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 set of measuring data	software science blogs / forums
doi	?

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)

swMATH
an information service for mathematical software

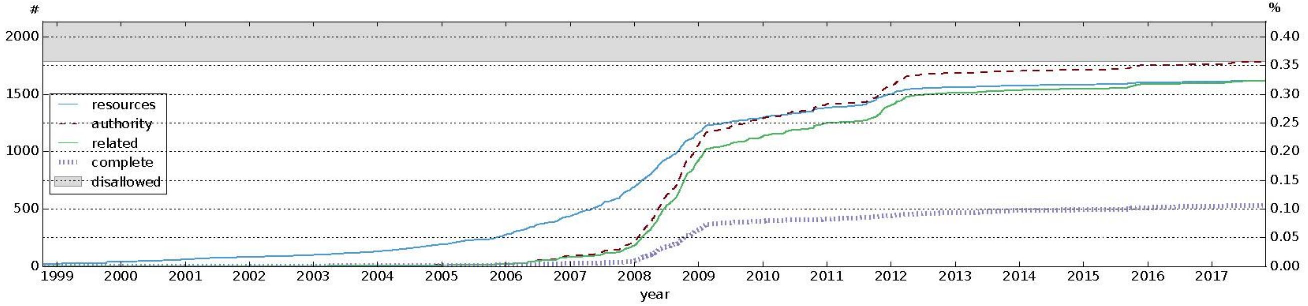
Case Studies

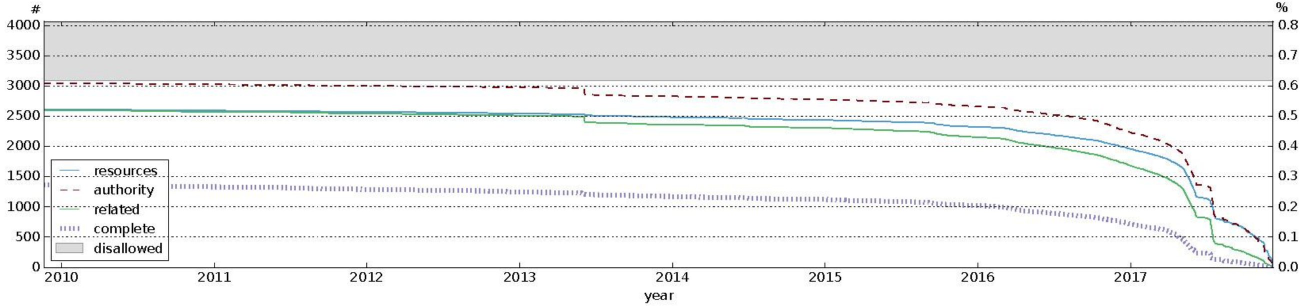
resources
an object as fraction of its archived resources

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

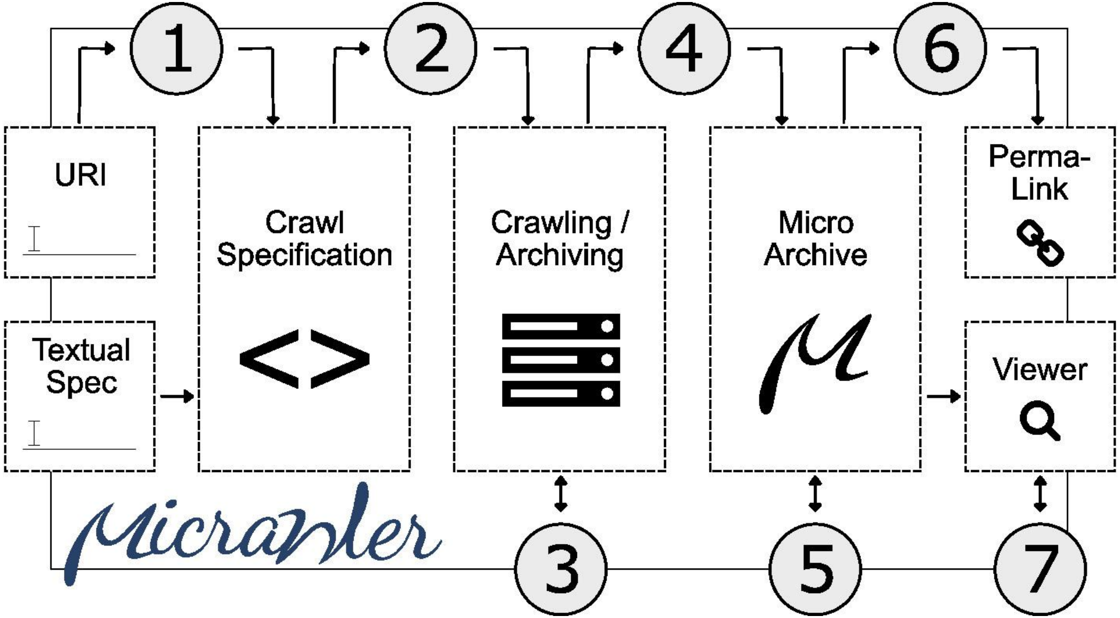
Crawling / Archiving

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

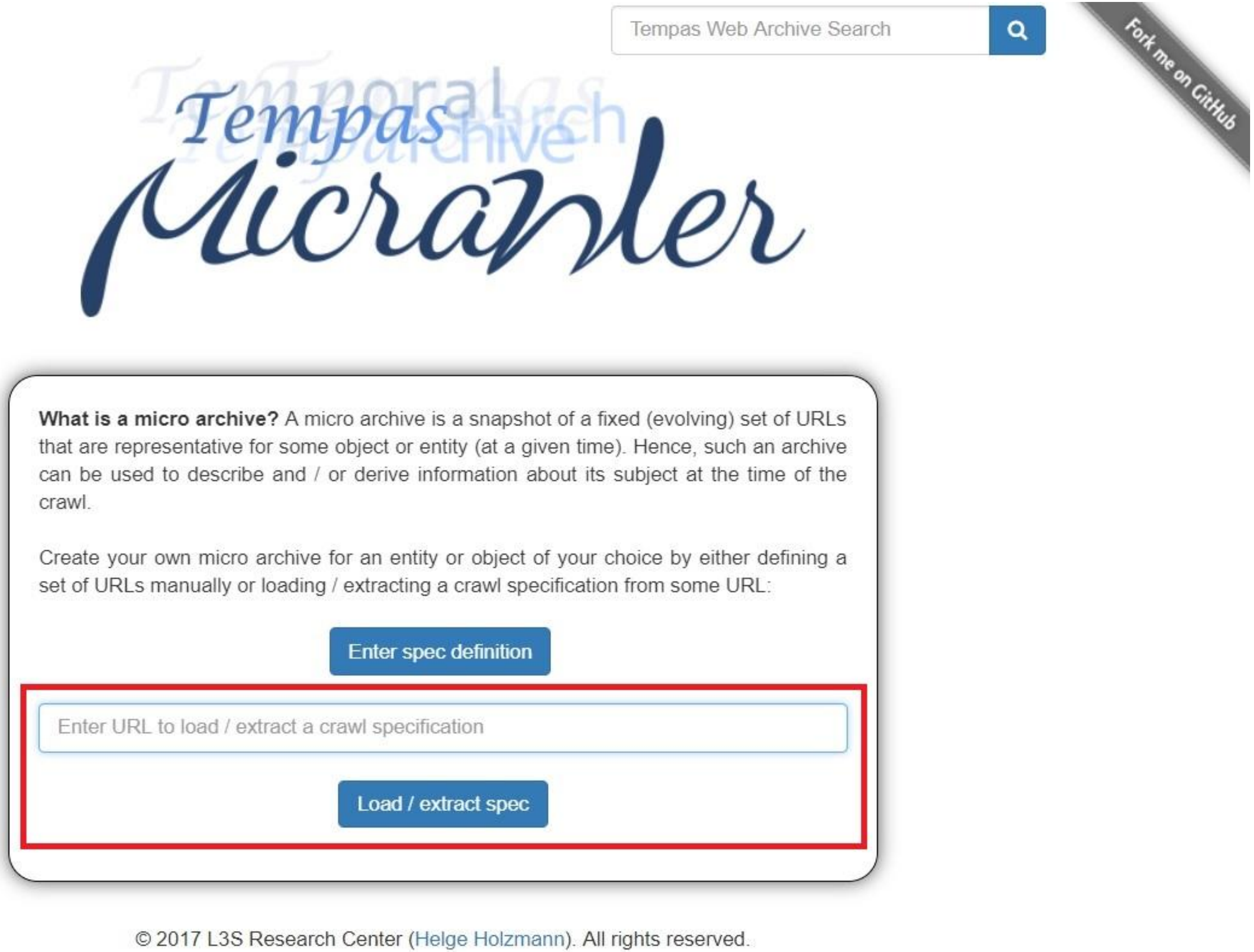


1. Spec Proxy
2. Crawl Queue
3. Archiving / Crawl Service
4. Archive Meta Service
5. Analysers
6. Persistence Provider
7. Viewer

Proof-of-concept prototype Micrawler

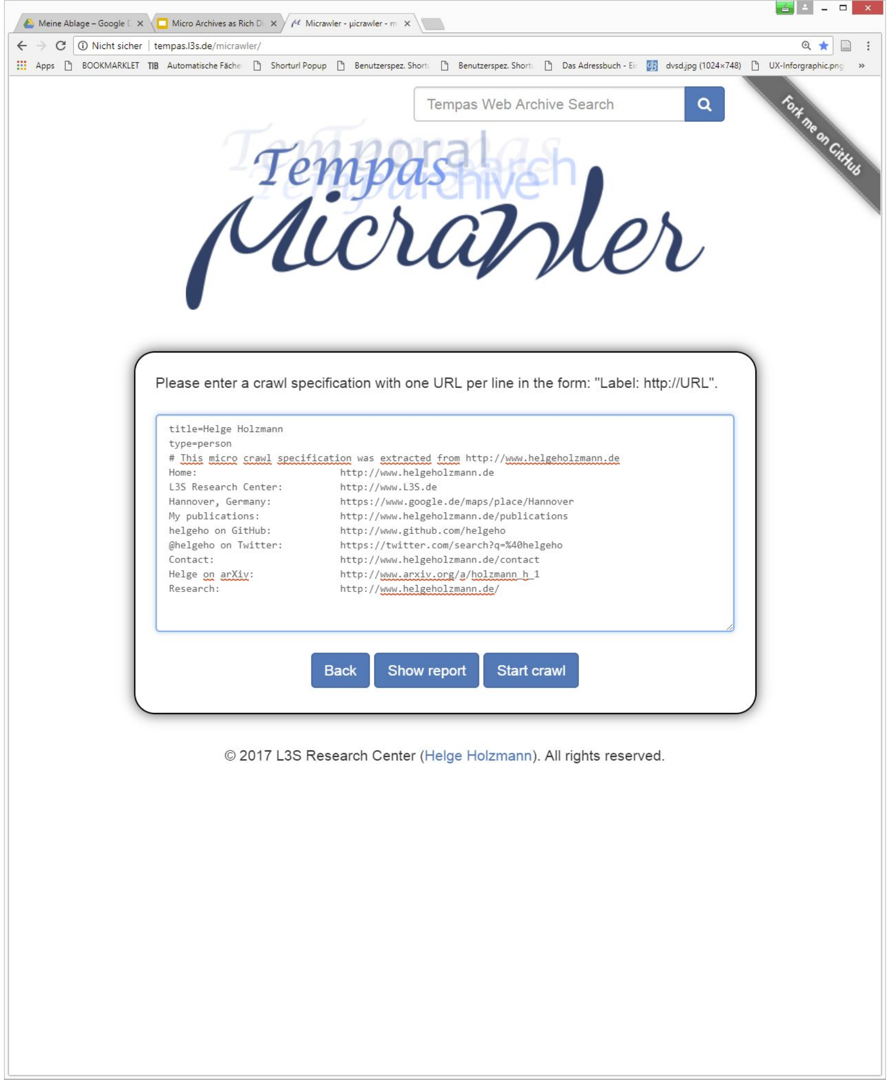
Demo

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.

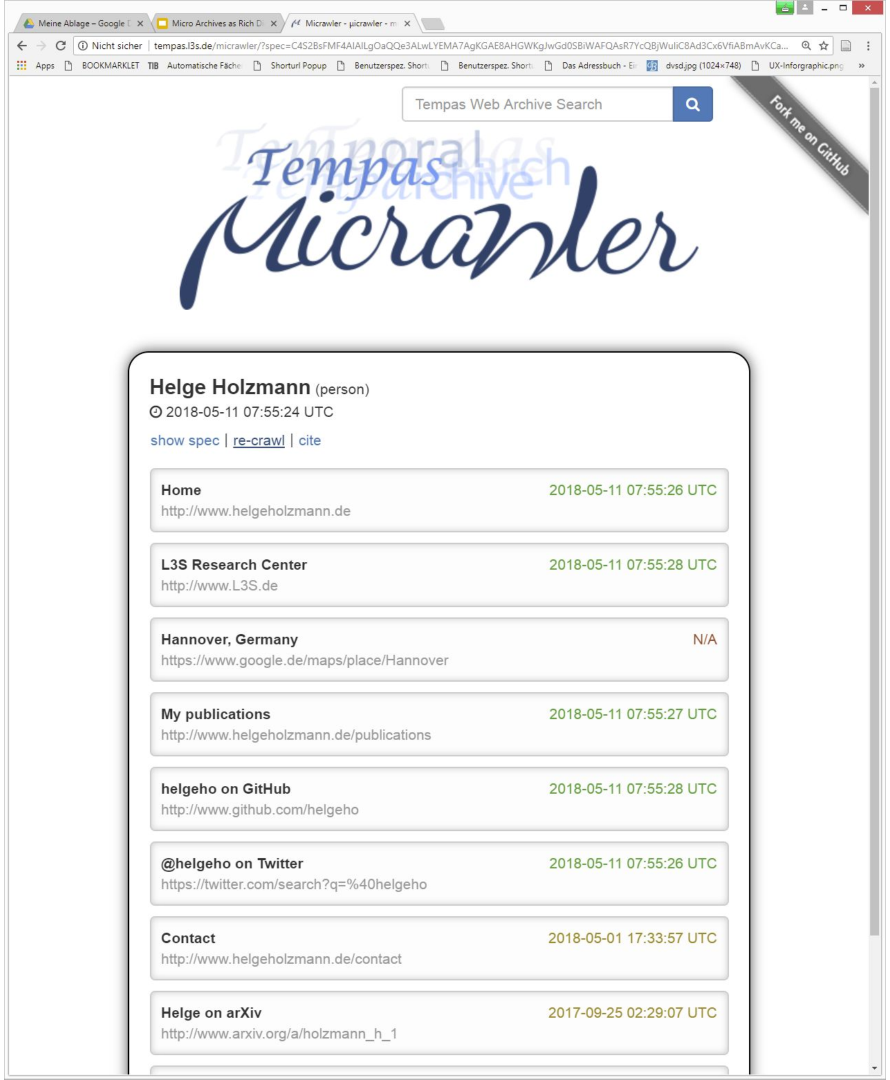
Create your own micro archive for an entity or object of your choice by either defining a set of URLs manually or loading / extracting a crawl specification from some URL:



Please enter a crawl specification with one URL per line in the form: "Label: http://URL".

```
title=Helge Holzmann
type=person
# This micro crawl specification was extracted from http://www.helgeholzmann.de
Home:                         http://www.helgeholzmann.de
L3S Research Center:          http://www.L3S.de
Hannover, Germany:            https://www.google.de/maps/place/Hannover
My publications:              http://www.helgeholzmann.de/publications
helgeho on GitHub:            http://www.github.com/helgeho
@helgeho on Twitter:          https://twitter.com/search?q=%40helgeho
Contact:                      http://www.helgeholzmann.de/contact
Helge on arXiv:               http://www.arxiv.org/a/holzmann_h_1
Research:                     http://www.helgeholzmann.de/
```

© 2017 L3S Research Center (Helge Holzmann). All rights reserved.



Helge Holzmann (person)

2018-05-11 07:55:24 UTC

show spec | re-crawl | cite

Home 2018-05-11 07:55:26 UTC
http://www.helgeholzmann.de

L3S Research Center 2018-05-11 07:55:28 UTC
http://www.L3S.de

Hannover, Germany N/A
https://www.google.de/maps/place/Hannover

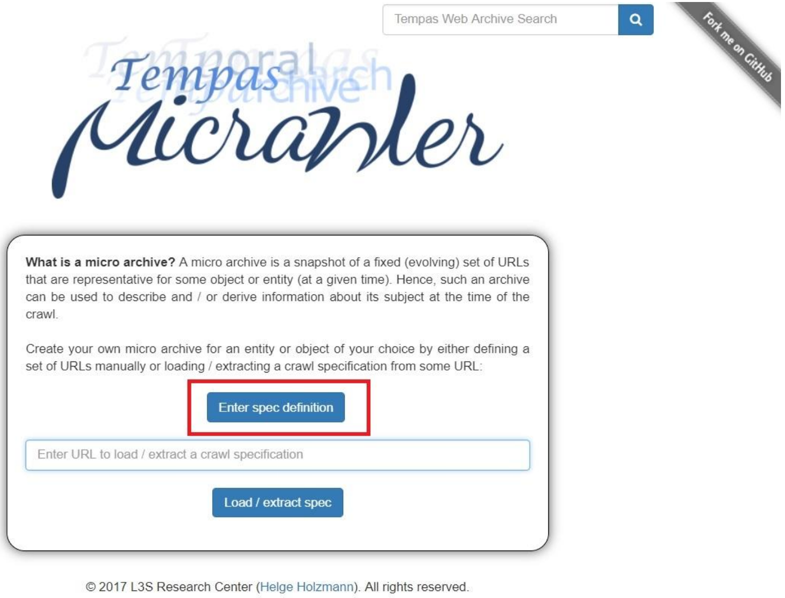
My publications 2018-05-11 07:55:27 UTC
http://www.helgeholzmann.de/publications

helgeho on GitHub 2018-05-11 07:55:28 UTC
http://www.github.com/helgeho

@helgeho on Twitter 2018-05-11 07:55:26 UTC
https://twitter.com/search?q=%40helgeho

Contact 2018-05-01 17:33:57 UTC
http://www.helgeholzmann.de/contact

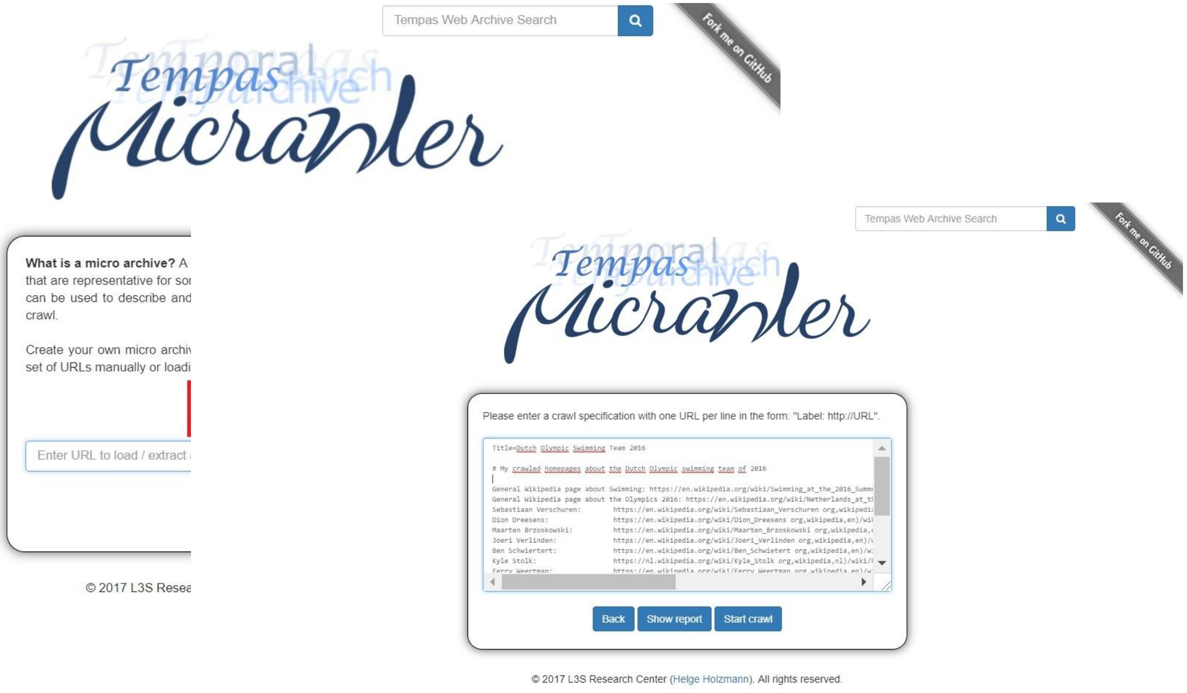
Helge on arXiv 2017-09-25 02:29:07 UTC
http://www.arxiv.org/a/holzmann_h_1


Tempas Web Archive Search
Fork me on GitHub
Tempas
Micrawler
Enter spec definition
Enter URL to load / extract a crawl specification
Load / extract spec

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:

© 2017 L3S Research Center (Helge Holzmann). All rights reserved.



What is a micro archive? A
that are representative for so
can be used to describe and
crawl.

Create your own micro archi
set of URLs manually or loadi

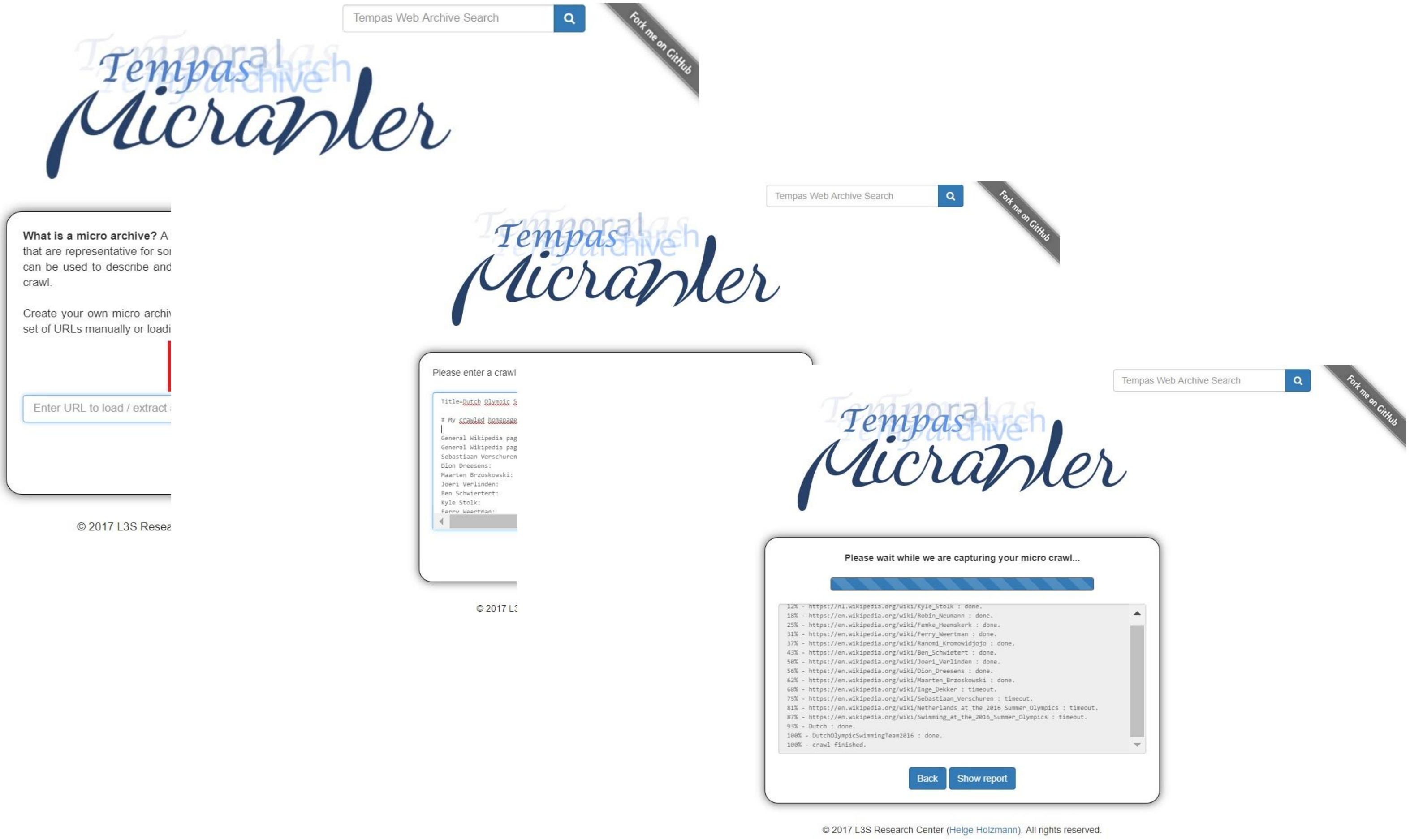
Enter URL to load / extract

© 2017 L3S Resea

Please enter a crawl specification with one URL per line in the form: "Label: http://URL".

```
Title=Dutch Olympic Swimming Team 2016

# My crawled homepages about the Dutch Olympic swimming team of 2016

General Wikipedia page about Swimming: https://en.wikipedia.org/wiki/Swimming_at_the_2016_Summ
General Wikipedia page about the Olympics 2016: https://en.wikipedia.org/wiki/Netherlands_at_tl
Sebastiaan Verschuren:        https://en.wikipedia.org/wiki/Sebastiaan_Verschuren org,wikipedi
Dion Dreesens:                https://en.wikipedia.org/wiki/Dion_Dreesens org,wikipedia,en)/wil
Maarten Brzoskowski:          https://en.wikipedia.org/wiki/Maarten_Brzoskowski org,wikipedia,
Joeri Verlinden:              https://en.wikipedia.org/wiki/Joeri_Verlinden org,wikipedia,en)/
Ben Schwiertert:              https://en.wikipedia.org/wiki/Ben_Schwietert org,wikipedia,en)/w
Kyle Stolk:                   https://nl.wikipedia.org/wiki/Kyle_Stolk org,wikipedia,nl)/wiki/
Ferry Weertman:               https://en.wikipedia.org/wiki/Ferry_Weertman org,wikipedia,en)/w
```

© 2017 L3S Research Center (Helge Holzmann). All rights reserved.



Tempas Web Archive Search
Fork me on GitHub
Tempas Micrawler

What is a micro archive? A
that are representative for so
can be used to describe and
crawl.

Create your own micro archi
set of URLs manually or loadi

Enter URL to load / extract

© 2017 L3S Resea

Tempas Web Archive Search
Fork me on GitHub
Tempas Micrawler

Please enter a crawl

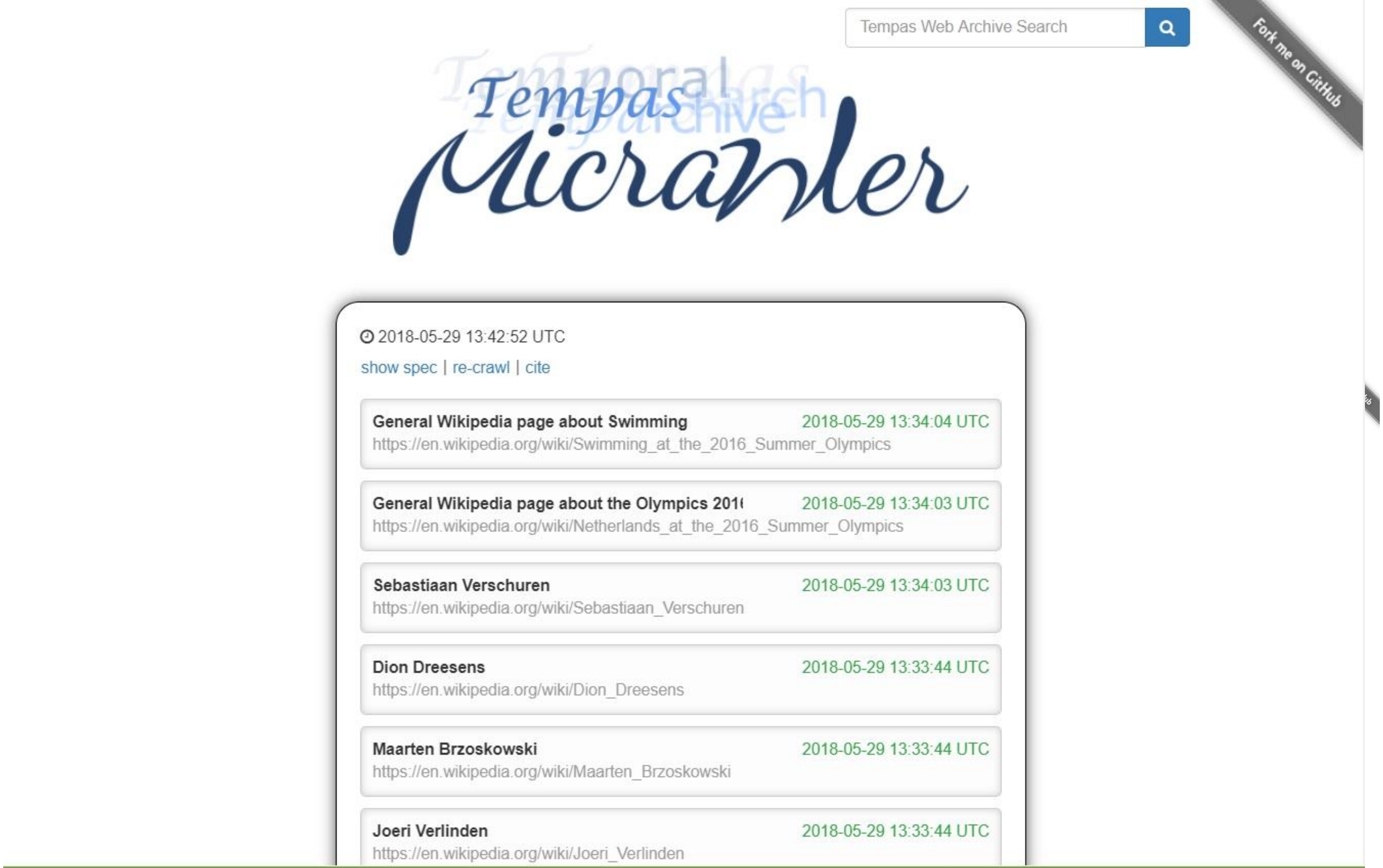
```
Title=Dutch Olympic S

# My crawled homepage
|
General Wikipedia pag
General Wikipedia pag
Sebastiaan Verschuren
Dion Dreesens:
Maarten Brzoskowski:
Joeri Verlinden:
Ben Schwiertert:
Kyle Stolk:
Ferry Weertman:
```

© 2017 L3

Please wait while we are capturing your micro crawl...

```
12% - https://nl.wikipedia.org/wiki/Kyle_Stolk : done.
18% - https://en.wikipedia.org/wiki/Robin_Neumann : done.
25% - https://en.wikipedia.org/wiki/Femke_Heemskerk : done.
31% - https://en.wikipedia.org/wiki/Ferry_Weertman : done.
37% - https://en.wikipedia.org/wiki/Ranomi_Kromowidjojo : done.
43% - https://en.wikipedia.org/wiki/Ben_Schwietert : done.
50% - https://en.wikipedia.org/wiki/Joeri_Verlinden : done.
56% - https://en.wikipedia.org/wiki/Dion_Dreesens : done.
62% - https://en.wikipedia.org/wiki/Maarten_Brzoskowski : done.
68% - https://en.wikipedia.org/wiki/Inge_Dekker : timeout.
75% - https://en.wikipedia.org/wiki/Sebastiaan_Verschuren : timeout.
81% - https://en.wikipedia.org/wiki/Netherlands_at_the_2016_Summer_Olympics : timeout.
87% - https://en.wikipedia.org/wiki/Swimming_at_the_2016_Summer_Olympics : timeout.
93% - Dutch : done.
100% - DutchOlympicSwimmingTeam2016 : done.
100% - crawl finished.
```

© 2017 L3S Research Center (Helge Holzmann). All rights reserved.



2018-05-29 13:42:52 UTC

show spec | re-crawl | cite

General Wikipedia page about Swimming 2018-05-29 13:34:04 UTC
https://en.wikipedia.org/wiki/Swimming_at_the_2016_Summer_Olympics

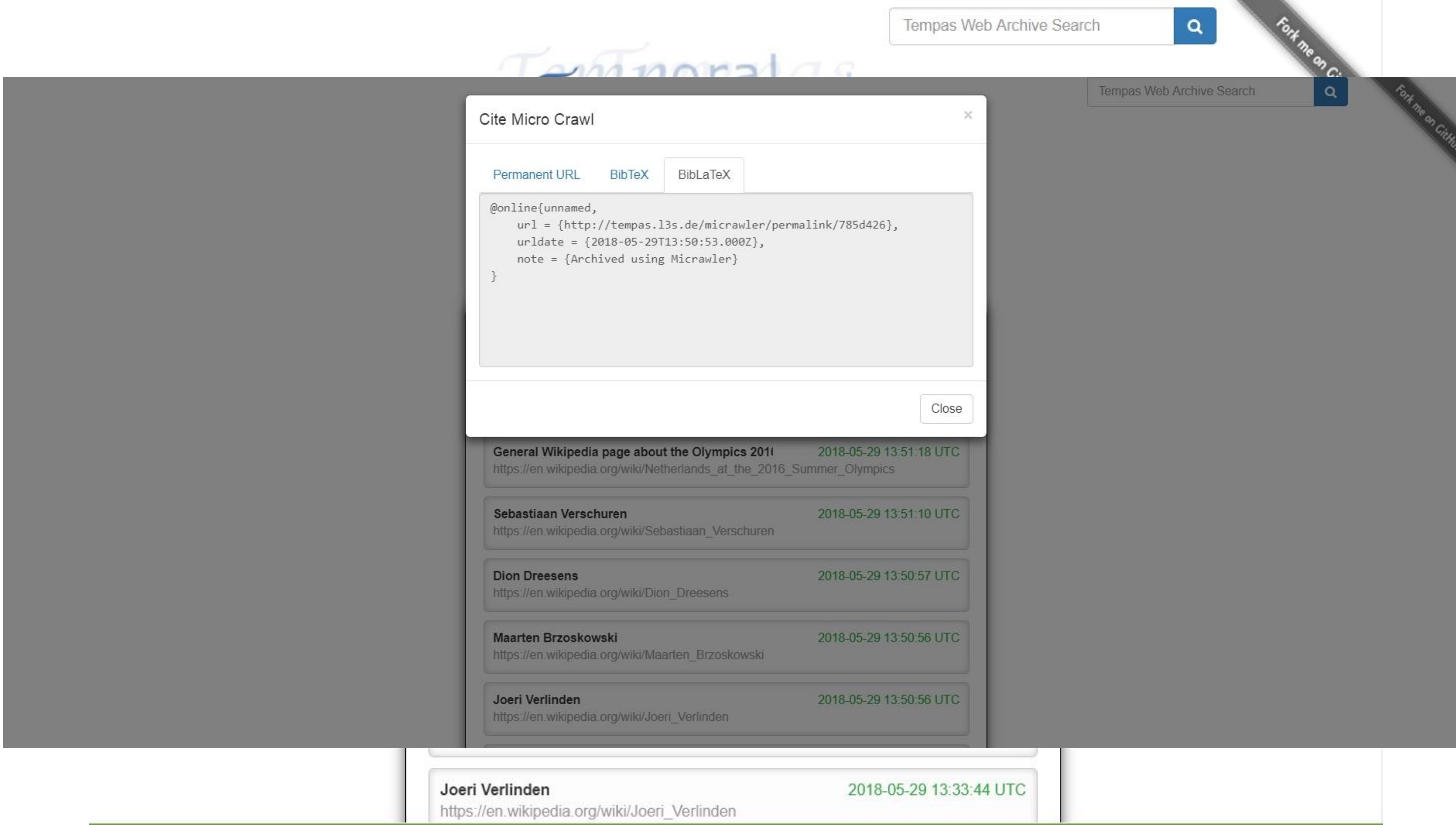
General Wikipedia page about the Olympics 201 2018-05-29 13:34:03 UTC
https://en.wikipedia.org/wiki/Netherlands_at_the_2016_Summer_Olympics

Sebastiaan Verschuren 2018-05-29 13:34:03 UTC
https://en.wikipedia.org/wiki/Sebastiaan_Verschuren

Dion Dreesens 2018-05-29 13:33:44 UTC
https://en.wikipedia.org/wiki/Dion_Dreesens

Maarten Brzoskowski 2018-05-29 13:33:44 UTC
https://en.wikipedia.org/wiki/Maarten_Brzoskowski

Joeri Verlinden 2018-05-29 13:33:44 UTC
https://en.wikipedia.org/wiki/Joeri_Verlinden



```
@online{unnamed,
    url = {http://tempas.l3s.de/micrawler/permalink/785d426},
    urldate = {2018-05-29T13:50:53.000Z},
    note = {Archived using Micrawler}
}
```

General Wikipedia page about the Olympics 201(2018-05-29 13:51:18 UTC
https://en.wikipedia.org/wiki/Netherlands_at_the_2016_Summer_Olympics

Sebastiaan Verschuren 2018-05-29 13:51:10 UTC
https://en.wikipedia.org/wiki/Sebastiaan_Verschuren

Dion Dreesens 2018-05-29 13:50:57 UTC
https://en.wikipedia.org/wiki/Dion_Dreesens

Maarten Brzoskowski 2018-05-29 13:50:56 UTC
https://en.wikipedia.org/wiki/Maarten_Brzoskowski

Joeri Verlinden 2018-05-29 13:50:56 UTC
https://en.wikipedia.org/wiki/Joeri_Verlinden

Joeri Verlinden 2018-05-29 13:33:44 UTC
https://en.wikipedia.org/wiki/Joeri_Verlinden

Outlook & opportunities

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

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE TREES TALL
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

holzmann@l3s.de

mila.runnwerth@tib.eu