Internet Archive

20 years / 20 petabytes
Internet Archive is a non-profit library of millions of free books, movies, software, music, and more.
Digital Storage

- Lots of hard drives
- 56.36 petabytes of disk space
- 21.93 petabytes of archival content
- Stored redundantly in different physical locations
DIY storage cluster

- Archival on-disk format
- Efficient (cost + energy)
Upload

- 8M Texts
- 2.4M Audio recordings
- 140K Concerts
- 1.9M Videos
- 100K Software items
- ... and your stuff!
Upload

- S3-compatible API
- Metadata API
- `pip install internetarchive`
- https://archive.org/help/abouts3.txt
Upload

Using IA S3 to back distributed art projects
Physical Storage
Upload
Wayback Machine

- Started in 1996 with Alexa crawls
- Updated in **realtime** at 2000 urls/sec
- 475,951,398,000 captures
- 10.5 petabytes of WARC data
- Intl Internet Preservation Consortium
- Open source on github.com/internetarchive
Realtime Wayback

Deleted within two hours:

http://vk.com/strelkov_info

1306 записей

Сводки от Стрелкова Игоря Ивановича
17.07.2014 17:50 (мск) Сообщение от ополчения.
"В районе Тореза только что сбили самолет An-26, является где-то за шахтой "Прогресс". Предупреждали же - не летать в "нашем небе". А вот и видео-подтверждение очередного "птичкапада". Птичка упала за террикон, жилой сектор не зацепила. Мирные люди не пострадали.
А также еще есть информация о втором сбитом самолете, вроде бы Су."
Capture a web page as it appears now for use as a trusted citation in the future. Only available for sites that allow crawlers.
“Brief overnight weekend stays” dropped precipitously during the days surrounding Valentine’s Day. The patterns are fascinating.
Wayback Under the Hood

- 10.5 petabytes of WARC data
- 2nd Level CDX Index: 20TB compressed
- 1st level: 13GB. Fits in core.
Wayback APIs

- Availability and CDX APIs: build your own tools on top of Wayback
- Open source tools to read/write WARCds
- WARC Support in `wget`:

  ```bash
  wget --warc-file
  ```
Books

- 2.5 million books with Scribe bookscanner
- 1000 books/day
- 35 scanning centers on 5 continents
- 800 million images
- Open-sourced gphoto Canon 1D/5D drivers
Books
Books

A general history of birds

124

EATING.


The only description I find, given by Forskal, mentions that the bird is wholly of a green colour; the throat yellow; the bill black and rather small; tongue black-red; tail consisting of twelve feathers, and even at the end; legs flesh-colour.

Inhabitants Egypt; the Arabic name is Chudjgle. The young are sold in the month of June.

C.—Johann Philippus, in: Z. 1. 120. 1. 120. 2. 120. Gen. Soc. 1. 120. 3. 120. 4. 120.

Length six inches and a half. Plumage the same with the others, excepting that the streak on each side of the head, and the band on the throat are blue; the two middle tail feathers do not exceed the others in length, but are narrower the whole way; and black. It is very probable, that the bird might have lost the two tail feathers, and been supplied with two new ones, not yet arrived at the full growth; in which case, this bird would be very little shorter than the first described, as the excess of length is that arose merely from the two middle feathers.

E.—Johann, Gen. Soc. 1. 120. 3. 120. 4. 120. 5. 120.

This variety has a yellow forehead, and a deep blue throat; the rest of the plumage as in the first described.

This is a common species in India, where they are often seen flying abroad; and are said to be fond of plants.
Television

About your Search

CANDIDATE
- Tom Corbett (con) 2
- Tom Wolf (con) 2
- John Kane (pro) 1
- Mike Fitzpatrick (pro) 1
- Tom Corbett (pro) 1
- Tom McGarrigle (con) 1

SPONSOR
- Tom Corbett for Governor 3
- Kane for Senate 2
- Philadelphia Federation of Teachers 2
- Fitzpatrick for Congress 1

Search Results 0 to 49 of about 389,622 (some duplicates have been removed)
First < 1 2 3 4 5 6 > Last

America Tonight
Al Jazeera America
Jan 30, 2014 4:00am EBT

The Rachel Maddow Show
MSNBC
Apr 21, 2015 1:00am PDT

The Journal Editorial Report
FOX News
Jan 2, 2015 11:00pm EBT

"President who revoked his passport president Obama" Wednesday on the hill the senate intelligence committee warned that he is a threat to this nation's
"President Obama did not miss more than half his"

"For the country, and for president Obama amid high unemployment, will macroeconomic recovery continue? and will Obama's declining approval ratings,
Edward Snowden has been nominated for the Nobel Peace Prize. He could share the honour as the president who revoked his passport President Obama. Wednesday on the Hill the Senate Intelligence Committee warned that he rose now a snapshot of stories making headlines - a bully on Capitol Hill, the cameras are rolling after New York reporter asked New York Congressman Michael Grimm’s Justice Department into his campaign’s finances. Things got ugly, Michael Grimm threatened to throw...
Internet Archive

- Archive-It
- Audio Books
- Bitcoin ATM
- Bittorrent seeds
- Bookmobile
- Book Reader
- Book Scanning
- CD archiving
- Community Wireless
- Credit Union
- DAISY / Print-Disabled
- Distributed Web
- Film preservation
- Foundation Housing
- Heritrix crawler
- IIPC
- JSMESS emulator
- Lending Library
- Listening Rooms
- Live Music Archive
- LP digitization
- Metadata API
- Microfilm scanning
- Microfiche
- Netlabels
- Newspaper digitization
- OPDS
- Open Content Alliance
- Open Library
- Petabox
- Physical Archive
- Reader Privacy
- Researcher Access
- S3-like API
- Scribe and TTScribe
- Software preservation
- Television Archive
- V2 site
- VHS digitization
- Wayback Machine
Questions?

- developers.archive.org
- we love volunteers!
- we are hiring engineers!
- archive.org/jobs