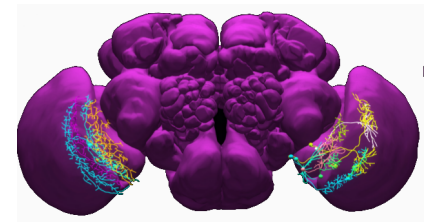


# The PeARS project

Low-resource web search applications from  
cognitive models

Aurelie Herbelot, Tanise Pagnan Ceron,  
Le Minh Nhut Truong



# What is PeARS?

[Search](#)[Indexer](#)[Pod management](#)[My orchard](#)[DB admin](#)[F.A.Q.](#)[Acknowledgments](#)[Search](#)

# A local search engine



## Distributed search engine - Wikipedia

Distributed search engine A distributed search engine is a search engine where there is no central s ...

[https://en.wikipedia.org/wiki/Distributed\\_search\\_engine](https://en.wikipedia.org/wiki/Distributed_search_engine)

## The Search Engine List | Comprehensive list of Search Engines

Home The Search Engine List is the web's ...

<http://www.thesearchenginelist.com/>

## Search Engines and Ethics (Stanford Encyclopedia of Philosophy)

Search Engines and Ethics First publishe ...

<https://plato.stanford.edu/entries/ethics-search/>

## Distributed Search Engines- And Why We Need Them In The Post-Snowden World | Techdirt

Distributed Search Engines- And Why We N ...

<https://www.techdirt.com/articles/20140701/03143327738/distributed-search-engines-why-we-need-them-post-snowden-world.shtml>

Peers



search engines



Me

# Index your Web

[Search](#)[Indexer](#)[Pod management](#)[My orchard](#)[DB admin](#)[F.A.Q.](#)[Acknowledgments](#)

Welcome to your orchard! Here, you can check out the number of pages you have indexed, ordered by keyword. You can also 'pick' a pod, that is, create a shareable version of your pod, available as either a CSV file or an image. It's simple: just click on your chosen pod and wait for your files to be created.

**Warning:** when you pick a pod, it gets deleted from your orchard. So only click when your pod is ready to share!

**generic**  
983

**computational  
linguists**  
751

**linguistics**  
2

**harrypotter**  
593

**syria**  
563

**backpain**  
373

**opera**  
258

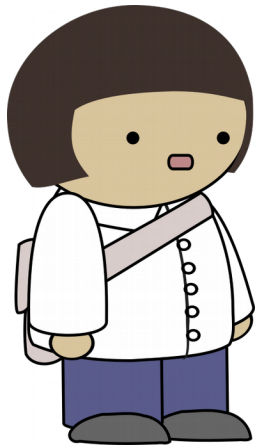
**brexit**  
360

**AI**  
359

**string theory**  
338

**DID**  
1

# Share your Web



What is string theory?

Check out my pods!

**Politics and society**

**Brexit**

624 pages about the Brexit process

**Anarchism**

1175 pages about anarchist movements

**The Black Panthers**

506 pages about the political organisation founded in 1966

**Science and culture**

**AI**

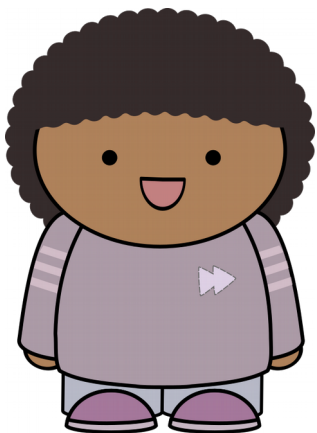
359 pages about AI and its relation to society

**String theory**

307 pages about string theory

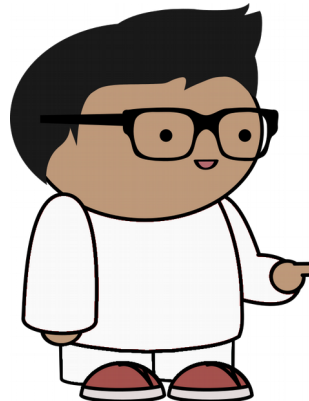
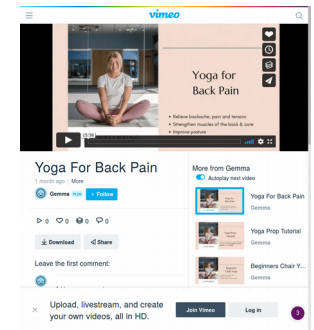
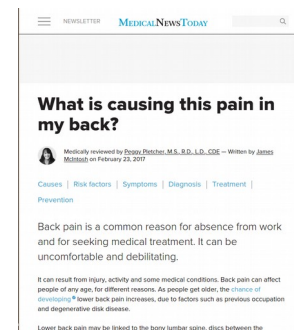
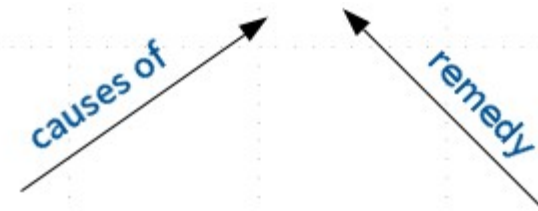
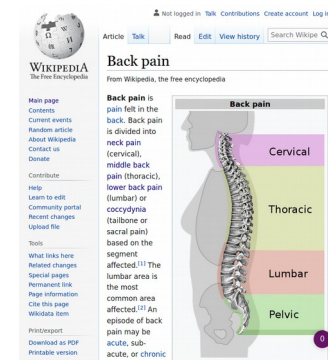
**Opera**

258 pages about opera



# Benefits

- Decentralised search
- Curate Web data locally, share your annotation!





# Can we help the curator?

<https://www.medicalnewstoday.com/articles/172943>

(Text: Causes of back pain)

<https://praxisinstitute.org/wp-content/uploads/2019/10/Anatomy-of-SC.png>

(Image: Anatomy of the spinal cord)

<https://vimeo.com/553448612>

(Video: Yoga for back pain)

Curation: lots of manual work if resources do not already contain structured data. The user might have a few thousands of URLs in their local pods. Many of those URLs won't be very informative with respect to the actual content of the resource, and they might span various formats, from text to images to videos.



# Can we help the curator?

7196759210defdc0 = **back pain, cause, injury, strain, spine**

(Text: Causes of back pain)

2638458674fhsgi4 = **spinal cord, spine, anatomy, description**

(Image: Anatomy of the spinal cord)

5782946738shsgd2 = **back pain, strain, remedy, yoga, exercise**

(Video: Yoga for back pain)

**What if we could hash the content of those resources into comparable semantic representations?**

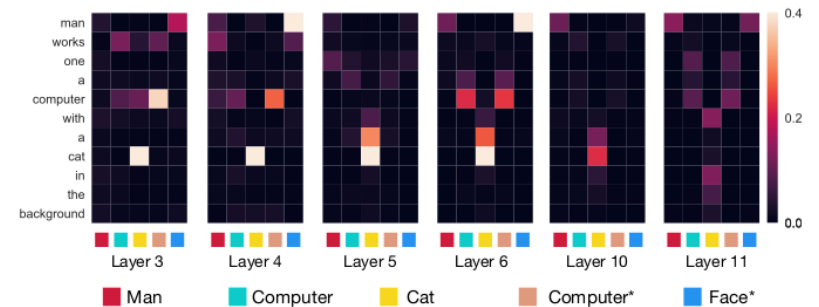
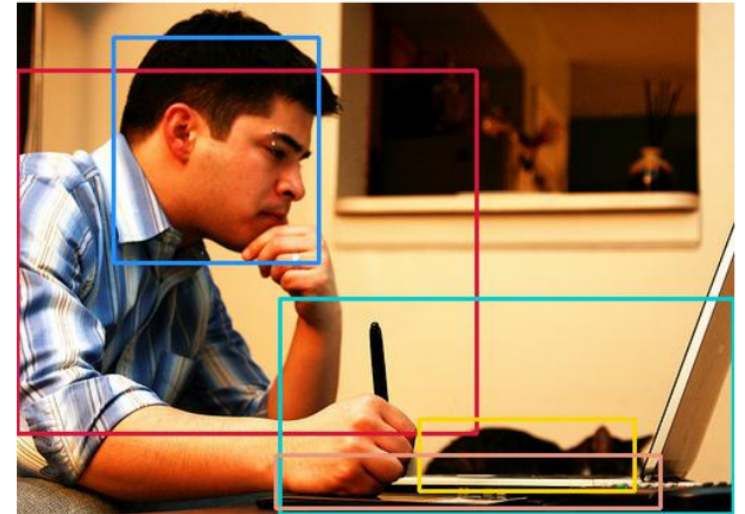
→ Cluster documents into subtopics, automatically suggest keywords, associate multimodal content with text fragments...  
and of course, search better!



# Semantics is expensive

Deep learning models can extract semantic representations from text and images, but they are too expensive to train for the average user.

- High levels of CO2 emissions (training pollutes!)
- Pretrained models are provided by large companies as ‘black boxes’ .  
No transparency.

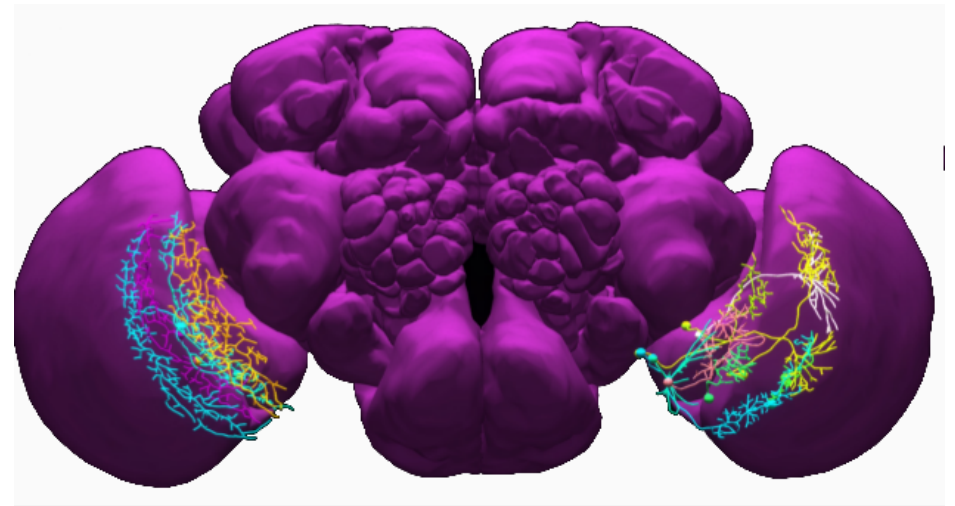


Li et al (2019),

<https://arxiv.org/abs/1908.03557>

# The Fruit Fly Algorithm

- The olfactory system of the fruit fly: a powerful hashing algorithm, a simple feedforward architecture.
- Already used for hashing word and image vectors.
- Can it encode the semantics of *any* URI, from scratch?



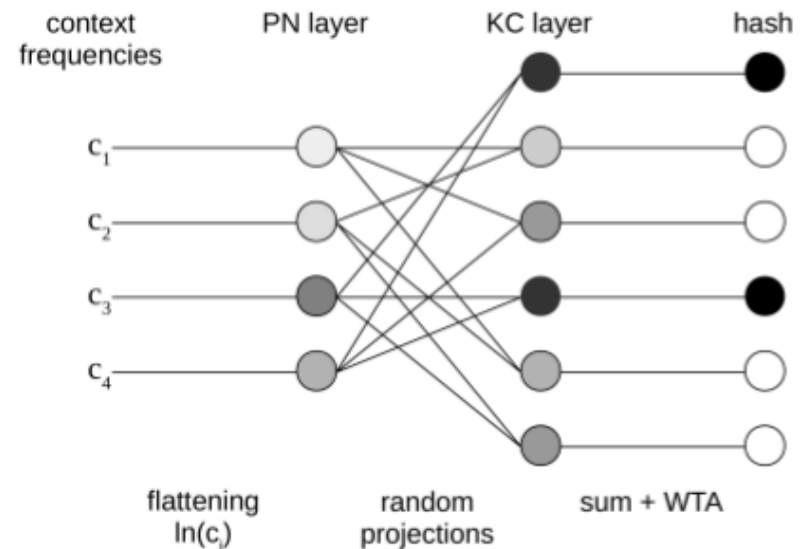
Ukani et al (2019), <https://doi.org/10.1101/580290>

# Spawning fruit flies

We are spawning thousands of fruit flies, with the idea to produce the best possible Web document presentations.

The most intelligent specimen will be integrated into the PeARS framework to support search and local annotation.

A typical fruit fly model can be stored in 100-200KB. The algorithm is highly efficient. The process can be reproduced by anybody on their personal laptop.



<https://www.flickr.com/photos/nasamarshall/15842201129>, CC BY-NC 2.0

# Read more about the project!

- Our website:

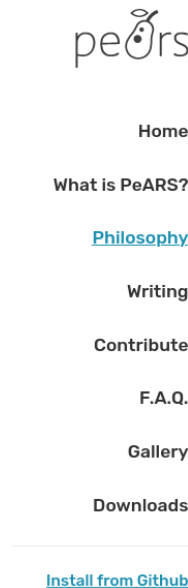
<https://pearsproject.org/index.html>

- Our GitHub:

<https://github.com/PeARSearch>

- The Fruit Fly project will be evolving at:

<https://github.com/PeARSearch/PeARS-fruit-fly>



## Our philosophy

“ We often think of Web data as being free. But information is only free if you can find it. People need to own the means of search.

### Distributed and collaborative

PeARS is a search engine run by all and for all. Data and algorithms are built and shared by the people who will use them. Everybody who browses the Web is a potential contributor.



### Lightweight

We want the search engine you use everyday to run efficiently on your laptop. No server farms needed!



### Open Source & Free

We are strong supporters of Free Software and Open Source philosophy. Our source code is open and wholeheartedly [welcome contributions](#).



# Acknowledgements



<https://openclipart.org/artist/anarres>  
for the cartoon characters in this  
presentation.

The Center for Mind/Brain Sciences,  
for hosting the project