ARCHIVE

INTERNET

Internet Archive: Universal Access. Open APIs.

Linuxcon NA 2015

Alexis Rossi   VM Brasseur

Director of Media & Access
Internet Archive
alexis@archive.org
@alexisrossi
http://archive.org

linuxcon@vmbrasseur.com
@vmbrasseur
IRC: vmbrasseur
http://vmbrasseur.com

Who we are (very, very briefly). Get your phones ready for the next slide, since I'll be sharing some helpful links.

Alexis Rossi   Vicky Brasseur

Director of Media & Access
Internet Archive
alexis@archive.org
@alexisrossi
http://archive.org

linuxcon@vmbrasseur.com
@vmbrasseur
IRC: vmbrasseur
http://vmbrasseur.com

Who we are (very, very briefly). Get your phones ready for the next slide, since I'll be sharing some helpful links.

Rex Features

Some Handy Links

- https://github.com/internetarchive

- https://www.zotero.org/groups/internet_archive_-
  _open_apis_and_examples/items

- https://archive.org/details/linuxconna2015-ia-apis

---

*   The Internet Archive github organization

*   The Zotero group where you can find all of the links and references we'll be mentioning

Don't worry, we'll be showing these links again at the end of the presentation.

A Tour of Internet Archive

Internet Archive

- 501(c)(3) non-profit, founded in 1996
- Registered library in California, USA
- Partnerships with hundreds of libraries, museums & universities
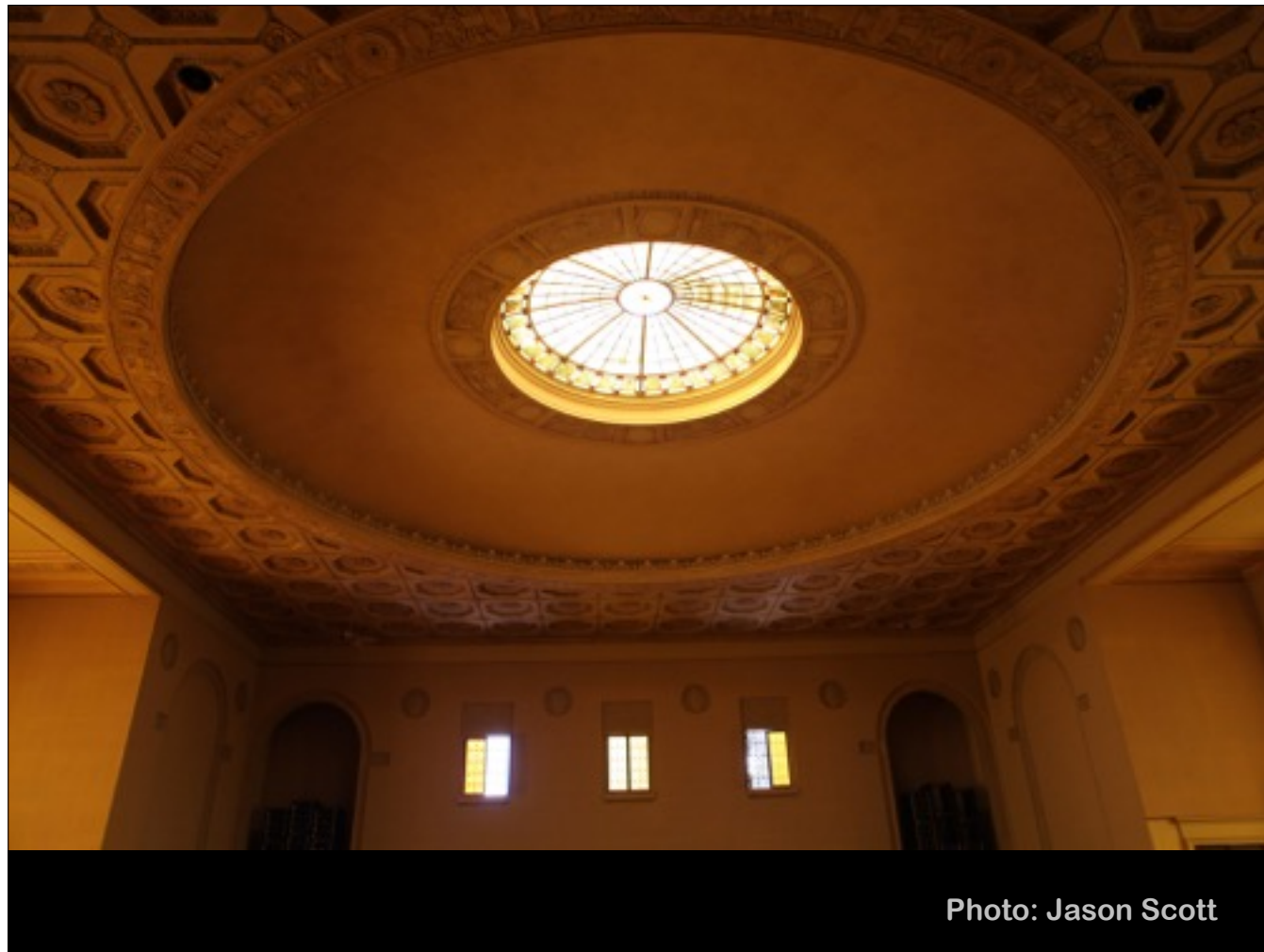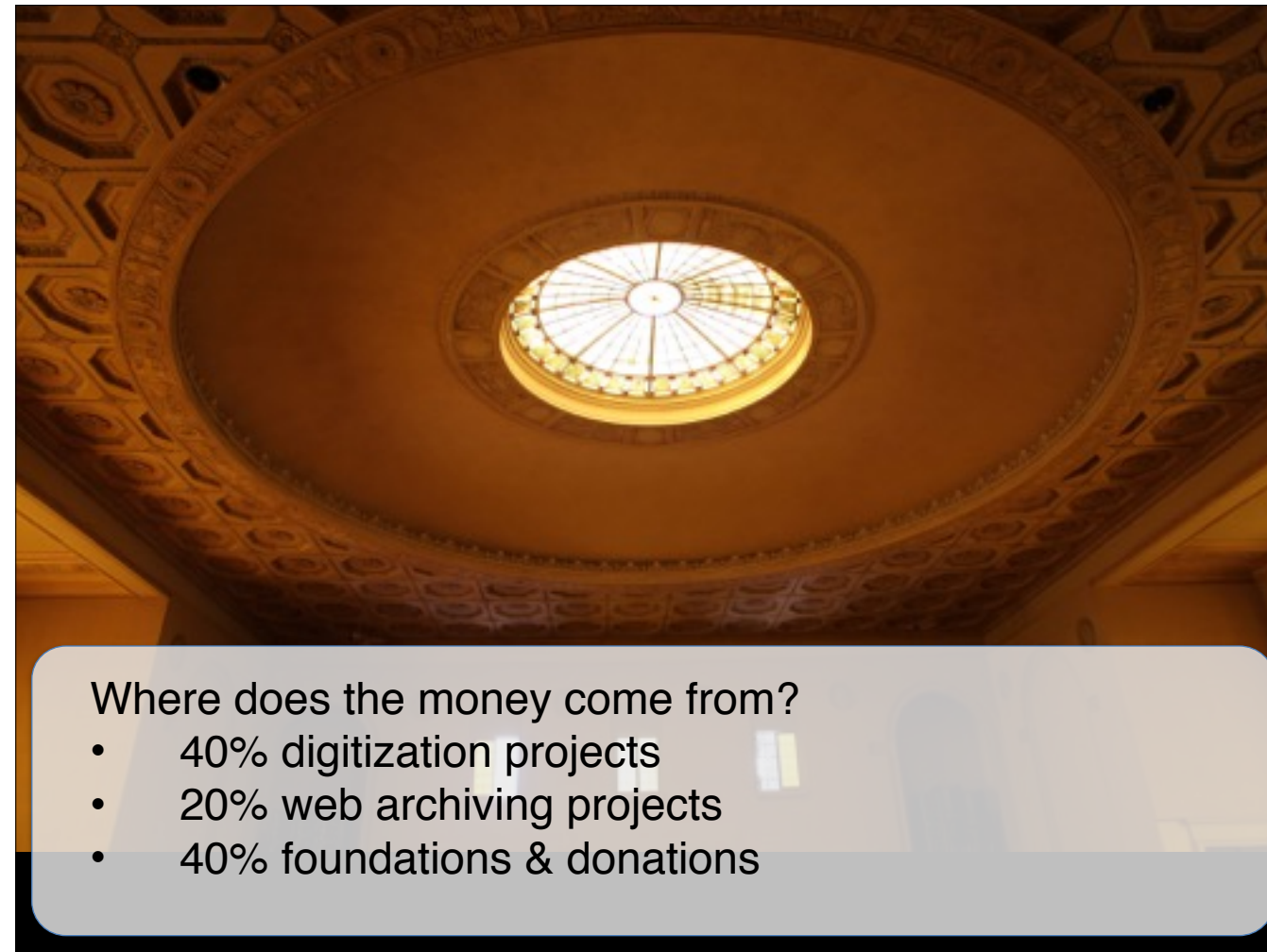- 24 petabytes of unique data

Photo: Jason Scott

Where does the money come from?
- 40% digitization projects
- 20% web archiving projects
- 40% foundations & donations

$12 million per year budget

Photo: Jason Scott

Who uses archive.org sites?
- In top 250 websites
- 2-3 million people per day
- 4 million downloads/streams per day
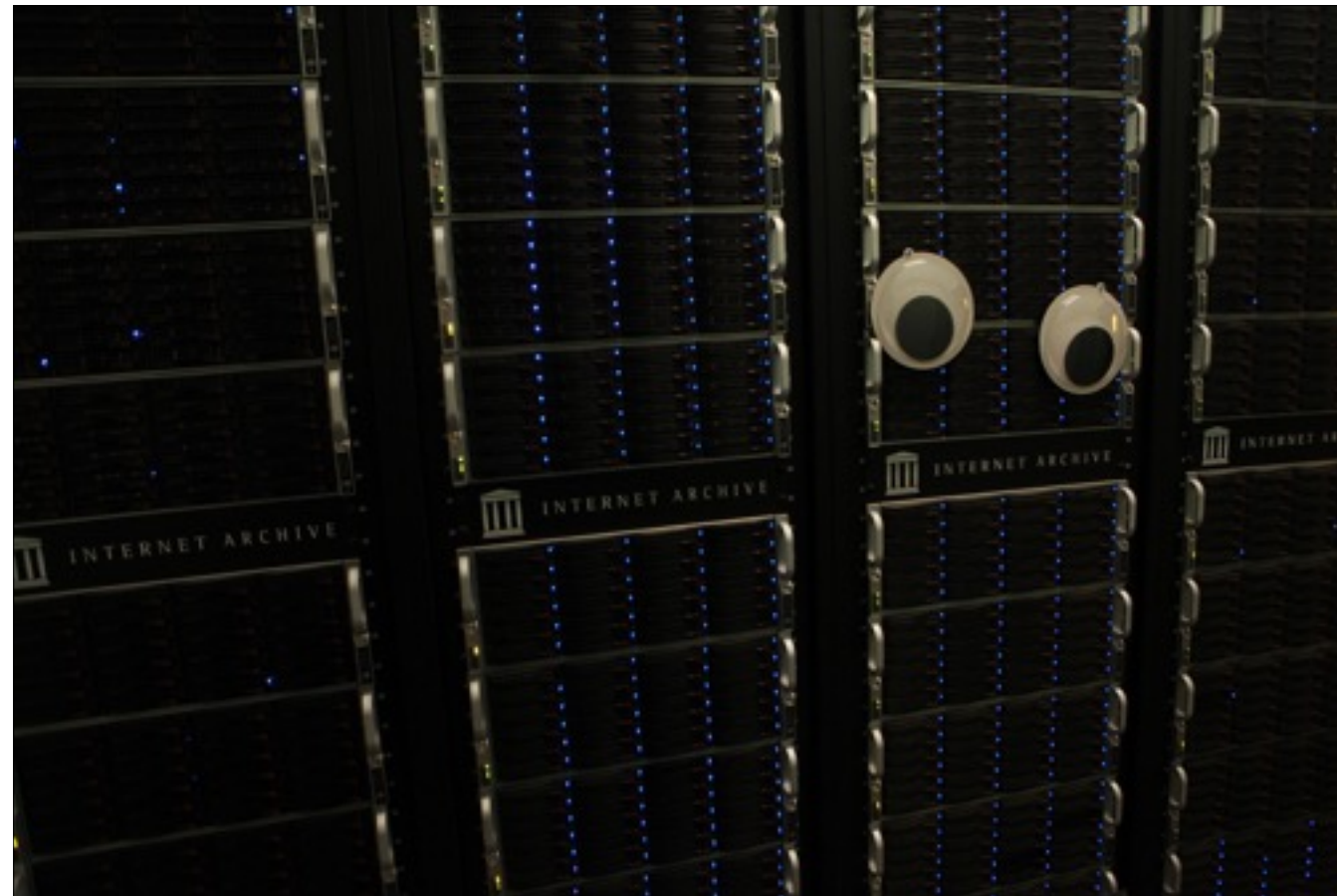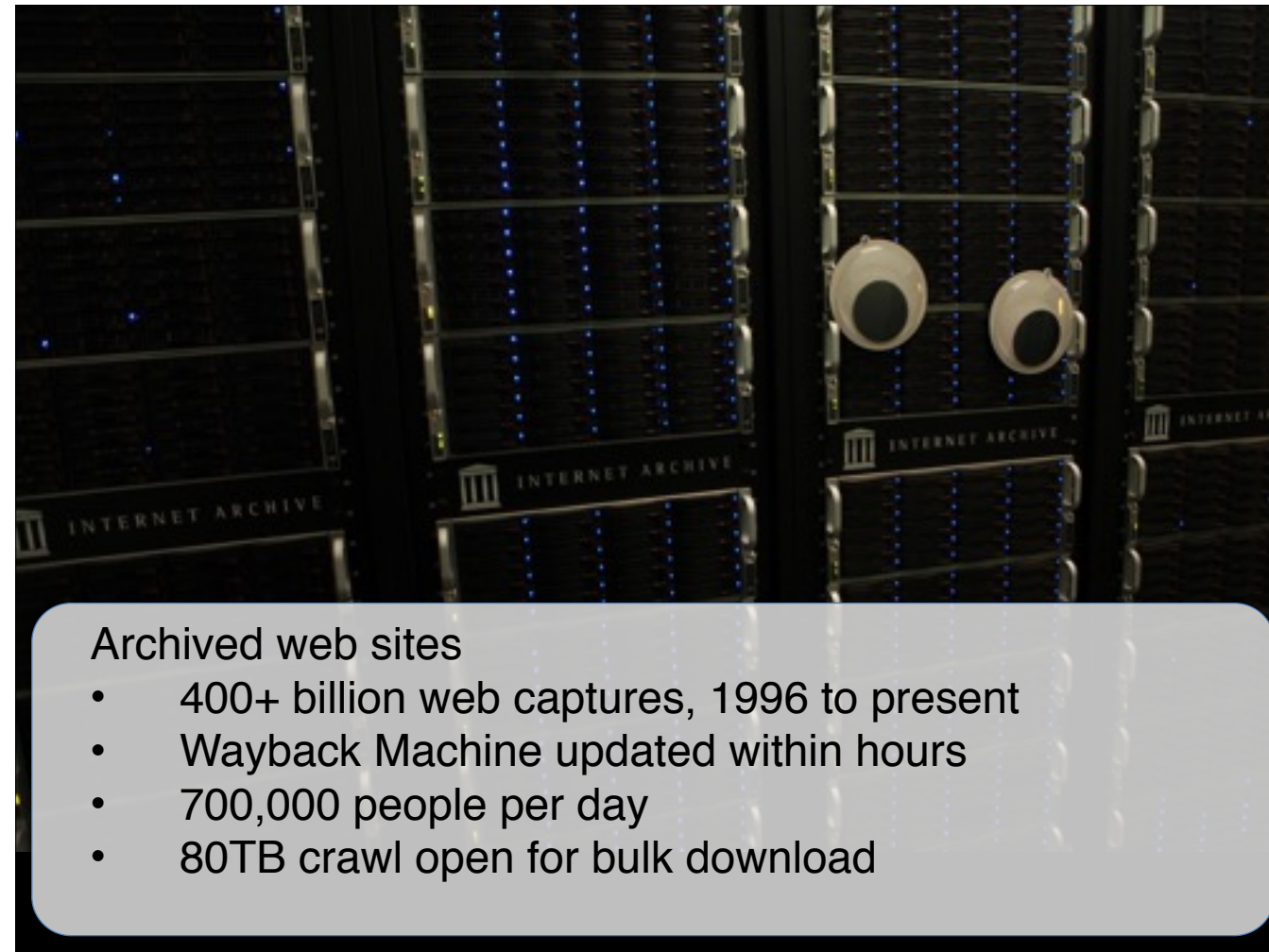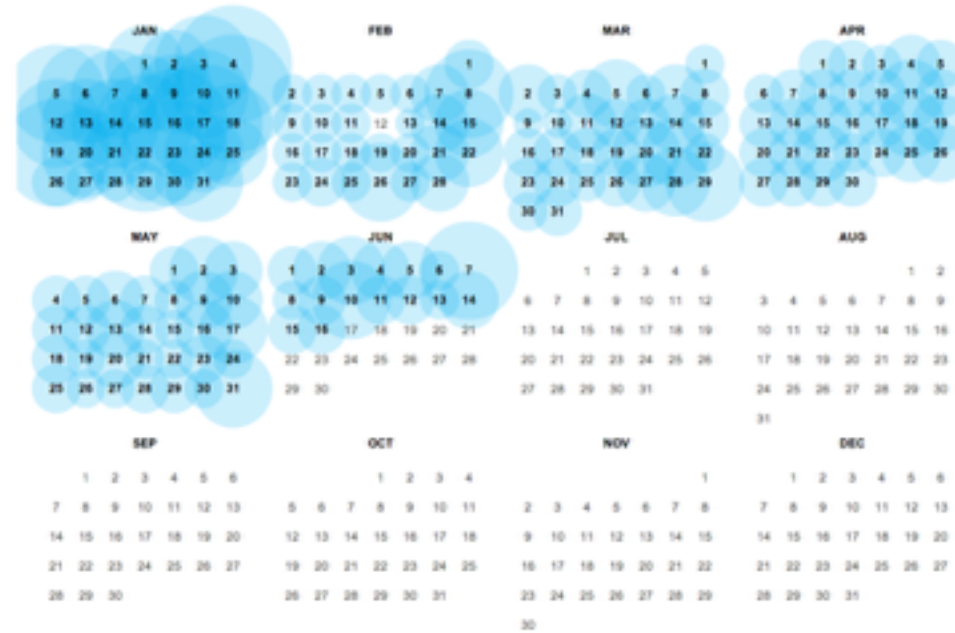- No IP addresses kept

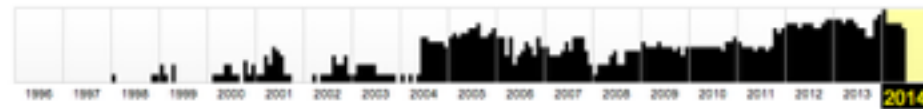Photo: Jason Scott

Photo: Jason Scott

Archived web sites
- 400+ billion web captures, 1996 to present
- Wayback Machine updated within hours
- 700,000 people per day
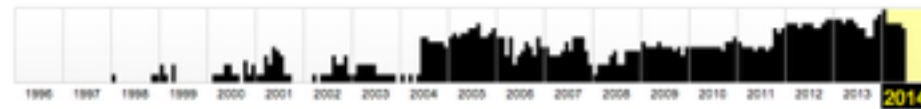- 80TB crawl open for bulk download

1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 **2014**

JAN

		1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31

FEB

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28

MAR

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31

APR

	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30

MAY

		1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

JUN

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30

JUL

	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31

AUG

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31

SEP

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30

OCT

		1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31

NOV

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30

DEC

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31

How do we decide what to archive?
- People pay us
- Organizations donate crawled content
- We crawl on our own behalf
  - Deep crawl on popular sites
  - Broad but shallow crawl on known domains
  - Targeted crawls
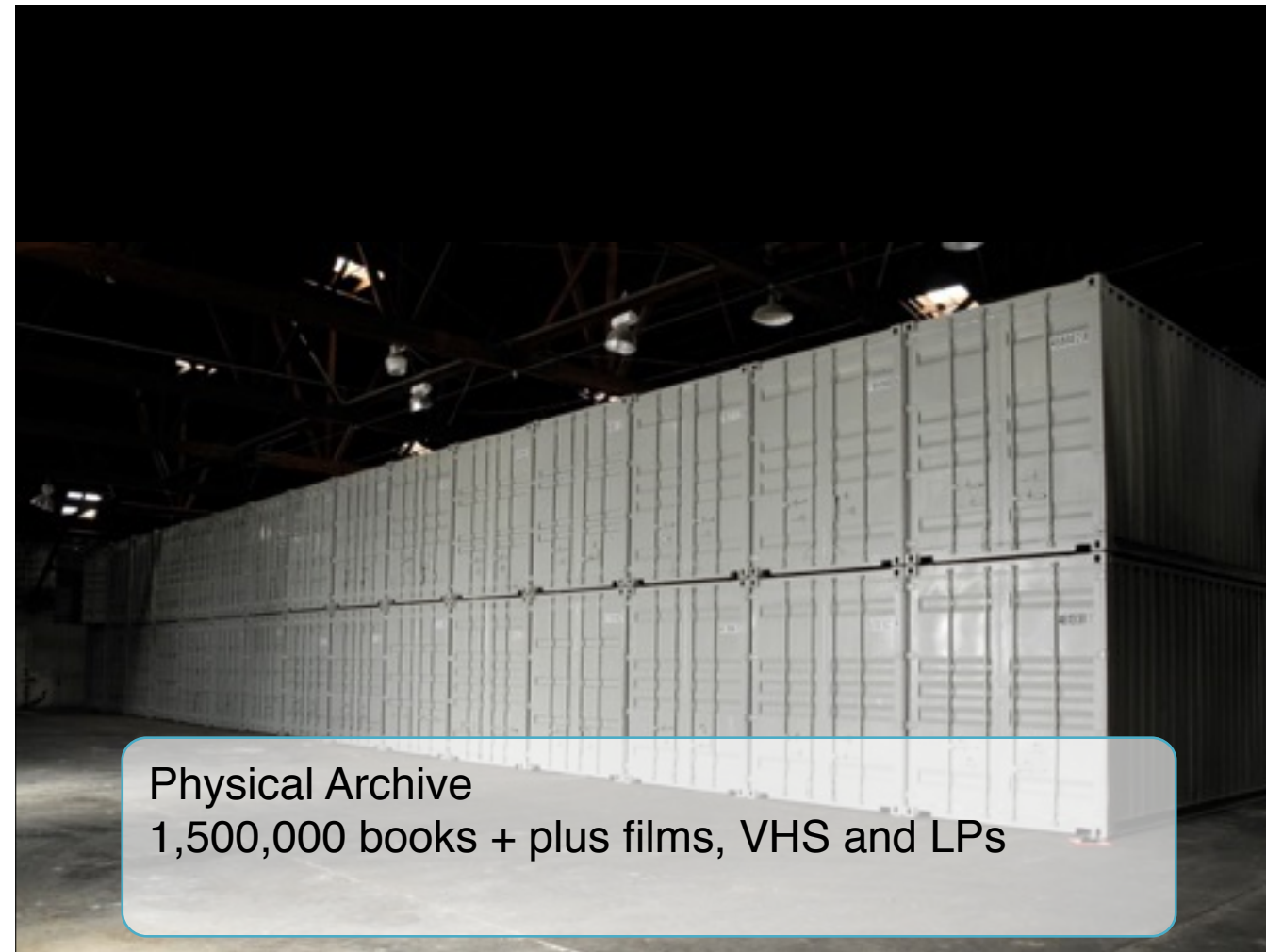
- 30 scanning centers in 8 countries
- 1,000 books per day

November 2013

Physical Archive
1,500,000 books + plus films, VHS and LPs
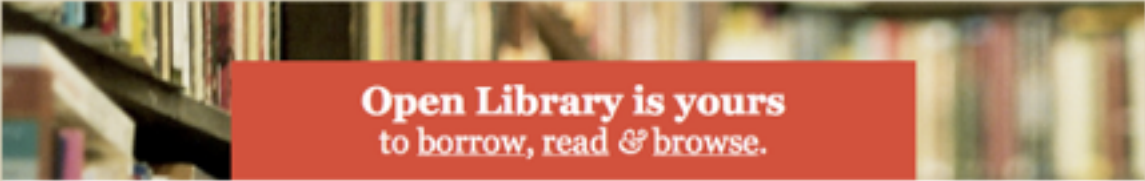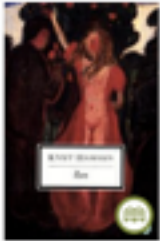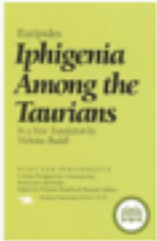
Borrow books at openlibrary.org

Video
- 2,000,000 items
- Feature films, documentaries, commercials, propaganda, stock footage, cartoons
- Archive 60 channels of TV 24 hours per day

U.S. TV News Archive: archive.org/tv

550,000 programs

Audio
- 2,500,000 items
- 150,000 live concerts from 6,000 bands
- 60,000 netlabel releases from 2,000 labels
- Mirrors of Jamendo, IUMA
- Audio books, radio shows, podcasts

IUMA: Internet Underground Music Archive

We are working on building a large collection of commercial music

Atari 2600: Frogger

Software
- 100,000 titles
- Classic PC games, MSDOS games, Arcade, Consoles
- Emulation in browser

Talk to Jason Scott, jscott@archive.org

Software: 10k cdroms, 19k titles at stanford. need clarification on DMCA

Why?

- We are good at storing and serving digital media and preserving it
- We care about the same things: knowledge, keeping information open, privacy
- We fight for what we care about
- We're not slick, but we are friendly!

Internet Archive APIs
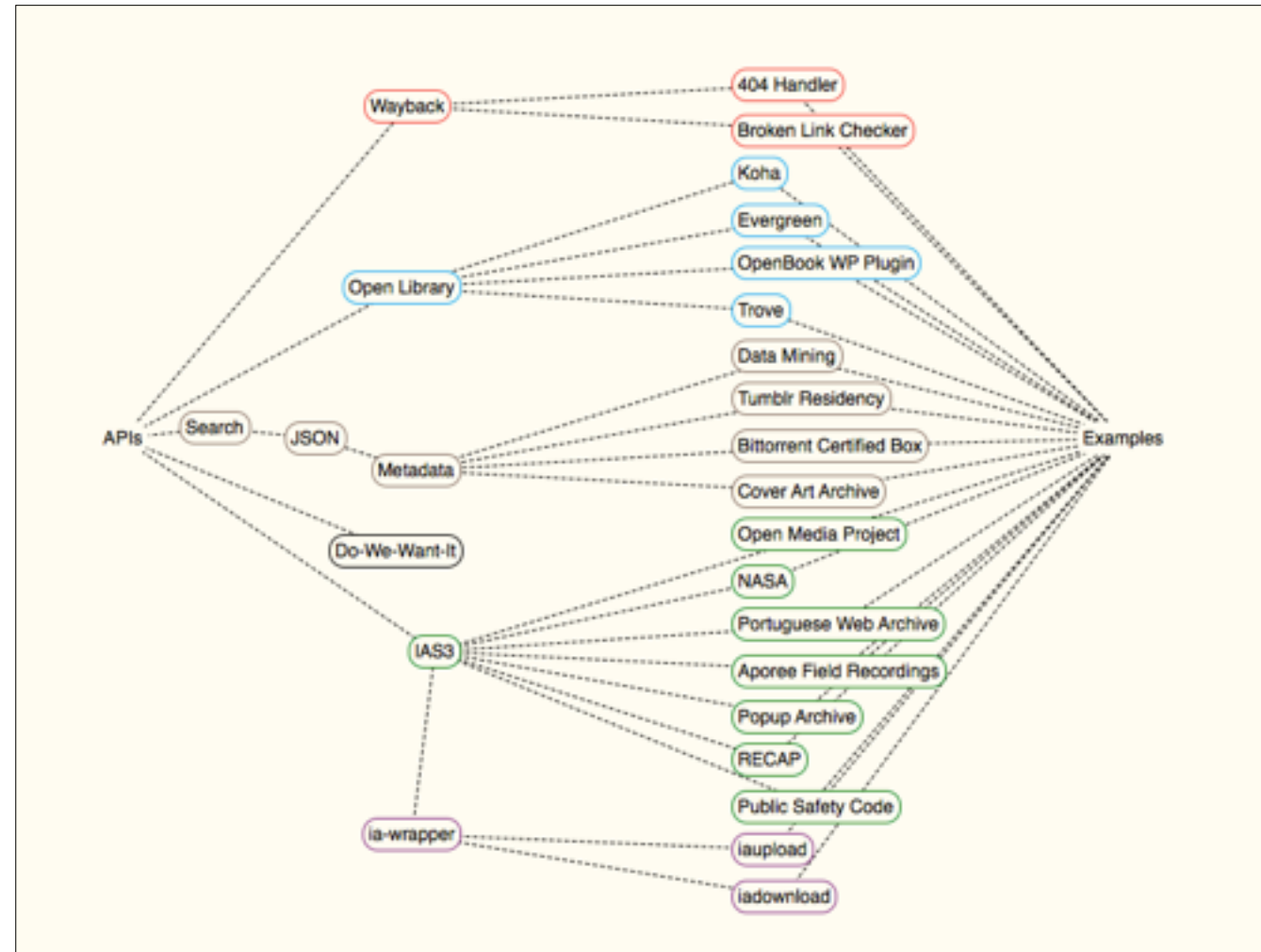
Let's set up some expectations. Internet Archive provides many different ways to access and contribute to those 18+ Petabytes of data. There's a lot to cover, so I'll only be giving you a **brief overview** of each rather than a **deep dive**. Complete information can be found at the Zotero link.

Wayback Machine API

`http://archive.org/help/wayback_api.php`

First up: The Wayback Machine API.

Wayback Machine API

- Is a URL archived?

- If so, is it available in the Wayback Machine?

This API is a study in simplicity & ease: exactly what you need w/o clutter.

Wayback Machine API

```
http://archive.org/wayback/available?
PARAMETERS
```

Simple URL-based API

Only three possible parameters (url, timestamp, callback) and two are optional

Returns JSON

Wayback Machine API

```
Parameters:

url=sub.domain.tld
timestamp=20030831060429 (YYMMDDhhmmss; optional)
callback=function_name (optional)
```

Parameters for the Wayback API

* url = what you're looking up (no protocol; http, etc.)
* timestamp = YYYYMMDDhhmmss; YYMMDD OK; Will return closest snapshot to this timestamp
* callback = for jsonp

We've had a bit of excitement recently about New Horizons & its fly-by of Pluto, so let's use that as the basis for our examples today.

Wayback Machine API

```
curl http://archive.org/wayback/
available?url=solarsystem.nasa.gov/
planets/profile.cfm?
Object=Dwarf&Display=Sats
```

Let's see whether the Wayback has their site on record yet. Just throw this to curl and see what it returns…

Wayback Machine API

```
{
  archived_snapshots: {
    closest: {
      available: true,
      url: "http://web.archive.org/web/
20150711221849/https://
solarsystem.nasa.gov/planets/profile.cfm?
Object=Dwarf",
      timestamp: "20150711221849",
      status: "200"
    }
  }
}
```

That's more like it.

Wayback Machine API

`{"archived_snapshots":{}}`

This is what you see when there is no entry for the URL in the Wayback. This API is in use in some pretty interesting places. One example is:

404 Handler

Free "404: File Not Found" Handler for Webmasters to Improve User Experience

Posted on October 24, 2013 by internetarchive

The Internet Archive today is launching a free service to help webmasters improve their user experience by augmenting their website's 404 Page Not Found page to link to the Wayback Machine in the case that it has it. Therefore users trying to get to any pages that might have been on a previous version of your website will now be given the option to go to the Wayback Machine.

To embed a link to the Wayback Machine on your site's 404 pages, just include this line in your error page:

```
<div id="wb404"/>
<script src="https://archive.org/web/wb404.js"> </script>
```

If an archived page is not found, then nothing will appear, if it is found, then your user will see:

INTERNET ARCHIVE Would you like to see an archived version of this page in the Internet Archive's Wayback Machine?

https://blog.archive.org/2013/10/24/web-archive-404-handler-for-webmasters/

The 404 Handler. Placed on your 404 page, if an archived version of the page is in the Wayback then it'll offer a link to it.

Open Library API

`https://openlibrary.org/developers/api`

Open web page for every book ever published. Think of it rather like Wikipedia for books. The library you can edit.

A full RESTful API.

Can return Open Library bibliographic & holdings records in JSON and RDF formats

Open Library API

- Query the Open Library database

- View record information

- Edit record information

- View record history

Features: most everything you can do on the Open Library website: Query the DB, look at information for books & authors, edit records and view edit history…

Open Library API

```
http://openlibrary.org/subjects/pluto.json
?limit=1
```

Let's see what Open Library has for the subject "pluto". I'll limit it to just one entry so it'll fit on the next slide.

Open Library API

```
{
  works: [
  {
    printdisabled: false,
    cover_id: null,
    ia_collection: [ ],
    has_fulltext: false,
    edition_count: 2,
    checked_out: false,
    title: "The planets",
    public_scan: false,
    lendinglibrary: false,
    lending_edition: "",
    overdrive: "",
    first_publish_year: null,
    key: "/works/OL2715443W",
authors: [
    {
       name: "Heather Couper",
       key: "/authors/
OL397296A"
    },

    {
       name: "Nigel Henbest",
       key: "/authors/
OL449922A"
    }
    ],
    ia: null,
    lending_identifier: "",
    subject: [
       "Solar system",
       "Juvenile literature",
       "Planets"
    ]
  }
  ],
  subject_type: "subject",
  work_count: 13,
  key: "/subjects/planets",
  name: "planets"
}
```

So I see there's "The Planets" by Heather Couper & Nigel Henbest.

Evergreen

http://evergreen-ils.org/

An example of this API in the wild: The Evergreen open source Integrated Library System uses the API to retrieve book covers and other information from OL.

Do-We-Want-It? API

`http://want.archive.org/`

So, how do all of those books get into Open Library? Well, a lot of them are donated by people like you. The Archive will accept your spare book, scan it & make it available online, and then save the book itself in its physical archive. But space is limited so they've provided a simple API to help you see whether they have a copy of that book yet.

Do-We-Want-It? API

```
http://want.archive.org/api?isbn=<isbn10 or isbn13>
```

This API works with anything which has an ISBN. It takes one argument and returns very easy to understand JSON…

Do-We-Want-It? API
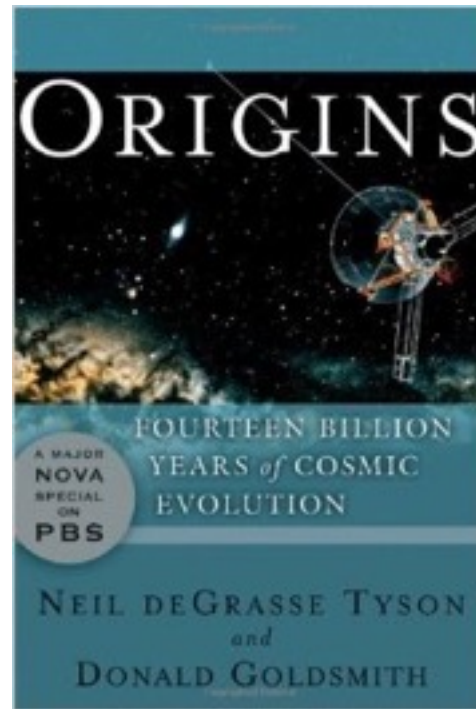

```
http://
want.archive.org/api?
isbn=978-0393350395
```

So let's see whether the Archive needs any copies of this book by Neil DeGrasse Tyson & Donald Goldsmith…

Do-We-Want-It? API

```
{
   status: "success",
   result: "1",
   description: "want_for_ia_pa",
   identifier: "-1"
}
```

Yup! We can see here that the Archive would like to have a copy of this book. It doesn't yet have it available.

Do-We-Want-It? API

```
Response keys:
  status
    fail    - We failed to process the request. The submitted ISBN was
              invalid.
    success - The request was serviced successfully.
  result
    -1 -  We failed to process the request. The submitted ISBN was
          invalid.
     0 -  The request was serviced successfully, but we have two copies,
          and do not want more.
     1 -  The request was serviced successfully, and we have no copies.
          We want it.
     2 -  The request was serviced successfully, and we have 1 copy
          already. We want it.

  description - "human-parsable" description, with respect to the above:
    failure
    do_not_want
    want_for_ia_pa
    want_for_alt_pa

  identifier - String when result = 2. This is the identifier already
  assigned on the cluster for that ISBN
    -1 - No identifier designated for this ISBN.
    <string> - Designated identifier for this ISBN.
```

The results are pretty easy to read, but here you can see all of the possibilities for data returned.

IA Search API

`https://archive.org/advancedsearch.php`

As Alexis has pointed out, the Archive has ALL THIS GREAT STUFF. But how do you find it? This isn't a documented API so much as an easily extrapolatable URL format.

IA Search API

Advanced Search

This form allows you to perform an advanced search. You only need to fill in one field below. This can be any field. If you select "not" as your match criteria, you must select one other field.

	Any field:	contains	
AND	Title:	contains	
AND	Creator:	contains	
AND	Description:	contains	
AND	Collection:	is	All collections
AND	Mediatype:	is	texts
AND	subject	contains	pluto
AND	Custom field	contains	
AND	Custom field	contains	
AND	Date:	YYYY MM DD	
AND	Date range:	YYYY MM DD TO YYYY MM DD	

Search

So let's continue my quest to learn more about planets, performing a search for all texts with a subject value of 'pluto.'

IA Search API

The search terms:
mediatype:(texts) AND subject:(pluto)

```
https://
archive.org/
search.php?
query=mediatype%3A
%28texts%29%20AND
%20subject%3A
%28pluto%29
```

It returns this URL. It has some HTML entity encoding going on, but otherwise makes it pretty obvious how to build a URL. OK, that's pretty cool, but…

IA JSON API

Advanced Search returning JSON, XML, and more

This will return results in the format of your choice.

Query:

mediatype:(texts) AND subject:(pluto) AND collection:(nasa)

Fields to return (pick one or more):
avg_rating
call_number
collection
contributor
coverage
creator
date
description
downloads
foldoutcount
format
headerimage
identifier
imagecount
language
licenseurl
mediatype
members
month
num_reviews
oai_updatedate
publicdate
publisher
reviewdate
rights
scanningcentre
source

(optional) Sort results by:

Number of results: 50

Page: 1

Indent response:

JSON format:
XML format:
save to file:
HTML table:
CSV format:
(show/hide help)
RSS format:
('Fields to return' ignored)

```
https://archive.org/
advancedsearch.php?
q=mediatype%3A
%28texts%29+AND
+subject%3A%28pluto
%29+AND+collection%3A
%28nasa%29&fl%5B
%5D=identifier&sort
%5B%5D=&sort%5B
%5D=&sort%5B
%5D=&rows=50&page=1&o
utput=json&callback=c
allback
```

…the fact that I can ask for the exact same data to be returned as JSON (or CSV, or XML, or…). As you can see, this URL is a bit busier, but comparing it against the form it's quite easy to see what's going on.

IA JSON API

```
callback(
  {
    responseHeader: {
      status: 0,
      QTime: 108,
      params: {
        json.wrf: "callback",
        wt: "json",
        rows: "50",
        qin: "mediatype:(texts) AND subject:(pluto)",
        fl: "identifier",
        start: "0",
        q: "mediatype:(texts ) AND subject:(pluto )"
      }
    },
    response: {
      numFound: 32,
      start: 0,
      docs: [
        ...
        {
          identifier: "nasa_techdoc_20050060913"
        },
        ...
      ]
    }
  }
)
```

But it's even easier to see when you view the JSON output. Note the identifier.

IA JSON API

- Lucene-based

  - Grouping

  - Fuzzy queries

  - Relevance boost

  - Date ranges

  - Etc.

You can do some pretty sweet things here. But wait, there's more!

IA Metadata API

`http://blog.archive.org/2013/07/04/metadata-api/`

The Internet Archive Metadata API allows you to DATA MINE that entire 18+ Petabytes of data. And it's WAY faster than it has any right to be.

IA Metadata API

`http://archive.org/metadata/`
`nasa_techdoc_20050060913/metadata`

I want to learn more about that item I retrieved with the JSON API, so let's call the Metadata API on its identifier. To make the response slightly shorter, I'm going to limit it to just the most metadata-est part of the information, rather than ALL of it.

IA Metadata API

```
{
  result: {
    identifier: "nasa_techdoc_20050060913",
    date: "2004",
    description: "Terra MODIS 250 m observations are being applied to a Suspended Sediment Concentrat
algorithm that is under development for coastal case 2 waters where reflectance is dominated by sedim
entrained in major fluvial outflows. An atmospheric correction based on MODIS observations in the 500
resolution 1.6 and 2.1 micron bands is used to isolate the remote sensing reflectance in the MODIS 25
resolution 650 and 865 nanometer bands. SSC estimates from remote sensing reflectance are based on ac
inherent optical properties of sediment types known to be prevalent in the U.S. Gulf of Mexico coasta
present our findings for the Atchafalaya Bay region of the Louisiana Coast, in the form of processed
over the annual cycle. We also apply our algorithm to selected sites worldwide with a goal of extendi
utility of our approach to the global direct broadcast community.",
    document-source: "CASI",
    documentid: "20050060913",
    nasa-center: "Goddard Space Flight Center",
    online-source: "http://wayback.archive-it.org/1792/20100127084754/http://hdl.handle.net/2060/2005
    original-nasa-rights: "Unclassified; No Copyright; Unlimited; Publicly available; Progress Report
    title: "Estimating Coastal Turbidity using MODIS 250 m Band Observations",
    updated-added-to-ntrs: "2008-06-02",
    year: "2004",
    collection: "nasa_techdocs",
    contributor: "NASA",
    language: "eng",
    licenseurl: "http://creativecommons.org/licenses/publicdomain/",
    mediatype: "texts",
    rights: "Public Domain",
…
```

And we get more JSON. There's a LOT of data here and I've truncated the output here.

IA Metadata API

It can write metadata, too!

If I had the authorizations to change this item's metadata. Which I don't. But if you do, then you can if you want. w00t.

But Wait!
There's more!

So right now you're probably thinking, "Sure, all this mining of 18+PB of data is neat and all, but how do I add to it?"

IAS3 API

`https://github.com/vmbrasseur/IAS3API`

This is the big daddy: The Internet Archive S3-like API. And now you know why I'm up here speaking to you today: I'm the maintainer of the documentation for this API, which you can find at this GitHub URL.

Reminder: You can upload ANYTHING to IA. For free. As much as you want. They'll serve it up & preserve it forever. For free.

IAS3 API

Create items in Internet Archive, upload files to those items, maintain the metadata for the items, and download from any publicly-available Internet Archive items.

Doesn't work for Open Library, but otherwise? Much of that stuff I just showed you? Can be handled with this one Big Daddy of an API.

IAS3 API

- Drop-in replacement for the Amazon S3 API

- Pick your favorite S3 library/client, change the server to s3.us.archive.org, and you're good to go.

This is a pretty involved API, so I'll only provide one brief and simple example. There are more in the documentation.

IAS3 API

```
curl --location --header 'x-amz-auto-make-bucket:1' \
     --header 'x-archive-meta01-collection:nasa_techdocs' \
     --header 'x-archive-meta-mediatype:movie' \
     --header \
     'x-archive-meta-title:Pluto Fly By' \
     --header "authorization: LOW $accesskey:$secret" \
     --upload-file new-horizon.mp4 \
http://s3.us.archive.org/pluto-new-horizon/new-horizon.mp4
```

Create a new item (aka bucket) on Internet Archive with the identifer pluto-new-horizon, Assign the item to the 'data' mediatype, then upload the file new-horizon.mp4 to the item

Can also download, change metadata, etc. A lot of people & organizations use this API, so I'll only highlight a very few of them.
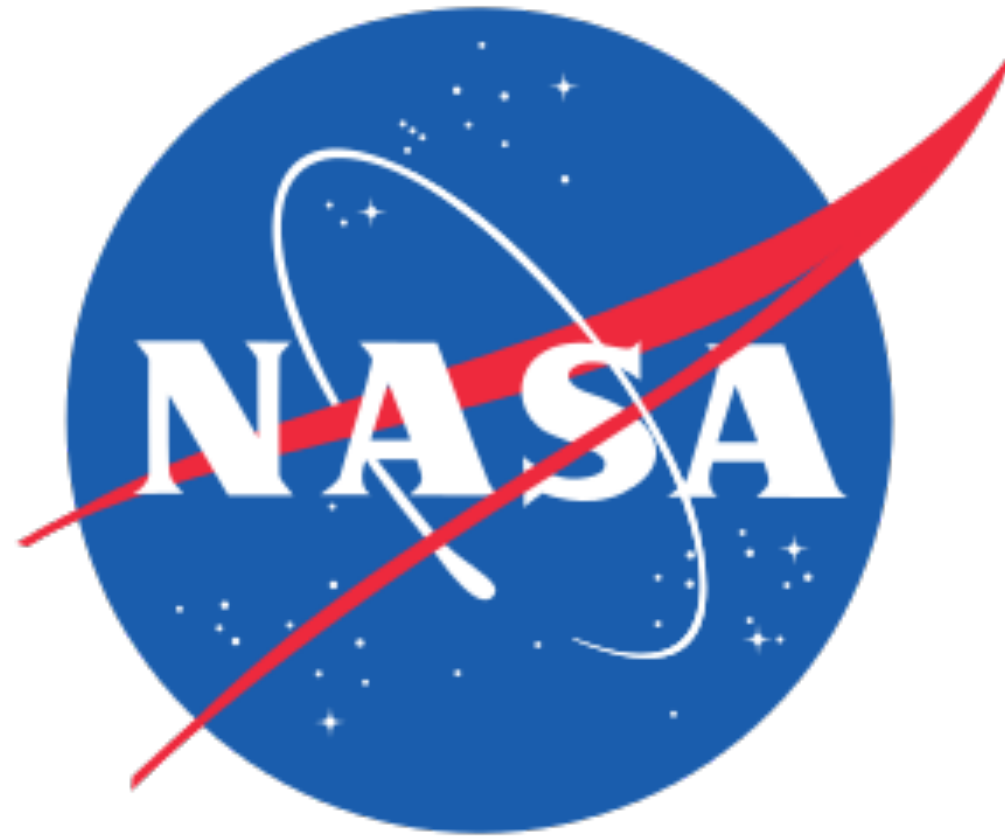
RECAP & Global Public Safety Codes

https://archive.org/details/usfederalcourts

https://archive.org/details/publicsafetycode

RECAP from Aaron Swartz & Global Public Safety Codes from Carl Malamud to free otherwise locked up public information.

https://archive.org/details/nasa

But, IMO, the most exciting use of IAS3API is by NASA.

Life is short.
What if I don't want to learn S3?

Sure, you think that's cool & all, but your time is valuable and you really don't want to spend it learning S3? OK, we can work with that.

ia-wrapper

https://github.com/jjjake/ia-wrapper

We have Jake Johnson at the Archive to thank for this little wonder. This is a Python wrapper around IAS3 and the metadata and search APIs. It includes utilities for everything you want to do, without all the mess of wrangling S3 API headers. As if that weren't good enough…

iadownload

```
#! /usr/bin/env python

# iadownload: Download all files in a collection or item
# Copyright 2014 VM Brasseur

import os
import sys
import internetarchive          ⟵
import pprint
import argparse
import json
```

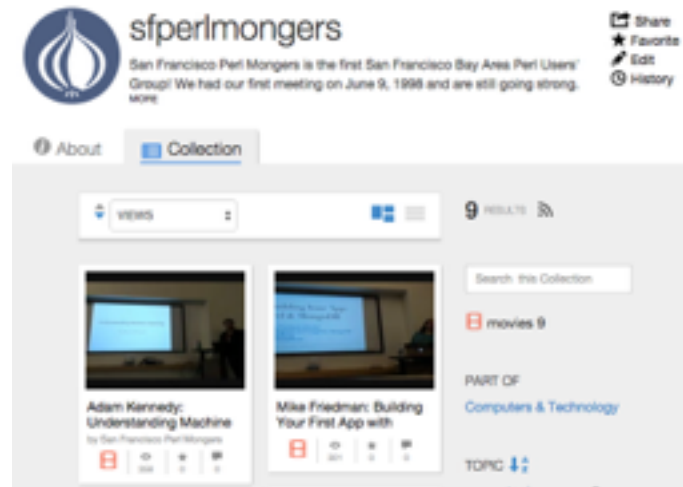https://github.com/vmbrasseur/iadownload

It also includes a Python library, so you can build your own utilities and services. As you can see, I use this myself. Not only do I use it for downloading…

iaupload

https://github.com/vmbrasseur/iaupload   http://archive.org/details/sfperlmongers

I also use it to upload. I'm the organizer of the San Francisco Perl Mongers user group. We now record all of our events and upload the videos to the Archive for all to see. But we don't really have a ton of material. To really put ia-wrapper through its paces we need to look to…

Saving All The Things

Source: Mirka23 on Flickr

Jason Scott. Internet Archive employee, Founder of Archive Team, and activist computer archivist. As of the writing of this talk, Jason has uploaded just shy of 300K items to the Archive. Most of the items contain several files. Jason uses and swears by ia-wrapper to help him archive as much of computer history as inhumanly possible.
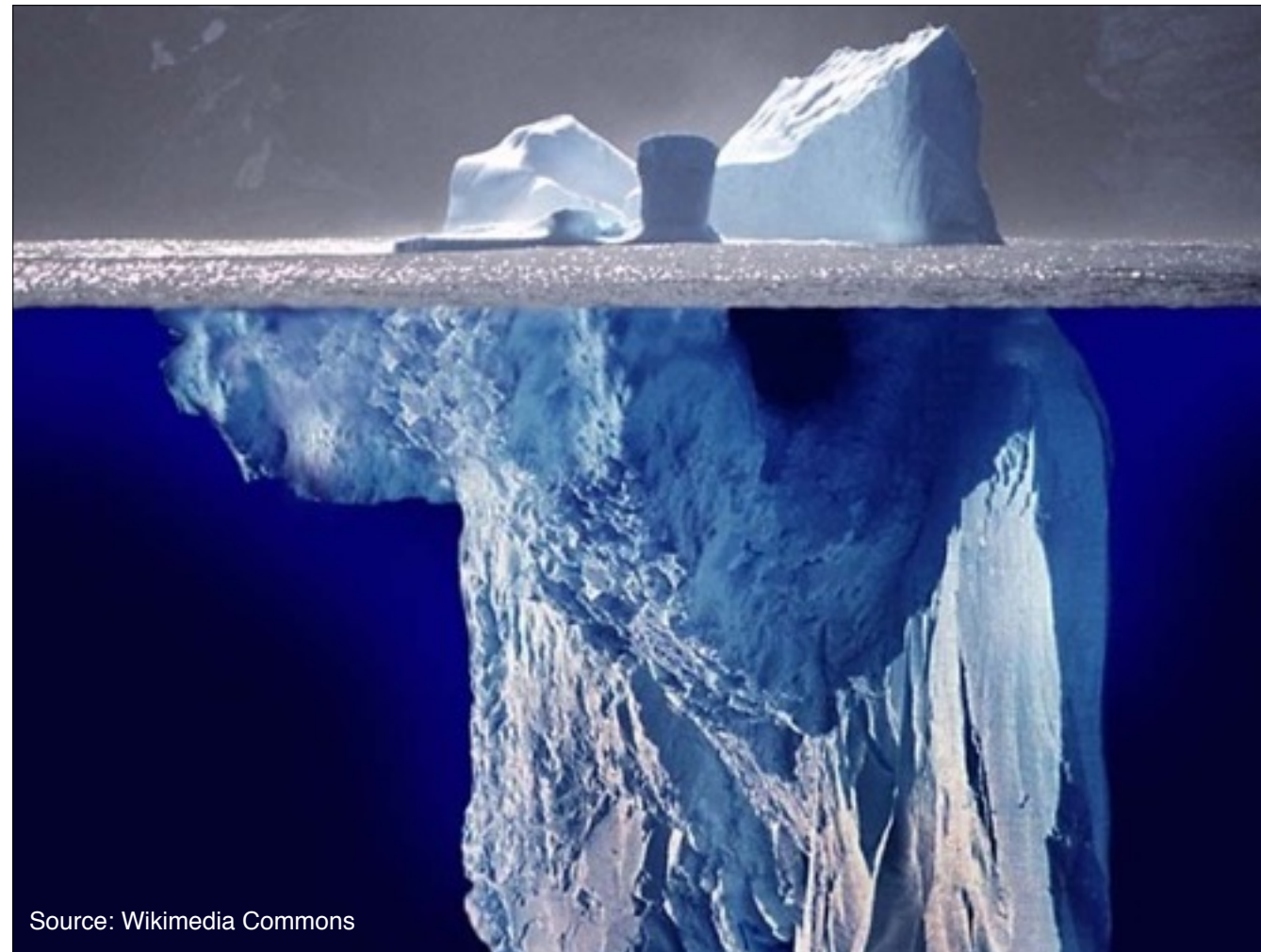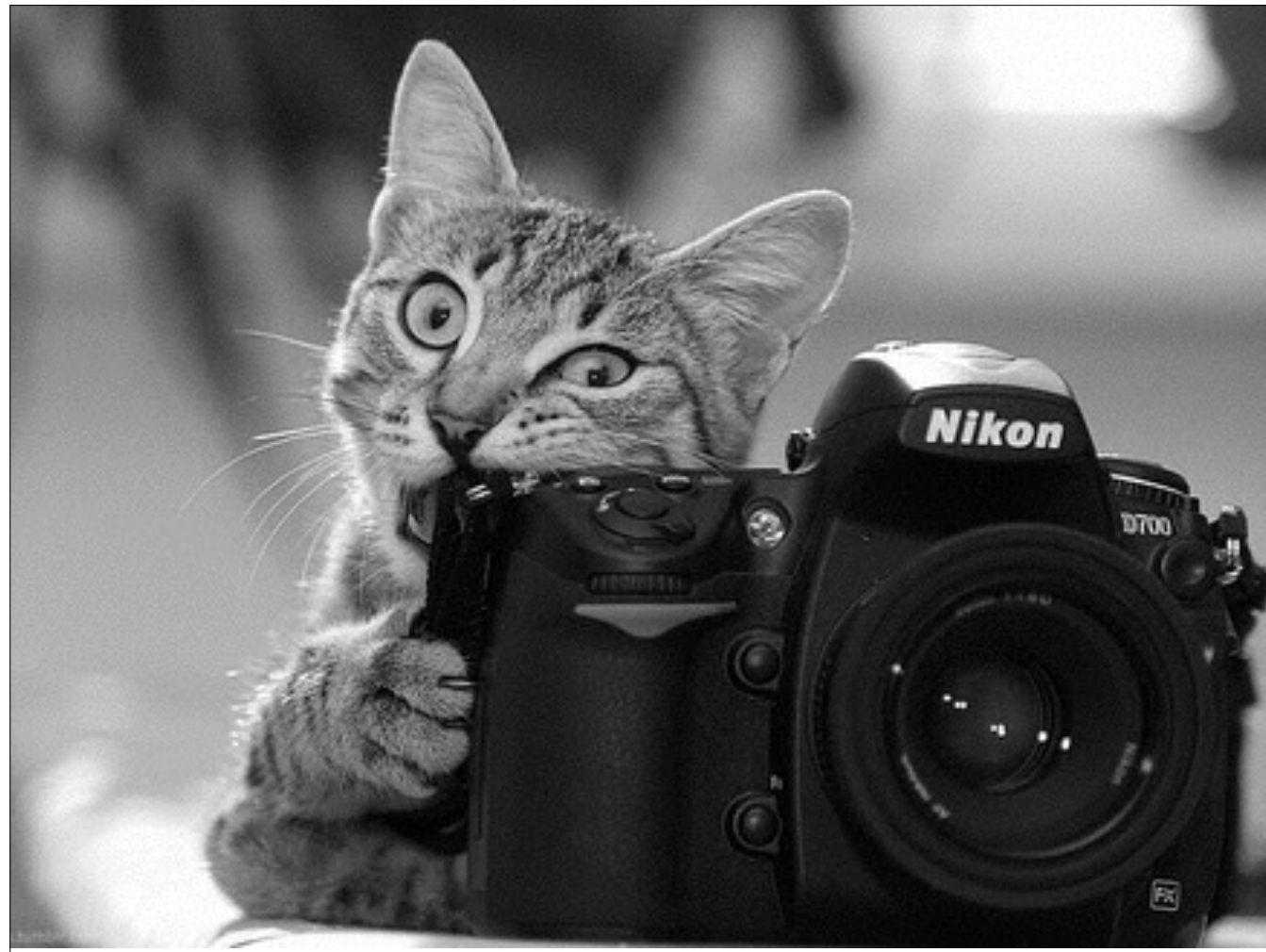
Source: Wikimedia Commons

As you can imagine. there's a WHOLE lot more which could be said on this topic. But now you at least know where to start looking for more information. Let's recap:

Recap

- Wayback API

- Open Library APIs

- Do-We-Want-It? API

- Search/JSON APIs

- Metadata API

- IAS3 API

- ia-wrapper

And, don't forget, if you want to learn more about any of these things…

Those Links Again…

- https://github.com/internetarchive

- https://www.zotero.org/groups/internet_archive_-
  _open_apis_and_examples/items

- https://archive.org/details/linuxconna2015-ia-apis

As promised, here are those links again. Snap a picture or find the slides at the IA item (last link). We have one more important link to share with you…

Donate to Internet Archive!

`http://archive.org/donate/`

Your support helps us build amazing services and keep them free for people around the globe.

THANK YOU!

Internet Archive is HIRING! (SF or Remote)

- *Senior Wayback Machine Engineer*
- *Senior Dev Ops Engineer*
- *Senior Cluster Storage & Computing Engineer*
- *Senior Python Engineer*
- *Web Application Developer*

We need programmers to help us change (and preserve)
the world!

VISIT US
Free lunch Fridays at noon
300 Funston Ave
San Francisco

alexis@archive.org

Questions?