

404 NOT FOUND

不定期闲聊

第 1 期

什么情况下会被 404?



此内容因违规无法查看

接相关投诉，此内容违反《互联网用户公众帐号信息服务管理规定》，查看详细内容

- 网站更新 URL 发生变化

- 如：http://old-domain.com/ -> http://new-domain.com/
- 如：http://a.com/page.php?id=10 -> http://a.com/page/10

- 网站关闭

- 如：http://www.wooyun.org/

- 内容被网站/创作者删除

- 如：高校的公示新闻（研究生复试名单公示）
- 如：微信公众号中某些文章



该内容已被发布者删除



已经被 404 怎么办？

- 唉，要是我当时把内容存起来就好了。

或许，那些重要的信息，
已经有人保存过了呢！？

已经被 404 怎么办？

- 档案馆

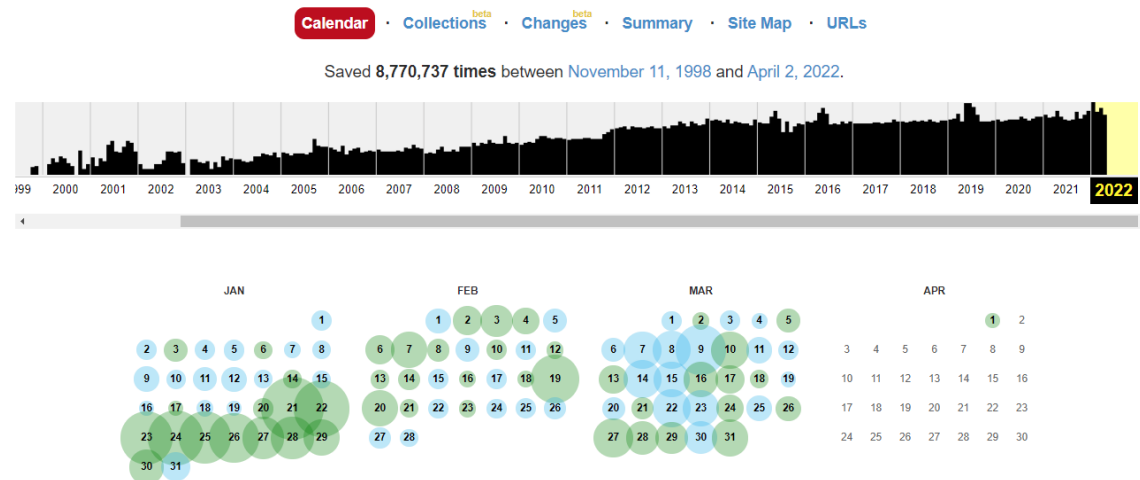
- 互联网档案馆 web.archive.org
- archive.today
- ArchiveBox（自建互联网档案馆）

- 缓存

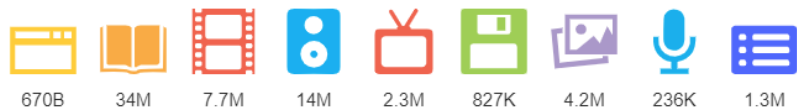
- 搜索引擎缓存 [cache:https://www.google.com/](https://www.google.com/cache:https://www.google.com/)
- RSS 缓存
- 搜索引擎找副本/转载

档案馆 - 互联网档案馆 web.archive.org

- 资料最全 (over **670 Billion** web pages)
- 定期自动保存 (收费) + 手动保存
- 可浏览**指定日期**的快照
 - 前提是那天曾经有人保存过该页面
- 有**浏览器插件**，直接点一下就能保存
- 严格遵循 robots.txt



Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.



Search

GO

Advanced Search

INTERNET ARCHIVE
WayBackMachine

DONATE

Explore more than 670 billion web pages saved over time

Enter a URL or words related to a site's home page

Results: 50 100 500



Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit [Archive-It](#) to build and browse the collections.



Collection Search

Enter any keyword

PDFs

SEARCH

This service is based on indexes of specific data from selected Collections.



Save Page Now

https://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

档案馆 - 互联网档案馆 web.archive.org

- 看看 1998 年时的谷歌长什么样子吧!

http://www.google.com/ NOV DEC JAN 1997 1998 2000

8,770,734 captures
11 Nov 1998 - 2 Apr 2022

Google!
B E T A

Search the web using Google!

Special Searches
[Stanford Search](#)
[Linux Search](#)

[Help!](#)
[About Google!](#)
[Company Info](#)
[Google! Logos](#)

Get Google!
updates monthly:

 [Archive](#)

Copyright ©1998 Google Inc.

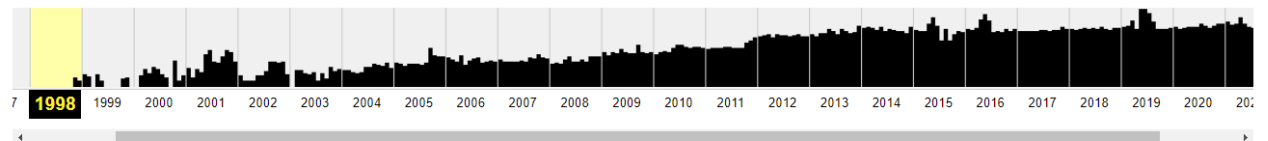
INTERNET ARCHIVE Explore more than 670 billion web pages saved over time

WayBackMachine

Results: 50 100 500

· [Collections](#) ^{beta} · [Changes](#) ^{beta} · [Summary](#) · [Site Map](#) · [URLs](#)

Saved 8,770,734 times between November 11, 1998 and April 2, 2022.

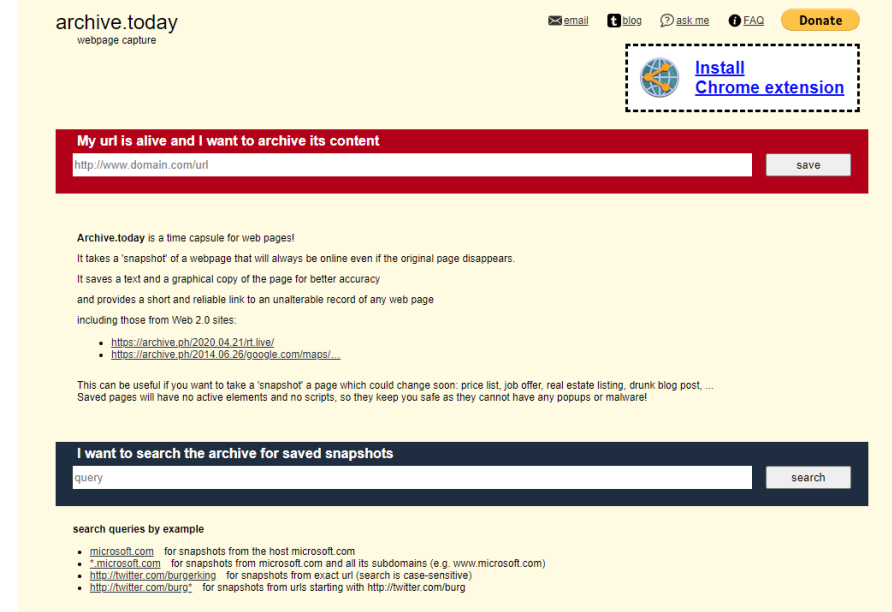


档案馆 - 互联网档案馆 web.archive.org

- 我的敏感信息**已经被缓存了**，怎么办？
 - 发送邮件说明并证明你的身份
- 我**不想以后再被缓存**，怎么办？
 - 在你的网站 robots.txt 中添加规则并邮件/发帖告知档案馆
 - 档案馆的爬虫会严格遵循 robots.txt

档案馆 - archive.today

- 对于**动态网页**（如：知乎，微博）支持较好
- 可**搜索**某域名下的所有快照
- 有**浏览器插件**，直接点一下就能保存



web.archive.org 无法显示评论



archive.today 可以正常显示评论

档案馆 - [ArchiveBox](#) (自建互联网档案馆)

- [开源](#)
- Django 写的, **搭建方便** (几条命令)
- 可以指定其**定期自动爬取**
- 可以**批量爬取**
- 接口丰富 CLI / WEB / RESTful / Desktop App



缓存 - 搜索引擎缓存

- Google
 - 无法选择指定日期的快照，仅能看到最新的网页版本
 - 一旦谷歌的爬虫重新爬了目标网页，则 404 也会被缓存下载
 - 使用方法：
 - `cache:https://www.baidu.com/`
- 不如档案馆靠谱，但是档案馆如果没找到的话，可以一试

缓存 – RSS 缓存

- 有些网站的 RSS 是包含全文而非只有摘要的
- 因此 RSS 阅读器（如：Feedly）如果爬了，就能留下全文



缓存 – 搜索引擎找副本/转载

- 重要的资料通常都会有很多人转载保存
- 直接搜索标题/作者

我们能做点什么呢？

前人栽树，后人乘凉。

那个后人极有可能就是未来的我们自己。

我们能做点什么呢？

- 对于还存活的珍贵资料，**顺手保存**
- 向身边的人**小范围传播**这些方法
 - 知道这些方法的人越多，那些宝贵的信息越有可能被保存下来

感谢倾听