌کند. در حالی که خطمی‌های سطح بالا اصولاً با قانون‌گذاری ملی وضع می‌شود، موسسات از انواع راهبردهای جمع‌آوری استفاده می‌کنند که از اهداف کسب‌وکار مرتبط و معیارهای انتخاب، تاثیر می‌پذیرند. با این حال، گاهی اوقات منابع درون محدوده را به واسطه محدودیت‌های فنی نمی‌توان به آرشیوهای وب افزود. دریافت و بازپخش منابع چندرسانه‌ای و منابع تعاملی، برای مثال چالش‌های قابل توجهی را برای جامعه آرشیوکننده وب مطرح می‌کند و اغلب مستلزم راه حل‌های گران و سفارشی است.

۱-۳ روش‌های جمع‌آوری

۱-۱-۳ مبانی فنی

رونوشتبرداری از منابع برخط یا برداشت از آن‌ها، روش اصلی جمع‌آوری منابع اینترنتی است. برداشت نیازمند استفاده از روبوت‌ها است که به صورت متوالی به نشانی‌های اینترنتی مراجعه می‌کنند و از منابع رونوشت تهیه و آن‌ها را ذخیره می‌کنند و منابع به دست آمده را برای دستیابی به نشانی‌های بیشتر تجزیه می‌کنند. نقطه شروع خزشگر، اغلب صفحه اصلی وب‌گاه است که دانه نامیده می‌شود. خزشگر مانند یک کاربر وب خودکار است و می‌تواند منابع اینترنتی به هم پیوسته^۱ را به صورت نامحدودی دنبال کند مگر این‌که محدوده جستجوی آن به عوامل خزشگر و یا تنظیمات خاصی، محدود شود. همچنین خزشگران می‌توانند هنگامی که با موانعی در فرایند برداشت مواجه می‌شوند، سهوا متوقف شوند.

عمق، پوشش و کیفیت مجموعه آرشیووب به وسیله مجموعه‌ای از تنظیمات فنی که قواعد برداشت نامیده شده‌اند، مشخص می‌شوند. منابعی که باید جمع‌آوری شوند به وسیله مکان آن‌ها (یو.آر.آل‌ها) به شکل فهرستی از نقاط شروع (دانه‌ها) و محدوده‌شان توصیف می‌شوند. محدوده و دامنه جستجو اساساً با توجه به تناسب و عمق برداشت بر جامعیت آرشیووب، تاثیر می‌گذارد.

1-Interlinked internet resources

۲-۱-۳ محدودیت‌ها

حجم عظیمی از اطلاعات با سرعت متغیر کنده‌ای در حال اضافه شدن به وب است. معمولاً موسسات با درنظرگرفتن منابع مرتبط با کارکنان، قدرت محاسبه و ظرفیت ذخیره سازی در مورد دامنه آرشیوکردن وب تصمیم می‌گیرند. علاوه بر این تصمیمات اتخاذ شده در طول فرایندهای انتخاب و برداشت باعث تمرکز موسسات بر منابع با ارزش و "خوب" می‌شود که برخلاف منابع تولیدشده خودکار کمارزش، نظیر داده ناخواسته و هرزنامه است که این خود یکی از منابع اصلی است که آرشیوهای وب باید از آن اجتناب کنند و نیازمند مدیریت فعال و اولویت‌بندی فرایند خوش است.

محدودیت‌های بسیاری وجود دارد که جمع‌آوری جامع منابع اینترنتی را چالش برانگیز می‌کند. برخی مربوط به فناوری و برخی معلول مقیاس و ماهیت اینترنت است. محدودیت‌های بیشتر ممکن است از طرف قوانین تحمیل شود.

الف- موضوعات ناشی از ساختار فعلی وب و فناوری‌های خوش

خوشگرهای آرشیوکنده قادر به ضبط محتويات تقریباً پایداری هستند که می‌توانند با درخواست یو.آر.ال‌ها ارائه شوند. وقتی یو.آر.ال‌ها به طور مشخص در اج.تی.ام.ال مورد ارجاع واقع نشده باشند ولی در نرم افزارهایی همچون جاوااسکریپت یا فلش جاگرفته یا به صورت پویا براساس نتایج تعامل با کاربران ایجاد شده باشد، خوشگران آرشیو کننده اغلب قادر به نسخه برگرفته شده از محتوای ارجاعی نخواهد بود. استخراج و تجزیه یو.آر.ال‌های اختیاری همیشه کار ساده‌ای نیست، زیرا که ترکیب یو.آر.ال‌ها می‌توانند بر هر نوع منبع شبکه‌ای نشانی‌دهی کنند و یا حتی فعال شوند و یو.آر.ال‌ها می‌توانند به صورت پویا فعال شوند. مضافاً ساختارهای یو.آر.ال‌های پیچیده ممکن است شامل متغیرهای فراوانی باشد که از علامت‌های "<"، ">"، """، "'"، شناسه‌های (IDs) نشست یا کاربری و نیز کدهایی برای پیگیری ارجاع تشکیل شده باشد. در برخی موارد، فایل‌های چند رسانه‌ای ارائه می‌شوند یا با کدهای تعبیه شده در برنامه‌های کاربردی وب که داده‌ها را از کارساز بازیابی می‌کنند، آغاز می‌شوند بدون اینکه مستقیماً نشانی فایل‌ها را در اج.تی.ام.ال قرار دهند.

فناوری برداشت موجود هنوز در حد کافی بزرگ نیست که به طور کامل وب را جمع‌آوری کند و برخی از انواع محتوا در وب را خارج از دسترس قرار می‌دهد. روش خوش بازگشته موجود بر پایه یو.آر.ال در جمع آوری قسمت اعظم وب، ناکام است و محتوای فرم پشت وب و واسطه‌های پرس و جو که معمولاً به «وب عمیق یا پنهان» معروفند، رسانه‌های جریانی و محتواهایی که بر روی پروتکل‌های غیر از اج.تی.پی و رسانه‌های اجتماعی تبادل می‌شوند را شامل می‌شود.

اساسی‌ترین چالش پیش رو در جوامع تغییر سریع وب با قالب‌ها، پروتکل و بن‌سازه‌های جدید است که سازمان‌های آرشیوکننده را ملزم می‌سازد پاسخگوی بهبود مستمر خود باشند و توانایی خود را برای آرشیوکردن محتواهای نو ظهرور، بهبود بخشد.

ب- مسائل مربوط به تناوب به روز شدن منابع وب

موضوع فنی دیگری که عموماً در ارتباط با خزش به آن اشاره می‌شود، گستگی موقتی است. اگر وب‌گاهی در زمانی که تحت خزش قرار دارد روزآمد شود، وجود همزمان صفحاتی از دوره‌های زمانی متفاوت می‌تواند به ناسازگاری لحظه‌ای^۱ منجر شود.

قوایین ممکن است باعث محدودیت‌های بیشتری در شیوه انجام آرشیوسازی وب شود. تصمیم کلیدی که تحت تاثیر قانون قرارمی‌گیرد این است که آیا بهتر است استانداردهای مستثنی‌سازی فایل‌های روباتس.تی.ایکس.تی در نظر گرفته شود یا نادیده انگاشته شود. این موضوع می‌تواند در دریافت یا از دست رفتن محتوای خاص، تفاوت قابل توجهی ایجاد کند.

محدودیت‌هایی که در بالا به آن‌ها اشاره شد، چالش‌های زیادی را برای فرآیند فعالیت‌های ارزشیابی آرشیوسازی وب با استفاده از معیارهای قبل مقایسه، ایجاد می‌کنند. رویکرد کلی که در این استاندارد از آن پیروی شده، پذیرش محدودیت‌ها و تمرکز بر آن چیزی است که شناخته شده و قابل مقایسه است.

۳-۱-۳ راهبردهای جمع‌آوری:

دو دسته اصلی از راهبردهای جمع‌آوری وجود دارند که از نظر سطح خودکار بودن درگیرشده و همچنین محدوده و مقیاس آرشیو وب حاصل شده، با یکدیگر تفاوت دارند.

- برداشت انبوه^۲، مانند برداشت دامنه ملی، که با هدف نسخه‌برداری لحظه‌ای از کل وضعیت کل دامنه (یا زیر مجموعه‌ای از آن نظیر یک دامنه ملی) در نقطه معینی از زمان انجام شده و منجر به ایجاد مجموعه‌های آرشیو وب در مقیاس بزرگ می‌شود. شناخته شده‌ترین آرشیو انبوhe، آرشیو اینترنت «ماشین راه به گذشته»^۳ است که با هدف محافظت از وب جهانی، ایجاد شده است. جمع‌آوری انبوhe فرآیندی است که تا حد قابل توجهی، خودکار است، اما به لحاظ مقیاس بهره‌برداری محدود است.

1-Distorted snapshot

2-Bulk harvesting

3-Wayback Machine

- در نظر گرفته شده است که این کار (برداشت انبوه) در دفعات کمتری انجام شود، اغلب فقط یک یا دو بار در سال. اگر تضمین کیفیت آن‌ها انجام شود، اغلب از طریق بررسی خودکار محتویات از قلم افتاده به کمک بازبینی کد وضعیت اج.تی.پی. صورت می‌پذیرد.
- آرشیوسازی انتخابی در مقیاسی بسیار کوچک‌تر انجام می‌شود، بسیار متمرکزتر است و در دفعات بیشتری اجرا می‌شود. فرآیند گزینش به منظور شناسایی وب‌گاه‌های مرتبط و براساس معیارهایی همچون موضوع، رخداد، قالب (مثلاً فایل‌های صوتی یا تصویری) و یا موافقتنامه با مالکان محتوا، صورت می‌گیرد. تضمین کیفیت، اصلی رایج در آرشیوسازی انتخابی است که در حال حاضر به شدت به مقایسه چشمی^۱، بازنگری برداشت‌های قبلی و ثبت وقایع خزش، وابسته است. همچنین ترجیح بر این است که آرشیو و ب انتخابی فراداده‌های توصیفی بیشتری داشته باشد، که اغلب در خلال فرآیند گزینش یا بعد از فرآیند برداشت توسط متقدیان، اضافه می‌شوند، که (فراداده‌ها) می‌توانند برای جستجو یا عملکرد مورگری غنی‌تر در واسطه کاربری آرشیوهای وب، مورد استفاده قرار گیرند.
- برخی از موسسات ترکیبی از راهبردهای بالا را به کار می‌گیرند. برخی وب‌گاه‌های اینترنتی به‌طور مرتب روزآمد می‌شوند و تغییرات آن‌ها را نمی‌توان فقط با تکیه بر برداشت‌های نامکرر دامنه‌ای یا برداشت‌های انبوه نسخه‌برداری کرد. غیرمعمول نیست که یک سازمان آرشیوکننده منفرد راهبردی اتخاذ کند که طی آن وب‌گاه‌های اینترنتی دارای اولویت بالاتر در دفعات بیشتری نسخه‌برداری، و وب‌گاه‌هایی که اولویت کمتری دارند تنها در زمان انجام برداشت‌های انبوه یا برداشت‌های دامنه‌ای، نسخه‌برداری شوند.

۴-۱-۳ معیارهای انتخابی

- معیارهای انتخابی اغلب توسط قانون و راهبرد گسترش مجموعه موسسه‌ای، هم‌راستا با مأموریت اصلی مؤسسات جمع‌آوری، ارائه می‌شوند. ملاحظات عملی یا محدودیت‌هایی نظیر کارکنان، منابع و تخصص، اغلب بر اجرای راهبرد، تأثیر می‌گذارد. معیارهای انتخابی، محدوده آرشیو و ب را تعریف می‌کند و می‌تواند به روش‌های مختلف بیان شود:
- از طریق نامهای دامنه‌ای که برای میزبانی منابع به کار می‌رود. برای مثال، از طریق دامنه‌های سطح بالا یا ملی نظیر fr. یا de. یا دامنه‌های سطح دومی که توسط ناشران مشخصی رزو شده‌اند نظیر gov. که برای انتشار نشریات دولتی به کار می‌رود. با وجود این، عنوان‌های دامنه دقیقاً نمی‌توانند به صراحة محتوای ملی را تعریف یا شناسایی کند زیرا اینترنت نظامی جهانی است و منابع در سراسر مرازهای جغرافیایی یا فیزیکی توزیع شده‌اند.

- از طریق مشخصه‌های منبع، برای مثال از طریق زمینه یا موضوع محتوایی وب‌گاه‌های اینترنتی، از طریق محبوبیت نزد مخاطبین یا زبان به کار گرفته شده، از طریق پروتکل‌های ارتباطی به کار رفته جهت ارائه منبع مانند اچ.تی.پی یا از طریق قالب‌های محتواها نظیر متن یا تصویر.
- از طریق شرایط دسترسی یا وضعیت قانون حق مؤلف^۱ منبع، مثلاً آیا منبع به طور رایگان در دسترس است یا از طریق خرید یا اشتراک قابل دستیابی است.
- از طریق توانی که یک سازمان می‌تواند در آرشیوسازی داشته باشد. ممکن است سازمان توانایی منابع مالی فقط برای پشتیبانی از دفعات محدودی برداشت را داشته باشد یا توانایی ارائه رویکرد نمونه برداری برای برداشت در موارد بسیار منتخب را داشته باشد.
- از طریق محدودیت‌های صریح یا استثنائی مرتبط با محتوا، برای مثال معیار انتخابی می‌تواند منابع شامل داده‌های حساس شخصی یا محتوای غیرقانونی را مستثنی کند.

همیشه کاملاً واضح نیست که دربرداشت، می‌توان برخی از انواع منابع خاص را شامل یا مستثنی کرد. در حالی که سازمانی تصمیم می‌گیرد که شبکه‌های اجتماعی، وبلاگ‌ها و بن‌سازه‌های تعاملی مشابه آن‌ها را آرشیو کند، سازمان‌های دیگری ممکن است آن‌ها را خارج از محدوده بدانند. این مسئله در اصل یک تصمیم مبنی بر سیاست‌گذاری است که می‌تواند به صورت یکسانی بر منابع نظری تبلیغ کردن برخط، محتوایات غیراخلاقی^۲ و منابعی که شامل یا تحت تاثیر بدافزارهای رایانه‌ای هستند، اعمال شود. نمونه‌برداری، روشی متدائل برای آرشیوسازی چنین منابعی است که ممکن است دارای ارزش برای برخی از پژوهشگران باشد اما پیش‌بینی این‌که در آینده چه کسی از آن استفاده خواهد کرد، در حال حاضر مشکل است.

۲-۳ روش‌های توصیف و دسترسی

۱-۲-۳ مبنای فنی: روش‌های توصیف

۱-۱-۲-۳ کلیات

جمع‌آوری منابع آرشیوشده و فراهم‌کردن دسترسی به آرشیوهای وب مستقل^۳ کاری رایج است. آن‌ها به همان روش وب زنده کار می‌کنند و اغلب دارای واسطه‌های کاربری اختصاصی هستند که به کاربران اجازه می‌دهد درون محدوده آرشیو، جستجو و حرکت کنند. یکی از نیازمندی‌های خاص که باید در طراحی واسط کاربری درنظر گرفته شود درنظر گرفتن بعد زمان است، به این

1-Copyright status

2-Pornography

3-Stand- alone

معنی که برای کاربران این امکان را فراهم می‌سازد که نسخه‌های متفاوت از منبعی واحد را که در زمان‌های مختلف دریافت شده است، پیدا کنند و به سادگی بین آن‌ها حرکت کنند تا تکامل آن‌ها را در طول زمان مشاهده کنند. رایج‌ترین راه برای مرور آرشیو یک وب از طریق یو.آر.آل است که می‌تواند با تاریخ دریافت همراه باشد.

۲-۱-۲ نمایه‌سازی^۱ از طریق یو.آر.آل (اجباری)

نمایه‌ها نقاط ورود به آرشیو وب را فراهم می‌کنند. آنها به عملیات جستجو و مرتب‌سازی سرعت می‌بخشند و تجربه کاربری بهتری را ارائه می‌دهند. بنیادی‌ترین نمایه، یو.آر.آل یا گونه اصلاح‌شده یو.آر.آل اصلی است که به کارساز میزبان آرشیو وب اشاره می‌کند. تاریخی که منبع مورد خرшу قرار می‌گیرد را می‌توان با یو.آر.آل ترکیب نمود تا از این طریق میان نسخه‌های متفاوت یک منبع تمایز قائل‌شده. رویکرد جایگزین دیگر پیاده‌سازی شناساگر ماندگاری برای هر منبع است که ممکن است شکل یک یو.آر.آل را به خود بگیرد، اما اصل کلیدی این است که موسسه جمع‌آوری تعهد می‌کند شناساگر ارجاع و روش دستیابی به آن منبع را به صورت همیشگی، فراهم کند.

۲-۱-۳ روش‌های دیگر نمایه‌سازی (اختیاری)

جستجوی تمام-متن^۲ روش دیگری از دسترسی به منابع است که آرشیوهای وب به‌طرز فزاً‌ینده‌ای آن را به کار می‌گیرند. این روش به نمایه تمام متن و موتور جستجو نیاز دارد. در عین حال که این روش راه حلی مقیاس پذیر برای دسترسی به منابع است، اما به لحاظ فنی پیاده‌سازی آن با چالش روبرو است. واژگان کلیدی و فراداده‌ها نیز می‌توانند به‌طور خودکار از منابع آرشیو شده استخراج شوند و برای ایجاد دسترسی به منابع به کارگرفته شوند. همچنین از سوی جوامع پژوهشی، نیاز فزاً‌ینده‌ای به‌داده و پیوندکاوی در آرشیوسازی‌های وب ابراز شده است. برخی پیشرفت‌های جدید حاکی از تغییر در تمرکز و توجه در آرشیوسازی وب از سطح سایتهاي اینترنتی یا منبع مستقل به سطح کل آرشیو وب است. با استفاده از فنون مصورسازی و تحلیل داده‌ها، فرصت‌هایی برای دستیابی به چشم‌اندازهای متفاوت از آرشیو وب، رمزگشایی الگوها و گرایشات نهفته، روابط و زمینه‌ها به دست می‌آید. تا زمانی که پیشرفت‌های اشاره شده در بالا به صورت گسترده‌ای در روش‌های عملی مورد استفاده قرار گیرد، در دسترس قراردادن وضعیت‌های پیشین منابع مستقل اینترنتی به عنوان روش‌های اصلی استفاده از آرشیوهای وب باقی خواهند ماند. تمرکز این استاندارد نیز بر این موضوع است.

۴-۱-۲-۳ فهرستنویسی^۳ (اختیاری)

1-Persistent identifier

2- Full text

2- Cataloguing

روش‌های سنتی مدیریت کتاب‌شناختی برای فهرست‌نویسی منابعی مانند کتاب‌های چاپی، مقالات و مجله‌ها می‌توانند در آرشیوهای وب نیز به کار گرفته شوند. یکپارچه سازی آرشیوهای وب در مجموعه‌های موجود در کتابخانه می‌تواند مفید باشد، زیرا همه آن‌ها از طریق جستجوی فهرست‌ها قابل شناسایی می‌شوند. اگرچه این رویکرد نیاز به منابع زیادی دارد و توسعه آن به آرشیوهای وب، به دلیل حجم زیاد اشیایی که در آن‌ها وجود دارد و همچنین چالش برانگیز بودن منابعی که باید فهرست شوند، بسیار دشوار خواهد بود. از فهرست‌نویسی می‌توان در سطح بالاتری از جزئیات استفاده کرد، به عنوان مثال در سطح مجموعه‌های خاص به جای وب‌گاه‌های اینترنتی مستقل.

۵-۱-۲-۳ ابزارهای کشف منابع با استفاده از فراداده‌ها (اختیاری)

دسترسی ممکن است به منابع از طریق اضافه کردن فراداده‌های مرتبط با آن فراهم شود. وب‌گاه‌های اینترنتی ممکن است توسط متقدیان یا از طریق ابزارهای خودکار به سلسله مراتب موضوعی یا مجموعه‌هایی حول رخدادها یا مجموعه‌های موضوع محور، طبقه‌بندی شوند. به کارگیری برچسب‌ها (واژگان کلیدی‌ای که کاربر اضافه می‌کند) که توسط متقدیان یا توسط کاربران عمومی اضافه شود، می‌تواند جزء جدایی‌ناپذیر از واسطه کاربری باشد.

۲-۲-۳ مبنای فنی: روش‌های دسترسی

۱-۲-۲-۳ کلیات

دستیابی به مطالب وب آرشیو شده از طریق به کارگیری نرم‌افزار تخصصی برای پیداکردن، بازیابی و بازپخش مطالب وب آرشیو شده فراهم می‌آید. این فرآیند ممکن است به وسیله بسته‌های نرم‌افزاری جداگانه‌ای که با هم کار می‌کنند، پیاده‌سازی شود. معمولاً مجموعه این نظام نرم‌افزاری را ابزار دسترسی می‌نامند.

ابزار دسترسی، صرف نظر از چگونگی طراحی و پیاده‌سازی، مجموعه‌ای از صفات رایج دارد که برخی اجباری و برخی دیگر اختیاری است.

۲-۲-۳-۲ ارائه (اجباری)

نرم‌افزار دسترسی باید قادر به شناسایی هر منبع با روشنی منحصر به فرد باشد (حتی اگر آن منبع در دفعات متعددی برداشت شده باشد) و بتواند آن منبع را از مخزن آرشیو بازیابی کند.

۳-۲-۳ بازنویسی یو.آر.آل (اجباری)

صفحات اج.تی.ام.آل که با نرم افزار دستیابی بازگردانده می شوند، باید نسبت به ظاهر اصلی شان، اصلاح شوند. توصیه می شود پیوندهای تعبیه شده (مطلق یا نسبی)، به مکان منبع در آرشیو رقمی و نه به مکان منابع اصلی، اشاره کنند. بازنویسی نشانی منابع اصلی می تواند به چند طریق به دست آید:

- این بازنویسی می تواند در زمان برداشت رخ دهد (ممکن است برای بازنویسی فوری یو.آر.آل در درون محتوا تصمیمی متصدی گرایانه گرفته شود و این محتوای اصلاح شده درون مخزن آرشیو جا داده شود).
- عمل محافظت از منابع آرشیو شده با هدفی که در بالا به آن اشاره شد، ممکن است در مراحل بعدی انجام شود. بنابراین، منبع به طور ضمنی پیوندها را به مکان جدیدشان مرتبط می سازد.
- زمان اجرای^۱ بازنویسی یو.آر.آل می تواند توسط رمز اجرا شده برروی کارساز و در هنگام درخواست یک منبع تولید شود یا با ارائه رونوشتی از منبع اصلی به کارفرما^۲ به همراه رمزی که باید کارفرما اجرا کند تا یو.آر.آل به صورت پویا بازنویسی شود، تولید شود.

۳-۲-۴ محدودیت‌ها

فرآیند برداشت و پردازش منابع اینترنتی آرشیو شده، شامل تبدیلاتی است که ممکن است بر ظاهر، رفتار و تجربیات کاربری منابع اصلی در هنگامی که به نقطه دسترسی مراجعه می شود، تاثیر بگذارد. رونوشت‌های آرشیو شده از منابع را باید به عنوان تصویر ثبیت شده از منابع در لحظه‌ای در نظر گرفت که تعامل پذیری نسخه‌های زنده را از دست داده‌اند. مثال‌هایی از این قبیل شامل تابلوهای پیام، انجمن‌های بحث و گفتگو، فرم‌های وب و جستجو است. همچنین ممکن است منابع به طرز مناسبی برداشت شود اما به وسیله قابلیت نرم افزار ارائه^۳ محدود شود و متعاقباً برای کاربران نهایی غیر قابل دسترس شود.

مشکل رایج در بازپخش منابع آرشیو شده، اصطلاحاً "نشت و بزنده" است که زمانی اتفاق می افتد که پیوندهای درون منبع آرشیو شده به جای اینکه به نسخه آرشیوی درون آرشیو و اشاره کنند، به نسخه موجود روی سایت زنده، پیوند می دهند. این موضوع معمولاً با بازنویسی نادرست یو.آر.آل

1-Runtime

2-Client

3-Rendering software

ایجاد می‌شود و اغلب نتیجهٔ پیوندهای قرار داده شده در جاوا اسکریپت^۱ است که ابزارهای دسترسی قابلیت شناسایی آن‌ها را ندارند.

۴-۲-۳ راهبردهای دسترسی

به رغم آنکه بسیاری از منابع اینترنتی به‌طور رایگان در دسترس هستند، معمولاً شامل قانون حق مؤلف نیز می‌شوند. موسسات جمع‌آوری، بسته به قوانین مرتبط (به زیریند^{۴-۴} مراجعه شود) و آنچه به لحاظ قانونی اجازه داده می‌شود، گسترهای از راهبردهای دستیابی را به کار می‌گیرند:

- آرشیو تاریک: مجموعه‌ها نمی‌تواند در دسترس کسی باشد (به‌جز بعضی وقت‌ها که در دسترس کارکنان و به منظور اهداف تصدی گرایانه قرار می‌گیرد).
- آرشیو خاکستری: فقط کاربران نهایی مجاز می‌توانند مجموعه‌ها را ببینند (برای مثال پژوهشگران)، و/یا به محوطه‌های سایت محدود می‌شود (برای مثال سالن مطالعه کتابخانه‌ها).
- آرشیو برخط: دسترسی برای همه کاربران و معمولاً از وب‌گاه‌های موسسه‌ها جمع‌آوری قابل دسترس است.

به عبارتی، یک آرشیو می‌تواند مدل ترکیب‌شده‌ای را با بخش‌هایی از آرشیو که یکی از راهبردهای بالا را می‌پذیرد، به کار ببرد. همچنین قابل ذکر است که توجه شود دسترسی به منابع برداشت شده بر مبنای منطق "بنا بر درخواست"^۳، صورت می‌پذیرد که در آن دسترسی مجوزها تنها به صورت پیش‌فرض و تلویحی و نه تصریحی است و می‌تواند در صورت درخواست مالک حقوق، سلب شود.

۳-۳ روش‌های محافظت^۳

۱-۳-۳ مبنای فنی^۴

محافظت از منابع پیوسته مثل کتاب‌ها و فایل‌های ضبط شده بر حفظ نسخه‌های اصلی منابع تمرکز دارد. در حالیکه محافظت رقمی با موضوعات بسیار متفاوت‌تری سرکار دارد. منابع رقمی در پایین‌ترین سطح شامل صفرها و یک‌ها (جریان بیت)^۵ هستند که مستقل از حامل داده^۶ یا رسانه‌ای هستند که بر روی آن ذخیره شده‌اند. رونوشت‌برداری از بیتها روی حامل دیگر بدون از دست رفتن اطلاعات و تهیه نسخه‌ای برابر با نسخه منبع یا نسخه اصلی امکان‌پذیر است. چون امکان زوال و منسوخ شدن حامل‌های داده وجود دارد لازم

1-Java Script

2-Opt out

3-Preservation method

4-Technical baseline

5-Bit stream

6- Data carrier

است که بیت‌ها به حامل جدید انتقال داده شده تا آن‌ها این نگهداشته شوند. در صورتی که رونوشت‌برداری مرتب انجام شود، می‌توان مطمئن بود جریان بیت بدون هیچ تلفاتی و به طور مداوم محافظت می‌شود.

علاوه بر حفظ بیت‌ها به نحو ایمن، چالش اصلی در محافظت رقمی، ذخیره‌سازی بیت‌ها به گونه‌ای قابل استفاده است. جریان بیتی تا زمانی که به‌وسیله نرمافزارها و محیط سخت‌افزاری مناسب ترجمه وارائه نشود، برای انسان قابل فهم نیست. به دلیل رشد سریع فناوری، اغلب سامانه‌های جدید با سامانه‌های قدیمی‌تر سازگار نیستند و نرمافزارهای ترجمه‌جذیدتر ممکن است قادر به ترجمه قالب فایل‌های قدیمی‌تر نباشند. حتی در صورت امکان ترجمه نرم افزارهای قدیمی‌تر در سامانه‌های موجود، کاربران فعلی ممکن است قادر به استفاده از آن‌ها نباشند هرچند که از کاربران انتظار می‌رود که با سامانه‌ها به روش‌های کاملاً متفاوتی، تعامل کنند.

توصیه می‌شود محافظت رقمی در تمامی مراحل آرشیوسازی وب، درنظر گرفته شود. در مقایسه با سایر منابع رقمی، چالش اصلی محافظت از آرشیووب، حجم بالای داده و تنوع قالب فایل‌ها و نوع رسانه‌های است. صفحات وب می‌توانند شامل تصاویر، ویدئوها، موسیقی، بازی‌ها، پایگاه‌های داده و انواع مختلفی از برنامه‌های کاربردی باشند. پیوند بین صفحات وب، ویژگی اصلی وب است که می‌تواند چالش‌های محافظت رقمی را به دلیل وابستگی‌های به وجود آمده از طریق پیوندها، ایجاد کند.

۲-۳-۳ محدودیت‌ها

آرشیوهای وب در حال حاضر هنوز شامل منابعی نه چندان قدیمی هستند و راهبردهای به اثبات رسیده با نتایجی قانع‌کننده و مطمئن وجود ندارد که ظرفیت لازم برای محافظت بلندمدت از منابع اینترنتی آرشیوشه را نشان دهنند. هدف این استاندارد، ارائه راه حل‌های کارآمد برای این موضوع نیست بلکه بر جسته کردن تلاش‌ها، استانداردها و رویه‌های فعلی در این خصوص است.

۳-۳-۳ راهبردهای محافظت

هدف محافظت رقمی در حداقل سطح، جلوگیری از گم شدن داده‌ها با حفظ یکپارچگی جریان بیت اصلی است. تدبیر اصلی برای محافظت جریان بیت، تکثیر و ایجاد نسخه پشتیبان است که شامل فعالیت‌هایی مانند ذخیره‌سازی داده به صورت موازی در مکان‌های مختلف از نظر فیزیکی، ایجاد نسخه پشتیبان منظم و بازبینی خطاهای خوانش^۱ است. همچنین لازم است داده‌ها از دسترسی‌های غیر مجاز مصون بمانند. محافظت از جریان بیت حداقل الزاماتی است که در خصوص همه منابع رقمی به کار می‌رود ولی در صورت اجرای محافظت جریان بیت، لازم است مقیاس آرشیو وب درنظر گرفته شود.

1-Read error

گذار^۱ و نمونه‌سازی^۲ راهبردهای محافظت پیچیده‌تری هستند که می‌توانند برای حفظ رفتار، عملکرد و تجربیات کاربری از منابع به کار گرفته شوند. به آن‌ها «محافظت منطقی»^۳ نیز گفته می‌شود و برای پیاده‌سازی آن‌ها نیاز به داده‌های منظم، قالب و تحلیل ریسک^۴ است. محافظت منطقی برای آرشیوهای وب با مقیاس بزرگ و قالب‌های فایل متعدد، به شدت چالش برانگیز است.

الف- گذار

گذار قالب یک فایل عبارت است از برگرداندن یک فایل به قالبی جدید، قبل از آنکه در محیط فنی کنونی غیرقابل استفاده شود. هر تبدیل به قالب فایل جدید، ممکن است باعث تغییر محتوا و آسیب احتمالی آن شود. درنتیجه، به منظور ارزیابی احتمال و پیامد از دست رفتن اطلاعات، یک فرآیند تحلیل ریسک لازم است.

گذار می‌تواند در زمانی که ریسک از منسخ شدن^۵ قالب فایل وجود داشته باشد یا در هنگام دسترسی به فایل‌ها (گذار در حال حرکت)^۶، انجام شود. هزینه گذار مستقیماً وابسته به تعداد فایل‌های تبدیلی است. گذار آرشیو وب در مقیاس بزرگ می‌تواند بسیار پر هزینه باشد. پیچیدگی و وابستگی منابع نیز بر چالش‌ها می‌افزاید و ص和尚 گذاری عملیات گذار را سخت‌تر می‌کند.

ب- نمونه سازی

نمونه‌سازی عبارت است از خلق دوباره کارکرد و رفتار محیط سامانه منسخ شده بر روی سامانه موجود با استفاده از نرم‌افزار تخصصی به اسم نمونه‌ساز. نمونه‌ساز، سامانه منسخ شده را تقلید نموده و دستیابی به منابع قدیمی بدون تغییر آن‌ها را ممکن می‌سازد. با این وجود، نمونه‌سازی تقریبی و بهندرت کامل است. خود نمونه‌ساز وابسته به یک محیط سامانه‌ای خاص و در معرض خطرهای محافظت است. توسعه نمونه‌سازها پرهزینه است اما نیاز به رسیدگی به رفع مشکلات اجزاء مستقل منابع را از میان بر می‌دارد. در آرشیوسازی وب، نیاز است که نمونه‌ساز کارکرد مرورگرهای رایج و نمایش دهنده‌های^۷ رسانه‌ها را در زمانی که صفحات وب برداشت می‌شوند، باز آفرینی کند.

1- Migration

2- Emulation

3-Logical preservation

4-Risk Analysis

5- Obsolete

6-On-the-fly

7-Player

بهتر است گذار و نمونه‌سازی بعنوان بخشی از طرح‌ریزی محافظت از سامانه‌های آرشیو رقمی، درنظر گرفته شود.

۴-۳-۳ فراداده محافظت

محافظت بلندمدت همچنین شامل حفظ ایمن فراداده‌های مربوط به منابع آرشیو و ب است که در فعالیت‌های محافظت، دسترسی و پشتیبانی از مدیریت مجموعه، نقش حیاتی دارد. انواع مختلفی از فراداده وجود دارد که می‌توانند در منابع تعبیه شده باشند، به صورت خودکار در طول فرآیندهای آرشیوسازی ایجاد شده و یا به صورت دستی توسط متقدیان اضافه شده باشند. استاندارد کدگذاری و انتقال فراداده^۱ (METS) پنج نوع فراداده را تعریف می‌کند که به صورت زیر در آرشیو وب مورد استفاده قرار می‌گیرد:

الف- فراداده توصیفی

موسساتی که آرشیو وب را فهرست یا فراداده را دستی اضافه می‌کنند، فراداده توصیفی بیشتری دارند. آنها یک آرشیوسازی وب در مقیاس بزرگ (خودکارشده) را انجام می‌دهند، متکی به استخراج فراداده تعبیه شده در منابع یا استفاده از دسته بندي یا خوشبندی خودکار برای به دست آوردن چنین فراداده‌ای هستند.

ب- فراداده ساختاری

منابع اینترنتی اغلب از اشیائی ساخته شده‌اند که مرکب از عناصر ساختار یافته و به هم پیوسته هستند. روابط ساختاری آن‌ها می‌تواند صریحاً بیان شود و در طرح‌واره‌های فراداده از قبیل "METS" ضبط شوند. چنین فراداده‌هایی می‌توانند در حالت گذار فایل، در جایی که لازم است ابرپیوندها به منظور قابلیت هدایتگری آرشیو تبدیل شوند، قابل استفاده باشند. برخی موسسات تصمیم می‌گیرند که این روابط اضافی یا تصریحی^۲ را ثبت نکنند، چرا که ذاتاً در منابع وجود دارند.

پ- فراداده اصالت^۳

فراداده اصالت به توصیف چرایی و چگونگی ایجاد یک منبع و اتفاقات چرخه حیات آن می‌پردازد. برخی فراداده‌های توصیفی که برای مثال به ضبط منطق^۴ مجموعه خاصی در آرشیو وب ضبط می‌کنند را می‌توان یک فراداده اصالت دانست. فراداده اصالت را نیز می‌توان در سطوح پایین‌تر یافت که شامل فایل‌هایی است که فعالیت‌های یک خزشگ آرشیوکننده را ضبط می‌کنند، نظریه

1-Metadata Encoding and Transmission Standard (METS)

2-Explicit

4-Provenance metadata

5-Rationale

فایل‌های پیکربندی، گزارش‌های خرز و فایل‌های ثبت وقایع و اطلاعات توصیف‌کننده تعامل بین کارساز وب و خزشگر مانند یو.آر.آل، تاریخ خرز و آدرس آی پی کارساز.

ت- فراداده فنی

فراداده فنی، ویژگی‌های فنی شیء^۱ رقمی را توصیف می‌کند و مشخص می‌کند که چگونه آن‌ها می‌توانند مورد دسترسی، اصلاح یا محافظت قرار گیرند. به این پدیده به عنوان نمایش اطلاعات در مدل مرجع سامانه اطلاعاتی دسترسی باز^۲ ارجاع داده می‌شود. قالب فایل مشخص شده با نوع ام.آی.ام.ای مثالی از فراداده فنی مرتبط با آرشیوهای وب است و یکی از آمارهای اصلی برای مشخصه‌یابی مجموعه^۳ است (برای جزئیات بیشتر به زیربند ۳-۲-۴ مراجعه شود).

ث- فراداده حقوق

این فراداده‌ها مالکیت و اجازه قانونی استفاده از منابع را مشخص می‌کنند. شرایط تعیین شده ممکن است در آینده مورد استفاده قرار گیرد. لازم است از این اطلاعات همراه با منابع آن برای جلوگیری از دسترسی‌های غیرمجاز، محافظت صورت گیرد.

به مجموعه فراداده‌ای اصالت، فراداده فنی و فراداده حقوق، داده‌های اجرایی گفته می‌شود.

۴-۳ بنیان قانونی^۴ برای آرشیوسازی وب

۱-۴-۳ کلیات

ایجاد آرشیوسازی وب با خطرات قانونی زیادی روبرو است. مهمترین آن‌ها مالکیت معنوی خصوصاً قانون حق مولف، حریم خصوصی و حافظت از داده‌های شخصی است. موسسات جمع‌آوری همچنین ممکن است مسئول انتشار مجدد مطالب غیر واقع و در اختیار داشتن و اشاعه منابع غیرقانونی باشند. قوانین ملی با محافظت ویژه قانونی از موسسات جمع‌آوری به چنین مخاطراتی به نحو اثر بخش، رسیدگی می‌کنند.

قانون‌گذاری ملی موثرترین چارچوب توانمندسازی و پشتیبانی از آرشیوسازی وب در مقیاس بزرگ است. آرشیوسازی وب ممکن است با استفاده از قوانین حق مولف و/یا سپرده‌های قانونی مرتبط و یا هر اقدامی که به‌طور خاص ماموریت‌ها وضعیت موسسه جمع‌آوری کننده را تعریف می‌کند، معرفی شود و این قانون مختص و مربوط به موسسات عمومی است که وضعیت و فعالیت‌شان از طریق قانون تعریف می‌شود همچون کتابخانه‌ها یا آرشیوهای ملی، موسسات تامین

1-Object

2-Open Access Information System Reference

3-Collection characterization

4-Legal basis

مالی حقوقی متخصص (به عنوان مثال، مؤسسات مربوط به محافظت از پخش‌ها یا فیلم‌ها)، آرشیوهای عمومی یا موزه‌ها. سازمان‌هایی بی‌بهره از این حمایت قانونی ممکن است همچنان عهده‌دار آرشیوسازی وب باشند که یا با صاحبان حقوق توافق‌های خاصی را مورد مباحثه قرار می‌دهند و یا دسترسی به منابع آرشیوشده را محدود می‌کنند تا مخاطرات قانونی را کنترل کنند. بعضی مؤسسات با دریافت و ارائه دسترسی به منابع اینترنتی براساس مجوزهای ضمنی، خود را به خطر می‌اندازند.

بعضی کشورها هنوز بنیان قانونی مشخصی برای آرشیوسازی وب ندارند و بعضی دیگر فقط چارچوب کلی دارند که برای تفسیر و قانونمندی کاربرد آن‌ها نیاز به قوانین ثانویه دارند. مؤسسات جمع‌آوری در بعضی کشورها به کمک سرمایه‌گذاری‌های داوطلبانه ناشرین، آرشیوسازی‌های وب اولیه را توسعه داده‌اند حتی در کشورهایی که قوانین ملی در جریان هستند، اقدام معمول این است که قوانین اغلب برای تفسیر باز است. مؤسسات جمع‌آوری لازم است رویکرد خود را در مورد ارزیابی خطر و تجربه در سطح کاربردی تعریف کنند.

۲-۴-۳ روش‌ها و محدوده جمع‌آوری

قانون‌گذاری آرشیوسازی وب، ممکن است به صراحت محتواهای خاصی را شامل و یا مستثنی کند. آن‌ها مرز یا حدود دامنه‌های ملی را مشخص می‌کنند و یا ممکن است تکرار یا عمق برداشت مجاز را نیز تعریف کنند.

عنصر کلیدی قانون‌گذاری این است که آیا باید قبل از جمع‌آوری منابع از صاحبان حقوق اخذ مجوز کرد. زیرا این امر تاثیر مهمی بر راهبردهای برداشت موسسات دارد. برداشت انبوه فقط در صورتی که نیازی به مجوز گرفتن نباشد امکان‌پذیر است درغیراینصورت برداشت انتخابی، الگوی مناسبتری است. رویکرد جایگزین برای مدیریت اخذ مجوز، الگوی به اصطلاح «صرفنظرکردن» یا "اهمیت دادن و حذف کردن" است که منابع بر اساس بدیهی و ضمنی فرض شدن اجازه و امکان استفاده برداشت می‌شوند و به هنگام تقاضای مالکان حق می‌توانند حذف شوند. "اجازه سرپوشی"^{۱۰} تنها از یک ناشر به منظور دسترسی به منابع متعدد، راه دیگری برای کم کردن هزینه‌های تواجهات متعدد با ناشران است.

قانون‌گذاری ممکن است، برای کمک به بهتر شدن کیفیت، مجوزهای لازم را به موسسه جمع‌آوری کننده، اعطای کند. نمونه‌هایی از این اطلاعات شامل فهرست نام یا رمز شناسایی دامنه‌های ملی و اطلاعات مدیریت حقوقی رقمی در مورد انتشارات هزینه‌دار است. قانون‌گذاری می‌تواند به طور خاص ناشران را مجبور به ارائه اطلاعات نماید و متخلفان را جریمه کند.

قانون گذاری، حتی ممکن است فنون جمع‌آوری خاصی را پیشنهاد یا اجبار کند. برای مثال، صراحتا در برخی موارد اجازه برداشت خودکار منابع اینترنتی به شرط موافقنامه دوطرفه بین موسسه و ناشر بر روی پروتکل برداشت را می‌دهد یا آن را ترغیب می‌کند.

قانون گذاری می‌تواند محدودیت‌هایی بر چگونگی آرشیوسازی وب را نیز اعمال کند. از دیگر محدودیت‌های اعمالی از طرف قانون گذاری اجرا یا صرفنظر کردن از استانداردهای محدودسازی روابط‌سازی‌ایکس‌تی است. در واقع چنین اقدامی می‌تواند تفاوت فاحشی بر حفظ یا حذف محتوی مشخص، پدید می‌آورد.

در کشورهایی که دستور قانونی برای جمع‌آوری منابع اینترنتی ملی وجود دارد، مسئولیت این امر به موسسه خاصی اعطا می‌شود. البته امکان دارد که این مسئولیت بین چند سازمان، تقسیم شود. برای مثال:

- بین مؤسسات ملی و موسسات محلی و منطقه‌ای
- بین کتابخانه‌های ملی و آرشیوهای ملی؛
- در شبکه‌ی انتلاقی از موسسات متخصص.

۳-۴-۳ دسترسی به آرشیوهای وب

دسترسی‌پذیری^۱ شرایط لازمی که بتوان از آرشیوب استفاده کرد را تعریف می‌کند و بعد مهمی از قانون گذاری است. به طور کلی شرایط دسترسی با قوانین برداشت سازگار است: به عبارتی، چنانچه به اخذ مجوز از صاحبان حقوق نیاز باشد، اجازه دسترسی برخط ممکن است بر حسب مورد صادر شود و در صورت برداشت اینبوه بدون اجازه، احتمال محدودیت دسترسی، بیشتر است.

محدودیت‌های قانون حق مولف مثل انتشار، استخراج، رونوشت برداری الکترونیکی یا بارگذاری می‌تواند در آرشیوهای وب، اعمال شود.

در صورتی که آرشیو وب به عنوان نوعی ذخیره قانونی اجرا شود، ممکن است قانون گذاری ممکن است انتشار فهرست^۲ ملی را برای آرشیوب، الزام کند که در مورد جمع‌آوری در مقیاس بزرگ، بسیار چالش‌برانگیز است. آرشیو وب بجای چاپ فراداده‌های توصیفی، عموماً واسطه‌های جستجو را به کاربران نهایی^۳ پیشنهاد می‌کند.

۴-۴-۳ حفاظت از آرشیو وب :

1-Accessibility

2-Bibliography

3-End users

حفظ اسناد ملی مدت محدود میراث فرهنگی است. بنابراین برای حصول اطمینان از دوام آرشیوهای وب، ممکن است راهنمایی یا دستوراتی در قانون گذاری درنظر گرفته شود. به ویژه حتی ممکن است اجزاء حذف منابع یا محافظت دائمی آنها تعیین شود. اغلب کتابخانه‌های ملی که منابع اینترنتی را با سرمایه‌گذاری قانونی جمع‌آوری می‌کنند، وظیفه حفظ آن برای آینده‌گان را برعهده دارند. در صورتی که هدف از جمع‌آوری، تهیه مجموعه داده‌ها برای پژوهش‌های کوتاه و میان‌مدت باشد، کتابخانه‌های پژوهشی و دیگر موسسات ممکن است ملزم به حفظ منابع به مدت نامحدود نباشند.

۵-۳ دلایل دیگر برای آرشیوسازی وب

۱-۵-۳ کلیات

انگیزه‌های دیگر برای آرشیوسازی وب معمولاً وابسته به سیاست و خط مشی است. این انگیزه‌ها چشم‌انداز راهبردی موسسه‌ای و نیز قرارداد و نگرش موسسه را نسبت به نوآوری‌های فرهنگی و فناوری بازتاب می‌دهد.

وب گستره متنوعی از منابع رقمی‌زاد و رقمی شده را میزبانی می‌کند. منابع رقمی شده برای ارائه اطلاعات چاپی (نظیر کتاب‌ها، گاهنامه‌ها و انتشارات دولتی) یا داده‌هایی که بر روی سایر رسانه‌های فیزیکی وجود دارند (نظیر فیلم، موسیقی، بازی، لوحی یا نوار)، استفاده می‌شود که بیشتر آن‌ها قبلاً فرآیند تبدیل یا گذار را پشت‌سر گذاشته‌اند. وب به سرعت تکامل می‌باید و بی‌دوام است. منابع ارزشمند به طور مرتبت ناپدید می‌شوند. محافظت از وب به عنوان وظیفه‌ای برای موسسه‌هایی با مأموریت بلندمدت در حیطه حفظ منابع میراث فرهنگی طبیعی و حیاتی، درنظر گرفته می‌شود.

آرشیوسازی وب، تداوم محتوای رقمی را تضمین می‌کند و برای جلوگیری از ایجاد خلاء در حوزه حافظه و دانش یک ملت یک فعالیت ضروری است. همچنین می‌تواند به حفظ دسترسی به تحقیقات ثبت شده کمک کند. این موارد به ویژه انگیزه بسیار قوی برای آرشیوها و کتابخانه‌های ملی است.

۲-۵-۳ تسهیل پژوهش‌های علمی

اینترنت، فضایی بسیار خلاقانه و مشارکتی است که افراد در آن با هم به تعامل و همکاری می‌پردازند. می‌توان گفت وب دانش اجتماعی جدید و ابزارهای پژوهشی تازه‌ای به وجود آورده است که بخشی از میراث ملی شمرده می‌شوند. هم‌اکنون ظهور روش‌ها و گروه‌های پژوهشی جدید ناشی از مطالعه وب زنده و آرشیوهای آن را می‌توان در همه‌جا مشاهده کرد.

آرشیوب فرست امکانات تحقیقاتی بی‌نظیری را برای پژوهشگران و دانشمندان اینترنت فراهم آورده است. آن‌ها نه تنها می‌توانند از نسخه‌های بازنگرانه یا تاریخی وب‌گاه‌ها بهره بگیرند بلکه می‌توانند با انجام داده کاوی و پیوند کاوی‌های وسیع به کشف الگوهای رمزگشایی مسائل علمی

کمک کنند. اگرچه هنوز داده‌کاوی و تحلیل داده‌ها در ابتدای مسیر پیشرفت است و اغلب مورد استفاده دانشمندان اجتماعی قرار می‌گیرد، اما احتمال توسعه آن در حیطه‌های علمی دیگر و مفید واقع شدن در بسیاری از رشته‌ها وجود دارد.

در سطح موسسات، آرشیوسازی وب می‌تواند راهی برای ارتقا و تأکید بر منابع رقمی خاص باشد. این روش به‌ویژه برای موسسه‌هایی چون دانشگاه‌ها که آثار دانشجویان و پژوهشگران خود را آرشیو می‌کنند، قابل استفاده است. آرشیوسازی وب اقدام ارزنده‌ای در جهت جمع‌آوری منابع برخط با اهمیت پژوهشی است و این امر انگیزه بالایی برای موسسات در جهت تولید این آرشیو به وجود می‌آورد.

۳-۵-۳ قابلیت پشتیبانی از انواع روش‌های استفاده توسط عموم کاربران

اینترنت، میزبان به اشتراک‌گذاری‌هایی است که در تمام جنبه‌های زندگی انسانی حضور دارد. برخلاف رسانه‌های چاپی، هر کسی توانایی انتشار روی وب را دارد. اگرچه ارزش آن‌ها می‌تواند یکسان نباشد اما این به اشتراک‌گذاری‌ها، مجموعه یکپارچه و بی‌نظیری را به وجود می‌آورد که بازتاب دهنده افکار و تعاملات فردی و اجتماعی است.

انواع خدمات تخصصی یا شخصی را می‌توان به کمک آرشیوهای وب ارائه داد. از آن‌ها می‌توان به عنوان شاهد در مسائل حق مؤلف استفاده کرد یا برای پژوهش‌های شخصی یا خانوادگی و شجره‌نامه رقمی از آن‌ها استفاده کرد. دسترسی آزاد و بلندمدت به منابع اینترنتی به ویژه برای نویسنده‌گان مطالب اینترنتی در حال و آینده توجیه محکمی برای ارائه پیشنهاد آرشیوهای وب به عنوان یکی از خدمات اجتماعی و عمومی است.

۴ آمارها

۱-۴ کلیات

آمارها داده‌هایی عینی هستند که مبنایی را برای تحلیل‌ها و تفسیرهای آینده ایجاد می‌کنند. شاخص‌های کیفیت، رتبه‌ای را برای داوری ارزش^۱ تعیین می‌کنند: نرسیدن به ملاک مورد نظر، نشانه برآورد منفی است. در این استاندارد آمارها با اعداد مطلق و نشانگرهای کیفی با اعداد نسبی و درصدها مشخص شده‌اند.

آمارها و شاخص‌های کیفی لازم است قابل اطمینان، آگاهی‌دهنده و قابل مقایسه و روش‌های دستیابی به آن‌ها لازم است کاربردی و انعطاف‌پذیر باشند. فنون شناخته شده روز کنونی در آرشیوسازی وب چنین است که تولید برخی آمارها متکی بر نرم‌افزارهای برداشت، نمایه‌سازی و یا جستجو است و تفاوت در انتخاب نرم‌افزار، ممکن است به اختلاف در نتایج منجر شود. بنابراین

1-Value Judgement

توصیه می‌شود که اگر هدف محکزنی^۱ است، آمارها براساس استفاده از نرم‌افزارهای یکسانی تولید شود. با توجه به آنکه آرشیوهای وب عموماً مقیاس وسیع دارند، اندازه‌گیری کیفیت آن‌ها با بهره‌گیری از آمارها، لازم، کاربردی و مقرر بود. شاخص‌های کیفی و آمارهای معرفی‌شده در این استاندارد، براساس رویه رایج در تهیه آرشیوسازی‌های وب تنظیم شده است. این رویه در بلندمدت دیدگاه کلیدی قابل اتكابی از آرشیوهای وب به دست می‌دهد و امکان مقایسه آن‌ها را فراهم می‌آورد.

این استاندارد آمارها و شاخص‌های کیفی عمومی را ارائه می‌دهد که البته همه آن‌ها قابل به کارگیری برای انواع آرشیوهای وب نیستند و همچنین با پیشرفت فناوری و آرشیوهای وب نیاز است که بعضی از آن‌ها روزآمد شوند.

این فصل از استاندارد به معرفی و توصیف تعدادی از آمارهای مربوط در هر بخش می‌پردازد. اما مجموعه کوچکتری از آمارهای اصلی ضروری درنظر گرفته می‌شود که در جداولی در آخر هر بخش همراه با مثال ارائه شده است.

۲-۴ آمارهای مربوط به توسعه مجموعه اطلاعات

۱-۲-۴ کلیات

آمارهایی که درادامه ارائه خواهد شد، رشد آرشیوب را با دنبال کردن برونداد کمی آن‌ها اندازه‌گیری می‌کند. این امر به طرح‌بازی و پایش گسترش مجموعه کمک می‌کند و امکان انجام تحلیل‌های دقیق را فراهم می‌سازد.

برخلاف مدارک آنالوگ، آرشیوب، حاوی منابع غیرخطی بهم‌پیوسته^۲ است. برخی از اسناد به قصد بازدیدهای مجدد توسط کاربران انسانی مورد استفاده قرار می‌گیرند و برخی دیگر فایل‌ها و فراداده‌های جداناپذیری^۳ هستند که قسمتی از منابع به حساب می‌آیند اما برای کاربران قابل رؤیت نیستند. بنابراین آمارهایی که ظرفیت آرشیوب را اندازه می‌گیرند قابل مقایسه با آمارهای اندازه‌گیری آرشیوهای فیزیکی نیستند.

اغلب این آمارها، مخصوص منابع آرشیو وب بوده و برای محتوای وب زنده، کاربردی نیست.

1-Benchmarking

2-Interlinked

3-Inseparable

۲-۲-۴ اندازه‌گیری اهداف جمع‌آوری: شمارش مقاصد آرشیو و نسخه‌های برگرفته شده

۱-۲-۲-۴ هدف

توصیه‌می‌شود موسسات جمع‌آوری اطلاعات قادر به بیان و ارزیابی اهداف آرشیوسازی وب بر حسب خطمشی‌هایشان باشند تا نتایج حاصل از جمع‌آوری آرشیووب با توجه به هدف قابل ارزیابی باشد و بتوان سطح موفقیت و اثربخشی فرایند جمع‌آوری را تعیین نمود.

روش یکسان و مشخصی برای بیان اهداف یا مقاصد آرشیو جمع‌آوری وجود ندارد. این فرآیند را هر موسسه براساس اهداف و خطمشی‌های خود، تعریف می‌کند. یک چارچوب کلی وجود دارد که در آن استفاده از مفاهیم "مقصد آرشیو" و "نسخه برگرفته شده از مقصد آرشیو"^۱ می‌تواند به اندازه‌گیری فعالیت کارکنان دخیل در انتخاب و مدیریت مقاصد آرشیو، کمک کند. این موضوع شاخصی برای میزان تلاش صورت گرفته جهت گزینش است که زمان صرف شده برای این که کدام وب‌گاه‌ها در آرشیووب قرار گیرند را نشان می‌دهد.

۲-۲-۲-۴ روش

هر مقصد آرشیو شامل یک یا چند نقطه شروع است و هر نقطه شروع هم مجموعه‌ای از تنظیمات خرزش دارد که محدوده آن را مشخص می‌کند. مقصد، مجموعه‌ای از منابع معنی‌دار قابل جمع‌آوری است و محدوده آن می‌تواند از منابع به هم پیوسته در یک دامنه که به عنوان یک وب‌گاه ارائه شده است تا یک منبع منفرد که با یک یو.آر.آل (مثالاً پی‌دی‌اف یا ویدئو) شناخته می‌شود یا کل یک دامنه سطح بالا گسترده باشد. هر مقصد آرشیو می‌تواند بیش از یک خرزش داشته باشد. هر خرزش یک نسخه برگرفته شده از مقصد آرشیو است.

مثال:

- خرزش روزانه صفحه اصلی وب‌گاه نیویورک تایمز
- نقاط شروع می‌توانند <http://global.nytimes.com> و <http://www.nytimes.com> باشند.
- محدوده می‌تواند "خرش صفحه اصلی و همه منابع با فاصله یک کلیک بر صفحه اصلی" باشد.
- تکرار می‌تواند "هر روز" باشد.
- مقصد آرشیو تمامی موارد بالاست.
- مجموعه خرزش منفرد و ذخیره‌سازی منابع مشخص شده در بالا، یک نسخه برگرفته شده از مقصد آرشیو محسوب می‌شود.

این روش عام، این امکان را به موسسات می‌دهد تا اهداف عملی را تعیین و بروزن داد را اندازه‌گیری کنند. هدف در مثال بالا می‌تواند جمع‌آوری ۳۶۵ نسخه برگرفته شده از مقصد آرشیو

مشخص شده در سال باشد. بنابراین، برای مشخص شدن این که دستیابی به هدف حاصل شده است، کافی است که تعداد نسخه‌های برگرفته شده از مقصد آرشیو در طول یک سال با تعداد مورد انتظار (۳۶۵ نسخه) را مقایسه کرد.

۳-۲-۴ محدودیت‌ها

مقایسه تعداد مقاصد آرشیو و تعداد نسخه‌های برگرفته شده از مقصد آرشیو بین موسسات تنها در صورتی معنادار است که از رویه‌ها و خط‌مشی‌های یکسانی پیروی کنند.

۳-۲-۴ اندازه گیری ابعاد آرشیو وب: شمارش یو.آر.آل‌ها

۱-۳-۲-۴ هدف

شمارش یو.آر.آل‌ها یکی از راه‌های اندازه گیری اندازه آرشیو وب است. یو.آر.آل‌ها متناظر با مکان منابع خرزش شده هستند و به وسیله خرزشگر آرشیو کننده برای شناسایی و درخواست منابع از کارساز وب مورد استفاده قرار می‌گیرند. کارساز وب گسترها از پاسخ‌های استاندارد را که با کدهای وضعیت شناسایی می‌شوند، باز پس می‌فرستد که نشانگر وضعیت منبع درخواست شده است. این پاسخ ممکن است تأییدی باشد که منبع درخواستی با موفقیت تحویل داده شده است، یا اینکه پاسخی باشد مبنی بر این که منبع درخواستی به مکان دیگر منتقل شده است (تغییر مسیر). همچنین پاسخ ممکن است یک پیام خطا باشد مبنی اینکه منبع درخواستی در دسترس نیست. بعضی پاسخ‌ها هم فراداده و هم محتوا را شامل می‌شوند و برخی تنها حاوی فراداده هستند و برخی فقط کد خطا را ارائه می‌دهند.

باید توجه داشت که همه یو.آر.آل‌ها همانند "مدارک" یا "اقلام" فیزیکی موجود در کتابخانه‌ها، منابع معنی‌دار و قابل خوانش برای انسان نیستند. حتی زمانی که کارساز وب در انجام درخواست‌ها ناموفق باشد باز یک پاسخ ارسال می‌کند که وضعیت تحویل درخواست مانند تغییر مسیر یا خطاها را نشان می‌دهد. چنین اطلاعاتی مسیر ممیزی^۱ برای فرآیند برداشت و اطلاعات اصل برای جمع‌آوری آرشیو وب را فراهم می‌آورد و می‌تواند برای اهداف دسترسی یا محافظت بسیار مفید باشد. بنابراین در این استاندارد توصیه می‌شود که همه پاسخ‌ها به عنوان بخشی از آرشیو وب، حفظ شوند.

یو.آر.آل‌ها به عنوان شناساگرهای منابع در وب و سامانه پیام اج.تی.تی.پی مورد استفاده قرار می‌گیرند. یو.آر.آل‌ها همچنین کوچک‌ترین واحد با محتوای خودشمول^۲ در آرشیو وب هستند و عموماً در سامانه‌های ذخیره و دسترسی برای منابع آرشیو شده در وب استفاده می‌شوند. بنابراین پیشنهاد می‌شود که یو.آر.آل‌ها برای شناسایی منابع و منابع متناظر بازفرستاده توسط کارساز وب

1-Audit trail

2-Self-contained

استفاده شوند. انواع مختلف کدهای وضعیت می‌تواند برای مرتبسازی و گروهبندی منابع در آرشیو وب استفاده شود.

۴-۳-۲ روش

جدول ۱ حاوی کدهای وضعیت مختلفی است. این کدها به صورت اعداد سه رقمی هستند و اولین رقم، کلاس پاسخ‌ها را مشخص می‌کند. هر کدووضعیت یک عبارت توضیحی^۱ دارد که توصیف کوتاهی از آن کدووضعیت را برای کاربران انسانی ارائه می‌دهد. جزئیات بیشتر در مورد کد وضعیت را می‌توان در آر اف سی^۲ ۲۶۱۶ یافت (به کتابنامه مراجعه شود).

جدول ۱: فهرست کدهای وضعیت اچ.قی.قی.پی

کد وضعیت	عبارت توضیحی
اطلاعاتی	1XX
ادامه	100
پرونکل‌های راه‌گزینی	101
موفقیت‌آمیز	2XX
صحیح	200
محتوٰ ساخته شد	201
پذیرفته شده اما فعلاً بر مبنای آن اجرا نشده است	202
اطلاعات غیر معتبر	203
بدون محتوٰ	204
تنظیم مجدد محتوٰ قبلًا فرستاده شده است	205
محتوٰ ناتمام	206
تغییرمسیر	3XX
انتخاب‌های چندگانه	300
انتقال دائمی	301
یافت شد	302
به دیگری رجوع شود	303
تغییرنیافته	304

1-Reason phrase

2-Request For Comments

کد وضعیت	عبارت توضیحی
به کارگیری پروکسی	305
تغییر مسیر موقت	307
خطای کارخواه	4XX
درخواست نامناسب	400
غیرمجاز	401
مستلزم پرداخت	402
ممنوع	403
یافت نشد	404
روش غیرمجاز	405
غیرقابل قبول	406
اصالت سنگی پروکسی مورد نیاز است	407
اتمام زمان درخواست	408
تعارض	409
ازبین رفته	410
طول مورد نیاز است	411
پیش شرط ناموفق	412
درخواست هستار بسیار بزرگ	413
درخواست یو.آر.آل بسیار بزرگ	414
نوع رسانه پشتیبانی نشده	415
محدوده درخواستی رضایت‌بخش نیست	416
برآورده نشدن انتظار	417
خطای کارساز	5XX
خطای کارساز داخلی	500
پیاده‌سازی نشده	501
درگاه نامعتبر	502
خدمت غیرقابل دسترس	503
پایان زمان مسیر درگاه	504
نسخه اج.تی.تی.پی پشتیبانی نمی‌شود	505

این استاندارد توصیه می‌کند که همه یو.آر.آل.ها، بدون درنظر گرفتن کدهای وضعیت مرتبط، درهنگام محاسبه تعداد کل منابع (برداشت‌شده)، در آرشیو وب بهشمار آیند.

البته مهم است که ماهیت و معنی کدهای وضعیت را بشناسیم چرا که می‌توانند در گروه‌بندی و پالیش منابع به هدف تحلیل اجزای خاص آرشیو وب استفاده شوند. برای مثال سریال‌های کد وضعیت xx2 نشانگر ارسال موقتی آمیز منبع درخواستی هستند و سریال‌های xx3 معمولاً فقط فراداده را بدون منبع درخواستی پس می‌فرستند. سری‌های xx5 برای مقاصد فنی مورد استفاده قرار می‌گیرند. سریال‌های xx2 از اهمیت خاصی برای موسساتی بخوردار است که بخواهند تعادلی بین مجموعه فیزیکی و آرشیو وبشان ایجاد کنند. توصیه می‌شود به خصوص یو.آر.آل.هایی با کد وضعیت‌های ۲۰۰، ۲۰۱، ۲۰۳، ۲۰۴ و ۲۰۵، شمارش شوند.

تعداد یو.آر.آل‌ها ممکن است قبل یا بعد از حذف تکرارها^۱ شمرده شود (اگر فرآیند حذف تکرارها انجام شود). حذف تکرارها در طول یک خرز زمانی روی می‌دهد که یک آدمواره موجود بودن یو.آر.آل در آرشیو را تشخیص می‌دهد و دوباره آن را مرور نمی‌کند. آدمواره ممکن است اطلاعاتی را تولید کند که نشان‌دهنده فرایند حذف تکرارها باشد که به آن پیشینه "بازدید مجدد"^۲ در دبلیو.ای.آر.سی گفته می‌شوند. هردوی این اعداد مفیدند:

- تعداد یو.آر.آل‌ها بعد از حذف تکرارها، نشانگر مقدار منابع موجود در آرشیو است. این عدد شماره قابل ارجاعی است که برای پیش‌بینی حجم حافظه مورد نیاز و محافظت دراز مدت قابل استفاده است.

- تعداد یو.آر.آل‌ها قبل از فرایند حذف تکرارها مخصوص کاربران انسانی است و از جنبه منطقی و محتوایی حائز اهمیت است. این موضوع تنها زمانی قابل توجه است که اطلاعات فرآیند حذف تکرارها هنوز در دسترس باشد (برای مثال وجود پیشینه‌های بازملاقات در فایل‌های دبلیو.ای.آر.سی). این اطلاعات در واقع نشان می‌دهند که یک یو.آر.آل حتی اگر آدمواره آن را خرز نکرده باشد، در یک تاریخ خاص هنوز فعل بوده است. توصیه می‌شود روش محاسبه (قبل و بعد از حذف تکرارها) همواره و به خصوص در هنگام مقایسه بین آرشیوها، مشخص باشد.

۳-۲-۴ محدودیت‌ها

وب در روزهای نخست بیشتر از صفحات اج.تی.ام.آل ایستا با منابع مستقیماً ارجاع داده شده، تشکیل شده بود. اما به سرعت تکامل یافته و در حال حاضر حجم انبوهای از محتواهای تعاملی و

1-De-duplication

2-Revisit

پویا را دربرمی‌گیرد که نیاز است به آن‌ها بارویکردی ورای نگاه سنتی به وب به عنوان مجموعه‌ای از مدارک و انتشارات اج.تی.ام.ال خود شمول نگریسته شود. تعداد یو.آر.ال‌ها در یک آرشیو وب با تعداد مدارک و انتشارات در یک کتابخانه به مفهوم سنتی معادل نیست. در محاسبه‌های آماری برای آرشیووب، درنظر گرفتن طبیعت وب به عنوان منبع برخط شبکه‌ای و بهم‌پیوسته، مهم است. توصیه می‌شود، آمارها تنها منابع قابل استفاده برای کاربران انسانی را شامل نشوند، بلکه فراداده‌ها و برنامه‌هایی که بخش جدایی‌ناپذیر آرشیووب هستند را نیز شامل شوند.

قابل ذکر است که همه کدهای وضعیت که از طرف کارساز برمی‌گردد، قابل اطمینان نیستند. در ادامه به نمونه‌هایی اشاره می‌شود.

الف- مفقودی ۴۰۴

بسیاری از کارسازهای وب، کد وضعیت ۴۰۴ مناسبی را در مقابل منبع درخواستی غیرقابل دسترس ارسال نمی‌کنند و به جای آن ممکن است پاسخ "OK 200" را به همراه بلاکی از محتوا ارسال کنند که توضیح می‌دهد منبع مورد نظر موجود نیست. در این صورت، راهی برای خزشگر آرشیوکننده به منظور تشخیص "پیدا نشدن ۴۰۴" وجود ندارد. بنابراین چنین پاسخی باید به عنوان پاسخ "خوب" یا "موفق" محسوب شود.

ب- تکرار با شناسه (ID) نشست‌های متفاوت^۱

بسیاری از کارسازهای وب، یو.آر.ال‌ها را به شکل پویا تولید می‌کنند که منجر به تکرار منابع در آرشیووب می‌شود. کارساز وب گاهی منابع را با پیوست یک شناساگر یکتا^۲ به هر یو.آر.ال به کاربر- عامل، برمی‌گرداند تا امکان ردیابی نشست وجود داشته باشد. برای مثال

کاربر- عامل ۱ چنین یو.آر.الی را می‌گیرد:

<http://www.example.cpm/id=12345/picture.jpg>

کاربر- عامل ۲، یو.آر.الی با شناسه مختلف^۳ متفاوت می‌گیرد:

<http://www.example.cpm/id=67890/picture.jpg>

این دو یو.آر.ال مربوط به منبع یکسانی هستند که در این مورد فایل تصویری است، اما از شناسه‌های مختلفی برای شناسایی کاربران استفاده می‌کنند. خزشگر آرشیو کننده امکان دارد یک منبع را با چندین یو.آر.ال مختلف جمع‌آوری کند و این نسخه‌ها به عنوان منابع واحد محسوب شوند.

1-Duplicate with different session ID

2-Unique Identifier

3-Different ID

پ- فقدان کدهای وضعیت

در آغاز فعالیتهای وب، معمول بود که کارساز فقط منبع درخواستی را بدون فراداده یا کد وضعیت بازگرداند. گاهی به آن اجتناب پیش نسخه ۰.۹ گفته می‌شد. بعضی کارسازها ممکن است هنوز از پروتکل‌های قدیمی استفاده کنند. فقدان کد وضعیت و فراداده به خصوص برای موسساتی که مجموعه‌های تاریخی طولانی مدت وب را جمع‌آوری می‌کنند، می‌تواند چالش برانگیز باشد.

۴-۲-۴ سنجش اندازه آرشیو وب: شمارش دامنه‌ها یا میزبان‌ها

۱-۴-۲-۴ هدف

شمارش دامنه‌ها یا میزبان‌ها روش دیگری برای سنجش اندازه آرشیو وب است. موسسات جمع‌آوری عموماً از این روش شمارش استفاده می‌کنند که در عمل اغلب به عنوان جایگزینی برای تعداد وب‌گاه‌ها است. وب‌گاه، برعکس، یک واحد منطقی مفهومی است که از مجموعه‌ای از صفحات وب به هم پیوسته تشکیل شده است و نمایانگر یک فرد یا گروه یا سازمان است. اما چیزی نیست که بتوان آن را از لحاظ فنی تعریف کرد یا از لحاظ کاربردی یا نظامی اندازه‌گیری نمود.

دامنه‌ها و میزبان‌ها به صورت نظامی قابل اندازه‌گیری هستند اما نباید آن‌ها را معادل خود وب‌گاه‌ها به شمار آورد زیرا آن‌ها صرفاً جهت نام‌گذاری و تعیین محل وب‌گاه‌ها به کاربرده می‌شوند. این آمارها می‌توانند برای توصیف دقیق جمع‌آوری یا تحلیل فنی، به منظور تعیین نوع منابع در یک آرشیو وب (مانند www.org) یا تشخیص این‌که محدوده مورد نظر توسط خزشگر پویش شده است یا نه، مفید واقع شوند.

۲-۴-۲-۴ روش

تعداد دامنه‌ها یا میزبان‌ها می‌تواند از طریق گزارش‌های خزشگر یا سایر روش‌های تحلیل فایل‌های ذخیره‌شده به صورت خودکار، محاسبه شود.

۳-۴-۲-۴ محدودیت‌ها

شمارش دامنه‌ها یا میزبان‌ها محدودیت‌هایی دارد. مانند شمارش یو‌آل‌ها، این فرایند نیز معمولاً تعدادی را برای منابع موجود در آرشیو گزارش می‌کند که بیشتر از تعداد منابع قابل روئیت و خواندن برای انسان است. همه دامنه‌ها به منابع فعل و معنی دار ارجاع ندارند و نام‌های مستعار و تکراری نیز وجود دارد.

الف- دامنه‌های غیرفعال

نام دامنه می‌تواند خریداری شود، اما غیرفعال باشد و به هیچ منبعی نیاجامد. برخی دامنه‌ها هم منبع دارند اما اغلب فقط به یک صفحه وب با پیشنهاد نام دامنه برای فروش می‌انجامند. مورد اول

از طریق کد وضعیت ۲۰۴ قابل شناسایی و درمورد بعدی اگر بخواهیم دقیق گفته باشیم، چون دامنه غیرفعال نیست، در شمارش دامنه‌ها محاسبه خواهد شد. اگرچه از چشم‌انداز متصدی گرایانه، چنین منابعی ارزش خاصی از لحاظ محافظتی ندارند. درهنگامی که برداشت به صورت انتخابی انجام می‌شود می‌توان چنین مواردی را در طول فرآیند انتخاب، حذف نمود. هرچند روش آسانی برای شناسایی خودکار و پالایش آن‌ها از برداشت‌های انبوه وجود ندارد مگر بازبینی دستی منابع با حجم بسیار اندک.

در صورت امکان توصیه می‌شود که دامنه‌های غیرفعال در آرشیو از طریق نمونه‌برداری پیگیری شود تا به اصالت جمع‌آوری و ارزیابی اثربخشی تضمین کیفیت کمک شود.

ب- نام‌های مستعار^۱

نام مستعار، جایگزینی برای نام دامنه است. نام مستعار این امکان را فراهم می‌سازد که یک وب‌گاه در یک دامنه میزبانی شود و نام دامنه دیگری روی آن گذاشته شود. نام‌های مستعار زیادی در اینترنت وجود دارند، زیرا صاحبان دامنه‌ها ممکن است تمایل داشته باشند از نام‌های زیادی استفاده کنند تا هرچه بیشتر در معرض دید کاربران باشند. نام‌های مستعار بیشتر از طریق تغییر مسیر به کارگرفته می‌شوند.

درگزارش‌های تولیدشده توسط خزشگرهای آرشیوکننده، نام‌های مستعار علی‌رغم ارجاع به منابع یکسان، دامنه‌های منحصر به فرد شمرده می‌شوند. پیداکردن نام‌های مستعار نیازمند مقایسه چشمی بین صفحات یک کارساز یا مقایسه جمع‌آزماهای^۲ آن‌ها است. پیداکردن نام‌های مستعار در آرشیوهای انتخابی آسان‌تر است، اما احتمال وجود آن‌ها در برداشت‌های انبوه، بیشتر است. وجود نام‌های مستعار، منجر به تکرار منابع در آرشیو می‌شود.

در صورت امکان بهتر است نام‌های مستعار کشف شده در آرشیو به منظور کمک به توصیف مجموعه و حذف تکرار منابع، ردیابی شوند.

۵-۲-۴ اندازه‌گیری آرشیو وب: شمارش بایت‌ها

۱-۵-۲-۴ هدف

اندازه آرشیووب را می‌توان بر حسب تعداد بایت‌های آن نیز اندازه‌گیری کرد. این کار، آمار مناسبی را برای طرح‌ریزی ذخیره و دیگر منابع فراهم می‌آورد. این روش با به کارگیری مقیاس خطی (متر) برای مدیریت قفسه‌های کتاب در کتابخانه قابل قیاس است.

1- Aliases
2-Checksums

۴-۵-۲- روش

اندازه یک آرشیو بر حسب بایت را می‌توان به صورت خودکار توسط اضافه کردن اندازه منابع خرزش شده از گزارش‌های خرزش یا از طریق دیگر روش‌های خودکار که فضای ذخیره‌سازی استفاده شده یک آرشیو را بررسی می‌کنند، محاسبه نمود.

آرشیوهای وب اغلب مقیاس بزرگی دارند. اندازه آن‌ها می‌تواند از مجموعه‌های کوچک چند صد گیگابایتی تا مجموعه‌های ملی چند صد ترابایتی باشد. در جمع‌آوری وب جهانی و داشتن یک آرشیو بلند مدت، اندازه آرشیو امکان دارد حتی به پتابایت هم برسد. یک راه ذخیره آرشیو وب فشرده‌سازی داده‌هاست. برای مثال، پیوست ت، "ویژگی قالب فایل دبلیو.ای.آر.سی" از استاندارد ISO 28500 که استاندارد آرشیو وب است، چگونگی استفاده از فشرده‌سازی جی.ذ.آی.پی برای دبلیو.ای.آر.سی را توضیح می‌دهد.

اندازه آرشیو وب می‌تواند هم به صورت فشرده وهم غیرفشرده، اندازه‌گیری شود. هنگام همترازی آرشیوها، استفاده از معیارهای یکسان حائز اهمیت است. به عبارت دیگر، توصیه می‌شود اندازه فشرده شده آرشیوب با اندازه غیرفشرده وب دیگر، مقایسه نشود.

- اندازه فشرده شده، مقدار فضای دیسک استفاده شده برای ذخیره‌سازی منابع را بازنمایی کند و می‌توان از آن به عنوان اندازه مرجع در ذخیره‌سازی و محافظت بلندمدت استفاده نمود.

- اندازه فشرده نشده، حجم منابع را بر روی وب زنده نشان می‌دهد. این مورد برای کاربران انسانی درنظر گرفته شده و از لحاظ منطقی و محتوایی حائز اهمیت است.

اندازه آرشیو وب به دلایل ذکر شده در زیربند ۴-۳-۲، می‌تواند قبل و بعد از حذف تکرارها نیز اندازه‌گیری شود. توصیه می‌شود روش محاسبه مجدداً به روشنی توضیح داده شود.

یکی از روش‌های متداول این است که فایل‌ها را در محفظه فایل‌ها^۱ مانند فایل‌های ای.آر.سی یا دبلیو.ای.آر.سی ذخیره کنند. محفظه‌بندی، فایل‌های متعدد را تجمیع کرده و باعث آسان شدن مدیریت و ذخیره‌سازی فایل‌ها از طریق کار با چند فایل بزرگ به جای تعداد بی‌شماری فایل‌های کوچک می‌شود. فایل‌های محفظه‌بندی شده امکان ذخیره فراداده‌ها به همراه منابع جمع‌آوری شده را فراهم می‌آورند. همچنین تعداد فایل‌های محفظه‌بندی شده آمارمی‌فیدی برای آرشیوسازی وب است زیرا اغلب از آن‌ها به عنوان واحدهای زیربنایی مدیریت ذخیره‌سازی، تبادل داده و گاهی در اهداف محافظت بلندمدت استفاده می‌شود.

1- Container files

۶-۲-۴ آمارهای اصلی برای توسعه جمع‌آوری

جدول ۲-آمارهای اصلی برای توسعه جمع‌آوری

مثال	هدف	آمارها
۸۰۰۰ مقصد آرشیو	هدف جمع‌آوری / بروندادهای کمی	تعداد مقاصد آرشیو
۱۴۰۰۰ نسخه‌های برگرفته شده از مقصد آرشیو	هدف جمع‌آوری /بروندادهای کمی	تعداد نسخه‌های برگرفته شده از مقصد آرشیو
۴ میلیارد یو.آل برداشت شده و ۱۰ میلیارد بعد از حذف تکرارها	بروندادهای کمی	تعداد یو.آل‌ها (قبل و بعد از حذف تکرارها)
۲ میلیون منبع با خوش موفقیت‌آمیز (کد ۲۰۰)	تعداد منابع بر حسب کد	توزیع یو.آل‌ها بر مبنای کدهای وضعیت
۳ میلیون نام دامنه	بروندادهای کمی	تعداد میزبان‌ها یا دامنه‌ها
۲۰۰ ترابایت غیر فشرده قبل از حذف تکرارها و ۱۶۰ ترابایت متراکم بعد از حذف تکرارها	بروندادهای کمی	اندازه در مقیاس بایت (غیر فشرده و فشرده‌سازی شده، قبل و بعد از حذف تکرارها)
۱۸۰۰۰ فایل دبلیو.ای.آر.سی	بروندادهای کیفی	تعداد دبلیو.ای.آر.سی یا هر فایل محفوظه‌بندی شده دیگر

۳-۴ مشخصه‌یابی مجموعه

۱-۳-۴ کلیات

آمارهای پیشنهادی در این بخش ویژگی‌های آرشیووب را توصیف می‌کند و به محدوده آرشیووب و همچنین به اخذ تصمیمات آگاهانه کتابداران^۱ کمک می‌کند. در حالی که بعضی آمارها مختص جمع‌آوری انتخابی یا جمع‌آوری انبوه است، سایر آمارها عام بوده و در آرشیوهای وبی کاربرد دارند که براساس هر دو راهبرد می‌باشند.

مقیاس آرشیو وب عموماً مانع از شمارش دستی می‌شود. برخی آمارها، مخصوصاً آمارهایی که مختص برداشت انبوه هستند را می‌توان فقط از طریق نمونه‌برداری جمع‌آوری شوند. جمع‌آوری دستی آمارها ممکن است برای آرشیوهای انتخابی امکان‌پذیر باشد، ولی بهتر است فقط بدون استفاده از منابع غیر تضمین‌شده، انجام گیرد.

۲-۳-۴ آمارهای معمول

۱-۲-۳-۴ توزیع براساس دامنه‌های سطح بالا و سطح دو^۱

۱-۱-۲-۳-۴ هدف

دامنه‌های سطح بالا (تی.ال.دی ها)، توزیع جغرافیایی منابع در آرشیو وب را نشان می‌دهند. و این آمار مورد توجه ویژه کتابخانه‌های ملی و آرشیوهایی است که باید تمامی بروندادهای معنوی کشور را حفظ کنند. دامنه‌های سطح دو، که خاص رده‌های به‌خصوصی از سازمان‌ها و حوزه‌های مورد نظر هستند، هم مفیدند، چرا که ماهیت گسترده منابع در آرشیو را نشان می‌دهند. به عنوان مثال، سازمان دولتی UK، منابع تحت نام دامنه gov.uk را انتشار می‌دهد.

۲-۱-۲-۳-۴ روش

توزیع تی.ال.دی ها و دامنه‌های سطح دو را می‌توان با استفاده از گزارش‌های تولید شده توسط خزشگرهای آرشیوکننده یا سایر روش‌های خودکار تحلیل دامنه‌ها، به طور خودکار محاسبه کرد و می‌توان آن‌ها را بر حسب اعداد مطلق یا بر حسب درصد، اندازه‌گیری کرد. فهرست کردن ۵ یا ۱۰ از تی.ال.دی هایی که در آرشیو وب جزو پر تکرارترین تی.ال.دی ها هستند، نیز مفید است.

- تعداد یا درصد تی.ال.دی ملی جمع‌آوری شده: ۷۰٪ یو.آر.ال.هایی که در آخرین خزش دامنه‌ای کتابخانه ملی فرانسه قرار دارند، بر روی نام دامنه fr. و سه درصد آن‌ها بر روی نام دامنه de، میزبانی شده‌اند.

- تعداد یا درصد دامنه‌های سطح دو جمع‌آوری شده: ۱,۵٪ یو.آر.ال.هایی که در آخرین خزش دامنه‌ای کتابخانه ملی فرانسه قرار دارند، بر روی نام دامنه gov.fr. میزبانی می‌شوند که سازمان‌های دولتی فرانسه این منابع را منتشر کرده‌اند.

آرشیو وبی که مقدار تی.ال.دی های ملی آن بیشتر از سایر دامنه‌ها است را می‌توان دارای محدوده ملی در نظر گرفت.

۳-۱-۲-۳-۴ محدودیت‌ها

برخی مؤسسات، منابعی که میزبان آن‌ها در خارج از تی.ال.دی های ملی شان قرار دارد را هم داخل حوزه درنظرمی‌گیرند. مثلاً www.lego.com (یک شرکت دانمارکی)، وب‌گاه دانمارکی درنظر گرفته می‌شود گرچه نام دامنه ای غیر از dk را به کار می‌برد.

1- Top and second level domains

۴-۳-۲-۲ توزیع بر اساس حجم منابع در هر دامنه (و/یا میزبان)

۴-۳-۲-۲-۱ هدف

تجزیه و تحلیل و گزارش اندازه منابعی که تحت هر دامنه و/یا میزبان، میزبانی می‌شوند و چگونگی توزیع آن درآرшиو وب نه تنها بینشی را از توصیف مجموعه گردآوری شده فراهم می‌آورد، بلکه به مدیریت فرآیند خوش نیز کمک می‌کند.

توزیع اندازه منابع موجود در آرشیو براساس دامنه‌ها و/یا میزبان‌ها، می‌تواند ویژگی‌های منابع میزبان تحت انواع خاصی از دامنه‌ها را آشکار نماید. این موضوع همچنین حاکی از توانمندی منابع در حال جمع‌آوری با اندازه‌های متفاوت، خصوصاً وب‌گاه‌های بزرگی^۱ است که از نظر فنی نسبت به خوش چالش دارند.

گروه‌بندی منابعی که اندازه‌های مختلفی دارند بر اساس دامنه و/یا میزبان، به پیکربندی و سازمان‌دهی فرآیند خوش نیز کمک می‌کند. معمول است که دامنه‌های با اندازه برابر در یک گروه قرار گرفته و تحت یک فرایند یا کار^۲ مجزا خوش شوند، زیرا که خوش چنین دامنه‌هایی، نیازمند تنظیمات مشابهی است و زمان

یکسانی برای کامل شدن آن‌ها مورد نیاز است. این امر منجر به استفاده بهینه از منابع ماشینی شده است و وظایف^۳ پایشگری و مدیریت را ساده می‌سازد.

۴-۳-۲-۲-۲ روش

حجم منابع در هر دامنه و/یا میزبان می‌تواند بر حسب مگابایت، گیگابایت یا ترابایت قابل اندازه‌گیری باشد:

< ۱۰ MB -

< ۱۰۰ MB -

۱۰۱ تا ۹۹۹ MB -

۱ GB -

> ۱ GB -

به روش جایگزین، تعداد یو.آر.ال‌ها در هر دامنه را نیز می‌توان شمرد:

< ۱۰ ۰۰۰ URL -

۱۰۰ ۰۰۰ تا ۵۰ ۰۰۰ URL -

> ۱۰۰ ۰۰۰ URL -

1- Large

2- Job

3- Task

تنها راه معین کردن حجم منابع در هر دامنه در وب زنده، استفاده از ارقام^۱ ارائه شده توسط موتورهای جستجو است. برای آرشیو وب، آمارهای بالا می‌تواند به‌طور خودکار از گزارش‌های خرز یا از طریق سایر روش‌های خودکار تحلیل فایل‌های ذخیره شده، محاسبه شود.

۴-۳-۲-۲-۳ محدودیت‌ها

همانند سایر آمارهای پیشنهادی در قسمت قبل، چنین آمارهایی تا حدودی تقریبی هستند و این آمارها وقتی که در طول زمان اندازه گیری شده باشند به‌طوری که مقایسه بتواند صورت پذیرد، مفیدتر خواهند بود.

۴-۳-۲-۳-۳ توزیع براساس نوع قالب^۲

۴-۳-۲-۳-۱ هدف

تحلیل و گزارش توزیع قالب‌های فایل‌ها در آرشیو وب، اقدامی کلیدی برای محافظت و نگهداری رقمی و همچنین عوامل مشخصه‌یابی آرشیو هست.

برای پایش و مدیریت خطرات محافظت و نگهداری مربوط به قالب‌های منسوخ شده، کسب دانش مربوط به انواع فایل‌های موجود در آرشیو وب، ضروری است.

اطلاعات قالب‌های فایل‌ها، با طبقه‌بندی سطح بالا و سنتی کتابخانه‌ها در رده‌بندی انتشارات، همچون تصویر، فیلم و صدای ضبط شده، قابل مقایسه است. جمع‌آوری این اطلاعات در طول زمان می‌تواند روندهای فناوری را نیز نشان دهد و به درک تکامل وب کمک می‌کند.

۴-۳-۲-۳-۴ روش

آمارهای قالب‌های فایل‌ها را می‌توان به طور خودکار از گزارش‌های خرز یا از طریق سایر روش‌های خودکار تحلیل فایل‌های ذخیره شده، محاسبه کرد.

توزیع قالب فایل‌ها را می‌توان به روش‌های مختلفی ساماندهی و محاسبه نمود، به عنوان مثال:

- نوع منابع: ۷۰٪ فایل‌ها، متنی (مثالاً اچ.تی.ام.ال) هستند، ۱۵٪ آن‌ها تصویری (از جمله جی.پی.ای.جی یا جی.آی.اف) و سه درصد آن‌ها ویدئویی (به عنوان مثال ام.پ.گ) هستند؛

- رایج‌ترین قالب‌های فایل (۵۰ مورد اول یا ۱۰۰ مورد اول): به عنوان مثال اچ.تی.ام.ال رایج‌ترین قالب فایل‌ها است؛

دنبال کردن قالب‌های خاص، شامل اولین زمان مشاهده یک قالب خاص در آرشیو و افزایش و کاهش تعداد وقوع آن در آرشیو در طول زمان؛

1- Figures

2- Format type

- کمترین نوع قالب مورد استفاده: برای مثال اگر قالب ویدئویی در آرشیووب کم اهمیت‌تر از وب زنده است، می‌تواند حاکی از ارائه کمتر این قالب فایل در آرشیو باشد.

۴-۳-۲-۳ محدودیت‌ها

بررسی پروفایل قالب‌های موجود در آرشیووب، معمولاً فهرست طولانی از قالب‌های فایل‌ها پدید می‌آورد که هزاران قلم را در برمی‌گیرد. پیشنهاد می‌شود که بر ۵۰ تا ۱۰۰ قالب متداول‌تر تمکز شود مگر اینکه تمایلی به یک قالب غیرمعمول خاص، وجود داشته باشد.

قالب‌های منابع خزش شده (نوع ام.آی.ام.ای) که از طریق کارساز وب گزارش می‌شوند یا در فایل‌های ثبت وقایع خزش ثبت می‌شوند، همیشه قابل اطمینان نیستند. کارسازهای وب ممکن است انواع ام.آی.ام.ای خط‌دار^۱ را برگردانند. بعضی مؤسسات از ابزارهای شناسایی قالب^۲ اضافه استفاده می‌کنند تا اطلاعات صحیح‌تری را به دست آورند.

۴-۲-۳-۴ مشخصه‌یابی زبانی

۱-۴-۲-۳-۴ هدف

تحلیل زبان‌های مورد استفاده در منابع آرشیو شده، بینشی از الگوهای زبانی آرشیووب را فراهم می‌آورد که می‌تواند برای درک ویژگی‌های این مجموعه ملی، از قبیل تنوع یا نزدیکی فرهنگی به سایر کشورها، مفید باشد. توصیف زبانی، بیش از همه به مللی که تنها به یک زبان گفتگو می‌کنند، مربوط می‌شود زیرا می‌تواند به تشخیص منابع مربوط به آن کشورها در سطح وب، کمک کند.

منابع موجود در دامنه ملی ممکن است شامل زبان‌های خارجی هم بشود، مثلًاً باسک در دامنه‌های فرانسه و عربی در دامنه‌های دانمارک. توصیف زبانی می‌تواند به تحلیل مسائل متعدد اجتماعی و فرهنگی و بازتاب آن‌ها در وب، کمک کند.

۲-۴-۲-۳-۴ روش

روش استاندارد یا فناوری خاصی برای تشخیص خودکار زبان در منابع آرشیو وب، وجود ندارد. ساختارهای خاص هر زبان طبیعی را می‌توان در تحلیل و تشخیص منابع همان زبان به کاربرد. چنانچه ممکن باشد، می‌توان از عناصر مرتبط با زبان در سرآیندهای اچ.تی.ام.ال و اچ.تی.پی نیز استفاده کرد. ابزارهای تشخیص و پردازش زبان طبیعی را نیز می‌توان برای این منظور استفاده کرد.

1- Faulty

2-Format identification tool

توزیع زبان‌های طبیعی در آرشیو و ب می‌تواند به صورت اعداد مطلق (تعداد کل صفحاتی که از یک زبان خاص استفاده می‌کنند) یا به صورت درصد (درصد صفحاتی که در آرشیو و ب از یک زبان خاص استفاده می‌کنند) بازنمایی شود. فهرست کردن پنج یا ده زبان اول که بیشتر بازنمود می‌یابد، نیز می‌تواند سودمند باشد.

۴-۳-۴ محدودیت‌ها

خیلی از منابع فاقد فراداده‌های زبانی هستند که باعث دشواری تشخیص خودکار زبان مورد استفاده جهت ایجاد این دسته از منابع می‌شود. همچنین برنامه‌های مترجم خودکار که امکان مشاهده صفحات وب را به چند زبان برای کاربران فراهم می‌سازند، می‌توانند این آمار را مخدوش کنند.

۴-۳-۵ پوشش زمانی

۴-۳-۵-۱ هدف

آرشیوهای وب را می‌توان از طریق پوشش زمانی آن‌ها توصیف کرد که به مدت زمان جمع‌آوری منابع برمی‌گردد. فرض عمومی این است که هرچه دوره زبانی طولانی‌تر باشد، احتمال فقدان منابع اصلی در وب رونوشت آن‌ها در آرشیو و ب وجود دارد، از وب زنده بیشتر است. این امر بر ارزش آرشیو می‌افزاید چرا که می‌تواند سوابق تاریخی منحصر به فردی را از دوره‌های زمانی یا رویدادهای تاریخی معین حفظ کند. پوشش زمانی آرشیو نیز اطلاعات اصلی است که برای طرح‌ریزی محافظت ضروری است. آرشیوهایی که دوره زمانی طولانی‌تر را پوشش می‌دهند، بیشتر در خطراز بین رفتن هستند.

۴-۳-۵-۲ روش

پوشش زمانی آرشیو و ب را می‌توان از طریق شمارش سال‌های بعد از اولین اخذ منبع، به سادگی محاسبه کرد. این آمار می‌تواند در ترکیب با سایر آمارها مثل اندازه آرشیو یا توزیع قالب فایل‌ها برای نشان دادن روند معمول یا بسط آرشیو در طول زمان، به کار رود.

۴-۳-۶ آمارهای آرشیوهای انتخابی

آمارهای زیر تنها در خصوص آرشیوهایی که به وسیله برداشت انتخابی جمع‌آوری شده‌اند، مرتبط است و کاربرد دارد.

۴-۳-۳-۱ مجوزها

۴-۳-۳-۱ هدف

در صورت نیاز به دریافت مجوز از ناشرین یا صاحبان حقوق قبل از گردآوری وب‌گاه‌های آن‌ها، آمارهای مربوط به مجوزها، نشانگر مهمی برای اثربخشی روند کار است. در قیاس با هزینه‌های نگهداری سیستم مدیریت مجوزها، به عنوان مثال، آمارهای مجوز می‌تواند بیانگر هزینه هر واحدی باشد که با موفقیت مجوز را کسب کرده است.

تعداد مجوزهای اخذ شده، علاقه ناشران و آگاهی آن‌ها از آرشیوکردن وب را نمایان می‌کند.

۴-۳-۳-۱ روش

شمارش تعداد مجوزهای اخذ شده و همچنین تعداد درخواست‌های ارسالی به صاحبان حقوق مفید است و می‌تواند به صورت دستی یا کارکردهای خودکار موجود در سیستم مدیریت مجوز یا ابزار متصدی انجام شود.

۴-۳-۳-۲ کاندیدها

۴-۳-۳-۱ هدف

کاندیدها منابع پیشنهادی برای شمول در آرشیووب هستند. کاندیدها می‌توانند از طرف گستره عظیمی از افراد همچون علاقمندان، کاربران حامی آرشیو وب و مردم عادی پیشنهاد شوند. بعضی آرشیوهای وب، با استفاده از هیئت داوری، شبکه‌های اجتماعی یا صفحات برخط معرفی کاندید، کاندیدها را به صورت فعالانه انتخاب می‌کنند. فرآیند انتخاب مشخص می‌کند که، بر اساس سیاست‌های توسعه مجموعه، آیا کاندید در حوزه آرشیووب قرار دارد یا خیر. انتخاب معمولاً توسط کتابداران مجموعه، آرشیوکنندگان یا متصدیان انجام می‌شود.

شمارش تعداد کاندیدها، تأثیر و آگاهی از فعالیت‌های آرشیو وب در بین ذی‌نفعان را نشان می‌دهد و نیز می‌تواند نحوه دخیل‌شدن فعالیت‌های این ذی‌نفعان با مجموعه آرشیوکننده را هدایت کند. این معیار همچنین می‌تواند بهره‌وری و اثربخشی مرتبط با فرآیند انتخاب را اندازه‌گیری کند و نیز می‌تواند در ارزیابی میزان تلاش هریک از متصدیان رقمی‌سازی (مثلاً تعداد کاندیدهای هر متصدی) مفید باشد. چنانچه موسسه‌ای راهبرد ترکیبی از هر دو نوع برداشت انتخابی وابوه را به کارگیرد، از این معیار می‌توان برای مقایسه آن‌ها استفاده کرد و هزینه‌های انتخاب دستی را در برابر رویکرد خودکار برداشت انبوه، برجسته نمود. به منظور بررسی چگونگی استفاده آرشیو وب، می‌توان آمارهای بیشتری تولید کرد که بررسی شود چگونه از آرشیو وب استفاده می‌شود. آیا محتوای برداشت مبتنی بر انتخاب متصدیان، بیشتر مورد استفاده قرار می‌گیرد و در نتیجه نسبت به محتوای برداشت شده به صورت خودکار از اهمیت بیشتری برخوردار است یا خیر.

۴-۳-۲-۲ روش

شمارش تعداد کاندیدها می‌تواند به صورت دستی جمع‌آوری یا با استفاده از کارکردهای خودکار ابزار متصدی انجام شود.

۴-۳-۲-۳ محدودیت‌ها

همه کاندیدها به مقصد اصلی یا محتوای مدنظر نمی‌رسند و عوامل بسیاری بر نتیجه تأثیرگذار است. در اخذ مجوز وب‌گاه ممکن است موقیت حاصل نشود، بنا به دلیل فنی ممکن است خزشگر قادر نباشد محتوای را به دست آورد یا موسسه‌ای ممکن است به دلیل محدودیت مالی موفق به برداشت منابع انتخابی نشود. حتی در صورت برداشت موفق منبع، هنوز مسائل دسترسی می‌تواند بازگرداندن آن از آرشیو وب را غیر ممکن سازد.

آمارهای ذکر شده در بالا، بیشتر بر تلاش در انتخاب تاکید دارد تا نتیجه.

۴-۳-۳-۳ موضوعات

۴-۳-۳-۱ هدف

پژوهش موضوعی آرشیو وب، محتوای آن را توصیف می‌کند و با کمک به رخ نمون و توازن مجموعه و مشخص کردن فقدان‌های احتمالی در آن، ارزش فوق العاده‌ای در توسعه مجموعه دارد. پژوهش موضوعی می‌تواند برای درک این‌که آیا آرشیو وب می‌تواند نیازهای محققین رشته‌های مختلف را برآورده سازد نیز به کار رود.

۴-۳-۳-۲ روش

استخراج اطلاعات موضوعی به روش‌های مختلفی امکان‌پذیراست. منابع برخی آرشیوهای وب‌صورت دستی به وسیله متصدیان در قالب واژگان موضوعی برچسب‌گذاری می‌شود که می‌توان در ادامه آن‌ها را گروه‌بندی، پالایش و تحلیل نمود. سامانه‌های طبقه‌بندی استاندارد، نظیر رده‌بندی دهدزی دیویی (دی.دی.سی) یا سرعونانه‌ای موضوعی کتابخانه کنگره (ال.سی.اس.اچ) عموماً برای توصیف منابع آرشیو شده، استفاده می‌شوند.

در اکثر آرشیوهای وب اصطلاحات موضوعی به صورت دستی اضافه نمی‌شوند. سپس اطلاعات موضوعی، اگر در دسترس باشند، از طریق استخراج داده‌های موضوعی که در فراداده‌های منابع تعبیه شده‌اند، نظیر مقدار فرانشانه برچسب "keyword"^۱ در سرآیند اچ.تی.تی.پی یا قسمت «موضوع» دوبلین‌کور، به دست می‌آید. خوشبندی یا طبقه‌بندی خودکار نیز می‌تواند برای تحلیل معنایی منابع در آرشیوسازی وب به کار رود و آن‌ها را به صورت خودکار براساس طرح‌های موضوعی، طبقه‌بندی کند. در حال حاضر فعالیت رایج و مورد توافقی در این خصوص در جامعه آرشیو وب، وجود ندارد.

^۱ نام یک فیلد در سرآیند http می‌باشد

در صد منابع تحت هر موضوع را، می‌توان برای کمک به درک چگونگی توزیع محتوای آرشیوب، در موضوعات مختلف فهرست کرد. تلاش‌های بعدی می‌تواند در جهت پرداختن به موضوعاتی باشد که کمتر ارائه شده‌اند. اطلاعات موضوعی می‌تواند در ترکیب با دیگر آمارها و اطلاعات نیز به کار رود به عنوان مثال، با فراداده‌های اصالت به منظور درک ارتباط بین حوزه‌های موضوعی و انواع ناشران.

۴-۳-۳-۳-۴ محدودیت‌ها

موضوع‌بندی اصطلاحات به صورت دستی، با دقیق‌بیشتری منابع آرشیوشده را توصیف و طبقه‌بندی می‌کند، اما هزینه اجرایی بالایی دارد. استخراج خودکار فراداده‌های تعبیه شده، متکی بر استفاده از فراداده‌های در دسترس است. خوش‌بندی یا دسته‌بندی خودکار هنوز روش توسعه یافته و مناسبی در حوزه آرشیوسازی وب محسوب نمی‌شود و از نظر فنی به ویژه برای آرشیوهای وب با مقیاس‌های وسیع، چالش‌برانگیز است.

۴-۳-۴ آمارهای اصلی مجموعه‌سازی

جدول ۳- آمارهای اصلی در توصیف مجموعه

مثال	هدف	آمار
ده میلیارد یو.آر.آل در تی.آل.دی مربوط به fr. در آرشیو	توزیع جغرافیایی	توزیع بر اساس دامنه‌های سطح ۱ و ۲
دو میلیون نام دامنه کمتر از ده یو.آر.آل دارند، در حالی که ۱۵۰۰۰ نام دامنه بیشتر از ۱۰۰۰۰ یو.آر.آل را میزبانی می‌کنند.	تحلیل دامنه	توزیع بر اساس حجم منابع در هر دامنه
پانصد میلیون یو.آر.آل از آخرین خوش‌انبوه در قالب اج.تی.ام.آل است.	مشخصه‌یابی قالب	توزیع بر اساس انواع قالب فایل‌ها
آرشیو شامل منابعی است که از سال ۱۹۹۶ تا حال حاضر جمع‌آوری شده است.	تحلیل زمانی	پوشش زمانی
مجوزهای اعطایی از ناشران نتیجه درخواست ۲۰۰۰۰ مجوز است.	بهره‌وری	تعداد مجوزهای اعطای شده
هر ساله ۱۰۰۰ کاندید جدید اضافه می‌شود.	بهره‌وری	تعداد کاندیدها

۴-۴ استفاده از مجموعه

۱-۴-۴ کلیات

همان‌طور که در زیربند ۲-۳ (روش‌های توصیف و دسترسی) بیان شد، شرایط توصیف و مشاوره منابع اینترنتی آرشیو شده بطبق قوانین ملی و خطمشی‌های موسسه‌ای متفاوت است و بنابراین برای انتخاب مناسب‌ترین شیوهٔ سنجش میزان استفاده از آرشیوهای وب، لازم است که مابین آرشیوهای سفید، خاکستری و تیره تمایز قائل شد.

در مورد آرشیوهای برخط، آمار استفاده از آرشیوب با همان روش‌ها و استاندارهایی که وب زنده به کارمی‌رود اندازه‌گیری می‌شود در اندازه‌گیری آمار استفاده از وب زنده به کار می‌رود. در مورد آرشیوهای خاکستری، آمار استفاده از آرشیوب از همان روش‌ها و استاندارهایی به دست می‌آید که در ارزیابی آمار استفاده از منابع الکترونیکی در کتابخانه‌ها به کار گرفته می‌شود. درنتیجه، این استاندار در اکثر موارد ارجاعاتی را به استاندارها و روش‌های موجود فراهم می‌کند، درحالی‌که تعاریف و اطلاعات فنی بیشتری را نیز برای آسان‌کردن درک بهتر از مفاهیم خاص مرتبط با دسترسی و کاربرد مجموعه‌های آرشیو وب، فراهم می‌کند.

۴-۴-۴ تعاریف و روش‌هایی برای اندازه‌گیری استفاده

الف- بازدیدکنندگان فیزیکی

قانون یا خطمشی موسسه‌ای ممکن است این امر را تحمیل کند که آرشیوهای وب به دسترسی‌های خاصی از مکان‌های جغرافیایی کنترل شده نظیر سالن‌های مطالعه در کتابخانه‌های ملی یا آرشیوها محدود شوند. در چنین مواردی، داده‌های اصلی دسترسی کاربران به سیستم می‌تواند از طریق داده‌های ثبت‌نام و سوابق ورود به سامانه رایانه‌ای جمع‌آوری شود. در این روش، اطلاعات استفاده از سیستم، با بالاترین کیفیت به دست می‌آید زیرا امکان پرسش درباره نیت و قصد مراجعه کنندگان، مستقیماً فراهم می‌شود.

ب- بازدیدکنندگان مجازی

آرشیوهای وبی که به عنوان وب‌گاه‌های در دسترس عموم منتشر شده‌اند، می‌توانند آمارهای استفاده خود را از طریق Web Analytics، که توسط انجمن تحلیل وب^۱ استاندارد سازی شده است، به دست آورند.

Web Analytics، حوزه‌ی رو به گسترش از تحلیل الگوهای استفاده از وب‌گاه‌های اینترنتی از طریق روش‌های زیر است.

1- Web Analytics association

بازدیدهای مجازی در استاندارد ISO 2789 این‌گونه تعریف شده است: چرخه‌ی مداوم فعالیت های کاربری در وب‌گاه کتابخانه توسط کاربرانی که نشانی آی‌پی آن‌ها خارج از فضای نشانی آی‌پی کتابخانه (معمولًا خارج از محل کتابخانه) است، بدون درنظر گرفتن تعداد صفحات یا چیزهایی که مشاهده شده است.

یادآوری ۱ - بازدیدکننده از وب‌گاه اینترنتی، یا یک برنامه مرورگر وب شناخته شده و منحصر به فرد است یا یک نشانی آی‌پی مشخص است که به صفحاتی از وب‌گاه کتابخانه، دسترسی یافته است.

یادآوری ۲ - برای آنکه مجموعه‌ای از درخواست‌ها، بازدید مجازی واحدی را نشان دهنده، فاصله بین هر دو درخواست متوالی معمولًا نباید بیش از یک وقفه باشد^۱ یا حداقل زمان مابین اولین و آخرین درخواست نیز نباید بیش از ۳۰ دقیقه باشد. زمان بیشتر از ۳۰ دقیقه، بازدید مجازی جدیدی را درخواست می‌کند.

یادآوری ۳ - کارسازهای وبی که خدماتی را ارائه می‌کنند که آمارشان در سایت دیگری گزارش می‌شود، باید از آمارهای سایت اینترنتی کتابخانه، مستثنی شوند.

پ - بازدیدکنندگان آدمواره

نوعی از بازدیدکنندگان مجازی منابع وب، عوامل غیرانسانی هستند که برای مقاصد خوش طراحی شده‌اند. این تعریف معمولًا به موتورهای جستجو اشاره دارد اما خزشگرهای آرشیو کننده نیز به طور مشخص در این دسته‌بندی جای می‌گیرند. به طور کلی بازدیدکنندگان آدمواره‌ای از طریق فیلترکردن نشانی آی‌پی، رشته مشخص‌کننده کاربر عامل، یا روش‌های پیچیده‌تری نظیر شناسایی الگوهای استفاده تکرارشونده (که در ادامه هر پیوند ذکر می‌شوند) که به طور عادی بازدیدکنندگان انسانی چنین الگوهایی را نشان نمی‌دهند، از گرددآوری آمارهای تحلیلی وب کنار گذاشته می‌شوند.

ممکن است تحلیل رفتار آدمواره مفید باشد، به‌ویژه اگر ابزار دسترسی استثنای مشخصی در ارتباط با آدمواره‌ها داشته باشد (روباتس.تی.ایکس.تی) که می‌تواند از فهرست شدن مطالب آرشیو شده توسط موتورهای جستجو جلوگیری نماید (و بنابراین به صورت بالقوه با وب زنده در موتورهای جستجو رقابت می‌کند)

ت - بازدیدکنندگان غیرعمدی^۲

ممکن است آمارهاری در مورد بازدیدکنندگان غیرعمدی جمع شود. این بازدیدکنندگان واقعاً قصد دیدن سایت را نداشته‌اند و بنابراین آمار آن‌ها جزء آمارهای مثبت کاذب^۳ است. نمونه‌هایی از این بازدیدکنندگان عبارتند از: درخواست‌های پویش رو به جلوی^۴ موتورهای جستجو یا برنامه‌های

1-Time-out period

2- Involuntary visitors

3- False positive

4- Forward page scanning

جانبی بررسی کننده ویروس برای مرورگرهایی که به طور خودکار پیوندها را قبل از اینکه کاربر آنها را درخواست دهد، پویش و درخواست می‌کنند. مرورگرهای مشخصی نیز هستند که از طریق فرستادن درخواست برای برخی از پیوندها نهان‌سازی^۱ نتایج دریافت شده و پیدیدآوردن تصاویری کوچک^۲ از آنها، از پیش انتخاب‌هایی را برای مرور کاربر تعیین می‌کنند. ارجاع خودکار به آرشیوها از طرف وب‌گاه‌های زنده نیز ممکن است با توجه به فعالیت‌های بعدی کاربر، در همین رده قرار گیرد.

ث- تحلیل فایل‌های ثبت وقایع

کارساز وب می‌تواند اطلاعات به دست آمده از سرآیندهای درخواست و پاسخ اج.تی.پی را ثبت کند. اطلاعاتی نظیر تاریخ زمان درخواست، نشانی آی.پی درخواست‌دهنده (حتی نام دامنه مرتبط با آی.پی)، منبع مورد تقاضا (یو.آر.آل)، منبع ارجاع دهنده و کد پاسخ را می‌توان برای هر تبادل درخواست/پاسخ ثبت کرد.

در تحلیل‌هایی که در اوایل دهه ۱۹۹۰ میلادی صورت می‌پذیرفت، شمارش هر مدخل ثبت شده، تعداد تک وردهای دریافتی واقعه را تعیین می‌کرد. با پیچیده‌تر شدن اج.تی.ام.آل و اضافه شدن عناصر تعبیه شده و ورود منابعی نظیر شیوه‌نامه‌ها و جاوا‌اسکریپت‌ها، از ارزش تعداد تک ورود به عنوان مقیاس سنجش کاسته شده و امروزه این معیار صرفاً برای سنجش میزان بارکارساز وب مفید است.

اما ابزارهای تحلیل‌گر سابقه، می‌توانند درک داده‌هایی را فراهم کنند که در فایل‌های ثبت وقایع موجودند و به صورت آمارهای تجمیعی قابل بیان هستند.

کاربرد گسترده پروکسی‌های وب و حافظه‌های نهان، درستی آمارهای به دست آمده از فایل‌های ثبت وقایع را مختل می‌کند. زیرا از یک سو، بسیاری از کاربران می‌توانند پشت نشانی آی.پی یکسان پنهان شوند و از سوی دیگر، درخواست‌های شبیه به هم، می‌توانند به جای آن که توسط کارساز وبی که ابزار دسترسی آرشیو وب را در اختیار قرار می‌دهد، پاسخ داده شود، توسط حافظه نهان وب پاسخگویی شود. حافظه‌های نهان وب همچنین می‌توانند برای افزایش سرعت کارایی ابزار دسترسی، به قیمت کاهش درستی آمار استفاده از آرشیو وب، به کار روند، زیرا این حافظه‌ها تلاش می‌کنند تا پیش از خود ابزار دسترسی حقیقی، به تقاضاها پاسخ دهند.

1- Cache

2- Thumbnail

ج- برچسب‌گذاری صفحه

برچسب‌گذاری صفحه، فرایندی است که از طریق آن اشیائی که در صفحات اچ.تی.ام.ال تعبیه شده‌اند در هنگام درخواست به وسیله جستجوگری که مدرک اچ.تی.ام.ال را بارگذاری می‌کند، برگردان‌های تماس^۱ پویایی با کارساز وب برقرار می‌کنند. نخستین کاربرد این روش، استفاده از شمارشگر صفحه تصویری در اوایل دهه ۱۹۹۰ بود که در آن، یک تصویر از سوی متن سی‌جی‌آی، مورد درخواست قرار می‌گرفت. این تصویر معمولاً حاوی مقدار عددی تعداد دفعاتی بود که متن اسکریپت مزبور اجرا شده بود و بنابراین تعداد بازدیدکنندگان متن را نیز نشان می‌داد. این روش امروزه با به کارگیری جاوا اسکریپت و کوکی‌هایی که اطلاعاتی را درباره کاربران و صفحه فراهم می‌کنند، پیشرفت کرده است. از آنجا که برگردان تماس از برچسب صفحه می‌تواند در هر مکانی (نه فقط کارساز وب میزبان) صورت پذیرد، این روش، رشد صنعت تحلیل وب برون سپاری شده را ممکن ساخته است.

ج- ثبت وقایع در سطح برنامه‌های کاربردی

از آنجایی که ابزار دسترسی، خود یک برنامه کاربردی تحت وب است امکان پذیر خواهد بود که معیارهای مرتبط با وقایع ثبت شده در خصوص استفاده از سامانه را مستقیماً از داخل خود این ابزار به دست آوریم.

ح- حریم خصوصی

جمع‌آوری اطلاعات با توجه به فعالیت‌های کاربر از طریق هر کدام از روش‌های بالا به رعایت خطمنشی حریم خصوصی مناسب نیازمند است که جاری و در دسترس کاربران سامانه باشد.

استفاده گسترده از برچسب‌گذاری صفحه (به بند ج، در بالا مراجعه شود)، می‌تواند چالش‌هایی برای آرشیوهای وبی که به صورت عمومی در دسترس هستند را ایجاد کند، زیرا ممکن است وب‌گاه‌های آرشیو شده، خود شامل برچسب‌های تعبیه شده باشند. بنابراین مشاهده صفحات آرشیو شده می‌تواند منجر به کوکی‌هایی شود که روی مرورگر کارفرما تنظیم می‌شوند و برگردان‌های تماسی با جمع‌آوری کننده‌های آمارهای تحلیلی صورت می‌دهند (نشت و ب زنده).

۴-۴-۳ آمارهای پایه برای سنجش میزان استفاده از آرشیو

جدول چهار، شامل آمارهایی است که به طور عمده توسط انجمن تحلیل وب معرفی شده‌اند. این آمارها به دو دسته آمارهای شمارشی ساده و آمارهای تجمیع پیچیده‌تر یا آمارهای ابعادی، تقسیم می‌شود. برخی معیارهای ویژه تبلیغات در جدول چهار حذف شده‌اند، زیرا به دامنه کاربرد این استاندارد ربطی ندارند و تعدادی از معیارهای مفید دیگر در انتهای جدول مزبور قرار گرفته‌اند.

1- Call backs

ستون چپ نشانگر اهمیت آثارهای ذکر شده است: زیاد برای اهمیت بالا، متوسط برای اهمیت متوسط، کم برای اهمیت پایین. در اکثر موارد به کار بردن آثارهای با اهمیت بالا برای آرشیو بوصیه می‌کنیم که این آثارها توسط اکثر برنامه‌های تحلیل‌گر، قابل ارائه هستند.

جدول ۴ - آثارهای پایه برای اندازه‌گیری میزان استفاده از آرشیو

نام	نوع	محاسبه	اهمیت
مشاهده صفحه	شمارشی	تعداد دفعاتی که صفحه دیده شده است.	زیاد- نمایش میزان استفاده از آرشیو به صورت پردازش نشده
بازدید (جلسات)	شمارشی	یک بازدید، تعاملی توسط فرد با وب‌گاه میباشد که شامل یک یا چند درخواست برای صفحه می‌باشد. اگر فردی در دوره‌ی زمانی مشخص، عمل دیگری انجام نداده باشد (معمولًاً دیدن صفحات دیگری از وبگاه)، با گذشت بازه زمانی وقفه، بازدید پایان خواهد یافت.	زیاد- تعداد پایه کاربران آرشیو
بازدیدکنندگان منحصر به فرد	شمارشی	تعداد استنباط شده از افراد مجازی (بدون درنظرگرفتن خوشگ ^۱ و آدمواره‌ها) که در یک بازه زمانی معین، فعالیتی شامل یک یا چند بازدید از سایت داشته‌اند. در بازه زمانی مزبور، هر فرد تنها یکبار در معیار منحصر به فرد بازدیدکنندگان، شمرده می‌شود.	متوسط
رویداد	اندازه و/یا شمارشی	هرگونه عمل ثبت شده که تاریخ و زمان مشخصی دارد که به وسیله مورگر یا کارساز به آن اختصاص یافته است.	کم
1-Crawler 2-Robots			

۴-۴-۴ آمارهای تجمیعی برای توصیف پیشرفتی میزان استفاده از آرشیو

جدول ۵- آمارهای انبوشه‌ی برای مشخصه‌یابی میزان استفاده از آرشیو

نام	نوع	محاسبه	اهمیت
صفحه ورود	بعد	اولین صفحه یک بازدید	متوسط
صفحه فرود	بعد	بازدید از صفحه به منظور شناسایی شروع تجربه کاربر حاصل از یک تلاش بازاریابی تعریف شده	کم
صفحه خروج	بعد	آخرین صفحه‌ای در سایت که در طول یک بازدید به آن دسترسی صورت پذیرفته است و پایان یک بازدید/نشست را نشان می‌دهد.	کم
مدت بازدید	شمارشی	طول زمانی یک نشست. محاسبه معمولاً بر اساس مهر زمانی آخرین فعالیت در نشست منهای مهر زمانی اولین فعالیت نشست، صورت می‌پذیرد.	زیاد
ارجاع دهنده	بعد	ارجاع دهنده یک اصطلاح عام است که منبع ترافیک ورودی به یک صفحه یا بازدید را توصیف می‌کند	متوسط
نام	نوع	محاسبه	اهمیت
ارجاع دهنده صفحه	اندازه	ارجاع دهنده صفحه منبع ترافیک ورودی به یک صفحه را توصیف می‌کند	متوسط
بازدیدکننده جدید	شمارشی	تعداد بازدیدکنندگان منحصر به فردی که فعالیت آنها شامل "اولین بار" بازدید از سایت در طول دوره گزارش دهی است. توجه داشته باشید که "اولین بار"، نسبت به زمانی است که جمع‌آوری داده‌ها با استفاده از ابزار کنونی، به درستی شروع شده است.	متوسط
بازدیدکننده برگشتی	شمارشی	تعداد بازدیدکنندگان منحصر به فردی که فعالیت آنها شامل بازدیدکننده بازگشتی از سایت در طول دوره گزارش دهی است و در آن یک بازدید کننده منحصر به فرد، پیش از دوره گزارش دهی نیز سایت را بازدید کرده است	متوسط
بازدیدکننده تکراری	شمارشی	تعداد بازدیدکنندگان منحصر به فردی که فعالیت آنها شامل دو یا بیشتر بازدید از سایت در طول یک دوره گزارش دهی است.	متوسط
بازدید از یک صفحه	بعد یا شمارشی	بازدیدی که شامل دیدن یک صفحه می‌باشد	متوسط
بازدیدکنندگان براساس مکان جغرافیایی	شمارشی	گزارش دهی مکان جغرافیایی بازدیدکنندگان براساس نشانی آی‌پی آنها	کم
عبارات جستجو جهت یافتن آرشیو	شمارشی	عبارات مورد استفاده در موتورهای جستجو به منظور یافتن وب‌گاه ابزار دسترسی	متوسط
عبارات جستجو مورد استفاده در آرشیو	شمارشی	عبارات جستجوی مورد استفاده در ابزار دسترسی به منظور یافتن نسخه‌های برگرفته شده از آرشیو شده	زیاد

جدول ۶- آمارهای اصلی برای استفاده از مجموعه

مثال	هدف	آمار
تعداد صفحات ۴۸۳۱۸ بین اول تا سیام ژوئن سال ۲۰۱۲ بازدید از آرشیو وب انجام شده است	گستره استفاده	تعداد صفحات دیده شده
تعداد صفحات ۱۱۴۱۵ بین اول تا سیام ژوئن سال ۲۰۱۲ بازدید از آرشیو وب انجام شده است	اندازه استفاده	تعداد بازدیدها
۹ ۴۳۴ بازدیدکننده منحصر به فرد از آرشیو وب انگلیس بین اول تا سیام ژوئن ۲۰۱۲ انجام شده است	گستره استفاده	تعداد بازدیدکنندگان منحصر به فرد
به طور متوسط، هر بازدید از آرشیو وب انگلیس بین اول تا سیام ژوئن ۲۰۱۲ به مدت (۲۵" و ۳") طول کشید	علاقة کاربران به آرشیو	مدت بازدید
بیشترین کلمه کلیدی مورد استفاده برای جستجوی آرشیو وب انگلیس بین اول تا سیام ژوئن ۲۰۱۲، "گوجی بری" ^۱ بود	رفتار کاربران	عبارات جستجوی مورد استفاده در آرشیو
۱- این میوه شبیه به ذغال اخته است..		

۴-۴ محافظت از آرشیو وب

۱-۵-۴ کلیات

بهتر است محافظت بلندمدت آرشیوهای وب، از چارچوبهای وسیع‌تر و قابل کاربردی که برای برای محافظت از تمامی منابع رقمی، جدا نشود. توصیه می‌شود موسسات جمع‌آوری ایده‌آل یک سامانه محافظت رقمی اختصاصی برای محافظت‌های رقمی خود، مطابق با استانداردهای همچون ISO14721 (سامانه‌های انتقال اطلاعات و داده‌های فضایی - سامانه اطلاعاتی آرشیو باز - مدل مرجع)، داشته باشند. مدل مرجع (او.ای.آی.اس) یک سامانه آرشیو ویژه محافظت و حفظ دسترسی به اطلاعات رقمی در طول زمان را مشخص کرده است.

این استاندارد به جزئیات مدل او.ای.آی.اس نمی‌پردازد ولی از مفاهیم و تعاریف پایه آن برای توصیف موضوعات محافظت مربوط به آرشیوهای وب، استفاده می‌کند. برخی از آمارهای

پیشنهادی را می‌توان به انواع دیگری از منابع رقمی تعمیم داد، سایر آمارها ویژه آرشیوهای وب می‌باشد.

چنان‌که در زیربند ۳-۳ (روش‌های محافظت) ذکر شد، محافظت رقمی می‌تواند در دو سطح انجام شود، در

سطح پایه برای نگهداری این بیت‌ها، و در سطح پیچیده‌تر با به کارگیری راهکارهایی چون گذار و نمونه‌سازی، برای محافظت از ظاهر، عملکرد، رفتار و حتی تجربه کاربر در منابع رقمی. سطح اول را "جريان بیت یا محافظت فیزیکی" می‌شود و سطح دوم را "محافظت منطقی" می‌نامند. هدف آمارهای ذکر شده در زیربند ۴-۵-۲، اندازه‌گیری کارایی فعالیت محافظت جريان بیت است. الگوی پیشنهاد شده در زیربند ۴-۵-۳، برای کمک به موسسات برای گزارش فراداده‌هایی است که انتظار می‌رود در آرشیوب محافظت شوند. آمارهای ذکر شده در زیر بند ۴-۵-۴، آمارهای مربوط به محافظت منطقی هستند.

۴-۵-۲ آمارهای مربوط به محافظت جريان بیت

۴-۵-۱-۱ حجم منابع ازدست‌رفته یا مخدوش شده

۴-۵-۱-۲-۱ هدف

بسیاری از موسسات پژوهشی از دست‌رفتن داده‌ها به واسطه خرابی ابزارهای فیزیکی را تجربه کرده‌اند، مانند خراب شدن ناگهانی سخت‌افزارها یا از کارافتاده شدن رسانه‌ها. برخی از داده‌ها ممکن است به طور تصادفی حذف شوند. منابع مخدوش شده، منابعی هستند که به طور کامل از دست رفته‌اند ولی صدمه دیده‌اند و امکان دسترسی کامل به آن‌ها وجود ندارد. اطلاعات درباره از دست رفتن داده‌ها عموماً در دسترس نیست، ولی نظارت و بازبینی میزان از دست رفتن منابع یا تخریب منابع، حائز اهمیت است زیرا این مسئله شاخصی مهم برای یکپارچگی آرشیوهای وب به حساب می‌آید.

۴-۵-۱-۲-۲ روش

میزان از دست رفتن و تخریب داده‌ها را می‌توان بر حسب بایت یا بر حسب تعداد یو. آر. ال. ها (با مقایسه منظم جمع آزمایش) اندازه‌گیری کرد.

۴-۵-۲-۲ حجم منابع رونوشت و توزیع شده

منابع پشتیبانی نشده یا کپی نشده در خطر از دست رفتن دائمی هستند و نمی‌توان آن‌ها را بازیابی کرد. شیوه رایج کپی کردن منابع در چندین مکان برای جلوگیری از ایجاد نقطه خطای واحد است. بنابراین حجم منابع کپی شده، شاخصی برای ایمنی مجموعه است.

با این حال، رونوشت^۳ منابع به خودی خود برای تضمین ایمنی مجموعه، کافی نیست. لازم است پارامترهای مختلف دیگری به منظور اثربخش بودن کپی داده‌ها، وضع شود. در اینجا موضوع‌های اساسی تنوع و یکپارچگی هستند که موارد زیر را دربرمی‌گیرند:

- تعداد پیکربندی‌های نرم‌افزاری و سخت‌افزاری متنوع برای نمونه‌های مختلف؛
- فاصله فیزیکی بین نمونه‌ها؛
- مقدار و شیوه‌ی بررسی یکپارچگی داده‌ها بین نمونه‌های داده؛

منابعی که از یکپارچگی آن‌ها با چنین سنجش‌هایی اطمینان حاصل شده است، کپی شده و توزیع شده تلقی می‌شوند.

هزینه‌ها در کپی و توزیع داده‌ها دخیل می‌باشند. تصمیم‌گیری برای اندازه‌ای که باید آرشیو و بکپی شود، باید با متعادل‌سازی تأثیر خطر، هزینه‌ها و پیچیدگی مدیریت منابع کپی شده، اتخاذ شود.

۴-۵-۳ آمارهای مربوط به محافظت فراداده‌ها

اهمیتِ محافظت از منابع به همراه فراداده‌های مربوط به آن در زیربند ۴-۳-۳ توضیح داده شد. توصیه می‌شود که موسسات جمع‌آوری با استفاده از جدول ۷، به طور منظم درباره ماهیت و حجم فراداده‌ها در آرشیو و بکپی گزارش دهند.

1-Single point of failure

2-Replication

جدول ۷- آمارهای مربوط به محافظت فراداده

توضیحات	درصد منابع حاوی فراداده	استاندارد مورد استفاده (در صورت وجود)	توصیف	نوع فراداده
هرگونه توضیحات مفید یا مرتبط	درصد منابع حاوی فراداده		توصیف فراداده	یکی از انواع فراداده توضیح داده شده در ۴-۳-۳ زیربنده، برای مثال، توصیفی
مثالها				
عبارت موضوعی به طور دستی توسط متقدیان تخصیص یافت و در ابزار متقدی وب ذخیره شد	۳۰٪	قوانین و مقررات فراداده دوبلین کور (دی.سی.ام.آی)؛ (ال.سی.اس.اج)	مجموعه عناصر فراداده دی.سی.ام.آی نام عبارت: موضوع. عنوان منبع	توصیفی
از فایل‌های پیکربندی خرمشاه در سال ۲۰۰۴ صرفنظر شد	۹۰٪		فایل‌های های پیکربندی	اصالت
همه فایل‌های برداشت شده قالب اطلاعاتی (ام.آی.ام.ای) را دارند که این موضوع البته می‌تواند غیرقابل اطمینان باشد	۱۰۰٪	توسعه چندمنظوره نامه‌های اینترنتی (ام.آی.ام.ای) بخش ۲: انواع رسانه ها	قالب‌های فایل‌های (انواع ام.آی.ام.ای)	فني
فقط برای مقاصد با دسترسی آزاد، مورد نیاز است	۱۰۰٪		مجوز آرشیوکردن و فراهمن کردن دسترسی برخط	حقوق

۴-۵-۴ آمارهای محافظت منطقی

محافظت جریان بیت، بایت‌ها را در رسانه‌های فیزیکی ایمن نگاه می‌دارد و محافظت منطقی تضمین می‌کند که منابع با گذشت زمان قابل استفاده باقی می‌مانند. این استاندارد، سه شاخص اصلی را برای فعالیت‌های محافظت منطقی، پیشنهاد می‌دهد.

۱-۴-۵-۴ توزیع براساس قالب‌های مشخص شده پروژه

۱-۱-۴-۵-۴ هدف

توزیع منابع آرشیووب براساس قالب‌های فایل، آماری است که جزئیات آن در جدول زیربند ۳-۲-۳-۴ به عنوان عنصری از مشخصات آرشیو تشریح شده است. دانش قالب فایل نیز برای محافظت رقمی مهم و حیاتی است. تعریف یک راهبرد محافظت برای هر منبعی بدون اطلاعات قالب آن منبع، غیرممکن است. آرشیوهای وب عموماً حاوی منابعی در طیف وسیعی از قالب‌ها هستند. انواع ام.آی.ام.ای برگشت داده شده به وسیله کارساز وب، تنها اطلاعات ساده در دسترس، درباره قالب فایل هستند که البته چندان هم قابل اطمینان نیستند. به منظور به دست آوردن اطلاعات دقیق‌تر درباره قالب فایل موجود در آرشیو وب برای اهداف محافظتی، استفاده از ابزارهای شناسایی قالب مورد نیاز است.

علاوه بر برنامه‌ریزی یک راهبرد محافظت رقمی، از اطلاعات قالب نیز می‌توان برای تعیین خطرات محافظت و اولویت‌بندی اقدامات محافظتی، استفاده کرد.

۲-۱-۴-۵-۴ روش

به مطلب "روش" که در زیربند ۳-۲-۳-۴ توضیح داده شده، مراجعه شود. وقتی هدف، محافظت باشد اطلاعات دقیق‌تری درباره قالب مورد نیاز است و این مهم می‌تواند با به کارگیری ابزارهای تشخیص و شناسایی قالب، حاصل شود. سپس می‌توان بروندادها را محاسبه و همانند راههای پیشنهادی در زیربند ۳-۲-۳-۴، سازمان‌دهی کرد.

نمونه‌هایی از ابزارهای شناسایی قالب فایل‌ها شامل دروید^۱ و جی‌هاو^۲ می‌باشند.

۳-۱-۴-۵-۴ محدودیت‌ها

ابزارهای شناسایی قالب، همیشه نمی‌توانند تمامی قالب‌ها را تشخیص دهند. قالب‌های جدید نمونه قابل ذکری است که ابزارها هنوز تشخیص نمی‌دهند.

1- DROID

2- Jhove

۴-۵-۲-۴ تعداد قالب‌های فایل با راهبرد محافظت تعریف شده

۴-۵-۴-۱ هدف

شناسایی قالب فایل‌ها، یک نقطه شروع خوب است. جهت اطمینان از طول عمر، لازم است یک راهبرد محافظت رقمی برای قالب‌های اصلی مورد استفاده در آرشیو وب تعریف شود. موسسات ممکن است در موقع مورد نیاز از گذار و نمونه‌سازی برای قالب‌های مختلف بسته به نوع قالب، هدف کاربرد و منابع مورد نیاز برای پیاده‌سازی، استفاده نمایند. برای مثال، موسسه‌ای ممکن است تصمیم بگیرد که برای فایل‌هایی از قبیل شوک ویو فلش^۱ که نیاز به بازآفرینی دوباره تجربه تعاملی کاربر دارند، راهبرد نمونه‌سازی مناسب‌تر است در حالی که برای مدارک مایکروسافت آفیس، راهبرد گذار موثرتر است. توجه به این نکته مهم است که بر اساس ارزیابی ریسک، گاهی "هیچ کاری نکن" ممکن است راهبردی درست باشد.

راهبرد محافظت برای قالب‌های کلیدی در یک آرشیو وب، تعهدات موسسه را برای محافظت بلندمدت نشان می‌دهد.

۴-۵-۴-۲ روش

قالب فایل‌هایی را که برای آن‌ها راهبرد محافظتی تعریف شده است، فهرست نمایید. این فهرست می‌تواند با آمارهای توزیع منابع در هر قالب فایل که بر حسب بایت یا بر حسب تعداد یو.آل.های اندازه‌گیری شده است، ترکیب شود.

۴-۵-۴-۳ حجم منابع در هر قالب با راهبرد محافظت فعال‌سازی شده

۴-۵-۴-۱ هدف

راهبردهای محافظت ممکن است برای قالب‌هایی که مورد استفاده هستند و هنوز منسوخ نشده‌اند، تعریف شود که در این صورت موسسات ممکن است ترجیح دهنده راهبردها را روی نمونه‌ای از منابع، به جای فعال‌سازی آن‌ها برای کل مجموعه، آزمایش کنند. محدودیت‌های منابع ممکن است مانع به کارگیری راهبرد شود. محاسبه حجم منابع با راهبرد فعال‌سازی شده یا پیاده‌سازی شده، نشان می‌دهد که تا چه حد منابع در ریسک، از راهبرد محافظت رقمی بهره برده‌اند. این محاسبه همچنین می‌تواند سطحی از تعهد موسسه به محافظت رقمی را نشان دهد. در مورد "راهبرد هیچ کاری نکن"، هر برسی درباره راهبرد همراه با تصمیم به ادامه آن، می‌تواند به عنوان فعال‌سازی درنظر گرفته شود، که اطمینان حاصل می‌کند که تصمیمات اطلاع داده شده در حال پیاده‌سازی هستند.

هیچ راهبرد محافظت "معینی" وجود ندارد. با پیشرفت فناوری، منابع ممکن است دوباره منسوخ شوند و ممکن است لازم باشد، راهبردهای جدید برای همان قالب تعریف و چندین بار فعالسازی شود.

۴-۵-۴-۳-۲ روش

قالب‌های فایلی را که راهبردهای مشخص برای آن‌ها فعال شده است و می‌توان آن‌ها را از سامانه محافظت رقمی محاسبه کرد، فهرست نمایید. این آمارها بعداً با توزیع منابع در هر قالب فایل، که بر حسب بایت یا بر حسب تعداد یو.آر.آل‌ها اندازه‌گیری شده است، ترکیب می‌شود.

جدول ۸ - آمارهای اصلی برای محافظت مجموعه

مثال	هدف	آمار
۱۵۰ تراپایت از آرشیوب دارای نسخه رونوشت است	ایمنی و انعطاف‌پذیری	حجم منابع دارای نسخه کپی شده
٪ ۶۰ از آرشیو به صورت اج.قی.ام.آل است	توانمندی محافظت	توزیع بر اساس قالب (مشخص شده) فایل‌ها
۵ قالب، یک راهبرد محافظت مشخص دارند: اج.قی.ام.آل، جی.پی.ای.جی، جی.ای.اف، پی.ان.جی، پی.دی.اف	توانمندی محافظت و تعهد	تعداد قالب‌هایی که برای آن‌ها راهبرد محافظت تعریف شده است

۶-۴ محاسبه هزینه‌های آرشیوسازی وب

هزینه‌های آرشیوسازی وب را می‌توان به‌روشی مشابه هزینه‌های ایجاد و محافظت، از دیگر مجموعه‌های رقمی (برای مثال رقمی‌سازی)، تعیین نمود. تنها نکته‌ای که باید به آن توجه کرد، این است که فعالیت‌های آرشیوسازی وب هنوز جدید هستند و برخی از جنبه‌های اثربخشی و هزینه تنها با گذشت زمان می‌تواند اندازه‌گیری شود. این مسئله به ویژه برای هزینه‌های مربوط به محافظت از آرشیو وب، صدق می‌کند.

۱-۶-۴ برونو سپاری

یک موسسه جمع‌آوری کننده، ممکن است تمامی یا برخی از فعالیت‌های آرشیوسازی وب را به یک فروشنده یا شخص ثالث، برونو سپاری کند. چنین خدمتی می‌تواند شامل برداشت، نمایه‌گذاری، دسترسی و ذخیره‌سازی داده‌ها باشد و همچنین ممکن است جمع‌آوری مجموعه‌های گذشته‌نگر یا تاریخی و یا فعالیت‌های توسعه نرم‌افزاری خاص را پوشش دهد.

در مورد برونو سپاری، محاسبه هزینه‌های برونو سپاری آرشیوسازی وب آسان‌تر است، زیرا عموماً مقدار کل پولی را شامل می‌شود که ارائه‌دهنده خدمت از موسسه مطالبه می‌کند. ممکن است هزینه‌های اضافی مربوط به انتخاب محتوا، مدیریت قرارداد و غیره وجود داشته باشد که عموماً

توسط موسسه به طور داخلی تأمین می‌شود. این هزینه‌ها، برای محاسبه هزینه کل آرشیوسازی، باید به هزینه‌های بروندسپاری و بضافه شود.

۴-۶-۴ آرشیوسازی وب داخلی

محاسبه هزینه‌های فعالیت آرشیوسازی وب داخلی تقریباً شبیه ارزیابی فرایندهای آشناتر مثل فهرست کردن کتاب‌ها در کتابخانه‌ها است. چهار گروه اصلی هزینه‌ای باید در نظر گرفته شود: سخت‌افزار، محاسبه، نرم‌افزار و نیروی کار.

۴-۶-۱ هزینه‌های سخت‌افزاری

هزینه‌های سخت‌افزاری شامل خریداری و نگهداری زیرساخت‌های مورد نیاز برای برداشت، نمایه‌سازی، وارد کردن، ذخیره‌سازی و فراهم‌سازی دسترسی به داده‌ها و دیگر منابع رقمی است.

۴-۶-۲ هزینه‌های محاسباتی

هزینه‌های محاسباتی شامل هزینه‌های مربوط به برق و شبکه (پهنه‌ای باند) است.

۴-۶-۳ هزینه‌های نرم‌افزاری

بسته به انتخاب نرم‌افزار، هزینه‌های تهیه مجوز ممکن است برای هزینه‌های آرشیوسازی و ب، کاربرد داشته باشد. در حال حاضر بسیاری از موسسات جمع‌آوری کننده از نرم‌افزارهای رایگان و متن‌بازی استفاده می‌کنند که توسط سازمان‌های بین‌المللی یا بنیادهای غیرانتفاعی^۱ ارائه و محافظت می‌شوند، استفاده می‌کنند. بسیاری نیز از راه حل‌هایی که از طریق همکاری‌های بین‌المللی تدوین شده‌اند، مانند کنسرسیوم بین‌المللی محافظت اینترنتی^۲ (آی.آی.پی.سی)، استفاده می‌کنند. این کار هزینه‌ها را برای توسعه‌های عمده داخلی یا پرداخت مخارج مجوز به شرکت‌های تجاری، کاهش می‌دهد.

پیاده‌سازی نرم‌افزار متن‌باز و ادغام آن با سامانه‌های داخلی نیازمند تخصص فنی است. منابع توسعه‌دهنده برای روزآمدی پیاده‌سازی محلی برای هر نرم‌افزار جدید باید در دسترس باشد. به علاوه، هر موسسه الزامات ویژه‌ای دارد، بدین معنا که ایجاد و عرضه نرم‌افزارها براساس سفارش‌ها صورت می‌گیرد. منابع انسانی برای کارهای فنی و ایجاد نرم‌افزار باید به عنوان هزینه‌های نیروی کار، در نظر گرفته شود.

1-Not for profit foundations
2-Internet Preservation Consortium (IIPC)

۴-۲-۶-۴ هزینه‌های نیروی کار

هزینه نیروی کار را می‌توان طبق معمول بر معادل تمام وقت (حسب اف.تی.ای)^۱ یا روزمزد، محاسبه کرد. به طور کلی سه گروه بزرگ از کارمندان در آرشیوسازی و ب دخیل هستند: متصدی، فنی و مدیریتی. آن‌ها به طور مشترک، گستره وسیعی از فعالیت‌ها از جمله توسعه مجموعه، کارهای فنی، برنامه‌ریزی و مدیریت عملکرد را انجام می‌دهند. تقسیمات معمول نیروی کار براساس تخصص کارکنان صورت می‌گیرد: برای مثال متصدی، بیشتر یک مجموعه را بررسی و توصیف می‌نماید و مهندس، خزشگر را به کارانداخته یا نرمافزار را ایجاد می‌کند. با این حال آرشیوسازی و ب اغلب مستلزم مهارت‌های ترکیبی است و در عمل وظایف اغلب در سراسر بخش‌های تخصصی، توزیع می‌شود. کارکنان باید درک اساسی و دانش وسیعی از آرشیوسازی و ب، فراتر از تخصص حرفه‌ای خود داشته باشند زیرا این امر موجب بهبود همکاری و دستیابی به عملکرد بهتر می‌شود.

کارها را می‌توان در بخش‌های مختلف توزیع کرد و بسیاری از کارکنان می‌توانند مشارکت نمایند. تعداد کارکنان استخدام شده در آرشیوسازی و ب با جمع زمان صرفشده توسط تمامی کارکنان دائمی و موقتی در آن بخش از جمله کارکنان پروژه‌ای، محاسبه می‌شود. چندین شیوه وجود دارد:

الف- برآورد: تعداد پست‌های معادل تمام وقت که مستقیماً به آرشیوسازی و ب تخصیصی داده شده-اند را محاسبه کنید. میانگین زمان مصرفشده توسط این کارمندان برای فعالیت‌هایی غیر از آرشیوسازی و ب را تخمین زنید و این زمان را از تعداد کم کنید. میانگین زمان مصرفشده توسط کارمندان دیگر بخش‌ها که به فعالیت‌های آرشیوسازی و ب اختصاص داده شده است را تخمین زنید و این زمان را با تعداد پست‌های معادل تمام وقت، جمع کنید.

مثال: معادل تمام وقت $\frac{3}{5}$ نفر به طور مستقیم به فعالیت‌های آرشیوسازی و ب اختصاص یافته است. در طول دوره گزارش، آن‌ها 10% از زمان خود را صرف وظایف دیگر کرده‌اند. کارکنان بخش‌های دیگر (با معادل تمام وقت ۸ نفر) 20% از وقت خود را در فعالیت‌های مرتبط با آرشیوسازی و ب سپری کرده‌اند. در این صورت مقدار کل معادل تمام وقت برای آرشیوسازی و ب برابر است با: $475 = \frac{1}{6} \times 35 + \frac{3}{5} \times 8$.

ب- ثبت زمانی: یک دوره نمونه‌گیری (به طور معمول یک یا دو هفته) را انتخاب نمایید که در طول آن آرشیوسازی و ب، بارکاری متوسطی را تجربه می‌کند. زمانی (به کمک برنامه‌های کاری) که کارکنان، از جمله کارکنان دیگر بخش‌ها، به فعالیت‌های مرتبط با آرشیو و ب اختصاص داده‌اند را ثبت کنید. نتیجه به دست آمده از این دوره نمونه‌گیری، تعداد نمونه‌ها با تعداد اف.تی.ای در مدت زمان گزارش دهی برابر می‌باشد.

هزینه‌ها را می‌توان نسبت به کار متصدی، فنی و مدیریت تفکیک کرد.

۴-۶-۵ سایر هزینه‌ها

دیگر هزینه‌ها می‌تواند شامل موارد زیر باشد:

- کسب فراداده، برای مثال خریداری فهرست نامهای دامنه از ثبت‌کننده‌های دامنه؛
- متخصصان قانونی: مشاوره قانونی، مقررات اقدام قانونی یا غرامت در پی رأی دادگاه. (برای مثال یک ناشر از یک موسسه جمع‌آوری، به منظور اعاده حیثیت و یا خسارت مالی شکایت می‌کند؛
- همکاری بین‌المللی: آرشیو سازی وب توسط یک انجمن جهانی اجراء و پشتیبانی می‌شود. مشارکت در همکاری‌های بین‌المللی ممکن است تحمیل هزینه‌های حق عضویت و مخارج مسافرت و معیشت را به دنبال داشته باشد.

جدول ۹ – آمارهای اصلی برای هزینه‌های مجموعه

مثال	هدف	آمار
هزینه تعویض زیرساخت ذخیره‌سازی پنجاه هزار یورو بود	هزینه‌های خریداری و نگهداری سخت‌افزار	هزینه‌های سخت‌افزاری
هزینه پهنه‌ای باند در هر سال ده هزار یورو است	هزینه‌های مربوط به برق و شبکه	هزینه‌های محاسباتی
توسعه ابزار متعدد جدید، با هزینه هشتاد هزار یورو برونو سپاری شده است	هزینه‌های خریداری، ادغام، توسعه، یا بهبود نرم‌افزار	هزینه‌های نرم‌افزاری
کارگروه آرشیو سازی وب شامل ۳ مهندس تمام وقت و ۴ متعدد تمام وقت می‌شود	هزینه‌های منابع انسانی (برای مثال، متعددان، مهندسان، ...). بر حسب معادل تمام وقت یا مقدار پول	هزینه‌های نیروی کار

۵ شاخص‌های کیفیت**۱-۵ کلیات**

کیفیت به عنوان "میزانی که مجموعه‌ای از ویژگی‌های ذاتی الزامات را برآورده می‌سازد"، تعریف شده است (استاندارد ملی ایران - ایزو ۹۰۰۰: سال ۱۳۸۷). این بند، شامل شاخص‌هایی است که ارزیابی میزانی که مجموعه‌ای از ویژگی‌های ذاتی یک برنامه آرشیو سازی وب، الزامات تعیین شده از سوی مدیریت و سهامدارانش را برآورده کرده است، امکان‌پذیر می‌سازد.

هدف از شاخص‌های کیفیت پیشنهاد شده در این بخش، کمک به موسسات جمع‌آوری برای پاسخ‌دادن به پرسش‌های بنیادی از قبیل موارد زیراست:

- آیا می‌دانیم چه چیزی را می‌خواهیم جمع‌آوری کنیم؟
- اگر نه، خطمشی شفاف و روشنی که محدوده آرشیو را تعریف می‌کند، مورد نیاز است.

- آیا ما آنچه را که می‌خواهیم، جمع‌آوری می‌کنیم؟
- اگر نه، لازم است از سازگاری بین منابع جمع‌آوری شده و هدف، اطمینان حاصل شود.
- آیا ما بهترین استفاده را از منابع خود می‌کنیم؟
- اگر نه، لازم است شیوه‌کار و گردش‌کاری^۱ را برای افزایش کارائی، بهبود بخشیم.
- آرشیو تا چه حد قابل‌دسترس و قابل‌جستجو است؟
- بهبود دائمی قابلیت استفاده از آرشیو مهم است.
- می‌توانیم تضمین کنیم که آرشیووب با گذشت زمان قابل‌دسترس خواهد ماند؟
- اگر نه، لازم است شیوه‌های محافظتی قابل‌اطمینانی را به کار ببریم.

شاخص‌های پیشنهادی برای ارزیابی میزان تغییرات در کیفیت خدمات سازمان در طول زمان، مناسب‌ترین هستند. توصیه می‌شود کیفیت به‌طور منظم اندازه‌گیری و ارزیابی شود. اگر شاخص‌ها در سازمان‌های مختلف به یک شکل به کار گرفته و تفسیر شوند، آن‌گاه ترازیابی بین موسسات از طریق این شاخص‌ها نیز امکان‌پذیر می‌شود. با این‌حال، توصیه می‌شود چنین مقایسه‌هایی همیشه با احتیاط انجام شود و تفاوت‌ها در بودجه موسسه، منابع و شیوه‌های کاری سازمان‌ها در نظر گرفته شود.

۲-۵ محدودیت‌ها

شاخص‌های کیفی پیشنهادی در این استاندارد، وضعیت‌کنونی فناوری در آرشیوسازی‌وب را از دو چشم‌انداز فناورانه و متصدی‌گرایانه، نشان می‌دهد. آن‌ها باید هم‌گام با توسعه در آرشیوسازی‌وب بازبینی روزآمد شوند.

تفسیر نتایج به کار گیری شاخص‌های کیفیت باید با احتیاط صورت‌پذیرد. خطاهای نمونه‌گیری و اندازه‌گیری ممکن است به تفسیر نادرست شود.

۳-۵ توصیف

۱-۳-۵ کلیات

شاخص‌های کیفی براساس ارتباط آنها با جنبه‌های کلیدی برنامه‌های آرشیوسازی‌وب (خطمشی، برداشت، دسترسی و محافظت)، در زیر فهرست شده‌اند:

وقتی محاسبات ممکن است بر مبنای یو.آر.آل‌ها یا بایت‌ها انجام شود، استفاده از محاسبه بر حسب یو.آر.آل‌ها، توصیه می‌شود. در حقیقت، هدف این شاخص‌ها محاسبه مقدار منابع است و یو.آر.آل‌ها نسبت به بایت‌ها شباهت بیشتری به منابع کتابخانه‌ای دارند.

- مدیریت

۱- هزینه برای هر یو.آر.آل جمع آوری شده

۲- درصد کارکنان دخیل در فعالیت آرشیوسازی وب

- کیفیت فرایند جمع آوری

۳- درصد منابع ناپدیدشده از وب زنده در طول یک دوره زمانی معین

۴- درصد منابع دریافت شده از محدوده مورد توقع

۵- درصد درخواست‌ها برای توافق یا مجوزهای اعطای شده از سوی دارندگان حقوق

- قابلیت دسترسی و استفاده

۶- درصد منابع در دسترسی کاربران نهایی

۷- درصد منابع نمایه‌سازی شده به صورت تمام‌متن

۸- درصد منابع فهرست‌شده

۹- درصد سالانه منابع مورد دسترس واقع شده

۱۰- درصد بازدید از کتابخانه که شامل بازدیدها از آرشیوب و نیز می‌شود

۱۱- تعداد صفحات دیده شده در هر بازدید

- محافظت

۱۲- درصد منابعی که حداقل یک رونوشت دیگر از آن‌ها وجود دارد

۱۳- درصد منابع از دست‌رفته یا مخدوش‌شده

۱۴- درصد منابع با قالب فایل مشخص‌شده

۱۵- درصد منابعی که قالب آن‌ها دارای راهبرد محافظتی مشخص است

۱۶- درصد منابعی که ویروس‌یابی شده‌اند

۵-۳-۲ مدیریت

شماره شاخص	۱
نام	هزینه هر یو.آل جمع آوری شده
هدف	ارزیابی اثربخشی فرآیندهای آرشیوسازی وب.
پیش‌نیازها	- هزینه کل آرشیوسازی وب همانگونه که در زیربند ۴-۶ توضیح داده شده است؛ - تعداد کل یو.آل‌های خوش شده.
روش	هزینه هر یو.آل جمع آوری شده عبارت است از: $\frac{A}{B}$ که در آن: A برابر با هزینه کل آرشیوسازی در یک دوره زمانی خاص و معین B تعداد کل یو.آل‌های خوش شده در دوره زمانی مشابه است به یک رقم اعشاری گردشود.
توضیحات	هزینه‌ی پایین هر یو.آل جمع آوری شده به طور کلی نشان‌دهنده اثربخشی بالای فرآیندهای آرشیوسازی است: هزینه‌های بیشتر هر یو.آل جمع آوری شده، همچنین می‌تواند نشان‌دهنده سطح بالای تصدی‌گری باشد. این شاخص برای مقایسه آرشیوهایی با اندازه و هدف یکسان مناسب است.

شماره شاخص	۲
نام	درصد کارکنان دخیل در آرشیوسازی وب
هدف	نشان دادن میزان تعهد موسسه‌ای به آرشیوسازی وب.
پیش‌نیازها	- تعداد کارکنان (معادل تمام وقت) دخیل در آرشیوسازی؛ - تعداد کل کارکنان کتابخانه (معادل تمام وقت).
روش	درصد تعداد کارکنان فعال در آرشیوسازی وب برابر است با: $\frac{A}{B} \times 100$ که در آن: A تعداد کارکنان کتابخانه (معادل تمام وقت) دخیل در بخش آرشیو و براحتی، فراهم‌سازی دسترسی و حفظ و نگهداری B تعداد کل کارکنان کتابخانه (معادل تمام وقت)، شامل کارکنان دائمی و موقت و همچنین کارمندان مرتبط با پروژه است. به نزدیک‌ترین عدد صحیح گرد شود. در مورد کارمندان تمام وقتی که تنها به صورت پاره وقت در بخش آرشیوسازی وب فعالیت می‌کنند، مدت زمانی که صرف فعالیت‌های مرتبط با آرشیوسازی می‌کنند، می‌تواند از طریق خوداظهاری و یا ثبت زمانی، محاسبه شود.
توضیحات	

۳-۳-۵ کیفیت فایند جمع‌آوری

شماره شاخص	۳
نام	درصد منابعی که از وب زنده طی یک مدت زمان معین، ناپدید می‌شوند
هدف	ارزیابی ارزش آرشیووب.
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد مقاصد در آرشیو؛ - تعداد مقاصد در آرشیو که از وب زنده ناپدید شده‌اند (پاک شده‌اند).
روش	<p>عدد مورد دوم را ممکن است به‌طور خودکار با دریافت نکردن پاسخ دی.ان.اس و یا پاسخ‌های ۴۰۴ یا به صورت دستی از طریق بررسی وب زنده، جمع‌آوری کرد.</p> <p>درصد منابعی که طی دوره‌ی زمانی معین از وب زنده ناپدید می‌شوند، به شکل زیر محاسبه می‌شود:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <ul style="list-style-type: none"> A تعداد مقاصد آرشیو ناپدید شده است و B تعداد مقاصد موجود در آرشیو است. <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	<p>تعیین ناپدید شدن یک مقصد آرشیو، دشوار است. مقصد آرشیوی که دامنه (آدرس اینترنتی) آن تغییرکرده است، الزاماً ناپدید نمی‌شود زیرا می‌تواند دامنه متفاوت جدیدی داشته باشد. در برخی مواقع تنها بخش‌هایی از یک مقصد ناپدید می‌شود (پاک می‌شود) که به دلایل کاربردی و عملی، به این حالت نمی‌پردازیم.</p> <p>در تعریف این شاخص، یک مقصد آرشیو زمانی ناپدید می‌شود که هیچ گونه پاسخ دی.ان.اسی وجود ندارد و یا آدرس اینترنتی آن، یک پاسخ ۴۰۴ را تولید می‌کند. درصورت امکان‌پذیربودن، قابل اطمینان‌ترین روش، بررسی دستی وب زنده است.</p>

شماره شاخص	۴
نام	درصد آرشیو شده از محدوده اجباری
هدف	ارزیابی این موضوع که آیا نتایج آرشیوسازی با محدوده اجباری برای آرشیو، مطابقت دارد یا خیر.
پیش‌نیازها	- تعداد مقاصد آرشیو برداشت شده توسط موسسه در هر سال؛ - تعداد مقاصد آرشیوی که مطابق با محدوده مورد انتظار، تعیین شده‌اند یا از آن بهدست می‌آید.
روش	درصد آرشیوشده از محدوده مورد انتظار به صورت زیر محاسبه می‌شود: $\frac{A}{B} \times 100$ که در آن: A تعداد مقاصد آرشیو برداشت شده از کتابخانه در هرسال و B تعداد مقاصد آرشیوی که مطابق محدوده مورد انتظار تعیین شده‌اند یا آز آن به دست می‌آید (برای مثال انبارش قانونی). به یک رقم اعشاری گرد شود.
توضیحات	محدوده اجباری برای آرشیوسازی، اغلب ملی و نیز موسسه‌ای است. هرگاه قوانین ملی در خصوص ذخیره وجود داشته باشد، معمولاً قلمرویی از دامنه‌های ملی تعریف می‌شود. برای منابع وب که با استفاده از تی.ال.دی‌های ملی میزبانی شده‌اند، این قلمرو به سادگی قابل تعریف است. اما تمامی منابعی که در قلمرو ملی قرارمی‌گیرند، همیشه تحت نامهای دامنه‌ای با مراجع مشخص، میزبانی نمی‌شوند. برای مثال، مطابق اطلاعات موجود در AFNIC (دفتر ثبت .fr)، تنها یک سوم از سایتها فرانسوی روی تی.ال.دی‌های .fr میزبانی می‌شوند. در اینجا به روش‌های دیگری، به جز توجه به تی.ال.دی‌های ملی، برای تشخیص قلمروهای مورد انتظار، نیاز است. از جمله می‌توان به کسب اطلاعات از ثبت‌کننده‌های نام دامنه اشاره کرد.

شماره شاخص	۵
درصد درخواست‌ها برای اخذ موافقت یا مجوزهای اعطای شده توسط صاحبان حقوق	نام
ارزیابی اثربخشی درخواست‌های مجوز	هدف
<ul style="list-style-type: none"> - تعداد درخواست‌ها برای اخذ موافقت یا مجوزهای ارسالی به صاحبان حقوق؛ - تعداد موافقت‌ها یا مجوزهای اعطای شده توسط صاحبان حقوق. 	پیش‌نیازها
<p>درصد درخواست‌ها برای اخذ موافقت‌ها یا مجوزهای اعطای توسط صاحبان حقوق، به صورت زیر محاسبه می‌شود:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد موافقت‌ها یا مجوزهای اعطای شده توسط صاحبان حقوق و</p> <p>B تعداد درخواست‌ها جهت اخذ موافقت یا مجوزهای ارسالی برای صاحبان حقوق است</p> <p>به یک رقم اعشاری گرد شود.</p>	روش
<p>رتبه بالا، نشان‌دهنده موفقیت روندِ فعالیت درخواست مجوز است.</p> <p>توصیه می‌شود که تعداد درخواست‌هایی که به صورت صریح رد شده‌اند و نیز تعداد درخواست‌های بدون پاسخ، ثبت شوند.</p> <p>برقراری یک ارتباط بهتر و حمایت‌گر از ایجاد کنندگان وب‌گاه‌های اینترنتی، می‌تواند راه حل مناسبی برای رسیدن به موفقیت باشد. یک طرح ارتباطی می‌تواند برای تعیین راهکارهای کلیدی جهت مقاعده‌سازی ایجاد کنندگان وب‌گاه‌ها و همچنین به منظور تصمیم‌گیری در زمینه انتخاب بهترین ماجرا (کانال) برای توزیع اطلاعات، طرح‌ریزی شود.</p>	توضیحات

۴-۳-۵ قابلیت دسترسی و کاربرد

شماره شاخص	۶
نام	درصد منابع قابل دسترسی برای کاربران نهایی
هدف	ارزیابی دسترسی‌پذیری به آرشیووب.
پیش‌نیازها	<p>- مجموع تعداد منابع موجود در بخش آرشیو،</p> <p>- تعداد منابع موجود برخط (آرشیو برخط)،</p> <p>- تعداد منابع موجود در سایت (آرشیو خاکستری).</p>
روش	<p>موارد بالا را می‌توان بر حسب یو.آر.ال‌ها یا بایت‌ها، محاسبه کرد.</p> <p>درصد منابع قابل دسترسی برای کاربران نهایی به صورت زیر محاسبه می‌شود:</p> $\frac{(A + A')}{B} \times 100$ <p>که در آن:</p> <p>A تعداد منابع در آرشیووب با قابلیت دسترسی به صورت برخط؛</p> <p>A' تعداد منابعی است که تنها روی آرشیووب قابل دسترسی هستند و</p> <p>B تعداد کل منابع موجود در آرشیووب است.</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	<p>رتبه بالا، نشان‌دهنده قابلیت رویت یا قابلیت دسترسی بالای آرشیو وب است.</p> <p>این شاخص را می‌توان تنها برای منابع قابل دسترسی به صورت برخط محاسبه کرد تا قابلیت دسترسی مستقیم به منابع برای کاربران نهایی، تعیین شود.</p> <p>واحد اندازه‌گیری مورد استفاده در محاسبه باید گزارش شود، به عنوان مثال یو.آر.ال‌ها یا بایت‌ها.</p>

شماره شاخص	۷
نام	درصد منابع نمایه‌سازی شده به صورت تمام متن
هدف	ارزیابی قابلیت جستجوی آرشیووب
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل منابع در آرشیووب؛ - تعداد منابعی که به صورت تمام متن نمایه‌گذاری شده‌اند. <p>این محاسبه می‌تواند بر حسب یو.آر.ال‌ها یا بایت‌ها، انجام گیرد.</p>
روش	<p>درصد منابع نمایه‌گذاری شده به صورت تمام متن عبارت است از:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد منابع نمایه‌سازی شده به صورت تمام متن و</p> <p>B تعداد کل منابع در آرشیووب است</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	جستجوهای تمام متن، به میزان زیادی سبب افزایش قابلیت دسترسی و قابلیت استفاده از آرشیووب می‌شوند. واحد اندازه‌گیری که برای محاسبه استفاده می‌شود، باید گزارش شود، به عنوان مثال، یو.آر.ال‌ها یا بایت‌ها.

شماره شاخص	۸
نام	درصد منابع فهرست شده
هدف	ارزیابی قابلیت جستجو و سطح تصدی گری آرشیووب.
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد مقاصد در آرشیووب؛ - تعداد مقاصد آرشیوی که دارای سابقه فهرست هستند.
روش	<p>درصد منابع فهرست شده عبارت است از:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد مقاصد آرشیوی است که سابقه فهرست دارد و</p> <p>B تعداد کل مقاصد در آرشیووب است</p> <p>به یک رقم اعشار گرد شود.</p>
توضیحات	<p>فهرست‌کردن منابع در آرشیووب به افزایش میزان دسترسی و قابلیت استفاده آرشیو کمک می‌کند. این کار همچنین به یکپارچگی منابع آرشیووب با دیگر منابع موجود در کتابخانه کمک می‌کند. البته فهرست‌کردن هزینه‌بر است و ممکن است در مورد آرشیوهای وب با مقیاس بزرگ که از طریق خرشنگ‌های اینبوه جمع‌آوری شده‌اند، امکان پذیر نباشد.</p> <p>توصیه می‌شود هنگام محاسبه این شاخص، راهبرد برداشت استفاده شده برای جمع‌آوری این منابع را نیز گزارش کنید.</p>

شماره شاخص	۹
نام	درصد منابع مورد دسترس واقع شده
هدف	ارزیابی میزان وسعت استفاده واقعی از آرشیووب
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل نامهای دامنه در آرشیووب؛ - تعداد نامهای دامنه که در طول سال حداقل یک صفحه از آن‌ها، مشاهده شده باشد.
روش	<p>درصد منابع مورد دسترس واقع شده برابر است با :</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد نامهای دامنه‌ای است که در طول سال حداقل یک صفحه از آن‌ها مشاهده شده باشد و</p> <p>B تعداد کل نامهای دامنه در آرشیووب است</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	<p>درصد بالا، نشان‌دهنده استفاده گسترده از آرشیووب است. دلیل به کارگیری نامهای دامنه به جای یو.آر.ال‌ها، این است که امکان دارد تنها منابعی که تحت یکسری دامنه‌های محدود هستند، به طور فعال استفاده شوند. محاسبه میزان استفاده برای هر دامنه، نشان‌دهنده سطح استفاده و جامعیت آرشیو است.</p> <p>برقراری ارتباط بهتر با پژوهشگران و همچنین با توده مردم راهی برای افزایش این درصد است. همان‌گونه که درمورد ساختار ۵ اظهار شد، یک طرح ارتباطی می‌تواند برای تعیین نمونه‌های متقاعد کننده‌تر از منابع با ارزش و همچنین به منظور تصمیم‌گیری در زمینه بهترین کانال‌های مربوط به توزیع اطلاعات، طرح‌ریزی شود.</p>

شماره شاخص	۱۰
نام	درصد بازدیدها از کتابخانه شامل بازدید از آرشیووب
هدف	ارزیابی میزان استفاده از آرشیووب توسط بازدیدکنندگان کتابخانه (از طریق سایت یا به صورت برخط).
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل بازدیدها از کتابخانه (بازدیدهای فیزیکی و مجازی)؛ - تعداد بازدیدهای آرشیووب.
روش	<p>درصد بازدیدکنندگان کتابخانه که از آرشیووب استفاده می‌کنند، برابر است با:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد بازدیدهای آرشیوهای وب (مثلاً در هر ماه یا هر سال) و</p> <p>B تعداد کل بازدیدهای کتابخانه (مثلاً در هر ماه یا هر سال) است</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	<p>رتیبه بالا، حاکی از استفاده فراوان کاربران دیگر خدمات‌های های کتابخانه از آرشیووب است. بازدیدهای فیزیکی، به معنای ورود کاربران به میانی کتابخانه است.</p> <p>بازدیدهای مجازی، به معنای بازدید از وب‌گاه کتابخانه است.</p>

شماره شاخص	۱۱
نام	تعداد صفحات مشاهده شده در هر بازدید
هدف	ارزیابی علاقه کاربران به آرشیو.
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل صفحات مشاهده شده در یک دوره گزارش دهی؛ - تعداد کل بازدیدهای آرشیو و ب درهمان دوره.
روش	<p>تعداد صفحات مشاهده شده در هر بازدید برابر است با:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد کل صفحات مشاهده شده در یک دوره گزارش دهی و</p> <p>B تعداد کل بازدیدهای آرشیو و ب درهمان دوره است.</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	<p>نسبت بالا، نشان‌دهنده استفاده فراوان از آرشیو و ب است، که به معنای این است که آرشیو حاوی مطالب مورد استفاده کاربران است.</p> <p>یکی از روش‌های افزایش این ضریب، توجه کردن به نیازها و انتظارات کاربران و پژوهشگران، به عنوان مثال از طریق انجام دادن پژوهش و بررسی‌ها است.</p>

۵-۳-۵ محافظت

شماره شاخص	۱۲
نام	درصد منابعی که حداقل یک رونوشت از آن‌ها وجود دارد
هدف	ارزیابی قابلیت محافظت جریان بیت است.
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل منابع در آرشیو و ب؛ - تعداد منابعی که حداقل یک رونوشت از آن‌ها وجود دارد. <p>این محاسبه را می‌توان بر حسب یو.آر.آل‌ها و یا بایت‌ها انجام داد</p>
روش	<p>درصد منابعی که حداقل یک رونوشت از آن‌ها وجود دارد برابر است با:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد منابعی است که حداقل یک رونوشت از آن‌ها وجود دارد و</p> <p>B تعداد کل منابع در آرشیو و ب است</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	واحد اندازه‌گیری مورد استفاده جهت محاسبه باید گزارش شود، مثلاً یو.آر.آل‌ها و یا بایت‌ها

۱۳	شماره شاخص
درصد منابع از بین رفته یا مخدوش	نام
ارزیابی امنیت ذخیره‌سازی در آرشیووب.	هدف
<ul style="list-style-type: none"> - تعداد کل منابع در آرشیووب؛ - تعداد منابع از بین رفته یا مخدوش. <p>این محاسبه، ممکن است بر حسب یو.آر.آل‌ها و یا بایت‌ها انجام شود.</p>	پیش‌نیازها
درصد منابع از بین رفته و یا مخدوش برابر است با:	روش
$\frac{A}{B} \times 100$	
<p>که در آن:</p> <p>A تعداد منابع از بین رفته یا مخدوش شده و</p> <p>B تعداد کل منابع موجود در آرشیووب است</p> <p>به یک رقم اعشاری گرد شود.</p>	
درصد کم، نشان‌دهنده امنیت زیاد است. واحد اندازه‌گیری مورد استفاده برای محاسبه باید گزارش شود، مانند یو.آر.آل‌ها یا بایت‌ها.	توضیحات

۱۴	شماره شاخص
درصد منابع با قالب پروژه شناسایی شده	نام
ارزیابی دانش سازمانی آرشیووب و قابلیت محافظت است.	هدف
<ul style="list-style-type: none"> - تعداد کل منابع در آرشیووب؛ - تعداد منابع با قالب پروژه شناسایی شده. <p>این محاسبه را می‌توان بر حسب یو.آر.آل‌ها یا بایت‌ها انجام داد.</p>	پیش‌نیازها
درصد آرشیو با قالب فایل مشخص برابر است با :	روش
$\frac{A}{B} \times 100$	
<p>که در آن:</p> <p>A تعداد منابع با قالب فایل شناسایی شده و</p> <p>B تعداد کل منابع موجود در آرشیووب است.</p> <p>به یک رقم اعشاری گرد شود.</p>	
واحد اندازه‌گیری مورد استفاده، باید گزارش شود به عنوان مثال یو.آر.آل‌ها یا بایت‌ها. این درصد ممکن است با توسعه بیشتر ابزارهای شناسایی (به منظور بهبود عملکرد آن‌ها) یا با بهبود فرایند شناسایی (اضافه کردن ماشین‌ها و غیره)، بهبود یابد	توضیحات

شماره شاخص	۱۵
نام	درصد منابعی که قالب آن‌ها راهبرد محافظتی مشخص دارد
هدف	ارزیابی تعهد سازمانی به محافظت منطقی از آرشیووب.
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل منابع در آرشیووب؛ - تعداد منابع با قالبی که راهبرد محافظتی مشخصی دارد. <p>این محاسبه را می‌توان بر حسب یو.آر.آل‌ها و یا بایت‌ها، انجام داد</p>
روش	<p>درصد منابعی که قالب آن‌ها راهبرد محافظتی مشخص دارد، برابر است با :</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد منابعی است که قالب آن‌ها راهبرد محافظتی مشخص دارد و</p> <p>B تعداد کل منابع موجود در آرشیووب است.</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	واحد اندازه‌گیری مورد استفاده جهت محاسبه باید گزارش شود، به عنوان مثال یو.آر.آل‌ها یا بایت‌ها.

شماره شاخص	۱۶
نام	درصد منابع ویروس‌یابی شده
هدف	ارزیابی میزان اینم بودن استفاده از آرشیووب برای دیگر مجموعه‌ها و نیز برای تجهیزات کاربران
پیش‌نیازها	<ul style="list-style-type: none"> - تعداد کل منابع موجود در آرشیووب؛ - تعداد منابع ویروس‌یابی شده. <p>این محاسبه را می‌توان بر حسب یو.آر.آل‌ها یا بایت‌ها، انجام داد</p>
روش	<p>درصد منابع ویروس‌یابی شده عبارت است از:</p> $\frac{A}{B} \times 100$ <p>که در آن:</p> <p>A تعداد منابعی که ویروس‌یابی شده است و</p> <p>B تعداد کل منابع موجود در آرشیووب است</p> <p>به یک رقم اعشاری گرد شود.</p>
توضیحات	هنگام اشکار شدن ویروس‌ها، ممکن است خط مشی موسسات این باشد که آن‌ها را از آرشیو پاک نکنند، بلکه امکان دسترسی به فایل‌های ویروسی شده را مسدود نمایند. واحد اندازه‌گیری مورد استفاده جهت محاسبه باید گزارش شود، مانند یو.آر.آل‌ها یا بایت‌ها.

۶ استفاده و مزایا

۱-۶ کلیات

ارائه آمارها و شاخص‌های کیفی در ارزیابی و هم‌سنگی عملکرد آرشیووب مفید است و توصیه می‌شود بخشی از طرح‌ریزی و گردش‌کاری منظم موسسه‌ای را تشکیل دهد. در ارائه نظامند ارقام آماری و ارزیابی کیفی یک آرشیووب، مزایای بسیاری وجود دارد:

- امکان تصمیم‌گیری آگاهانه را فراهم می‌سازد.
- کمک شایانی به مدیریت عملیات آرشیوسازی وب از طریق ارائه نکات ضروری برای نظارت می‌کند.
- امکان گفتگو بین موسسات آرشیوسازی وب، نهادهای پشتیبانی و تامین‌کننده بودجه و جامعه کاربران را تسهیل می‌سازد.
- سطح آگاهی عمومی از آرشیوسازی وب را به‌طور کلی بالا می‌برد و بهترین رویه‌ها را تقویت می‌سازد.
- در تشخیص کمبودها و مشکلات متداول مفید خواهد بود، به گونه‌ای که می‌توان به این موارد به صورت جمعی و از طریق همکاری بین موسسات جمع‌آوری، رسیدگی کرد.
- شواهدی از پیشرفت را ارائه می‌دهد و به نشان دادن ارزش آرشیوهای وب، کمک می‌کند.

۲-۶ کاربردها و مخاطبین مورد نظر

از اصطلاحات آماری و شاخص‌های کیفی که در این استاندارد تعریف و توصیف شده‌اند، می‌توان به منظور ارزیابی کارایی آرشیوهای وب، بهره گرفت. همچنین می‌توان از آن‌ها برای همترازی و مقایسه آرشیوهای وب استفاده کرد.

ارقام آماری داده‌های خنثی و عینی به شمار می‌روند که مبنایی را برای تحلیل‌ها و تفسیرهای بیشتر فراهم می‌سازند. شاخص‌های کیفی داوری مبتنی بر ارزش را به نمایش می‌گذارند: نتایج به کارگیری شاخص، خوب یا بد بودن کارایی را نشان می‌دهد.

علاوه بر این، این استاندارد به ارائه یک مرور کلی و همچنین یک توصیف تفصیلی از انواع فرایندهای فنی می‌پردازد. اگر چه تمام این استاندارد برای همه مخاطبین مفید می‌باشد، اما جدول ۱۰ محتوای اصلی برای هر یک از سه گروه کاربر اصلی مورد نظر را به تفکیک نشان می‌دهد که توصیه می‌شود بخش‌های خاصی از استاندارد را با جزئیات کامل مطالعه کنند.

جدول ۱۰- کاربردها و مخاطبان موردنظر

دسترسی و محافظت	برداشت و گردآوری	گروه کاربران
۱،۳،۴،۳،۵،۴،۲،۱،۴،۲،۶،۴،۳،۱،۴،۶،۵،۱	۱،۳،۴،۳،۵،۴،۴،۱،۴،۴،۵،۴،۵،۱،۴،۵،۵،۴،۶،۵،۱	مدیران
۳،۱،۳،۴،۴،۲،۵،۴،۳	۳،۲،۳،۳،۳،۴،۴،۴،۴،۵،۳،۴،۵،۴،۵	متصدیان
۳،۱،۴،۲،۴،۳	۳،۲،۳،۳،۴،۴،۴،۵	مهندسان

۳-۶ مزایا برای گروههای کاربری

این استاندارد می‌تواند برای تصمیم‌گیرندگان (مدیران) درون و یا خارج از موسسات جمع‌آوری، برای متصدیانی که وب‌گاهها را انتخاب و تصدی‌گری می‌کنند (کتابداران و آرشیوسازان) و برای کارکنان بخش فنی که این سامانه‌ها را مدیریت و زیرساختها را حفظ و نگهداری و عملیات فنی را پشتیبانی می‌کنند (مهندسان)، مفید واقع شود.

الف- برای مدیران

آمارها و شاخص‌های کیفی مربوط به جمع‌آوری و دسترسی، امکان طرح‌ریزی مناسب برای منابع را فراهم می‌کند و به ارزیابی میزان موفقیت آرشیوسازی با توجه به مأموریت‌های موسسه‌ای کمک می‌کند. شاخص‌های مربوط به هزینه‌ها، به ارزیابی و اولویت‌بندی توسعه مطابق با اولویت برنامه کمک می‌کنند. شاخص‌های کیفی همچنین امکان نمایش دستاوردها در بخش آرشیوسازی را با به کارگیری اصطلاحات رایج، فراهم می‌سازد.

ب- برای متصدیان

آمارهای مربوط به جمع‌آوری به متصدیان کمک می‌کند تا درانتخاب‌ها روی حوزه‌های کلیدی تمرکز کنند و امکان مقایسه فرایند انتخاب را با آنچه در دیگر سازمان‌ها انجام می‌شود، داشته باشند. شاخص‌های کیفی مربوط به دسترسی، به شناخت نیازهای آتی کمک می‌کنند. شاخص‌های کیفی مربوط به محافظت برای تصمیم‌گیری در مورد جمع‌آوری فراداده‌ها، مفید هستند. شاخص‌های کیفی همچنین امکان ارزیابی عملکرد متصدیان را فراهم می‌کنند.

پ- برای مهندسان

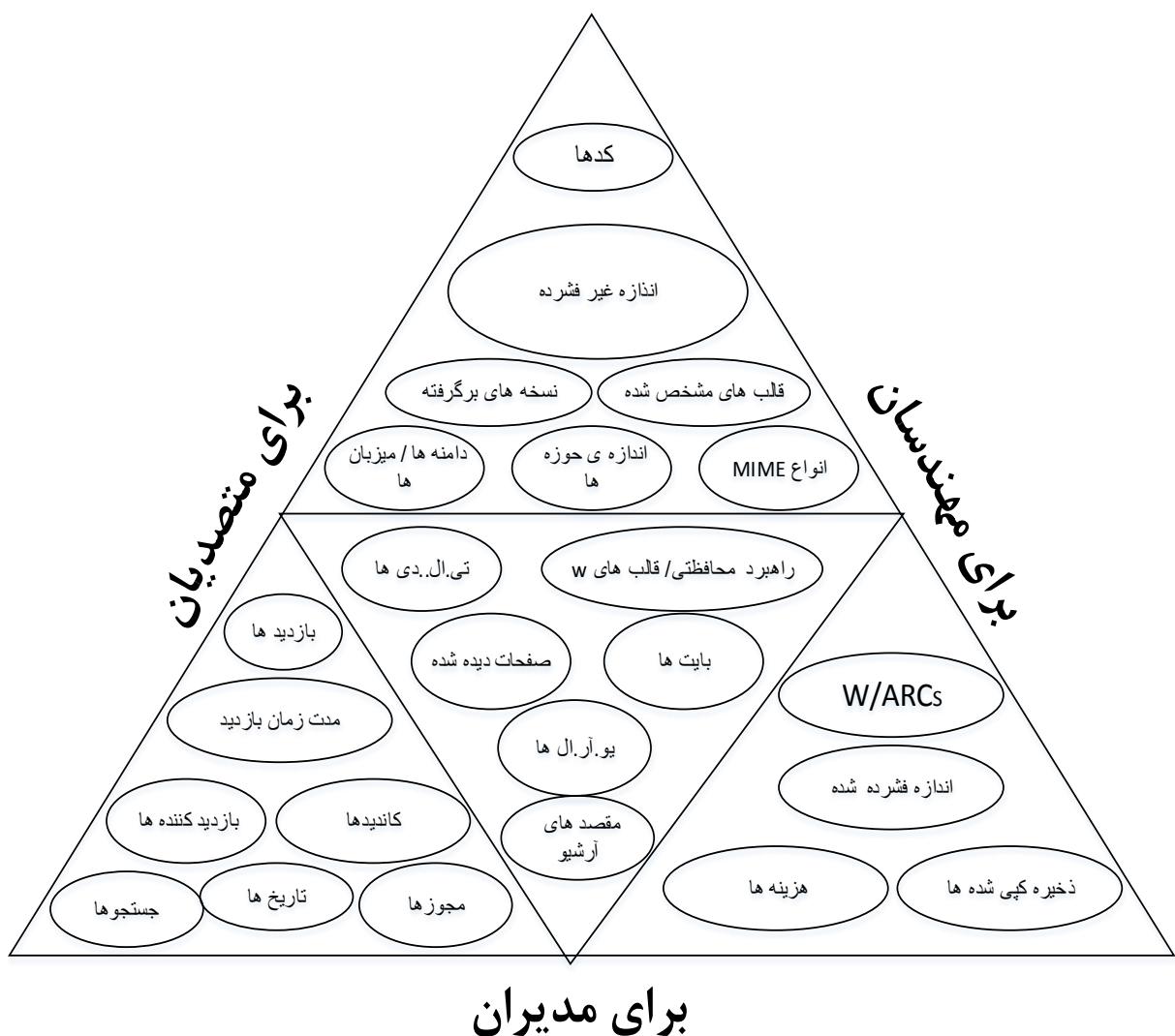
آمارها و شاخص‌های کیفی این امکان را به مهندسان می‌دهند تا درباره جنبه‌های مختلف فنی سامانه‌های آرشیوسازی وب، از جمله کارایی آن‌ها، به کسب اطلاعات و دانش بپردازند. ضروری

است به طور منظم به مهندسان گزارش داده شود تا آن‌ها نسبت به بهینه سازی خدمات و عملیات، اقدام کنند. آمارها و شاخص‌های کیفی مربوط به محافظت ضروری هستند و می‌توانند کمک کنند که سامانه‌هایی طراحی شوند تا از دسترسی پایدار به آرشیو وب، اطمینان حاصل شود.

۴-۶ به کارگیری آمارهای پیشنهادی توسط گروه کاربران

تمامی آمارها و شاخص‌های پیشنهاد شده در این استاندارد، به مخاطبین موردنظر آن مرتبط است. اما نمودار زیر، مجموعه اصلی^۱ برای برقراری ارتباط و بحث میان گروه‌های اصلی کاربران را بر جسته می‌سازد.

آمارهای ارائه شده در کناره‌های مثلث، ارتباط زیادی با دو گروه کاربری که اسمی آن در کنار ضلع مثلث نوشته شده دارد. آمارهای موجود در مثلث میانی ارتباط زیادی با تمامی گروه کاربران دارد. برای مثال، آمارهای موجود در ضلع بالایی این مثلث، بسیار زیاد به ارتباط مابین متصدیان و مهندسان ربط دارد.



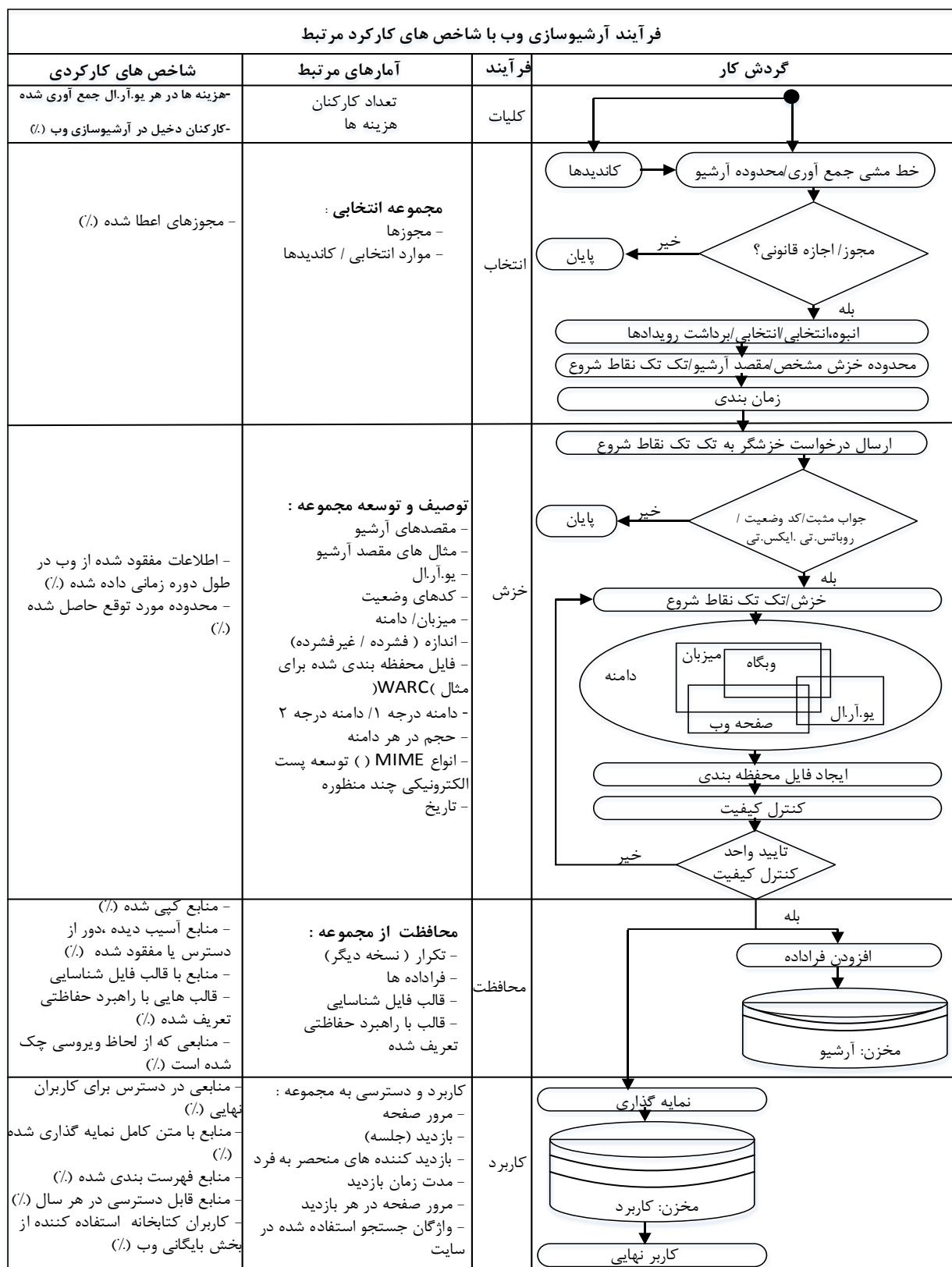
شکل ۱- اصطلاحات آماری مورد استفاده در شکل ۱ در جدول ۱۱ ارائه شده است

جدول ۱۱- اصطلاحات مورد استفاده در تصویر

نوع	آمار
بایت‌ها	اندازه آرشیو، فشرده یا غیرفشرده
نسخه‌های برگرفته	تعداد نسخه‌های رونوشت از مقصد آرشیو
کد‌ها	توزیع یو.آر.آل ها (رشته متنی که محل فایل یا برنامه‌ای را در اینترنت مشخص می‌کند) توسط کد وضعیت، و بهویژه سری ۲XX
اندازه فشرده شده	اندازه فشرده شده آرشیو بر حسب بایت
هزینه‌ها	سخت افزار، محاسبه، نرم افزار، هزینه‌ای کار
تاریخ‌ها	پوشش زمانی
دامنه‌ها / میزبان‌ها	تعداد دامنه‌ها یا میزبان‌ها
اندازه دامنه‌ها	توزیع حجم منابع در هر دامنه یا هر میزبان
قالب‌ها با راهبرد محافظت	تعداد قالب‌های فایل بر اساس راهبرد محافظتی تعیین شده
قالب‌های مشخص شده	توزیع بر اساس قالب‌های مشخص فایل
انواع ام.آی.ام.ای	توزیع براساس انواع قالب
کاندیدا	تعداد کاندیدا
صفحات دیده شده	تعداد صفحات دیده شده
مجوزها	تعداد مجوزها
ذخیره رونوشت‌شده‌ها	حجم منابع رونوشت‌شده
جستجو	رایج ترین واژه کاربردی در آرشیو
مقصد آرشیو	تعداد مقاصد آرشیو
تی.آل.دی‌ها	توزیع براساس دامنه‌های سطح ۱ و ۲
اندازه غیرفشرده	اندازه غیرفشرده آرشیو بر حسب بایت‌ها
یو.آر.آل‌ها	تعداد کلی یو.آر.آل‌ها (یا پاسخ‌ها)
مدت زمان بازدید	متوجه مدت زمان بازدید
بازدیدکنندگان	تعداد بازدیدکنندگان منحصر به فرد
بازدیدها	تعداد بازدیدها
دبليو/ اي.آر.سي‌ها	تعداد فایل‌های دبليو/ اي.آر.سي یا فایل‌های محفوظه‌بندی شده

۶-۵ فرایند آرشیو سازی وب با شاخص های کارایی مربوطه

آمارها و شاخص های کیفی پیشنهاد شده در این استاندارد، برروال گردش کار عمومی آرشیو سازی وب نگاشت شده اند تا ارتباط ما بین هر یک از آن ها با هر یک از فرایندهای آرشیو سازی، مشخص شود.



شکل ۲- فرآیند آرشیوسازی وب با شاخص های عملکردی مرتبط

کتاب نامه

- [۱] استاندارد ملی ایران- ایزو شماره ۹۰۰۰: ۱۳۸۷، سیستم های مدیریت کیفیت- مبانی و واژگان
- [۲] استاندارد ملی ایران شماره ۱۰۰۴۷: ۱۳۸۶، اطلاعات و مستندسازی- مدیریت سوابق- بخش اول: کلیات
- [۳] استاندارد ملی ایران شماره ۲۰۷۴۶: ۱۳۹۴، اطلاعات و مستندات، روش ها و رویه های ارزیابی تاثیر کتابخانه ها

[4] ISO 2789, Information and documentation — International library statistics

یادآوری- استاندارد ملی ایران ۹۱۰۴: ۱۳۸۶، اطلاعات و مستند سازی - آمار بین المللی کتابخانه ای با استفاده از ISO 2789:2006 تدوین شده است

[5] ISO 11620, Information and documentation — Library performance indicators

یادآوری- استاندارد ملی ایران شماره ۸۸۹۴: ۱۳۹۳، اطلاعات و دبیش(مستند سازی)- شاخص های عملکرد کتابخانه با استفاده از ISO 11620:2014 تدوین شده است.

[6] ISO 14721:2012, Space data and information transfer systems — Open archival information system (OAIS) — Reference model

[7] ISO 28500:2009, Information and documentation — WARC file format

[8] Digital R.A.M.B.O.R. A. (DRAMBORA), available from: <http://www.repositoryaudit.eu/>

[9] Global W.M. S. available from: <http://www.ifabc.org/>

[10] Transfer Protocol H. HTTP/1.1, available from: <http://www.ietf.org/rfc/rfc2616.txt>

[11] Encoding M., & Standard T. (METS), available from: <http://www.loc.gov/standards/mets/METSOversight.html>

[12] Multipurpose Internet Mail Extension (MIME) Part Two: Media Types, available from: <http://www.ietf.org/rfc/rfc2046.txt>

[13] Nestor Catalogue of Criteria for Trusted Digital Repositories. available from: http://files.d-nb.de/nestor/materialien/nestor_mat_08_eng.pdf

[14] Revised Guidelines For Statistical Measures Of Usage Of Web-Based Information Resources. available from: <http://iclc.net/statement/revised-guidelines-statistical-measuresusage-web-based-information-resources>

[15] Trusted Repositories Audit & Certification (TRAC), available from: http://www.crl.edu/sites/default/files/attachment/pages/trac_0.pdf

[16] Identifiers U.R. (URI), available from <http://www.ietf.org/rfc/rfc2396.txt>

[17] Web Analytics Association. available from: <http://www.webanalyticsassociation.org/?page=standards>

[18] Website metric definitions, available from: <http://www.jicwebs.org/standards.php>

[19] BALL. A. 2010. Web Archiving, available from: <http://www.dcc.ac.uk/sites/default/files/documents/report/sarwa-v1.1.pdf>

[20] BERMES. E. and ILLIEN, G. 2009. Metrics and Strategies for Web Heritage Management and Preservation, available from: <http://conference.ifla.org/past/ifla75/92-bermes-en.pdf>

[21] BRÜGGER. N. 2005. Archiving Websites. General Considerations and Strategies, available from: http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf

[22] DOUGHERTY. M., MEYER, E.T., MADSEN, C., VAN DEN HEUVEL, C., THOMAS, A. and WYATT, 2010.

Researcher Engagement with Web Archives: State of the Art. Report. London: JISC, available from: <http://ssrn.com/abstract=1714997>

- [23] DROID (Digital Record Object Identification). available from: <http://sourceforge.net/projects/droid/>
- [24] IIPC Access Working Group. 2006: Use cases for Access to Internet Archives, available from: <http://www.netpreserve.org/resources/use-cases-access-internet-archives>
- [25] JACOBSEN. G. 2007. Webarchiving Internationally: Interoperability in the Future? Results of a survey of Web archiving activities on national libraries. Paper published on the IFLANET prior to the World Library and Information Congress: 73rd IFLA General Conference and Council held in Durban, South Africa, available from: http://netarkivet.dk/publikationer/InteroperabilityInTheFuture_IFLA2007.pdf
- [26] JHOVE - JSTOR/Harvard Object Validation Environment. available from: <http://jhove.sourceforge.net/>
- [27] MASANÈS. J. (ed.). 2006. Web Archiving, Springer, Berlin.
- [28] MASANÈS. J. 2002. Towards Continuous Web Archiving. In: D-Lib Magazine 8 (12), available from: <http://www.dlib.org/dlib/december02/masanes/12masanes.html>
- [29] MEYER. E., THOMAS, A. and SCHROEDER, R. 2011: Web Archives: The Future(s), available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1830025
- [30] PINSENT, E., DAVIS, R., ASHLEY, K., KELLY, B., GUY, M. and HATCHER, J. 2010. PoWR: The Preservation of Web Resources Handbook, available from: <http://www.jisc.ac.uk/publications/programmerelated/2008/powrhandbook.aspx>
- [31] AUBRY. S. 2010: Introducing Web Archives as a New Library Service: the Experience of the National Library of France. In: Liber Quarterly, 2010, vol. 20, no. 2, available from: <http://liber.library.uu.nl/index.php/lq/article/view/7987>
- [32] BAILEY. S. and THOMPSON, D. 2006: UKWAC: Building the UK's First Public Web Archive. D-Lib. 2006, 12 (1). Available at: <http://www.dlib.org/dlib/january06/thompson/01thompson.html>
- [33] GLENN. V. 2007: Preserving Government and Political Information: The Web-at-Risk Project. First Monday. 2007, 12 (7). Available at: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fma/rticel/eview/1917/1799>
- [34] HOCKX-YU. H. 2011: The Past Issue of the Web. In: Proceedings of the ACM WebSci'11, WebscienceTrust, June 17, 2011, available from: <http://www.websci11.org/fileadmin/websci/Papers/PastIssueWeb.pdf>
- [35] HOCKX-YU. H., CRAWFORD, L. ROGER, C., JOHNSON, S. 2010: Capturing and Replaying StreamingMedia in a Web Archive – a British Library Case Study. In: Proceedings of iPRES 2010, September 2010, available from: <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf>
- [36] ILLIEN. G. 2008: L'archivage d'Internet, un défi pour les décideurs et les bibliothécaires: scénarios d'organisation et d'évaluation; l'expérience du consortium IIPC et de la BnF. In: Actes du 74e congrès de la Fédération internationale des associations de bibliothécaires et d'institutions (IFLA), Québec, Canada, available from: <http://archive.ifla.org/IV/ifla74/papers/107-Illien-fr.pdf>
- [37] ILLIEN. G. and STIRLING, P. 2011: The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In: Proceedings of the 77th IFLA congress, San Juan, Porto Rico, available from: <http://conference.ifla.org/past/ifla77/193stirling-en.pdf>
- [38] JACOBSEN. G. 2008: Web Archiving: Issues and Problems in Collection Building and Access. In: Liber Quarterly, Volume 18, Nr. 3/4 (2008), available from: <http://liber.library.uu.nl/index.php/lq/article/view/7936/8202>

- [39] LASFARGUES. F., OURY C. and WENDLAND B. 2008: Legal deposit of the French Web: harvestingstrategies for a national domain. In: Proceedings of the 8th International Web Archiving Workshop,Aarhus, Danemark, available from: [ht tp://iwaw.net/08/IWAW2008-LASFARGUES.pdf](http://iwaw.net/08/IWAW2008-LASFARGUES.pdf)
- [40] OURY C.. PEYRARD S. 2011: From the World Wide Web to digital library stacks: preserving theFrench Web archives. In: Proceedings of iPRES 2011, p. 231-241, available from: [ht tp://halshs.archives-ouvertes.fr/halshs-00868729](http://halshs.archives-ouvertes.fr/halshs-00868729)
- [41] POPE. J. and BERESFORD, Ph. 2007: IIPC Web Archiving Toolset Performance Testing at theBritish Library. In: Ariadne, no. 52 (2007), available from [ht tp://www.ariadne.ac.uk/issue52/pope-beresford/](http://www.ariadne.ac.uk/issue52/pope-beresford/)
- [4 2] RAUBER.A .,ASCHENBRENNER ,A. ,WITVOET,O . , B R U CKNER,R .and KAISER,M.2002: Un c o v e r i n g Information Hidden in Web Archives: A Glimpse at Web Analysis Building on Data Warehouses D-Lib. 2002, 8 (12). Available at: [ht tp://www.dlib.org/dlib/december02/rauber/12rauber.html](http://www.dlib.org/dlib/december02/rauber/12rauber.html)
- [43] SMITH. J., and NELSON, M. 2008: Creating Preservation-Ready Web Resources. D-Lib. 2008, 14 (1/2). Available at: [ht tp://www.dlib.org/dlib/january08/smith/01smith.html](http://www.dlib.org/dlib/january08/smith/01smith.html)
- [44] SPENCER. A., SHERIDAN, J., THOMAS, D. and PULLINGER, D. 2009: UK Government WebContinuit y: Persisting Access through Aligning Infrastructures. International Journal of DigitalCuration.2009,4(1).Availableat:<http://www.ijdc.net/index.php/ijdc/article/view/106>