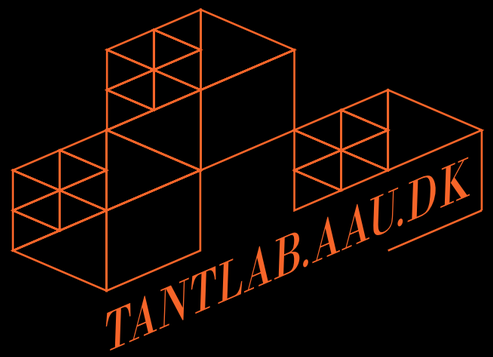


What is a web entity?

Hyphe’s specific way to group pages

Mathieu Jacomy
Aalborg University TANTLab

Web entities are based on the URL shape

https://en.wikipedia.org/wiki/Snail

https://fr.wikipedia.org/wiki/Escargot

https://en.wikipedia.org/wiki/Slug

https://en.wikipedia.org/wiki/Slug#Behavior

https://en.wikipedia.org/wiki/Lettuce

https://en.m.wikipedia.org/w/index.php?title=Lettuce

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php ?title=Lettuce

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php ?title=Lettuce

TLD

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php ?title=Lettuce

Domain

TLD

Web entities are based on the URL shape

https://		en.	wikipedia	.org	/wiki	/Snail	
https://		fr.	wikipedia	.org	/wiki	/Escargot	
https://		en.	wikipedia	.org	/wiki	/Slug	
https://		en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://		en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en. m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Path

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail		
https://	fr.	wikipedia	.org	/wiki	/Escargot		
https://	en.	wikipedia	.org	/wiki	/Slug		
https://	en.	wikipedia	.org	/wiki	/Slug		#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce		
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Path

Query

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en. m.	wikipedia	.org	/w	/index.php	?title=Lettuce

Subdomains

Domain

TLD

Path

Query

Fragment

Web entities are based on the URL shape

https://	en.	wikipedia	.org	/wiki	/Snail	
https://	fr.	wikipedia	.org	/wiki	/Escargot	
https://	en.	wikipedia	.org	/wiki	/Slug	
https://	en.	wikipedia	.org	/wiki	/Slug	#Behavior
https://	en.	wikipedia	.org	/wiki	/Lettuce	
https://	en.	m.	wikipedia	.org	/w	/index.php	?title=Lettuce	

Protocol

Subdomains

Domain

TLD

Path

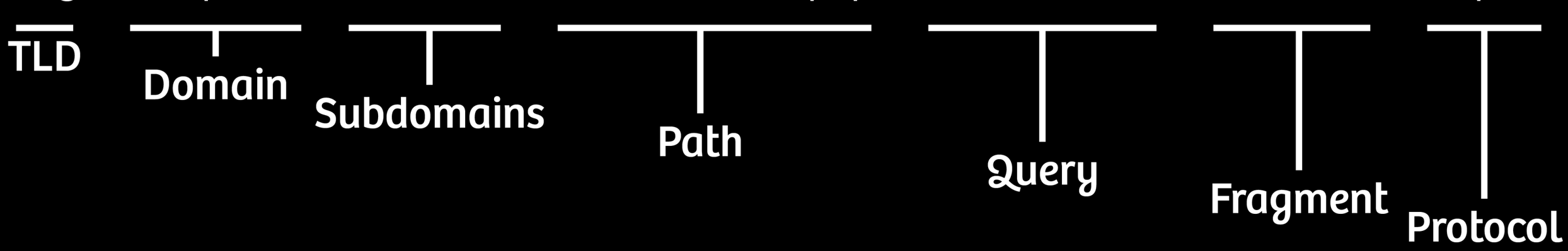
Query

Fragment

Web entities are based on the URL shape

.org	wikipedia	en.	/wiki	/Snail		https://
.org	wikipedia	fr.	/wiki	/Escargot		https://
.org	wikipedia	en.	/wiki	/Slug		https://
.org	wikipedia	en.	/wiki	/Slug	#Behavior	https://

.org	wikipedia	en.	/wiki	/Lettuce		https://
.org	wikipedia	m. en.	/w	/index.php	?title=Lettuce	https://



Web entity examples

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **WIKIPEDIA**

Rule: must be prefixed by "**.org** | **wikipedia**"

Web entity examples

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **ENGLISH WIKIPEDIA**

Rule: must be prefixed by "**.org** | **wikipedia** | **en.**"

Web entity examples

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **ENGLISH WIKIPEDIA** (IMPROVED)

Rule: must be prefixed by "**.org** | **wikipedia** | **en.**" OR "**.org** | **wikipedia** | **m.** | **en.**"

Web entity examples

.org	wikipedia	en.		/wiki	/Snail			https://
.org	wikipedia	fr.		/wiki	/Escargot			https://
.org	wikipedia	en.		/wiki	/Slug			https://
.org	wikipedia	en.		/wiki	/Slug		#Behavior	https://
.org	wikipedia	en.		/wiki	/Lettuce			https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce		https://

LOOKING FOR **THE SLUG PAGE**

Rule: must be prefixed by "**.org** | **wikipedia** | **en.** | **/wiki** | **/Slug**"

Web entity examples

.org	wikipedia	en.	/wiki	/Snail		https://
.org	wikipedia	fr.	/wiki	/Escargot		https://
.org	wikipedia	en.	/wiki	/Slug		https://
.org	wikipedia	en.	/wiki	/Slug	#Behavior	https://
.org	wikipedia	en.	/wiki	/Lettuce		https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce	https://

ALL THE RULES CAN **COEXIST**

Example: **WIKIPEDIA** + **SLUG PAGE**

The rules apply in order: every page is in **ONE and ONLY ONE** web entity

Web entity examples

.org	wikipedia	en.	/wiki	/Snail		https://
.org	wikipedia	fr.	/wiki	/Escargot		https://
.org	wikipedia	en.	/wiki	/Slug		https://
.org	wikipedia	en.	/wiki	/Slug	#Behavior	https://
.org	wikipedia	en.	/wiki	/Lettuce		https://
.org	wikipedia	m.	en.	/w	/index.php	?title=Lettuce	https://

YOU CAN DEFINE **EACH WIKI PAGE** AS A DIFFERENT ENTITY

Hyphe can learn **automatic rules** that apply to future crawls

Thank you for your attention

@jacomyma
reticular.hypotheses.org
Mathieu.Jacomy@gmail.com