

# Archivage des données du web : sauvegarde et citabilité pour la recherche

Jean-Christophe Peyssard (CNRS – MMSH UAR 3125)  
Ecole d'été DISTAM Juillet 2022



Maison méditerranéenne  
des sciences de l'homme  
UAR3125



1. Courte introduction aux archives du Web
2. Des outils pour la collecte

# The World Wide Web by numbers

- “The [first-ever website \(info.cern.ch\)](http://info.cern.ch) was published on August 6, 1991 by British physicist Tim Berners-Lee while at CERN, in Switzerland”
- In the **June 2022** survey we received responses from **1,146,976,964** sites across **273,010,403** unique domains and **12,224,786** web-facing computers. This reflects a loss of 8.75 million sites and 583,000 domains, but a gain of 155,000 computers. [from previous 30th May, 2022], <https://news.netcraft.com/archives/2022/05/30/may-2022-web-server-survey.html>
- “The Indexed Web contains **at least 5.36 billion pages**” (Monday, 04 July, 2022, <https://www.worldwidewebsite.com/>)

# URL : Uniform Resource Locator

The structure of URLs on the World Wide Web (www):

**protocol://subdomain.domain.top-level domain/path/page/**

Ex:

**<https://distam.hypotheses.org/category/ecole-dete>**

<https://en.wikipedia.org/wiki/URL>

<http://dac.au.dk/forskning/forskningsprogrammer/> p. 51

## Life and death of Web pages

- “The average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999. [...] As ubiquitous as the Web seems to be, it is also ephemeral, and much of today's Web will have disappeared by tomorrow.” (Lyman, 2002 p. 38)
- “40% of the material on the Internet disappears within a year, while another 40% has been changed, which is why today we can only expect to find 20% of the material that was on the Internet one year ago.” (Brügger, 2005 p. 15)
- “We now know that Web pages only last about 100 days on average before they change or disappear.” (Kahle, 2015)
- In 2013, the average life span of a URL is 9.3 years (Musiani at al., 2019, <https://books.openedition.org/oep/8743>)
- In 2019, according to the Wayback Machine team the average life span of a Web page is 92 days (2018)
- “Since January 2013, 66.5% of the links pointing to the 2,062,173 websites we sampled have rotted. We found another 6.45% with temporary errors. We don't know if they're still there or not.” (Stox, 2022, <https://ahrefs.com/blog/link-rot-study/>)

# When the Web is broken

## A problem for research and scholarly communication

- Error 404, Broken links, Link rot, Reference rot, Infosuicide, digital ruins, content drift, zombie media,..
- Shut down & take down, mergers and acquisitions:
  - On March 18, 2019, it was revealed that MySpace lost all of their user content from 2016 and earlier in "a server migration gone wrong". It was widely reported that over 50 million songs and 12 years worth of content was permanently lost, and there was no backup (<https://en.wikipedia.org/wiki/Myspace>)
- History :
  - Yougoslavia (.yu) breakup (now Serbia and Montenegro, .rs and .me)
  - Czechoslovakia (.cs) dissolution (now Czech Republic and Slovakia, .cz and .sk)
- **Reference rot**, a combination of:
  - **Content decay**: The content of the linked resource may change over time and, as a result, the degree to which that content remains representative of the content that was intended to be linked to may decrease over time.
- **Link rot**: The linked resource may disappear altogether. (Thoughts on Referencing, Linking, Reference Rot <http://mementoweb.org/missing-link/>)
- The integrity of research is at risk ! (James G. Neal, <http://library.ifla.org/id/eprint/907>)

## Why Web Archiving?

- To maintain our digital cultural heritage
- To stabilize and preserve web materials as a research object
- To be able to document and illustrate claims based on analyses of web materials (whether the web itself is the research object or a source of knowledge about other research objects).

Nielsen, J. (2016). *Using web archives in research: an introduction*. Retrieved from <http://www.worldcat.org/oclc/960018046> p. 7

## A definition of Web archiving

- Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use. (IIPC Web site, <http://netpreserve.org/web-archiving/>)
- “Web archiving is the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research.” (Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18 (3/4). <https://doi.org/10.1045/march2012-niu1>)

## A short chronology of Web archives

**1537** Legal deposit in France (1619 Spain, 1710 United Kingdom)

**1989** The World Wide Web was invented by Tim Berners-Lee

**1996** Internet Archive is founded by Brewster Kahle

**1996** Kulturarw3 in Sweden for archiving the .se top level domain name

**1998** Google is launched

**2001** The Wayback Machine (Internet Archive)

**2003** UNESCO Charter on the Preservation of Digital Heritage

**2003** the International Internet Preservation Consortium is formally chartered at the National Library of France with 12 participating institution

**2005** Youtube is launched

**2006** [The] Facebook and Twitter are launched

**2006** National Library of France is in charge of collecting and preserving the “French Internet” (new French Copyright Law) alongside with the National Audiovisual Institute (Ina)

**2013** EU Web Archive

# Different strategies & methods for Web archiving

- Macro archiving
- Micro archiving
- Thematic or selective archiving
- Bulk or snapshot harvesting, or broad crawls
- “Exhaustivity” of a National domain name archiving (.se in Sweden)
- Event and ‘real-time’ institutional archiving (after 2015 terrorist attacks and Notre Dame fire in 2019 France)
- Shared archiving among institutions (in France btwn BnF and Ina)
- ...



**BnF** **Dépôt légal web BnF**   
@DLwebBnF Abonné

Merci à [@BNE\\_biblioteca](#) et à toutes les institutions membres de [@NetPreserve](#) qui ont participé à cette collecte par le biais d'une sélection de sites web !

**Biblioteca Nacional de España**  [@BNE\\_biblioteca](#)  
Estamos colaborando con la Biblioteca Nacional de Francia en una recolección web sobre el incendio de Notre Dame  
Afficher cette discussion

17:02 - 25 avr. 2019

8 Retweets 15 J'aime             

  8  15 

# Different ways of accessing Web archives

- Finnish Web Archive (since 2006) <https://www.kansalliskirjasto.fi/en/collections-and-content-online#finnish-web-archive>

The contents of the archive can be only accessed from special legal deposit workstations that are available in selected libraries within Finland (including The National Library of Finland).

- Portugal (since 2008) accessible in Open Access at [Arquivo.pt](https://arquivo.pt)
- The Wayback Machine (since 2006) accessible in Open Access at <https://archive.org/web/>
- EU Web Archive : <https://op.europa.eu/en/web/euwebarchive/>

# There is a lot of different types of Web archives

As much as for other kind of archives, one must know the history of a Web archive and how it was constructed to better understand it and use it in research work. What you see is a reconstruction, not a copy of the site

- “What is harvested is both a point in time (the time of harvesting) and a period of time (the period up to the time of harvesting).” (Brügger, 2008 p. 158)
- “On the one hand the archive does not look like the internet as it actually was in the past (we have lost something), but on the other hand the archive might look like the internet as it never was in the past (we get something different).” (Brügger, 2001 p. 6)

Web archiving projects often needs to gather diverse and multiples expertises and skills : archivists and librarians, researchers, legal officers, IT and computer specialists and... users and stakeholders

# Legal and ethical issues

As for any other kind of archives one must act lawfully and ethically when archiving and using Web archives:

- The materials in the web archives are protected by copyright law as they were on the live web
- There is “tensions around the archival principles of preserving the public record vs the individual’s expectation of the right to be forgotten” (<http://netpreserve.org/ga2018/programme/abstracts/#paper21>)
- The processing of personal data is submitted to laws and even more to the research project ethics
- New laws to take into account ex. General Data Protection Regulation (GDPR, <https://gdpr-info.eu/>)

# IIPC: International Internet Preservation Consortium



INTERNATIONAL  
INTERNET  
PRESERVATION  
CONSORTIUM

[HOME](#) [ABOUT IIPC](#) [WEB ARCHIVING](#) [EVENTS](#) [BLOG](#) [JOIN US](#)



The web is a unique  
and dynamic resource that  
is of high value to current  
and future researchers



[Learn about the value of our work](#)

## IIPC Members

Members are organizations from over 45 countries, including national, university and regional libraries and archives.

## Working groups

IIPC members join **working groups** that engage in short and long-term **projects** to advance the practice of **web archiving**.

## Events

Our community comes together annually to share experiences and present solutions during the **Web Archiving Conference** and the **General Assembly**.

<http://netpreserve.org/>

# New research fields and projects

À propos Événements Pratiques professionnelles Réflexion scientifique Expérimentation Publications Projet : les membres Crédits

## Réseau de partenaires pour l'analyse et l'exploration de données numériques

SUIVRE :

RESPADON RÉSEAU DE PARTENAIRES POUR L'ANALYSE ET L'EXPLORATION DE DONNÉES NUMÉRIQUES

PLUS

- À propos
- Événements
- Cycle d'ateliers
- Journées d'étude
- Pratiques professionnelles

SOUTENU PAR LE GIS COLLEX-PERSÉE ET PORTÉ PAR L'UNIVERSITÉ DE LILLE ET LA BIBLIOTHÈQUE NATIONALE DE FRANCE (BNF), EN PARTENARIAT AVEC SCIENCES PO ET LE CAMPUS CONDORCET, LE PROJET RESPADON SE FONDÉ SUR UNE ANALYSE DES USAGES DES ARCHIVES DU WEB ET DES CORPUS NUMÉRIQUES CONSTITUÉS PAR LA BNF POUR EXPÉRIMENTER DE NOUVELLES MODALITÉS D'ACCÈS ET D'EXPLOITATION DE CES COLLECTIONS NUMÉRIQUES.

<https://respadon.hypotheses.org/>

## RESAW

A Research infrastructure for the Study of Archived Web materials

home About RESAW Events Projects Participants Web Archives Resources Forum Contact

## Home

### RESAW

A Twitter list by @resaw\_en  
Members tweet about...

#### UK Web Archive

@UKWebArchive

Web Archiving Roundup: April, 2019

[webarchiving.rtp.wordpress.com/2019/04/29/web-...](http://webarchiving.rtp.wordpress.com/2019/04/29/web-...) via @WebArch\_RT

#### Web Archiving Roundup: April, 2019

The Library of Congress Digital Collections Development Coordinator graciously invited me to the Archives Unleashed a training webinar on Web Archiving Systems API

Stories and News

Peter Webster, Web Historian, UK

A. S. Byatt's faded church

Ian Milligan, Web Historian, Canada

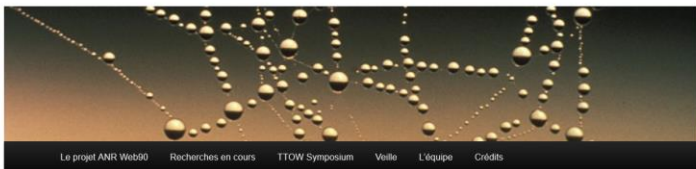
New Grant: "Continuing Education to Advance Web Archiving"

<http://resaw.eu/>

## Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990

"Technological progress has merely provided us with more efficient means for going backwards." — Aldous Huxley, Ends and Means

Recherche



Le projet ANR Web90 Recherches en cours TIOW Symposium Veille L'équipe Crédits

Publié le 20 août 2018 par Valérie Schafer

Recherche



jahendier

@jahendier



Tim Berners-Lee and I had a paper rejected because reviewer said we didn't really understand how the web worked.

#### PRÉSENTATION

Où sont l'histoire de l'internet et du Web aujourd'hui? Quelle place les archives du Web et plus globalement le patrimoine numérique (Web, groupes de discussion, RSS, etc.) occupent-ils dans

<https://web90.hypotheses.org/>

Welcome to the Archives Unleashed Project



The Archives Unleashed Project

Home

Welcome  
Contact Us

About the Project

Archives Unleashed Toolkit

Archives Unleashed Cloud

Archives Unleashed Notebooks

Warlight

Get Involved

Events

Publications

Welcome to the Archives Unleashed Project



Welcome

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by a grant from the Andrew W. Mellon Foundation, we are developing web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web.

<https://archivesunleashed.org/>

# Challenges and setbacks in Web archiving

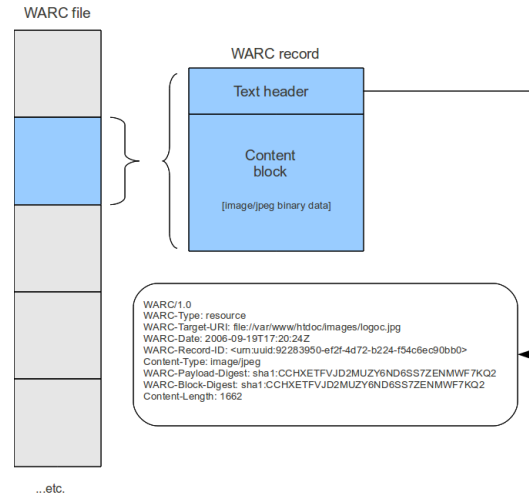
- Robots.txt
- Captcha (ie Completely Automated Public Turing-test to tell Computers and Humans Apart)
- User interaction needed
- Password protected content
- Technologies and dynamic content : Flash, java scripts,...
- Distant content
- Temporal inconsistencies
- Bot traps
- ...

## Method and tools for Web archiving

- Define a strategy
- Use a log and write throughout the life of the project
- You may need to use additional methods and tools
  
- Screen capture and screen recording
- Link crawling
- On demand archiving
- ...

# A standard format for Web archives

- WARC file format = Web ARChive archive format
- ARC was accepted as an international standard in 2009 (ISO 28500:2009)
- WARC is now recognised by most national library systems as the standard to follow for web archival



[https://en.wikipedia.org/wiki/Web\\_ARChive](https://en.wikipedia.org/wiki/Web_ARChive)

[https://wiki.archivematica.org/Significant characteristics of websites](https://wiki.archivematica.org/Significant_characteristics_of_websites)

<https://wiki.archivematica.org/File:WARCdiagram.png>

# Internet Archive's Wayback Machine

The screenshot shows the homepage of the Internet Archive's Wayback Machine. At the top, a dark grey banner contains the text "Search the history of over 351 billion web pages on the Internet." Below this is a search bar with the "WaybackMachine" logo and a placeholder "enter URL or keywords". A navigation bar below the search bar includes icons for various media types (books, movies, software, music, images) and a "SIGN IN" button. Below the navigation bar is a menu with links: ABOUT, CONTACT, BLOG, PROJECTS, HELP, DONATE, JOBS, VOLUNTEER, PEOPLE. The main content area features a large icon of a classical building on the left. To its right, a text block states: "Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more." Below this text is a row of icons representing different media types with their respective counts: 351B (books), 20M (movies), 5.9M (software), 5.2M (music), 1.8M (websites), 417K (images), 3.3M (audio), 292K (video), and 419K (text). A search bar with a "GO" button and a link to "Advanced Search" is positioned below the icons. On the right side of the main content area, there is an "Announcements" section with several links to recent news items, including "Official EU Agencies Falsely Report More Than 550 Archive.org URLs as Terrorist Content", "Boston Public Library's 78rpm Records Come to the Internet: Reformatting the Boston Public Library Sound Archives", and "Google Plus (or Minus) and the Ephemerality of Community". A "SEE MORE" link is located at the bottom of the announcements section. At the very bottom of the page, a small link for "Terms of Service Dec 31, 2014" is visible.

- Launched in 2001
- 357 billion archived Web pages so far
- Archived content going back to 1996

<https://archive.org/>

Structure of URLs in the Internet Archive's  
Wayback Machine:

Wayback Machine URL/collection/time shown as  
yyymmddhhmmss/URL

Ex:

<https://web.archive.org/web/20220707064851/>

<https://distam.hypotheses.org/2656>

# Internet Archive / Wayback Machine extension for browsers



Wayback Machine

Disponible sur Chrome

Sélection

★★★★★ 209 | Outils de recherche | 100 000+ utilisateurs



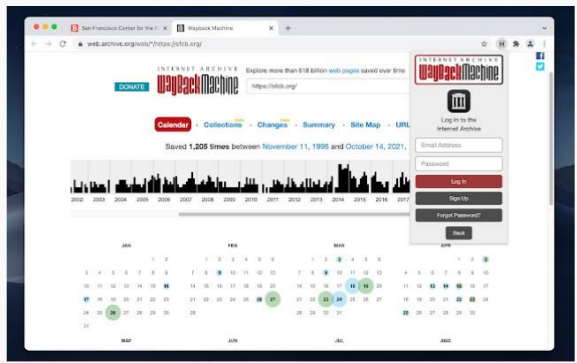
## Wayback Machine by Internet Archive

⚠ This add-on is not actively monitored for security by Mozilla. Make sure you trust it before installing. [Learn more](#)

Welcome to the Official Internet Archive Wayback Machine Browser Extension! Go back in time to see how a website has changed through the history of the Web. Save websites, view missing 404 Not Found pages, or read archived books & papers. [Remove](#)



- Présentation
- Pratiques en matière de confidentialité
- Avis
- Assistance
- Articles similaires

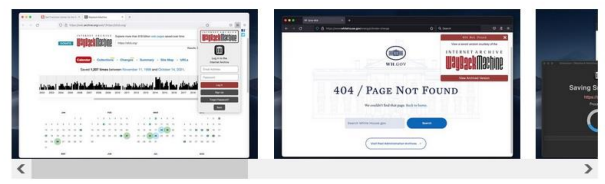


### Rate your experience

How are you enjoying Wayback Machine?  
[Log in to rate this extension](#)  
[Report this add-on for abuse](#)

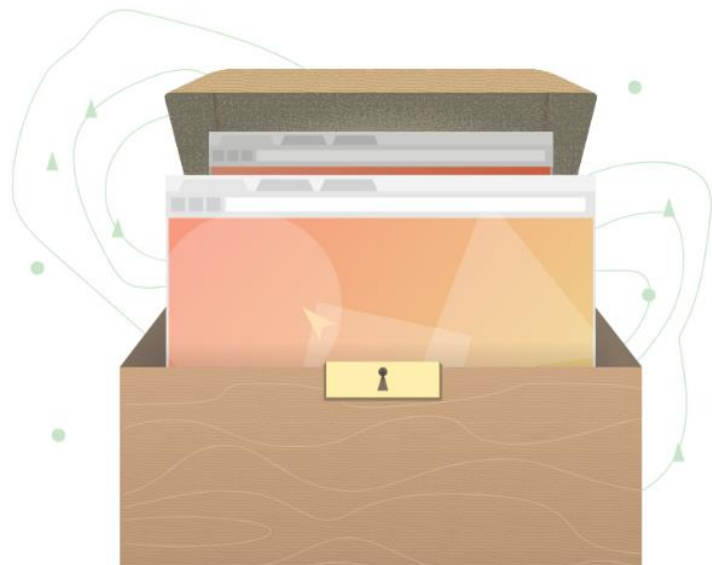
[Read all 236 reviews](#)

### Screenshots



<https://chrome.google.com/webstore/detail/wayback-machine/fpnmgdkabkmnadcjpehmlllkndpkmiak?hl=fr>

[https://addons.mozilla.org/en-US/firefox/addon/wayback-machine\\_new/](https://addons.mozilla.org/en-US/firefox/addon/wayback-machine_new/)



# Conifer

Collect and revisit web pages.

Conifer is a web archiving service that creates an interactive copy of any web page that you browse, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth.

[Create a Free Account](#)

[Existing Users Login](#)

Free accounts with 5GB of storage. Get more and support this project by [becoming a supporter](#).

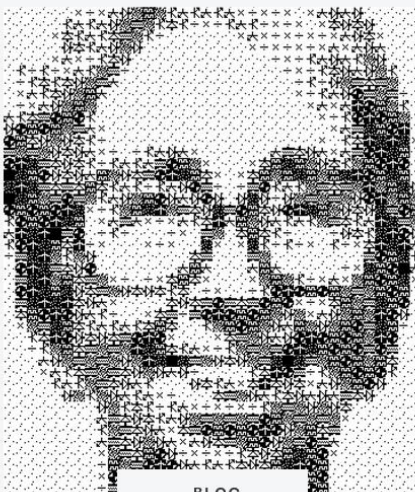


. BLOG

**You're Invited:  
Rhizome's 2022  
Benefit Honoring  
Rachel Rossin &  
Julie Martin**

---

RHIZOME



. BLOG

**Studies in  
Perception IV: Julie  
Martin**

---

MICHAEL CONNOR

. BLOG

**Off: Endgame**

---

RHIZOME



# Webrecorder

*Web archiving for all!*

[Blog](#) [Tools](#) [Community](#) [About](#) [Contact](#) [FAQ](#) [Jobs](#)

---

Webrecorder provides a **suite of open source tools and packages** to **capture** interactive websites and **replay** them at a later time as **accurately as possible**. Learn more about our key tools and efforts:

**ArchiveWeb.page**



**ReplayWeb.page**



**pywb**



**Browsertrix  
Crawler**



**Browsertrix Cloud**



<https://webrecorder.net/>



Interactive browser-based web archiving, a [Webrecorder](#) project.

ArchiveWeb.page is the latest tool from Webrecorder to turn your browser into a full-featured interactive web archiving system!

ArchiveWeb.page is available as an extension for any Chrome or Chromium based browsers. (A standalone app version is also in development.)

To create web archives, the extension (or app) will be needed. Once created, the archives can be viewed in any modern browser using [ReplayWeb.page](#) -- no extension required!

 [View the ArchiveWeb.page User Guide](#)

✘ Sorry, the ArchiveWeb.page browser Extension requires Chrome or a Chromium based browser. Please try ArchiveWeb.page in Chrome or check back soon for additional options.

<https://archiveweb.page/>

# Bibliography

zotero

Groups Documentation Forums Get Involved Log In

Q Title, Creator, Year

Upgrade Storage

Other Libraries

- peyssard's Library
  - 404
  - A lire
  - About Zotero
  - Acquisitions\_suggestions
  - Aldébaran
  - AO
  - archives de la recherche
  - assesment
  - ated-tunis-09-2016
  - bad science
  - BAH
- archive du Web
- communication
- Internet
- Internet Archive
- patrimoine numérique
- science de l'information
- technologie
- temporalité
- temps
- web

Filter Tags

Title	Creator	Date
An Overview of Web Archiving	Niu	2012
Archiver les traces numériques en Méditer...	Gebeil	
Archiving the World Wide Web	Lyman	2002
Archiving websites: general considerations...	Brügger	2005
ASAP – Archives sauvegarde attentats Pari...		
Conservative party deletes archive of spee...		
Digital Humanities Institute – Beirut		2019-04-28
In Jordan, the "Invisible Hand" Blocks Inter...		
La sélection de sites web dans une bibliot...	Bonnel and Oury	2014-07-30
Legal deposit of the French Web: harvesti...	Lasfargues et al.	2008-09-18
LibGuides: Preserving Web Sites using Win...	Benalayt	
List of Web archiving initiatives		2019-04-12
Locking the Web Open, a Call for a Distrib...	Kahle	2015-11-02
Methods of collecting facebook material a...	Brügger and Sandvik	2013
More than 9 million broken links on Wikip...		
Official EU Agencies Falsely Report More T...		
Peut-être n'est-il pas nécessaire d'archiver...		
Qu'est ce qu'une archive du web ?	Muriani et al.	2019-01-28

Info Notes Tags Attachments Related

Item Type Book

Title The Archived Web

Author Brügger, Niels

Publisher MIT Press

Date September 2018

# of Pages 200

Language en

ISBN 978-0-262-03902-4

URL <http://www.worldcat.org/oclc/105...>

Accessed 14/05/2019 à 16:39:44

Date Added 14/05/2019 à 16:39:44

Date Modified 14/05/2019 à 16:44:24

Abstract

An original methodological framework for approaching the archived web, both as a source and as an object of study in its own right.

As life continues to move online, the web becomes increasingly important as a source for understanding the past. But historians have yet to formulate a methodology for approaching the archived

<https://www.zotero.org/peyssard/collections/CNGX47Z3>

Thank you / Merci / شكراً

## Ecole d'été DISTAM 2022

[jean-christophe.peyssard@univ-amu.fr](mailto:jean-christophe.peyssard@univ-amu.fr)



Maison méditerranéenne  
des sciences de l'homme  
UAR3125

