www.Ghaemiyeh.com
www.Ghaemiyeh.org
www.Ghaemiyeh.net
www.Ghaemiyeh.ir

آمارها و شاخص‌های کیفیت در آرشیو وب

استاندارد ایزو / آی‌ای‌اس ۲۸۷۳


بسم الله الرحمن الرحیم

آمارها و شاخص های کیفیت در آرشیو وب

نویسنده:

معاونت پژوهش برنامه ریزی و فناوری اداره کل برنامه ریزی و توسعه سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

ناشر چاپي:

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

ناشر ديجيتالي:

مرکز تحقيقات رايانه‌اي قائميه اصفهان

فهرست

آمارها و شاخص های کیفیت در آرشیو وب

مشخصات کتاب

عنوان و نام پدیدآور:آمارها و شاخص های کیفیت در آرشیو وب (استاندارد ایزو/ تی آر 14873 : 2013) ویراست نخست (01/12/2013)/ [سازمان بین المللی استاندارد]؛ برگردان به فارسی فرزانه شادان پور؛ ویراستار آرزو تجلی؛ [برای] سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران، گروه پژوهش های توسعه ای فناوری اطلاعات.

مشخصات نشر:تهران : سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران، 1397.

مشخصات ظاهری:115 ص.: مصور، جدول.

شابک:150000 ریال 978-964-446-560-4 :

وضعیت فهرست نویسی:فاپا

یادداشت:پشت جلد به انگلیسی:. ... Information and documentation statistics and quality

یادداشت:کتابنامه: ص. [113] - 115.

موضوع:استاندارد ایزو/ تی.آر. 14873

ISO/TR 14873 Standard

آرشیوهای وب -- استاندارد ها

Web archives -- Standards

شناسه افزوده:شادان پور، فرزانه، 1344-، مترجم

شناسه افزوده:سازمان بین المللی استاندارد

شناسه افزوده:International Organization for Standardization

شناسه افزوده:سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران. اداره کل پژوهش و آموزش. گروه پژوهش های توسعه ای فناوری اطلاعات

شناسه افزوده:سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

رده بندی کنگره:Z701/3 /آ4 آ8 1397

رده بندی دیویی:025/284

شماره کتابشناسی ملی:5611474

دسترسی و محل الکترونیکی:http://dl.nlai.ir/UI/b9f8da72-9c58-4276-8195-91a8dee914b4/Catalogue.aspx

اطلاعات رکورد کتابشناسی:فاپا

خیراندیش دیجیتالی : انجمن مددکاری امام زمان (عج) اصفهان

ویراستار کتاب : خانم مهدیه نیلی خواجو

ص: 1

اشاره

بِسْمِ اللّٰهِ الرَّحْمنِ الرحيم

ص: 2

آمارها و شاخص های کیفیت در آرشیو وب

(استاندارد ایزو تی آر (14873 : 2013) ویراست نخست (2013/12/01)

برگردان به :فارسی فرزانه شادان پور

گروه پژوهش های توسعه ای فناوری اطلاعات

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

ص: 3

فهرستنویسی پیش از انتشار کتابخانه ملی جمهوری اسلامی ایران

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

آمارها و شاخص های کیفیت در آرشیو وب

برگردان به فارسی : فرزانه شادان پور

ویراستار: آرزو تجلی

ناشر : انتشارات سازمان اسناد و کتابخانه جمهوری اسلامی ایران

شمارگان 500 نسخه / چاپ اول 1397

چاپ و صحافی انتشارات سازمان اسناد و کتابخانه جمهوری اسلامی ایران آصف آصفی

طرح جلد شهره خوری صفحه آرایی : علیرضا خورسندی

ناظر فنی : پرویز بختیاری

بهاء : 110000 ریال

نشانی :(تهران بزرگراه شهید حقانی غرب به شرق) بلوار کتابخانه ملی تلفن فروشگاه : 81623315 دورنگار: 81623301

وبگاه : www.nlai.ir پست الکترونیک انتشارات: Publication@nlai.ir

1397

ص: 4

فهرست

سخن مترجم

آرشیو وب سال هاست که به عنوان بخشی از انتشارات ملی در کتابخانه های ملی کشورهای ،جهان و اغلب با همکاری سازمان های علمی و پژوهشی مرتبط با موضوع محور پژوهش ها اقدامات فناورانه متنوعی قرار گرفته .است در کشور ما این مهم به واسطه مشکلاتی مانند تغییرات مدیریتی و فقدان بودجه کافی هنوز تا تحقق راه درازی در پیش دارد؛ اما چند سالی است که دغدغه سازمان اسناد و کتابخانه ملی شده است.

و شاید این سوال به ذهن خطور کند که وقتی هنوز آرشیو وبی در سازمان اسناد و کتابخانه ملی وجود ندارد ترجمه استانداردی در مورد آن چه وجهی میتواند داشته باشد؟ سوالی بجاست اما مگر می شود تنها به این علت که نتوانسته ایم آرشیو وب را محقق کنیم از پرداختن به ابعاد مختلف تخصصی و فناورانه آن خودداری کنیم و به کنارش بگذاریم؟

مهمترین علت ترجمه این استاندارد دانش خوبی است که در مورد آرشیو وب به دست می دهد. ترجمهٔ حاضر از نسخه چاپ سال 2013 صورت پذیرفته و قطعا نکته سنجی و تیزبینی ویراستار توانمند سرکار خانم آرزو تجلی در زدودن اغلاط و اصلاح نواقص و نقایص نقشی انکار ناپذیر داشته است از ایشان سپاسگزارم و نیز از سرکار خانم فروزان رضایی نیا که زحمت

بازخوانی چندباره متن را بر خود هموار کردند از شورای محترم انتشارات سازمان اسناد کتابخانه ملی جناب آقای دکتر رضا خانی پور مدیر کل آموزش و پژوهش سازمان و جناب آقای یوسف امیری مدیر محترم انتشارات سازمان بینهایت سپاسگزارم

و زحمات همکاران خدوم انتشارات سازمان سرکار خانم ها طاهره معمار و شهره خوری و جناب آقایان پرویز ،بختیاری شهرام ،چوپان آصف ،آصفی مهدی ،ادیبی و علیرضا خورسندی را ارج می نهم

هیچ اثری و هیچ برگردانی بی نقص .نیست دیدگاههای خوانندگان را با دیده منت پذیرا

خواهم بود.

فرزانه شادان پور

1397

پیشگفتار

ایزو (1) فدراسیونی متشکل از سازمان های ملی استاندارد در سراسر جهان موسوم به سازمان های عضو (ایزو است استانداردهای بین المللی) را معمولاً کمیته های فنی ایزو تهیه میکنند هر سازمان عضو که به موضوعی علاقه مند ،باشد در صورتی که برای آن موضوع کمیته ای در ایزو تشکیل ،شود میتواند در آن کمیته حضور یابد سازمان های بین المللی دولتی و غیر دولتی که با ایزو ارتباط دارند نیز در امور مشارکت میکنند ،ایزو در همه موضوعات مربوط به استانداردسازی در حوزه الکتروتکنیک با «کمیسیون بین المللی الکتروتکنیک (آی ای سی) » (2) همکاری نزدیکی دارد .

رویه های معمول در تدوین این سند با رویه هایی که برای بازنگریهای بعدی در نظر گرفته شدهاند در « دستورالعمل ایزو I آی ای سی » (3) بخش اول شرح داده شده اند. ضوابط لازم برای تصویب سندهای ایزو باید به طور ویژه مورد ملاحظه قرار گیرند. پیش نویس این سند مطابق با قواعد هیئت تحریریه دستورالعمل ایزولای ای ،سی بخش دوم نگارش شده است

نگاه (:کنید www.iso.org/directives).

توجه کنید که ممکن است بعضی از عناصر این سند موضوع حقوق ثبت اختراع باشند مسئولیت شناسایی این حقوق بر عهده ایزو نیست جزئیات حقوق ثبت اختراع که در حین تدوین سند شناسایی ،شوند در مقدمه یا در فهرست اعلانات حقوق ثبت اختراع ذکر می شوند .

نگاه (:کنید www.iso.org/patents).

نام های تجاری اطلاعاتی هستند که برای استفاده راحت تر کاربران ذکر میشوند و نباید 2.

1- International Organization for Standardization

2- International Electrotechnical Commission (IEC)

3- ISO/IEC Directives, Part 1.

به منزله تاییدی برای آن ها تلقی شود.

برای توضیح واژه ها و اصطلاحات خاص که برای ارزیابی میزان انطباق به کار میروند و نیز اطلاعاتی درباره تبعیت ایزو از اصول سازمان تجارت جهانی (1) در موانع فنی تجارت (2) به این منبع مراجعه کنید

Foreword – Supplementary Information

این سند را کمیته فنی شماره 46 ایزو برای اطلاع رسانی و ،دبیزش کمیته فرعی ، برای

ارزیابی آمارها و عملکرد (3) تهیه کرده است. 1.

1- World Trade Organization

2- Technical Barriers to Trade

3- ISI/TC 46, Information and Documentation, SC 8, Quality- Statistics and Performan ce Evaluatio

مقدمه

این گزارش در پاسخ به درخواست های جهانی برای دریافت رهنمودهایی درباره مدیریت و ارزیابی فعالیت ها و محصولات آرشیو وب تدوین شده است. آرشیو کردن وب یعنی انتخاب، گیر انداختن (1) ، ذخیره (2) ، حفاظت، و مدیریت دسترسی به رونوشت هایی از منابع اینترنتی در طول .زمان این رویداد در اواخر دهه 90 آغاز شد و مبتنی بر این بینش بود که در آینده آرشیو منابع اینترنتی پیشینهای حیاتی برای پژوهش و تجارت و دولت خواهد بود منابع ،اینترنتی بخشی از میراث فرهنگی به حساب میآیند و بنابراین همانند میراث انتشاراتی چاپی حفاظت میشوند موسسات بسیاری دست اندرکار آرشیو وب شدهاند و آن را بخشی از وظیفه بلندمدت خود در حفظ میراث ملیشان می انگارند

این فعالیت در بسیاری از کشورها با چارچوب های ،قانونی نظیر واسپاری ،آثار تأیید شده و به کار افتاده است طیف گسترده ای از منابع در اینترنت وجود دارد؛ مانند ،متن ،تصویر فیلم ،صوت و سایر قالب های چندرسانه ای در اینترنت علاوه بر صفحاتی که با ابر پیوند به یکدیگر متصل می شوند خبرنامه و گروه خبری و وبلاگ و خدمات تعاملی مانند بازی و پروتکل های مورد استفاده در انتقال و ارتباطات در دسترس قرار دارند در آرشیوهای ،وب رونوشت هایی از منابع موجود در اینترنت گرد هم آمده اند که با نرم افزار برداشت (3) در فواصل زمانی منظم گردآوری شده.اند هدف این است که منابع حتی الامکان به گونه ای که در محیط اصلیشان بودهاند با حفظ روابط ذاتی،شان مثالا با استفاده از پیوندهای

1- Capturing

2- Preserving

3- Harvest

موجود در ،ابرپیوندها نشان داده شوند.

هدف اولیه از آرشیو کردن ،وب حفظ پیشینه ای ماندگار و تا حد امکان نزدیک به شکل ،اولیه از ،وب برای اهداف دانشگاهی و حرفه ای و خصوصی است.

آرشیو کردن وب فعالیتی نوظهور ولی در حال گسترش است که مداوماً رویکردها ابزارهای جدیدی را برای هماهنگ و همگام ماندن با تغییرات سریع فناورانه در وب طلب می کند برحسب اهمیت استراتژیک آرشیو وب برای موسسه آرشیوگر و بسته به روش های در دسترس و گاه نیز بنا بر الزامات ،قانونی رویکردهای مختلفی برای آرشیو کردن منابع اینترنتی در پیش گرفته شده است از رونوشت برداری از صفحات وب شخصی گرفته تا همه وبگاه های تحت دامنه های سطح بالا . (1) از منظر سازمانی نیز آرشیو وب در سطوح مختلف پختگی قرار دارد در بعضی از سازمان ها به صورت فعالیت تجاری معمولی درآمده است و سازمان های دیگری به تازگی برنامه های آزمایشی برای بررسی آن را آغاز کرده.اند برحسب دامنه

و هدف ایجاد مجموعه نیز دو استراتژی (راهبرد اصلی) در آرشیو وب دیده میشود برداشت پشته ای (2) و برداشت انتخابی (3) در برداشت پشت های در مقیاس بزرگ مانند برداشت وبگاه های تحت نام دامنه ملی همه دامنه یا زیر مجموعه ای از آن رونوشت برداری می شود برداشت انتخابی . در مقیاس بسیار کوچکتری صورت میگیرد متمرکز است و مکرراً بر اساس معیارهایی مانند موضوع رخداد قالب شنیداری فایل های ویدئویی و مانند (آن یا توافق با صاحبان) محتوا در وب انجام می شود تفاوت اصلی این دو استراتژی در دو موضوع نهفته است؛ یکی سطح کنترل ،کیفیت و دیگری ارزیابی وبگاه های گردآوری شده با این رویکرد که آیا استانداردهای از پیش تعریف شده رعایت شدهاند یا خیر برداشت در وسعت نام دامنه مقایسه دستی و چشمی میان منابع برداشت شده و منابع موجود در وضعیت زنده وب (4) را ناممکن میسازد که البته روشی معمول برای اطمینان از کیفیت در شیوه برداشت انتخابی است.

این گزارش فنی نشان می دهد که چگونه میتوانیم آرشیوهای وب را به عنوان بخشی از مجموعه ،میراث مشابه و مطابق با شیوه ای مبتنی بر گردش کار سنتی کتابخانه اندازه گیری و مدیریت .کنیم گزارش حاضر، اطلاعاتی درباره گسترش مجموعه و سرشت نمایی (5) توصیف

1- Top-Level Domain

2- Selective harvest

3- Bulk harvest

4- Live Web

5- Characterization

حفاظت اسـتناد و ساختار سازمانی ارائه میکند و نشان می دهد که در اصل بیشتر وجوه گردش کار سنتی در مدیریت مجموعه برای آرشیو وب نیز معتبر هستند که البته در عمل تطبیق و تنظیم آن ها ضروری است .

این گزارش فنی مروری اجمـالی بر وضـعیت فعلی آرشـیو وب دارد و بحث اصـلی آن معطوف به تعریف و کـاربرد آمارهـا و شاخصـهای کیفیت در آرشـیو وب است تولیـد بعضـی از آمارهـا وابسـته به نرم افزار برداشت نمایه سازی و مرور است و ممکن است انتخاب نرم افزار متفـاوت منجر به اختلاـف نتایـج شود مع الوصف این گزارش فنی هیـچ نرم افزار خاصـی را تاییـد یا توصـیه نمیکنـد؛ بلکه مجموعه ای از شاخصـها را در اختیار میگذارد که میتوانند در اندازه گیری عملکرد و کیفیت عموم آرشـیوهای وب یاری گر باشند این گزارش فنی اثری در حال پیشرفت است و احتمال دارد در آینده بعضی از قسمت های محتوای آن در ایزو 2789 و ایزو 11620 قرار گیرند.

ص: 13

اطلاعات و دبیزش آمارها و شاخص های کیفیت در آرشیو وب

اشاره

ص: 15

1- دامنه

در این گزارش فنی ،آمارهـا اصـطلاحات و معیارهای کیفیت در آرشـیو وب تعریف می شونـد. در گزارش نیازها و عملکردهای جاری در طیف گسترده ای از سازمان ها ماننـد کتابخانه ها ،آرشـیوها موزه ها مراکز پژوهشـی و موسـسات میراثی در نظر گرفته شده اند. نمونه های ذکر شده از میان کتابخانه های بزرگ برگزیده شدهاند؛ چراکه کتابخانه ها به ویژه کتابخانه های ملی کار آرشیو کردن وب را در چارچوب و اسـپاری قانونی آثار بر عهـده گرفته انـد؛ البته بـدان معنی نیست که اهمیت موسـساتی غیر از کتابخانه ها اندک انگاشـته شود نیز از قابلیت کاربست این گزارش فنی در موسسات میراثی و نزد حرفه مندان آرشیو نمی کاهد .

این گزارش فنی برای حرفه مندانی تدارک دیده شده است که بی واسـطه و اغلب در گروه های متشـکل از متصدیان کتابخانه یا آرشیو دست اندرکار آرشـیو کردن وب هسـتند. همچنین برای سرمایه گذاران موسسات آرشیو وب و ذینفعان بیرونی این موسسات مفید است تلاش بر این که اصطلاحات به کار رفته در این گزارش در عین رعایت توازن میان علوم کامپیوتر و

بوده .

مـدیریت و ،کتابـداری بازتابی از طیف گسترده علایق و تخصـص های مخاطبان آن باشـد این گزارش فنی برای مدیریت منابع الکترونیکی دانشگاهی و تجاری مانند مجله و کتاب و روزنامه های الکترونیکی که معمولا جداگانه و با استفاده از سامانه های مدیریتی متفاوتی ذخیره و پردازش میشونـد در نظر گرفته نشـده است. این هـا منـابع اینترنتی محسوب می شونـد، ولی در این گزارش به عنوان یـک محتوای برخط متمایز با سایر منابع اینترنتی لحاظ نمی شوند بعضـی از سازمان ها منابع الکترونیکی خاصـی را گردآوری میکنند که ممکن است از طریق سامانه های واسپاری و انباره های الکترونیکی ناشران در وب به دست .آیند این منابع نیز از دامنه پوشش

این گزارش خارج انـد . اصول و فنون به کار رفته در این نوع گردآوری به طور مسـلم با آرشـیو وب بسـیار متفاوت است؛ آمارها و شاخص های کیفیت مربوط به یک روش لزوماً برای روش دیگر مناسب نیسـتند سـرانجام این گزارش فنی اساساً معطوف به اصول و روشهای آرشیو وب است و شامـل راه هـای جایگزین در گردآوری منابع اینترنتی نمیشود در واقع، بعضـی از منابع اینترنتی، به ویژه آنها که در وب منتشـر نشـده انـد مانند خبرنامه هایی که به صورت نامه های الکترونیکی توزیع می شوند با فناوری های آرشـیو وب برداشت نمیشوند و گردآوری آن ها به روش های دیگری صورت میگیرد که در این گزارش فنی تشریح و تحلیل نمی شوند.

2- اصطلاحات و تعاریف

اصطلاحات و تعاریف زیر در محتوای این سند به کار برده می شوند :

1.2

دسترسی

پرس و جوی موفق خدمت برخطی که توسط کتابخانه ارائه می شود

یادداشت : دسترسـی حلقه ای از فعالیت های کاربر است که نوعاً با اتصال کاربر به یک خدمت برخط که توسط کتابخانه ارائه میشود آغاز شده و با یک فعالیت پایانی عامدانه (خروج ارادی کاربر از پایگاه داده یا از) (سامانه یا غیر عامدانه اتمام) وقت به علت غیر فعال بودن

(کاربر) به پایان میرسد .

یادداشت :2 دسترسی به وبگاه کتابخانه بازدید مجازی (1) محسوب می شود .

یادداشت : پرس و جوهای ارسالی به یک ورودی عمومی یا وبگاه در این تعریف قرار

نمی گیرند .

یادداشت :4 پرس و جوهایی که از طریق موتورهای جست وجو انجام میشوند در صورت

-امکان از این محدوده خارج اند .

[منبع ایزو :2798 2013، تعریف 2. 2. 1].

2.2

ابزار دسترسی (2)

Virtual visit –1

Access tool –2

نرم افزار (1) خبره برای یـافتن و بازیابی و بازپخش منابع اینترنتی آرشـیو شـده یادداشت : ممکن است به وسـیله تعـدادی بسـته نرم افزاری جداگانه که باهم کار می کنند .

پیاده شود.

3.2

فراداده مدیریتی (2)

اطلاعات لازم برای مدیریت مطلوب اشیای دیجیتالی در یک انباره

یادداشت :1 فراداده مدیریتی به سه دسته زیر تقسیم می شود :

–فراداده منشا (3) یا زمینه (4) چرخه حیات یک منبع را تا نقطه ای توصیف میکند از جمله

هستارها (5) و پردازش های مرتبط مانند پیکربندی (6) و کارنامه ورود به سامانه؛ (7)

– فراداده فنی (8) ویژگی های فنی یک شی دیجیتالی را توصیف میکند مثلاً قالب آن را؛

–فراداده حقوق (9) مالکیت و مجوز قانونی استفاده از یک شیء را تعیین میکند .

4.2

آرشیو

آرشیو وب

دسته ای از منابع به طور کامل که در طول زمان با خزش از وب به دست ،آید شامل یک

یا چند مجموعه

5.2

داده بیتی (10)

1- Specialist software

2- Administrative metadata

3- Provenance metadata

ردیفی از دیجیت های 0 و 1 که یک فایل دیجیتالی را تشکیل می دهند .

6.2

توان (خزش)

حد و حدود خزش یا تک هسته ها (1) که می تواند تعدادی فایل داده یا زمان مصرف شده در هر خزش باشد که در تنظیمات خزشگر تعیین شدهاند.

7.2

خزش پشته ای (2)

برداشت پشته ای

خزشی که با هدف گردآوری کل یک یا چند دامنه سطح بالا یا زیر مجموعه انجام میشود.

یادداشت 1: خزش های پشته ای در مقایسه با خزش های انتخابی دامنه گسترده تری دارند

و نوعا بسامد تکرار آنها کمتر است.

یادداشت 2: خزش های پشت های معمولاً آرشیوهای وب بزرگ مقیاس را ایجاد می کنند کنترل کیفیت دقیق را غیر ممکن می سازند کنترل کیفیت معمولاً با نمونه گیری انجام می.شود

8.2

گیر انداختن

نمونه

رونوشتی از یک منبع که در نقطه ای از زمان خزش شده است.

یادداشت : اگر یک منبع سه بار در زمانهای مختلف با خزش به دست آمده باشد، سه

وب گرفت (3) از آن موجود خواهد بود.

9.2

مجموعه (4)

1- Seeds

2- Bulk crawl

3- Capture

4- Collection

مجموعه آرشیو وب

منابع منسجمی که در قالب یک گروه ارائه شده اند.

یادداشت : هر مجموعه ممکن است به گونه ای خاص پیش از اقـدام به برداشت انتخاب شـده باشد مثلا (یک رخداد یا یک موضوع) یا به صورت گذشته نگر (1) از منابع در دسترس و موجود در آرشیو گرد هم آمده باشد.

یادداشت :2 هر آرشیو وب ممکن است از یک یا چند مجموعه تشکیل شده باشد.

10.2

خزش

برداشت

فرایند مرور (2) و رونوشت گیری (3) از منابع با استفاده از خزشگر

یادداشت : خزش ها را میتوان به دو دسته خزش پشته ای و خزش انتخابی تقسیم کرد.

11.2

تنظیمات خزش پارامترهای خرش

تعیین اینکه کدام منابع باید گردآوری شوند و بسامد و عمق مجموعه هسته ها چقدر باشد.

یادداشت 1 :تنظیمات خزش همچنین شامل ادب (4) خزشـگر تعداد پرس وجو در ثانیه یا دقیقه که به سرور میزبان منبع موردنظر ارسال می شود هماهنگی با استاندارد منع خزش (5) و

پالایه (6) ها برای مصونیت از تله های خزش (7) است.

12.2

خزشگر

Retrospective -1

Browse -2

Copying -3

برداشتگر

خزشگر آرشیوسازی

اصطلاح منسوخ : اسپایدر

نرم افزاری که نشانی های وب (1) را با موفقیت پرس وجو می کند و منبعی را که به دست می آید برای یافتن نشانی های وب دیگر که در آن موجودند تجزیه (2) می کند .

یـادداشت : ممکن است منابع ذخیره شونـد و نشانی های وب بر طبق مجموعه قواعـد از پیش تعیین شـده از نوبت کار خزش خارج شوند نگاه کنید به تنظیمات خزشگر (11.2) و دامنه

(خزش) (42)].

13.2

تله خزش

صفحات (وب یا) دسته ای از (آن ها) که یا باعث میشوند خزشگر از کار بیفتد یا آن را به منابع

بی پایانی ارجاع میدهند که کم ارزش یا فاقد ارزش اند

یادداشت 1 : تله های خزش ممکن است عامدانه کار گذاشته شوند تا مانع خزشگرها

در برداشت منابع اطلاعاتی شوند تله خزش ممکن است به صورت ناخواسته رخ دهد مثلاً

هنگامی که خزشگر تاریخ های یک تقویم بی پایان را تعقیب می کند .

2. 14

نرم افزار آرشیوگر (3)

نرم افزاری کاربردی که پیشتر از خزشگر به کار می افتد و فرایند برداشت اطلاعات را پشتیبانی میکند

یادداشت 1 : یک کار کرد ،اصلی مدیریت منابع هدف (4) و فراداده های مدیریتی و توصیفی

همراه آنهاست که ممکن است شامل بخش هایی برای زمان بندی (5) و کنترل کیفیت باشد.

2. 15

1- URLs: Uniform Resource Locators

2- Parse

3- Curator tool

4- Targets

5- Scheduling

داده کاوی (1)

پردازش محاسباتی که با تجزیه و تحلیل کمی دادهها از چشم اندازها و ابعاد متفاوت با طبقه بندی آن ها و تلخیص روابط و تاثیر بالقوه آن ها الگوهایی را انتزاع می کند [منبع ایزو :16439 تعریف 13.3]

2, 16

وب عمیق (2)

اصطلاح منسوخ وب پنهان (3)

اصطلاح منسوخ وب نامرئی (4)

بخشی از وب که موتورهای جست وجو نمیتوانند آن را خزش و نمایه سازی کنند و به ویژه

از منابعی تشکیل شده که به صورت پویا ایجاد میشوند یا با گذرواژه (5) محافظت میشوند.

17.2

فراداده توصیفی

اطلاعاتی که محتوای فکری یک شیء دیجیتالی را توصیف می کند.

18.2

نام دامنه (6)

رشته کاراکترهای تعیین هویت که نشان دهنده قواعد و رویه های DNS (سامانه نام دامنه) (7) در مورد قلمرو خود مختار اداری حاکمیتی یا نظارتی بر اینترنت است.

19. 2. سامانه نام دامنه

DNS

1- Data mining

2- Deep Web

3- Hidden Web

سامانه نام گذاری (1) توزیع شده و سلسله مراتبی جهانی برای شناسایی هستارهایی که به اینترنت

متصل اند.

یادداشت 1: دامنه های سطح بالا در بالاترین جایگاه در سلسله مراتب قرار دارند .

2.2

همگون سازی (2)

خلق مجدد کارکرد و رفتار یک سامانه خارج از رده با استفاده از یک نرم افزار که همگون ساز نامیده می شود بر رایانه های موجود

یادداشت : همگون سازی یک استراتژی کلیدی در حفاظت است

21.2

میزبان (3)

قطعه ای از یک شناسگر منبع و بی (4) که منبع محتوا را در شبکه نامگذاری می کند

یادداشت1 : میزبان نوعا نام دامنه ای مانند WWW.ARCHIVE.ORG یا زیر دامنه ای

مانند ORG.WEBARCHIVE است.

22.2

HTML

زبان نشانه گذاری ابرمتن (5)

مهمترین زبان نشانه گذاری صفحات ،وب از عناصری تشکیل شده است که برای افزودن اطلاعات ساختاری و معنایی به متن خام به کار می رود

23.2

HTTP

1- Naming

پروتکل انتقال ابرمتن (1)

پروتکل ارتباط میان کاربر خدمتگر که برای انتقال اطلاعات در وب به کار می رود.

24.2

ابر پیوند (2)

پیوند

ساختاری رابطه ای که برای پیوند دادن اطلاعات در اینترنت به کار می رود.

25.2

دورریز (3)

هرزنامه (4)

محتواهای ناخواسته ای که نامربوط یا فاقد ارزش بلندمدت تلقی میشوند. یادداشت : هرزنامه های تعمدی معمولاً برای دستکاری نمایه موتورهای جست وجو به کار میروند. دورریزها نیز به گونه ای ناخواسته هنگامی پدید میآیند که خزشگر در تله خزش

گرفتار میشود یادداشت : موسسات گردآورنده محتوای وب عموماً میکوشند تا از گردآوری دورریزها و هرزنامه ها پیشگیری کنند تا امکانات برای برداشت منابع «خوب» به کار روند. مع الوصف، بعضی از این موسسات نمونه کوچکی از این گونه منابع را به عنوان بخشی از پیشینه وب نگه میدارند

26.2

پیوند کاوی (5)

پردازش و تجزیه و تحلیلی که معطوف به انتزاع الگوها و اکتشاف از ابرپیوندهاست مانند ترسیم

گراف شبکه

27.2

1- Hypertext Transfer Protocol

2- Hyperlink

تراوایی به وب زنده (1)

مشکل شایعی در ارائه منابع آرشیو است و هنگامی رخ می دهد که پیوندها در یک منبع آرشیو شده به جای باز شدن به رونوشت موجود منبع مورد پیوند در انباره همان آرشیو به منبع اولیه در وب زنده باز می شوند.

یادداشت :1 تراوایی به وب زنده هنگامی رخ می دهد که اسکریپت های صفحات وب آرشیو شده در زمان ارائه محتوای آرشیو به ارجاع به منابع وب زنده ادامه می دهند و پرس و جوهای موقتی نیز ارسال می.کنند این امر باعث میشود منابعی مانند فیدهای وب یا ویدئوهای رسانه های اجتماعی (2) از وب زنده در صفحه آرشیو شده ظاهر شوند.

28.2

کارنامه ورود به سامانه

فایلی است که خدمتگر به طور خودکار از فعالیت های خود نگه میدارد .

29.2

فراداده

داده ای که بافتار (3) محتوا و ساختار اشیاء دیجیتالی و مدیریت آنها را در طول زمان توصیف میکند

منبع: ایزو :15498: 2001، تعریف 2 12

یادداشت : فراداده را میتوان به سه دسته ،توصیفی ،ساختاری و اداری تقسیم کرد

30.2

مهاجرت (4)

تبدیل (5) قالب های فایل (6) خارج از رده یا قدیمی تر به قالب های جدیدتر یا زنده به منظور حفظ

دسترس پذیری شیء دیجیتالی

یادداشت : مهاجرت یک استراتژی کلیدی حفاظت است.

1- Live Web leakage

2- Social media

[منبع ایزو :14489 2001، تعریف 3 13]

31.2

نوع فایل (1)

نوع رسانه اینترنت

نوع محتوا

شناسگر دو بخشی برای انواع فایل در اینترنت

یادداشت :1 نوع فایل سرایند نوع محتوا را به کار میبرد که از نوع و زیر- نوع (2) تشکیل شده

و قالب یک منبع را نشان میدهد مثل IMAGE/JPG

32.2

نامزدی (3)

منبعی که نامزد ورود به آرشیو وب است

33.2

صفحه (4)

صفحه وب

منبع ساختاریافته ای که علاوه بر محتوای قابل خواندن برای انسان (5) با یک نشانی وب شناسایی میشود و ممکن است با چند منبع دیگر ارتباط داشته باشد یا با هیچ منبع دیگری در ارتباط نباشد.

34.2

مجوز

اجازه خزش یک وبگاه زنده و / یا اجازه نمایش محتوای آن در یک آرشیو وب

یادداشت :1 مجوز ممکن است با یک گواهی رسمی از صاحب حقوق اعلام یا به موجب

(MIME Type (MIME: Multipurpose Internet Mail Extension –1

Sub–type –2

Nomination –3

Page –4

Human–readable –5

واسپاری قانونی لزوم اخذ آن منتفی شود.

35.2

کاربر ثبت نام شده

شخص یا سازمانی که در یک کتابخانه ثبت نام کرده تا از مجموعه و / یا خدمات آن در فضای کتابخانه یا خارج از آن استفاده کند.

یادداشت : ممکن است کاربران خودشان درخواست عضویت بدهند یا هنگام ثبت نام در

موسسه به طور خودکار در آرشیو نیز عضو شوند.

یادداشت :2 عضویت در فواصل ،منظم دست کم هر سه سال یکبار پایش و کاربران

غیر فعال حذف میشوند [منبع ایزو :3789 2013 ، تعریف 2 282]

36.2

پرس و جو

پیامی در قالب HTTP که یک سامانه (مانند مرورگر یا یک خزشگر) برای پرس و جوی منبع خاصی که با نشانی وب شناسایی میشود به یک خدمتگر دور (1) میفرستد .

37.2

پاسخ

پاسخ یک خدمتگر دور به یک پرس وجو با پروتکل HTTP که این پاسخ یا منبع درخواستی را با خود دارد یا به نشانی دیگری در وب هدایت می کند یا پاسخی منفی (خطا (2)) با خود دارد که ذکر میکند چرا منبع درخواست شده قابل ارائه نیست

38.2

که پاسخ (3)

کد وضعیت (4)

Remote server -1

Error -2

Response code –3

Status code –4

شمارهای سه رقمی که وضعیت منبع درخواستی را به خدمتگری که پرس و جویی را ارسال کرده

نشان می دهد

یادداشت : کدی که مثلاً با عدد 4 (4xx) آغاز میشود نشان میدهد که منبع مورد

درخواست در دسترس نیست

39.2

txt.robots

استاندارد منع خزش (1)

پروتکلی که برای پیشگیری از دسترسی خزشگرها به همه یا بخشی از یک وبگاه به کار میرود

یادداشت : txt.robots هنوز اجباری نشده است.

یادداشت : این استاندارد همچنین ممکن است برای به حداقل رساندن زمان تاخیر میان درخواسته ای پی در پی یا حتی برای فراهم کردن پیوند به نقشه سایت (2) به کار رود تا خزش

بهتر سایت میسر شود .

40.2

دامنه (خزش)

مجموعه پارامترهایی که گسترهٔ خزش را تعیین میکنند مانند حداکثر تعداد پرش (3) یا حداکثر عمق مسیری که خزشگر باید تعقیب کند

یادداشت : دامنه خزش میتواند همه یک نام دامنه سطح بالا باشد (مثلاً de) یا به یک

فایل محدود شود.

41.2

دامنه (آرشیو) (وب)

گستره آرشیو یا مجموعه ای آرشیو شده از وب آن طور که اختیارات قانونی مؤسسه یا سیاست

مجموعه سازی تعیین میکند

1- Robots exclusion standard

2- Site map

3- Hop

42.2

دامنه سطح دوم (1)

زیر بخش های داخل دامنه سطح بالا برای سطوح ویژه ای از سازمانها یا حوزههای مورد علاقه (مانند .ir.gov برای وبگاه های دولتی ir.asso برای وبگاه های انجمن ها.)

43.2

هسته

نشانی های وب هدف

نشانی وب مربوط به مکان منبع خاصی که باید خزش شود و خزشگر آن را نقطه آغاز کار

خود قرار دهد.

44.2

انتخاب

فرایند تصمیم گیری تصدی گرانه (2) که در سنجش با سیاست گسترش مجموعه تعیین می کند آیا مجموعه منابع ارزشمندی در دامنه آرشیو وب قرار میگیرد یا خیر.

45.2

خزش گزیده (3)

برداشت گزیده؟ (4)

خزش با هدف گردآوری منابع برگزیده بر اساس چند معیار

یادداشت : خزش گزیده در مقایسه با خزش پشتهای دامنهٔ محدودتری دارد و نوعاً با

بسامد بیشتری انجام می شود .

یادداشت : خزش های گزیده ،مستمر خزش هایی هستند که با هدف گردآوری منابعی

Second Level Domain –1

Curatorial –2

Selective crawl –3

Selective harvest –4

صورت می گیرند که بر اساس چندین معیار برگزیده شده اند؛ مانند اهمیت علمی ربط به یک موضوع یا بسامد روزآمدسازی مداوم منبع

یادداشت : خزش های گزیـده ،مناسـبتی خزش در زمانی محـدود است که در تاریخی خاص پایان می یابـد و هدف از آن گردآوری منابع مرتبط با رخدادی منحصربه فرد نظیر انتخابات رخدادهای ورزشی و حوادث است.

46.2

فراداده ساختاری (1)

اطلاعاتی که شرح میدهد چگونه اشیاء مرکب با یکدیگر آمیخته شده اند تا واحدهای منطقی

را بیارایند

47.2

هدف (2)

مجموعه ارزشمندی از منابع که با تعیین یک یا چند هسته و تنظیمات مربوط به آن باید

گردآوری شوند.

48.2

دامنه سطح بالا

بالاترین سـطح دامنه ها در سامانه نام ،دامنه شامل دامنه های سطح بالای کد کشور (مانه) (fr) که طبق کدهای دو حرفی نام کشورها در کوته نوشتهای اسامی کشورها، مندرج fir, d در ایزو 3166 تعیین شدهاند و دامنه های سطح بالای عام اند (.com, .net, .org, .paris .)

یادداشت :1 به جز موارد خاص این اصطلاح در این گزارش به معنای نام دامنه های سطح

بالای کد کشور به کار میرود .

49.2

شناسگر منبع وبی

URI

رشته کاراکترهای گسترش پذیر (3) مورد استفاده برای شناسایی یا نامگذاری یک منبع در اینترنت

1- Structural metadata

2- Target

3- Extensible

50.2

نشانی وب

URL

زیر مجموعه ای از شناسگر منبع وبی (URI) که مکان یک منبع و پروتکل بازیابی آن را

مشخص میکند .

51.2

قالب WARC و (1)

قالب فایل که روش ترکیب منابع مختلف دیجیتالی را برای ایجاد یک فایل آرشیوی مجتمع (2) همراه با اطلاعات مربوط معین می کند .

یادداشت : قالب WARC از سال 2009 استاندارد ایزو ایزو 28500 (2009) شناخته شده است.

52.2

وبگاه

صفحات

وبی

که از نظر قانونی و / یا انتشاراتی به هم متصل اند

یادداشت : معمولا وبگاهها نماینده مؤسسات ،رسمی سازمان ها ،شرکت های خصوصی و

صفحات خانگی هستند

53.2

وب

برنامه کاربردی اصلی نشر در اینترنت که با سه استاندارد کلیدی شناسگر منبع ،وبی HTTP و HTML فعال شده است.

3.روشها و اهداف آرشیو کردن وب

اشاره

شکل و محتوای آرشیوهای وب آن گونه که امکانات فنی آن را سیاست های موسسه متولی

تعیین می.کند سیاست ها در سطوح بالا را قانون گذاری ملی تنظیم میکند اما موسسات

Web ARChive -1

Aggregate archival file -2

طیف متنوعی از استراتژی های گردآوری را به کار می گیرند که از اهداف تجاری و معیارهای انتخاب برخاسته اند ،باوجود این منابعی که در دامنه گردآوری قرار می گیرند گاهی به علت محدودیت های فنی قابل افزوده شدن به آرشیو نیستند برای نمونه گیر انداختن و بازپخش منابع چندرسانه ای و منابع تعاملی (1) چالش بزرگی فراروی جامعه آرشیو وب بوده و اغلب مستلزم راه حل های (2) گران قیمت و بومی شده هستند.

1.3. روش های گردآوری

1.1.3

پایه فنی

رونوشت برداری از منابع برخط یا برداشت آن ها روش اصلی در گردآوری منابع اینترنتی است برای برداشت باید از خزشگر استفاده کرد که پیاپی منابع را با ارسال نشانی وب آن ها پرس وجو می،کند از آن ها رونوشت میگیرد ذخیرهشان می کند و برای یافتن نشانی های ابر پیوند شده موجود در محتوای منابع برای خزش های بعدی آن ها را تجزیه می.کند نقطه آغاز به کار یک خزشگر – که اغلب صفحه خانگی یک وبگاه –است هسته نامیده می شود .

خزشگر همانند یک کاربر خودکار در وب عمل میکند و میتواند به گونه ای پایان ناپذیر منابع اینترنتی را که به هم پیوند دارند تعقیب کند؛ مگر اینکه دامنه خزش منابع با پارامترها یا تنظیمات خاصی برای آن تعیین یا محدود شده .باشد ،خزشگر همچنین ممکن است هنگام مواجهه با موانع در طول فرایند برداشت اطلاعات ناخواسته دچار توقف شود ،پوشش عمق و کیفیت کلی یک مجموعه آرشیو وب در ارتباطی تنگاتنگ با مجموعه ای از تنظیمات فنی تعریف میشوند این تنظیمات فنی با عنوان قواعد برداشت بیان شده اند منابعی که باید گردآوری شوند با مکان نشانی وب در شکل یک فهرست هسته و نیز به وسیله دامنه توصیف می شوند .

دامنه عمدتا با بسامد عمق برداشت تعریف میشود که بر جامعیت (3) یک آرشیو تاثیر میگذارد

1- Interactive resources

2- Solutions

3- Comprehensiveness

2.1.3

محدودیت ها

مقـادیر عظیم اطلاعات با سـرعتی شـگفت آور به وب افزوده می.شود مؤسـسات معمولاً با در نظر گرفتن منابع در ارتباط با ،کارکنان توان محاسـبات و ظرفیت ذخیره سـازی در مورد دامنه آرشـیو وب تصـمیمگیری میکننـد علاـوه بر این انتخـاب هایی که حین گزینش و فرایند برداشت صورت می گیرد به موسـسات امکان می دهد بر منابع خوب و ارزشمند متمرکز باشند؛ برعکس محتواهایی که به صورت خودکار تولیـد شدهانـد و ارزش انـدکی دارند نظیر دورریزها و هرزنامه ها ،این مهمترین منبع اختلال (1) در آرشـیو وب است که باید از آن حذر کرد؛ و لازمه حذر کردن از آن مدیریت فعال و اولویت بندی فرایند خزش است.

محـدودیت هایی وجود دارنـد که گردآوری جامع منابع اینترنتی را دچار چالش میکنند بعضـی از آن ها مربوط به فناوری،اند بعضـی دیگر معلول مقیاس و طبیعت اینترنت.اند محدودیت های دیگری نیز ممکن است در اثر قوانین تحمیل شوند.

(الف) مسائل ناشی از معماری فعلی و فناوری زنده خزش

خزشگرهای آرشیوگر قادرند محتوای ایستا (2) را که با پرس و جوی یک نشانی وب در اختیار

قرار می گیرد به مقـداری که باید گیر بیندازند وقتی نشانی های وب به طور مسـتقل درون حد HTML با فناوری ارجاع ابر متن (3) ارجاع دار (4) نشده اند بلکه در فایل هایی مانند جاوا اسـکریپت یا فلش جاسازی شده یا به طور پویا (5) بر پایه تعامل با کاربران تولید شدهاند؛ خزشگرهای آرشیوگر اغلب در گیر انداختن محتوای ارجاعی شکست می خورند

استخراج (6) و تجزیه نشانی های وب که مطابق قواعد تولید شده اند – همیشه کار آسانی نیست چراکه نحو (7) نشانی وب ممکن است به گونه ای مطول شده باشد که تقریباً همه نوع منبع در شبکه را نشانی دهی کند و ممکن است نشانی وب به طور پویا تولید شده باشد.

ساختـار فوق العـاده پیچیـده نشـانی هـای وب شامل متغیرهای متعـددی است که با واژه نگار یا همان آمپرسـند علائمی ماننـد و / و « »، علامت

،مساوی شناسه های ،کاربری شناسه های ،نشست و نیز کدهای ردیابی ارجاع (8) نشانه گذاری

1- Noise

2- Static

3- Hreference (Hypertext reference)

4- Referenced

5- Dynamically

6- Extract

می شونـد در مواردی فایل های چندرسانه ای با کمک برنامه های کاربردی تحت وب و جاسازی شده باز میشوند یا در اختیار قرار میگیرند این برنامه ها داده ها را بدون اینکه در تگ های HTML قرار گرفته باشند از خدمتگر اصلی بازیابی می کنند .

فنـاوری فعلی برداشت اطلاعـات هنوز راه زیادی در پیش دارد تا برای کار با همه وب تناسب پیـدا کنـد شـیوه فعلی در ،خزش یعنی تکرار خزش مبتنی بر نشـانی وب بـه گردآوری مقـادیر روبه افزایش ،وب از جمله محتـوای پشت فرم هـای وب و رابـط هـای کـاربری پرس وجو معروف به وب پنهـان یـا ،عمیق رسـانه های در ،جریان محتوای تحویلی با پروتکل هایی غیر از HTTP، و رسانه های اجتماعی قـد نمی دهد ،باوجوداین مهمترین چالش برای دست اندرکاران آرشـیو وب سـرعت تغییر در وب با قالب ها و پروتکل ها و سـکوهای (1) جدید است که سازمان های آرشـیوگر را ملزم می کنـد هنگامی که نیاز به توسـعه مداوم و بهبود توان آرشـیو کردن انواع محتوا پدید می،آید به آن پاسخ دهند

ب) مسائل ناشی از بسامد روزآمد کردن منابع وب

مسئله فنی دیگری که عموماً در مورد خزش ذکر میشود فقدان انسجام زمانی است.

روزآمـدسازی یـک وبگاه هنگام ،خزش ممکن است منجر به حصول تصویرهای لحظه ای (2) به هم ریخته ای شود که صـفحات وب با طول عمر (3) متفاوتی در آن حضور دارند. قانون گذاری نیز میتواند موجب بروز محدودیت هایی در ایجاد آرشیو وب شود.

یکی از تصمیمات مهم که از قوانین تاثیر میپذیرد این است که آیا استاندارد منع خزش باید رعایت شود .

یـا خیر این ،مسـئله در گیر انـداختن یا چشم پوشـی از بعضـی محتواها تفاوت مهمی ایجاد می کنـد محـدودیت هایی که ذکر شـد فرایند ارزیابی فعالیت های آرشـیو کردن وب با سنجه های (4) قابل مقایسه را دسـتخوش چالش می.کند رویکردی که در این گزارش عموماً در پیش گرفته شده است اذعان به محدودیت ها و تمرکز بر چیزهایی است که معلوم و قابل مقایسه اند.

3.1.3استراتژی های گردآوری

برای گردآوری دو استراتژی اصلی وجود دارد که هر کدام سطح خودکاری و دامنه آرشیو وبی

1- Platform

2- Snapshot

3- Life span

4- Measures

متفاوتی به دست می دهند.

برداشت ،پشته ای مانند برداشت دامنه ملی که هدف آن گیر انداختن تصویر لحظه ای از یک دامنه کامل یا یک زیر مجموعه مانند دامنه ملی در زمانی معین است که به پدید آمدن مجموعه های بزرگ مقیاس آرشیو وب منتهی می.شود شناخته شده ترین آرشیو با این استراتژی Way back machine متعلق به « اینترنت آرکایو » (1) است که با هدف حفاظت از همه وب ایجاد شده است برداشت پشت های فرایندی نسبتاً خودکار است که اندازه عملیات آن را محدود می.کند این نوع گردآوری با بسامد اندک اغلب یک بار در سال انجام می.شود تضمین کیفیت اگر انجام شود معمولاً بر کنترل خودکار محتوای ناقص از طریق وارسی کدهای (2) وضعیت HTTP ممکن است. آرشیو انتخابی در مقیاس کوچکتری انجام می،شود متمرکزتر است و بسامد بیشتری دارد. فرایند انتخاب برای شناسایی وبگاه های مرتبط بر اساس معیارهایی مانند موضوع رخداد قالب مثلاً فایل های ویدئویی یا (صوتی) یا توافق با صاحبان محتوا انجام می شود.

تضمین ،کیفیت، عنصر مشترکی در آرشیو انتخابی است که در حال حاضر عمدتاً بر مقایسه چشمی (3) بازبینی برداشت های قبلی و کارنامه خزش استوار .است در آرشیوسازی انتخابی گرایش بیشتری به برداشتن فراداده توصیفی وجود دارد معمولا متصدی آرشیو حین انتخاب یا بعد

از فرایند برداشت اطلاعات این فراداده ها را الصاق میکند که برای کارکردهای جست وجو و مرور با نتایج بهتر در رابط کاربری آرشیو وب مورد استفاده قرار میگیرد. بعضی از موسسات رویکردی مرکب از موارد بالا در پیش می گیرند. بعضی از وبگاه ها مرتباً روزآمد می شوند و تغییرات تنها با برداشت پشت های ندرتی یا برداشت دامنه گیر انداخته نمیشوند اینکه یک سازمان آرشیوگر خود به گونه ای استراتژی را تعریف کند که وبگاه های دارای اولویت بالاگیر انداخته شوند و وبگاه های کم اهمیت تر فقط با برداشت دامنه یا برداشت پشته ای گیر انداخته شوند مورد نادری نیست

4.1

معیارهای انتخاب

معیارهای انتخاب معمولاً با قوانین و استراتژی گسترش مجموعه موسسه و هماهنگ با

1- Internet Archive

2- HTTP status code

3- Visual comparison

ماموریت های اصـلی موسسه تنظیم می.شود ملاحظات عملی یا محدودیت هایی مانند استخدام کارکنان و امکانات و تخصص اغلب بر تحقق استراتژی تاثیر میگذارند معیارهای انتخاب دامنه آرشیو وب را تعیین میکنند و میتوانند به طرق گوناگونی بیان شوند

از طریق نام دامنه هایی که میزبانی منابع را بر عهـده دارنـد مانند نام دامنه ملی یا دامنه های سـطح بالا مثل .fr یا de ؛ یا دامنه های سطح دوم که به بعضی ناشران اختصاص دارد مثل .gov برای انتشارات دولتی ،باوجوداین نام دامنه ها نمیتوانند به طور قطعی موجب شـناسایی یا تعیین محتوای ملی ،باشـند چراکه اینترنت سامان های جهانی است و منابع در اقصـی نقاط فیزیکی و جغرافیایی پراکنده اند؛ با سرشت ،منابع مانند موضوع یا مضـمون محتوای وبگاه محبوبیت نزد مخاطبان یا زبان پروتکل های ارتباطی مورد استفاده برای تحویل منبع مانند HTTP یا قالب ها مانند متن

یا ویـدئو از طریق شـرایط دسترسـی یا وضـعیت حق مولف ماننـد اینکه به رایگان در دسترس قرار می گیرد یا با خریـداری یا پرداخت حق اشتراک؛ از طریق بودجه ای که سازمان می توانـد برای آرشـیو هزینه کنـد یک سازمان ممکن است فقط منابع مالی کافی برای پشتیبانی از برداشت اطلاعات در دفعات محدود را در اختیار داشته باشد

یا بتوانـد برای رویکرد برداشت نمونه ای بسـیار برگزیـده از محتوای وب هزینه کند. از طریق محدودیت ها یا اسـتثنائات ،محتوایی به عنوان مثـال خـارج کردن منـابع حـاوی مطـالب مشـخص داده های حساس یا غیرقانونی میتوانـد از معیارهای انتخاب باشـد. همیشه نمیتوان به روشنی دانست که آیا باید بعضی منابع را در فهرست انتخاب گنجاند یا از آن خارج کرد یک سازمان تصمیم می گیرد شبکه های اجتماعی و وبلاگ ها و سکوهای تعاملی (1) (2) مشابه را آرشیو ،کند حال آنکه سازمان دیگر ممکن است این دسته از محتوا را از حوزه کار آرشیو خود خارج کنـد این ها تصـمیمات مربوط به سـیاست گـذاری است که در مورد منابعی ماننـد تبلیغات برخط ،پورنوگرافی و منابع حاوی ویروس یا معیوب یا خود ویروس ها بایـد اتخاذ شود نمونه گیری راهی است برای آرشـیو کردن این نوع منابع که ممکن است برای بعضـی پژوهشگران دارای اهمیت باشند و پیش بینی احتمال استفاده از آن ها در آینده در زمان حال دشوار است

-1

-2

2.3. دسترسی و روش های توصیف

1.2.3

پایه فنی روش های توصیف

1.1.2.3

کلیات

گرد هم آوردن منابع آرشیو شده و فراهم کردن دسترسی به آن ها از طریق یک آرشیو وب به تنهایی فعالیتی متداول است آرشیوهای وب همانند وب زنده عمل می کنند و اغلب رابط کاربرهایی دارند که به کاربران امکان جست وجو و ناوبری (1) در محدوده آرشیو وب را می دهند. ویژگی مهمی که در طراحی رابط کاربر ضروری است بعد زمانی است که باید در نظر گرفته ،شود به گونه ای که کاربر بتواند هر یک از ویرایش های مختلف یک صفحه را به همان حالتی در زمان های مختلف گیر انداختن و رونوشت برداری بوده است و بتواند از میان این رونوشت ها به آسانی ناوبری کند و تغییرات یک صفحه را ببیند متداول ترین روش مرور در یک آرشیو وب مرور از طریق نشانی وب است که میتواند با تاریخ گیر انداختن صفحه تلفیق شود.

ببیند .

1.2.2.3

نمایه سازی نشانی وب (الزامی)

نمایه ها شناسه هایی (2) را از آرشیو وب در اختیار قرار می دهند نمایه ها جست وجو و ترتیب (3) را سرعت میبخشند و تجربه کاربری بهتری را میسر می.سازند نمایه نشانی وب یا شکل دستکاری شده ای از نشانی ،اولیه اصلی ترین نمایه در آرشیوهای وب و نقطه ورود به خدمتگر است که آرشیو را میزبانی می.کند تاریخی که منبع خزش شده است میتواند با نشانی

وب تلفیق شود تا بتوان میان ویرایش های مختلف یک منبع تمایز قائل شد. رویکرد جایگزین پیاده کردن شناسگر دائمی (4) برای هر منبع است که ممکن است شکل یک نشانی وب را داشته باشد؛ ولی نکته اصلی آن این است که موسسه گردآورنده تضمین میکند که شناسگر ارجاع و دسترسی به منبع را بدون محدودیت تامین کند.

1- Navigate

2- Entry points

3- Sorting

4- Persistent identifier

3.1.2.3

سایر انواع نمایه سازی (اختیاری)

جست وجوی تمام متن روشی در دسترسی به منابع است که آرشیوهای وب به گونه ای روزافزون برمی گزینند. لازمه آن نمایه تمام متن و موتور جست وجو است این روش هرچند راه حل مقیاس پذیرتری (1) برای دسترسی است اجرای آن از حیث فنی چالش برانگیز است. کلیدواژه فراداده نیز به صورت خودکار از منابع آرشیو شده استخراج و برای فراهم کردن دسترسی مورد استفاده قرار میگیرند جامعه پژوهشگران نیازی رو به فزونی را برای داده کاوی و پیوند کاوی در آرشیو وب ابراز کرده اند بعضی از توسعه های جدید حاکی از آن هستند که در آرشیوهای وب تمرکز از سطح منابع تکی یا وبگاه ها به کل آرشیو وب تغییر جهت داده است.

کاربرد فنون مصور سازی (2) و تحلیل داده (3) با گشودن الگوها و گرایش ها ،روابط و زمینه های جاسازی شده فرصت هایی برای دسترسی به نماهای (4) متفاوتی از آرشیو وب فراهم کرده است. تا پیش از آنکه پیشرفت های مذکور به طور وسیع برای اجرا برگزیده شوند در دسترس قرار دادن وضعیت پیشین منابع تکی اینترنتی اصلیترین مکانیسم دسترسی به آرشیوهای وب بود این گزارش فنی به این موضوع نیز معطوف خواهد بود.

فهرست نویسی (اختیاری)

روش های سنتی مدیریت کتابشناسی ها با فهرستنویسی ،منابع همانند فهرستنویسی کتاب های چاپی و مقالات قابل کاربرد در آرشیو وب هستند این روش خوبی برای یکپارچه کردن آرشیوهای وب با مجموعه های موجود در کتابخانه هاست تا آن ها از طریق جست وجو در فهرست کتابخانه ها نیز قابل اکتشاف شوند؛ ولی این رویکرد منبع محور است و به علت مقدار زیاد اشیایی که آرشیوهای وب در خود دارند و چالش در تعیین منابعی که باید فهرستنویسی شوند، به دشواری برای کاربرد در آرشیوهای وب قابل تطبیق است فهرستنویسی را میتوان در سطح بالاتری از دانه بندی (5) قرار داد مثلا در سطح مجموعه های ویژه و نه در سطح وبگاه ها

1- Scalable

2- Visualization

3- Data analytics

4- Views

5- Gramularity

5.1.2.3

ابزارهای اکتشاف منبع با استفاده از فراداده (اختیاری)

دسترسی را میتوان از طریق افزودن فراداده به منبع فراهم کرد آرشیوگر میتواند وبگاه ها را با استفاده از روش های خودکار به سلسله مراتب های موضوعی یا مجموعه های مربوط به یک رخداد یا موضوع دسته بندی کند . تگ ها (کلید واژه های) کاربران که آرشیوگر یا کاربران به منبع می دهند نیز میتواند به رابط کاربری افزوده شود.

2.2.3

پایه :فنی روش های دسترسی

1.2.2.3

کلیات

دسترسی به منابع آرشیو شده با استفاده از نرم افزار خبره در ،یافتن بازیابی و باز پخش منابع وبی آرشیو شده میسر میشود که ممکن است با چند بسته نرم افزاری جداگانه که با هم کار میکنند اجرا شود مجموعه سامانه نرم افزاری که معمولاً ابزار دسترسی نامیده می شود صرفنظر از اینکه یک ابزار دسترسی چگونه طراحی و پیاده سازی می شود دارای ویژگی هایی است که بعضی اختیاری و بعضی دیگر اجباری اند .

اجرا (1) (اجباری)

نرم افزار دسترسی باید قادر باشد منابع را به صورت یکتا شناسایی حتی اگر بعضی منابع به دفعات برداشت شدهاند و شیء مذکور را از انباره (2) آرشیو بازیابی کند.

3.2.2.3

بازنویسی نشانی وب (اجباری)

صفحات HTML که نرم افزار دسترسی آن ها را باز میگرداند باید از شکل اولیه شان تغییر

1- Rendering

2- Repository

یابنـد نشانی های وب که جاسازی شده اند (چه مطلق (1) باشند و چه نسبی (2)) باید به مکان منبع درون آرشیو دیجیتالی اشاره کنند و نه به مکان اولیه منبع در وب زنده این کار به چند روش قابل اجراست .

-بازنویسـی در زمـان برداشت اطلاعـات رخ دهـد (آرشـیوگر بایـد برای بازنویسـی فوری نشـانی وب در محتوا و وارد کردن این محتوای دستکاری شده به انباره آرشیو تصمیم گیری کند.)

-می توان اجرای عملکرد حفاظت برای منابع آرشیو شده را که برای تحقق هدف پیش گفته انجام میشود تا مرحله بعدی به تعویق انداخت.

پیوندهای موجود در منبع به مکان جدید در انباره آرشیو باز میشوند.

-زمان اجرای (3) بازنویسـی نشانی وب را میتوان با اجرای کد بر خدمتگر در هر پرس و جوی منبع انجام داد و یا با تهیه رونوشتی از منابع اصلی برای کاربر همراه با کدی که باید توسط او اجرا شود تا نشانی وب به طور پویا بازنویسی شود.

3.2.3

محدودیت ها

فرایند برداشت و پردازش منابع اینترنتی آرشیو شده مستلزم تغییراتی است که هنگام بازپخش در اثنای دسترسی بر ظاهر و رفتار منابع اولیه و تجربه کاربر از آنها تاثیر می گذارد. رونوشت های آرشیو شده را باید نوعی تصویر لحظه ای در نظر گرفت که در نقطه ای

از زمان منجمـد شـده اند تعامل با وب زنده را از دست می دهند پیام های ارسال شده در تالارهای گفت و گو (4) و گروه های بحث (5) فرم های (6) وب و جست وجو از آن جمله انـد همچنین ممکن است منبعی به طور کامل برداشت شود ولی توان نرم افزار اجرا در نمایش آن محدودیت ایجاد کند و در نتیجه برای کاربر غیر قابل دسترسی باقی بماند یک مسئله متداول هنگام باز پخش منابع آرشیو شده پدیده ای است که به آن تراوایی به وب زنده میگویند و هنگامی رخ می دهد که پیوندهای یک منبع آرشـیو شده به جای آنکه به نسـخه آرشیو شده در آرشـیوی که در آن قرار دارند بازشوند به اصل همان منبع در وب زنده باز می.شوند علت این پدیده معمولاً بازنویسـی نادرست نشانی های وب

1- Absolute

2- Relative

3- Runtime

4- essage boards

5- Discussion forums

6- Web forms

است و از پیوندهای جاسازی و نوشته شده در زبان جاوا اسکریپت ناشی میشود که با ابزار

دسترسی کشف نشده اند.

4.2.3

استراتژی های دسترسی

منـابع ،اینترنتی بـا وجود اینکه آزادانه در دسترس همگـان قرار دارنـد عمومـاً تحت حمـایت حقوق مـالکیت معنوی هسـتند . موسـسات گردآورنده محتوای وب بسته به قوانین مربوط (نگاه کنید :

404) و اینکه چه بخش هایی قانوناً مجازند طیفی از استراتژی ها را در این خصوص به کار می برند.

-آرشیو تیره (1) مجموعه هایی که هیـچ کس نبایـد به آن ها دسترسـی داشـته باشد (به جز مواقعی برای کارکنان برای رتق و فتق امور اداره آرشیو)

-آرشیو خاکستری (2) مجموعه هایی که فقط کاربرانی که مجوز دارند میتوانند آنها را ببینند (مانند پژوهشـگران) یا دسترسی به آنها محدود به حضور در محل آرشیو (مانند تالارهای

مطالعه کتابخانه) است؛

-آرشیو برخط (3) دسترسی برای همه کاربران میسر است و معمولاً از وبگاه یک موسسه گردآورنده قابل دسترسی است.

آرشیو به گونه ای دیگر میتواند مدل آمیخته (4) را به کار گیرد که در آن هر بخش از آرشـیو به طور جداگانه یکی از استراتژی های گفته شده را بر می گزیند شایان ذکر است که دسترسـی به آن دسـته از منابع که به صورت غیر « مشارکتی » (5) برداشت می شوند و مجوز برداشت آن ها صراحتا داده نشده مفروض یا ضمنی ،است در صورتی که صاحبان حقوق درخواست کنند قطع می شود .

3.3. روش های حفاظت

1.3.3

پایه فنی

حفاظت منابع آنالوگ مانند کتاب ها یا پیشینه ها بر حفاظت ماده اصلی معطوف است حال آنکه

حفاظت دیجیتالی با مسائل متفاوتی سروکار دارد.

Dark archive –1

Grey archive –2

Online archive –3

Mixed –4

Opt–Out –5

در لایه زیرین منابع دیجیتالی از صفر و یک تشکیل شده اند (جریان بیت ها) از حامل داده یا رسانه ای که بر روی آن ذخیره می شوند مستقل اند . می توان بدون آنکه اطلاعاتی گم شود بیت ها را در حامل داده دیگری رونوشت برداری کرد و نسخه ای مشابه منبع یا منبع اصلی پدید آورد. از آنجا که حامل های داده (معیوب خراب) یا از رده خارج می شوند، لازم است برای سالم نگه داشتن بیت ها آنها را به حامل های جدید منتقل کرد اگر رونوشت برداری به صورت قاعده مند انجام ،شود منطقی است که فرض کنیم جریان بیت بدون گم شدن به طور دائم تحت حفاظت قرار دارد علاوه بر سالم نگه داشتن بیت ها قابل استفاده نگه داشتن آن ها نیز چالشی حقیقی برای حفاظت دیجیتالی است.

جریان ،بیت تا وقتی با استفاده از محیط نرم افزاری و سخت افزاری اولیه اجرا نشود برای انسان قابل درک نیست با تحول سریع ،فناوری مقایسه سامانه های جدید با سامانه های قدیمی و نرم افزارهای اجرای جدید که قادر به تفسیر (1) فایل فرمت های قدیمی تر هستند، دیگر ممکن نیست حتی اگر بتوان نرم افزارهای قدیمیتر را بر سامانه های فعلی اجرا کرد ممکن است کاربران فعلی قادر به استفاده از آن ها نباشند؛ بلکه انتظار میرود کاربران با آن به شیوه های کاملا متفاوت تعامل داشته باشند

حفاظت دیجیتالی در هر مرحله از گردش کار (2) در آرشیو وب باید در نظر گرفته شود. چالش اصلی در حفاظت آرشیوهای وب در مقایسه با سایر منابع دیجیتالی حجم زیاد داده و تنوع قالب های فایل و انواع رسانه است صفحات وب ممکن است حاوی ،تصویر فیلم موسیقی ،بازی پایگاه داده و انواع برنامه های کاربردی باشند ویژگی اصلی ،وب وجود پیوند میان صفحات است که ممکن است چالش های مربوط به حفاظت دیجیتالی را به علت وابستگی هایی که

پیوندها ایجاد میکنند پیش بکشد

2.3.3

محدودیت ها

آرشیوهای وب منابع متاخر راکد را در خود دارند و در مقابل استراتژی هایی که با نتایج متقاعد کنند یا با اطمینان به اثبات رسیده باشند وجود ندارند که بتوانند توان جامعه آرشیو را در حفاظت بلندمدت از منابع اینترنتی آرشیو شده نشان دهند.

هدف این گزارش فنی ارائه راه حلهای عملی نیست بلکه به عملکردهای فعلی استانداردها

و مسائل این حوزه میپردازد .

Interpret -1

Workflow -2

3.3.3

استراتژی های حفاظت

هـدف حفاظت ،دیجیتالی در حـد ،کمینه پیشـگیری از گم شـدن داده با حفظ تمامیت (1) جریـان بیت اولیه است استراتژی اصـلی برای جریـان بیت یـا حفـاظت فیزیکی رونوشت گیری و تهیه نسـخه پشتیبان، (2) شامـل عملیـاتی مثل ذخیره موازی داده (3) در مکـان های فیزیکی جداگانه تهیه منظم نسخه پشتیبان و وارسی منظم برای یافتن خطاهاست.

لازم است امنیت داده ها نیز حفظ شود تا بتوان از دسترسـی های غیر مجاز پیشـگیری کرد. حفاظت جریان بیت کمترین الزامی است که شامل همه منابع دیجیتالی می شود هرچند

هنگام اجرای برنامه حفاظت برای جریان بیت اندازه آرشیو وب نیز باید در نظر گرفته شود.

مهاجرت و همگون سازی استراتژی های پیچیده تر برای حفاظت هستند که از کارکرد ،رفتار و تجربه کاربر از منابع نیز معطوف اند.

این دسـته استراتژیها را «حفاظت ، منطقی» (4) مینامنـد و مسـتلزم تحلیل منظم و مسـتمر داده و قالب ها و خطرات برای اجرا هسـتند به علت بزرگی و قالب های گوناگون فایل موجود در آرشیوهای وب حفاظت منطقی برای آن ها بسیار چالش برانگیز است.

(الف) مهاجرت

مهاجرت قالب های فایل مسـتلزم برگردان فایل به قالب های جدیـد است پیش از آنکه در چارچوب محیط فناورانه فعلی غیر قابل اسـتفاده شوند برگردان داده به قالب فایل جدید محتوا را تغییر میدهد و ممکن است باعث آسـیب به آن شود بنابراین لازم است به منظور سـنجش از پیش تحلیل خطر صورت گیرد امکان و تاثیر گم شدن داده .

مهـاجرت را می توان هنگـامی که در یک قالب فایل خطر کهنه شـدن بروز می کنـد یا هنگام دسترسـی مهاجرت حین کار (5) انجام داد هزینه مهـاجرت نسـبت مسـتقیمی با تعـداد فایل های در دست مهاجرت دارد برای آرشـیوهای وب بزرگ مقیاس مهاجرت ممکن اسـت بسیار پرهزینه باشد پیچیدگی و وابستگی فایل ها به هم نیز به مشکل افزوده می شود و اعتبار مهاجرت را دشوار می سازد .

ب همگون سازی

همگون سازی خلق دوباره کارکردها و رفتار محیط یک سامانه خارج از رده در سامانه فعلی

1- integrity

2- Back up

3- Parallel data storage

4- Logical preservation.

با استفاده از نرم افزار خبرهای است که همگون ساز نام دارد همگون ساز رفتار سامانه از رده خارج شده را تقلید میکند و دسترسی به منابع تاریخ گذشته (1) را بـدون تغییر دادن آن ها میسـر می سازد ،باوجوداین همگون سازی به دشواری بی عیب انجام میشود و حسن عملکرد آن تقریبی است همگون ساز ،خود به محیط دیگری وابسته و در معرض خطرات حفاظتی است. تدوین همگون سازها هزینه بر است ولی نیـاز به پرداختن به تک تک اجزا را مرتفع میسازد در آرشـیو ،وب نرم افزار همگون ساز بایـد کار کرد مرورگرهای متـداول و ابزارهای بخش رسانه ها (2) را در وضعیت زمانی ای که صفحات وب آرشیو شده اند بازآفرینی کند

مهاجرت و همگون سازی را باید به عنوان بخشی از برنامه حفاظت سامانه های آرشیوی دیجیتال در نظر داشت.

4.3.3

فراداده حفاظت

حفـاظت بلندمـدت شامـل حفـظ امنیت فراداده هـای همراه منـابع در آرشـیو وب نیز می شود که برای فعالیت های پشتیبانی از مـدیریت ،مجمـوعه دسترسـی و حفـاظت حیاتی.انـد فراداده هـای مختلفی وجـود دارد که می شود آن هـا را درون منـابع جاسـازی کرد به صورت خودکار در اثنای فرایند آرشیو کردن تولید کرد یا به صورت دستی توسط کارکنان به منابع افزود استاندارد متس (3) پنج نوع فراداده را تعیین میکند که به شرح زیر در آرشیو وب به کاربرده می شوند

الـف فراداده توصـیفی موسـساتی که منـابع آرشـیو وب را فهرستنویسـی یـا فراداده را به صورت دسـتی به این منـابع الصـاق کننـد معمولا فرادادههای توصیفی بیشتری دارند موسساتی که دست اندرکار آرشیو وب در مقیاس بزرگ به صورت خودکار هستند باید فراداده جاسازی شده درون منابع را برداشت کنند یا روش های خوشه بندی یا دسته بندی خودکار را برای به دست آوردن چنین فراداده هایی به کار ببرند.

ب فراداده ساختاری منابع ،اینترنتی اغلب اشـیایی ترکیبی هستند که از عناصر ساختارمند و هم پیوند درست شده اند روابط ساختاری این منـابع به صورت واضـح در فرانماهـایی نظیر (4) متس قابـل بیـان و قابـل ضـبط است. چنین فراداده هـایی در مـوارد مهـاجرت فایـل که ابرپیوندهای آن نیز لازم است

1- Out of date

2- Media players

3- Metadata Encoding and Transmission Standard (Mets)

4- Metadata Schemes

همراه فایـل مهاجرت کننـد تا ناوبری در آرشـیو میسـر ،شود مفیـد .است برخی موسـسات تصـمیم می گیرنـد که چنین روابطی را به طور جداگانه و افزون بر بقیه به پیشینه نیفزایند چراکه ذاتا

درون منابع وجود دارند.

ج) فراداده منشا

فراداده منشا چگونگی و چرایی پدیـد آمدن یک منبع و اینکه در طول عمرش چه بر سـر آن آمده را توضـیح می دهد بعضـی از فراداده های توصـیفی مانند ثبت توضیح منطق و علت اصلی وجود یک مجموعه در آرشیو وب را نیز میتوان فراداده منشا به حساب آورد. فراداده منشا را می توان در سـطوح پـایین تر از قبیل فعالیت های ضـبط فایل توسط خزشـگر آرشـیو (ماننـد فایل های ،پیکربنـدی گزارش های خزش و کارنـامه هـای ورود به سامـانه) و اطلاعـات مربوط به تعامـل میان خـدمتگر وب و خزشـگر شامل نشانی ،وب تاریـخ خزش و نشانی IP خدمتگر یافت.

د) فراداده فنی

فراداده فنی ویژگی های فنی یک شـیء دیجیتالی را توصـیف میکنـد و معلوم می کند که چگونه میتوان به آن دسترسـی داشت دسـتکاری یا حفاظتش کرد این نوع فراداده در مدل مرجع آرشـیو باز (1) بازنمون اطلاعات (2) نامیـده می شود قالب فایل که نوع فایل آن را نشان دهد نیز مثالی از فراداده فنی مربوط به آرشیوهای وب و از آمارهای هسته برای شناساندن

مجموعه است (برای توضیح نگاه کنید به 3.2.3.4)

ه) فراداده حقوقی

فراداده حقوقی مالکیت و اسـتفاده مجاز و قانونی از منابع را تعیین می کند. شرایط را میتوان برای زمانی در آینده به کاربرد این اطلاعات باید همراه منابع حفظ شود تا از دسترسی غیر مجاز پیشگیری شود.

فراداده های منشا ،فنی و حقوقی مجموعاً فراداده مدیریتی نامیده م یشوند.

4.3. مبانی قانونی آرشیو وب

1.4.3

کلیات

ابتکارات آرشیو وب با مخاطرات قانونی بسیاری مواجه هستند مهمترین آن ها به مالکیت ،فکری

1- OAIS(Open Access Information System)

به ویژه حق ،مولف حریم خصوصی و محافظت از داده های شخصی (1) مربوط می شود . موسسات گردآورنده همچنین ممکن است مسئول بازنشر (2) محتوای افترا آمیز و پردازش و توزیع مواد غیرقانونی باشند قانون گذاری ملی با حمایت قانونی از موسسات گردآورنده به گونه ای موثر با این خطرات مقابله می کند .

قانون گذاری ملی موثرترین و باکفایت ترین چارچوب برای فعال کردن و حمایت از آرشیو وب است آرشیو کردن وب ممکن است با قانون گذاری درباره حق مؤلف و ایا واسپاری آثار یا هر قانون دیگری که به طور خاص ماموریت ها و جایگاه شان و موقعیت یک موسسه گردآورنده را تعیین می کند آغاز شود این قانون گذاری به طور خاص مربوط به موسسات عمومی است که معمولا جایگاه و عملکردشان را قانون تعیین میکند مانند کتابخانه ها یا آرشیو ملی موسسات خاص واسپاری آثار مانند موسساتی که وظیفه آنها حفاظت از فیلم یا پخش رسانه ای است آرشیوها یا موزه های عمومی موسسات فاقد

وظیفه قانونی نیز ممکن است وب را آرشیو کنند این دسته از موسسات یا با صاحبان حقوق به تفاهم میرسند یا دسترسی به منابع آرشیو را با هدف مدیریت خطرات مربوط به قانون محدود می کنند .

بعضی از موسسات نیز با آرشیو و در دسترس قرار دادن منابع اینترنتی بر مبنای مجوز ضمنی با این پیش فرض که در دسترس همگان قرار دارند خطراتی را پذیرا میشوند بعضی از کشورها هنوز مینای قانونی روشنی برای آرشیو وب ندارند. دسته دیگر چارچوب عامی دارند که مستلزم قانون گذاری ثانویه ای برای تفسیر و تنظیم اجرای آن است موسسات آرشیو وب در بعضی کشورها ابتکارات آرشیو وب خود را بر اساس واسپاری داوطلبانه منابع توسط ناشران توسعه دادهاند حتی در کشورهایی که قانون گذاری ملی در این زمینه صورت پذیرفته است معمولاً راه برای تفسیر قانون باز است برای موسسات آرشیو وب،

تعیین رویکردشان به آرشیو وب از جمله ارزیابی خطر و تجربه اندوزی در مرحله اجرا ضروری است.

2.4.3

دامنه و روش های گردآوری

در قانون گذاری برای آرشیو وب ممکن است به گونه ای مُصرح بعضی از محتواها در محدوده گردآوری وارد یا از آن خارج شود این قوانین مرزها یا قلمرو جغرافیایی یک دامنه ملی را تعیین و بسامد یا عمق برداشت مجاز محتوا را مشخص میکنند

Privacy -1

Republish -2

عنصـر کلیـدی در قـانون گـذاری این است که بایـد پیش از برداشت از صاحبـان حقوق اجازه گرفته شود این امر تاثیر مهمی در استراتژی گردآوری موسـسه دارد . برداشت پشته ای فقط هنگامی امکان پذیر است که مجوزی لازم نیست در غیر این صورت برداشت انتخابی مدل مناسب تری است.

رویکردهای دیگر در مدیریت مجوز مدل «تاریخ گذشـته» یا « درخواست حذف » (1) هستند که در آن ها منابع بر اساس مجوز تلویحی یا مفروض برداشت می شوند و در دسترس قرار می گیرند و با درخواست صاحب حقوق از آرشیو حذف می شوند.

اخذ « مجوز کلی » (2) برای برداشت منابع متعدد از هر ،ناشر راه دیگری برای کاهش هزینه های مربوط به مدیریت توافق با ناشران متعدد است . ممکن است در ،قانون حقوق دریافت اطلاعات حفاظت شده برای کمک به ارتقاء کیفیت و تمامیت مجموعه به موسـسه اعطا شود .

فهرست نام دامنه های ملی یا کدهای شـناسایی و اطلاعات مدیریت حقوق دیجیتالی برای انتشارات تجاری از نمونه های این حقوق است . ممکن است ،قانون ناشـران را به طور خاص جبور به تحویل اطلاعات کنـد و برای عـدم تابعیت از ،قانون مجازات کیفری در نظر بگیرد. ،همچنین ممکن است در ،قـانون متون گردآوری ویژه ای توصـیه یـا احیا شود. در بعضـی قوانین به طور مشـخص برداشت خودکار منابع اینترنتی تـا انجام توافق دوجانبه میان موسـسه و ناشـران در خصوص پروتکل برداشت اطلاعات تشویق یا مجاز دانسـته شـده است. قانون میتوانـد محـدودیت هـای بیشتری دربـاره راه هـای آرشـیو کردن وب اعمـال کنـد تصـمیم که متـاثر از قانون است رعایت کردن یا نکردن استاندارد منع خزش است این امر تفاوت معناداری در گیر انداختن یا حذف محتوا ایجاد می کند .

در کشورهایی که قانون اجبار به واسپاری آثار برای گردآوری همه منابع اینترنتی وجود دارد معمولا یک موسسه موظف میشود که به تنهایی این کار را انجام دهد این وظیفه ممکن است توسط سازمان های مختلف به طور مشترک انجام شود مانند

مهمی

میان موسسات ملی یا فدرال و موسسات منطقه ای و محلی

میان کتابخانه های ملی و آرشیوهای ملی

در چارچوب یک شبکه یا کنسرسیوم از موسسات تخصصی

1- Notice and taken down

2- Blanket Permission

3.4.3

دسترسی به آرشیوهای وب

دسترس پـذیری وجه مهم قانون گذاری برای آرشـیو وب است و معین میکند که تحت چه شـرایطی میتوان از آرشـیو اسـتفاده کرد شـرایط ،دسترسی با تنظیمات برداشت اطلاعات هماهنگ :هستند اگر اخذ مجوز از صاحب حقوق الزامی است دسترسی بر خط برحسب آن

اجازه داده شود؛ اگر مود؛ اگر برداشت پشته ای بدون مجوز انجام میشود احتمال بیشتری برای

محدود شدن دسترسی وجود دارد .

محـدودیت های حق ،مولف مانند ،چاپ برداشت بخش هایی از محتوا رونوشت برداری ،الکترونیکی و بارگیری (1) در مورد آرشیوهای وب نیز به کار میروند هرگاه آرشـیو وب به عنوان شکلی از واسپاری قانونی اجرا ،شود ممکن است قوانین انتشار کتابشناسی برای آرشیوهای وب را الزامی کننـد . این امر برای مجموعه هـای بزرگ چـالش برانگیزترین مسـئله است معمولاًـ آرشـیوهای وب به جـای انتشـار فراداده توصیفی رابط کاربرهای جست وجو برای کاربر فراهم می کنند .

4.4.3

حفاظت آرشیوهای وب

حفـاظت بلندمـدت الزام و توجیه کلیـدی برای گردآوری منـابع میراث فرهنگی است بنـابراین قـانون بایـد شامل اشاره یا اجبار به تضـمین دیرپایی (2) آرشـیو وب باشـد. در قانون بایـد به ویژه مشـخص شـده باشـد که آیا پاک کردن منابع مجاز است یا آن ها بایـد برای همیشـه نگهداری .شوند بیشتر کتابخانه های ملی که منابع اینترنت را بر اساس واسپاری آثار گرد می آورند ملزم به حفظ آن برای آیندگان می شوند. اگر هدف از گردآوری فراهم کردن مجموعه های داده برای پژوهش های کوتاه یا میان مدت ،باشد کتابخانه های پژوهشی و سایر موسسات ممکن است

ملزم به حفظ منابع در بینهایت زمان نباشند .

5.3. علت های دیگر برای آرشیو وب

1- Downloading

2- Longevity

1.5.3

کلیات

سیـاست نیز میتوانـد انگیزه ای برای آرشـیو کردن وب .باشـد این انگیزه هـا بازتابی از بینش استراتژیک موسـسه و روش و منش آن در قبال نوآوری هـای فرهنگی و فنـاوری هسـتند. وب طیف متنوعی از منـابع دیجیتال زاد (1) و دیجیتالی شـده را میزبانی میکنـد این دسـته اخیر معمولاً چاپ می شوند ،کتاب ،پیایند انتشارات دولتی و مانند آن یا بر رسانه مادی سوار می شوند (،فیلم، موسـیقی بازی بر دیسک یا نوار) (2) و بسـیاری از آن ها تاکنون از قالب های مختلف مهاجرت کرده.انـد وب به سـرعت تغییر میکند و ناماندگار است به طور منظم منابع ارزشـمندی ناپدیـد می.شونـد حافظت از وب برای موسـساتی که رسالت دائمی حفاظت از منابع فرهنگی میراث را دارند هم طبیعی و هم قابل بحث است.

آرشیوسازی وب استمرار (3) دیجیتالی را تضـمین می کند و عملی ضـروری برای پیشگیری از ایجاد حفره های (4) دیجیتالی در دانش و حافظه یک ملت است می تواند دسترسی را به پژوهش های مورد ارجاع حفظ کند این انگیزه به ویژه برای کتابخانه های ملی و آرشیوهای

ملی انگیزه ای قوی است.

2.5.3

کمک به پژوهش دانشگاهی

اینترنت فضـایی پر مشارکت و نوآورانه است که مردم در آن باهم ارتباط و همکاری دارنـد ممکن است چنین اسـتدلال شود که وب دانش اجتماعی و محصولاتی با ارزش پژوهشی تازهای خلق میکند که با میراث ملی مرتبط اند .

ما شاهد پدیدار شدن رویه ها و جوامع جدید پژوهشی هستیم که به مطالعه وب زنده و بالقوه به آرشیو آن اختصاص یافته است. آرشیوهای وب امکانات پژوهشی بی همتایی برای دانشمندان و پژوهشگران اینترنت فراهم می کنند .

این آرشـیوها نه فقط ویراست های تاریـخ وبگاه ها را مطالعه میکنند بلکه داده کاوی و پیوندکاوی های بزرگ مقیاسـی انجام میدهند که به انتزاع الگوها و گرایش ها و گشودن دانش نهفته یاری می رساند. داراوی یا تحلیل داده هنوز دوران کودکی خود را سپری می کند و

1- Born- digital

2- Tape

3- Contenting

4- Black hole

معمولاً دانشـمندان علوم اجتماعی به آن مبادرت میکننـد اما به نظر میرسـد می تواند به سایر حوزه های دانشـگاهی تسـری یابد و در بیشتر رشته ها مفید واقع شود.

آرشیو کردن وب در سطح ،موسسات راهی برای ترویج یا برجسته کردن منابع خاص دیجیتالی است؛ به ویژه در موسساتی نظیر دانشگاه ها که انتشارات اعضای هیئت علمی و دانشـجویان را منتشـر می.کنند آرشـیو کردن وب تلاشی است ارزشمند در گردآوری منابع بر خط که حاوی و حائز ارزش پژوهشی ،هستند و انگیزهای قانع کننده برای بسیاری از موسسات

آرشیو وب است.

3.5.3

پشتیبانی از کاربردهای مختلف توسط عموم مردم

اینترنت گذشـته هایی از همه جنبه های زنـدگی را میزبانی می کنـد . برخلاف منابع چاپی هر کسـی می توانـد در وب مطلب منتشـر ،کند گرچه ارزش مطـالب باهم متفاوت است. کل مطالب مجموعه واحـدی از منابع را تشـکیل می دهـد که گزارش هایی است از خاطرات و تعاملات جامعه و افراد.

برپایه آرشـیوهای وب خـدمات حرفه ای و خصوصـی متنوعی میتوان تدارک دید همچنین از آرشـیوهای وب می توان به عنوان شاهدی در مناقشات مربوط به حقوق مولف یا پژوهش های خانواده و شـجره شناسـی دیجیتالی (1) بهره برد ایجاد دسترسـی آزاد و بلندمدت به منابع اینترنتی به ویژه به شـخص پدیدآورندگان محتوا در همین زمان و نسل های بعدی ایشان در آینده استدلال محکمی در اهدای آرشیو وب به عنوان یک فهرست عمومی است.

4. آمار

1.4 کلیات

آمارهـا داده هایی عینی هسـتند که پایه تحلیل ها و تفسـیرهای دیگر قرار میگیرنـد شاخص های کیفی نشان از درجه ای از داوری ارزشـی هسـتند عـدم تطـابق بـا معیارهـا نشـانه منفی بودن ارزیابی است. در این گزارش ،فنی آمارها به صورت قـدر مطلق (2) گرفته می شونـد و شاخص های کیفیت

به صورت اعداد نسبی و درصد.

آمارها و شاخص های کیفیت باید پایا (3) ، حاوی بار ،اطلاعاتی و قابل مقایسه باشند و روش های

1- Digital genealogy

2- Absolute numbers

به دست آوردن آنها نیز باید عملی و انعطاف پذیر باشد.

وضـعیت فعلی آرشیو کردن وب نشان می دهد که تولید بعضی آمارها باید متکی بر نرم افزار ،برداشت نمایه سازی یا مرور باشد و انتخاب نرم افزار متفاوت به تفاوت نتایج منتهی می شود بنابراین توصیه میشود اگر هدف از اندازه گیری مقایسه ،است برای آمارگیری از نرم افزار

واحدی استفاده شود.

آمارگیری و سنجش کیفیت در آرشیوهای وب با توجه به بزرگی اندازه اکثر آن ها عملی و مقرون به صرفه است .

شاخص های آماری کیفیت که در این گزارش معرفی می شوند مبتنی بر عملکردهای متداول در آرشیوسازی وب هستند و مروری اجمالی و قابل اعتنا از آرشیو وب نیز ارائه و مقایسه میان آن ها را امکان پذیر می کنند .

این گزارش ها آمارهای عمومی (1) و شاخص های کیفی را پیشنهاد می کنند همه آن ها در انواع

مختلف آرشـیوهای وب قابل اسـتفاده .نیسـتند علاوه بر ،این از آنجاکه فناوری پیشـرفت می کار آرشـیو کردن وب حرکتی رو به پیش دارد بعضی از آمارها و شاخص ها نیاز به روزآمد

شدن خواهند داشت.

این فصـل از گزارش فنی تعـدادی از آمارهـای مربوط در هر بخش را مطرح و توصـیف کنـد ،باوجوداین مجموعه کوچک تری از آمارهای هسته به عنوان اساس در نظر گرفته

می شوند که در قالب جداولی در انتهای هر بخش همراه با مثال ارائه می شوند .

2.4. آمارهای گسترش مجموعه

1.2.4

کلیات

آمارهای زیر رشد مجموعه آرشیو وب را با ردگیری خروجی های کمی آن اندازه می گیرد. این کار به برنامه ریزی و پایش گسترش مجموعه کمـک می کنـد و تحلیـل تفضـیلی هزینه ها را ممکن می سازد . آرشـیوهای وب برخلاف منابع ،آنالوگ دارای چنـد منبع غیر خطی (2) هسـتند که با هم پیوند دارند بعضـی از این منابع برای کاربران بازپخش میشوند بعضی دیگر فایل ها و فراداده های غیر قابل تفکیکی هستند که بخشی از منابع اند ولی کاربر آن ها را نمی بیند ،بنابراین

1- Generic Statistics

2- Non linear

آمارهایی که حجم آرشیو وب را اندازه میگیرند با آن هایی که منابع فیزیکی را اندازه می گیرند نباید مقایسه شوند .

بیشتر این آمارها مختص آن دسته از منابع و بی هستند که آرشیو شده اند نمی توان آن ها را برای اندازه گیری محتوای وب زنده به کار برد

2.2.4

اندازه گیری وبگاه های هدف شمارش وبگاه های هدف و گیر انداختن ها

1.2.2.4

هدف

موسسات گردآورنده باید بتوانند اهداف آرشیو کردن وب را در پرتو سیاسته ای خود بیان ارزیابی کنند به عبارتی باید بتوانند نتایج گردآوری در آرشیو وب را – ضمن مقایسه با اهداف تعیین شده با ذکر میزان تحقق و کارآمدی فرایند گردآوری ارزیابی کنند هیچ روش سرراست و یکسانی برای بیان اهداف و مقاصد گردآوری وجود ندارد این کار عملا توسط هر موسسه و برمبنای سیاست و اهدافش تعیین می.شود چارچوب عامی وجود دارد که استفاده از مفهوم « هدف » (1) و « گیر انداختن هدف » (2) را پیشنهاد می کند و می تواند فعالیت مجریان گزینش و مدیریت وبگاههای هدف را اندازه بگیرد این شاخص مختص ارزیابی انتخاب و اندازه گیری مدت زمان صرف شده برای تعیین وبگاه هایی است که باید جزء

آرشیو قرار بگیرند

2.2.2.4

روش

یک هدف از یک یا چند هسته تشکیل می شود و هر یک از هسته ها دارای مجموعه تنظیمات خزش است که دامنه آن را تعیین می.کند هدف مجموعه منابع معناداری است که باید گردآوری شوند دامنه هدف متغیر است از طیفی از منابع موجود در یک نام دامنه که به هم پیوند شده و به عنوان « یک وبگاه » شناخته میشوند تا تک منبعی که با یک نشانی وب (مثلاً یک فایل پی دی اف یا یک ویدئو) شناخته می شود یا حتی تمام یک دامنه سطح (3) بالا یک

1- Traget

2- Target captures

3- Top Level Domain

هدف را می توان بیش از یکبار خزش کرد هر خزش گیر انداختن یک هدف است.

مثال خزش روزانه صفحه نخست وبگاه نیویورک تایمز:

- هسته ممکن است این نشانی وب باشد ؛

/http://www.ngtimes.com/, http://global.nytimes.com

دامنه میتواند دستور « صفحه نخست و همه منابع آن را با یک کلیک (در یک سطح درون

صفحه) خزش کن » باشد؛

بسامد خزش ممکن است هر روز باشد؛

هدف همه موارد بالاست؛

تک مجموعه منابع خزش و ذخیره شده در دستور بالا یک وب گرفت از هدف است.

این روش عام به موسسات امکان می دهد اهداف عملی و ارزیابی خروجی را تنظیم کنند در مثال ،بالا ممکن بود هدف گردآوری در سال باشد با این ،حساب، می توان تعداد فعلی وب گرفت های هدف موردنظر را بر مبنای سال مقایسه کرد تا بتوان دستیابی به اهداف

مجموعه سازی را ارزیابی نمود

3.2.2.4

محدودیت ها

مقایسه تعداد منابع هدف و وب گرفت ها میان موسسات تنها هنگامی معنادار است که موسسات سیاست ها و عملکردهای گزینش یکسانی داشته باشند .

3.2.4

اندازه گیری آرشیو وب : شمارش نشانی های وب

1.3.2.4

هدف

شمردن نشانی های وب تنها راه سنجش اندازه آرشیو وب است نشانی های وب مربوط به مکان منابعی است که خزش می شوند و خزشگرهای آرشیوگر از آنها برای شناسایی و درخواست منابع از خدمتگرهای وب استفاده می کنند . خدمتگر وب پاسخ های استانداردی

را که با کد وضـعیت شناسایی می،شوند با ذکر وضعیت منبع درخواستی به آرشیو باز میگرداند ممکن است پاسخ تایید تحویل موفق منبع درخواستی یا حاوی نشانه ای از جابه جا شدن منبع به محل دیگر

ص: 54

(تغییر مسیر (1)) باشد .،همچنین ممکن است خدمتگر با پیغام خطا پاسخ دهد، نشان از اینکه منبع درخواستی در دسترس نیست. بعضی از پاسخ ها هم محتوا و هم فراداده را در اختیار قرار می دهند و بعضی دیگر فقط فراداده و دسته دیگر نیز تنها یک کد خطا را . (2)

دانستن این نکته مهم است که هر نشانی وب لزوماً قابل تطبیق به یک منبع انسان- خوان معنادار نیست؛ به گونه ای که بتوان آن را معادل « اقلام » یا « منابع فیزیکی » در مجموعه های سنتی یک کتابخانه در نظر گرفت. حتی هنگامی که یک خدمتگر وب در پاسخ به پرس وجوها ناموفق است باز پاسخی تحویل می دهد که نشان دهنده وضعیت تحویل یک درخواست باشد؛ مانند تغییر مسیر یا پیغام خطا اطلاعاتی از این نوع سند ممیزی (3) از فرایند برداشت و منشا اطلاعات در مجموعه یک آرشیو وب فراهم میکنند و میتوانند برای اهداف دسترسی یا حفاظت مفید واقع شوند پس این گزارش فنی توصیه میکند که همه پاسخ ها نگه داری و بخشی از آرشیو وب تلقی شوند.

نشانی های وب به عنوان شناسگر در وب و در سامانه پیام HTTP به کار می روند. نشانی های ،وب همچنین کوچک ترین واحد محتوای « خود شامل » (4) در یک آرشیو وب هستند عموما از آن ها برای ذخیره سازی و دسترسی به منابع آرشیو شده وب استفاده می شود . بنابراین پیشنهاد می کنیم از نشانی های وب برای شناسایی منابع و پاسخ های بازگشتی استفاده شود. از انواع کدهای وضعیت میتوان برای ردیف کردن یا دسته بندی گروه های منابع در آرشیو

وب استفاده کرد.

2.3.2.4

روش

جدول 1 مشتمل بر کدهای وضعیت مختلفی است که اعداد سه رقمی هستند و رقم اولشان دسته پاسخ ها را نشان می دهد هر کد وضعیت یک عبارت توضیحی (5) دارد که برای کاربران انسانی در نظر گرفته شده و در متنی کوتاه کد وضعیت را توضیح می دهد. جزئیات بیشتر درباره کد وضعیت را میتوان در(RFC 1616) یافت نگاه کنید به منابع انتهای کتاب

1- Redirection

2- Error code

3- Audit trail

4- Self- contained

5- Reason phrase

جدول -1 فهرست کدهای وضعیت HTTP

جدول ۱- فهرست کدهای وضعیت HTTP

کد وضعیت	عبارت توضیح
1xx	پاسخ موقت
100	ادامه
101	در حال تعویض پروتکل‌ها
2xx	موفقیت‌آمیز
200	موفق
201	محتوا ایجاد شده است
202	پذیرفته شده، ولی هم‌اکنون پاسخ نمی‌دهد
203	اطلاعات نامعتبر
204	بدون محتوا
205	تنظیم مجدد محتوای ارسال‌شده
206	محتوای ناقص
3xx	تغییر مسیر
300	گزینه‌های متعدد
301	انتقال دائمی
302	یافته شده
303	به محل دیگر مراجعه کنید
304	اصلاح‌نشده
305	استفاده از پراکسی
307	تغییر مسیر موقت
4xx	خطای کاربر
400	درخواست ناصحیح
401	غیرمجاز
402	پرداخت لازم

ص: 56

جدول

403	ممنوع
404	پیدا نشد
405	روش غیرمجاز
406	غیرقابل‌پذیرش
407	نیاز به تصدیق پراکسی
408	وقفه درخواست
409	ناسازگاری
410	رفته
411	طول [محتوای فیلد سرآیند] موردنیاز است
412	شکست پیش‌شرط
413	منبع مورد درخواست بیش‌ازحد بزرگ است
414	نشانی بیش‌ازحد طولانی است
415	نوع رسانه پشتیبانی نمی‌شود
416	محدودهٔ مورد درخواست در دسترس سرور نیست
417	شکست پاسخ مورد انتظار
5xx	خطای خدمتگر
500	خطای داخلی خدمتگر
501	اجرا نشده است
502	درگاه خراب
503	خدمت در دسترس نیست
504	وقفهٔ درگاه
505	نسخهٔ HTTP پشتیبانی نمی‌شود.

این گزارش فنی توصیه می‌کند که همهٔ نشانی‌های وب هنگام محاسبهٔ تعداد کل منابع (برداشت شده) در یک آرشیو وب، بدون در نظر گرفتن کد وضعیت هریک، به‌حساب آیند. باوجوداین، درک معنا و ماهیت کدهای وضعیت از این حیث اهمیت دارد که می‌توان از آنها

این گزارش فنی توصیه می کند که همه نشانی های وب هنگام محاسبه تعداد کل منابع (برداشت شده) در یک آرشیو وب بدون در نظر گرفتن که وضعیت هریک به حساب آیند. ،باوجوداین درک معنا و ماهیت کدهای وضعیت از این حیث اهمیت دارد که می توان از آن ها

ص: 57

برای دسته بندی یا پالایش منابع به منظور تحلیل بخش های خاص آرشیو وب استفاده کرد به عنوان نمونه دسته کدهای وضعیت 2xx نشانه تحویل موفق منابع مورد درخواست اند و با پاسخ های دسته xxx معمولا فراداده ها بدون منابع درخواستی آورده می شوند. دسته 5xx کاربرد فنی .دارند در صورتی که موسسه ای مایل به حفظ همانندی میان مجموعه های فیزیکی

آرا شیو وب ،باشد دسته 2xx اهمیت ویژه ای پیدا می کند . توصیه میشود نشانی های وب با

کدهای وضعیت ،200 ، 201 ، 203، 205 و 206 به طور خاص مدنظر قرار گیرند.

اگر فرایند مکرر زدایی (1) رخ داده باشد، تعداد نشانی های وب را می توان پیش یا بعد از آن شمرد مکررزدایی طی ،خزش هنگامی رخ می دهد که خزشگر در می یابد که یک نشانی وبی که در شرف خزش شدن است قبلاً گردآوری شده و در آرشیو قابل دسترسی است؛ پس خزشگر را دوباره خزش نمی کند .

ممکن است خزشگر اطلاعاتی درباره فرایند مکررزدایی تولید کند .

که در « استاندارد WARC از آن به بازبینی » (2) یاد می شود . هر دو عدد به دست آمده مفیدند: تعداد نشانی های وب پس از مکررزدایی کل منابع آرشیو را نشان می دهد و عدد مرجعی است که در تمهید اندیشیه ای ذخیره سازی و حفاظت بلند مدت کاربرد دارد؛ تعداد نشانی های وب پیش از مکررزدایی برای کاربران انسانی در نظر گرفته شده و از منظر محتوا و فکر معنادار است. این عدد فقط تا هنگامی که اطلاعات درباره فرایند مکرر زدایی در دسترس است مفید است به عنوان مثال به علت وجود پیشینه های بازبینی در فایل های WARC این اطلاعات در واقع نشان می دهد که نشانی مکرر زدایی شده تا تاریخ خاصی هنوز برخط بوده هرچند خزشگر آن را خزش نکرده است شیوه محاسبه پیش یا پس از (مکرر زدایی) همواره باید ذکر ،شود به ویژه هنگام مقایسه

میان آرشیوهای وب

3.3.2.4

محدودیت ها

وب از دوران آغازین ،خود عمدتاً از صفحات ایستای HTML با منابعی که به طور مشخص

ارجاع دار بودند تشکیل می شد ،وب به سرعت تحول یافت و مقادیر رو به افزایشی از محتوای تعاملی و تولید شده به طریق پویا در آن وجود دارد که ما را به اندیشیدن فراسوی مدل سنتی

1- De- duplication

2- Revisit

وب به عنوان مجموعه ای از « مـدارک » (1) یا « انتشارات » (2) شامل HTML ملزم می کند تعداد نشانی های ،وب در یک آرشـیو وب معـادل تعـداد «مـدرک» یـا «انتشارات» به معنایی که این عبارات معمولا در بافت یک کتابخانه افاده می کننـد .نیست هنگام محاسـبات آماری در آرشـیوهای وب در نظر گرفتن سـرشت وب و اندیشـیدن درباره آن به عنوان منابعی برخط که باهم پیوند دارند مهم .است آمارها ،بایـد هم منـابعی را در بر بگیرنـد که برای کاربران انسانی در نظر گرفته شـده انـد و هم فراداده های همراه آنها و برنامه هایی که بخش های جدایی ناپذیر آرشیوی وب هستند.

درک این نکته نیز بسیار مهم است که همه کدهای وضعیت که خدمتگرهای وب

باز می گردانند معتبر یا قابل اعتماد نیستند در زیر نمونه هایی ذکر شده اند.

(الف) گمشدگی کد 404

خدم محمد

بعضی از خدمتگرهای وب زمانی که منبعی در دسترس نیست که وضعیت 404 را به درستی بازنمی گردانند خدم این شرایط ممکن است به جای کـد مذکور پاسخ OK 900 را با محتوایی مسدود و توصـیح اینکه منبع درخواستی وجود ندارد .بفرسـتد خزشگر آرشیوگر به هیچ طریقی نمی توانـد بفهمد که این را باید کد « Not Found 404 » لحاظ کنـد بنابراین پاسخ را « خوب » یا « موفقیت » آمیز محسوب می کند .

ب)تکرارهایی با شناسه نشست (3) متفاوت

بسـیاری از خدمتگرهای وب به طور پویا به تولید نشانی وب میپردازند که منجر به بروز منابع تکراری در آرشیو وب می شود . گاهی اوقات خدمتگر وب هنگام تحویل منابع به عامل کاربری (4) یک شناسگر یکتا به هر نشانی وب الصاق میکند تا رد یک نشست را حفظ کند مثل

عامل کاربری 1 یک نشانی وب را که به نشانی زیر شبیه است میگیرد

http// :www.example.com/id/12345-picture.jpg

عامل کاربری ،2 یک نشانی وب را با شناسه متفاوت میگیرد

http// :www.example.com/id/67890-picture.jpg

هر دو نشـانی به یـک منبع ختم میشونـد که در این مثـال یک تصویر jpg است ولی شناسـگرهای متفاوتی دارند در این صورت خزشـگر آرشیوگر می بایست یک منبع را با

Documents -1

Publications –2

Session –3

User agent –4

نشانی های متفاوت بارها گردآوری می کرد . این منابع مکرر یک منبع واحد محسوب می شوند .

ج فقدان که وضعیت

در ابتدای پیدایش ،وب خدمتگرها معمولا فقط منابع درخواستی را بدون کد وضعیت فراداده ارسال می کردند .

از این پدیـده گاهی به صورت HTTP ... یاد می شود. بعضـی از خدمتگرها ممکن است هنوز از پروتکل قدیمی اسـتفاده کنند فقدان که وضعیت فراداده میتواند به طور ویژه برای موسساتی که دارای مجموعه های و بی تاریخی قدیمی هستند موضوعیت داشته باشد.

4,2,4

اندازه گیری یک آرشیو :وب محاسبه دامنه ها یا خدمتگرها

1,42,4

هدف

شمارش دامنه ها یا خدمتگرها شـیوه ای گویا در اندازه گیری آرشیو وب است. موسسات گردآورنده عموماً از این روش استفاده می.کنند بر خلاف ،این یک وبگاه واحـدی مفهومی و فکری است که از مجموعه ای از صـفحات وب تشـکیل میشود که به هم متصل انـد و هر کدام نماینده یک شخص گروه یا سازمان هستند که از حیث فنی قابل تعریف نیست و اجازه اندازه گیری

سیستماتیک یا محلی را نمی دهد .

نامه ای دامنه و خدمتگرها به گونه ای نظام مند قابل اندازه گیری اند . ولی نباید آن ها را معادل خود وبگاه ها در نظر گرفت؛ زیرا صـرفا برای نامیدن و مکان یابی وب سایت ها مورد اسـتفاده قرار می.گیرند این آمارها برای فهم سـرشت مجموعه به گونه ای ،تفصیلی یا تحلیل فنی به منظور تعیین نوع منابع و بی آرشیو شده مناسباند مانند com یا org یا سنجش اینکه آیا خزشگر دامنه مورد نظر را تعقیب کرده است یا خیر

2.4.2.4

روش

تعـداد دامنه ها یا خدمتگرها به طور خودکار از گزارش های خزش یا به روش های خودکار دیگر که می توان فایل های ذخیره شده را با آن ها تحلیل کرد محاسبه می شوند .

3.4.2.4

محدودیت ها

همان طور که در شـمارش نشانی های وب محـدودیت هایی وجود دارد در شـمردن دامنه ها یا خـدمتگرها هم تمایلی در به دست آوردن تعداد بیش از اندازه منابع قابل خواندن برای انسان در یک آرشـیو وب وجود دارد همه دامنه ها به منابع فعال یا معنادار منتهی نمی شوند نام های مستعار (1) و مکررها نیز از موارد دیگر است.

(الف) دامنه های غیر فعال

ممکن است دامنه خریـداری شود ولی غیر فعال بماند و به هیچ منبعی باز نشود دامنه های توقفگاهی (2) نیز وجود دارند که حاوی منابعی هسـتند ولی اغلب از یک صـفحه وب تشـکیل شده اند که نام دامنه برای فروش ارائه می کند . اولین دسـته با کد وضـعیت 204 شناسایی می.شود دومی به معنـای واقعی یـک غیر فعال نیست و هنگام شـمارش دامنه ها به حساب میآیـد با وجود این چنین منابعی از دیـد متولی آرشیو حاوی ارزش مهمی تلقی نمی شوند.

در وضعیت برداشت انتخابی میتوان این دامنه ها را به طور مؤثر در اثنای فرایند گزینش از دور کار خارج نمود و آن ها را برداشت نکرد ولی روش آسانی برای شناسایی و پالایش خودکار این دسته از منابع از برداشت پشته ای وجود ندارد به جز اینکه منابع با اندازه بسیار کوچک به طور دستی

وارسی شوند.

در صورت ،امکان پیشنهاد می شود که دامنه های غیر فعال با نمونه گیری در آرشیو وب ردگیری شوند تا بتوان سرشت مجموعه را مشخص و کارآمدی تضمین کیفیت را ارزیابی کرد

ب نامهای مستعار

نام دامنه ،مسـتعار جایگزینی برای نام دامنه است . نام دامنه ،مستعار میزبانی یک وبگاه را ر یک دامنه ممکن و سایر نام دامنه ها را به آن باز می کند .

نام های مسـتعار بسیاری در اینترنت وجود دارد زیرا ممکن است صاحبان دامنه بخواهند برای آنکه تا سر حد امکان برای کاربر رویت (3) پذیر شوند از نام دامنه های متعددی برای یک محتوا استفاده کنند نام های مستعار عمدتاً از طریق تغییر مسیرها اجرا می شوند .

در گزارشـهایی که خزشگرهای آرشیوگر تولید می کنند نام های مستعار دامنه های منحصر به فردی هستند که به منابع واحدی باز می شوند .

Aliases -1

Parked domains −2

Visible −3

کشف (1) نام های مستعار مستلزم مقایسه چشمی میان صفحه هایی است که از یک خدمتگر میآیند و با مقایسه چک سام ها (2) میسر است نام های مستعار را راحت تر میتوان کشف کرد در آرشیو انتخابی کمتر وجود دارند و بیشتر احتمال دارد در برداشت پشت های وارد آرشیو شوند بودن آن ها نشان می دهد که در آرشیو منابع تکراری وجود دارد پیشنهاد می شود در صورت امکان نام های مستعار کشف شده در آرشیو وب ردگیری شوند تا بتوان به خاص بودن مجموعه و مکررزدایی آن کمک کرد

5.2.4

اندازه گیری یک آرشیو وب شمارش بایت ها

1.5.2.4

هدف

آرشیو وب را از طریق بایت های آن هم میتوان اندازه گیری کرد که آمار مفیدی است که به برنامه ریزی برای ذخیره سازی و سایر منابع کمک کند این کار با سنجش طول خطی مخازن یک کتابخانه بر حسب متر یا مایل برای مقاصد مدیریتی قابل مقایسه است.

205.2.4

روش

اندازه یک آرشیو برحسب ،بایت به صورت خودکار با افزودن اندازه منابع خزش شده برگرفته از گزارش های خزش یا با سایر روش های خودکار که فضای اشغال شده دیسک را در آرشیو وارسی میکنند برآورد می شود .

آرشیوهای وب اغلب بزرگ مقیاس اند . اندازه آن ها از یک مجموعه کوچک شامل چند صد گیگابایت تا مجموعه های ملی چند صد فرابایتی متغیر است اندازه آرشیوهایی که وب را در مقیاس جهانی گردآوری می کنند و دیرپا هستند حتی به پتابایت هم میرسد روش متداول در ذخیره آرشیو وب فشرده سازی (3) داده هاست به عنوان مثال ضمیمه D در ایزو 28500 ویژگی های قالب فایل WARC که قالب آرشیوی استاندارد برای آرشیوهای وب است چگونگی استفاده از روش فشرده سازی GZIP را برای قالب WARC بیان می.کند اندازه یک آرشیو وب را میتوان برحسب داده فشرده یا غیر فشرده تعیین کرد. باوجوداین هنگام مقایسه

Detection -1

Checksum -2

Compress -3

آرشـیوها اسـتفاده از معیارهای واحد اهمیت دارد بدین معنی که نباید اندازه داده فشـرده یک آرشیو را با داده غیر فشرده آرشیو دیگر مقایسه کرد.

اندازه آرشـیو در حالت فشرده نمایانگر فضای اشغال شده دیسک است. این اندازه مرجع برای تمهیدات ذخیره سازی و حفاظت بلندمدت است؛

اندازه غیر فشرده نمایانگر حجم منابع به همان صورتی است که در وب زنده بوده اند. این آمار برای اطلاع کاربران انسانی تهیه می شود و از منظر محتوا یا ابعاد فکری معنادار است. اندازه گیری آرشیو وب را همچنین میتوان به عللی که در 2030204 ذکر شد پیش یا پس از

مکرر زدایی انجام داد اما باز روش محاسبه باید به طور روشن ذکر شود.

رویه معمول این است که فایـل ها را در فایل های حاملی ماننـد ARC و WARC ذخیره می کننـد . حامل گذاری (1) فایل ها را باهم مجتمع می کند . ذخیره و رسـیدگی به چند فایل بزرگ در مقایسه با تعداد زیادی فایل کوچک آسان تر است فایل های حامل (2) معمولاً ذخیره فراداده را همراه با منابع برداشت شده میسر می.کنند تعداد فایل های حامل نیز آمار مفیدی برای آرشیو

وب ،است چراکه اغلب به عنوان پایه ای ترین واحدهای مدیریت در اهداف ذخیره تبادل داده

گاهی حفاظت بلند مدت به کار می روند

6.2.4

آمارهای هسته برای گسترش مجموعه

جدول –2 آمارهای هسته برای گسترش مجموعه

آرشیوها استفاده از معیارهای واحد اهمیت دارد، بدین معنی که نباید اندازهٔ دادهٔ فشردهٔ یک آرشیو را با دادهٔ غیرفشردهٔ آرشیو دیگر مقایسه کرد.

- اندازهٔ آرشیو در حالت فشرده نمایانگر فضای اشغال شدهٔ دیسک است. این اندازهٔ مرجع برای تمهیدات ذخیره‌سازی و حفاظت بلندمدت است؛
- اندازهٔ غیرفشرده نمایانگر حجم منابع به همان صورتی است که در وب زنده بوده‌اند. این آمار برای اطلاع کاربران انسانی تهیه می‌شود و از منظر محتوا یا ابعاد فکری معنادار است.

اندازه‌گیری آرشیو وب را همچنین می‌توان به عللی که در ۲.۴ .۲.۳ ذکر شد پیش یا پس از مکررزدایی انجام داد، اما باز روش محاسبه باید به‌طور روشن ذکر شود.

رویهٔ معمول این است که فایل‌ها را در فایل‌های حاملی مانند ARC و WARC ذخیره می‌کنند. حامل‌گذاری[1] فایل‌ها را باهم مجتمع می‌کند. ذخیره و رسیدگی به چند فایل بزرگ در مقایسه با تعداد زیادی فایل کوچک آسان‌تر است. فایل‌های حامل[2] معمولاً ذخیرهٔ فراداده را همراه با منابع برداشت شده میسر می‌کنند. تعداد فایل‌های حامل نیز آمار مفیدی برای آرشیو وب است، چراکه اغلب به‌عنوان پایه‌ای‌ترین واحدهای مدیریت در اهداف ذخیره، تبادل داده و گاهی حفاظت بلندمدت به کار می‌روند.

۶.۲.۴

آمارهای هسته برای گسترش مجموعه

جدول ۲- آمارهای هسته برای گسترش مجموعه

آمارها	اهداف	نمونه ها
تعداد منابع هدف	اهداف گردآوری/ خروجی‌های کمّی	۸۰۰۰ هدف
تعداد اهداف گیرانداخته‌شده	اهداف گردآوری/ خروجی‌های کمّی	۱۴ هزار هدف گیرانداخته‌شده
تعداد نشانی‌های وب (پیش و پس از مکررزدایی)	خروجی‌های کمّی	۱۴ میلیون نشان وب برداشت‌شده، ۱۰ میلیون پس از مکررزدایی

1. Containerisation
2. Container files

ص: 63

Container files -1

Containerisation -2

3.4. سرشت نهایی مجموعه

عکس

توزیع نشانی های وب برحسب کدهای وضعیت	تعداد منابع برحسب نوع	۲ میلیون منبع که با موفقیت خزش شده اند (کد ۲۰۰)
تعداد دامنه‌ها یا میزبان‌ها	خروجی‌های کمّی	۳ میلیون نام دامنه
اندازه برحسب بایت (غیرفشرده و فشرده، پیش و بعد از مکررزدایی)	خروجی‌های کمّی	۲۰۰ ترابایت غیرفشرده پیش از مکررزدایی، ۱۶۰ ترابایت فشرده و پس از مکررزدایی
تعداد فایل‌های حامل WARC یا هر حامل دیگری	خروجی‌های کمّی	۱۸ هزار فایل WARC

۳.۴
سرشت نهایی مجموعه
۱.۳.۴
کلیات

آمارهایی که در این بخش پیشنهاد می‌شوند ویژگی‌های آرشیوهای وب را توصیف می‌کنند، در تعیین دامنهٔ آنها یاری می‌رسانند و منجر به تصمیم‌گیری‌های آگاهانهٔ متولیان آرشیو می‌شوند. بعضی از آمارها مختص برداشت انتخابی یا پشته‌ای هستند، و بعضی دیگر عمومی‌اند و برای آرشیوهای وبی که با استفاده از هردو استراتژی ایجاد شده‌اند قابل کاربرد هستند.

اندازهٔ آرشیوهای وب عموماً مانعی برای شمارش دستی است. بعضی از آمارها – به‌ویژه آنها که به برداشت پشته‌ای مربوط می شوند– فقط با نمونه گیری به دست می‌آیند. گردآوری دستی آمارها برای آرشیو انتخابی نیز میسر است، اما صرف منابع برای آن باید توجیه داشته باشد.

۲.۳.۴
آمارهای متداول
۱.۲.۳.۴
توزیع برحسب دامنهٔ سطح بالا یا سطح دوم
۱.۱.۲.۳.۴
هدف

دامنه‌های سطح بالا (TLDS) نشانگر توزیع جغرافیایی منابع در یک آرشیو وب هستند. این

13.4

کلیات

آمارهایی که در این بخش پیشنهاد می شوند ویژگی های آرشیوهای وب را توصیف می کنند در تعیین دامنه آن ها یاری میرسانند و منجر به تصمیم گیری های آگاهانه متولیان آرشیو می شوند . بعضی از آمارها مختص برداشت انتخابی یا پشت های هستند و بعضی دیگر عمومی انـد و برای آرشـیوهای وبی که با اسـتفاده از هر دو استراتژی ایجاد شـده اند قابل کاربرد هسـتند. اندازه آرشـیوهای وب عموماً مانعی برای شمارش دستی است .

بعضـی از آمارهـا – به ویژه آن ها که به برداشت پشت های مربوط می شونـد فقط با نمونه گیری به دست می آینـد. گردآوری دستی آمارها برای آرشیو انتخابی نیز میسر است اما صرف منابع برای آن باید توجیه داشته باشد .

2.3.4

آمارهای متداول

1.2.3.4

توزیع برحسب دامنه سطح بالا یا سطح دوم

1,1,2.3.4

هدف

دامنه های سطح بالا (TLDS) نشانگر توزیع جغرافیایی منابع در یک آرشیو وب هستند این

ص: 64

آمار برای کتابخانه ها و آرشیوهای ملی که ماموریت حفاظت کل خروجی فکری یک کشور را بر عهده دارند اهمیت ویژه ای دارد دامنه های سطح دوم که برای دسته های ویژه ای از سازمان ها از حیث حوزه های مورد علاقه در نظر گرفته شده نیز ،مفیدند چراکه گستردگی منابع در یک آرشیو را آشکار می سازند به عنوان نمونه منابع ذیل نام دامنه .uk.gov را یک سازمان دولتی انگلستان در وب منتشر می کند .

2.1.2.3.4

روش

توزیع دامنه های سطح بالا ـ و دامنه های سطح دوم را میتوان به طور خودکار با استفاده از گزارش هایی که خزشگرهای آرشیوگر تولید میکنند یا سایر روش های خودکار در تحلیل دامنه ها محاسبه کرد اندازه ها را میتوان با اعداد مطلق یا درصد نشان داد فهرست کردن 5 یا 10 نام دامنه سطح بالا که بیشترین رخداد 1 را در آرشیو وب دارند نیز مفید است. تعداد یا درصد نام دامنه سطح بالای ملی گردآوری شده 70 درصد از نشانی های موجود در آخرین خزش دامنه کتابخانه ملی فرانسه در نام دامن میزبانی می شوند

3 درصد همان نشانی ها در نام دامنه de میزبانی می شوند تعداد یا درصد دامنه های سطح دوم گردآوری شده 1/5 درصد از نشانی های وب موجود در آخرین خزش دامنه کتابخانه ملی فرانسه بر نام دامنه های gouv.fr میزبانی می شوند و منابعی هستند که موسسات دولتی فرانسه منتشر کرده اند.

آرشیو وبی که نسبت به سایر دامنه ها دارنده مقادیر بزرگتری از نام دامنه های ملی باشد .

آرشیوی دارای دامنه ملی محسوب میشود.

3.1.2.3.4

محدودیت ها

بعضی از موسسات منابعی را که خارج از نام دامنه سطح بالای ملی شان میزبانی شود در دامنه کار خود قرار می دهند برای مثال www.lego.com یک شرکت (دانمارکی یک وبگاه دانمارکی محسوب می شود) هرچند از نام دامنه ای غیر از dk استفاده می.کند این نشان می دهد که نام

؟؟؟؟؟ (1)

1- Occurring

دامنه های سطح بالا همیشه برای تعیین دامنه یا مرزهای یک دامنه ملی کافی نیستند.

2.2.3.4

توزیع برحسب مقدار (1) منابع بر حسب دامنه (و) یا (میزبان)

1.2.2.3.4

هدف

تحلیل و ارائه گزارش درباره اندازه منابعی که در هر دامنه میزبانی میشوند و یا میزبان هستند و چگونگی توزیع آن در آرشیو وب ضمن آنکه بینشی درباره سرشت مجموعه به دست می دهد به مدیریت فرایند خزش نیز یاری میرساند توزیع اندازه بر حسب نامه ای دامنه وا یا میزبان ها در آرشیو میتواند ویژگی های منابعی را آشکار کند که در انواع دامنه ها میزبانی می.شوند ،همچنین توان آرشیو در گردآوری منابع با اندازه های مختلف را نشان می دهد به ویژه وبگاه های پرحجمی که از حیث فنی خزش آن ها دشوار است.

دسته بندی منابعی که اندازه های مختلفی دارند برحسب دامنه و / یا میزبان همچنین به پیکربندی و ساماندهی فرایند خزش کمک می.کند دسته بندی و خزش دامنه های با اندازه مشابه در فرایندها یا کارهای (2) جداگانه مرسوم است زیرا این کار تنظیمات مشابهی می طلبد زمان مشابهی برای تکمیل باید آن صرف کرد این کار باعث می شود از منابع ماشینی بهترین استفاده صورت گیرد و پایش و مدیریت وظیفه ها (3) را آسان می کند.

2.2.2.3.4

روش

مقدار منابع برحسب دامنه و / یا میزبان را میتوان با واحد MB/GB/TB اندازه گرفت

- 10MB>

- 100MB>

- MB 101-999

- 1GB

- 1GB<

Job –2

Task –3

به جای آن می توان تعداد نشانی های وب را در هر دامنه شمرد

- <10000URL

- URL 50000-100000>

- URL 100000<

تنهـا راه معلوم کردن مقـدار منابع در هر دامنه در وب زنـده اسـتفاده از تصاویری است که موتورهای جست وجو ارائه می.کننـد برای یک آرشیو وب آمارهای مذکور در بالا با محاسبه خودکار گزارش های خزش یا با سایر روش های خودکار در تحلیل فایل های ذخیره شده به

دست می آید.

3.2.2.3.4

محدودیت ها

همانند بعضی آمارهایی که در بخش های پیشین پیشنهاد شدند سطحی از تقریب به این آمار منضم می شود انجام این اندازه گیری در طول زمان مفیدتر ،است زیرا میتوان برمنیای آن ها

مقایسه هایی انجام داد.

3.2.3.4

توزیع بر اساس نوع قالب

1.3.2.3.4

هدف

تحلیل و گزارش دهی درباره توزیع قالب های فایل در یک آرشـیو وب فعالیتی کلیـدی در حفاظت دیجیتالی و نیز عنصـری از ویژگی های آرشیوی است به منظور پایش و مدیریت خطرات حفاظتی مربوط به کهنگی ،قالب دانستن انواع فایل موجود در آرشیو مهم است اطلاعات مربوط به قالب ها با رده بنـدی فخیم و معظمی که به طور سـنتی در کتابخانه ها طبقه بندی انواع ،انتشارات نظیر تصویر ،فیلم صوت انجام می شد قابل مقایسه است گردآوری این اطلاعات در طول زمان میتواند گرای شهای فناورانه را آشکار سازد و به درک ما از تحول وب یاری

می رساند.

203.2.3.4

روش

آمار قالب ها به طور خودکار از گزارش های خزش یا روش های خودکار دیگر در تحلیل فایل های ذخیره شده به دست میآید.

توزیع قالب ها ممکن است با محاسبه و ساماندهی به روش های مختلفی صورت گیرد .

به عنوان نمونه

از طریق انواع منـابع 7 درصـد از فایـل ها(متن) مثل (15 ،(HTML درصـد تصویر مثل jpg و (3 (gif درصـد صوتی (مثل mpeg) هستند.

از طریق متداول ترین قالب های فایل بالاترین 50 یا 100 مثلاً HTML فراوان ترین قالب فایل است.

از طریق ردگیری بعضـی قـالب ها که از اولین شـکل پیـدایش آن ها در آرشـیو آغاز میشود و کاهش یا افزایش فراوانی را در طی زمان نشان بدهد؛

از طریق قالبی که کمترین فراوانی را دارد مثلاً اگر قالب های ویـدئویی در یک آرشـیو وب کمتر از وب زنده وجود دارند ممکن است نشانه ای از بازنمایی اندک آن ها در آرشیو باشد .

3.3.2.3.4

محدودیت ها

تحلیل رخ نمای (1) قالب در یک آرشیو وب به طور طبیعی منجر به ایجاد فهرست بلندی از هزارها قالب فایل می شود .

پیشـنهاد می شود که آمار بر 50 یا 100 قالبی که رایـج ترند معطوف ،شود مگر اینکه علاقه ای به قالب غیر متداول خاصـی وجود داشـته باشـد قالب های منابع خزش شده (انواع فایل) که خدمتگر وب گزارش کرده و در گزارش های خزش ثبت شده است همیشه قابل اعتماد نیستند ممکن است خدمتگرهای وب انواع قالب را به اشتباه برگردانند. بعضی ،موسسات از ابزارهای افزوده مخصوص تشخیص قالب برای کسب اطلاعات صحیح تر

استفاده می کنند .

Profile -1

4.2.3.4

ویژگی زبان

1.4.2.3.4

هدف

تحلیل زبان های به کار رفته در منابع آرشیو ،شده دیدگاه هایی درباره الگوهای زبان شناختی یک آرشیو وب به دست می دهد و برای درک ویژگی های یک مجموعه ملی از جمله تنوع آن یا قرابت های فرهنگیاش با سایر کشورها مفید است. ویژگی زبانی برای ملت هایی که به

زبان های منحصر به فردی تکلم می کنند مهم تر است زیرا می تواند به شناسایی منابعی در وب

که مرتبط به این کشورها باشد یاری رساند.

در دامنه های ملی منابعی به زبان های بیگانه هستند مانند باسک در دامنه های فرانسه عربی در دامنه های دانمارکی ویژگی زبانی می تواند به تحلیل موضوعات متنوع اجتماعی و فرهنگی و بازتاب آن ها در وب کمک کند.

2,4,2,3.4

روش

روش ها یا فناوری استانداردی برای کشف خودکار زبان در منابع آرشیو وب وجود ندارد. ساختارهای ویژه یک زبان طبیعی را میتوان برای کمک به تحلیل و شناسایی منابعی که به آن زبان نوشته شده اند به کار گرفت همچنین میتوان از عناصر مرتبط با زبان در سرآیندهای HTML و HTTP – درصورتی که در دسترس باشند استفاده کرد ابزارهای کشف زبان یا پردازش زبان طبیعی نیز وجود دارند که می توان از آن ها هم برای این منظور استفاده کرد. توزیع زبان های طبیعی در آرشیو وب را میتوان با اعداد مطلق مانند جمع کل صفحاتی که به یک زبان هستند؛ یا با درصد مثل درصد صفحاتی که از زبان خاصی هستند نشان داد نیز ارائه فهرستی از 5 یا 10 زبانی که بیشتر کاربرد دارند مفید است .

3.4.2.3.4

محدودیت ها

بسیاری از منابع فاقد فراداده مربوط به زبان هستند که شناسایی خودکار زبان هایی را که برای ساختی منابع مورد استفاده قرار گرفته است دشوار می.کند علاوه بر این برنامه های

ترجمه خودکار که به کاربر امکان میدهند صفحه وب را به زبان های مختلف ببیند این آمار را

دستخوش تحریف میکنند

5.2.3.4

پوشش زمانی

1.5.2.3.4

هدف

سرشت آرشیو وب با پوشش زمانی آن نیز نمایانده می شود که ارجاع به بازه زمانی خاصی است که طی آن منابع گردآوری شده اند فرض کلی این است که هر چه بازه زمانی طولانی تر باشد احتمال اینکه منابع اصلی که رونوشت آن ها در آرشیو وب وجود دارد از وب زنده ناپدید شده باشند بیشتر است. این امر به ارزش آرشیو میافزاید چراکه ممکن است پیشینه های تاریخی مختصر به فردی از دوره های خاص یا حوادث تاریخی را در خود داشته باشد. پوشش زمانی آرشیو حاوی اطلاعات مهمی است که برای برنامه ریزی حفاظت ضروری است آرشیوهایی که زمان های طولانی تری را پوشش می دهند بیشتر در معرض خطر کهنگی قرار دارند.

2.5.2.3.4

روش

پوشش زمانی یک آرشیو وب با شمارش سال های بعد از تاریخی که اولین منبع گیر انداخته شده است به سادگی قابل اندازه گیری است این آمار را میتوان با سایر آمارها نظیر اندازه آرشیو یا توزیع قالب های فایل ترکیب کرد و روند تشکیل آرشیو یا توسعه آن را در طول زمان نشان داد

3.3.4

آمارهای آرشیو انتخابی

آمارهای زیر مربوط به آرشیوهایی هستند که با برداشت انتخابی گردآوری شده اند و در آن ها

قابل کاربردند.

1.3.3.4

مجوزها

1.1.3.3.4

هدف

هنگـامی که اخـذ مجوز از ناشـران یـا صاحبـان حقوق پیش از برداشت داده از وبگاه هایشان لازم ،باشـد آمار مجوزها شاخص مهمی از کارآمدی این گردش کار خواهد بود.

در مقایسه با هزینه های نگهـداری سامانه مدیریت ،مجوزها آمار مجوزها به عنوان نمونه میتواند هزینه اخذ موفقیت آمیز هر واحد مجوز را نشان دهد تعداد مجوزها نیز بازتاب علاقه ناشران به آرشیو وب و آگاهی عمومی آن ها از آن است.

2.1.3.3.4

روش

تعداد مجوزهای اعطا شده و تعداد درخواست های ارسالی به صاحبان حقوق برای اخذ مجوز را می توان شمرد این کار را میتوان دستی یا با استفاده از تمهیدات خودکار تعبیه شده در سامانه مدیریت مجوزها یا ابزار آرشیوگر انجام داد

2.3.3.4

نامزدی

1.2.3.3.4

هدف

نامزدها منابعی هسـتند که برای ورود در آرشـیو وب انتخاب و پیشـنهاد می شونـد و از طیف گسترده ای از منابع منشا می گیرنـد؛ از جمله هواداران کاربران و حامیان آرشیو وب، و عموم .مردم بعضی از آرشیوهای وب فعالانه نامزدها را با یک هیئت تحریریه شبکه های اجتماعی یا نامزدی برخط تقاضا می کنند در فرایند ،انتخاب با توجه به سیاست گسترش مجموعه تعیین می شود که آیا یک نامزد در دامنه آرشیو وب قرار می گیرد یا خیر گزینش را معمولاً کتابداران موضوعی، دیجیتال آرشیوگران یا متصدیان آرشیو انجام می دهند.

تعداد نامزدی ها تاثیر فعالیت ها و آگاهی رسانی آرشیو وب میان ذینفعان (1) را آشکار می کند و میتوان از آن به عنوان راهنمای فعالیت های نامزدی استفاده کرد این آمار همچنین ثمربخشی و کارآمدی مربوط به فرایند گزینش را اندازه می گیرد و میتواند برای ارزیابی عملکرد و تلاش فرد فرد متصدیان آرشیو به کار رود به عنوان مثال تعداد نامزدی ها برحسب هر (متصدی) در صورتی که موسس های استراتژی آمیخته برداشت پشت های و انتخابی را به کار گرفته باشد مقایسه هردو میتواند با تاکید بر هزینه های مربوط به گزینش دستی در مقابل رویکرد اتوماتیک برداشت پشت های صورت گیرد آمارهای بیشتری برای بررسی نحوه استفاده از آرشیو نیز میتوان استخراج ،کرد ضمن آنکه محتوایی که متصدی انتخاب میکند بیشتر مورد استفاده کاربران قرار می گیرد؛ ،پس از محتوایی که به طور خودکار انتخاب می شود ارزشمندتر است.

2.2.3.3.4

روش

آمار تعداد نامزدی ها را میتوان به صورت دستی یا خودکار با استفاده از نرم افزار آرشیوگر

گردآوری کرد.

3.2.3.3.4

محدودیت ها

همه نامزدی ها منجر به هدف گزینی (2) یا گیر انداختن محتوا نمی شود. فاکتورهای بسیاری بر خروجی تاثیر می گذارند ممکن است اخذ مجوز لازم برای آرشیو کردن وبگاه با موفقیت همراه نباشد؛ ممکن است خزشگر به عللی حتی قادر به گیر انداختن محتوا نباشد؛ ممکن است موسسه ای نتواند هزینه برداشت همه منابع برگزیده را به علت محدودیت های مالی تامین کند. حتی در موردی که منبعی با موفقیت برداشته شده است مسائل مربوط به دسترسی می تواند ارائه آن را از آرشیو وب ناممکن کند آمار مذکور بیشتر تلاش های فرایند انتخاب را اندازه می گیرد تا خروجی آن را

Stakeholders -1

Targeting -2

3.3.3.4

موضوعات

13.3.3.4

هدف

پوشش موضوعی یک آرشیو وب ویژگی های محتوای آن را مشخص میکند و برای گسترش مجموعه حائز ارزشی فوق العاده تلقی می شود؛ چراکه به ایجاد رخ نمای مجموعه کمک میکند و آن را متعادل میسازد و شکاف های (1) ممکن در محتوا را آشکار می.کند پوشش موضوعی را همچنین میتوان برای درک اینکه آیا یک آرشیو وب پاسخگوی نیازهای پژوهشگران رشته های مختلف هست یا خیر و چگونه به کاربرد

2.3.3.3.4

روش

راه های مختلفی برای استخراج اطلاعات مربوط به موضوع وجود دارد. در بعضی از آرشیوهای وب به صورت دستی عبارتی به عنوان موضوع به منابع داده می شود که بعداً قابل دسته بندی و پالایش و تحلیل خواهد بود استفاده از نظام های رده بندی استاندارد مانند رده بندی دیویی و سرعنوان های موضوعی کتابخانه کنگره آمریکا برای توصیف منابع آرشیو شده متداول است.

در بیشتر آرشیوهای ،وب واژگان موضوعی به صورت دستی افزوده نمی شوند در صورتی که اطلاعات مربوط به موضوع منابع موجود باشد با استخراج فراداده جاسازی شده در منابع به دست میآید؛ مانند مقادیر متاتگ « Keyword در سرآیند HTTP یا از ناحیه subject » در دوبلین کور (2) همچنین می توان روش خوشه بندی (3) یا دسته بندی خودکار را برای تحلیل معناشناختی منابع آرشیو وب به کار برد و آن ها را بر اساس فرانمای موضوعی به گونه خودکار رده بندی کرد در حال حاضر، جامعه آرشیوگران وب توافق یا روش یکسانی برای این کار ندارند درصد منابع در هر موضوع به درک توزیع موضوعی محتوای آرشیو وب کمک میکند تلاش ،بعدی پرداختن به موضوعاتی است که به قدر کافی منعکس نشده اند همچنین میتوان از اطلاعات موضوعات در ترکیب با سایر آمار و اطلاعات استفاده کرد مثلاً با ترکیب موضوع با فراداده منشا میتوان رابطه میان حوزه های موضوعات و نوع ناشران را دریافت

Gap -1

Dublin core -2

Clustering -3

3.3.3.3.4

محدودیت ها

با موضوع دادن به صورت دستی منابع آرشیو شده با دقت بیشتری توصیف یا رده بندی می شوند ولی اجرای آن پرهزینه است استخراج خود فراداده جاسازی شده نیز مستلزم این است که فراداده موجود باشد. خوشه بندی یا رده بندی خودکار حیطه ای توسعه نایافته در آرشیو وب و از حیث فنی به خصوص برای آرشیوهای وب بزرگ مقیاس چالش برانگیز است.

4.3.4

آمارهای هسته برای توصیف ویژگی های مجموعه

جدول –3 آمارهای هسته برای سرشت نمایی مجموعه

۳.۳.۳.۳.۴

محدودیت ها

با موضوع دادن به‌صورت دستی، منابع آرشیو شده با دقت بیشتری توصیف یا رده‌بندی می‌شوند؛ ولی اجرای آن پرهزینه است. استخراج خود فرادادهٔ جاسازی‌شده نیز مستلزم این است که فراداده موجود باشد. خوشه‌بندی یا رده‌بندی خودکار، حیطه‌ای توسعه‌نایافته در آرشیو وب و از حیث فنی به‌خصوص برای آرشیوهای وب بزرگ‌مقیاس، چالش‌برانگیز است.

۴.۳.۴

آمارهای هسته برای توصیف ویژگی‌های مجموعه

جدول ۳- آمارهای هسته برای سرشت‌نمایی مجموعه

آمار	هدف	نمونه
توزیع برحسب دامنهٔ سطح بالا یا سطح دوم	توزیع جغرافیای	۱۰ بیلیون نشانی وب در دامنهٔ سطح بالای fr در آرشیو
توزیع برحسب مقدار منابع در دامنه	تحلیل دامنه	۲ میلیون دامنه کمتر از ۱۰ نشانی وب دارند، درحالی‌که ۱۵۰هزار نام دامنه میزبان بیش از ۱۰هزار نشانی وب هستند.
توزیع برحسب نوع فایل	ویژگی قالب	۵۰۰ میلیون نشانی وب در آخرین خزش پشته ای در قالب HTML هستند
پوشش زمانی	تحلیل زمانی	آرشیو، منابعی از سال ۱۹۹۶ را در خود دارد
تعداد مجوزهای اعطاشده	بهره‌وری	۲۰ هزار درخواست مجوز منجر به اعطای مجوز از طرف ناشران شده است
تعداد نامزد	بهره وری	۱۰۰۰ نامزدی جدید در سال افزوده شده‌اند

4.4. استفاده از مجموعه

1.4.4

کلیات

همـان گونه که در بخش 203 دسترسـی و روش های توصـیف گفته ،شد شـرایط توصـیف و استفاده از منابع آرشیو شده اینترنت برحسب قوانین ملی و سـیاست هـای سازمـانی متفـاوت است و تمایز میان آرشـیوهای سـفید خاکستری و تیره می توانـد برای انتخاب روش های مناسب اندازه گیری میزان استفاده مفید واقع شود.

روش ها و اسـتانداردهای تهیه آمار در مورد میزان اسـتفاده از آرشـیو خاکستری همان هایی هستند که برای ارزیابی میزان استفاده از منابع الکترونیکی در کتابخـانه هـا به کـار میرونـد در نتیجه در این گزارش ،فنی در عین ارائه اطلاعات فنی و تعاریف فنی بیشتر برای ایجاد درک بهتری از موضوعات مربوط به دسترسی و استفاده از مجموعه های آرشیو غالباً به روش ها و استانداردهای موجود ارجاع داده می شود .

2.4.4

تعاریف و روش های اندازه گیری استفاده

الف) مراجعه کنندگان حضوری

ممکن است الزامـات قـانونی و سـیاست سازمـانی اسـتفاده از آرشـیو وب را به مکان های جغرافیایی ویژه با دسترسـی کنترل –شـده مانند تالارهای مطالعه در کتابخانه ها و آرشـیوهای ملی محدود .کند در این موارد داده های اولیه مربوط به دسترسـی را میتوان از کاربران آن ها هنگامی که ثبت نام می کننـد و از داده های ورود به سامانه گردآوری کرد با این کار اطلاعات اسـتفاده با بالاترین کیفیت به دست می آید؛ چراکه میتوان مستقیماً از قصد کاربران سوال کرد.

ب) مراجعه کنندگان مجازی

آرشـیوهای وبی که برای عموم دسترس پذیرند می توانند آمارهای استفاده را با به کار بردن ابزارهای تحلیل وب (1) گردآوری کنند درست مانند آنچه « انجمن تحلیل وب » (2) در قالب اسـتاندارد درآورده است تحلیل ،وب پهنه ای رو به رشد در زمینه تحلیل الگوهای استفاده از وبگاه ها با روش های زیر است

1- Web Analytics

2- Web Analytics Association

بازدیدهای مجازی در ایزو 2789 به عنوان یک دور کامل از استفاده کاربرانی از وبگاه کتابخانه است که نشانی آن ها خارج از فضای IP کتابخانه است اغلب بیرون از محوطه کتابخانه بدون در نظر گرفتن تعداد صفحات یا عناصر دیده شده

یادداشت : یک بازدیدکننده وبگاه یا یک مرورگر و بی شناخته شده است یا یک نشانی

IP قابل شناسایی که به صفحه از طریق پایگاه یا کتابخانه دسترسی دارد. یادداشت :2 فاصله میان دو درخواست پیاپی معمولا طولانی تر از وقفه زمانی یا 30 دقیقه نیست. اگر این ها قرار است به عنوان بخشی از یک بازدید تلقی شوند فاصله طولانی تر بازدید تازه ای را آغاز می کند .

یادداشت : خدمتگرهای وبی که خدماتی فراهم می آورند که آمارهای شان در وبگاه دیگر

گزارش می شود از آمارهای وبگاه کتابخانه خارج می شوند .

ج) عامل های ماشینی (1)

دسته ای از بازدید کنندگان مجازی منابع اینترنتی عامل های غیرانسانی اند که برای اهداف خزش طراحی شده اند .

این ها معمولاً همراه موتورهای جست وجو هستند و خزشگرهای آرشیوگر نیز جزء آنها قرار می گیرند معمولاً عامل های ماشینی بازدید کننده از آمارهای تحلیل وب با پالایش نشانی های IP رشته علائم (2) عامل کاربری ماشینی یا با روش های پیچیده تری مانند شناسایی الگوهای تکرار شونده استفاده با(دنبال کردن هر) پیوند که یک بازدید کننده انسانی معمولا نمی تواند عامل آن ها ،باشد کنار گذاشته می شوند .

تحلیل رفتار عامل ماشینی میتواند مفید واقع شود به ویژه اگر نرم افزار دسترسی استاندارد منع خزش را فعال کرده باشد تا آرشیو کردن منابعی را که در فهرست کار موتورهای جست وجو قرار دارند متوقف کند بنابراین بالقوه با وبگاه زنده در

فهرست های جست وجو رقابت می کند .

د بازدیدکنندگان ناخواسته

آمار را می توان از بازدید کنندگان ناخواسته نیز گردآوری کرد ولی چون کاربر قصد بازدید وبگاه را نداشته این مورد به عنوان مثبت کاذب (3) عمل می کند . نمونه هایی از آن عبارت اند از پویش (4) صفحه ای که فرستاده (5) می شود یا افزایه های (6) وارسی ویروس در مرورگرها که

1- Robot visitors

2- String

3- False positive

به طور خودکـار پیونـدها را بـدون اینکه کـاربر واقعاً آن ها را درخواست کرده باشـد درخواست و پالایش می کننـد . برخی مرورگرها پیش دستی می کننـد و پیونـد مربوط به انتخاب های کاربران را می فرسـتند و نسـخه بندانگشتی (1) تولید می کنند تغییر مسـیرهای خودکار از وبگاه های زنده به آرشیو نیز ممکن است در این دسته قرار گیرد که به فعالیت بعدی کاربر بستگی دارد.

ه) تحلیل کارنامه ورود به سامانه

هر تراکنش HTTP از طریق خدمتگر وب با اطلاعاتی که در سـرآیند پرس وجوی HTTP وجود دارد و نیز سرآیند پاسخ HTTP که خود خـدمتگر تولید می کند می تواند وارد سامانه شود. اطلاعاتی مانند تاریخ زمان پرس وجو نشانی IP درخواست کننده حتی نام دامنه تجزیه شده ، منبع درخواستی نشانی وب منبع ارجاع دهنده و کد پاسخ میتوانند برای هر تراکنش پرس وجو پاسخ وارد سامانه شوند.

در تحلیلهای اوایل دهه 1990 شـمارش هر ورود به سامانه به عنوان یک «کوبه» (2) که وبگاه دریافت کرده است به حساب می آمـد از آنجاکه HTML با افزوده شدن اشـیاء جاسازی شده و انضـمام منابعی نظیر سبک برگه (3) یا جاوا اسـکریپت پیچیده تر ،شده کوبه ها به عنوان ابزار سنجش اعتبار خود را از دست داده اند و امروز فقط برای اندازه گیری بار (4) خدمتگر وب به کار می آیند .

بـاوجوداین نرم افزارهـای تحلیـل کارنـامه ورود به سامانه میتواننـد آگاهی هایی از داده های موجود در فایل های ورود به سامانه با توجه به مجموع ،آمارها به دست دهند. اسـتفاده گسترده از پراکسـی های وب (5) نهانگاه وب در دقت آمارهای ورود به سامانه از هم گسـیختگی (6) ایجـاد می کنـد به این علت که تعـداد زیادی کاربر در پوشـش یک نشانی IP قرار میگیرنـد و ممکن است نهانگـاه وب محلی به جای خدمتگر و بی که نرم افزار دسترسی آرشیو وب را اجرا می کند مرجع دریافت پرس و جوهای پی در پی قرار گیرد

،همچنین ممکن است نهانگاه های وب برای شـتاب دادن به عملکرد نرم افزار دسترسـی به بهای کاهش دقت آمارهای میزان استفاده مورد استفاده قرار گیرند؛ چراکه نهانگاه وب در نرم افزار

دسترسی خواهد کوشید پرس وجوها را پیش از خود نرم افزار دسترسی واقعی پاسخ دهد.

1- Thumbnail

2- Hit

3- Style sheet

4- Load

5- Caching

6- Disrupt

ه) برچسب زدن صفحات (1)

برچسب زدن صفحات فرایندی است که در آن اشیاء جاسازی شده HTML هنگام

بارگیری یک سند HTML توسط مرورگر به صورت پویا مکرراً درخواست میشوند این روش اولین بار هنگام کاربرد شمارنده صفحات (2) حاوی تصویر در اوایل دهه 90 به کار رفت؛ یعنی زمانی که میشد تصویری را از یک CGI درخواست کرد که معمولاً حاوی ارزش عددی تعداد دفعات اجرای اسکریپت و بنابراین تعداد بازدیدکنندگان صفحه بود

این روش به کاربرد جاوا اسکریپت و کلوچک ها (3) برای فراهم کردن اطلاعاتی درباره کاربر و صفحه پیشرفت کرده است از آنجا که برگردان پرس وجو از برچسب صفحه می تواند به هر مکانی باشد نه فقط خدمتگر وب ،(میزبان) این روش منجر به رشد برون سپاری صنعت تحلیل وب شده است.

و) ورود به سامانه در سطح برنامه کاربردی

از آنجا که نرم افزار دسترسی یک برنامه کاربردی تحت وب است این امکان وجود دارد که

سنجه های ورود به سامانه برای استفاده را مستقیما از خود برنامه تهیه کرد.

ز حریم خصوصی گردآوری اطلاعات درباره فعالیت های کاربر با هر یک از روش های تشریح شده بالا مستلزم سیاست گذاری مناسب درباره حریم خصوصی و در دسترس بودن آن برای کاربران سامانه است . کاربرد گسترده برچسب زدن صفحات نگاه (کنید به شرط)

ه- معرف چالش های دسترسی همگانی به آرشیو وب است چراکه و بگاه های آرشیو شده خودشان می توانند حاوی برچسب های جاسازی شده باشند. بنابراین ممکن است دیدن صفحات آرشیو منجر به ایجاد کلوچک هایی در مرورگر کاربر و تکرار پرس وجوها از تجمیع گرهای تحلیلی شود تراوایی به وب زنده

3.4.4

آمارهای پایه که استفاده از آرشیو را اندازه می گیرند .

جدول 4 شامل آمارهایی است که عمدتا انجمن تحلیل وب آنها را شناسایی کرده است این آمارها به آمارهای شمارش ساده و آمارهای پیچیده تجمعی یا آمار بعد (4) تقسیم می شوند

1- Page tagging

2- Page counters

3- Cookie

4- Dimension

بعضی از سنجه های ویژه آگاهی بخش به علت نامربوط بودن به دامنه این گزارش فنی حذف و تعدادی از سنجه های مفید اضافی در انتهای جدول جای گرفته اند.

در ستون سمت چپ، ،حروف اهمیت هر آمار پیشنهادی را نشان می دهند H برای ،بالا M برای ،متوسط و L برای .کم در غالب مواقع کاربرد آمارهای بسیار مهم برای آرشیو وب را توصیه می کنیم که میتواند به عنوان بخشی از استاندارد در آن لحاظ شود و در بیشتر برنامه های تحلیل گزارش شده اند

جدول 4 – آمارهای پایه که استفاده از آرشیو را اندازه می گیرند

جدول

بعضی از سنجه‌های ویژهٔ آگاهی‌بخش به‌علت نامربوط بودن به دامنهٔ این گزارش فنی حذف و تعدادی از سنجه های مفید اضافی در انتهای جدول جای گرفته‌اند. در ستون سمت چپ، حروف، اهمیت هر آمار پیشنهادی را نشان می‌دهند: H برای بالا، M برای متوسط، و L برای کم. در غالب مواقع کاربرد آمارهای بسیار مهم برای آرشیو وب را توصیه می کنیم که می‌تواند به‌عنوان بخشی از استاندارد در آن لحاظ شود و در بیشتر برنامه‌های تحلیل گزارش شده‌اند.

جدول ۴- آمارهای پایه که استفاده از آرشیو را اندازه می گیرند

نام	نوع	محاسبه	اهمیت
دیدن صفحه[1]	شمارگان	تعداد دفعاتی که یک صفحه دیده شده است	H= بالا- نشانهٔ استفادهٔ خام[2] از آرشیو
بازدید (نشست‌ها)	شمارگان	بازدید تعامل میان یک شخص یا وبگاه است که از یک یا چند پرس‌وجو برای یک صفحه تشکیل می‌شود. اگر شخص کار دیگری در یک وبگاه و در زمان مشخص انجام ندهد(یعنی صفحات دیگری را نبیند)، بازدید بدون زمان وقفه پایان می‌یابد.	H= بالا- شمارش پایهٔ کاربران آرشیو
بازدیدکنندگان یکتا[3]	شمارگان	تعداد تک کاربران (خزشگرها و عوامل پالایش شده‌اند) در یک بازهٔ زمانی گزارش دهی، با فعالیتی متشکل از یک یا چند بازدید از یک وبگاه. هر شخص در این آمار فقط یک بار برای دورهٔ مورد گزارش شمارش می‌شود.	M= متوسط
رخداد[4]	بُعد و/ یا شمارگان	هر ورود به سامانه یا عمل قابل ضبطی مشخصی که توسط خدمتگر یا سرور تاریخ و زمانی به آن تخصیص یافته است.	L= اندک

1.
2. Usage
3. Unique visilors
4. Event

(1) .1

(2) .2

(3) .3

(4) .4

1- ؟

2- usage ؟

3- Unique visilors

4- Event

4.4.4

آمارهای تجمعی برای ترسیم پیشرفته ویژگی های استفاده از آرشیو وب

جدول –5 آمارهای تجمعی برای ترسیم پیشرفته ویژگی های استفاده از آرشیو وب

۴.۴.۴

آمارهای تجمعی برای ترسیم پیشرفتهٔ ویژگی‌های استفاده از آرشیو وب

جدول ۵- آمارهای تجمعی برای ترسیم پیشرفتهٔ ویژگی های استفاده از آرشیو وب

نام	نوع	محاسبه	اهمیت
صفحهٔ ورودی[1]	بُعد	اولین صفحهٔ یک بازدید.	M= متوسط
صفحهٔ فرود[2]	بُعد	دیدن صفحه باهدف تشخیص نقطهٔ آغاز تجربهٔ کاربر منتج از یک تلاش تعریف شدهٔ بازاریابی.	L= کم
صفحهٔ خروج[3]	بُعد	آخرین صفحه در یک وبگاه که طی یک بازدید مورد دسترسی قرار می گیرد، به معنای پایان یک بازدید/ نشست.	L = کم
طول مدت بازدید	شمارگان	طول زمان در یک نشست. محاسبه نوعاً با استفاده از برچسب زمان مربوط به آخرین فعالیت در نشست، منهای برچسب زمان در اولین فعالیت نشست.	H= بالا
ارجاع دهندهٔ[4]	بُعد	ارجاع دهنده اصطلاح عامی است که منبع آمدورفت به یک صفحه یا بازدید را توصیف می کند.	M= متوسط
ارجاع دهندهٔ صفحه	بُعد	ارجاع دهندهٔ صفحه منبع آمدورفت به یک صفحه را توصیف می کند.	M= متوسط
بازدیدکنندهٔ جدید	شمارگان	تعداد بازدیدکنندگان یکتا یا فعالیتی متشکل از یک بازدید کاملاً تازه از یک وبگاه در طول دورهٔ مورد گزارش. به یاد داشته باشید که «اولین بار»[5] با توجه به زمانی است که گردآوری داده‌را با استفاده از نرم افزار معمول آغاز کرده‌اند.	M= متوسط

1. Entry Page
2. Landing Page
3. Exit Page
4. Refferre
5. First ever

(1) .1

2. (2)

3. (3)

4. (4)

5. (5)

ص: 80

1- Entry Page

2- Landing Page

3- Exit Page

4- Refferre

5- First ever

عکس

بازدیدکننده بازگشتی[1]	شمارگان	تعداد بازدیدکنندگان یکتا با فعالیتی متشکل از یک بازدید از یک وبگاه در طول مدت گزارش‌دهی درجایی که بازدیدکننده یکتا همچنین وبگاه را پیش از دورهٔ گزارش دهی بازدید کرده باشد.	M= متوسط
بازدید کنندهٔ مکرر[2]	شمارگان	تعداد بازدیدکنندگان یکتا با فعالیتی متشکل از دو بازدید یا بیشتر از یک وبگاه در طی دورهٔ گزارش دهی.	M= متوسط
بازدید یک صفحه (پرش)[3]	بُعد یا شمارگان	بازدیدی که از یک‌بار دیدن صفحه تشکیل‌شده است.	M= متوسط
بازدیدکنندگان برحسب مکان جغرافیایی	شمارگان	IP جغرافیایی که گزارش از نشانی IP درخواست‌کننده می‌دهد.	L= کم
واژگان جست‌وجو برای یافتن آرشیو	شمارگان	واژگان جست‌وجوی به‌کاررفته در موتورهای جست‌وجو برای یافتن وبگاه نرم‌افزار دسترسی.	M= متوسط
واژگان جست‌وجوی به‌کاربرده شده در آرشیو	شمارگان	واژگان جست‌وجوی به‌کاررفته در نرم‌افزار دسترسی آرشیو برای یافتن وب‌گرفت‌های آرشیو شده.	H= بالا

۵.۴.۴

آمارهای هسته برای استفاده از مجموعه

جدول ۶- آمارهای هسته برای استفاده از مجموعه

آمار	هدف	نمونه
تعداد صفحات دیده‌شده	گسترهٔ استفاده	۴۸/۳۱۸ صفحه در آرشیو وب انگلستان بین اول تا سی‌ام ژوئن ۲۰۱۲ دیده شده است.
تعداد بازدید	گسترهٔ استفاده	۱۱/۴۱۵ بازدید از آرشیو وب انگلستان بین اول تا سی‌ام ژوئن ۲۰۱۲ صورت گرفته است.

1. Returning Visitor
2. Repeative Visitor
3. Bounce

5.4.4

آمارهای هسته برای استفاده از مجموعه

جدول –6 آمارهای هسته برای استفاده از مجموعه

1. (1)

2. (2)

3. (3)

ص: 81

1- Returning Visitor

2- Repeative Visitor

3- Bounce

عکس

تعداد بازدیدکنندگان یکتا	گسترهٔ استفاده	۹۴۳۴ بازدیدکنندهٔ یکتا بین اول تا سی‌ام ژوئن ۲۰۱۲ از آرشیو وب انگلستان بازدید کرده‌اند.
طول مدت بازدید	علاقهٔ کاربران به آرشیو	بین اول تا سی ام ژوئن ۲۰۱۲ هر بازدید از آرشیو وب انگلستان ۳ دقیقه و ۲۵ ثانیه طول کشیده است.
واژگان جست‌وجو درون آرشیو	رفتار کاربر	پراستفاده ترین کلیدواژهٔ به‌کاررفته برای جست‌وجو در آرشیو وب انگلستان بین اول تا سی ام ژوئن ۲۰۱۲ «goji berry» بوده است.

۵.۴
حفاظت از آرشیو وب
۱.۵.۴
کلیات

حفاظت بلندمدت آرشیوهای وب نباید از چارچوب کلی‌تری که برای حفاظت همهٔ منابع دیجیتالی به کار گرفته می‌شود جدا شود. به‌گونه‌ای آرمانی بجاست که مؤسسات گردآورندهٔ سامانهٔ حفاظت دیجیتالی مختص موجودی‌های دیجیتالی‌شان ایجاد کنند که منطبق بر استانداردهایی مثل ایزو ۱۴۷۲۱(سامانه های اطلاعاتی آرشیو باز = OAIS) باشد.

مدل مرجع OAIS ویژگی‌های سامانهٔ آرشیوی مختص حفاظت و در دسترسی نگه داشتن اطلاعات دیجیتالی در طول زمان را توصیف می‌کند. این گزارش فنی دربارهٔ جزئیات مدل OAIS بحث نمی‌کند، بلکه از مفاهیم پایه و تعاریف آن برای توصیف موضوعات مرتبط با آرشیو وب بهره می‌برد. بعضی از آمارهای پیشنهادی به سایر انواع منابع دیجیتالی قابل‌تعمیم‌اند و بعضی دیگر مختص آرشیو وب هستند.

همان‌گونه که در قسمت ۳.۳ روش‌های حفاظت توضیح داده شد، حفاظت دیجیتالی را می‌توان در دو سطح انجام داد: در سطح پایه بیت‌ها باید سالم نگه‌داشته شوند؛ و در سطح پیچیده‌تر با استفاده از استراتژی‌هایی مثل مهاجرت و همگون‌سازی ظاهر، عملکرد، رفتار و حتی تجربهٔ کاربر با منبع دیجیتالی حفظ می‌شود. از اولی به‌عنوان «حفاظت فیزیکی یا جریان بیت» و از دومی با عنوان «حفاظت منطقی»[1] یاد می‌کنند.

آمارهایی که در ۲. ۵.۴ تشریح شده‌اند اندازه‌گیری میزان کارآمدی فعالیت های حفاظت

1. Logical preservation

5.4. حفاظت از آرشیو وب

1.5.4

کلیات

حفاظت بلندمدت آرشیوهای وب نباید از چارچوب کلیتری که برای حفاظت همه منابع دیجیتالی به کار گرفته می شود جدا شود . به گونه ای آرمانی بجاست که موسسات گردآورنده سامانه حفاظت دیجیتالی مختص موجودی های دیجیتالی شان ایجاد کنند که منطبق بر استانداردهایی مثل ایزو 14721 سامانه های اطلاعاتی آرشیو باز = OATS باشد. مدل مرجع OAIS ویژگی های سامانه آرشیوی مختص حفاظت و در دسترسی نگه داشتن اطلاعات دیجیتالی در طول زمان را توصیف می.کند این گزارش فنی درباره جزئیات مدل DAIS بحث نمی کند بلکه از مفاهیم پایه و تعاریف آن برای توصیف موضوعات مرتبط با آرشیو وب بهره میبرد بعضی از آمارهای پیشنهادی به سایر انواع منابع دیجیتالی قابل تعمیماند و بعضی دیگر مختص آرشیو وب هستند.

همان گونه که در قسمت 3.3 روش های حفاظت توضیح داده شد حفاظت دیجیتالی را می توان در دو سطح انجام داد در سطح پایه بیت ها باید سالم نگه داشته شوند؛ و در سطح پیچیده تر با استفاده از استراتژی هایی مثل مهاجرت و همگون سازی ،ظاهر عملکرد رفتار و حتی تجربه کاربر با منبع دیجیتالی حفظ می.شود از اولی به عنوان حفاظت فیزیکی یا جریان «بیت و از دومی با عنوان حفاظت منطقی» (1) یاد می کنند .

آمارهایی که در 2.5.4 تشریح شدهاند اندازه گیری میزان کارآمدی فعالیت های حفاظت

1- Logical preservation

جریان بیت را مدنظر .دارند در 3.5.4 قالبی (1) پیشنهاد شده است که موسسات را در گزارش فراداده هایی که انتظار میرود در آرشیو وب حفاظت شوند یاری میرساند. آمارها در 4.5.4 آن دسته ای هستند که به حفاظت منطقی مشهورند.

2.5.4

آمارهای حفاظت جریان داده

1.2.5.4

مقدار منابع گم شده یا آسیب دیده

1.1.2.5.4

هدف

بسیاری از موسسات ،میراثی گم شدگی داده به علت خرابی فیزیکی رسانه را تجربه کرده اند درایوهای دیسک سخت ناگاه خراب و رسانه ها از رده خارج می شوند و بعضی از داده ها تصادفاً پاک می شوند . منابع آسیب دیده منابعی هستند که محتوای آن ها کاملا از دست نرفته است ولی تمامیت آن ها دچار خدشه شده و در نتیجه دسترسی یا ارائه محتوای کامل آن ها ممکن نیست اطلاعات درباره گم شدگی داده معمولاً موجود نیست ولی پایش مقدار منابع گمشده یا معیوب دارای اهمیت است چراکه شاخص مهمی برای تمامیت آرشیوهای وب به شمار می رود .

2.1.2.5.4

روش

مقدار داده گمشده یا معیوب را میتوان برحسب بایت یا تعداد نشانی های وب با مقایسه منظم چک سام به دست آورد.

2.2.5.4

مقدار منابع تکراری و توزیع شده

منابعی که از آن ها نسخه پشتیبان تهیه نشده و نسخه تکراری هم ندارند در معرض خطر گم شدگی همیشگی قرار دارند و نمیتوان آنها را احیا کرد روش معمول این است که نسخه مکرر منابع در مکان های مختلفی نگهداری شود تا از ایجاد نقطه شکست (2) پیشگیری کنند.

1- Template

2- Point of failure

،پس مقدار منابع مکرر نشانه ایمنی مجموعه است.

با وجود ،این داشـتن نسـخه های مکرر از دارایی ها به تنهایی برای تعیین ایمنی مجموعه کافی نیست پارامترهای متنوع دیگری بایـد در نظر گرفته شوند تا تهیه نسخه مکرر از داده ها کارآمد باشد در اینجا مهمترین مسئله تنوع و تمامیت است که شامل موارد زیر می شود .

تعداد پیکربندی های متنوع نرم افزارها برای اقلام مختلف

فاصله فیزیکی میان اقلام

سرعت و شیوه وارسی تمامیت در بین اقلام داده ای

منابعی که تمامیت آن ها با چنین اطلاعاتی تضـمین شود مکرر شـده و توزیع شده در نظر گرفته می شوند لازمه تهیه نسـخه مکرر و توزیع داده صـرف هزینه است. تصـمیم گیری درباره گستره اینکه تا چه حد از منابع یک آرشـیو وب باید نسـخه تکراری تهیه شود با ایجاد تعادل میان تاثیر خطرات بها و پیچیدگی مدیریت مجموعه منابع تکراری انجام می شود .

3.5.4

آمارهای مربوط به حفاظت فراداده

اهمیت حفـاظت منابع همراه با فراداده مربوط به هر کـدام در 4.3.3 توضـیح داده شده است. توصـیه می شود موسسات گردآورنده به طور منظم درباره سرشت و مقدار فراداده در یک آرشیو وب با استفاده از جدول 7 گزارش ارائه کنند.

جدول – آمارهای مربوط به حفاظت فراداده

پس، مقدار منابع مکرر نشانهٔ ایمنی مجموعه است.

با وجود این، داشتن نسخه‌های مکرر از دارایی‌ها به‌تنهایی برای تعیین ایمنی مجموعه کافی نیست. پارامترهای متنوع دیگری باید در نظر گرفته شوند تا تهیهٔ نسخهٔ مکرر از داده‌ها کارآمد باشد. در اینجا مهم‌ترین مسئله تنوع و تمامیت است که شامل موارد زیر می‌شود:

- تعداد پیکربندی‌های متنوع نرم‌افزارها برای اقلام مختلف؛
- فاصلهٔ فیزیکی میان اقلام؛
- سرعت و شیوهٔ وارسی تمامیت در بین اقلام داده‌ای؛

منابعی که تمامیت آنها با چنین اطلاعاتی تضمین شود مکرر شده و توزیع‌شده در نظر گرفته می‌شوند. لازمهٔ تهیهٔ نسخهٔ مکرر و توزیع داده صرف هزینه است. تصمیم‌گیری دربارهٔ گسترهٔ اینکه تا چه حد از منابع یک آرشیو وب باید نسخهٔ تکراری تهیه شود با ایجاد تعادل میان تأثیر خطرات، بها و پیچیدگی مدیریت مجموعهٔ منابع تکراری انجام می‌شود.

۳.۵.۴

آمارهای مربوط به حفاظت فراداده

اهمیت حفاظت منابع همراه با فرادادهٔ مربوط به هرکدام در ۴.۳.۳ توضیح داده‌شده است. توصیه می شود مؤسسات گردآورنده به‌طور منظم دربارهٔ سرشت و مقدار فراداده در یک آرشیو وب با استفاده از جدول ۷ گزارش ارائه کنند.

جدول ۷- آمارهای مربوط به حفاظت فراداده

نوع فراداده	توصیف	استاندارد مورداستفاده (در صورت موجود بودن)	درصد منابعی که حاوی این فراداده هستند	توضیحات
یکی از انواع فراداده، مذکور در ۴.۳.۳ مانند فراداده توصیفی	توصیف فراداده		در صد منابعی که فراداده را در خود دارند	هر توضیح مفید یا مناسب

عکس

توصیفی	مجموعه عناصر فرادادهٔ DCMI[1]، نام اصطلاح: موضوع. موضوع منبع	DCMI؛ LCSH[2]	۳۰	عبارت موضوع به‌صورت دستی توسط آرشیوگر به منبع تخصیص داده می شود و در سامانه WCT[15] ذخیره می‌شود.
منشأ	فایل های پیکربندی		۹۰	فایل های پیکربندی خزش‌های سال ۲۰۰۴ از مجموعهٔ فایل‌هاوچین شده‌اند.
فنی	قالب‌های فایل	MIME: Part two: Media Types[16]	۱۰۰	همهٔ فایل های برداشت شده اطلاعات قالب فایل را دارند، ولی ممکن است قابل‌اعتماد نباشند.
حقوق	مجوز آرشیو و دسترسی بر خط		۱۰۰ درصد	الزامی فقط برای وبگاه‌های هدفی که دسترسی به آنها آزاد است.

۴.۵.۴

آمارهای حفاظت منطقی

حفاظت جریان بیت، سعی در ایمن نگه داشت بایت‌ها بر مادهٔ رسانه و حفاظت منطقی تضمین قابل‌استفاده ماندن منابع در طول زمان است. در این گزارش فنی سه شاخص اصلی برای فعالیت‌های حفاظت منطقی مطرح‌شده است.

1. Dublin Core Metadata Initiative
2. Library of congress Subject Headings

آمارهای حفاظت منطقی

حفاظت جریان ،بیت سعی در ایمن نگه داشت بایت ها بر ماده رسانه و حفاظت منطقی تضمین قابل استفاده ماندن منابع در طول زمان .است در این گزارش فنی سه شاخص اصلی برای فعالیت های حفاظت منطقی مطرح شده است.

1. (1)

2. (2)

ص: 85

1- Dublin Core Metadata Initiative

2- Library of congress Subject Headings

1.4.5.4

توزیع برحسب قالب های [شناسایی شده] فایل

1.1.4.5.4

هدف

توزیع منابع آرشیو وب برحسب قالب فایل به عنوان عنصری از سرشت نمایی آرشیو آماری است که به تفضیل در 3020304. شرح داده شده است. شناخت قالب های فایل در حفاظت دیجیتالی بسیار حساس .است تعیین استراتژی حفاظت برای هر منبعی بدون اطلاعات مربوط به قالب غیر ممکن است آرشیوهای وب نوعاً حاوی منابعی در طیف وسیعی از قالب ها هستند.

اطلاعات نوع فایلی (Type) MIME که خدمتگرهای وب در پاسخ به پرس و جوی منابع باز می گردانند تنها اطلاعات قابل اعتماد درباره قالب فایل ها هستند که در دسترس.اند کاربرد نرم افزارهای شناسایی قالب برای دریافت اطلاعات دقیق تر درباره قالب ها در آرشیو وب برای اهداف حفاظت الزامی است .

علاوه بر صورت بندی و تدوین یک استراتژی حفاظت ،دیجیتالی اطلاعات قالب می تواند .

برای شناسایی خطرات حفاظت و اولویت بندی عملکردهای حفاظتی مورد استفاده قرار گیرد .

2.1.4.5.4

روش

به «روش» مذکور در 3020304 نگاه کنید برای حفاظت اطلاعات دقیق تری درباره قالب ها مورد نیاز است و این با کاربرد ابزارهای شناسایی قالب قابل انجام است خروجی ها را می توان

به شیوه هایی که در 3020304 پیشنهاد شده است محاسبه و سازماندهی کرد.

DROIT و Jhove مثال هایی از نرم افزارهای شناسایی قالب فایل هستند .

3.1.4.5.4

محدودیت ها

نرم افزارهای شناسایی قالب همیشه همه قالب ها را نمیتوانند شناسایی کنند مثال خوبی در

این مورد قالب های جدیدی است که نرم افزارها هنوز نمی توانند آن ها را بشناسند.

2.4.5.4

تعداد قالب های فایل با استراتژی مشخص حفاظت

1.2.4.5.4

هدف

استراتژی های حفاظت برای قالب هایی تعین می شوند که هنوز مورد استفاده اند و از رده خارج نشده اند. موسسات میتوانند به جای فعال کردن استراتژی ها برای همه مجموعه آن ها را بر نمونه ای از منابع آزمایش کنند محدودیت های منابع ممکن است مانع از کاربرد استراتژی ها در مورد آن ها شود .

محاسبه اندازه منابعی که استراتژی ها در موردشان فعال شده و یا اجرا می شود نشان می دهد که منابع در معرض خطر تا چه حد از استراتژی حفاظت دیجیتالی بهرهمند شده اند .

همچنین میزان تعهد سازمان را در قبال حفاظت دیجیتال آشکار می سازد. در صورتی که « استراتژی هیچ کاری نکن » (1) حاکم بود هر بار بازبینی استراتژی با تصمیم ادامه کار با آن می تواند فعال کردن آن تلقی شود چیزی که اطمینان می دهد تصمیمات آگاهانه اخذ می شوند استراتژی قطعی و صریحی برای حفاظت وجود ندارد همان گونه که فناوری متحول می شود .

منابع ممکن است همچنان از رده خارج شوند و به تعریف و اجرای مکرر استراتژی های تازه ای برای همان قالب ها نیاز باشد.

2.3.4.5.4

روش ها

از قالب های فایلی که استراتژی حفاظتی برایشان تعیین و فعال شده است فهرستی تهیه کنید که می تواند از سامانه حفاظت دیجیتالی گرفته و محاسبه شود این ،آمار ،سپس با توزیع منابع در هر قالب فایل که برحسب بایت یا نشان های وب اندازه گرفته شده است ترکیب می شود .

5.5.4

آمارهای هسته برای حفاظت مجموعه

جدول 8 – آمارهای هسته برای حفاظت مجموعه

۲.۴.۵.۴
تعداد قالب‌های فایل با استراتژی مشخص حفاظت
۱.۲.۴.۵.۴
هدف

استراتژی‌های حفاظت برای قالب‌هایی تعین می‌شوند که هنوز مورد استفاده‌اند و از رده خارج نشده‌اند. مؤسسات می‌توانند به‌جای فعال کردن استراتژی‌ها برای همهٔ مجموعه آنها را بر نمونه‌ای از منابع آزمایش کنند. محدودیت‌های منابع ممکن است مانع از کاربرد استراتژی‌ها در مورد آنها شود. محاسبهٔ اندازهٔ منابعی که استراتژی‌ها در موردشان فعال‌شده و یا اجرا می شود نشان می‌دهد که منابعِ در معرض خطر تا چه حد از استراتژی حفاظت دیجیتالی بهره‌مند شده‌اند. همچنین میزان تعهد سازمان را در قبال حفاظت دیجیتال آشکار می‌سازد.

درصورتی‌که «استراتژی هیچ کاری نکن»[1] حاکم بود، هر بار بازبینی استراتژی با تصمیم به ادامهٔ کار با آن می‌تواند فعال کردن آن تلقی شود، چیزی که اطمینان می‌دهد تصمیمات آگاهانه اخذ می‌شوند. استراتژی قطعی و صریحی برای حفاظت وجود ندارد. همان‌گونه که فناوری متحول می‌شود، منابع ممکن است همچنان از رده خارج شوند و به تعریف و اجرای مکرر استراتژی‌های تازه‌ای برای همان قالب‌ها نیاز باشد.

۲.۳.۴.۵.۴
روش‌ها

از قالب‌های فایلی که استراتژی حفاظتی برایشان تعیین و فعال شده است فهرستی تهیه کنید که می‌تواند از سامانهٔ حفاظت دیجیتالی گرفته و محاسبه شود. این آمار، سپس، با توزیع منابع در هر قالب فایل که برحسب بایت یا نشان‌های وب اندازه گرفته شده است، ترکیب می‌شود.

۵.۵.۴
آمارهای هسته برای حفاظت مجموعه

جدول ۸- آمارهای هسته برای حفاظت مجموعه

آمار	هدف	مثال
مقدار منابع مکرّر	ایمنی و ترمیم‌پذیری[2]	۱۵۰ترابایت از منابع آرشیو وب نسخهٔ مکرر هستند

1. Do- nothing strategy
2. Resilience

2. (2)

ص: 87

1- Do– nothing strategy

2- Resilience

عکس

توزیع پراکندگی برحسب قالب‌های (شناسایی شدهٔ) فایل	توانمندی حفاظت	۶۰ درصد از آرشیو در HTML است
تعداد قالب‌هایی که برایشان استراتژی حفاظت تعیین شده است	توانمندی و تعهد حفاظت	برای ۵ قالب استراتژی حفاظت تعیین شده است: HTML، Gif، jpeg، PNG، PDF

۶.۴
سنجش هزینه‌های آرشیو وب

هزینه‌های آرشیو وب، مانند هزینه‌های ایجاد و حفاظت سایر مجموعه‌های دیجیتالی (مانند دیجیتالی کردن)، به روش‌های مختلفی قابل‌سنجش است. تنها چیزی که باید در خاطر داشت این است که فعالیت‌های آرشیو وب هنوز تازه‌اند و بعضی از ابعاد کارآمدی و هزینه‌ها را فقط با گذشت زمان می‌توان سنجید. این نکته به‌خصوص در مورد هزینه‌های مربوط به حفاظت آرشیو وب صدق می‌کند.

۱.۶.۴
برون‌سپاری

یک‌سازمان گردآورنده می‌تواند بعضی یا همهٔ فعالیت‌های مربوط به آرشیو وب را به یک پیمانکار یا شخص ثالث واگذار کند. چنین خدمتی می‌تواند شامل برداشت داده، نمایه‌سازی، دسترسی یا ذخیره باشد. این خدمت همچنین ممکن است فراهم‌آوری محتوای پیشین و تاریخی مجموعه‌ها و فعالیت‌های توسعهٔ نرم‌افزار خاصی را در برگیرد.

برون‌سپاری سرراست ترین راه برای محاسبهٔ هزینه‌های آرشیو وب است، چراکه این هزینه‌ها کل مبلغ پولی را که باید مؤسسه برای پیمانکار بپردازد شامل می‌شود. ممکن است هزینه‌های مربوط به انتخاب محتوا، مدیریت مناقصه و سایر موارد نیز به آن اضافه شود که معمولاً مؤسسه به‌صورت داخلی آنها را تأمین می‌کند. برای محاسبهٔ کل هزینه‌های آرشیو باید این موارد اخیر نیز به هزینه‌های برون‌سپاری افزوده شود.

6.4. سنجش هزینه های آرشیو وب

هزینه های آرشیو وب مانند هزینه های ایجاد و حفاظت سایر مجموعه های دیجیتالی (مانند دیجیتالی کردن) به روش های مختلفی قابل سنجش است.

تنهـا چیزی که بایـد در خاطر داشت این است که فعالیت های آرشـیو وب هنوز تازه انـد و بعضـی از ابعاد کارآمـدی و هزینه ها را فقط با گذشت زمان میتوان سنجید این نکته به خصوص در مورد هزینه های مربوط به حفاظت

آرشیو وب صدق میکند

1.6.4

برون سپاری

یک سازمان گردآورنـده میتوانـد بعضـی یا همه فعالیت های مربوط به آرشـیو وب را به یک پیمانکار یا شـخص ثالث واگـذار کنـد . چنین خدمتی می تواند شامل برداشت داده نمایه سازی دسترسی یا ذخیره باشد این خدمت همچنین ممکن است فراهم آوری محتوای پیشین و تاریخی مجموعه ها و فعالیت های توسعه نرم افزار خاصی را در برگیرد برون سپاری سرراست ترین راه برای محاسبه هزینه های آرشیو وب است چراکه این هزینه ها کل مبلغ پولی را که باید موسسه برای پیمانکار بپردازد شامل می شود ممکن است هزینه های مربوط به انتخاب محتوا مـدیریت منـاقصه و سایر موارد نیز به آن اضافه شود که معمولا موسسه به صورت داخلی آن ها را تامین می.کنـد برای محاسـبه کل هزینه های آرشیو باید این موارد اخیر نیز به هزینه های برون سپاری افزوده شود

2.6.4

ایجاد آرشیو وب در موسسه

ارزیابی هزینه های ایجاد آرشیو وب توسط موسسه چالش برانگیزتر از محاسبه فرایندهای آشناتری مانند فهرستنویسی در کتابخانه ها .نیست چهار دسته اصلی از مخارج باید در نظر گرفته شوند سخت افزار رایانش (1) نرم افزار و نیروی انسانی

1.2.6.4

هزینه های سخت افزار

هزینه های سخت افزار شامل فراهم آوری و نگهداری زیر ساخت لازم برای برداشت نمایه سازی، پردازش (2) ذخیره و ایجاد دسترسی به داده ها و سایر منابع دیجیتالی است.

2.2.6.4

هزینه های رایانش

هزینه های رایانش شامل هزینه های مربوط به توان (3) و شبکه پهنای باند است.

3.2.6.4

هزینه های نرم افزار

بسته به اینکه چه نرم افزاری برای آرشیو وب انتخاب می شود بهای مجوز کاربرد پیدا می کند. بسیاری از موسسات گردآورنده آرشیو وب هم اکنون از نرم افزارهای کد منبع باز و رایگانی استفاده می کنند که سازمان های بین المللی یا بنیادهای غیر انتفاعی تدوین کرده و توسعه و نگهداریشان را بر عهده دارند بعضی موسسات نیز از راه حل هایی استفاده می کنند که طی همکاری های بین المللی تهیه شدهاند مانند کنسرسیوم بین المللی حفاظت اینترنت (4) این کار منجر به کاهش هزینه های تهیه نرم افزار در سازمان یا پرداخت هزینه مجوز به شرکتهای تجاری می شود . پیاده کردن نرم افزار کد منبع باز و یکپارچه کردن آن ها با سامانه های داخلی موسسات مستلزم تخصص فنی است . منابع توسعه دهندگان نرم افزار باید برای روزآمدسازی

1- Computing

2- Ingest

3- Power

4- International Internet Preservation Consortium

در دسترس باشند هر موسسه الزامات ویژه خود را دارد که مستلزم توسعه های نرم افزار برای

بومی سازی است.

426.4

هزینه نیروی انسانی

هزینه نیروی انسانی را میتوان به طور معمول به صورت نیروی کار تمام وقت یا روزمزد محاسبه کرد . عموماً سه دسته کلی از کارکنان در آرشیو وب دست اندر کارند: آرشیوگر، فنی و مدیریتی.

این کارکنان با همکاری هم طیف گسترده ای از عملکردها از جمله گسترش مجموعه عملیات ،فنی برنامه ریزی و مدیریت اجرایی را انجام می دهند انواع تقسیمات نیروی انسانی برمبنای تخصص وجود دارد؛ به عنوان مثال یک آرشیوگر بیشتر احتمال دارد که به دامنه مجموعه و توصیف آن بپردازد و یک مهندس خزشگر را به کار بیندازد یا نرم افزار بنویسد با وجود این آرشیو کردن وب اغلب مستلزم مهارت های چندگانه است و وظایف معمولاً برحسب حوزه های تخصصی توزیع و انجام می شوند .

کارکنان باید درک پایه و دانش گستردهای از آرشیو علاوه بر تخصص حرفه ای خودشان داشته باشند چرا که موجب بهتر شدن همکاری می شود و به ارائه عملکرد بهتر یاری می رساند وظایف را می توان میان کارکنان واحدهای کاری مختلف توزیع کرد و بسیاری از کارکنان می توانند در انجام آن ها شرکت کنند تعداد کارکنانی را که باید در آرشیو وب به کار گرفته شوند با افزودن زمان صرف شده توسط همه کارکنان دائمی و موقت و کارکنان طرحی که در آرشیو وب به کار گرفته می شوند میتوان محاسبه کرد روش های مختلفی برای این کار وجود دارد

الف) برآورد (1) تعداد شغل های تمام وقت را که مستقیماً به آرشیو وب اختصاص یافته است محاسبه کنید میانگین صرف شده توسط کارکنان آرشیو وب را که در سایر خدمات نیز مشغول به خدمت هستند بگیرید و نتیجه را از زمان قبلی کم کنید . میانگین زمانی را نیز که کارکنان سایر بخش ها در آرشیو وب کار میکنند بگیرید و این عدد را به عدد قبل بیفزایید مثال: 3/5 کارمند تمام وقت فقط در آرشیو وب کار میکنند طی دورهٔ تهیه گزارش 10 درصد از وقتشان را در سایر وظایف صرف می کنند . کارکنان سایر بخش ها (8 کارمند تمام وقت) 20 درصد از وقتشان را در آرشیو وب صرف میکنند مجموع نیروی انسانی تمام وقت برای آرشیو وب چنین برآورد می شود :

3/5-0/35+1/6-4/75

Estimate -1

ب) کارنامه حضور و غیاب (1) دوره نمونه ای را انتخاب کنید (معمولاً یک یا دو هفته) که طی

آن آرشیو وببار کاری میانگین نسبت به اوقات دیگر دارد.

زمان کاری ،کارکنان کارکنان سایر واحـدها که اوقاتی را هم در آرشـیو وب کار میکنند را از یادداشت های ثبت روزانه ساعت کاری ثبت کنید ،نتیجه تعداد کارکنان تمام وقت را برای

دوره مورد گزارش را نشان می دهد .

ج) همچنین میتوان هزینه ها را برحسب بخش های تصدی گری امور فنی و مدیریتی

از هم متمایز کرد .

5.2.6.4

سایر هزینه ها

سایر هزینه ها میتواند شامل موارد زیر باشد :

فراهم آوری ،فراداده مثلاًـ خریـد فهرست نامه ای دامنه از موسـسات ثبت دامنه؛ کارشناسـی :حقوق مشاور حقوقی تـدارک اقـدام قانونی یا پرداخت غرامت در پی رای دادگاه به عنوان مثال ناشری علیه یک موسسه آرشیو وب به علت مسائل حیثیتی و یا ضرر مادی

طرح شکایت می کند .

همکاری بین المللی آرشـیو وب را جامعه ای جهانی پدیـد آورده و پشتیبانی کرده است وارد شدن در همکاری بین المللی ممکن است به علت هزینه های عضویت و مخارج سفر و اقامت لازم شده باشد.

جدول –9 آمارهای هسته برای هزینه های مجموعه

ب) کارنامهٔ حضوروغیاب[1]: دورهٔ نمونه ای را انتخاب کنید (معمولاً یک یا دو هفته) که طی آن آرشیو وب بار کاری میانگین نسبت به اوقات دیگر دارد.

زمان کاری کارکنان، کارکنان سایر واحدها که اوقاتی را هم در آرشیو وب کار می کنند را از یادداشت‌های ثبت روزانهٔ ساعت کاری ثبت کنید. نتیجه، تعداد کارکنان تمام‌وقت را برای دورهٔ مورد گزارش را نشان می‌دهد.

ج) همچنین می‌توان هزینه‌ها را برحسب بخش‌های تصدی گری، امور فنی، و مدیریتی از هم متمایز کرد.

۵.۲.۶.۴

سایر هزینه ها

سایر هزینه‌ها می‌تواند شامل موارد زیر باشد:

- فراهم آوری فراداده، مثلاً خرید فهرست نام‌های دامنه از مؤسسات ثبت دامنه؛
- کارشناسی حقوق: مشاور حقوقی، تدارک اقدام قانونی یا پرداخت غرامت در پی رأی دادگاه (به‌عنوان‌مثال ناشری علیه یک مؤسسهٔ آرشیو وب به علت مسائل حیثیتی و یا ضرر مادی طرح شکایت می کند)؛
- همکاری بین‌المللی: آرشیو وب را جامعه‌ای جهانی پدید آورده و پشتیبانی کرده است. وارد شدن در همکاری بین‌المللی ممکن است به علت هزینه‌های عضویت و مخارج سفر و اقامت لازم شده باشد.

جدول ۹- آمارهای هسته برای هزینه‌های مجموعه

آمار	هدف	مثال
هزینه های سخت افزار	هزینه های خرید و نگهداری سخت افزار	هزینهٔ جا به جایی زیرساخت ذخیره‌سازی ۵۰ هزار یورو بوده است
هزینه‌های رایانش	هزینه‌های توان و شبکه	هزینه و پهنای باند ۱۰هزار یورو در سال
هزینه‌های نرم افزاری	هزینه‌های خرید، یکپارچه‌سازی، تدوین یا ارتقاء نرم افزار	تدوین نرم افزار جدید آرشیوگری با مبلغ ۸۰ هزاریورو برون‌سپاری شد

1. Time logging

1- Time logging

عکس

هزینهٔ نیروی انسانی	هزینه‌های منابع انسانی (مثل آرشیوگران، مهندسان و غیره) به‌صورت تمام‌وقت یا درازای مبلغ پرداختی	گروه آرشیو وب شامل ۳ مهندس تمام‌وقت و ۴ متصدی تمام‌وقت است

۵. شاخص‌های کیفیت

۱.۵

کلیات

کیفیت «میزان تحقق الزامات توسط ویژگی‌های ذاتی» تعریف شده است (ایزو ۹۰۰۰: ۲۰۰۵). این شرط شامل شاخص‌هایی است که امکان می دهند ارزیابی کنیم که مجموعهٔ ویژگی‌های ذاتی یک برنامهٔ آرشیو وب تا چه درجه مجموعهٔ الزاماتی را که مدیران یا ذینفعان آن تعیین کرده‌اند برآورده می کند.

هدف از شاخص‌های کیفیت که در این بخش پیشنهاد می‌شوند کمک به مؤسسهٔ آرشیو وب در پاسخ به سؤال‌هایی است مانند آنچه در زیر می‌آید:

- آیا می‌دانیم چه چیزی را گردآوری می کنیم؟
- اگر نه، سیاست روشنی لازم است که در آن دامنهٔ آرشیو تعریف‌شده باشد.
- آیا در کار گردآوری آن چیزی هستیم که می‌خواهیم؟
- اگر نه، لازم است از همگونی میان منابع گردآوری شده و منابع هدف اطمینان حاصل کنیم.
- آیا ما بهترین استفاده را از منابعمان می بریم؟
- اگر نه، به بهبود روش‌ها و گردش کار برای افزایش کارآمدی نیاز داریم.
- آرشیو به چه میزانی در دسترس و قابل‌جست‌وجوست؟
- بهبود مداوم کار قابلیت کاربرد آرشیو اهمیت دارد.
- آیا می توانیم تضمین کنیم که آرشیو وب در طول زمان قابل‌دسترسی باقی بماند؟
- اگر نه، لازم است اقدامات قابل‌اعتماد حفاظت را اجرا کنیم.

شاخص‌های پیشنهادی، مناسب‌ترین‌ها برای ارزیابی کیفیت خدمات فراهم شده توسط یک سازمان در طول زمان است. توصیه می شود کیفیت به‌طور منظم اندازه‌گیری و ارزیابی شود. مقایسهٔ مؤسسات در صورتی میسر است که شاخص‌ها به یک روش به کار گرفته و تفسیر شوند. باوجوداین، ضمن در نظر داشتن تفاوت وظایف و مأموریت‌های سازمانی و نیز تفاوت سازمان‌ها در منابع و رویه‌ها، در چنین مقایسه هایی باید محتاط بود.

5. شاخص های کیفیت

1.5. کلیات

کیفیت میزان تحقق الزامات توسط ویژگی های ذاتی تعریف شـده است (ایزو 9000: 2005). این شرط شامل شاخص هایی است که

امکان می دهند ارزیابی کنیم که مجموعه ویژگی های ذاتی یک برنامه آرشیو وب تا چه درجه مجموعه الزاماتی را که مدیران یا ذینفعان آن تعیین کرده اند برآورده می کند.

هدف از شاخص های کیفیت که در این بخش پیشنهاد می شوند کمک به موسسه آرشیو

وب در پاسخ به سوالهایی است مانند آنچه در زیر می:آید

آیا میدانیم چه چیزی را گردآوری میکنیم؟

– اگر نه سیاست روشنی لازم است که در آن دامنه آرشیو تعریف شده باشد.

آیا در کار گردآوری آن چیزی هستیم که میخواهیم؟

– اگر نه لازم است از همگونی میان منابع گردآوری شده و منابع هدف اطمینان حاصل کنیم.

آیا ما بهترین استفاده را از منابعمان میبریم؟

– اگر نه به بهبود روش ها و گردش کار برای افزایش کارآمدی نیاز داریم

آرشیو به چه میزانی در دسترس و قابل جست و جوست؟

– بهبود مداوم کار قابلیت کاربرد آرشیو اهمیت دارد

آیا می توانیم تضمین کنیم که آرشیو وب در طول زمان قابل دسترسی باقی بماند؟

– اگر ،نه لازم است اقدامات قابل اعتماد حفاظت را اجرا کنیم

شاخص های ،پیشنهادی مناسبترین ها برای ارزیابی کیفیت خدمات فراهم شده توسط یک سازمان در طول زمان است توصیه میشود کیفیت به طور منظم اندازه گیری و ارزیابی

مقایسه موسسات در صورتی میسر است که شاخص ها به یک روش به کار گرفته و تفسی– .شوند ،باوجوداین ضمن در نظر داشتن تفاوت وظایف و ماموریت های سازمانی و نیز تفاوت سازمانها در منابع و رویه ها در چنین مقایسه هایی باید محتاط بود.

شود.

2.5. محدودیت ها

شاخص های کیفیتی پیشنهاد شده در این گزارش فنی بازتابی از وضعیت فعلی آرشیو وب از هر دو دیدگاه فنی و آرشیوگری هستند این شاخص ها پابه پای پیشرفت هایی که در آرشیو وب صورت میگیرد باید بازبینی و روزآمد .شوند تفسیر نتایج کاربرد شاخص های کیفیت باید با احتیاط صورت گیرد ممکن است خطاهایی در نمونه گیری و اندازه گیری رخ دهد و منجر به

بی دقتی شود .

3.5. توصیف

1.3.5

کلیات

در زیر شاخص های کیفیت برمبنای ربطشان به ابعاد کلیدی برنامه های آرشیو وب فهرست شده:اند سیاست گذاری ،برداشت و دسترسی و حفاظت

هر چند محاسبات را می توان برحسب نشانی های وب یا بایت انجام داد محاسبه با نشانی های وب توصیه می شود در واقع با این شاخصها مقدار منابع محاسبه میشود و نشانی های وب بیشتر از بایت ها به منابع کتابخانه ای شبیه اند

مدیریت

1) هزینه بر حسب نشانی های وب گردآوری شده

2) درصد کارکنان دست اندرکار آرشیو وب

کیفیت فرایند گردآوری

1) درصد منابعی که در دوره زمانی مشخصی از وب زنده ناپدید شده اند

2) درصد تحقق دامنه اجباری

3) درصد درخواست ها برای تفاهم نامه یا مجوزهای اعطا شده توسط صاحبان حقوق

دسترس پذیری و استفاده

1) درصد منابع قابل دسترسی برای کاربر

2) درصد منابعی که به صورت تمام متن نمایه شده اند

3) درصد منابع فهرستنویسی شده

4) درصد منابع مورد دسترسی در سال

5) درصد بازدید از کتابخانه شامل بازدید از آرشیو وب

6) تعداد صفحاتی که در هر بازدید دیده شده اند.

حفاظت

1) درصد منابعی که دست کم یک نسخه مکرر از آن ها

2) درصد منابع گم شده یا معیوب

3) درصد منابع با قالب های فایل شناخته شده

وجود دارد

4) درصد منابعی که برای قالب آن ها استراتژی حفاظت تعیین شده است

5) درصد منابعی که از حیث ویروس وارسی شده اند

2,3,5

مدیریت

دسترس‌پذیری و استفاده

۱) درصد منابع قابل‌دسترسی برای کاربر،
۲) درصد منابعی که به‌صورت تمام متن نمایه شده‌اند،
۳) درصد منابع فهرستنویسی شده،
۴) درصد منابع مورد دسترسی در سال،
۵) درصد بازدید از کتابخانه شامل بازدید از آرشیو وب،
۶) تعداد صفحاتی که در هر بازدید دیده شده‌اند.

حفاظت

۱) درصد منابعی که دست‌کم یک نسخهٔ مکرر از آنها وجود دارد،
۲) درصد منابع گم‌شده یا معیوب،
۳) درصد منابع با قالب‌های فایل شناخته شده،
۴) درصد منابعی که برای قالب آنها استراتژی حفاظت تعیین شده است،
۵) درصد منابعی که از حیث ویروس وارسی شده‌اند.

۲.۳.۵

مدیریت

شمارهٔ شاخص	۱
نام	هزینهٔ برحسب نشانی های وب گردآوری شده
هدف	ارزیابی کارآمدی فرایندهای آرشیو وب
پیش	- جمع هزینهٔ آرشیو وب آن‌چنان‌که در ۶.۴ ذکر شده است - تعداد کل نشانی‌های وب که خزش شده اند.
روش	هزینه برحسب نشانی‌های وب گردآوری شده است: A/B که در آن: A هزینهٔ کل آرشیو وب در یک دورهٔ زمانی خاص است. B تعداد نشانی های وب خزش شده در یک دوره زمانی. گرد کردن تا یک رقم اعشار

عکس

توضیحات	هزینهٔ پایین برحسب نشانی‌های وب گردآوری شده عموماً حاکی از کارآمدی فرایندهای آرشیو وب است؛ هزینه‌های بالا برحسب نشانی‌های وب گردآوری شده ممکن است نشان دهندهٔ سطح بالایی از آرشیوگری باشد. بهترین کاربرد این شاخص مقایسهٔ مجموعه هایی با اندازه و اهداف مشابه است.

شمارهٔ شاخص	۲
نام	درصد کارکنان دست‌اندرکار آرشیو وب
هدف	نشان دادن تعهد سازمانی به آرشیو وب
پیش‌نیازها	- تعداد کارکنان تمام‌وقت دست‌اندرکار آرشیو وب - کل تعداد کارکنان کتابخانه
روش	درصد کارکنان دست‌اندرکار در آرشیو وب: A/B×۱۰۰ که در آن: A تعداد کارکنان کتابخانه که تمام‌وقت دست‌اندرکار آرشیو وب هستند: گزینش، برداشت، دادن دسترسی، حفاظت. B کل تعداد کارکنان تمام‌وقت کتابخانه، شامل کارکنان دائمی و موقت و کارکنان طرحی است: گرد کردن تا نزدیک ترین عدد صحیح برای کارکنان تمام‌وقت دست اندرکار آرشیو وب. در مورد کارکنان پاره‌وقت می‌توان زمانی را که در آرشیو صرف کرده اند، با محاسبهٔ زمان خود اظهاری یا از طریق بررسی حضوروغیاب به دست آورد.

کیفیت فرایند گردآوری

شمارهٔ شاخص	۳
نام	درصد منابعی که در دورهٔ زمانی مشخص از وبِ زنده ناپدید شده‌اند
هدف	ارزیابی ارزش آرشیو وب
پیش‌نیازها	تعداد اهداف [وبگاه‌ها] در آرشیو تعداد اهداف [وبگاه‌ها] در آرشیو که از وبِ زنده ناپدید شده‌اند. رقم دوم را می توان به‌صورت خودکار در غیاب پاسخ DNS یا پاسخ ۴۰۴ یا به‌صورت دستی و از طریق وارسی وبِ زنده به دست آورد.

کیفیت فرایند گردآوری

ص: 95

عکس

روش	درصد منابعی که طی مدت زمان مشخص از وبِ زنده ناپدید شده‌اند این‌گونه به دست می‌آید: A/B×۱۰۰ که در آن A تعداد [وبگاه‌های] هدف ناپدید شده است. B تعداد [وبگاه‌های] هدف در آرشیو است. گرد کردن تا یک رقم اعشار
توضیحات	تعیین ناپدیدشدگی یک [وبگاه] هدف دشوار است. هدفی که دامنهٔ (نشانی اینترنتی) آن تغییر کرده لزوماً ناپدید نشده و ممکن است دامنهٔ جدید و متفاوتی پیدا کرده است. در بعضی مواقع فقط بخشی از یک هدف ناپدید می‌شود. این مورد به علل اجرایی از بحث خارج شده است. در مفهوم این شاخص، یک هدف، زمانی ناپدید می‌شود که دیگر پاسخ DNS وجود ندارد یا نشانی اینترنتی آن پاسخ ۴۰۴ را ایجاد می‌کند. در صورت امکان قابل‌اعتمادترین روش، چک کردن دستی وب است.

شمارهٔ شاخص	۴
نام	درصد تحقق دامنهٔ اجباری
هدف	ارزیابی اینکه آیا آرشیو وبی که به‌دست‌آمده با آنچه تعیین شده و باید به دست می‌آمد تطابق دارد
پیش‌نیازها	- تعداد [وبگاه‌های] هدف برداشت شده توسط مؤسسه در یک سال - تعداد [وبگاه‌های] هدف تعیین شده توسط دامنهٔ اجباری یا برگرفته از آن
روش	درصد تحقق دامنهٔ اجباری عبارت است از: A/B×۱۰۰ که در آن: A تعداد [وبگاه‌های] هدف برداشت شده توسط کتابخانه در یک سال است. B تعداد [وبگاه‌های] هدف درون دامنه[1] تعیین شده توسط دامنهٔ اجباری یا برگرفته از آن(برای مثال، واسپاری قانونی) گرد کردن تا یک رقم اعشار

1. In - Scope

1. (1)

ص: 96

In- Scope -1

عکس

توضیحات	الزام برای آرشیو وب اغلب هم ملی و هم سازمانی است. اگر قانون واسپاری آثار در کشور وجود دارد، معمولاً در آن حدود جغرافیایی یک دامنهٔ ملی تعریف شده است. برای منابع اینترنتی که با دامنه‌های سطح بالای ملی میزبانی می‌شوند، این تعریف واضح است؛ ولی منابعی که در دامنهٔ خاص وظایف ملی قرار می‌گیرند، همیشه به‌طور مشخص در نام دامنه‌های دارای ارجاع جغرافیایی میزبانی نمی‌شوند. برای نمونه، بر اساس گزارش AFNIC[1] که متولی ثبت دامنهٔ .fr است، فقط یک‌سوم وبگاه‌های فرانسوی در نام دامنه‌های سطح بالای .fr میزبانی می‌شوند. روش‌های دیگری غیر از نام دامنه‌های سطح بالا برای تعیین حیطهٔ جغرافیایی وبگاه‌ها لازم است. این امر ممکن است شامل کسب اطلاعات از مؤسسات متولی ثبت دامنه نیز بشود.

شمارهٔ شاخص	۵
نام	درصد درخواست‌های تفاهم نامه یا مجوزهای اعطا شده توسط صاحبان حقوق
هدف	ارزیابی کارآمدی درخواست‌های مجوز
پیش نیازها	- تعداد درخواست‌های تفاهم‌نامه یا مجوز ارسال شده به صاحبان حقوق - تعداد تفاهم‌نامه ها یا مجوزهای اعطا شده توسط صاحبان حقوق
روش	درصد درخواست‌ها برای تفاهم‌نامه یا مجوزهای اعطا شده توسط صاحبان حقوق این گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد تفاهم نامه‌ها یا مجوزهای اعطا شده توسط صاحبان حقوق B تعداد درخواست‌های تفاهم نامه یا مجوزهای ارسالی به صاحبان حقوق گرد کردن تا یک رقم اعشاری
توضیحات	نرخ بالا نشان دهندهٔ موفقیت فعالیت درخواست مجوز است. نیز توصیه می‌شود تعداد رد درخواست‌های مشخص و تعداد عدم پاسخ ثبت شود. ارتباط و مدافعهٔ بهتر از پدیدآورندگان وبگاه می‌تواند راهی در بهتر شدن میزان موفقیت باشد. می توان برنامه‌ای برای ارتباطات تهیه و استدلال‌های کلیدی را که می‌توانند پدیدآورندگان را متقاعد کنند شناسایی کرد و در مورد بهترین مجاری توزیع اطلاعات تصمیم گیری نمود.

1. AFNIC: Association Francaisa Pour le nommage Internet en Cooperation

1. (1)

ص: 97

AFNIC: Association Francaisa Pour le nommage Internet en Cooperation -1

عکس

۴.۳.۵
قابلیت دسترسی و استفاده

شمارهٔ شاخص	۶
نام	درصد منابع قابل‌دسترسی برای کاربر
هدف	ارزیابی قابلیت دسترسی آرشیو وب
پیش‌نیازها	- تعداد کل منابع در آرشیو وب - تعداد منابع در دسترس به‌صورت برخط (آرشیو برخط) - تعداد منابع در دسترس در مؤسسه (آرشیو خاکستری) موارد بالا برحسب نشانی وب یا بایت قابل‌اندازه‌گیری هستند.
روش	درصد منابع قابل‌دسترسی برای کاربر این گونه قابل‌محاسبه است: $(A+A')/B\times 100$ که در آن: A تعداد منابع در آرشیو وب است که به‌صورت برخط قابل‌دسترسی هستند؛ A′ تعداد منابع در آرشیو وب است که فقط در مؤسسه قابل‌دسترس هستند؛ B تعداد کل منابع در آرشیو وب است. گرد کردن تا یک رقم اعشاری
توضیحات	درصد بالا نشان دهندهٔ رؤیت‌پذیری[1] یا قابلیت دسترسی بالای آرشیو وب است. شاخص را می توان برای ارزیابی قابلیت دسترسی مستقیم منابع برای کاربر به‌تنهایی برای منابع در دسترس به‌صورت برخط محاسبه کرد. واحد اندازه گیری برای محاسبه باید گزارش شود، مثلاً نشانی وب یا بایت.

شمارهٔ شاخص	۷
نام	درصد منابع نمایهٔ تمام متن
هدف	ارزیابی قابلیت جست‌وجوی منابع آرشیو وب
پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابعی که به‌صورت تمام متن نمایه‌سازی شده‌اند. محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.

1. Uisiblity

1. (1)

ص: 98

Uisiblity -1

عکس

روش	درصد منابع نمایه تمام متن بدین گونه به دست می آید: A/B×۱۰۰ که در آن: A تعداد منابعی است به‌صورت تمام متن نمایه‌سازی شده‌اند B تعداد کل منابع آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	جست‌وجوی تمام متن قابلیت دسترسی و استفاده از آرشیو وب را تا حد زیادی افزایش می‌دهد. واحد اندازه‌گیری مورداستفاده برای محاسبه باید گزارش داده شود، یعنی نشانی وب یا بایت.

شمارهٔ شاخص	۸
نام	درصد منابع فهرستنویسی شده
هدف	ارزیابی قابلیت جست‌وجو و سطح مدیریت[1] آرشیو وب
پیش‌نیازها	- تعداد وبگاه‌های هدف در آرشیو وب؛ - تعداد وبگاه‌های هدف که فهرستنویسی شده‌اند و در فهرست‌ها پیشینه‌ای دارند.
روش	درصد منابع فهرستنویسی شده این گونه به دست می آید: A/B×۱۰۰ که در آن: A تعداد وبگاه‌های هدف که فهرستنویسی شده و در فهرست‌ها پیشینه‌ای دارند B تعداد کل وبگاه‌های هدف در آرشیو وب گرد کردن تا یک رقم اعشاری
توضیحات	فهرستنویسی منابع در آرشیو وب قابلیت دسترسی و استفاده از آنها را افزایش می‌دهد. این کار به یکپارچه شدن منابع آرشیو وب با سایر منابعی که در کتابخانه نگه داری می‌شوند یاری می‌رساند. هرچند فهرستنویسی هزینه بر است و برای آرشیوهای وب بزرگ‌مقیاس که با خزش پشته‌ای گردآوری می شوند ممکن نیست. هنگام محاسبهٔ این شاخص توصیه می‌شود استراتژی برداشت مورداستفاده در گردآوری منابع نیز گزارش شود.

شمارهٔ شاخص	۹
نام	درصد منابع در دسترس قرار گرفته

1. Curation

1. (1)

ص: 99

Curation -1

عکس

هدف	ارزیابی گسترهٔ[1] فعلی استفاده از آرشیو وب
پیش‌نیازها	- تعداد کل نام دامنه‌ها در آرشیو وب؛ - تعداد نام های دامنه‌ای که دست‌کم یک صفحه از آنها در سال دیده شده است.
روش	درصد منابعی که در دسترس قرارگرفته‌اند بدین گونه محاسبه می‌شود: A/B×۱۰۰ که در آن: A تعداد نام‌های دامنه‌ای که دست‌کم یک صفحه از آنها در سال دیده شده است. B تعداد کل نام های دامنه در آرشیو وب. گرد کردن تا یک رقم اعشاری
توضیحات	درصد بالا نشانهٔ استفادهٔ گسترده از آرشیو وب است. دلیل استفاده از نام‌های دامنه به‌جای نشانی‌های وب این است که ممکن است فقط منابعی که در تعداد محدودی نام دامنه قرار دارند به‌طور فعال مورداستفاده قرار گیرند. محاسبهٔ استفاده برحسب دامنه گسترهٔ استفاده و جامعیت آرشیو را نشان می دهد.

شمارهٔ شاخص	۱۰
نام	درصد بازدید از کتابخانه، شامل بازدید از آرشیو وب
هدف	ارزیابی میزان استفاده از آرشیو وب توسط بازدیدکنندگان از کتابخانه (برخط یا در محل)
پیش‌نیازها	- تعداد کل بازدیدها از کتابخانه (بازدیدهای حضوری و مجازی)؛ - تعداد کل بازدیدها از آرشیو وب.
روش	درصد بازدیدکنندگان از کتابخانه که از آرشیو وب استفاده می کنند این‌گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد بازدیدها از آرشیوهای وب است (مثلاً در ماه یا در سال) B تعداد کل بازدیدها از کتابخانه (مثلاً در ماه یا سال) گرد کردن تا یک رقم اعشاری

1. Breadlth

1. (1)

ص: 100

Breadlth -1

عکس

توضیحات	درصد بالا استفادهٔ بالای کاربرانِ سایر خدمات کتابخانه‌ای از آرشیو وب را نشان می‌دهد. بازدیدهای حضوری مربوط به کاربرانی است که به محوطهٔ کتابخانه وارد می‌شوند. بازدیدهای مجازی بازدید از وبگاه کتابخانه هستند.

شمارهٔ شاخص	۱۱
نام	تعداد صفحات دیده شده در هر بازدید
هدف	ارزیابی میزان علاقهٔ کاربران به آرشیو
پیش‌نیازها	- تعداد کل صفحات دیده شده در هر دورهٔ گزارش؛ - تعداد کل بازدیدها از آرشیو وب در همان دوره.
روش	تعداد صفحات دیده شده در هر بازدید این‌گونه به دست می‌آید: $A/B \times ۱۰۰$ که در آن: A تعداد کل صفحات دیده شده در یک دورهٔ گزارش‌دهی است B تعداد کل بازدیدها در همان دوره است. گرد کردن تا یک رقم اعشاری
توضیحات	نرخ بالا نشان‌دهندهٔ استفادهٔ زیاد از آرشیو وب است و بدین معناست که آرشیو محتوای مناسبی برای کاربران دارد. در نظر داشتن نیازها و توقعات کاربران و پژوهشگران، مثلاً از طریق پیمایش نظراتشان، راهی برای افزایش نرخ استفاده است.

شمارهٔ شاخص	۱۲
نام	درصد منابعی که دست‌کم یک نسخهٔ مکرر دارند
هدف	ارزیابی توان حفاظت از جریان داده
پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابع با دست‌کم یک نسخهٔ مکرر. محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.

عکس

روش	درصد منابعی که دست‌کم یک نسخهٔ مکرر دارند این‌گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد منابع با دست‌کم یک نسخهٔ مکرر است B تعداد کل منابع در آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	واحد اندازه گیری مورداستفاده باید ذکر شود؛ یعنی نشانی وب یا بایت.

شمارهٔ شاخص	۱۳
نام	درصد منابع گم‌شده یا معیوب
هدف	ارزیابی امنیت ذخیره‌سازی در آرشیو وب
پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابع گم‌شده یا معیوب؛ محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.
روش	درصد منابع گم‌شده یا معیوب این‌گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد منابع گم‌شده یا معیوب است B کل تعداد منابع در آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	نرخ پایین نشان‌دهندهٔ پایین بودن امنیت است. واحد اندازه‌گیری به‌کاررفته برای محاسبه باید ذکر شود، یعنی نشانی وب یا بایت.

شمارهٔ شاخص	۱۴
نام	درصد منابع با قالب‌های فایل شناخته‌شده
هدف	ارزیابی دانش سازمانی دربارهٔ آرشیو وب و توان حفاظت
پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابع با قالب‌های فایل شناسایی‌شده؛ محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.

ص: 102

عکس

روش	درصد منابع با قالب فایل شناخته شده در آرشیو این گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد منابع با قالب فایل شناخته شده است B تعداد کل منابع در آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	واحد اندازه‌گیری مورداستفاده برای محاسبه باید در گزارش ذکر شود، یعنی نشانی وب یا بایت. این درصد را می‌توان با توسعهٔ بهتر ابزارهای شناسایی (به‌منظور بهبود عملکردشان) یا با بهبود فرایند شناسایی قالب‌ها با افزودن ماشین‌ها و مانند آن افزایش داد.

شمارهٔ شاخص	۱۵
نام	درصد منابعی که برای قالب آنها استراتژی حفاظت تعیین شده است
پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابع در یک قالب دارای استراتژی حفاظت؛ محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.
روش	درصد منابعی که برای قالب آنها استراتژی حافظت تعیین شده است بدین گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد منابعی است که برای قالبشان استراتژی حافظت تعیین شده است B تعداد کل منابع در آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	واحد اندازه‌گیری به‌کاررفته در محاسبه باید در گزارش ذکر شود، یعنی نشانی وب یا بایت.

شمارهٔ شاخص	۱۶
نام	درصد منابعی که از حیث ویروس وارسی شده اند
هدف	ارزیابی استفادهٔ ایمن از آرشیو وب، سایر مجموعه‌ها و دستگاه‌های کاربر

عکس

پیش‌نیازها	- تعداد کل منابع در آرشیو وب؛ - تعداد منابع وارسی شده از حیث ویروس؛ محاسبه را می‌توان برحسب نشانی وب یا بایت انجام داد.
روش	درصد منابعی که از حیث ویروس وارسی شده‌اند این‌گونه به دست می‌آید: A/B×۱۰۰ که در آن: A تعداد منابعی است که از حیث ویروس وارسی شده‌اند B تعداد کل منابع در آرشیو وب است گرد کردن تا یک رقم اعشاری
توضیحات	ممکن است سیاست مؤسسات در کشف ویروس‌ها، پاک نکردن آنها از آرشیو وب، و در مقابل مسدود کردن دسترسی به منابع آلوده باشد. واحد اندازه‌گیری باید در گزارش ذکر شود، یعنی نشانی وب یا بایت.

۶ استفاده و منافع

۱.۶

کلیات

آمارها و شاخص‌های کیفیت به مقایسه و ارزیابی عملکرد آرشیوهای وب یاری می‌رسانند و باید بخشی از برنامه‌های منظم و گردش کار ارزیابی مؤسسات باشند. منافع فراوانی بر تهیهٔ نظام‌مند آمارها و سنجش کیفیت آرشیو وب مترتب است:

- تصمیم‌گیری آگاهانه را ممکن می‌کند؛
- با فراهم کردن نقاط وارسی لازم به مدیریت عملکردهای آرشیو وب یاری می‌رساند؛
- گفت‌وگو میان مؤسسات آرشیو وب، حامیان نهادهای سرمایه‌گذار، و جامعهٔ کاربران را تسهیل می‌کند؛
- آگاهی از آرشیو وب به‌طور عام را افزایش می‌دهد و بهترین عملکردها را تشویق می‌کند؛
- به شناسایی شکاف‌ها و مسائل مشترک کمک می‌کند تا این مسائل با همکاری مؤسسات آرشیوگر به‌طور جمعی بررسی شوند؛
- شاهدی است بر اجرا و به اثبات ارزش آرشیو وب یاری می‌رساند.

6. استفاده و منافع

1.6. کلیات

آمارها و شاخص های کیفیت به مقایسه و ارزیابی عملکرد آرشیوهای وب یاری میرسانند باید بخشی از برنامه های منظم و گردش کار

ارزیابی موسسات .باشند منافع فراوانی بر تهیه نظام مند آمارها و سنجش کیفیت آرشیو وب مترتب است:

- تصمیم گیری آگاهانه را ممکن می کند؛

با فراهم کردن نقاط وارسی لازم به مدیریت عملکردهای آرشیو وب یاری می رساند؛ گفت وگو میان موسسات آرشیو وب حامیان نهادهای سرمایه گذار و جامعه کاربران را

تسهیل میکند؛

آگاهی از آرشـیو وب به طور عام را افزایش می دهد و بهترین عملکردها را تشویق میکند به شناسایی شکاف ها و مسائل مشترک کمک می کند تا این مسائل با همکاری موسسات

رشیوگر به طور جمعی بررسی شوند؛

شاهدی است بر اجرا و به اثبات ارزش آرشیو وب یاری میرساند.

ص: 104

2.6. استفاده ها و کاربران مورد نظر

اصـطلاحات آماری و شاخص های کیفیت که در این گزارش فنی تعریف و توضـیح داده شد را میتوان برای ارزیابی عملکرد آرشـیو وب به کار برد این شاخص ها به مقایسه آرشیوهای وب نیز میتوانند کمک کنند

آمارها بی طرفاند و میتوان از داده های عینی به عنوان پایه ای برای تحلیل ها و تفسـیرهای بیشتر استفاده کرد شاخص های کیفیت به معنای نوعی داوری ارزشی هستند با به کار بردن یک شاخص عملکرد خوب یا بد نشان داده می شود.

این گزارش ،فنی همچنین شامل بررسی کلی و توصیف تفصیلی فرایندهای فنی متنوعی ،است گرچه کل گزارش برای همهٔ خوانندگان مفید است و در عین اینکه خوانـدن بخش های تفصـیلی مربوط توصـیه میشود در جدول 10 محتوای اصـلی برای هر یـک از سه گروه کاربر به صورت متمایز ارائه شده است.

گروه های کاربر

مدیران

جدول –10 استفاده ها و خوانندگان مورد نظر

۲.۶
استفاده ها و کاربران موردنظر

اصطلاحات آماری و شاخص‌های کیفیت -که در این گزارش فنی تعریف و توضیح داده شد- را می‌توان برای ارزیابی عملکرد آرشیو وب به کار برد. این شاخص‌ها به مقایسهٔ آرشیوهای وب نیز می‌توانند کمک کنند.

آمارها بی طرف‌اند و می‌توان از داده‌های عینی به‌عنوان پایه‌ای برای تحلیل‌ها و تفسیرهای بیشتر استفاده کرد. شاخص‌های کیفیت به معنای نوعی داوری ارزشی هستند: با به کار بردن یک شاخص، عملکرد خوب یا بد نشان داده می‌شود.

این گزارش فنی، همچنین شامل بررسی کلی و توصیف تفصیلی فرایندهای فنی متنوعی است، گرچه کل گزارش برای همهٔ خوانندگان مفید است و در عین اینکه خواندن بخش‌های تفصیلی مربوط توصیه می‌شود، در جدول ۱۰ محتوای اصلی برای هریک از سه گروه کاربر به‌صورت متمایز ارائه شده است.

جدول ۱۰- استفاده ها و خوانندگان موردنظر

گروه‌های کاربر	برداشت اطلاعات	دسترسی و حفاظت
مدیران	۱.۲.۴. ۳.۵. ۳.۴. ۱. ۴.۳.۴. ۱.۳.۴. ۶.۲.۴.۱.۵. ۶.۴	۱.۴.۴. ۵.۳. ۴.۳. ۱. ۵.۵.۴. ۱.۵.۴. ۵.۴.۴. ۱.۵. ۶.۴
آرشیوگران	۲.۴ .۴.۳. ۱.۳.۵. ۳.۴	۴.۴. ۴.۳. ۳.۳. ۲.۳.۵. ۴.۵.۴. ۳.۵.۴
مهندسان	۲.۴.۳ .۳.۴	۴.۴. ۳.۳. ۲.۳.۵.۴

۳.۶
بهره‌مندی گروه‌های کاربر

این گزارش برای تصمیم‌گیرندگان (مدیران) درون و بیرون از مؤسسات گردآورنده، برای کارکنان آرشیو که وبگاه‌ها را انتخاب و اداره می‌کنند (کتابداران و آرشیوگران)، و کارکنان فنی که سامانه‌ها را مدیریت و زیرساخت را نگه‌داری و از عملیات فنی پشتیبانی می‌کنند (مهندسان) مفید است.

3.6. بهره مندی گروه های کاربر

این گزارش برای تصمیم گیرندگان (مدیران) درون و بیرون از موسسات گردآورنده برای کارکنان آرشیو که وبگاه ها را انتخاب و اداره میکنند کتابداران) و آرشیوگران و کارکنان فنی که سامانه ها را مدیریت و زیر ساخت را نگهداری و از عملیات فنی پشتیبانی می کنند (مهندسان) مفید است.

الف) برای مدیران

آمارها و شاخص های کیفیت مربوط به گردآوری و ،دسترسی برنامه ریزی درست برای منابع را ممکن میکند و به اندازه گیری میزان موفقیت آرشیو وب در جهت ماموریت های سازمانی یاری می رساند شاخص های مرتبط با هزینه ها به ارزیابی و اولویت بندی توسعه برنامه ریزی شده کمک می.کند شاخص های کیفیت همچنین تحقق اهداف در آرشیو وب را با استفاده از اصطلاحات متداول نشان می دهند .

ب کارکنان آرشیو آمارهای مجموعه به کارکنان آرشیو کمک می کند تا به حوزه های کلیدی در انتخاب منابع معطوف شوند و مقایسه فرایند انتخاب را با سایر موسسات ممکن می سازد. شاخص های کیفیت مرتبط با دسترسی به شناسایی الزامات آتی کمک می.کنند شاخص های کیفی ارزیابی عملکرد آرشیوگرها را نیز پشتیبانی می کنند .

ج) برای مهندسان

آمارها و شاخص های کیفیت به مهندسان فرصت حصول دانش در ابعاد مختلف سامانه های آرشیو وب از جمله عملکرد آنها را میدهد این آمارها و شاخصها باید به طور منظم گزارش شوند تا خدمات و عملیات را بهبود بخشند آمارها و شاخص های کیفیت مرتبط با حفاظت مهماند و میتوانند در طراحی سامانه هایی که دسترس پذیری پشتیبانی شده آرشیو وب را تضمین کنند یاری رسان باشند .

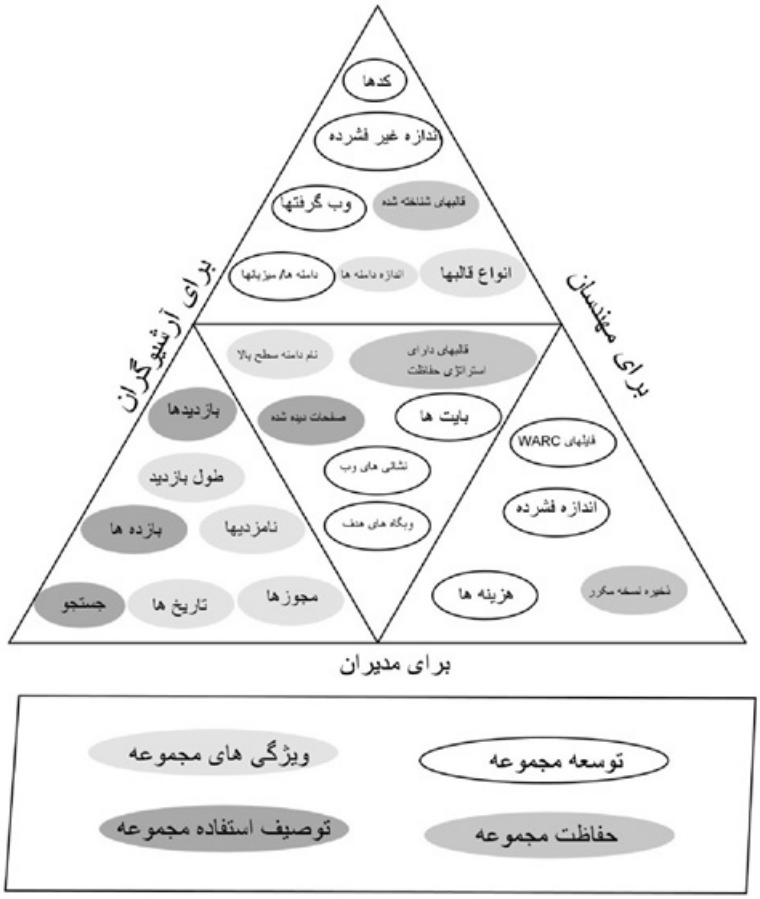
4.6. استفاده از آمارهای پیشنهاد شده توسط گروه های کاربری

همه آمارها و شاخص های مطرح شده در این گزارش فنی مربوط به خوانندگان مورد نظر است .

تصویر 1 تنظیمات اصلی برای ارتباط و مباحثه میان گروه های اصلی را مشخص کرده است. آمارهای کنار مثلث به دو گروه از کاربران که اسامیشان در طول ضلع نوشته شده کاملاً مربوط هستند آمارهای میانهٔ مثلث برای همه گروه های کاربری است.

به عنوان مثال، آمارهای بالای مثلث مربوط به رابطهٔ میان آرشیوگران و مهندسان است .

عکس



تصویر: استفاده از آمارها توسط گروه های کاربر اصطلاحات آماری مورد استفاده در شکل ۱ در جدول۱۱ توضیح داده شده اند.

تصویر استفاده از آمارها توسط گروه های کاربر اصطلاحات آماری مورد استفاده در شکل 1 در جدول 11 توضیح داده شده اند.

ص: 107

عکس

جدول ۱۱- اصطلاحات به‌کاررفته در شکل ۱

نوع	آمار
بایت	اندازهٔ آرشیو، چه فشرده و چه غیر فشرده
وب‌گرفت ها	تعداد وبگاه‌های هدف که گیرانداخته شده‌اند
کد	توزیع نشانی‌های وب برحسب کدهای وضعیت و به‌ویژه ردیف 2xx
اندازهٔ فشرده	اندازهٔ فشردهٔ آرشیوها برحسب بایت
هزینه ها	هزینه‌های سخت افزاری، رایانشی، نرم افزاری، نیروی انسانی
تاریخ ها	پوشش زمانی
دامنه‌ها/ میزبان‌ها	تعداد دامنه‌ها و میزبان‌ها
اندازهٔ دامنه‌ها	قالب‌های دارای استراتژی حفاظت
قالب‌های شناخته‌شده	توزیع برحسب قالب‌های فایل شناسایی‌شده
انواع قالب	توزیع برحسب نوع قالب
نامزدی‌ها	تعداد نامزدی‌ها
صفحات دیده‌شده	تعداد صفحات دیده‌شده
مجوزها	تعداد مجوزها
ذخیرهٔ نسخه‌های مکرر	مقدار منابع مکرر
جست‌وجو	پرکاربردترین اصطلاحات مورداستفاده در آرشیو
وبگاه‌های هدف	تعداد وبگاه‌های هدف
نام دامنه های سطح بالا	توزیع برحسب نام دامنه‌های سطح اول و سطح دوم
اندازهٔ غیرفشرده	اندازهٔ آرشیو به شکل غیر فشرده برحسب بایت
نشانی‌های وب	کل تعداد نشانی‌های وب (یا پاسخ ها)
طول مدت بازدید	متوسط زمان یک بازدید
بازدیدکنندگان	تعداد افراد بازدیدکنندگان
بازدیدها	تعداد بازدیدها
فایل‌های WARC	تعداد فایل‌های WARC یا سایر فایل‌های حامل

5.6. فرایند آرشیو وب با شاخص های عملکرد مربوط

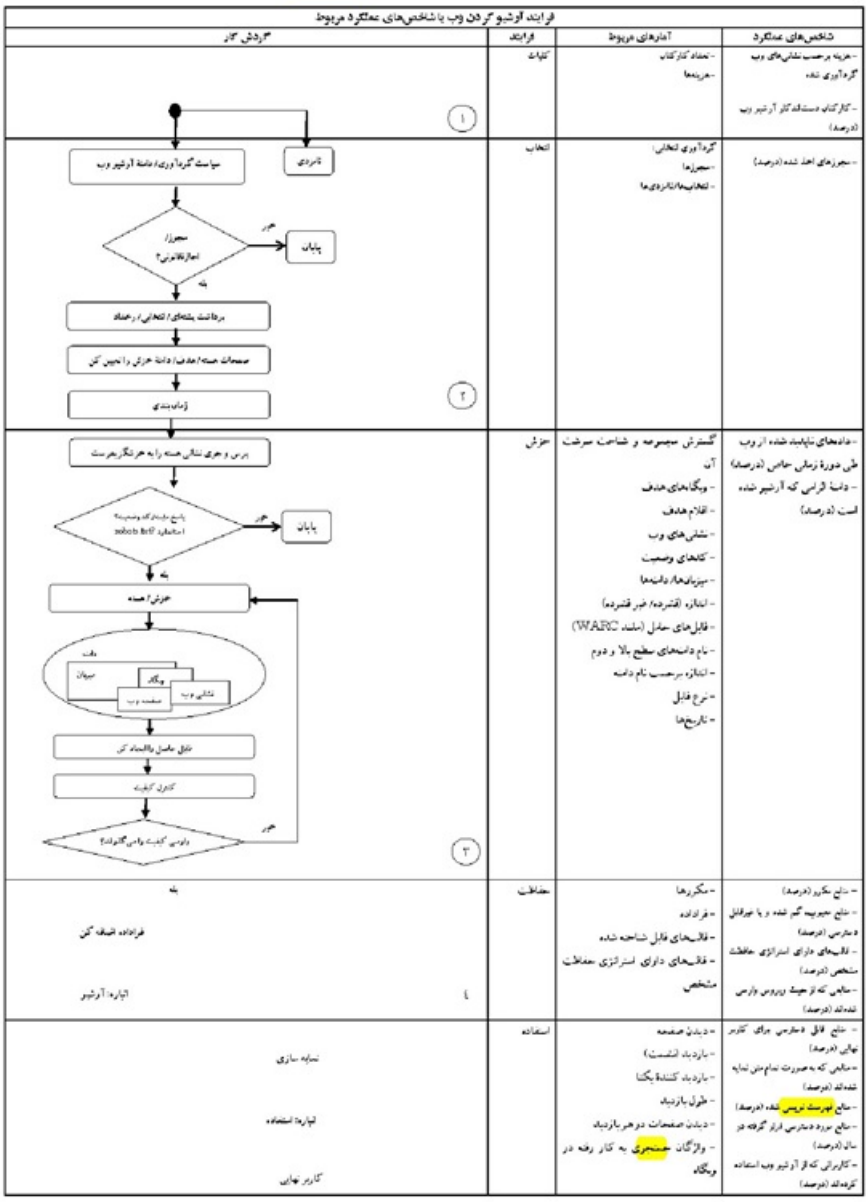
شکل ،2 آمارها و شاخص های کیفیت پیشنهاد شده در این گزارش –فنی برای نشان دادن اینکه

کدام یک به کدام فرایند آرشیو وب مربوط هستند در گردش کار در یک آرشیو وب نوعی را نشان داده می دهد.

ص: 109

عکس

شکل ۲. فرایند آرشیو کردن وب با شاخص های عملکرد مربوط



شکل ۲. فرایند آرشیو کردن وب با شاخص های عملکرد مربوط

شکل .2 فرایند آرشیو کردن وب با شاخصهای عملکرد مربوط

ص: 110

منابع

ISO 2789, Information and documentation –International library statistics [1]

ISO 9000:2005, Quality management systems–Fundamentals and vocabular [2]

ISO 11620, Information and documentation – Library performance indicators [3]

ISO 14721:2012, Space data and information transfer systems–Open archival information system [4]
(OAIS)–Reference model

ISO 15489–1:2001, Information and documentation–Records management–Part 1: General [5]

ISO 16439:–1), Methods and procedures for assessing the impact of libraries [6]

ISO 28500:2009, information and documentation– WARC file format [7]

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

/GLOBAL W.M.S.available from: http://www.ifabc.org [9]

TRANSFER PROTOCOL H. HTTP/1.1, available from: http://www.ietf.org/rfc/rfc2616.txt [10]

/ENCODING M., STANDARD T. (METS), available from: http://www.loc.gov/standards [11]

mets/METSOverview.html

//:MULTIPURPOSE Internet Mail Extension (MIME) Part Two: Media Types, available from: http [12]

www.ietf.org/rfc/rfc2046.txt

NESTOR CATALOGUE OF CRITERIA FOR Trusted DigitAL REPOSITORIES. available from: http://
files.d–nb.de/nestor/materialien/nestor_mat_08_eng.pdf

REVISED GUIDELINES FOR STATISTICAL MEASURES OF USAGE OF WEB–BASED INFORMATION RESOURCES. available from: http://icolc.net/statement/revised–guidelines–statistical–measures– usage–web–based–information–resources

/Trusted Repositories Audit Certification (TRAC), available from: http://www.crl.edu/sites [15]

default/files/attachments/pages/trac_0.pdf

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

.WEB ANALYTICS ASSOCIATION [17]

org/?page=standards

.available from: http://www.webanalyticsassociation

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

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

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

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

DOUGHERTY. M., MEYER, E.T., MADSEN, C., VAN DEN HEUVEL, C., THOMAS, A. and [22] WYATT, 2010. Researcher Engagement with Web Archives: State of the Art. Report. London: JISC, available from: http://ssrn.com/abstract=1714997

.(DROID (DIGITAL RECORD OBJECT IDENTIFICATION [23]

/net/projects/droid

.available from: http://sourceforge

:IIPC ACCESS WORKING GROUP. 2006: Use cases for Access to Internet Archives, available from [24]

http://www.netpreserve.org/resources/use-cases-access-internet-archives

JACOBSEN. G. 2007. Webarchiving Internationally: Interoperability in the Future? Results of [25]

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

JHOVE – JSTOR/HARVARD OBJECT VALIDATION ENVIRONMENT. available from: [26]
.http://jhove

/sourceforge.net

.MASANÈS. J. (ed.). 2006. Web Archiving, Springer, Berlin [27]

MASANÈS. J. 2002. Towards Continuous Web Archiving. In: D-Lib Magazine 8 (12), available [28]

from: http://www.dlib.org/dlib/december02/masanes/12masanes.html

:MEYER. E., THOMAS, A. and SCHROEDER, R. 2011: Web Archives: The Future(s), available from [29]

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1830025

PINSENT, E., DAVIS, R., ASHLEY, K., KELLY, B., GUY, M. and HATCHER, J. 2010. PoWR: The [30]
Preservation of Web Resources Handbook, available from: http://www.jisc.ac.uk/publications/
programmerelated/2008/powrhandbook.aspx

AUBRY. S. 2010: Introducing Web Archives as a New Library Service: the Experience of the National [31]
Library of France. In: Liber Quaterly, 2010, vol. 20, no. 2, available from: http://liber.
library.uu.nl/index.php/lq/article/view/7987

BAILEY. S. and THOMPSON, D. 2006: UKWAC: Building the UK's First Public Web Archive. D- [32]
.Lib

Available at: http://www.dlib.org/dlib/january06/thompson/01thompson.html .(1) 12 ,2006

GLENN. V. 2007: Preserving Government and Political Information: The Web-at-Risk Project. First [33]
Monday. 2007, 12 (7). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.
php/fm/article/view/1917/1799

HOCKX-YU. H. 2011: The Past Issue of the Web. In: Proceedings of the ACM WebSci'11, [34]
Webscience Trust, June 17, 2011, available from: http://www.websci11.org/fileadmin/websci/Papers/
PastissueWeb.pdf

HOCKX-YU. H., CRAWFORD, L.ROGER, C., JOHNSON, S. 2010: Capturing and Replaying [35]
Streaming Media 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/ipres 2010/papers/hockxyu-44.pdf

ILLIEN. G. 2008: L'archivage d'Internet, un dé fi pour les dé cideurs et les bibliothé caires: scé narios [36]
d'organisation et d'é valuation; l'expé rience du consortium IIPC et de la BnF. In: Actes du 74e congrè s de la
Fé dé ration internationale des associations de bibliothé caires et d'institutions (IFLA), Qué bec, Canada,
available from: http://archive.ifla.org/IV/ifla74/papers/107-Illien-fr.pdf

ILLIEN. G. and STIRLING, P. 2011: The state of e-legal deposit in France: looking back at five years [37]
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/ifla77/193- stirling-en.pdf

JACOBSEN. G. 2008: Web Archiving: Issues and Problems in Collection Building and Access. In: [38]
Liber Quarterly, Volume 18, Nr. 3/4 (2008), available from: http://liber.library.uu.nl/index.
php/lq/article/view/7936/8202

LA SFARGUES. F., OURY C. and WENDLAND B. 2008: Legal deposit of the French Web: [39] harvesting strategies for a national domain. In: Proceedings of the 8th International Web Archiving Workshop, Aarhus, Danemark, available from: http://iwaw.net/08/IWAW2008-Lasfargues.pdf

OURY C.. PEYRARD S. 2011: From the World Wide Web to digital library stacks: preserving the [40] French Web archives. In: Proceedings of iPRES 2011, p. 231-241, available from: http://halshs. archives-ouvertes.fr/halshs-00868729

POPE. J. and BERESFORD, Ph. 2007: IIPC Web Archiving Toolset Performance Testing at the British [41] /Library. In: Ariadne, no. 52 (2007), available from http://www.ariadne.ac.uk/issue52/ pope-beresford

RAUBER. A.ASCHENBRENNER,A,,WITVOET, O.,BRUCKNER, R. and KAISER, M. 2002: [42] Uncovering Information Hidden in Web Archives: A Glimpse at Web Analysis Building on Data Warehouses. D-Lib. 2002, 8 (12). Available at: http://www.dlib.org/dlib/december02/rauber/12rauber.html

SMITH. J., and NELSON, M. 2008: Creating Preservation-Ready Web Resources. D-Lib. 2008, 14 [43] (1/2). Available at: http://www.dlib.org/dlib/january08/smith/01smith.html

SPENCER. A., SHERIDAN, J., THOMAS, D. and PULLINGER, D. 2009: UK Government Web [44] Continuity: Persisting Access through Aligning Infrastructures. International Journal of Digital Curation. 2009, 4 (1). Available at: http://www.ijdc.n

درباره مرکز

بسمه تعالی

جَاهِدُواْ بِأَمْوَالِكُمْ وَأَنفُسِكُمْ فِي سَبِيلِ اللّهِ ذَلِكُمْ خَيْرٌ لَّكُمْ إِن كُنتُمْ تَعْلَمُونَ

با اموال و جان های خود، در راه خدا جهاد نمایید، این برای شما بهتر است اگر بدانید.

(توبه : 41)

چند سالی است که مرکز تحقیقات رایانه‌ای قائمیه موفق به تولید نرم‌افزارهای تلفن همراه، کتاب‌خانه‌های دیجیتالی و عرضه آن به صورت رایگان شده است. این مرکز کاملا مردمی بوده و با هدایا و نذورات و موقوفات و تخصیص سهم مبارك امام علیه السلام پشتیباني مي‌شود.

براي خدمت رسانی بیشتر شما هم می توانید در هر کجا که هستید به جمع افراد خیراندیش مرکز بپیوندید.

آیا می‌دانید هر پولی لایق خرج شدن در راه اهلبیت علیهم السلام نیست؟

و هر شخصی این توفیق را نخواهد داشت؟

به شما تبریک میگوییم.

شماره کارت :

6104-3388-0008-7732

شماره حساب بانک ملت :

9586839652

شماره حساب شبا :

IR390120020000009586839652

به نام : (موسسه تحقیقات رایانه ای قائمیه)

مبالغ هدیه خود را واریز نمایید.

آدرس دفتر مرکزی:

اصفهان -خیابان عبدالرزاق - بازارچه حاج محمد جعفر آباده ای - کوچه شهید محمد حسن توکلی -پلاک 129/34- طبقه اول

وب سایت: www.ghbook.ir

ایمیل: Info@ghbook.ir

تلفن دفتر مرکزی: 03134490125

دفتر تهران: 88318722 ـ 021

بازرگانی و فروش: 09132000109

امور کاربران: 09132000109


مرکز تحقیقات رایانه‌ای
قائمیه
اصفهان
www
برای داشتن کتابخانه های‌تخصصی
دیگر به سایت این مرکز به نشانی
www.Ghaemiyeh.com
www.Ghaemiyeh.net
www.Ghaemiyeh.org
www.Ghaemiyeh.ir
مراجعه و برای سفارش با ما تماس بگیرید.
۰۹۱۳ ۲۰۰۰ ۱۰۹