HOW CAN WE USE WEB ARCHIVE?

A brief overview of WARP and how it is used at the National Diet Library, Japan

HARVESTING AND ACCESS

Legal Deposit

The National Diet Library Law, Article 25-3, allows the NDL to harvest websites of public agencies, including those of the national government, municipal governments, public universities, and independent administrative agencies. WARP uses web crawlers to harvest content periodically and has thus far accumulated a total of 1 petabyte of archived content.

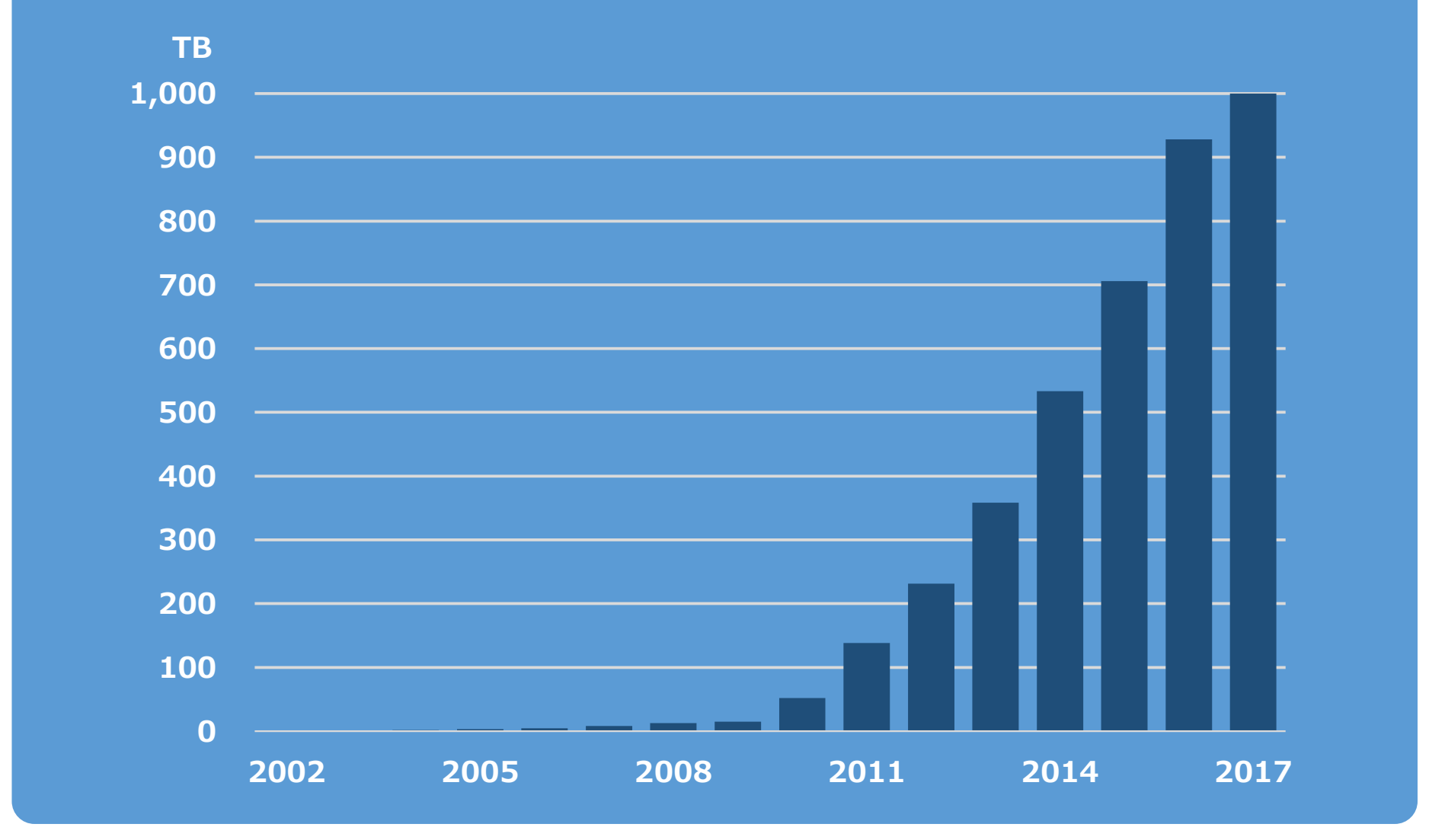
Permission

Legal deposit does not allow the NDL to harvest websites of private organizations, so we need to receive permission from the copyright holder beforehand. We target primarily the websites of foundations, associations, political parties, cultural and international events, private museums and academic institutions.

1 Petabyte

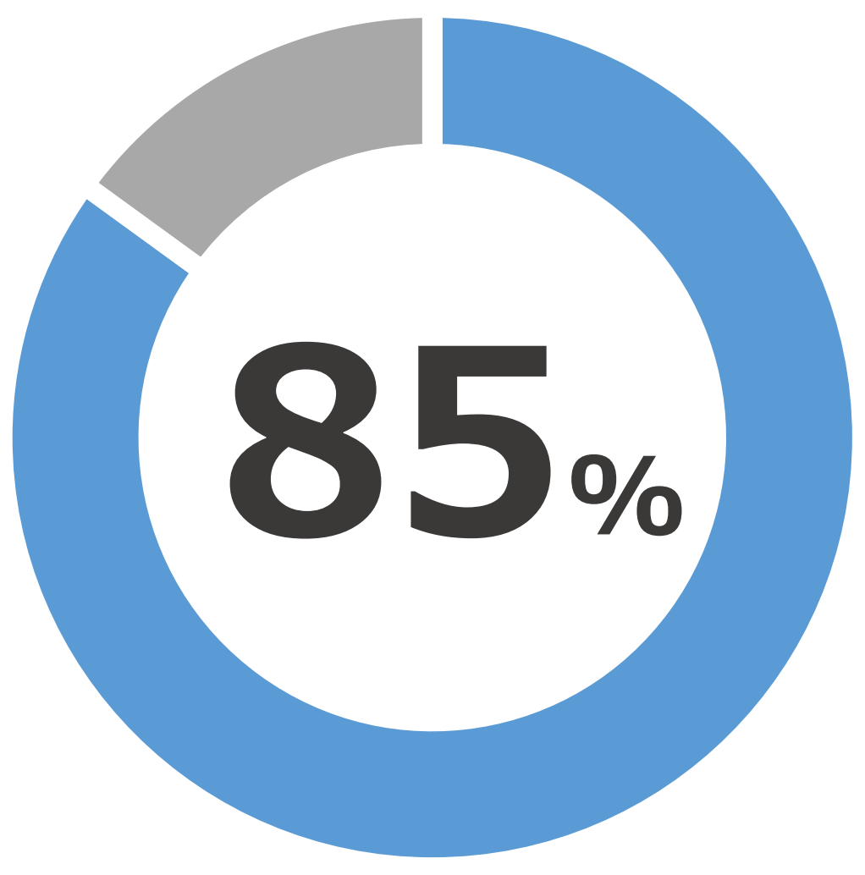
5 Billion Files

130,000 Captures



Open Access

85% of the archived websites are freely available to the public via the Internet, and the remainder can be accessed at the NDL. WARP provides a variety of search methods, including URL, full text, metadata, and by category.



Technologies

WARP uses standard web archiving technologies, such as Heritrix for web-crawling, WARC file format for storage, OpenWayback for playback, and Apache Lucene Solr for full text search. With the exception of Solr, these technologies were developed by the International Internet Preservation Consortium and are distributed freely as open source software.

HERITRIX

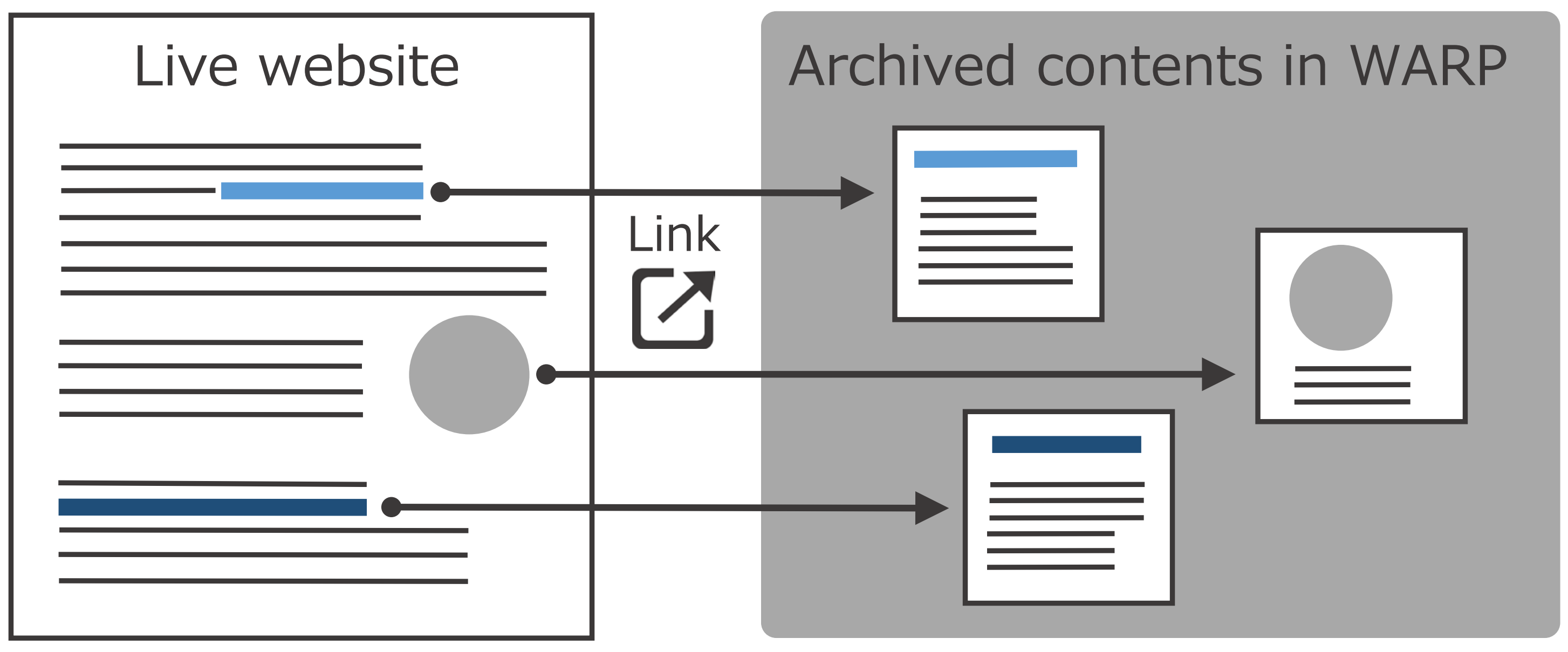
OpenWayback

Solr

WARC file format
(ISO 28500:2009)

USE CASES

Linking from live websites

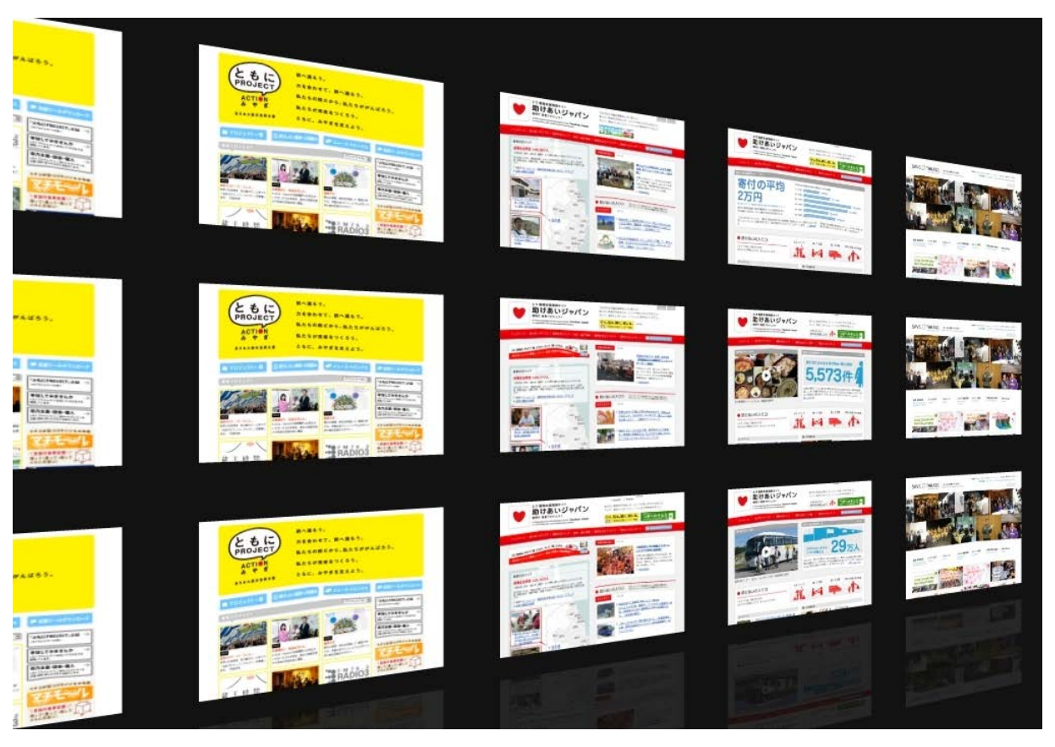


WARP comprehensively harvests and archives the websites of public agencies under the legal deposit system. A significant quantity of content is posted, updated, and deleted on these websites every day. Many of these agencies use WARP as a backup database. Before deleting content from their websites, they add a link to content that is archived by WARP. Doing this enables these websites to keep archived content seamlessly available while also reducing the operating costs of their own web servers.

Curation

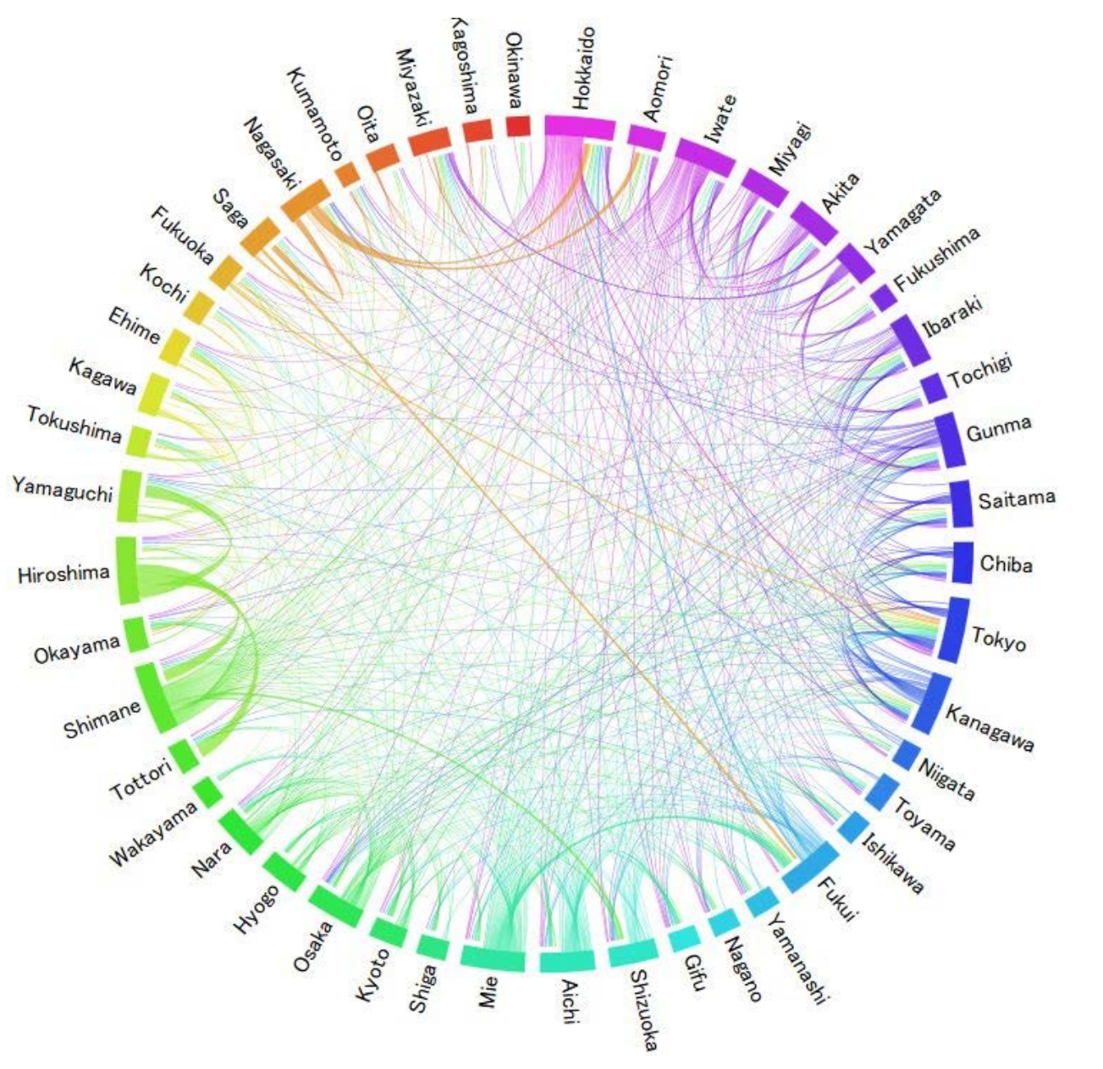
Patrons can use a variety of search methods to find content of interest archived in WARP, but it is not easy for them to gauge the full extent of archived content. The NDL curates archived contents for a variety of subjects and provides visual representations that could provide patrons with unexpected discoveries.

Search by region for obsolete websites of defunct municipalities.

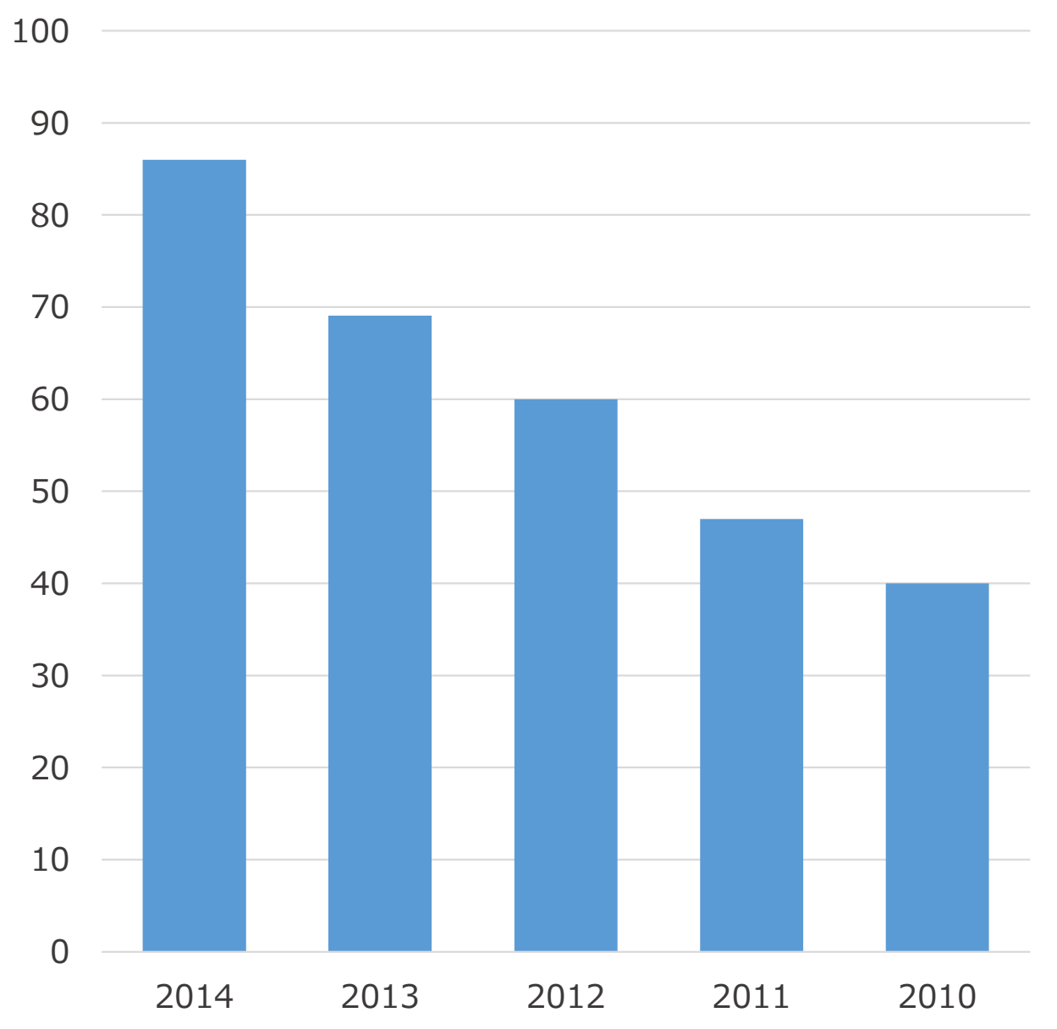


3D wall for the collection of the Great East Japan Earthquake in 2011

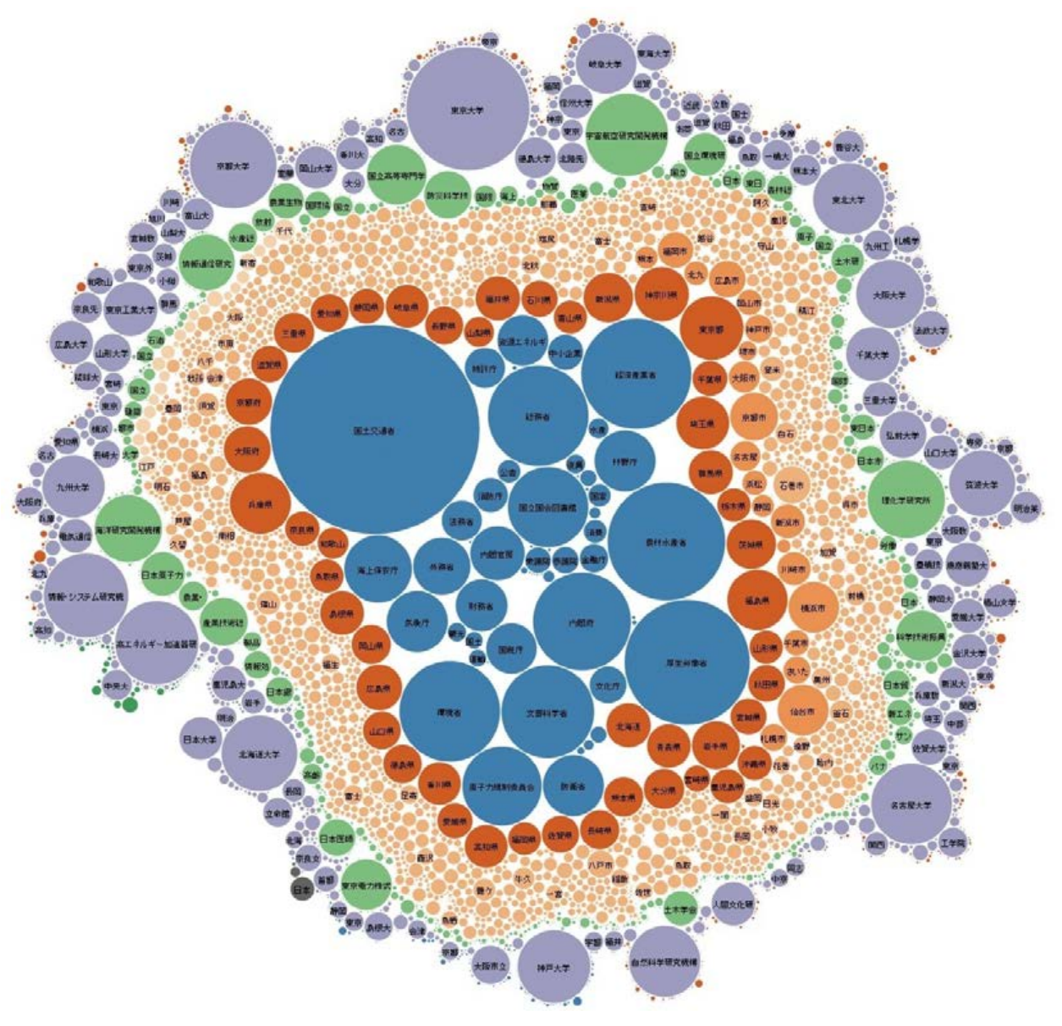
Analysis and Visualizations



The circular graph illustrates link relations between websites in Japan's 47 prefectures, thereby showing the extent of their interconnection on the Web.

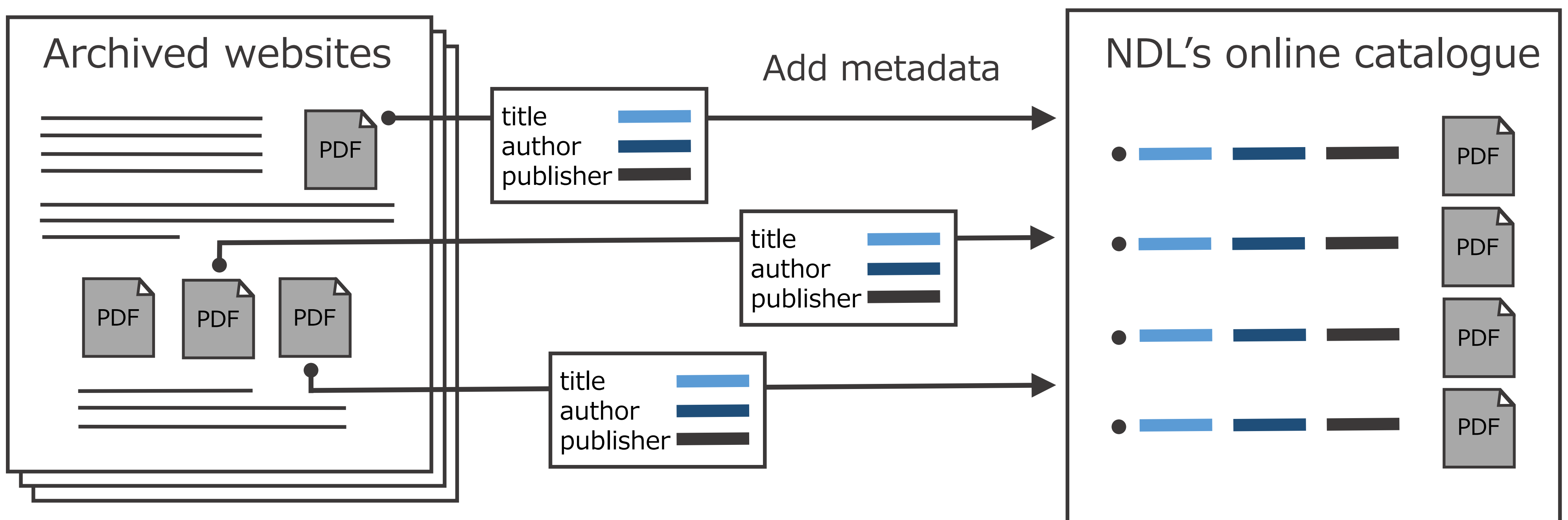


The percent of URLs on websites of the national government that were live in 2015. 60% of the URLs that existed in 2010 gave 404 errors during 2015.



The bubble chart shows the relative size of data accumulated from each of the 10,000 websites archived in WARP.

Uncovering PDF Documents



The websites that are archived in WARP contain many PDF files of books and periodical articles. The NDL searches for these publications and adds metadata in DC-NDL format, which is a significantly enhanced version of Dublin Core. These PDF files with metadata are then integrated into the NDL's online catalogue, so that patrons can find them using conventional search methods.

1 Million Records

1.4 Million Files

FUTURE CHALLENGES

Data Mining

Web archives have tremendous potential for use in big data analysis. Archived data is one form of cultural property that contains a vast output from intellectual activities, and analysis of this data could help uncover how human history has been recorded in cyberspace. The NDL is studying how to make data sets suitable for data mining and how to promote engagement with researchers.

Robust Search

Full-text searches of web archives can be a significant challenge, because of poor indexing performance and the large numbers of files that yield "noisy" search results. WARP provides full-text search with Apache Lucene Solr, and has already indexed 2.5 billion files in the creation of indexes totaling 17 terabytes. But we are not satisfied with search results, which contain duplicate material archived at different times and other "noise." We need to develop a robust and accurate search engine specialized for web archives that uses temporal elements.

The National Diet Library, Japan