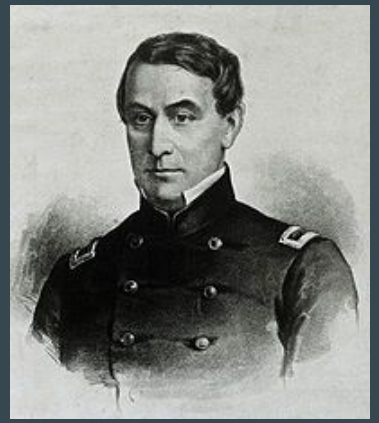
~~Civil War~~ Star Wars TriviaBot



Team Rogue Squadron

Roadmap

- Motivation and Goal
- Research
- Initial Attempts
- TriviaBot Architecture
- Data Gathering Methods
- Front End Development
- Demo

Motivation and Goal

Motivation: Can we gather unstructured information on the web and turn it into an ontology?

Goal: Scrape Wikipedia in order to create an ontology that can be queried with natural language questions.

Background

- MediaWiki - wiki “software”
- Wikia: "the rest of the library and magazine rack" to Wikipedia's encyclopaedia. - Gil Penchina, Former Wikia CEO
  - Wiki hosting service
- Wookiepedia: the Star Wars Wiki.

Initial Attempt

Battle	Date	State	CWSAC	Outcome
Battle of Fort Sumter	April 12–14, 1861	South Carolina	A	Confederate victory: Beauregard takes Charleston Federal fort, first battle of American Civil War.
Battle of Sewell's Point	May 18–19, 1861	Virginia	D	Inconclusive: Union gunboats fight inconclusive battle with Confederate artillery.
Battle of Aquia Creek	May 29 – June 1, 1861	Virginia	D	Inconclusive: Confederate artillery hit by naval bombardment, later withdrawn.

Battle of Wilson's Creek

Part of the American Civil War

Battle of Wilson's Creek by Kurz and Allison.

Date	August 10, 1861
Location	Greene County and Christian County, Missouri
Result	Confederate States victory

Belligerents

United States	Missouri (Confederate) Confederate States

Commanders and leaders

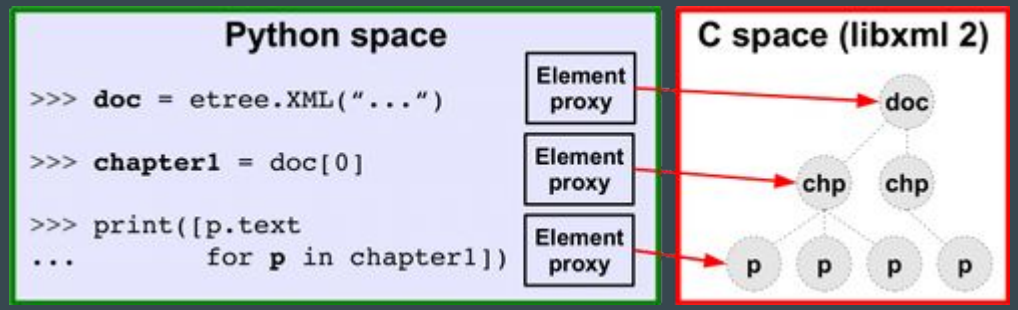
Nathaniel Lyon † Franz Sigel Samuel Sturgis	Sterling Price Ben McCulloch Nicholas Pearce

Units involved

Army of the West	Missouri State Guard Western Army 1st Division, Army of Arkansas

Web Scraping - lxml

- Lxml: XML and HTML with Python
- The lxml XML toolkit is a Pythonic binding for the C libraries libxml2 and libxslt.
  - Fully-featured, maintained, and fast.
  - Issues: low-level code, under-documented and segfault issues.
- Build tree from html source.
- Traverse the tree for desired data.



Initial Attempt - XPath

//*[@id="mw-content-text"]/table[1]/tbody/tr[8]/td[1]/a[1]

Battle of Wilson's Creek	
Part of the American Civil War	
Battle of Wilson's Creek by Kurz and Allison.	
Date	August 10, 1861
Location	Greene County and Christian County, Missouri
Result	Confederate States victory
Belligerents	
United States	Missouri (Confederate) Confederate States
Commanders and leaders	
Nathaniel Lyon † Franz Sigel Samuel Sturgis	Sterling Price Ben McCulloch Nicholas Pearce
Units involved	
Army of the West	Missouri State Guard Western Army 1st Division, Army of Arkansas

Initial Attempt - XPath Problems

- Difficult to automate
  - Table paths different between pages
  - Some elements missing between pages
  - Meta tags injected in markup by Wikipedia difficult to strip
- Difficult to get right
  - Trial and error required for all pages to get correct paths

Additional Foregone Research

- Neo4j Sparql Plugin.
  - Linked to other pages
- Graphipedia: tool for creating a Neo4j graph database of Wikipedia pages and the links between them.
  - Worked well! But didnt use Neo4j
- AutoSPARQL: create SPARQL queries over RDF knowledge bases from natural language with low effort.
  - Java 7, Maven, difficult setup

Lexico-Syntactic Patterns for Automatic Ontology Building

Carmen Klaussner
University of Nancy 2
carmen@wordsmith.de

Desislava Zhekova
University of Bremen
zhekova@uni-bremen.de

No.	Pattern
1.	NP_0 *including* NP_{1+i}
2.	NP_0 *such as* NP_{1+i}
3.	*by such* NP_0 *as* NP_{1+i}
4.	NP_0 *like* NP_{1+i}
5.	NP_0 *except* NP_{1+i}
6a.	NP_0 *e.g.* NP_{1+i}
6b.	NP_0 *i.e.* NP_{1+i}
7a.	NP_0, *(a) kind(s) | type(s) | form(s) of* NP_{1+i}
7b.	NP_0*: (a) kind(s) | type(s) | form(s) of* NP_{1+i}
8.	NP_0 *other than* NP_{1+i}
9.	*There (are | is) (could | would) be two types of* NP_0 *(: | ,)* NP_{1+i}
10a.	NP_0 *especially* NP_{1+i}
10b.	NP_0 *notably* NP_{1+i}
10c.	NP_0 *particularly* NP_{1+i}
10d.	NP_0 *usually* NP_{1+i}
10e.	NP_0 *mostly* NP_{1+i}
10f.	NP_0 *mainly* NP_{1+i}
10g.	NP_0 *principally* NP_{1+i}

Table 1: *Patterns for the acquisition of definitions*

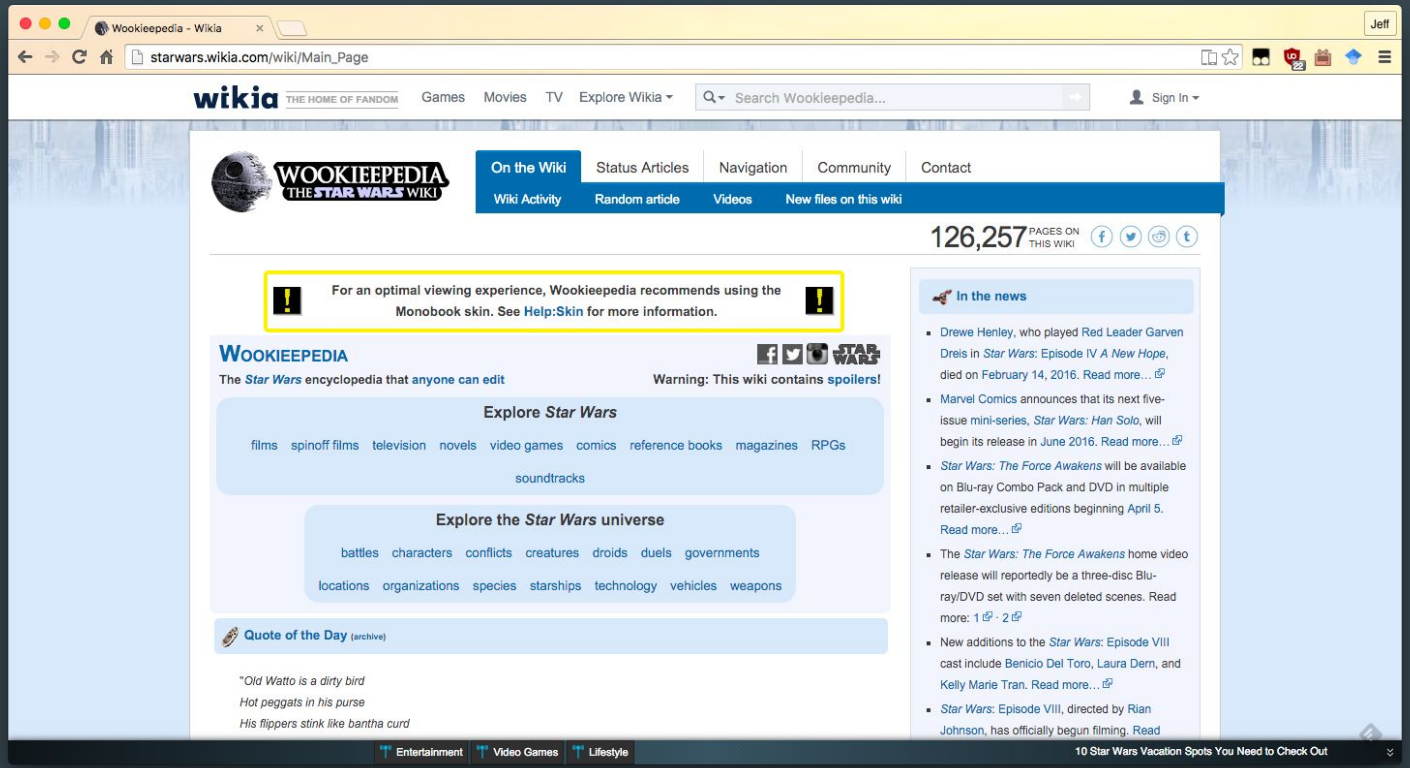
No.	Overall occurrence	% of success	one-directional
1.	601	409 (68%)	No
2.	2389	2107 (88.2%)	Yes
3.	9	9 (100%)	Yes
4.	401	330 (82%)	Yes
5.	18	10 (56%)	Yes
6a.	170	134 (79%)	Yes
6b.	no occur.	nil	nil
7a.	48	31 (65 %)	Yes
7b.	7	6 (85%)	Yes
8.	19	16 (84 %)	Yes
9.	4	4 (100%)	Yes
10a.	61	9 (89%)	Yes
10b.	22	13 (59%)	Yes
10c.	29	23 (79%)	Yes
10d.	9	7 (78%)	Yes
10e.	5	4 (80%)	Yes
10f.	3	2 (67%)	Yes
10g.	no occur.	nil	nil

Table 2: *Pattern success rates*

Web Scraping - BeautifulSoup

- Sits atop an HTML or XML parser (lxml), providing Pythonic idioms for iterating, searching, and modifying the parse tree.
- Beautiful Soup parses anything you give it, and does the tree traversal stuff for you.
- User can quickly specify commands to:
  - Find all links.
  - Find all links of a specified class.
  - Find all links which match a specified string.

Wookiepedia



Wookiepedia

C-3PO

126,257 PAGES ON THIS WIKI

View source | Talk 144 | LEGENDS

CANON | LEGENDS

[Show]

[Show]

"I am C-3PO, human-cyborg relations."
—C-3PO[src]

C-3PO

C-3PO, sometimes spelled **See-Threepio** and often referred to as **Threepio**, was a bipedal, humanoid protocol droid designed to interact with organics, programmed primarily for etiquette and protocol. He was fluent in over six million forms of communication, and developed a fussy and worry-prone personality throughout his many decades of operation. After being destroyed and discarded on the planet Tatooine before 32 BBY, C-3PO was rebuilt; his salvaged nature gave him special qualities that distinguished him from similar droid models. Along with his counterpart, the astromech droid R2-D2, C-3PO constantly found himself directly involved in pivotal moments of galactic history, and aided in saving the galaxy on many occasions. C-3PO considered various droids and organics to be friends of his, and

Wookiepedia

"I am C-3PO, human-cyborg relations."
—C-3PO[src]

C-3PO, sometimes spelled **See-Threepio** and often referred to as **Threepio**, was a bipedal, humanoid protocol droid designed to interact with organics, programmed primarily for etiquette and protocol. He was fluent in over six million forms of communication, and developed a fussy and worry-prone personality throughout his many decades of operation. After being destroyed and discarded on the planet Tatooine before 32 BBY, C-3PO was rebuilt; his salvaged nature gave him special qualities that distinguished him from similar droid models. Along with his counterpart, the astromech droid R2-D2, C-3PO constantly found himself directly involved in pivotal moments of galactic history, and aided in saving the galaxy on many occasions. C-3PO considered various droids and organics to be friends of his, and was very dedicated to them, as well as to any master that he served.

Originally activated on Affa in 112 BBY, C-3PO had served as a protocol droid to the emissary of the Manakron system. Nearly eighty years later, he was gutted and discarded on the streets of Mos Espa, a city on the Outer Rim world of Tatooine. After being rebuilt by the Human slave Anakin Skywalker, C-3PO served Skywalker and his mother Shmi for over ten years, performing household chores and helping Skywalker earn his freedom by winning a pod race. Skywalker left Tatooine but returned in 22 BBY when his mother passed away, and C-3PO was given to Skywalker, now a Jedi Padawan, by Shmi's stepson Owen Lars. C-3PO, Skywalker, R2-D2, and the Naboo Senator Padmé Amidala immediately became embroiled in the Clone Wars, a galaxy-wide conflict between the Galactic Republic and the

C-3PO	
Production information	
Homeworld	Tatooine[1]
Date created	112 BBY, Affa[2]
Date destroyed	3 ABY, Bespin (temporarily dismantled, rebuilt)[3]
Creator	Anakin Skywalker[4]
Manufacturer	Cybot Galactica[1]
Model	3PO-series protocol droid[1]
Class	Protocol droid[5]
Technical specifications	
Height	1.67 meters[1][6]

Wookiepedia: Categories

Category:Browse - Wookie

Jeff

starwars.wikia.com/wiki/Category:Browse

wikia THE HOME OF FANDOM Games Movies TV Explore Wikia Search Wookieepedia... Sign In

WOOKIEEPEDIA THE STAR WARS WIKI

On the Wiki | Status Articles | Navigation | Community | Contact | Contribute

Wiki Activity | Random article | Videos | New files on this wiki

Browse

Category page

126,257 PAGES ON THIS WIKI

Edit | Talk 0

This is a starting point that can be used to access any article on Wookieepedia.

In other languages

Dansk | Deutsch | Español | Français | Italiano | Polski | Português | Русский | Suomi

Subcategories

This category has the following 3 subcategories, out of 3 total.

- [−] Browse (3 C)
  - [+] Browse (3 C)
  - [+] Star Wars (4 C, 1 P)
  - [+] Wookieepedia (28 C, 47 P)
- [−] Star Wars (4 C, 1 P)
  - [+] Articles by canonicity (3 C)
  - [×] Disambiguation pages (2,680 P)
  - [−] In-universe articles (12 C, 9 P)
    - [+] Artifacts (2 C, 95 P)
    - [+] Awards (3 C, 59 P)

Recent Wiki Activity

- The rebellion
  edited by AnilSerifoglu 2 seconds ago
- Star Wars: Episode VII The Force Awakens
  edited by AdamDeanHall 4 minutes ago
- The Mystery of Chopper Base
  edited by AnilSerifoglu 6 minutes ago
- Bloodline (novel)
  edited by AnilSerifoglu 12 minutes ago

See more >

Entertainment | Video Games | Lifestyle

10 Star Wars Vacation Spots You Need to Check Out

Wookiepedia: Categories

Clones

Category page

126,262 PAGES ON THIS WIKI

Edit | Talk 0

This category is for clones.

In other languages

Español | Nederlands | Русский

Subcategories

This category has the following 2 subcategories, out of 2 total.

- [+] Human clones (2 C, 45 P)

K

- [×] Khommites (8 P)

Pages in category "Clones"

The following 35 pages are in this category, out of 35 total. Perform a category intersection.

A

- Aleksin

B

- Blind berserker

C

- Chewbaacca

M

- Maulkiller
- Mitth'raw'nuruodo (clone)
- Morgukai Shadow Army

O

- Ohali Two

U cont.

- Unidentified clone 2 (cremation center)
- Unidentified clone 3 (cremation center)
- Unidentified clone medical officer
- Unidentified clone medical officer

Wookiepedia: Categories

Clones

Category page

126,262 PAGES ON THIS WIKI

Edit | Talk 0

This category is for clones.

In other languages

Español | Nederlands | Русский

Subcategories

This category has the following 2 subcategories, out of 2 total.

- [+] Human clones (2 C, 45 P)

K

- [×] Khommites (8 P)

Pages in category "Clones"

The following 35 pages are in this category, out of 35 total. Perform a category intersection.

A

- Aleksin

B

- Blind berserker

C

- Chewbaacca

M

- Maulkiller
- Mitth'raw'nuruodo (clone)
- Morgukai Shadow Army

O

- Ohali Two

U cont.

- Unidentified clone 2 (cremation center)
- Unidentified clone 3 (cremation center)
- Unidentified clone medical officer
- Unidentified clone medical officer

Wookiepedia: Categories

Clones

Category page

126,262 PAGES ON THIS WIKI

Edit | Talk 0

This category is for clones.

In other languages

Español | Nederlands | Русский

Subcategories

This category has the following 2 subcategories, out of 2 total.

- [+] Human clones (2 C, 45 P)

K

- [×] Khommites (8 P)

Pages in category "Clones"

The following 35 pages are in this category, out of 35 total. Perform a category intersection.

A

- Aleksin

B

- Blind berserker

C

- Chewbaacca

M

- Maulkiller
- Mitth'raw'nuruodo (clone)
- Morgukai Shadow Army

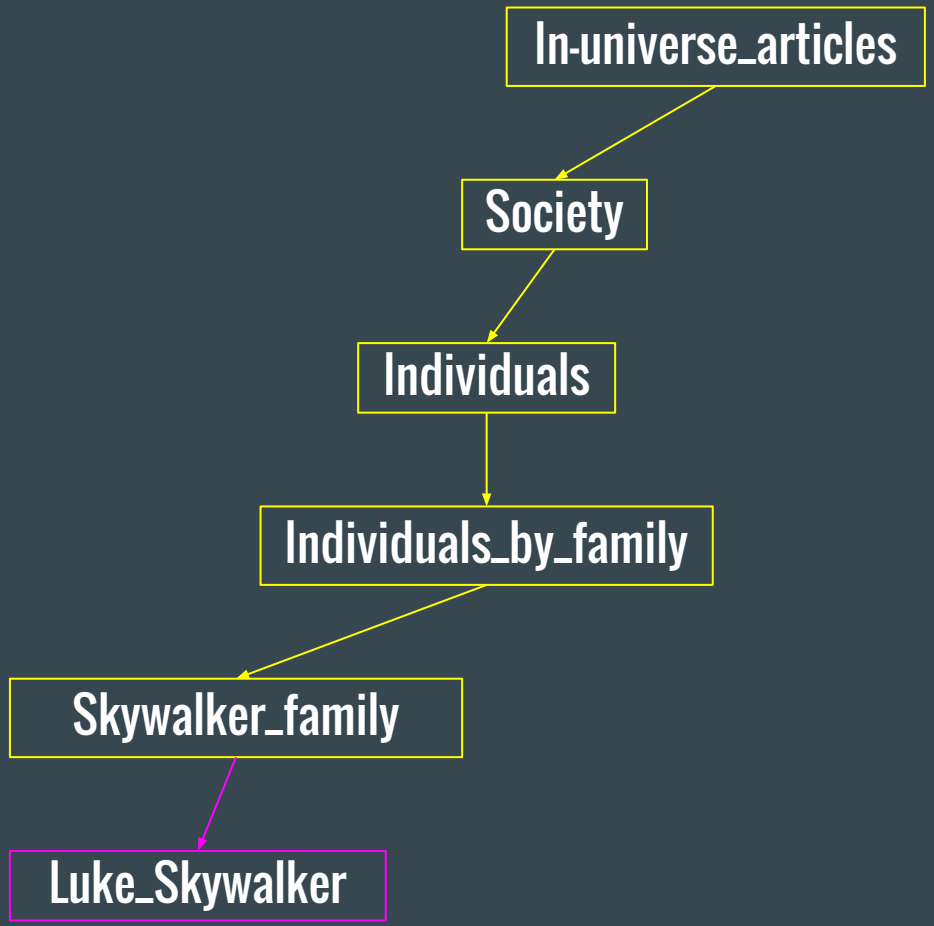
O

- Ohali Two

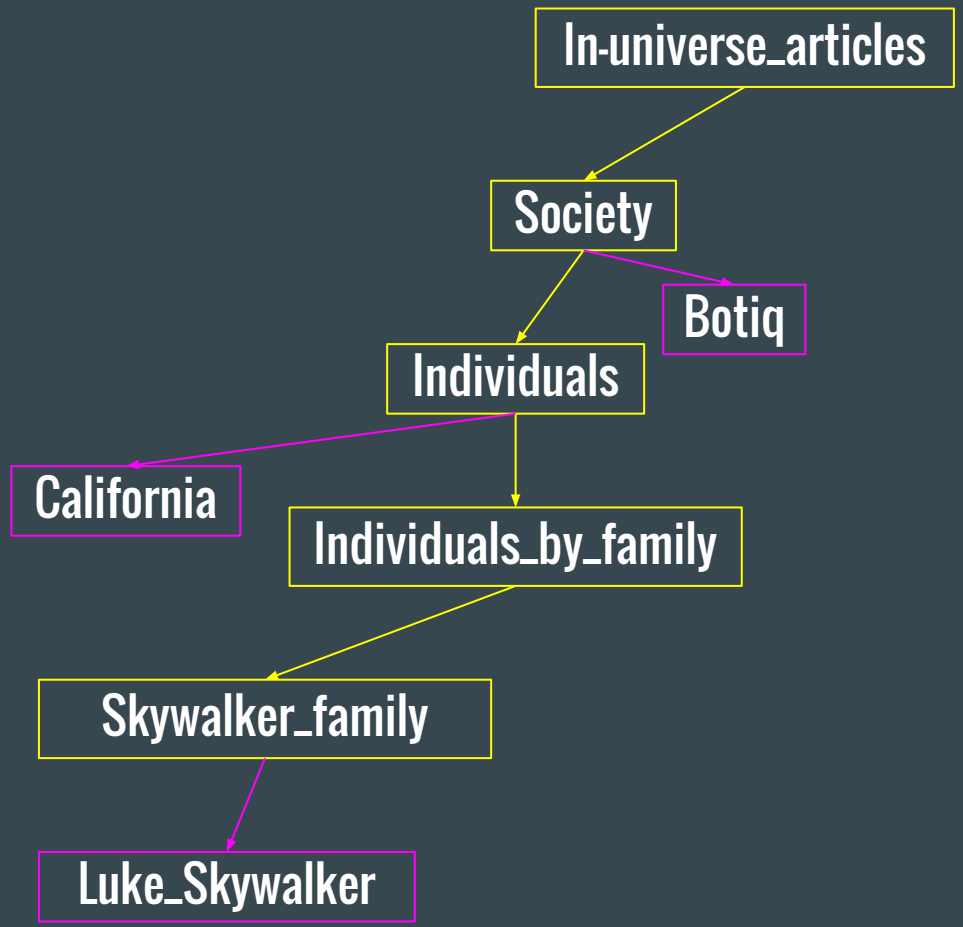
U cont.

- Unidentified clone 2 (cremation center)
- Unidentified clone 3 (cremation center)
- Unidentified clone medical officer
- Unidentified clone medical officer

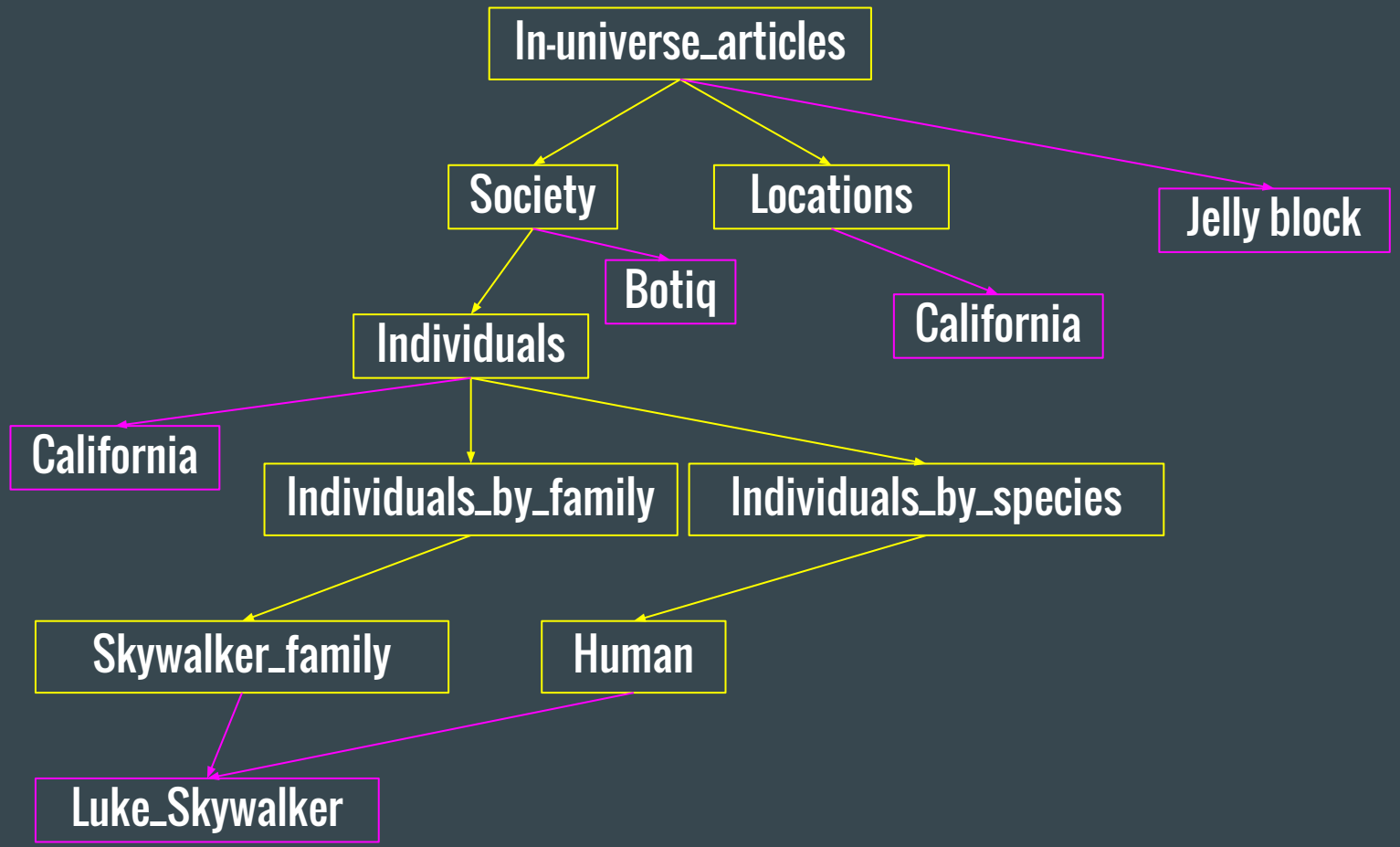
Wookiepedia: Categories



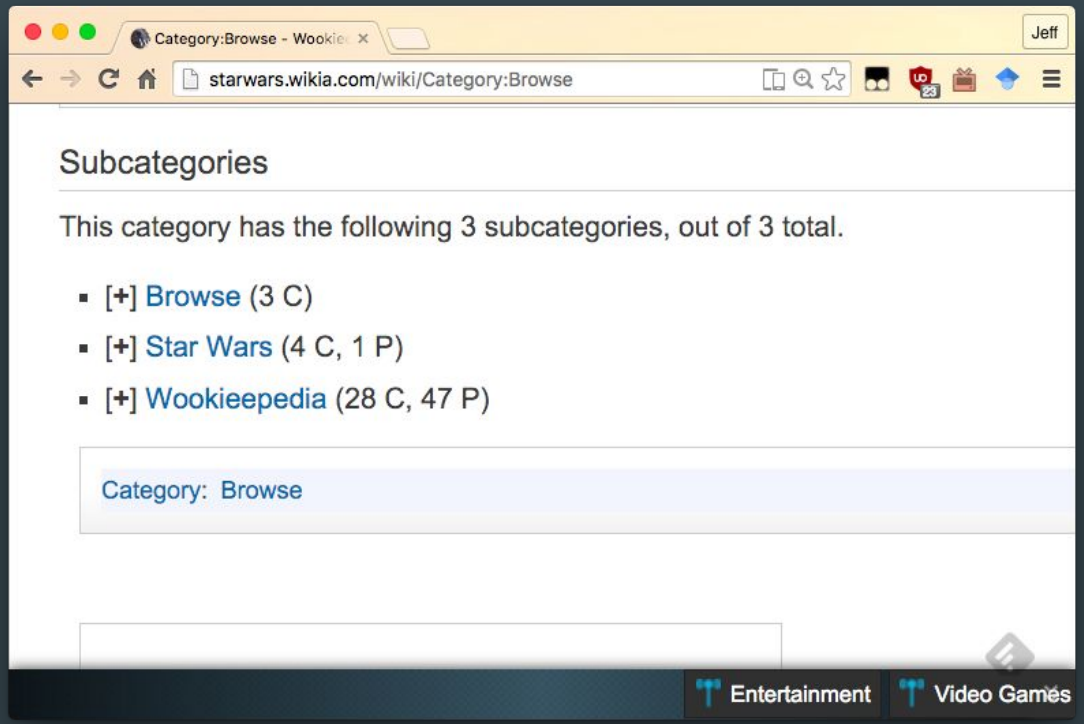
Wookiepedia: Categories



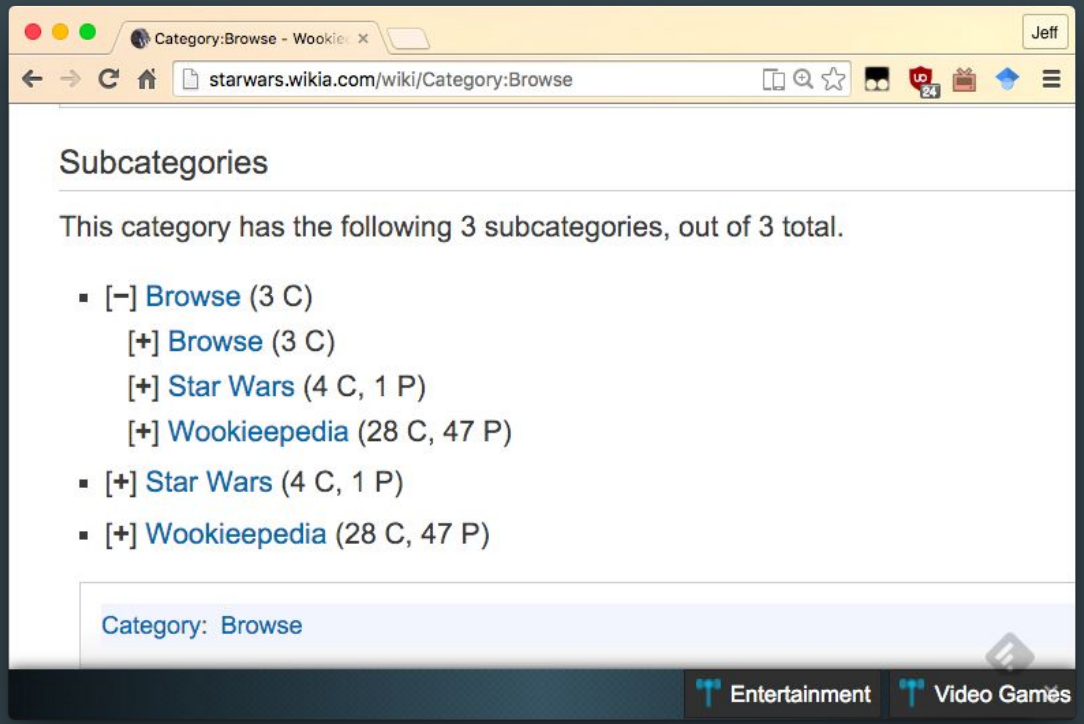
Wookiepedia: Categories



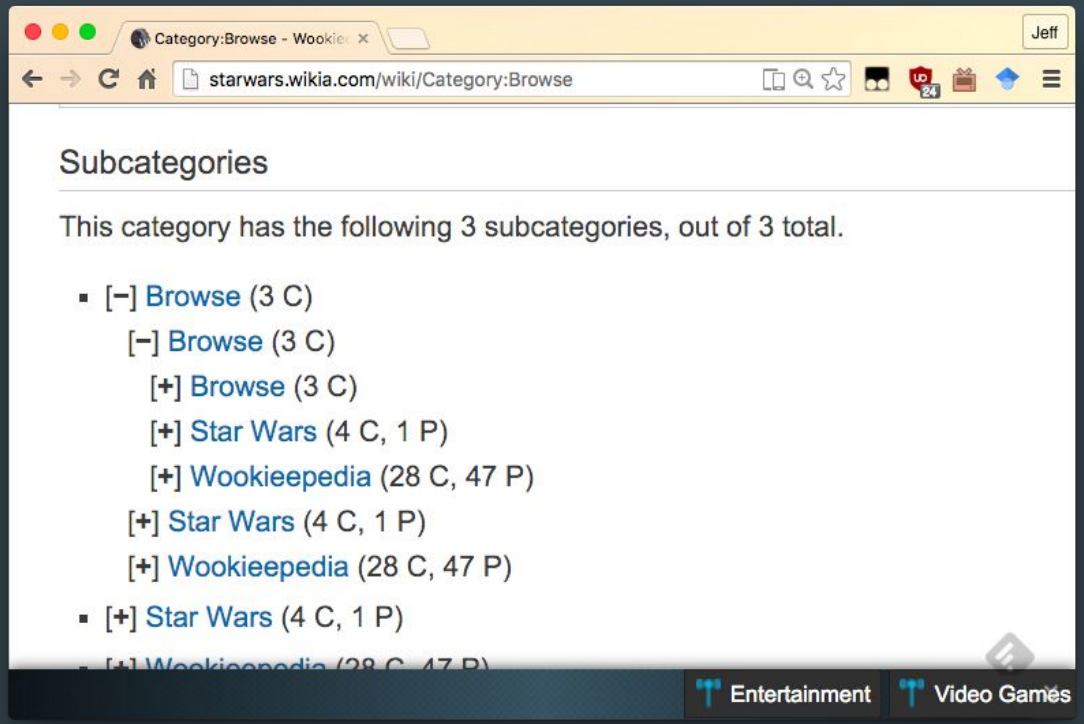
Wookiepedia: Traversal Problems



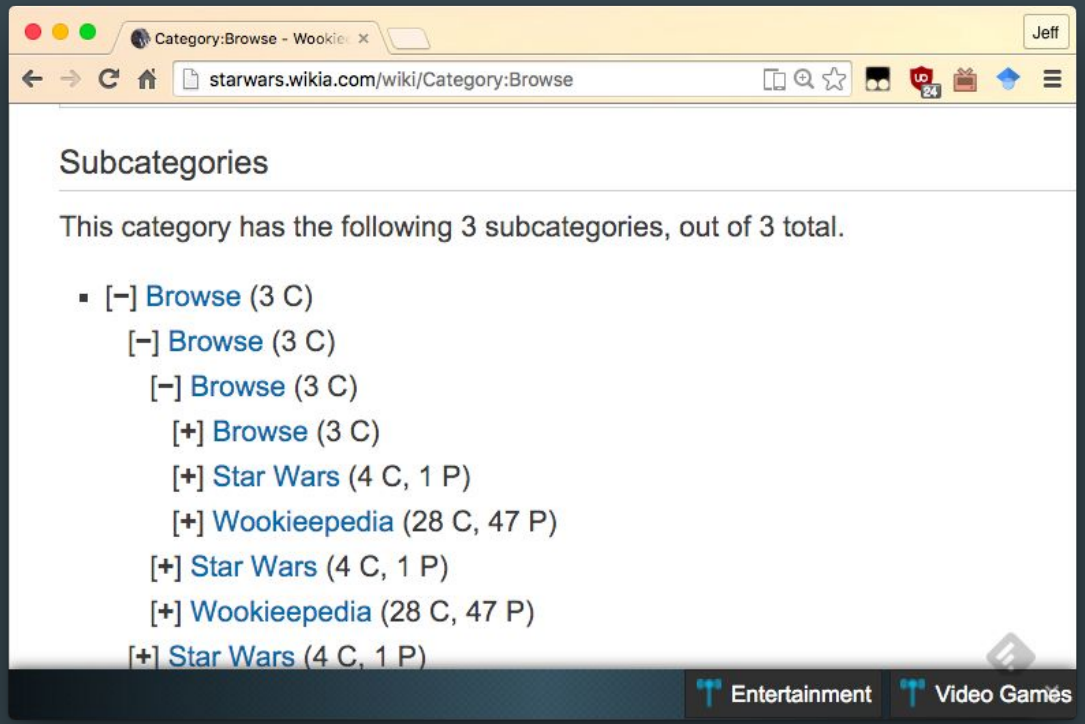
Wookiepedia: Traversal Problems



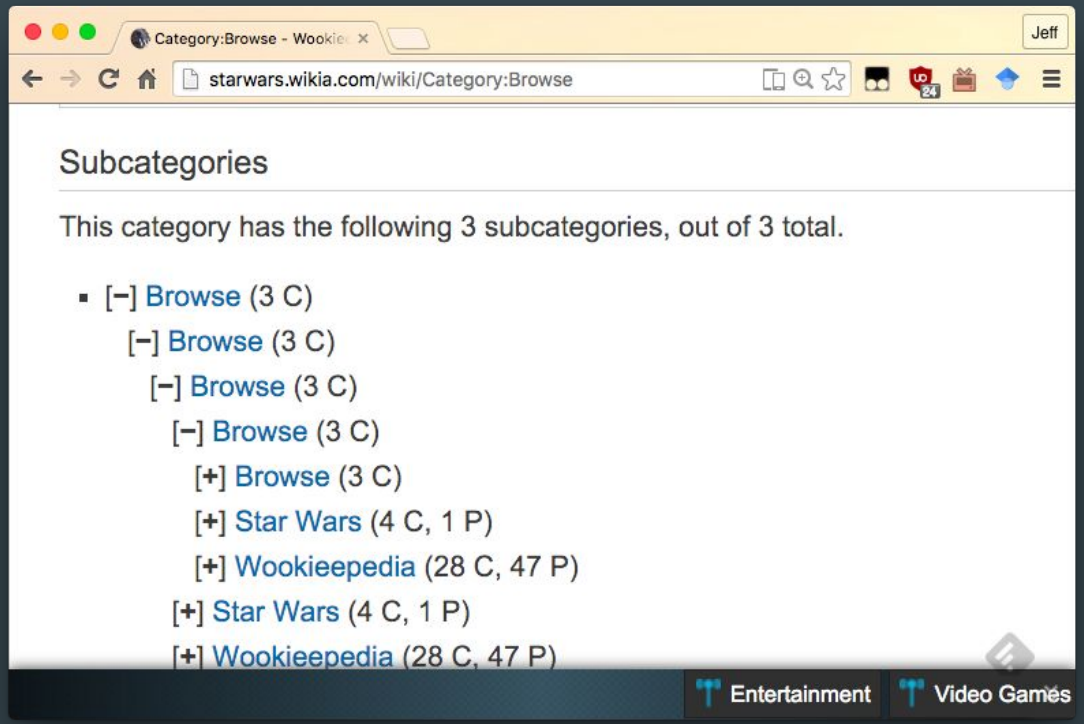
Wookiepedia: Traversal Problems



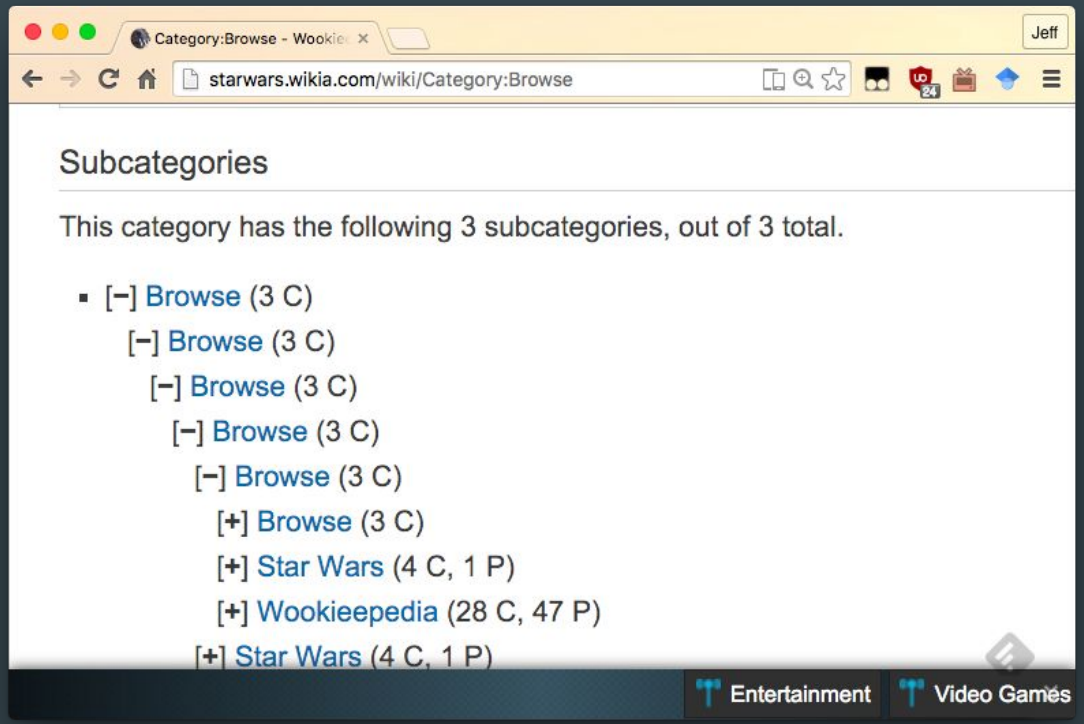
Wookiepedia: Traversal Problems



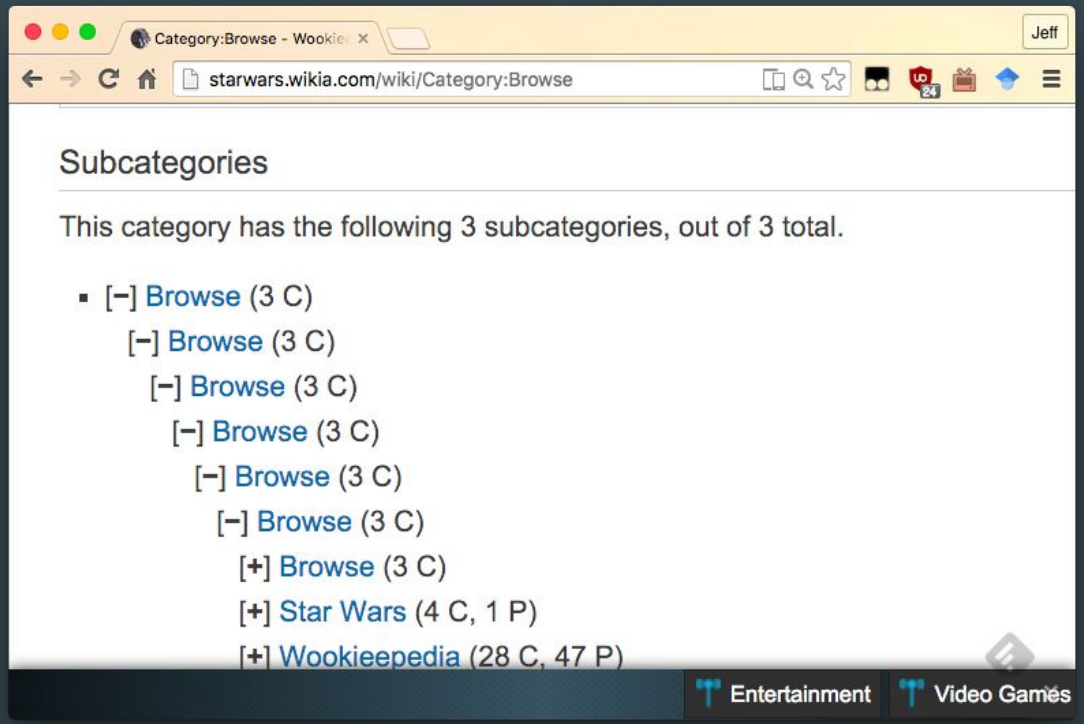
Wookiepedia: Traversal Problems



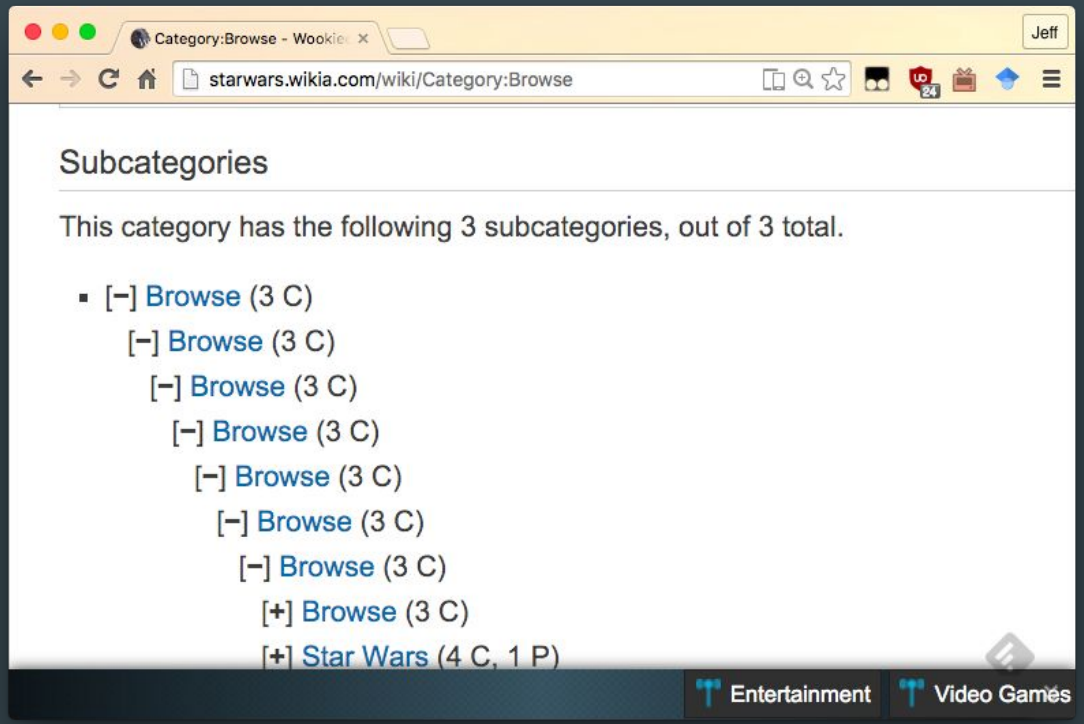
Wookiepedia: Traversal Problems



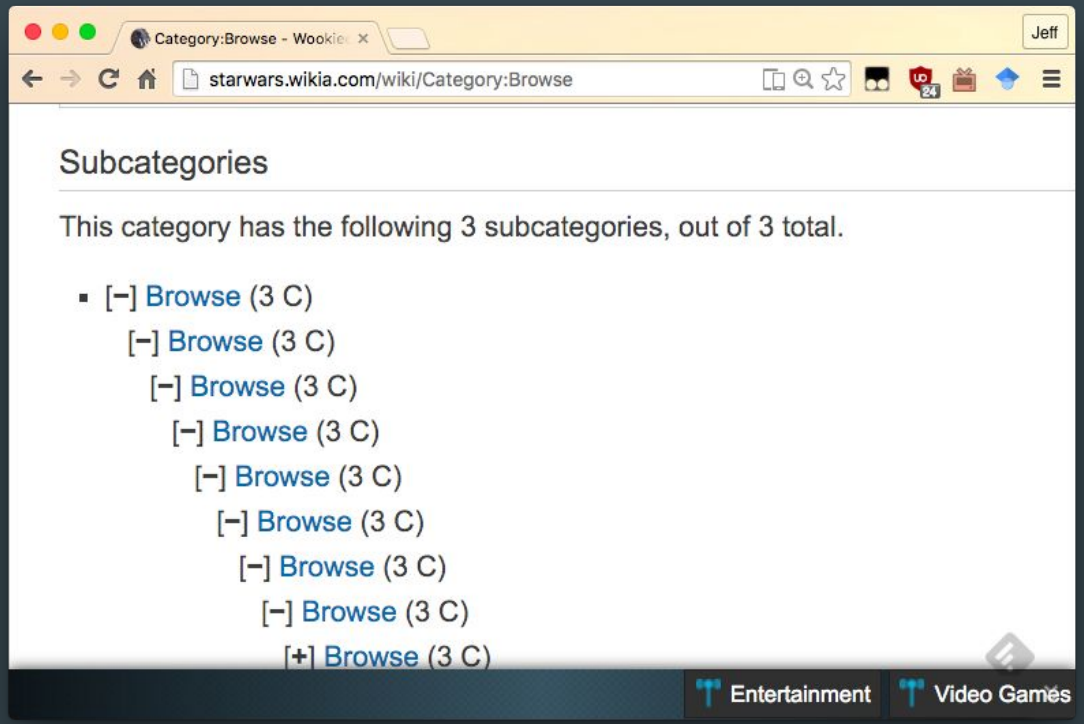
Wookiepedia: Traversal Problems



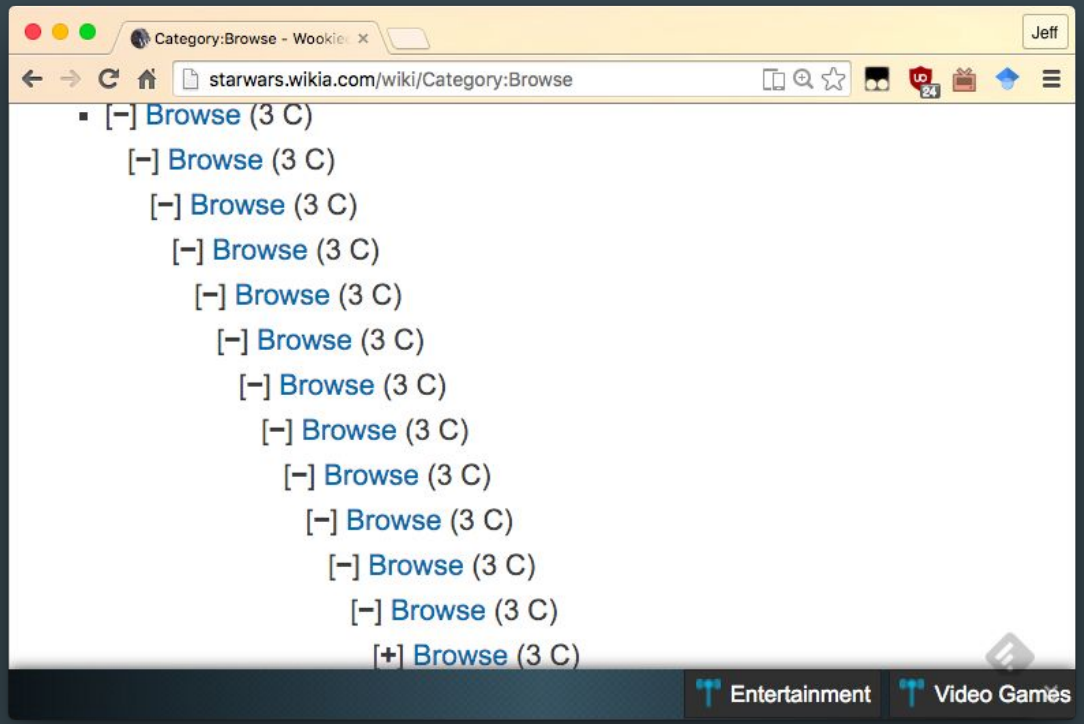
Wookiepedia: Traversal Problems



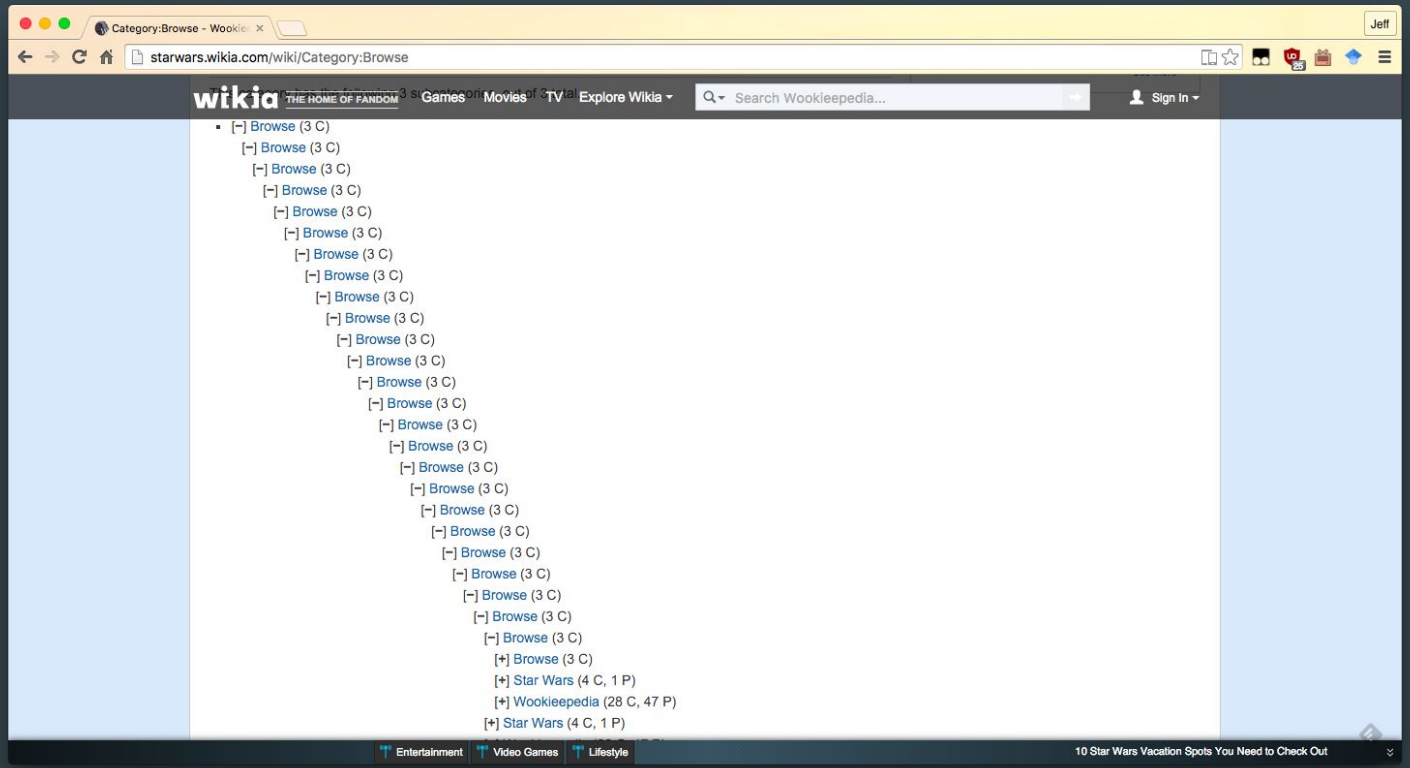
Wookiepedia: Traversal Problems



Wookiepedia: Traversal Problems

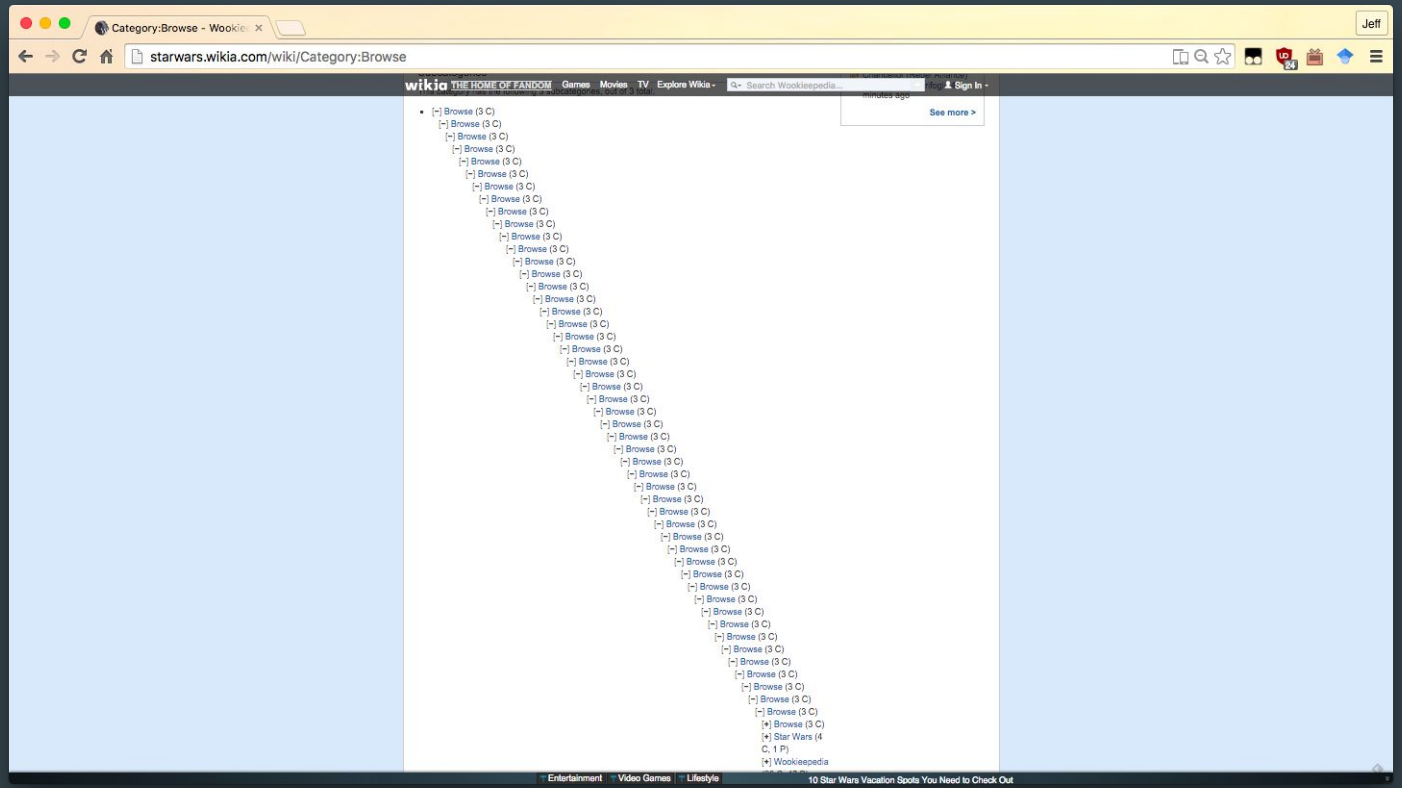


Wookiepedia: Traversal Problems



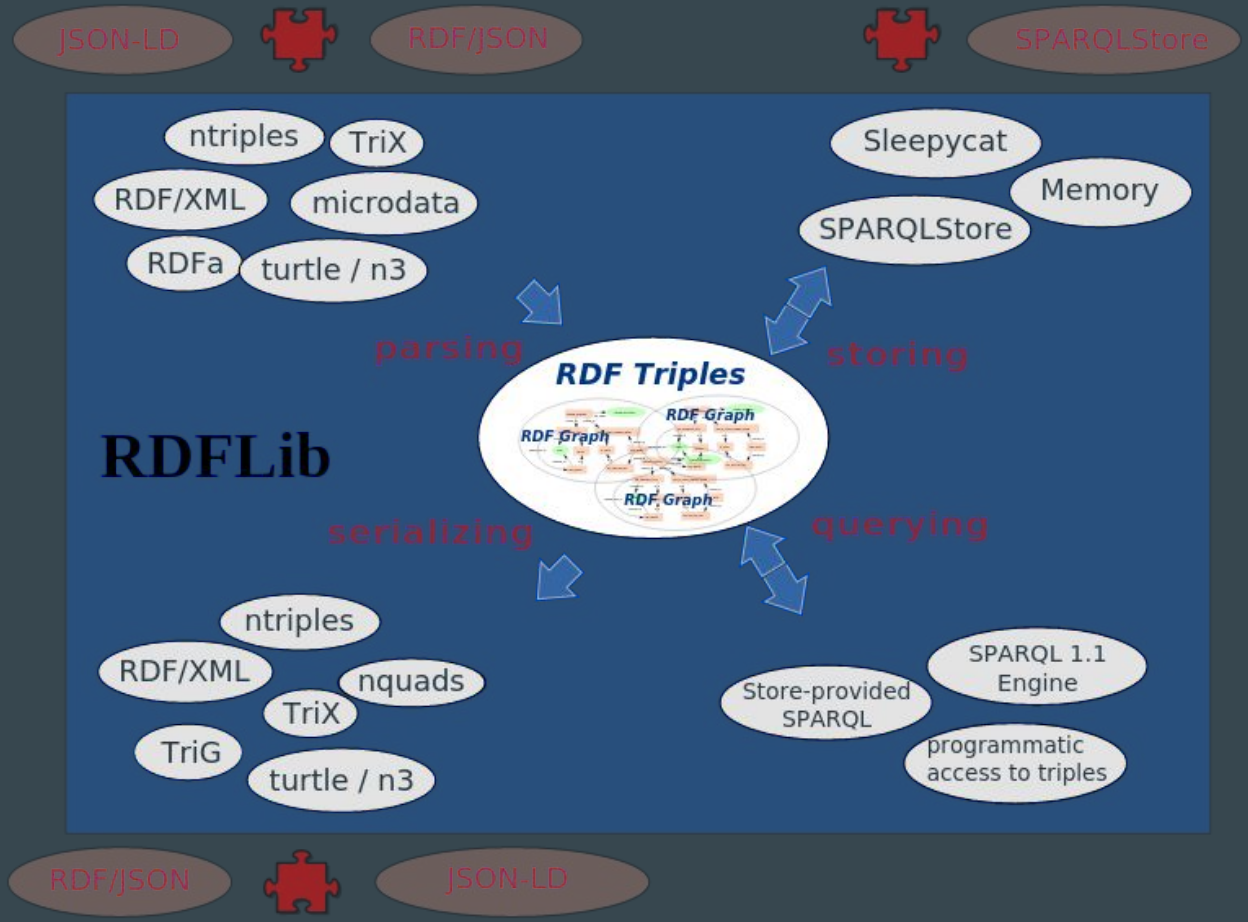
- [−] Browse (3 C)
  - [−] Browse (3 C)
    - [−] Browse (3 C)
      - [−] Browse (3 C)
        - [−] Browse (3 C)
          - [−] Browse (3 C)
            - [−] Browse (3 C)
              - [−] Browse (3 C)
                - [−] Browse (3 C)
                  - [−] Browse (3 C)
                    - [−] Browse (3 C)
                      - [−] Browse (3 C)
                        - [−] Browse (3 C)
                          - [−] Browse (3 C)
                            - [−] Browse (3 C)
                              - [−] Browse (3 C)
                                - [−] Browse (3 C)
                                  - [−] Browse (3 C)
                                    - [−] Browse (3 C)
                                      - [−] Browse (3 C)
                                        - [−] Browse (3 C)
                                          - [−] Browse (3 C)
                                            - [−] Browse (3 C)
                                              - [−] Browse (3 C)
                                                - [−] Browse (3 C)
                                                  - [+] Browse (3 C)
                                                  - [+] Star Wars (4 C, 1 P)
                                                  - [+] Wookieepedia (28 C, 47 P)
                                                - [+] Star Wars (4 C, 1 P)

Wookiepedia: Traversal Problems



RDFLib – Building a Graph

- Python library for working with RDF.
- Simple API for generating graphs of RDF triples:
  - BNodes
  - Literals
  - URIRefs
- Tools for parsing, storing, serialization, and querying.


JSON-LD
RDF/JSON
SPARQLStore
ntriples
TriX
RDF/XML
microdata
RDFa
turtle / n3
Sleepycat
Memory
SPARQLStore
parsing
storing
RDF Triples
RDF Graph
RDF Graph
RDF Graph
RDFLib
serializing
querying
ntriples
RDF/XML
nquads
TriX
TriG
turtle / n3
SPARQL 1.1 Engine
Store-provided SPARQL
programmatic access to triples
RDF/JSON
JSON-LD

RDFLib - Terminology

- BNode: a blank node representing a resource for which a URI or literal is not given. The resource represented by a blank node is also called an anonymous resource.
- Literal: attribute values in RDF, for instance, a person's name, date of birth, height, etc. Literals can have datatypes or a language tags.
- URIRef: a Unicode string representing an absolute URI.

RDFLib - Adding Triples

- RDF triples: subject, predicate, and object.
  - Subject denotes the resource, using URIRef keyword.
  - Predicate expresses a relationship between the subject and the object.
  - Object: a URI reference, a literal or a bnode.
- graph.add(

  URIRef(“Grandon_Holleck”),

  RDFS.subClassOf,

  URIRef(“Governers_Of_The_Galactic_Empire”)

)

RDFLib - Querying with SPARQL

```
query_result = graph.query(
    '''SELECT ?subject
      WHERE {
        ?subject rdfs:subClassOf* sww:Governers_Of_The_Galactic_Empire
      }''')
```

Front End Development

- Front end developed in Python and FLASK
- Original implementation attempted to use Quepy
  - Natural language to SPARQL query converter
  - Unreliable, used REGEX matching for query types
- Final implementation manually translated specific queries to SPARQL - ie:

What are types of Governers of the Galactic Empire?

```
SELECT ?title WHERE { ?subject rdfs:subClassOf swb:Governers_Of_The_Galactic_Empire . ?subject owl:title ?title .}
```

Front End Development

- Publically available ontology:
  - jamesbilous.com/static/test.owl
- Pickled (serialized) for quick load into rdf graph
- Conversion from ontology to graph takes ~ 5-6 min
- Supported Queries:
  - Immediate Parents of X
  - Subcategories and Instances Of X
  - Parents of X
  - Children of X
  - Generic SPARQL queries
- Demo