DRASTIC Measures

Designing Scalable Cyberinfrastructure for Metadata Extraction in Billion-Object Archives

Gregory Jansen
Richard Marciano

Coalition for Networked Information
Washington, DC
Dec 13, 2016

UNIVERSITY OF MARYLAND

dcic digital curation innovation center

NCSA

The Next Twenty Minutes..

- Approaching 1 Billion files
- New DRAS-TIC Repository
- NCSA’s Brown Dog Service
- Automatic Feature Extraction & Curation
- Digging into Collections with Elasticsearch
- Projects & Opportunities

We are accumulating “Format Debt”

- Discovery, access, and reuse are limited by format
- Collections of **unstructured and un-curated** digital data
  - No plain text
  - No useful file or folder names
  - Minimal metadata
- Many media types and hundreds of file formats
- Depending on legacy software for access
- Investment required to unlock formats
- Pay now, pay later, or... your users must pay

Approaching Billions at 1/10 Scale

100 Million files
72 Terabytes of data
Hundreds of file formats
Unique file formats

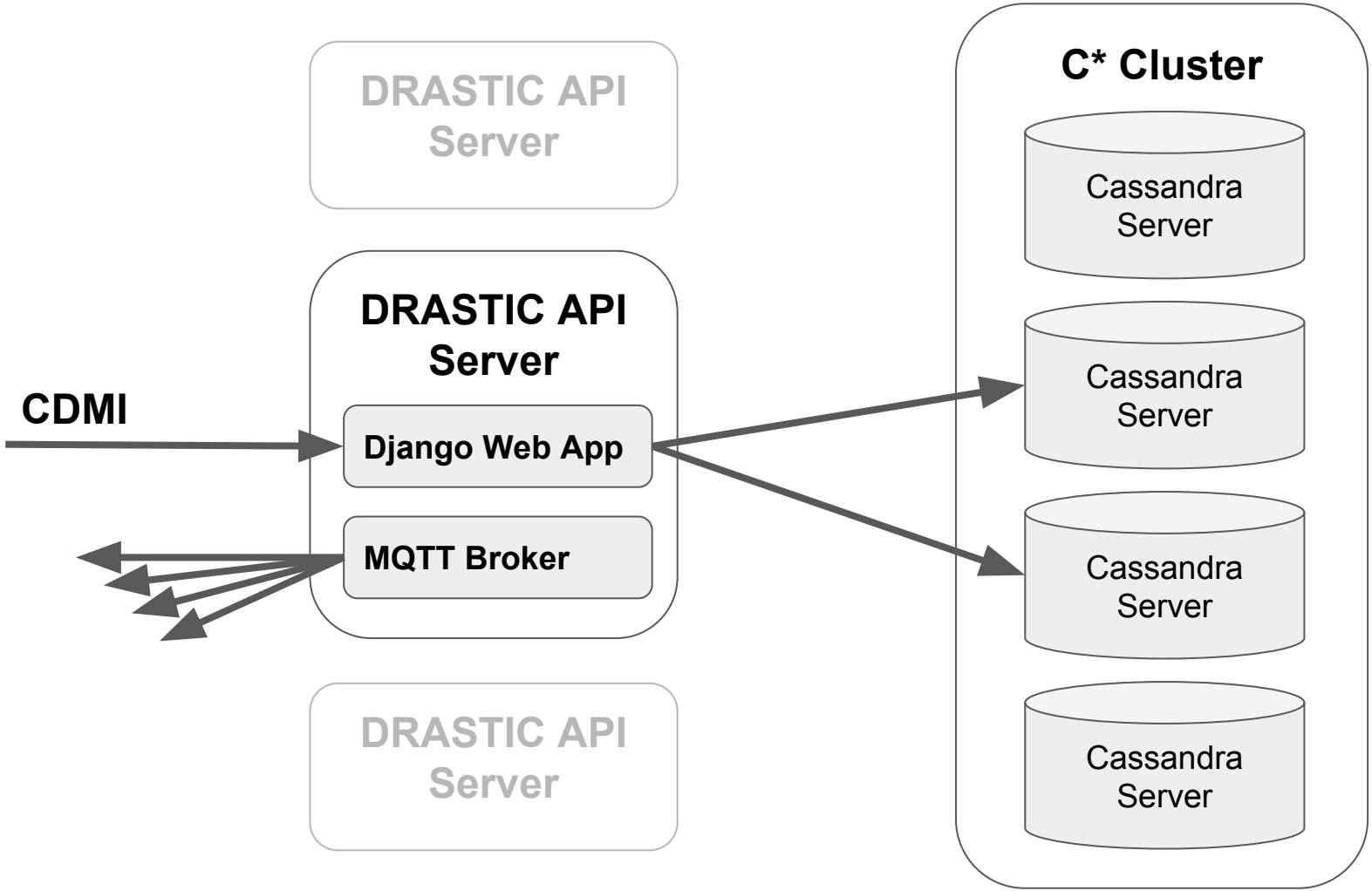
4 x 32 core servers
15 trays of hard drives
180 4 Terabyte drives
720 Terabytes raw storage

NEW

DRAS·TIC

Digital Repository at Scale that Invites Computation

- Product of 2 year startup by partners, Archival Analytics
- Horizontal scaling to billions of files and beyond
- Web UI and command-line client
- Industry standard REST storage API (CDMI)
- Key-value metadata
- Eventing over MQTT message system
- Python source on GitHub (Open AGPL license)
- Based on Apache Cassandra


DRASTIC API Server
DRASTIC API Server
CDMI
Django Web App
MQTT Broker
DRASTIC API Server
C* Cluster
Cassandra Server
Cassandra Server
Cassandra Server
Cassandra Server

DRAS-TIC

RG 029 - Records of the Bureau of Census

Edit Delete

Search ... Go!

Archive
Users
Groups
Activity

Add new collection Add new item

- 2006 Census Operational Photos
- A Profile Of Older Workers In West Virginia
- acs
- acs2002
- acs2003
- acs2004

DRASTIC Measures (next software steps)

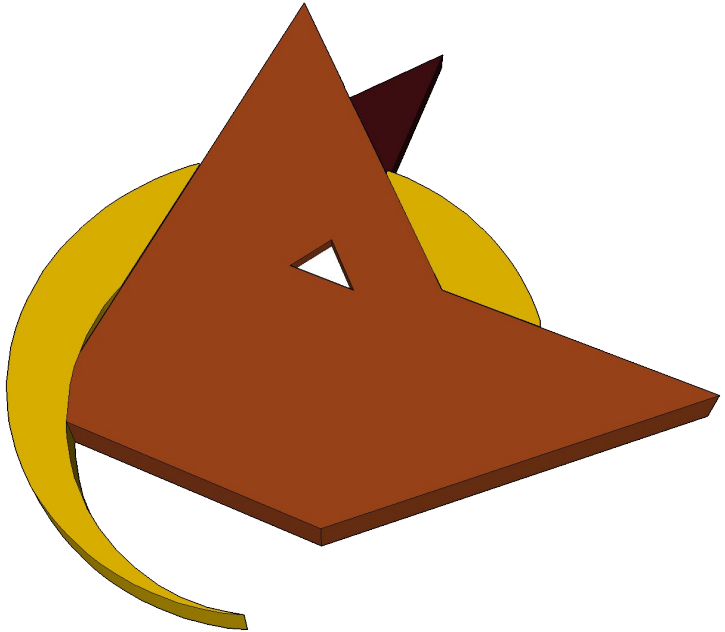
- Integrate with Fedora repository API
- Distributed computing to analyze extracted metadata and full text
- Operationalize DRASTIC for production use:
  - Quickstart installs
  - Backup-recovery scripts
  - Multi-datacenter replication
  - Cloud deployments

NCSA Brown Dog
“The Super Mutt”

Public API for

- Format Migration
- Feature Extraction

Web Scale



Brown Dog Project

- NSF Data Infrastructure Building Blocks (DIBBs) Award to NCSA ($10.5 M, 2013 -2018)

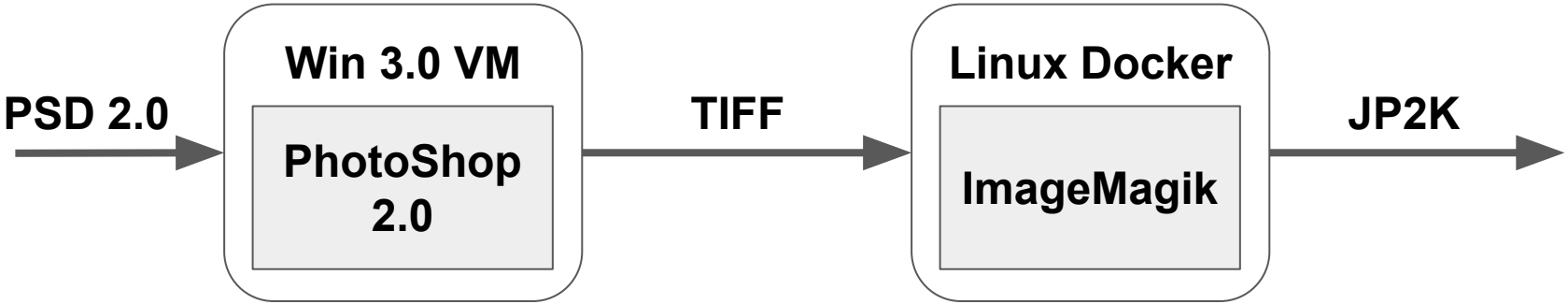
- Collaboration between NCSA, University of Illinois at Urbana-Champaign, University of Maryland, Boston University, Southern Methodist University

Brown Dog REST APIs and Client Tools

- Enable access to file contents irrespective of format
- Extract metadata to enable index and search
- Reuses existing conversion, extract, and analysis tools
- Community contributes extractors and converters
- Dynamic scaling - brings more VMs and HPC online
- Easy to use – provides uniform interface

File Format Conversion

- Image to image or text or PDF format
- AutoDesk's DXF to (svg, jpg, png, tif, pdf, XML)
- A/V format (avi, flv, wav, mp3, mp4) to other A/V formats
- May chain multiple conversion steps using different tools



Metadata/Feature Extraction

- Extract metadata or derived products from a file's content
- File in, JSON-LD out
- Face extraction from image
- Text extraction using OCR
- Data table from a pdf
- Extraction of river paths from historical river maps
- Extraction of vegetation patterns from LIDAR images

Contribute, Share and get Credit for your Tool

- Tools Catalog is a web application that
  - Allows addition of new information by users on new tools they build or any existing tools
  - Share the tools with the community
  - Get credit and citation for your work

Brown Dog Services- Software Components, Cloud/HPC Resources

Clowder

Versus

RabbitMQ

mongoDB

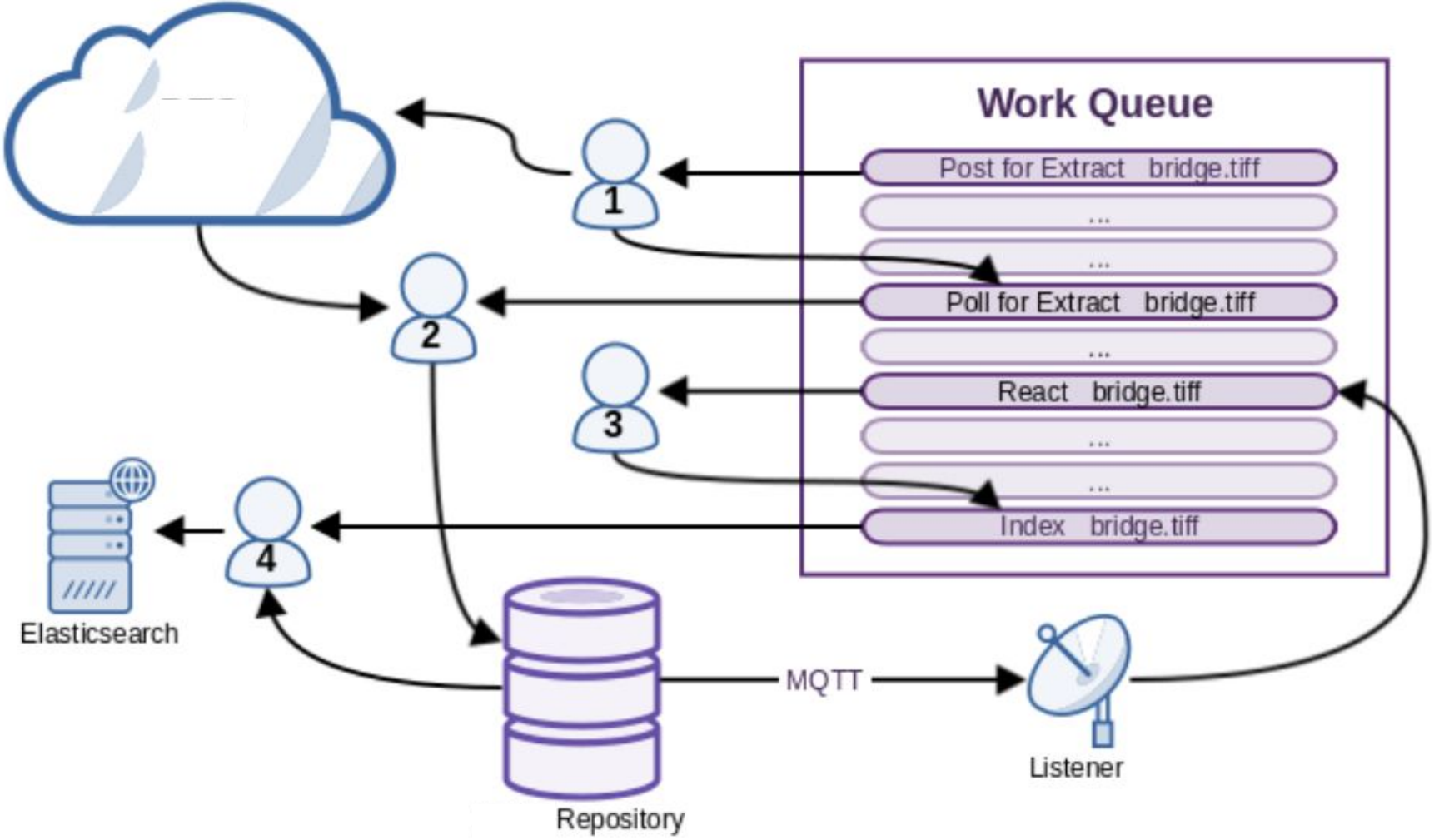
openstack™
CLOUD SOFTWARE

NCSA
Nebula
An OpenStack Cloud Resource

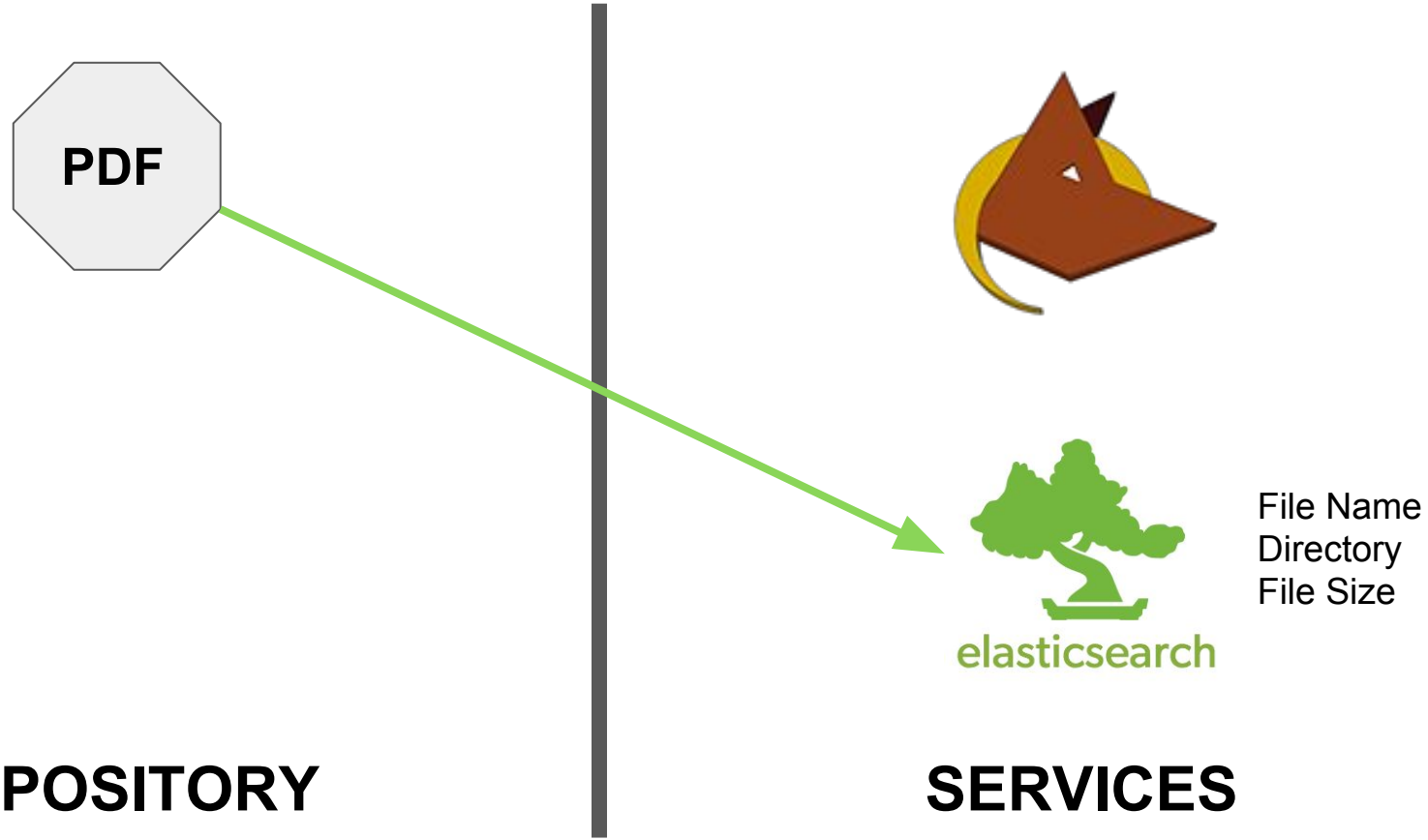
docker

XSEDE
Extreme Science and Engineering Discovery Environment

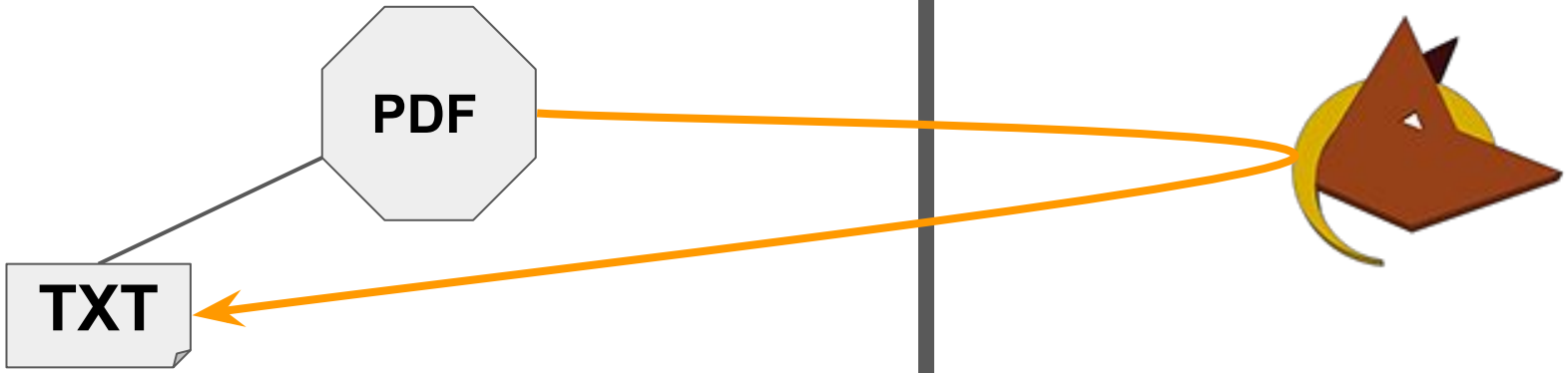
Project website:
http://browndog.ncsa.illinois.edu/


Work Queue
Post for Extract bridge.tiff
...
...
Poll for Extract bridge.tiff
...
React bridge.tiff
...
...
Index bridge.tiff
1
2
3
4
Elasticsearch
Repository
MQTT
Listener

Workflow for a Digital Object



Text Format Conversion (PDF to TXT)

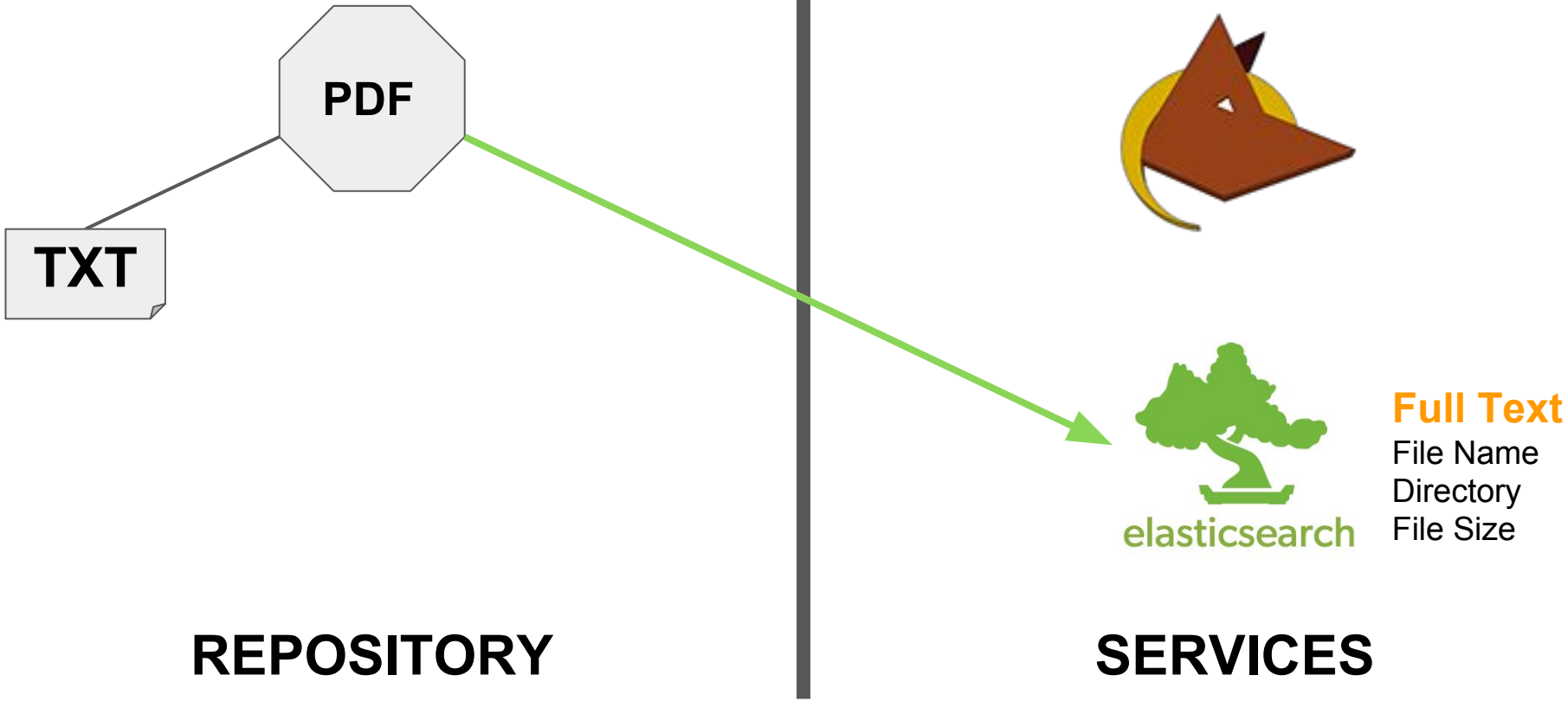


File Name
Directory
File Size

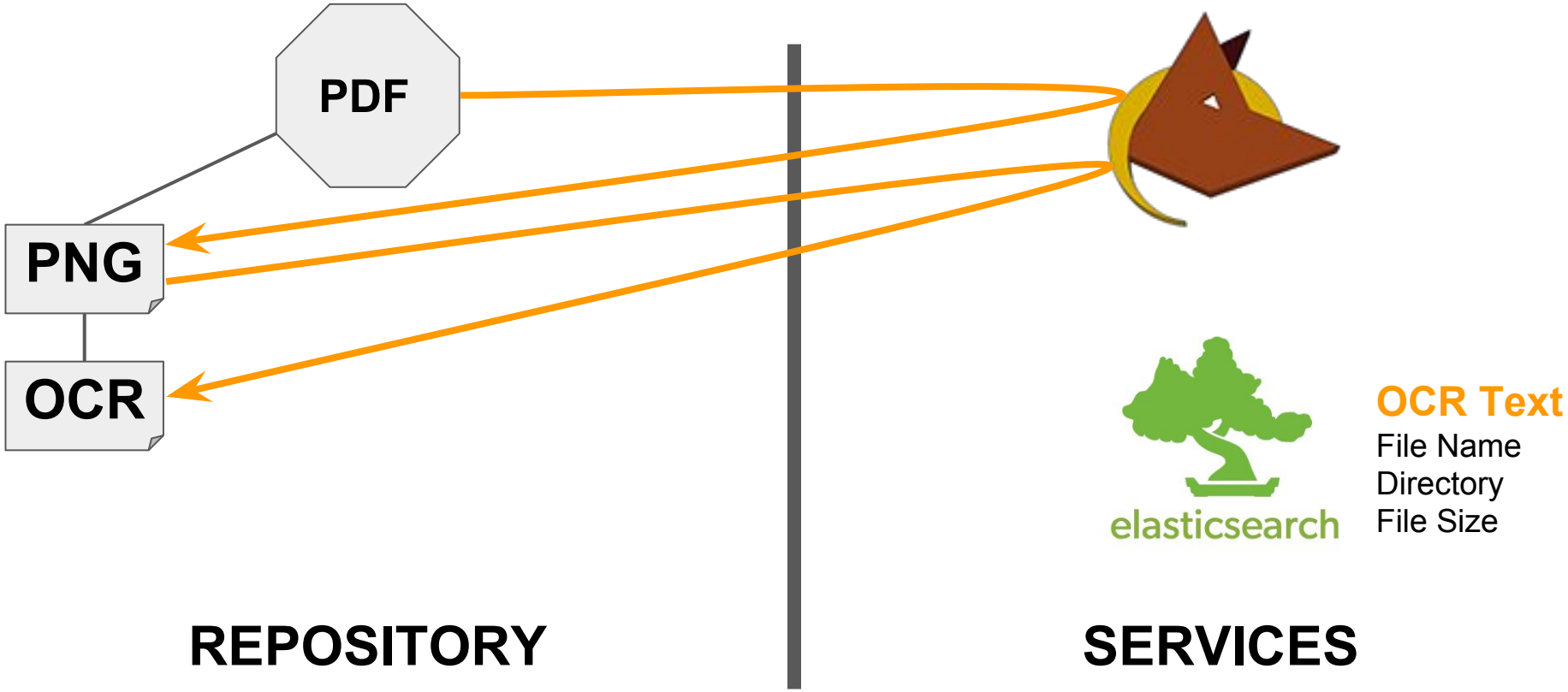
REPOSITORY

SERVICES

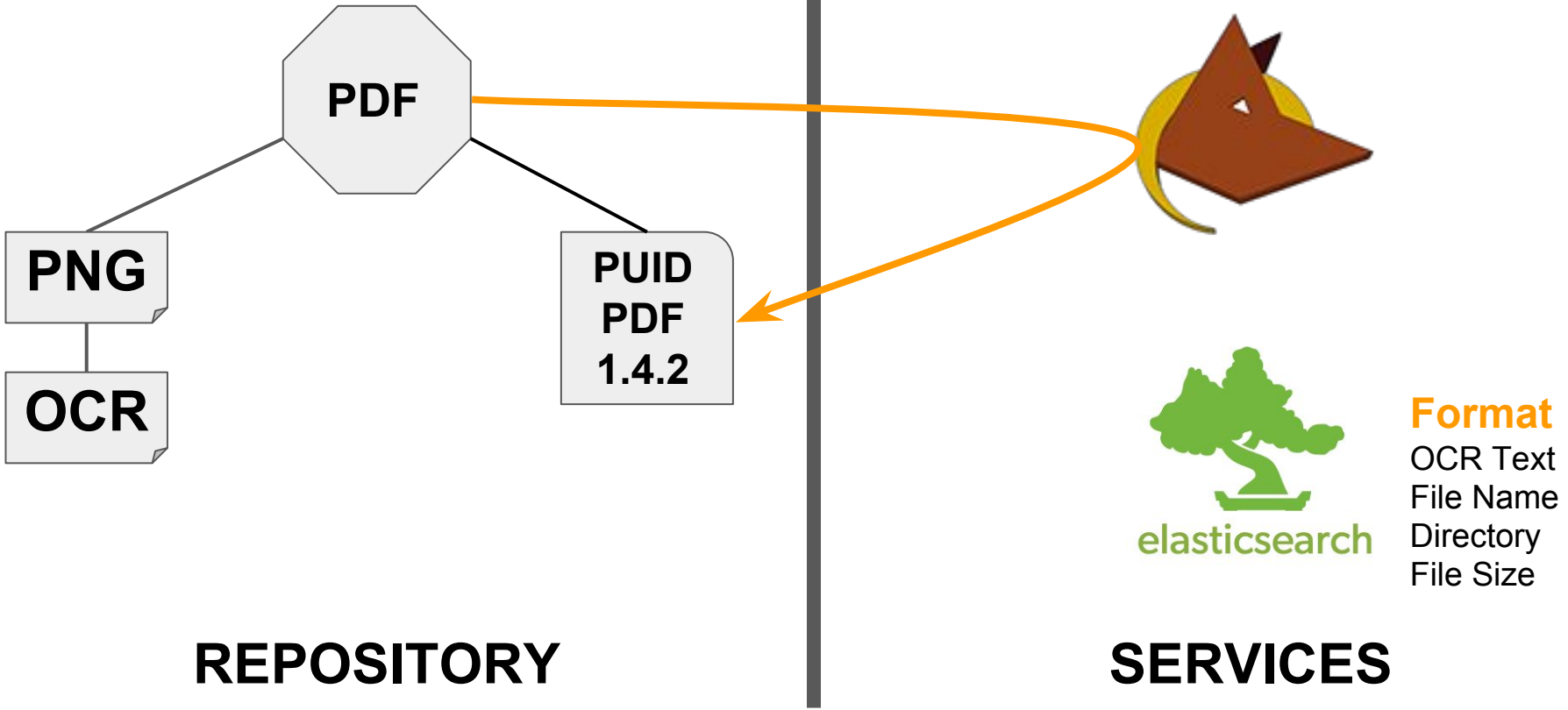
Now we have a full text index..



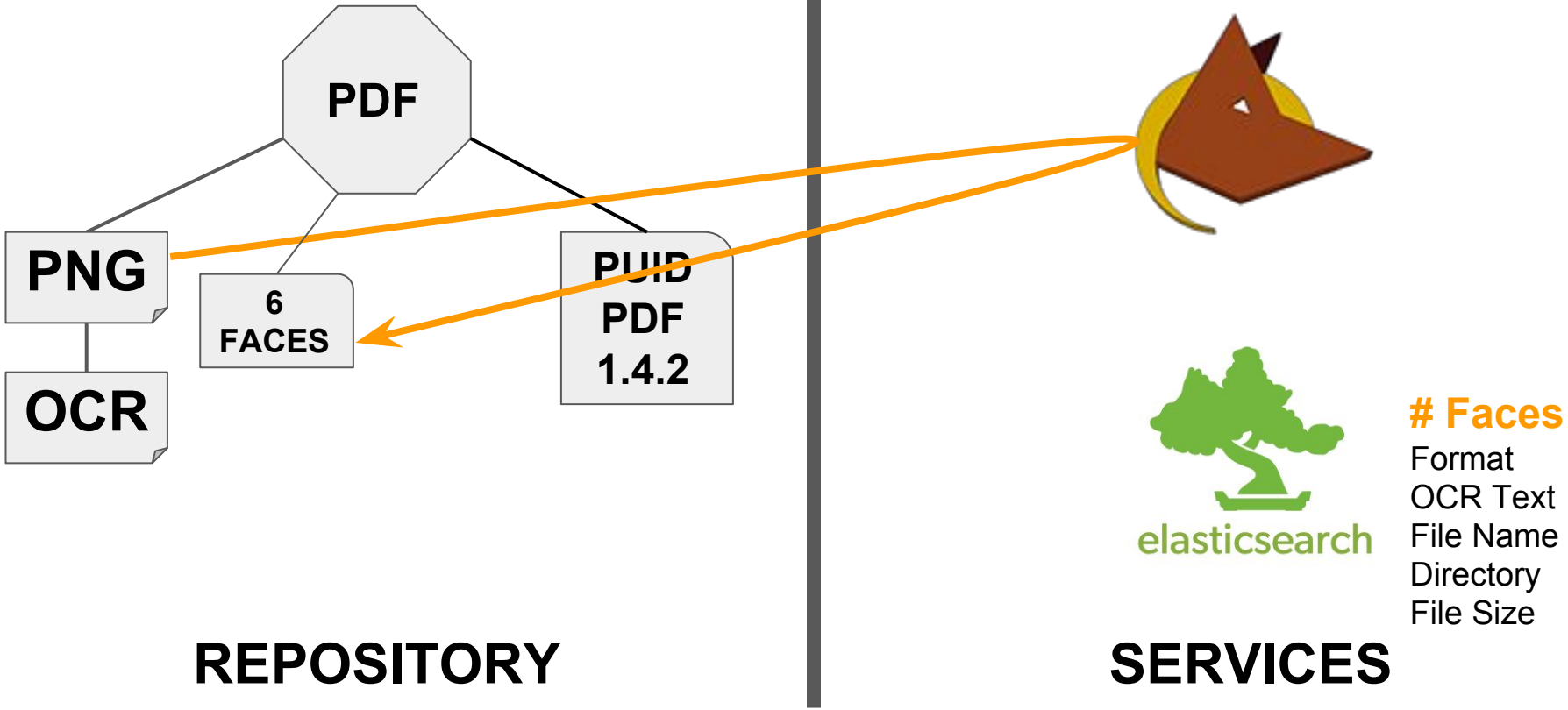
Optical Character Recognition (OCR) Extractor



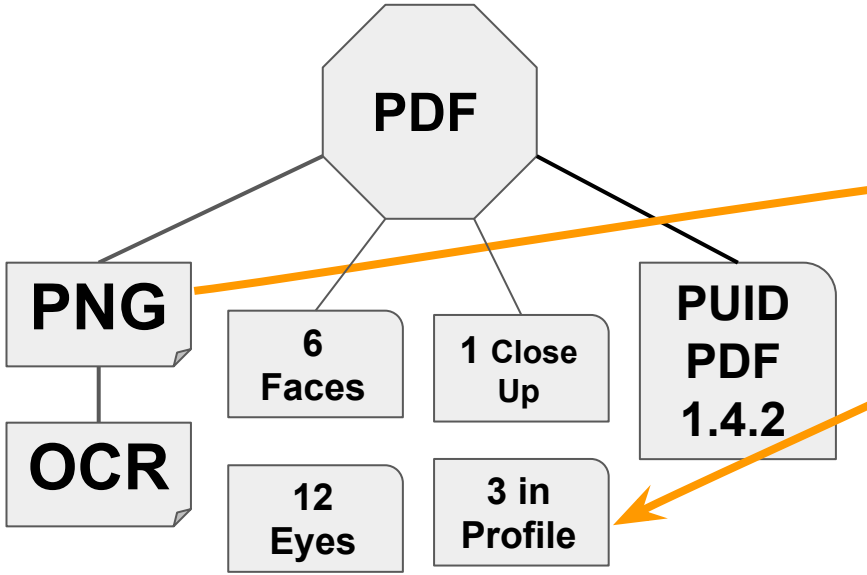
Format Recognition (Siegfried PRONOM Extractor)



Facial Recognition (Computer Vision Extractors)



Facial Recognition (Computer Vision Extractor)



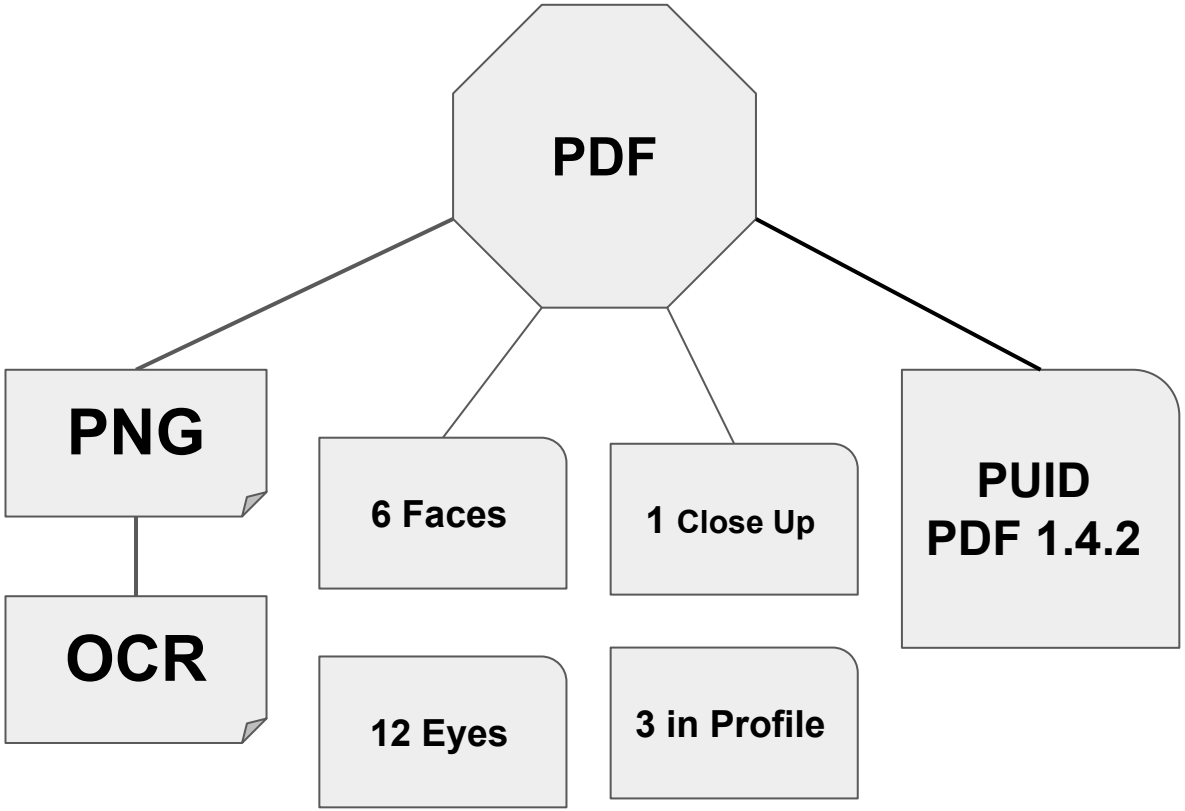
 # Faces
 # Eyes
 # Close Ups
 # Profiles

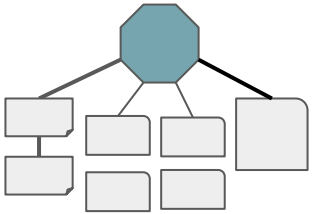
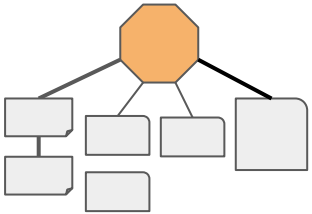
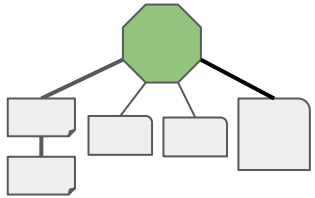
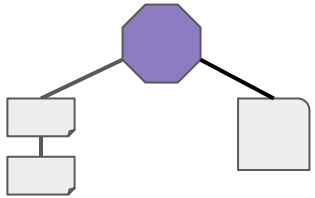
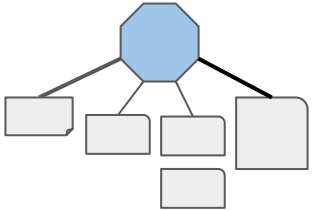
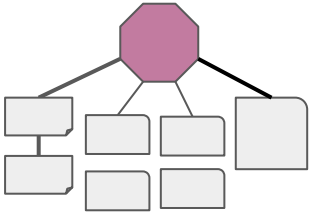
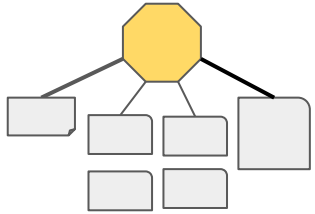
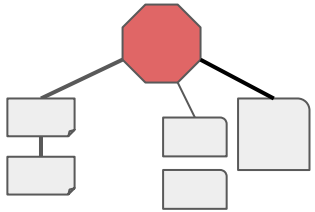
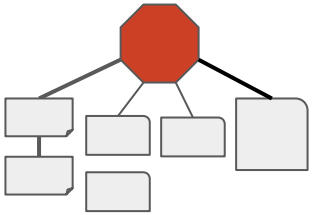
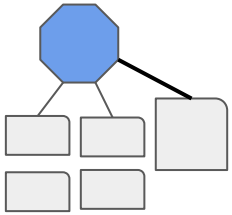
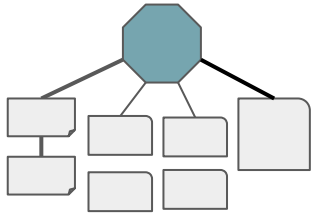
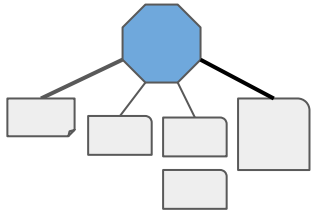
Format
OCR Text
File Name
Directory
File Size

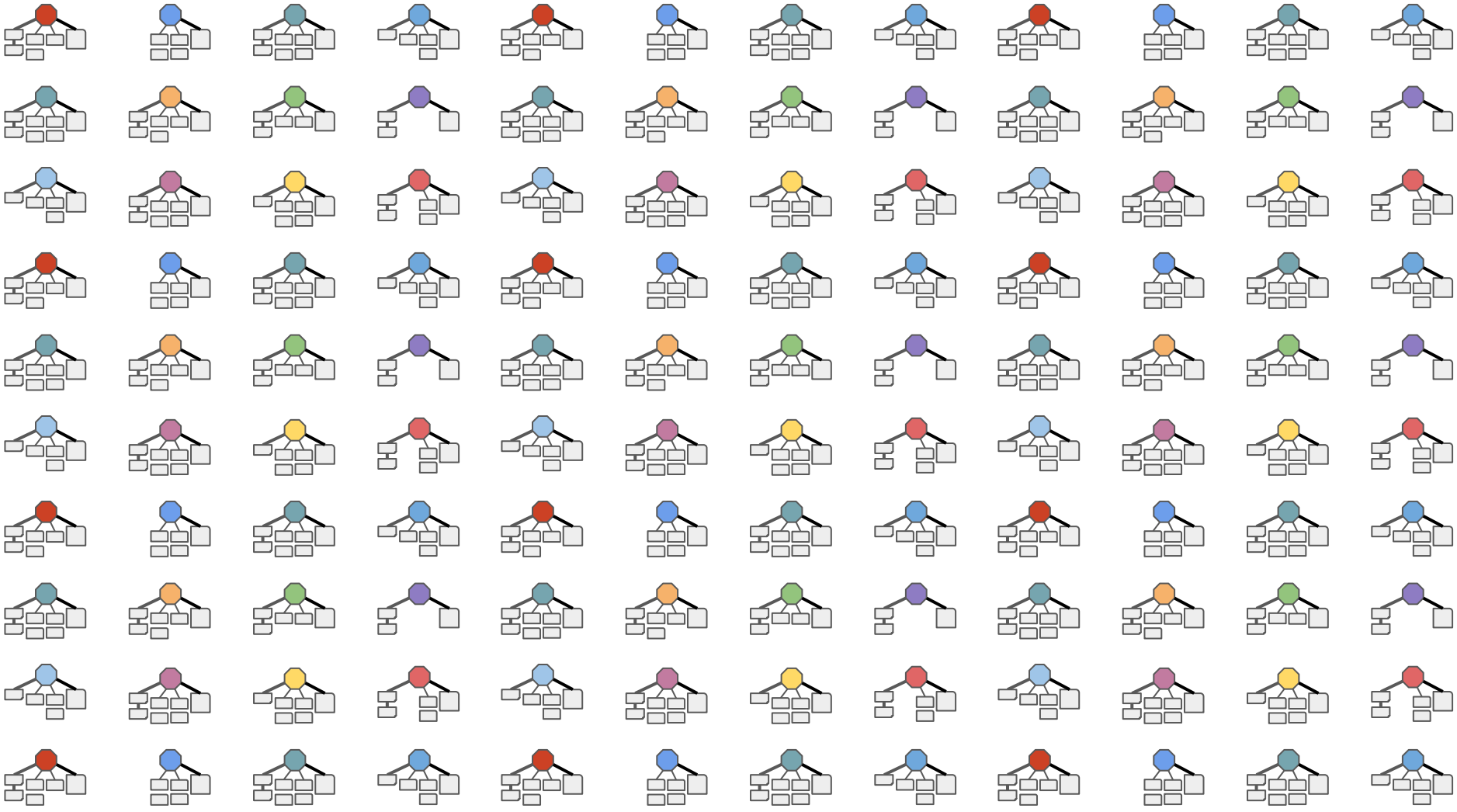
REPOSITORY

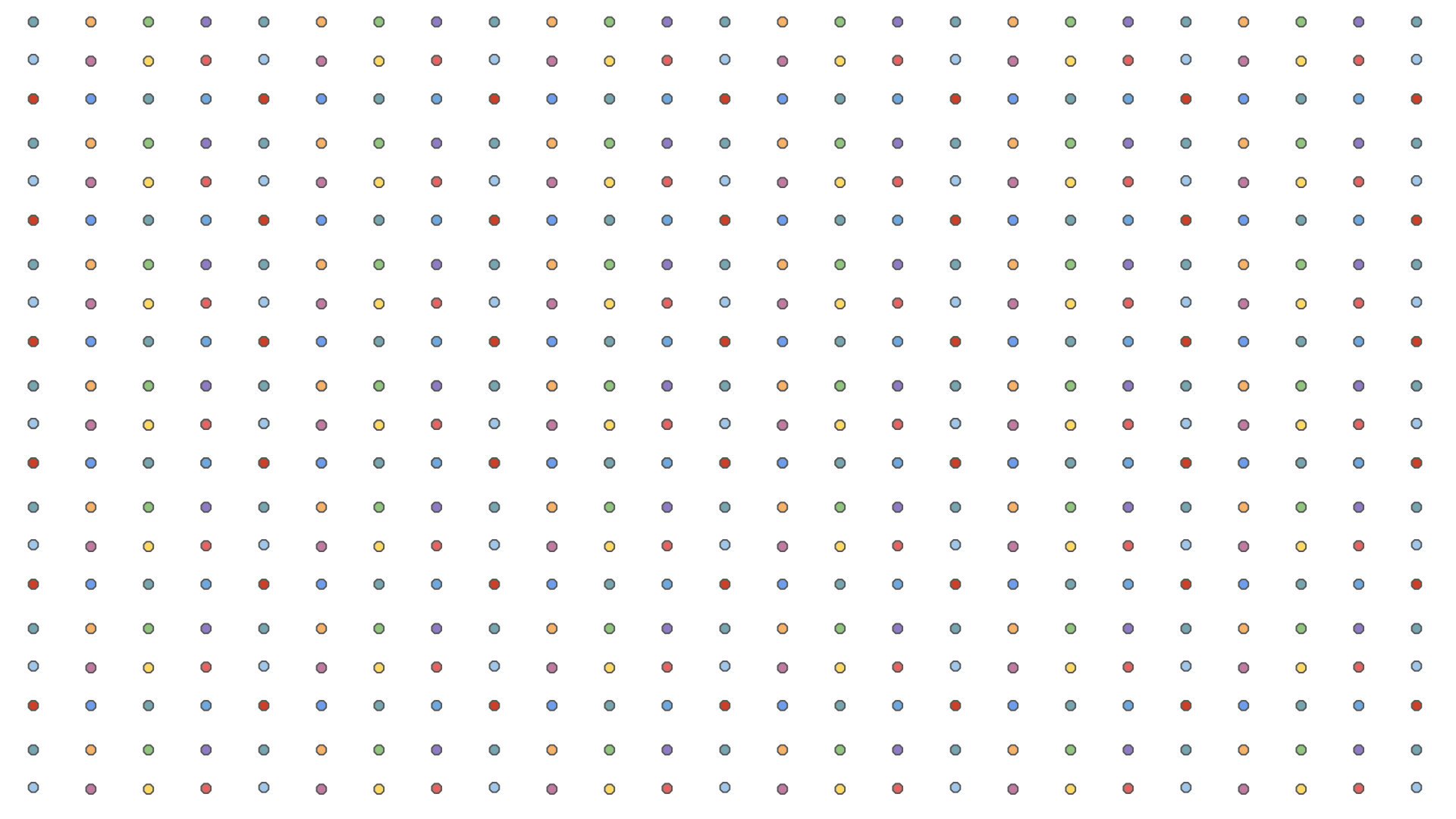
SERVICES

PDF Object Enhanced with Extracted Metadata



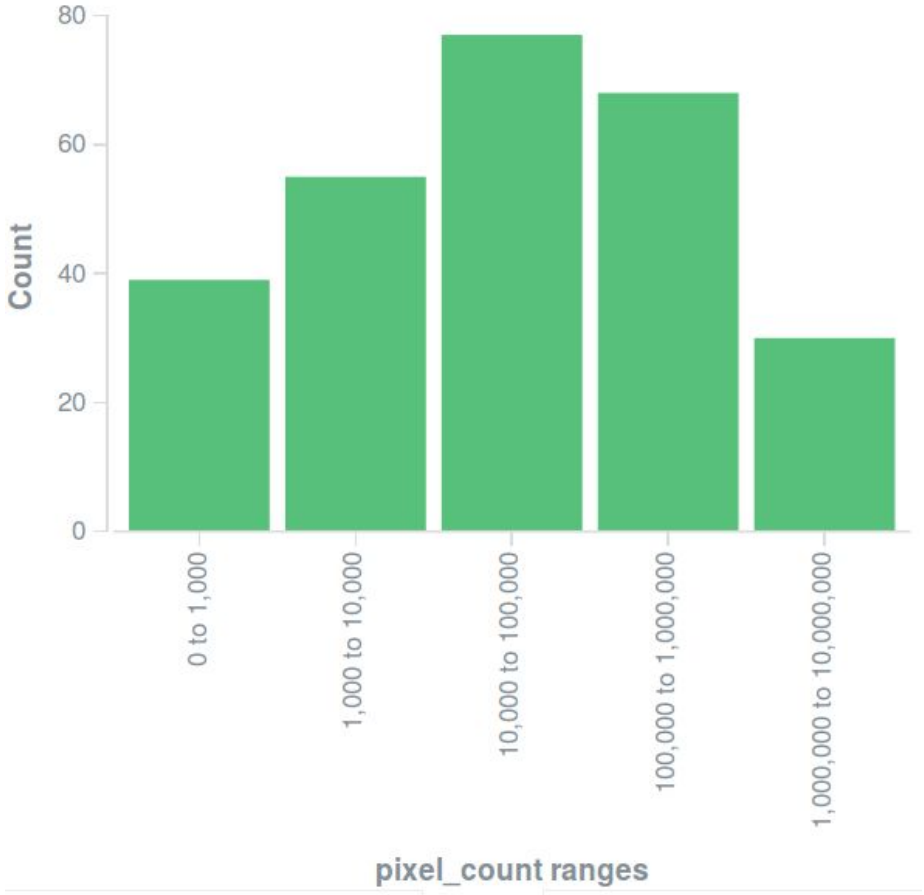




DON’T PANIC

Elasticsearch + Kibana

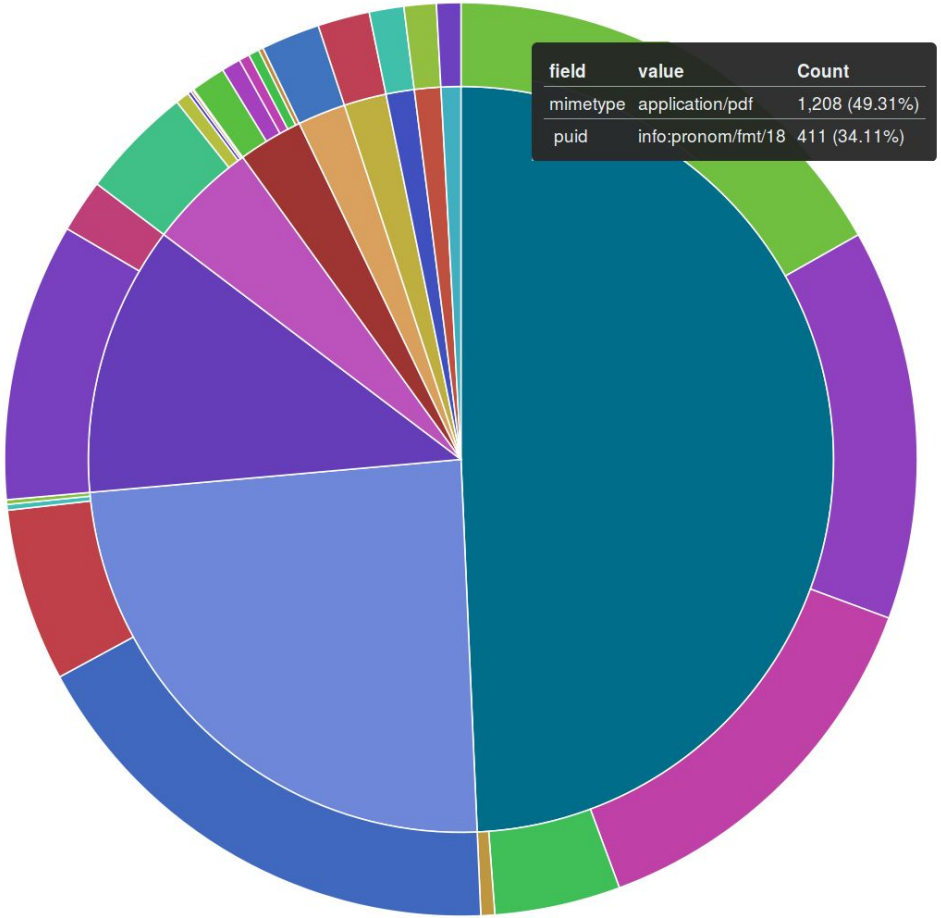
- Free plugin for Elasticsearch
- Gives shape to an Elasticsearch index
- Write queries visually and interactively



Lots of ways to explore the data

Files Formats

- Concentric Pie Chart
- Inner: Mimetype
- Outer: PRONOM PUID



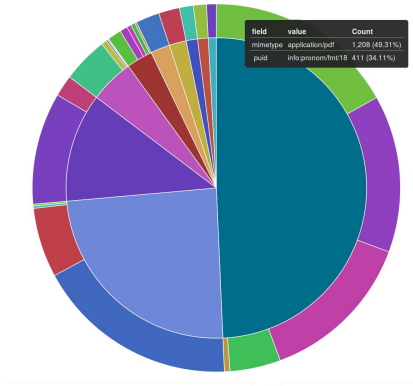
Charts can be added to data dashboards..

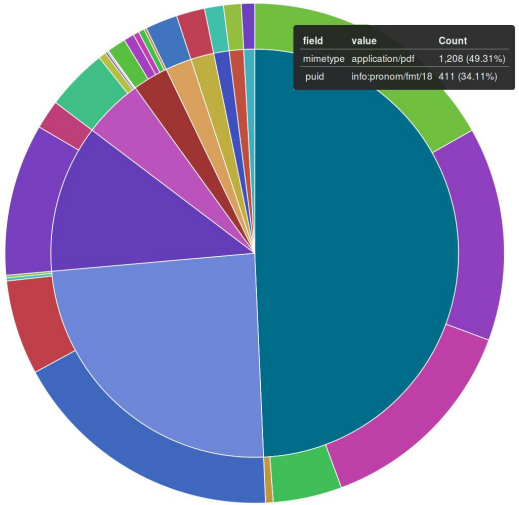
Arrangement can be used as a Facet

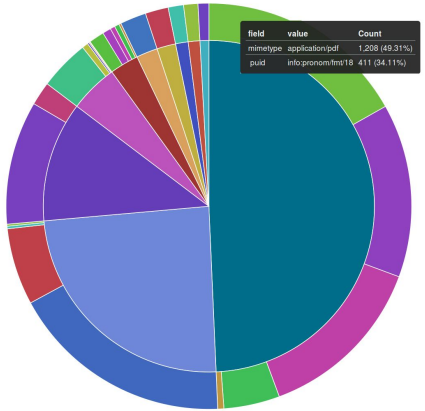
As you browse the hierarchy...

The entire dashboard is redrawn to reflect the particular record group, series or folder under study.

“Drill down” or zoom in and out of your collections.





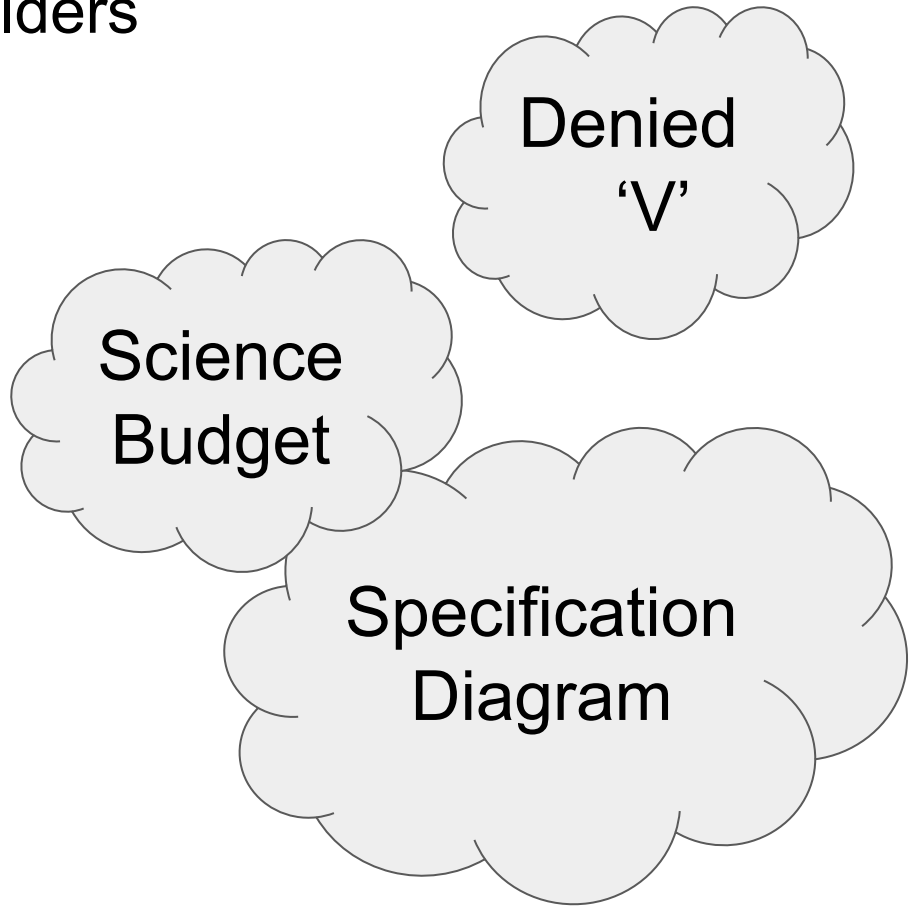


Text Comparison between Folders

Significant Terms are based on full text.

They are significant within overall scope of query.

Significant Terms can be used to distinguish neighboring folders or documents.



parentURI: Descending	Top 2 unusual terms in fulltext	Count
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	denied	606
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	v	670
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	science	305
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	budget	288
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis/lpm6/structural_frame_schema/lexical/	specification	314
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis/lpm6/structural_frame_schema/lexical/	diagram	284

DRAS-TIC

Institutional R&D Partners
Use cases for Parallel Compute
Fedora Sprinters

Brown Dog

Try it on your Scientific Data
Become an Early Adopter of the API
Contribute Extractors & Converters

UMD iSchool

Partner with the DCIC on Projects
Digital Curation Certificate Program
Computational Archival Science

Discuss!

http://dcicblogs.umd.edu
http://github.com/UMD-DRASTIC
http://browndog.ncsa.illinois.edu

jansen@umd.edu
marciano@umd.edu