Micro Archives as Rich Digital Object Representations

Helge Holzmann, Mila Runnwerth
WebSci ‘18
Referencing digital objects in scientific publications: Stable & dynamic data

| stable data                      | dynamic data                          |
|---------------------------------|---------------------------------------|
| journal articles                | software                              |
| set of measuring data           | science blogs / forums                |
| ![DOI](doi.png)                 | ![Question Mark](question_mark.png)   |
Referencing digital objects in scientific publications: The Web as infrastructure

Blog and software pages show the latest status of development or participation.

The Internet Archive unsystematically archives snapshots so web pages can be retraced over time.

Question: How complete and coherent is the archived web with respect to related resources linked on the corresponding web pages?
Case Studies: Data sets

Blogs: TREC Blogs ‘08

- 28,488,766 permalinked blog posts
- from 1,011,733 homepages

Software: swMATH

- 21,947 software packages of mathematical software with corresponding URLs (May 29th)
Case Studies

resources
an object as fraction of its archived resources

authority
authority pages only

related
fraction of resources only if authority is archived

complete
Micro Archives Use Case Scenario

Specifying Micro Archives

Defining a digital representation of an object with certain specification

Crawling / Archiving

According to the specification crawling & archiving is executed at the same time in the specified depth, including the specified resources.

Presentation / Citing

The newly created *micro archive* is a semantic collection anchored in time that may be assigned a URL or, more specifically, a DOI for citing.
Proof-of-concept prototype Micrawler

1. Spec Proxy
2. Crawl Queue
3. Archiving / Crawl Service
4. Archive Meta Service
5. Analysers
6. Persistence Provider
7. Viewer
Proof-of-concept prototype Micrawler

Demo

https://github.com/helgeho/Micrawler
What is a micro archive? A micro archive is a snapshot of a fixed (evolving) set of URLs that are representative for some object or entity (at a given time). Hence, such an archive can be used to describe and/or derive information about its subject at the time of the crawl.

Create your own micro archive for an entity or object of your choice by either defining a set of URLs manually or loading/extracting a crawl specification from some URL:

Enter spec definition

Enter URL to load/extract a crawl specification

Load/extract spec

© 2017 L3S Research Center (Helge Holzmann). All rights reserved.
What is a micro archive? A micro archive is a snapshot of a fixed (evolving) set of URLs that are representative for some object or entity (at a given time). Hence, such an archive can be used to describe and/or derive information about its subject at the time of the crawl.

Create your own micro archive for an entity or object of your choice by either defining a set of URLs manually or loading/extracting a crawl specification from some URL:

- Enter spec definition
- Enter URL to load / extract a crawl specification
- Load / extract spec
@online{unnamed,
  url = [http://tempas.l3s.de/micrawler/permalink/785d426],
  urldate = {2018-05-29T13:50:53z},
  note = {Archived using Micrawler}
}
Outlook & opportunities

- Supporting web archives
- Temporally relevant collections
- Structuring the Web
- Rich information
