Discovery of digital forensic dataset characteristics with CASE-Corpora

By:
Alex Nelson (National Institute of Standards and Technology)

From the proceedings of

The Digital Forensic Research Conference

DFRWS USA 2022

July 11-14, 2022

Discovery of digital forensic dataset characteristics with CASE-Corpora

Alex J. Nelson, Ph.D.
Computer Scientist, NIST
Technical Steering Committee Vice-Chair, Cyber Domain Ontology Project
Ontology Committee Chair, Unified Cyber Ontology

DFRWS-USA
2022-07-11

Disclaimer

The views and opinions expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of any agency of the U.S. government. Any mention of a vendor or product is not an endorsement or recommendation. Logos and trademarks are copyright their respective owners.

We’ll always need test data…

And what luck, we keep getting it.

Ample motivations exist:

- Professional training and certification data sets
- Academic papers’ exemplar samples
- Forensic contests
- Capture the Flags
- Investigation-relevant controlled reproductions
- Other controlled demonstrations

…but how do we find it when we need it?

How do YOU find your test data?

- Infosec Twitter?
- Conferences? Journals?
- Build it yourself?
  - …And release it? Data re-use and release was low circa 2017:
    2017, Grajeda et al., "Availability of datasets for digital forensics – And what is missing”
    https://doi.org/10.1016/j.diin.2017.06.004
- How do you find data years after its publication?
  - Corpus management remains a significant challenge.
    2009, Garfinkel et al., "Bringing science to digital forensics with standardized forensic corpora”
    https://doi.org/10.1016/j.diin.2009.06.016
  - Corpus distribution and discovery is harder.

CASE-Corpora is a forensic data catalog.

CASE-Corpora indexes forensic dataset metadata.

- Extends general data-catalog language with forensic concepts.
- Adds chain of custody details.
- Encodes authors’ ground truth descriptions for search, discovery, and cross-verification.
- CASE general-purpose tools help analyze and maintain data quality.

Outline

- Background
  - Graphs
  - Ontologies and data models
- Ontologies used in CASE-Corpora
- Provenance
- Usage of CASE-Corpora

Background

Ontologies used in CASE-Corpora

Provenance

Usage of CASE-Corpora

The data in CASE-Corpora is written as RDF.

RDF – Resources Data Framework
(see also "Semantic web")

Used to define a *graph* of:

- Individuals
  (E.g. Paul Erdös, Kevin Bacon, Hank Aaron)
- Classes
  (E.g. Mathematicians, Film Stars, Baseball Players)
- Properties
  (E.g. Co-authored with, co-starred in)

Graphs are defined with ontologies,
which are models of reality.

RDF serializes interchangeably in several formats, including:

- XML
- JSON-LD ("JSON Linked Data")
- Turtle

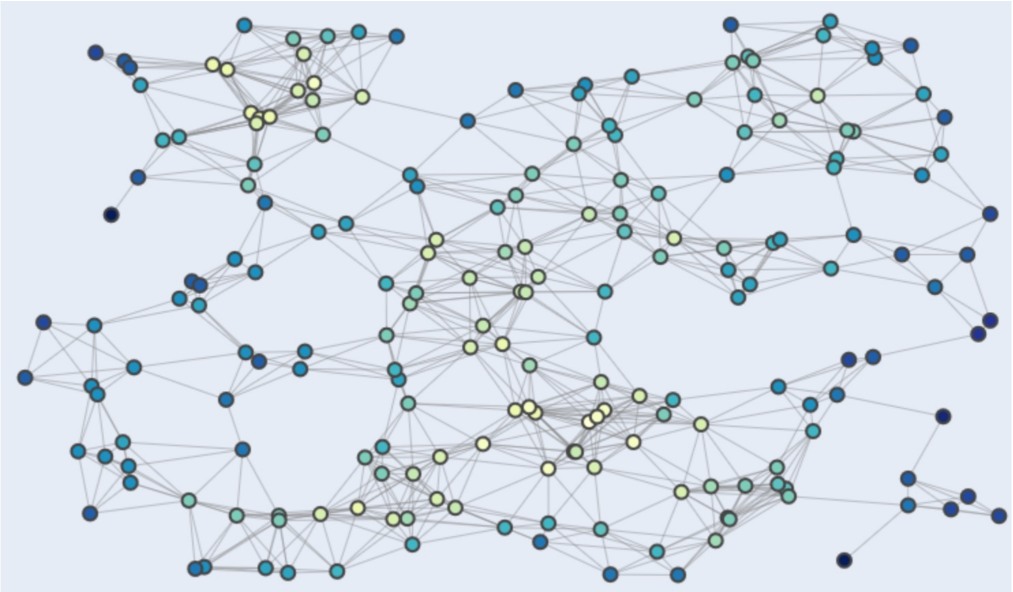


Figure: A random graph.

Figure source: https://plotly.com/python/network-graphs/

RDF querying uses SPARQL.

SPARQL is a graph query language, similar in purpose to SQL.

SPARQL is suited for:

- Path queries (What’s person X’s Erdös-Bacon Number?)
- Arbitrary relationships (In what ways does X relate to Erdös or Bacon?)

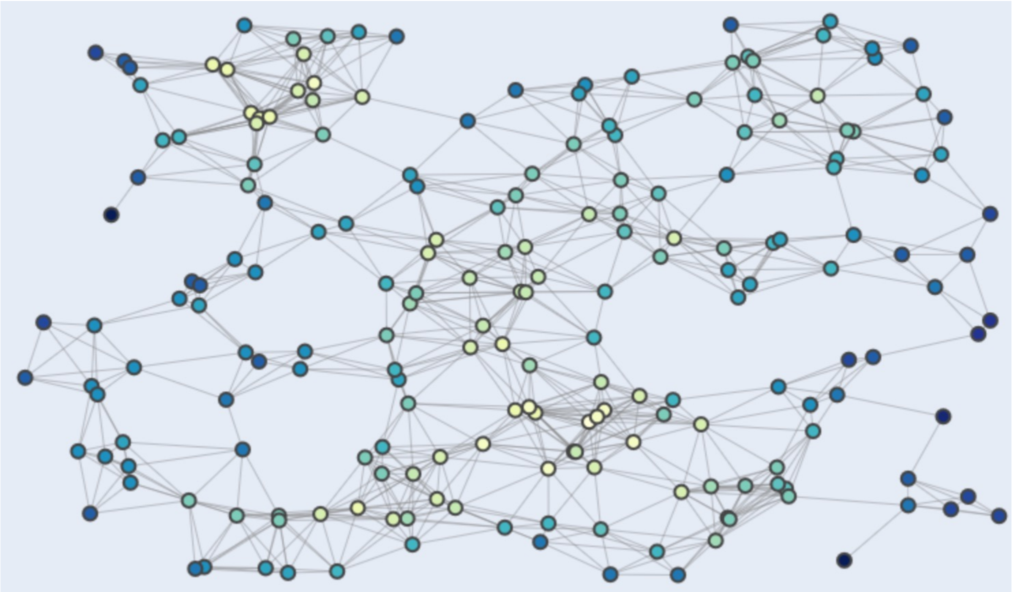


Figure: A random graph.

Figure source: https://plotly.com/python/network-graphs/

Background

Ontologies used in CASE-Corpora

Provenance

Usage of CASE-Corpora

The CDO ontologies in CASE-Corpora

CDO – The Cyber Domain Ontology Project
A series of Linux Foundation Projects, LLC
Established January, 2022
Umbrella project for CASE & UCO (both first drafted 2016), and future ontological communities of interest.

- *UCO – Unified Cyber Ontology*
  A middle-level ontology providing
  cross-domain cyber concepts
- *CASE – Cyber-investigation Analysis Standard Expression*
  An ontology community of interest,
  extending UCO into investigations

Version 1.0.0 scheduled for August 30, 2022.

Non-CDO ontologies in CASE-Corpora

DCAT (and DCAT-US) - Data Catalog Vocabulary

A RDF-based data model describing data sets

- Includes what resources are in a dataset,
  where to download them,
  and other *publication-level* metadata.
- Underpins https://data.gov/

PROV-O – Provenance Ontology

An OWL implementation of histories of objects

Background

Ontologies used in CASE-Corpora

Provenance

Usage of CASE-Corpora

PROV-O represents and illustrates provenance.

PROV-O is built upon:

- Activities
- Agents
- Entities

Time (and logical ordering) illustrates flowing downward.

CASE practice:
Provenance chains link back to initial evidence submission.

“Initial” = derived from nothing

“Nothing” = the PROV-O empty set

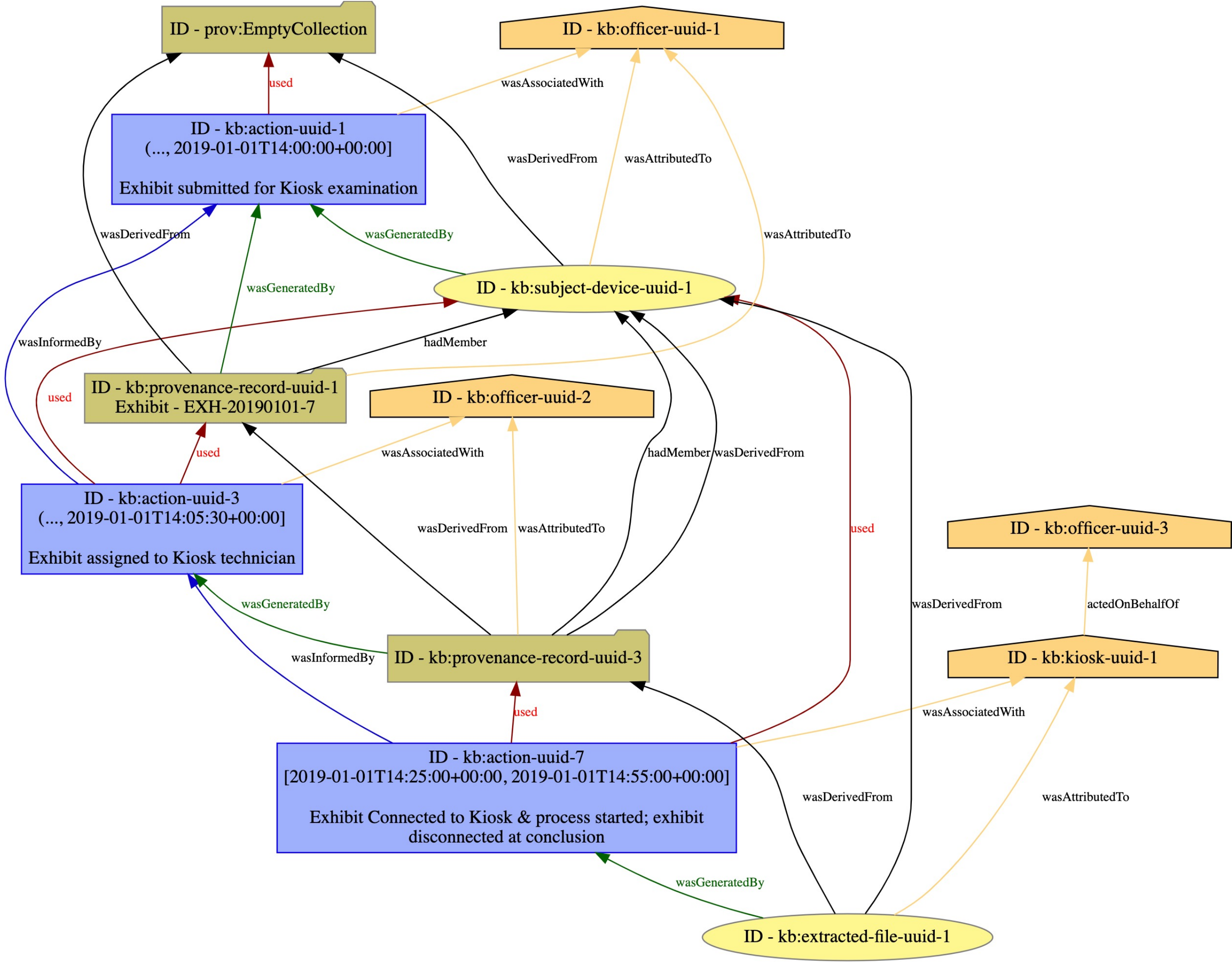
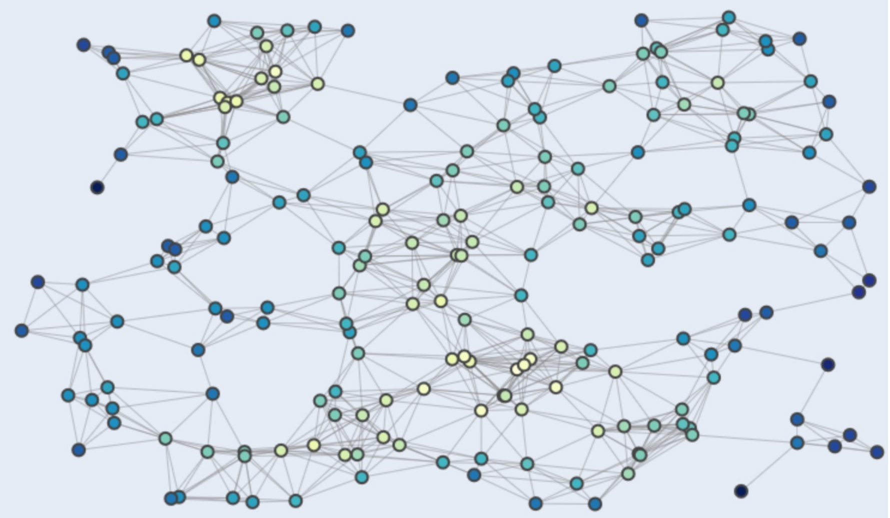
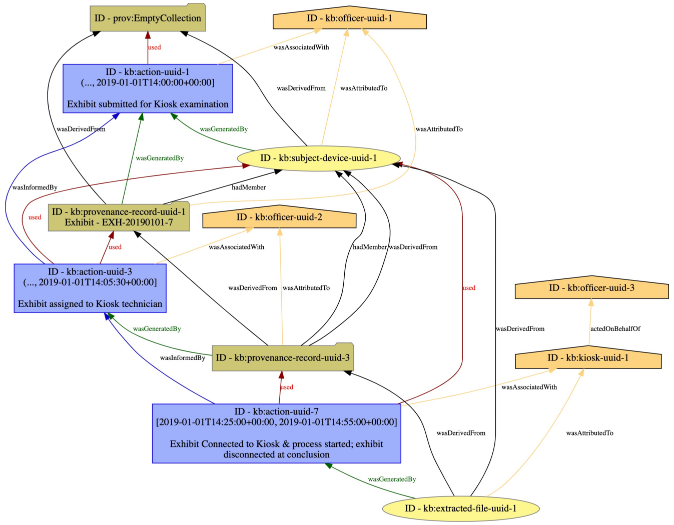


Figure: “Urgent Evidence” narrative, history of one extracted JPEG

CASE projects into PROV-O.



`case_prov_rdf`



ID - prov:EmptyCollection

ID - kb:officer-uuid-1

ID - kb:action-uuid-1
(..., 2019-01-01T14:00:00+00:00]

Exhibit submitted for Kiosk examination

ID - kb:subject-device-uuid-1

ID - kb:provenance-record-uuid-1
Exhibit - EXH-20190101-7

ID - kb:officer-uuid-2

ID - kb:action-uuid-3
(..., 2019-01-01T14:05:30+00:00]

Exhibit assigned to Kiosk technician

ID - kb:officer-uuid-3

ID - kb:provenance-record-uuid-3

ID - kb:kiosk-uuid-1

ID - kb:action-uuid-7
[2019-01-01T14:25:00+00:00, 2019-01-01T14:55:00+00:00]

Exhibit Connected to Kiosk & process started; exhibit disconnected at conclusion

ID - kb:extracted-file-uuid-1

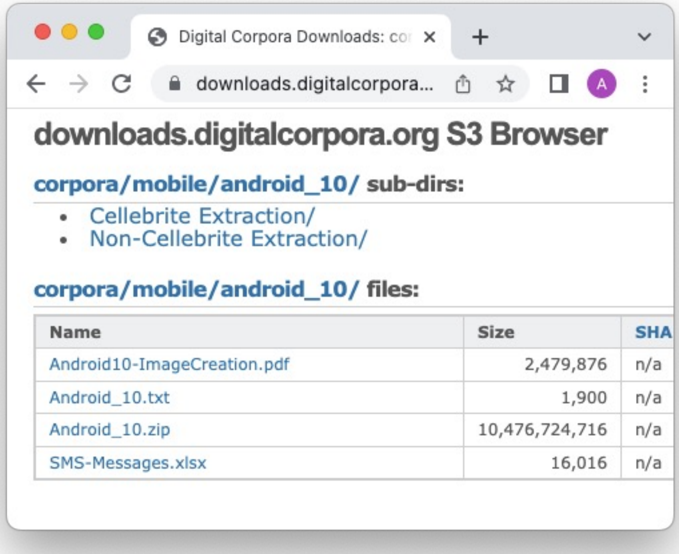
A graph of a CASE investigation’s provenance chain …

…maps directly to PROV-O.

Left figure source: https://plotly.com/python/network-graphs/

Datasets need provenance review.

Digital Corpora
“Android 10” dataset



downloads.digitalcorpora.org S3 Browser

corpora/mobile/android_10/ sub-dirs:

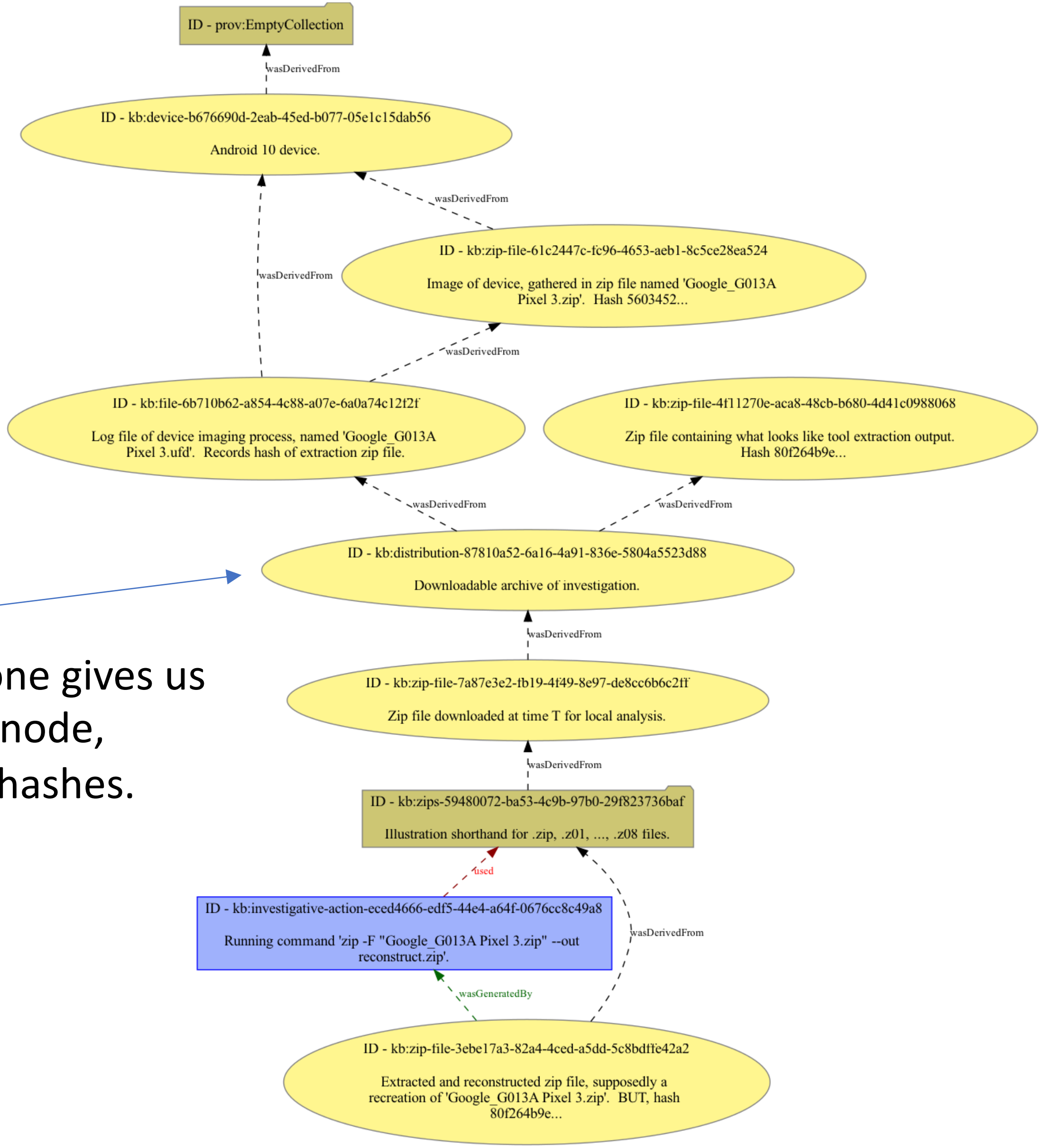
- Cellebrite Extraction/
- Non-Cellebrite Extraction/

corpora/mobile/android_10/ files:

Name	Size	SHA
Android10-ImageCreation.pdf	2,479,876	n/a
Android_10.txt	1,900	n/a
Android_10.zip	10,476,724,716	n/a
SMS-Messages.xlsx	16,016	n/a

This provenance was *sketched* from log:
`Google_G013A Pixel 3.ufd`

DCAT alone gives us this one node, without hashes.



ID - prov:EmptyCollection

ID - kb:device-b676690d-2eab-45ed-b077-05e1c15dab56

Android 10 device.

ID - kb:zip-file-61c2447c-fc96-4653-aeb1-8c5ce28ea524

Image of device, gathered in zip file named 'Google_G013A Pixel 3.zip'. Hash 5603452...

ID - kb:file-6b710b62-a854-4c88-a07e-6a0a74c12f2f

Log file of device imaging process, named 'Google_G013A Pixel 3.ufd'. Records hash of extraction zip file.

ID - kb:zip-file-4f11270e-aca8-48cb-b680-4d41c0988068

Zip file containing what looks like tool extraction output. Hash 80f264b9e...

ID - kb:distribution-87810a52-6a16-4a91-836e-5804a5523d88

Downloadable archive of investigation.

ID - kb:zip-file-7a87e3e2-fb19-4f49-8e97-de8cc6b6c2ff

Zip file downloaded at time T for local analysis.

ID - kb:zips-59480072-ba53-4c9b-97b0-29f823736baf

Illustration shorthand for .zip, .z01, ..., .z08 files.

ID - kb:investigative-action-eced4666-edf5-44e4-a64f-0676cc8c49a8

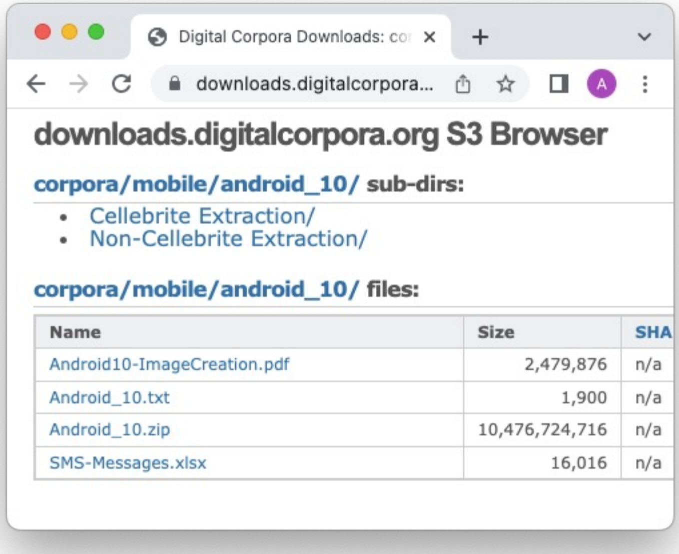
Running command 'zip -F "Google_G013A Pixel 3.zip" --out reconstruct.zip'.

ID - kb:zip-file-3ebe17a3-82a4-4ced-a5dd-5c8bdffe42a2

Extracted and reconstructed zip file, supposedly a recreation of 'Google_G013A Pixel 3.zip'. BUT, hash 80f264b9e...

Datasets need provenance review.

Digital Corpora “Android 10” dataset



downloads.digitalcorpora.org S3 Browser

corpora/mobile/android_10/ sub-dirs:

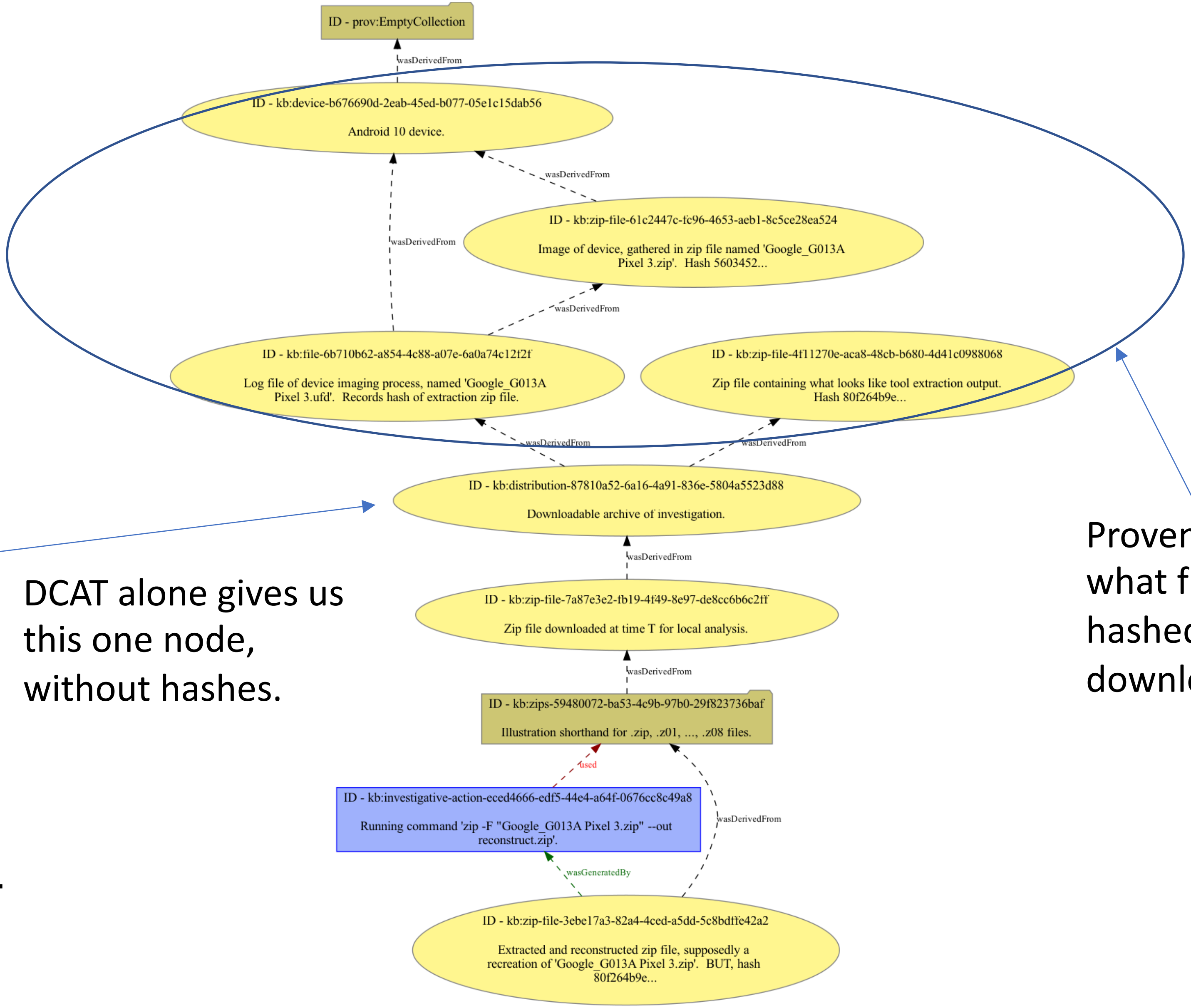
- Cellebrite Extraction/
- Non-Cellebrite Extraction/

corpora/mobile/android_10/ files:

Name	Size	SHA
Android10-ImageCreation.pdf	2,479,876	n/a
Android_10.txt	1,900	n/a
Android_10.zip	10,476,724,716	n/a
SMS-Messages.xlsx	16,016	n/a

This provenance was *sketched* from log:

`Google_G013A Pixel 3.ufd`



ID - prov:EmptyCollection

ID - kb:device-b676690d-2eab-45ed-b077-05e1c15dab56

Android 10 device.

ID - kb:zip-file-61c2447c-fc96-4653-aeb1-8c5ce28ea524

Image of device, gathered in zip file named 'Google_G013A Pixel 3.zip'. Hash 5603452...

ID - kb:file-6b710b62-a854-4c88-a07e-6a0a74c12f2f

Log file of device imaging process, named 'Google_G013A Pixel 3.ufd'. Records hash of extraction zip file.

ID - kb:zip-file-4f11270e-aca8-48cb-b680-4d41c0988068

Zip file containing what looks like tool extraction output. Hash 80f264b9e...

ID - kb:distribution-87810a52-6a16-4a91-836e-5804a5523d88

Downloadable archive of investigation.

ID - kb:zip-file-7a87e3e2-fb19-4f49-8e97-de8cc6b6c2ff

Zip file downloaded at time T for local analysis.

ID - kb:zips-59480072-ba53-4c9b-97b0-29f823736baf

Illustration shorthand for .zip, .z01, ..., .z08 files.

ID - kb:investigative-action-eced4666-edf5-44e4-a64f-0676cc8c49a8

Running command 'zip -F "Google_G013A Pixel 3.zip" --out reconstruct.zip'.

ID - kb:zip-file-3ebe17a3-82a4-4ced-a5dd-5c8bdffe42a2

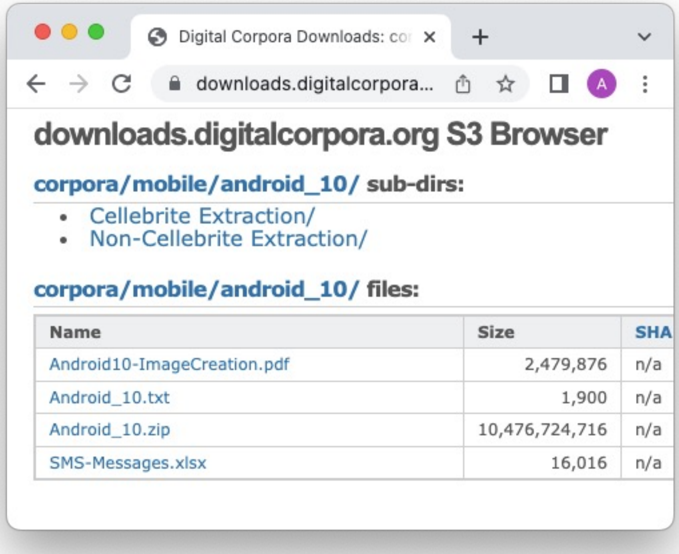
Extracted and reconstructed zip file, supposedly a recreation of 'Google_G013A Pixel 3.zip'. BUT, hash 80f264b9e...

DCAT alone gives us this one node, without hashes.

Provenance shows what files were hashed, and led to downloadable link.

Datasets need provenance review.

Digital Corpora "Android 10" dataset



downloads.digitalcorpora.org S3 Browser

corpora/mobile/android_10/ sub-dirs:

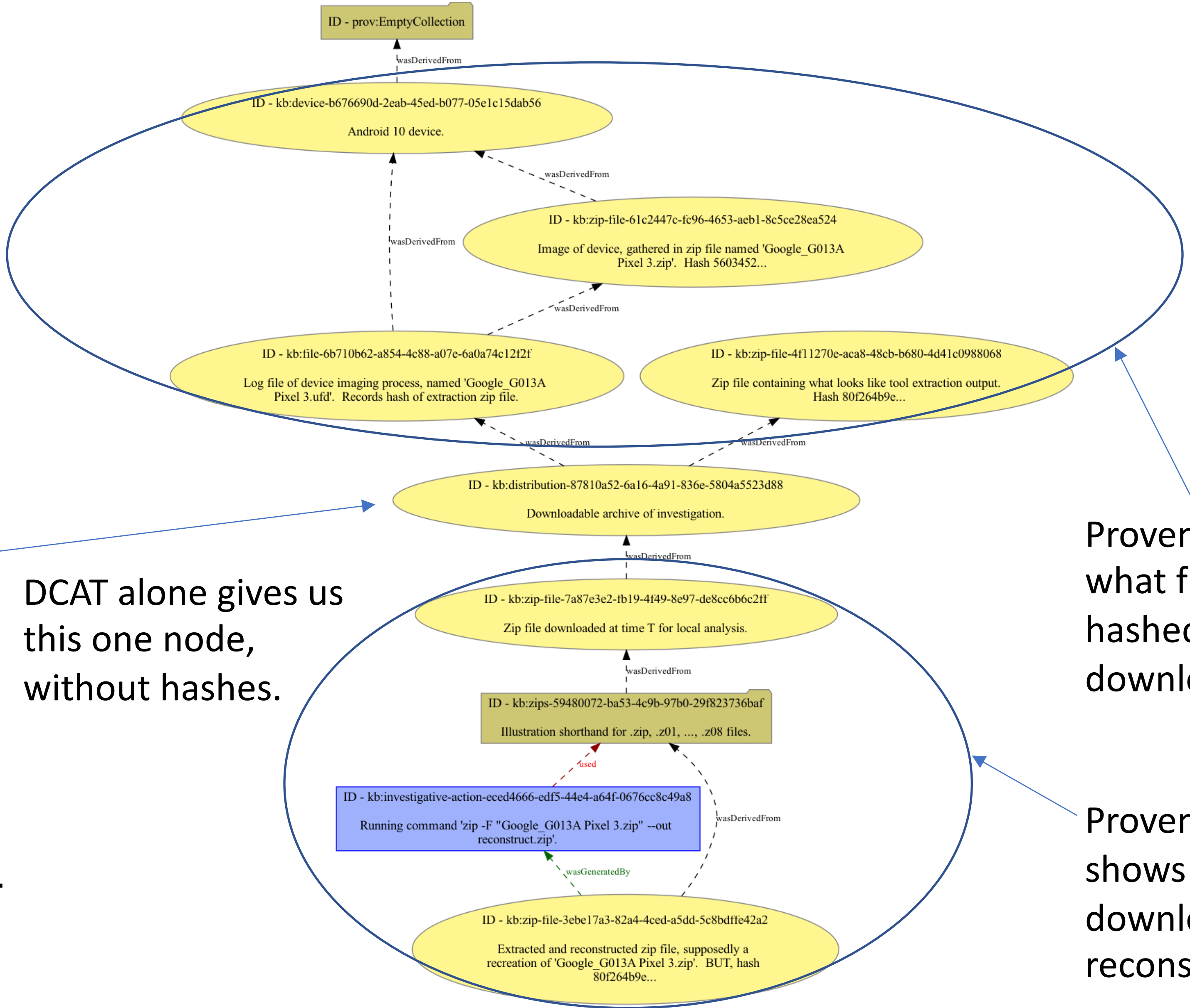
- Cellebrite Extraction/
- Non-Cellebrite Extraction/

corpora/mobile/android_10/ files:

Name	Size	SHA
Android10-ImageCreation.pdf	2,479,876	n/a
Android_10.txt	1,900	n/a
Android_10.zip	10,476,724,716	n/a
SMS-Messages.xlsx	16,016	n/a

This provenance was *sketched* from log:

`Google_G013A Pixel 3.ufd`



ID - prov:EmptyCollection

ID - kb:device-b676690d-2eab-45ed-b077-05e1c15dab56

Android 10 device.

ID - kb:zip-file-61c2447c-fc96-4653-aeb1-8c5ce28ea524

Image of device, gathered in zip file named 'Google_G013A Pixel 3.zip'. Hash 5603452...

ID - kb:file-6b710b62-a854-4c88-a07e-6a0a74c12f2f

Log file of device imaging process, named 'Google_G013A Pixel 3.ufd'. Records hash of extraction zip file.

ID - kb:zip-file-4f11270e-aca8-48cb-b680-4d41c0988068

Zip file containing what looks like tool extraction output. Hash 80f264b9e...

ID - kb:distribution-87810a52-6a16-4a91-836e-5804a5523d88

Downloadable archive of investigation.

ID - kb:zip-file-7a87e3e2-fb19-4f49-8e97-de8cc6b6c2ff

Zip file downloaded at time T for local analysis.

ID - kb:zips-59480072-ba53-4c9b-97b0-29f823736baf

Illustration shorthand for .zip, .z01, ..., .z08 files.

ID - kb:investigative-action-eced4666-edf5-44e4-a64f-0676cc8c49a8

Running command 'zip -F "Google_G013A Pixel 3.zip" --out reconstruct.zip'.

ID - kb:zip-file-3ebe17a3-82a4-4ced-a5dd-5c8bdffe42a2

Extracted and reconstructed zip file, supposedly a recreation of 'Google_G013A Pixel 3.zip'. BUT, hash 80f264b9e...

DCAT alone gives us this one node, without hashes.

Provenance shows what files were hashed, and led to downloadable link.

Provenance also shows later download and reconstruction.

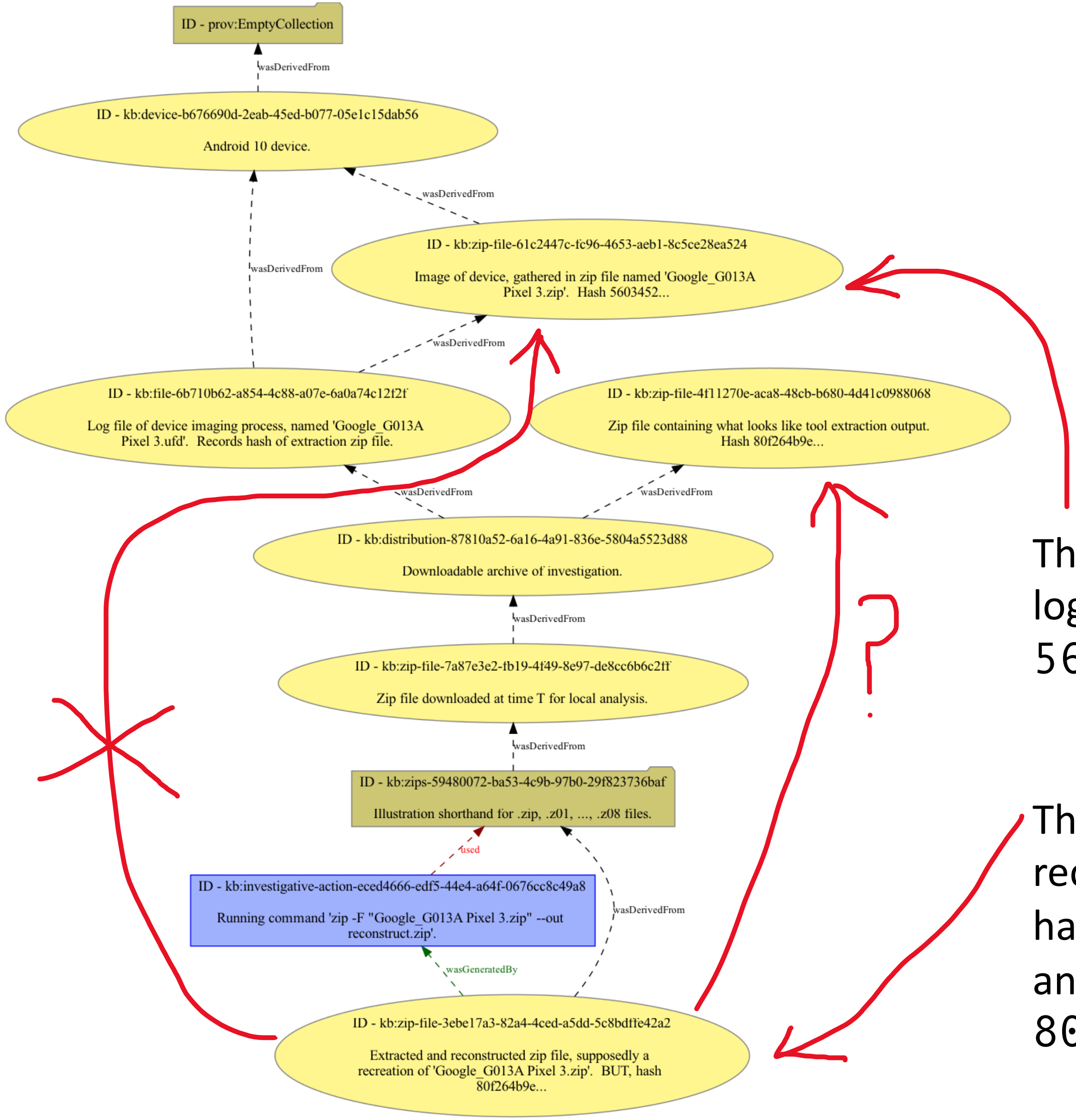
Provenance review can show discrepancies.

Digital Corpora "Android 10" dataset

This provenance was *sketched* from log: `Google_G013A Pixel 3.ufd`

A split zip in dataset reconstructs with an unrecognized hash.

Was this tool error?
Dataset author error?
Downloader error?



ID - prov:EmptyCollection

ID - kb:device-b676690d-2eab-45ed-b077-05e1c15dab56

Android 10 device.

ID - kb:zip-file-61c2447c-fc96-4653-aeb1-8c5ce28ea524

Image of device, gathered in zip file named 'Google_G013A Pixel 3.zip'. Hash 5603452...

ID - kb:file-6b710b62-a854-4c88-a07e-6a0a74c12f2f

Log file of device imaging process, named 'Google_G013A Pixel 3.ufd'. Records hash of extraction zip file.

ID - kb:zip-file-4f11270e-aca8-48cb-b680-4d41c0988068

Zip file containing what looks like tool extraction output. Hash 80f264b9e...

ID - kb:distribution-87810a52-6a16-4a91-836e-5804a5523d88

Downloadable archive of investigation.

ID - kb:zip-file-7a87e3e2-fb19-4f49-8e97-de8cc6b6c2ff

Zip file downloaded at time T for local analysis.

ID - kb:zips-59480072-ba53-4c9b-97b0-29f823736baf

Illustration shorthand for .zip, .z01, ..., .z08 files.

ID - kb:investigative-action-eced4666-edf5-44e4-a64f-0676cc8c49a8

Running command 'zip -F "Google_G013A Pixel 3.zip" --out reconstruct.zip'.

ID - kb:zip-file-3ebe17a3-82a4-4ced-a5dd-5c8bdffe42a2

Extracted and reconstructed zip file, supposedly a recreation of 'Google_G013A Pixel 3.zip'. BUT, hash 80f264b9e...

This file and hash is logged in tool output. `5603452...`

This file is reconstructed, but hash not logged anywhere. `80f264b...`

Background

Ontologies used in CASE-Corpora

Provenance

Usage of CASE-Corpora

CASE-Corpora is downloadable here:

https://github.com/casework/CASE-Corpora/

The total data graph is contained in data/kb-all.ttl.

CASE general-purpose, offline commands are available from PyPI:

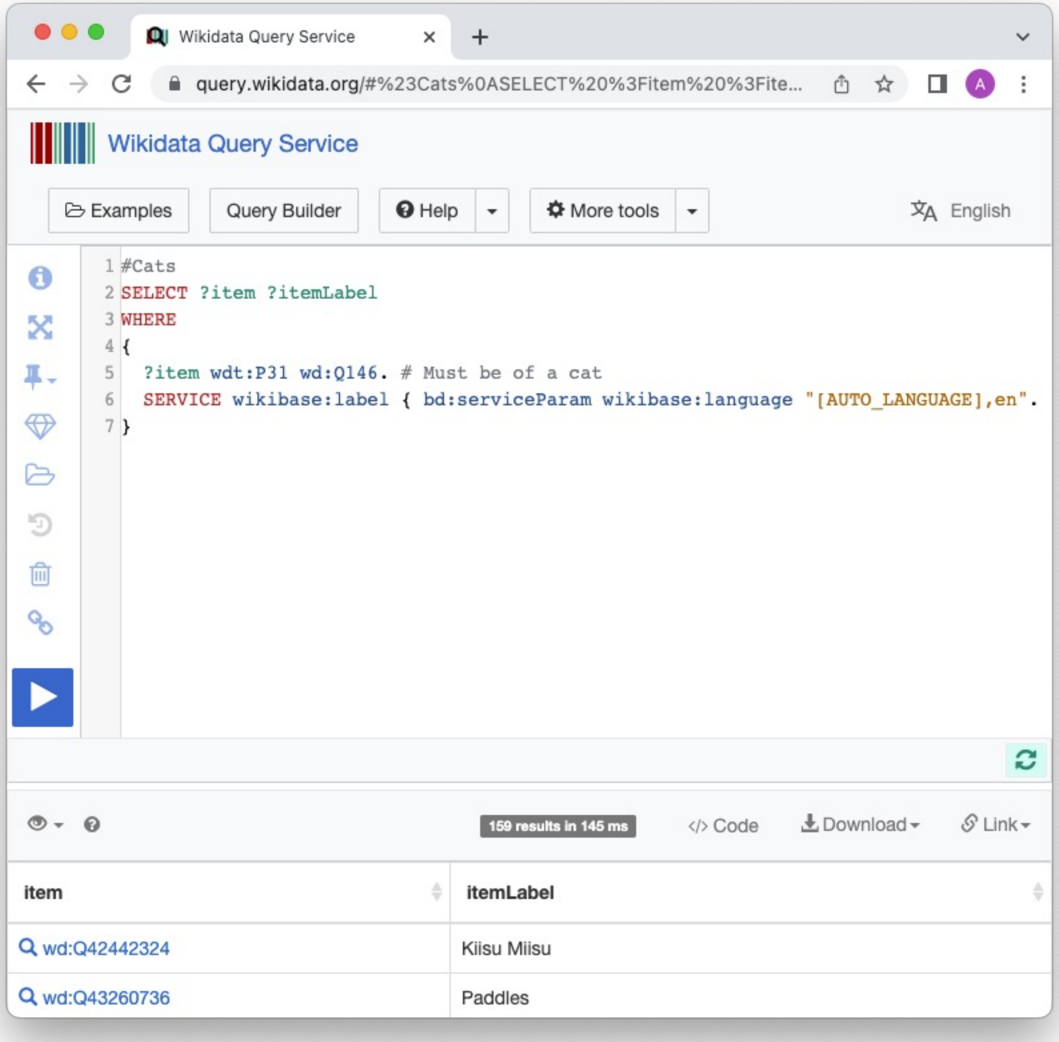
- `case_sparql_select` - run a query
- `case_validate` - validate conformance of used CASE concepts
- `case_prov_check` - review constructed provenance
- `case_prov_dot` - illustrate provenance

Running `pip install case-prov` makes these available.

How to explore CASE-Corpora

Comfort with the pattern-matching query language SPARQL is beneficial.

- Try https://query.wikidata.org/ to learn the basics. Try the example “Cats” (of Wikipedia).
- In CASE-Corpora, see the `reports/*.sparql` query files to see CASE and UCO vocabulary demonstrations.
- Start with a question you somewhat know the answer to, and then try expanding it.
- Documentation of CASE and UCO are at: https://ontology.caseontology.org/ https://ontology.unifiedcyberontology.org/



```
#Cats
SELECT ?item ?itemLabel
WHERE
{
  ?item wdt:P31 wd:Q146. # Must be of a cat
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".
}
```

159 results in 145 ms

item	itemLabel
wd:Q42442324	Kiisu Miisu
wd:Q43260736	Paddles

Conclusion

CASE-Corpora is an index of forensic metadata.

Immediate pragmatic value to the community is:

- Aggregating dataset existence
- Chain of custody details, for downloads and their analysis files

Other research value to the community is expanding the discovery language for relevant forensic datasets.

CASE-Corpora is intended to be a community project. Please consider helping the community highlight relevant data.

Contact

CASE-Corpora is downloadable here:

https://github.com/casework/CASE-Corpora/

Dataset requests, query forms - all manners of input welcome as Github Issues.

Joining CDO to improve CASE and UCO:
https://cyberdomainontology.org/contact.html

Other questions?

alexander.nelson@nist.gov