Web Archiving Workshop
Mois des archives 2023

The web is ephemeral and ever changing. With a growing emphasis on digitisation in all aspects of society, more and more information is solely published online, making its distribution and accessibility faster and easier. On the other hand, the lack of physical copies of publications accentuates the dangers of data loss. Relevant information that are is printed or preserved in any other shape or form could be lost, depriving future generations of the sources of knowledge available to us today.

404 Not Found, is the error code which we encounter when links to websites don't work anymore. A common misconception is that the contents of the Internet will stay available for eternity, while in reality, 80 % of web pages are not available in their original form after 1 year, 13 % of web references in scholarly articles disappear after 2 years and 11 % of social media resources are lost after 1 year. Images, videos, documents, most of the information online, will eventually vanish within a few years.

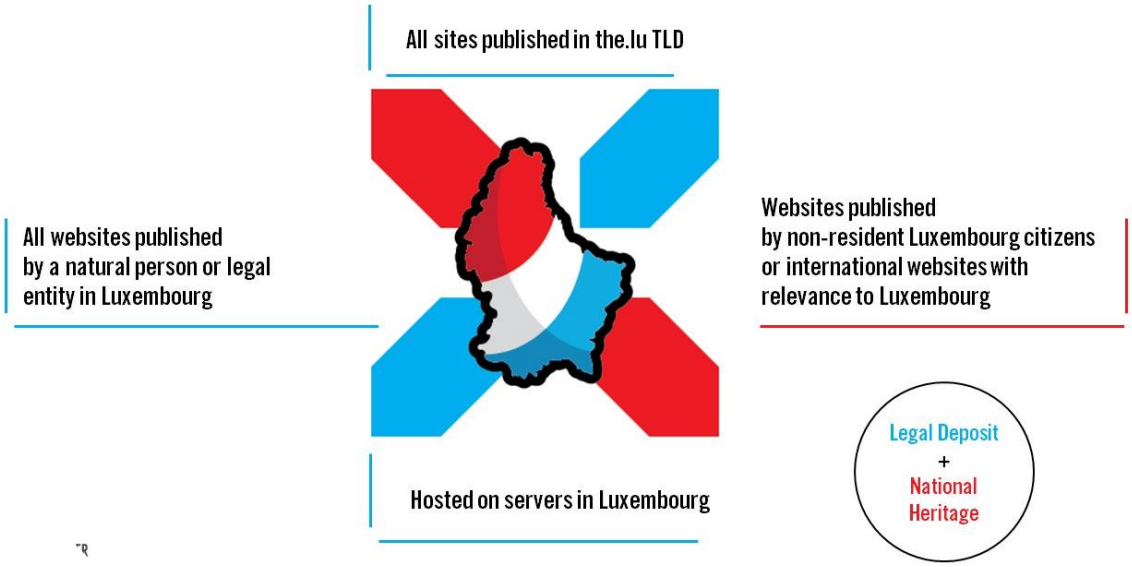
Preserving the web and legal deposit

Web archiving is a relatively new field in digital preservation. The BnL has operated the first .lu domain crawl in 2016. While many countries don't have a national web archive yet, they are most often operated by the country's national library, as is also the case in Luxembourg.

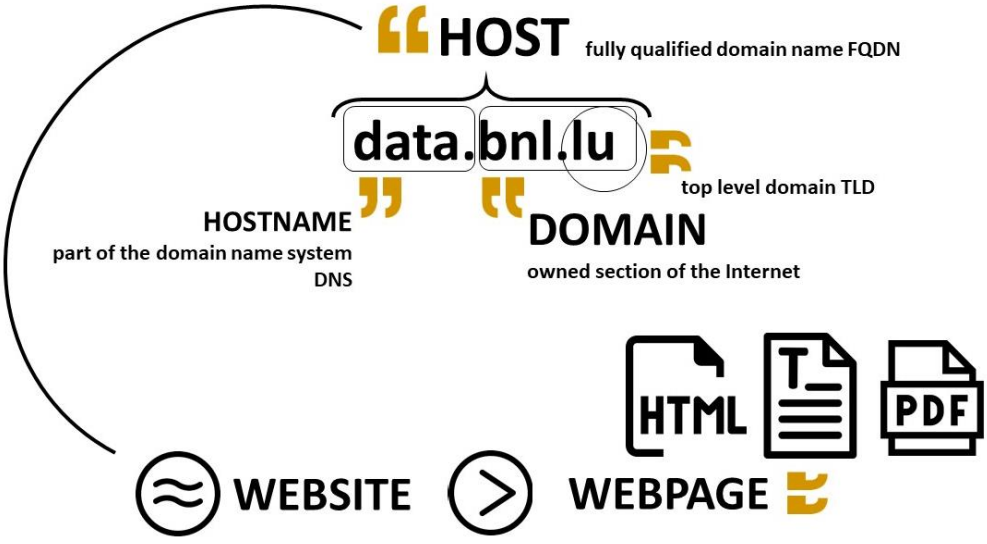
The legal deposit and preservation of all types of publications, is at the core of the BnL's activities. The legal deposit for the Luxembourg web is an extension of the same legal basis for traditional publications on paper and is defined to include the following areas of interest:

- All websites in the ".lu" top level domain
- Websites published in Luxembourg
- Websites hosted on servers in Luxembourg
- Websites published by non-resident Luxembourg citizens or websites in relation with Luxembourg

Although the last category of websites is not part of the Luxembourg legal deposit, the National Library might choose to include websites that are relevant to the national digital heritage.

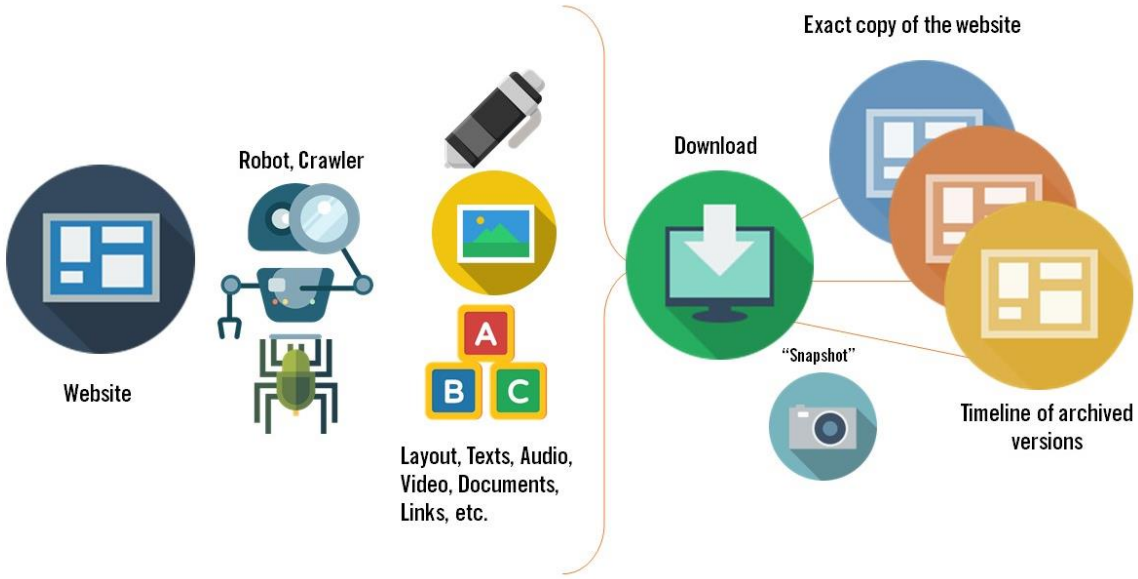


It is important to know, that the Luxembourg Web Archive is preserving the *web*, which is only a part of the *Internet*. More specifically: we are looking at the surface web, which is websites that are freely accessible to everyone in the world, not protected by authentication and indexed by search engines. There are different definitions to what a *website* is, but the most common understanding is that a domain can include several hosts, which can be individual websites. Every website is made up of one or more webpages.



Methods and collections

In order to preserve the contents of the web, we need to do more than just saving screenshots of webpages. Archiving the web means to create an exact copy of the original website. The web crawler will run over a *seed* (an address serving as a starting point for the crawler), to look for all elements of that website: documents, links, media and layout. Everything is downloaded to create an exact copy of the original at the moment of capture: a snapshot in time. By creating regular captures of the same website, we build a timeline of archived versions, which allow us to travel back and forth in time to follow the changes made to a website.



Even if the Luxembourg Web Archive is only concerned with a small portion of the Internet, it is clear that we cannot archive every website for every second of every day. Crawling websites takes time and every capture is linked to costs in harvesting tools and storage.

In order to achieve the best possible coverage of our areas of interest and preserve as much information as possible, we are using different methods to harvest the Luxembourg web:

Domain crawls

Domain crawls are operated twice a year and create a snapshot of the all ".lu" addresses plus additional lists of websites determined by the Luxembourg Web Archive.

These crawls cover a large number of websites at once, but can be tardy in capturing sites that are changing at a rapid pace or may have disappeared between two harvests.

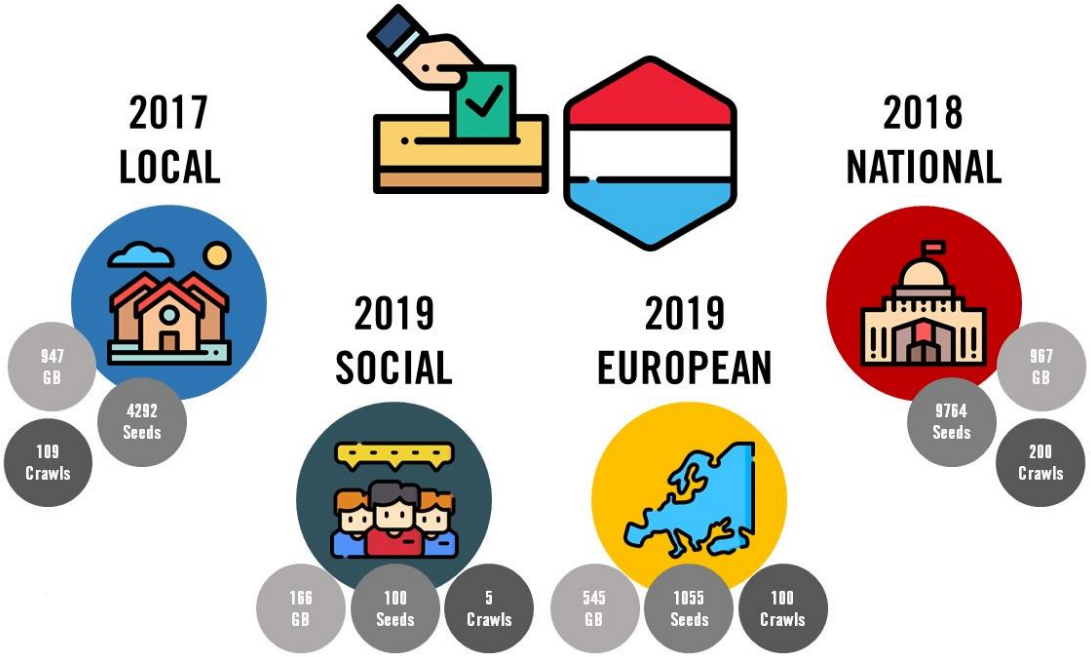
Event collections

Event collections try to harvest as much information as possible about a certain event over a limited time frame. The seed lists for event crawls are restrictive, but the frequency of captures will likely be higher. There is always a start and end date to event crawls, which could be determined in advance, (e.g. for elections), or could depend on the urgency of surprising events (e.g. natural catastrophes or Covid-19).

Thematic collections

Thematic collections cover a specific topic or field of interest, with a higher priority to the Luxembourg Web Archive. This could be linked to the pace of changing information, or the importance of the topic. The seed lists will expand over time and have additional harvests, complementing the coverage by domain crawls and event collections.

Over the past years, we have been focusing on event collections, surrounding the different election campaigns from 2017 to 2019.



Collaborative collecting and contributions

In the near future, we would like to focus more on thematic collections, which will evolve over time and emphasize the importance of collaborating with subject experts, website owners and content creators. At the same time, the combination of different methods and collection types, will help to improve our knowledge and coverage of the Luxembourg web. Starting from event collections, we use the research from the events to expand the seed list for domain crawls and therefore, also add another layer of coverage to those events. These seed lists also serve as a base to extract the subject areas for thematic collections. In turn, the thematic collections will serve as a base to start event collections, and save time on research. In time, event collections will help with a more intense coverage for the subjects of thematic collections and the latter will capture information before and after the topic of an event collection. For example, the seed list from an election crawl can serve as a basis for the thematic collection “Politics & Society”. The continued coverage and expansion from this collection will serve as an improved basis for a seed list, once the next election campaign comes around. Moreover, both types of collections will help in broadening the scope of domain crawls and achieve better coverage of the Luxembourg web.

While the preparations for these thematic collections are still ongoing, we highly encourage all suggestions and contributions to our seed lists. We rely on website owners and content creators to let us know about their addresses, especially if their websites are registered in a different top level domain as .lu. On webarchive.lu, you’ll find a very simple form un der *Contact,* and we hope that the participants of this year’s Summer School will also help us out in preserving the Luxembourg digital heritage.

Links and further resources:

Internet Archive:
https://archive.org/

Save Page Now :
https://web.archive.org/save

Portuguese Web Archive :
https://arquivo.pt/

Webrecorder :
https://archiveweb.page/

International Internet Preservation Consortium :
https://netpreserve.org/

List of web archiving initiatives around the world:
https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives