TECHNICAL REPORT

ISO/TR 14873

First edition
2013-12-01

Information and documentation — Statistics and quality issues for web archiving

Information et documentation — Statistiques et indicateurs de qualité pour l'archivage du web

Reference number
ISO/TR 14873:2013(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 46, *Information and documentation*, Subcommittee SC 8, *Quality - Statistics and performance evalutation*.

Introduction

This Technical Report was developed in response to a worldwide demand for guidelines on the management and evaluation of Web archiving activities and products.

Web archiving refers to the activities of selecting, capturing, storing, preserving and managing access to snapshots of Internet resources over time. It started at the end of the 1990s, based on the vision that an archive of Internet resources would become a vital record for research, commerce and government in the future. Internet resources are regarded as part of the cultural heritage and therefore preserved like printed heritage publications. Many institutions involved in Web archiving see this as an extension of their long standing mission of preserving their national heritage, and this is endorsed and enabled in many countries by legislative frameworks such as legal deposit.

There is a wide range of resources available on the Internet, including text, image, film, sound and other multimedia formats. In addition to interlinked Web pages, there are newsgroups, newsletters, blogs and interactive services such as games, made available using various transfer and communication protocols. Web archives bring together copies of Internet resources, collected automatically by harvesting software, usually at regular intervals. The intention is to replay the resources including the inherent relations, for example by means of hypertext links, as much as possible as they were in their original environment. The primary goal of Web archiving is to preserve a record of the Web in perpetuity, as closely as possible to its original form, for various academic, professional and private purposes.

Web archiving is a recent but expanding activity which continuously requires new approaches and tools in order to stay in sync with rapidly evolving Web technology. Determined by the strategic importance perceived by the archiving institution, means available and sometimes legal requirements, diverse approaches have been taken to archive Internet resources, ranging from capturing individual Web pages to entire top-level domains. From an organisational perspective, Web archiving is also at different levels of maturity. While it has become a business as usual activity in some organisations, others have just initiated experimental programmes to explore the challenge.

Depending on the scale and purpose of collection, a distinction can be made between two broad categories of Web archiving strategy: bulk harvesting and selective harvesting. Large scale bulk harvesting, such as national domain harvesting, is intended to capture a snapshot of an entire domain (or a subset of it). Selective harvesting is performed on a much smaller scale, is more focused and undertaken more frequently, often based on criteria such as theme, event, format (e.g. audio or video files) or agreement with content owners. A key difference between the two strategies lies in the level of quality control, the evaluation of harvested Websites to determine whether pre-defined quality standards are being attained. The scale of domain harvesting makes it impossible to carry out any manual visual comparison between the harvested and the live version of the resource, which is a common quality assurance method in selective harvesting.

This Technical Report aims to demonstrate how Web archives, as part of a wider heritage collection, can be measured and managed in a similar and compliant manner based on traditional library workflows. The report addresses collection development, characterization, description, preservation, usage and organisational structure, showing that most aspects of the traditional collection management workflow remain valid in principle for Web archiving, although adjustment is required in practice.

While this Technical Report provides an overview of the current status of Web archiving, its focus is on the definition and use of Web archive statistics and quality indicators. The production of some statistics relies on the use of harvesting, indexing or browsing software, and a different choice of software may lead to variance in the results. This Technical Report however does not endorse nor recommend any software in particular. It provides a set of indicators to help assess the performance and quality of Web archives in general.

This Technical Report should be considered as a work in progress. Some of its contents are expected to be incorporated in the future into ISO 2789 and ISO 11620.

Information and documentation — Statistics and quality issues for web archiving

1 Scope

This Technical Report defines statistics, terms and quality criteria for Web archiving. It considers the needs and practices across a wide range of organisations such as libraries, archives, museums, research centres and heritage foundations. The examples mentioned are taken from the library sector, because libraries, especially national libraries, have taken up the new task of Web archiving in the context of legal deposit. This should in no way be taken to undermine the important contributions of institutions which are not libraries. Neither does it reduce the principal applicability of this Technical Report for heritage institutions and archiving professionals.

This Technical Report is intended for professionals directly involved in Web archiving, often in mixed teams consisting of library or archive curators, engineers and managerial staff. It is also useful for Web archiving institutions' funding authorities and external stakeholders. The terminology used in this Technical Report attempts to reflect the wide range of interests and expertise of the audiences, striking a balance between computer science, management and librarianship.

This Technical Report does not consider the management of academic and commercial electronic resources, such as e-journals, e-newspapers or e-books, which are usually stored and processed separately using different management systems. They are regarded as Internet resources and are not addressed in this Technical Report as distinct streams of content of Web archives. Some organisations also collect electronic documents, which may be delivered through the Web, through publisher-based electronic deposits and repository systems. These too are out of scope for this Technical Report. The principles and techniques used for this kind of collecting are indeed very different from those of Web archiving; statistics and quality indicators relevant for one kind of method are not necessarily relevant for the other.

Finally, this Technical Report essentially focuses on Web archiving principles and methods, and does not encompass alternative ways of collecting Internet resources. As a matter of fact, some Internet resources, especially those that are not distributed on the Web (e.g. newsletters distributed as e-mails) are not harvested by Web archiving techniques and are collected by other means that are not described nor analysed in this Technical Report.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
access
successful request of a library-provided online service

Note 1 to entry: An access is one cycle of user activities that typically starts when a user connects to a library-provided online service and ends by a terminating activity that is either explicit (by leaving the database through log-out or exit) or implicit (timeout due to user inactivity).

Note 2 to entry: Accesses to the library website are counted as virtual visits.

Note 3 to entry: Requests of a general entrance or gateway page are excluded.

Note 4 to entry: If possible, requests by search engines are excluded.

[SOURCE: ISO 2789:2013, definition 2.2.1]