PID Graphs

Muggle Scientists Develop Harry Potter “Marauder’s Map” Technology

Luc Boruta — Thunken
luc@thunken.com — @thunkenizer

Bethesda, 2020/02/11

http://gph.is/XI8Wen

THUNKEN

Data Discovery?

- Discoverability of **documents/resources**
  - “Articles on...”
- Discoverability of **assertions**
  - “Articles that support the hypothesis that...”
- Discoverability of **attention/impact**
  - “Resources that discuss articles that...”

Information Repositories

Both specialized and generalist **repositories** give the illusion of uniformity, completeness, and persistence.

Information repositories are **not** closed systems.

The web is decentralized and volatile by nature.

On the Impermanence of (Textual) Metadata

Research objects should not be described using free text:

- Natural languages are **ambiguous**
  - 47k people in the US named John Smith
- Names, titles, addresses, etc. are **not persistent**
  - Univ. Paris 7, Univ. Paris Diderot, Univ. de Paris

https://cobaltmetrics.com/docs/page/searching-on-cobaltmetrics

The Marauder’s Map

https://www.youtube.com/watch?v=vNc43oKqQzg

The Marauder’s Map

“The Marauder's Map was a magical document that revealed all of Hogwarts School of Witchcraft and Wizardry. [It] showed **every inch of the grounds**, as well as **all the secret passages** [...]. It was also capable of **accurately identifying each person**, and was not fooled by [...] invisibility cloaks; even the Hogwarts ghosts were not exempt.”

https://harrypotter.fandom.com/wiki/Marauder%27s_Map

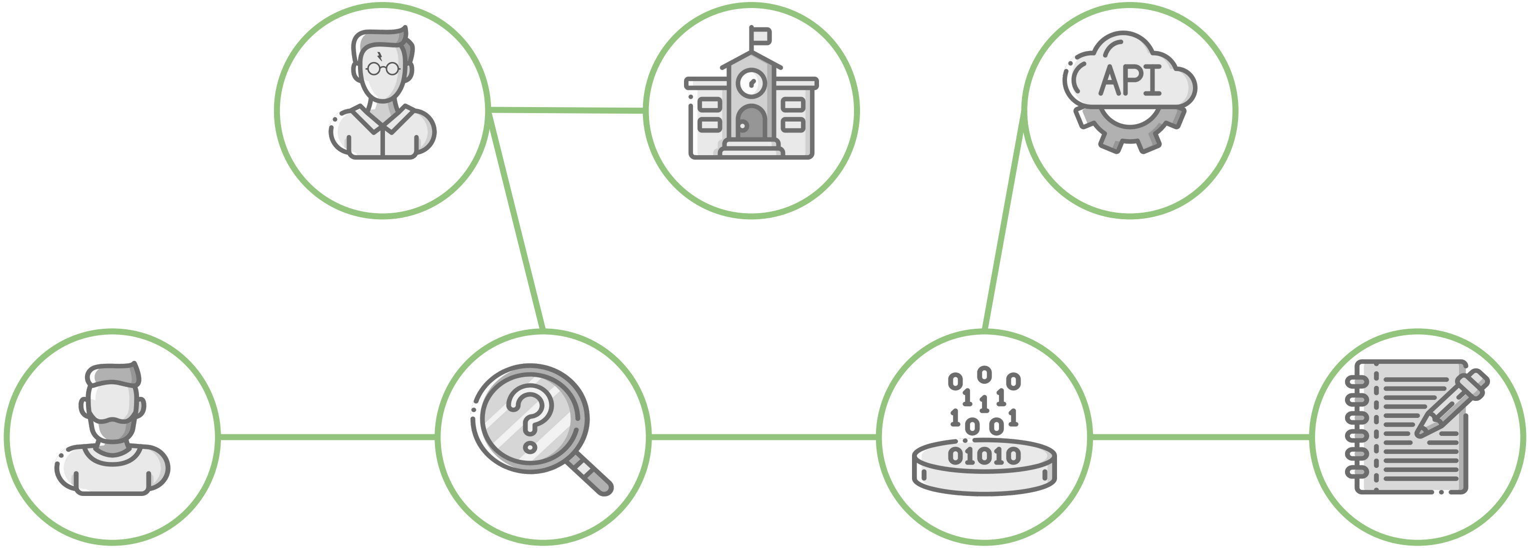
PID Graphs: Muggles' Marauder's Map

- **FREYA**'s PID graph
- **Cobaltmetrics**' knowledge graph (PID+URI)
- ResearchGraph
- SciGraph
- Wikidata

Cobaltmetrics × FREYA

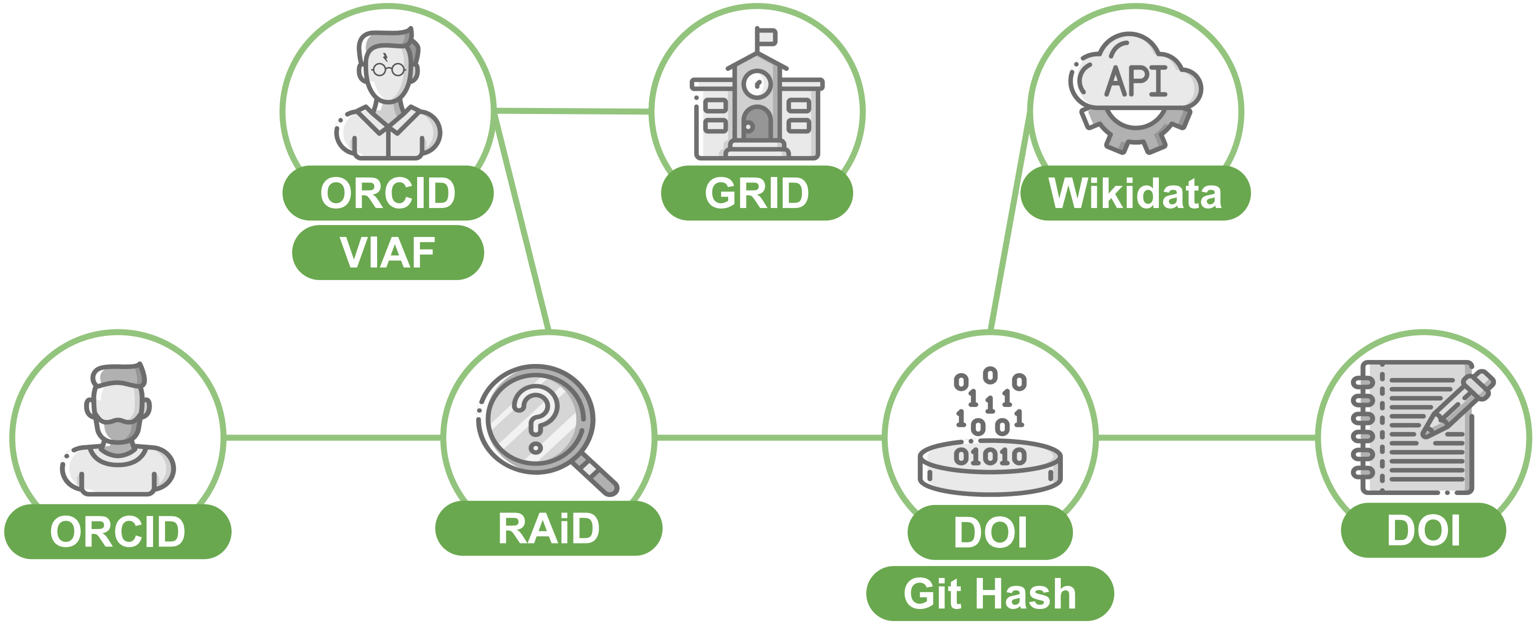
- FREYA: **PID graph**
  - Focus on scholarly entities
  - Many relations between entities
- Cobaltmetrics: **PID+URI graph**
  - Scholarly entities, but not only
  - Only one relation between entities ≈ owl:sameAs

FREYA's PID Graph



https://www.flaticon.com/authors/smashicons
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

FREYA’s PID Graph



https://www.flaticon.com/authors/smashicons
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

Better a URL Today Than a PID Tomorrow

The ideal identifier should be **persistent**, findable, accessible, interoperable, and reusable...

...we all **copy-paste from the address bar** of our browser.

The web is not FAIR (and will most likely never be) and **that is just fine**.

PIDs Are Contracts, Not Silver Bullets

There are **billions of documents**
that will never get DOIs or any other fancy PID:
old documents, grey literature, and **the rest of the web**.

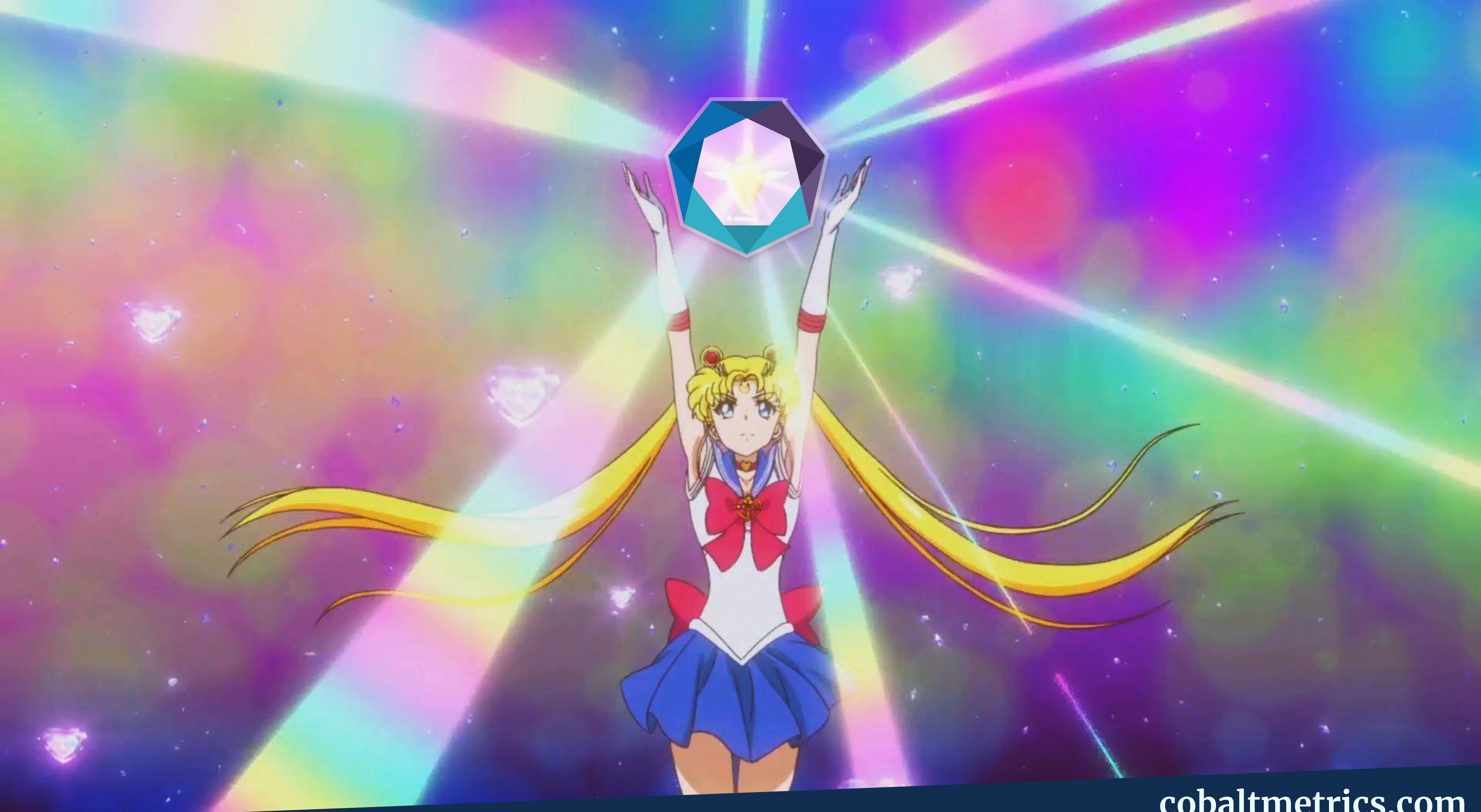
There are tons of documents with PIDs that are cited
with no mention of their PIDs.

Compact IDs vs. Good Old URLs

Cobaltmetrics' citation index (February 2019):

- HTTP+HTTPS+FTP: 256 million URLs (98%)
- Every other scheme: 4 million IDs

https://doi.org/10.6084/m9.figshare.9970265.v1

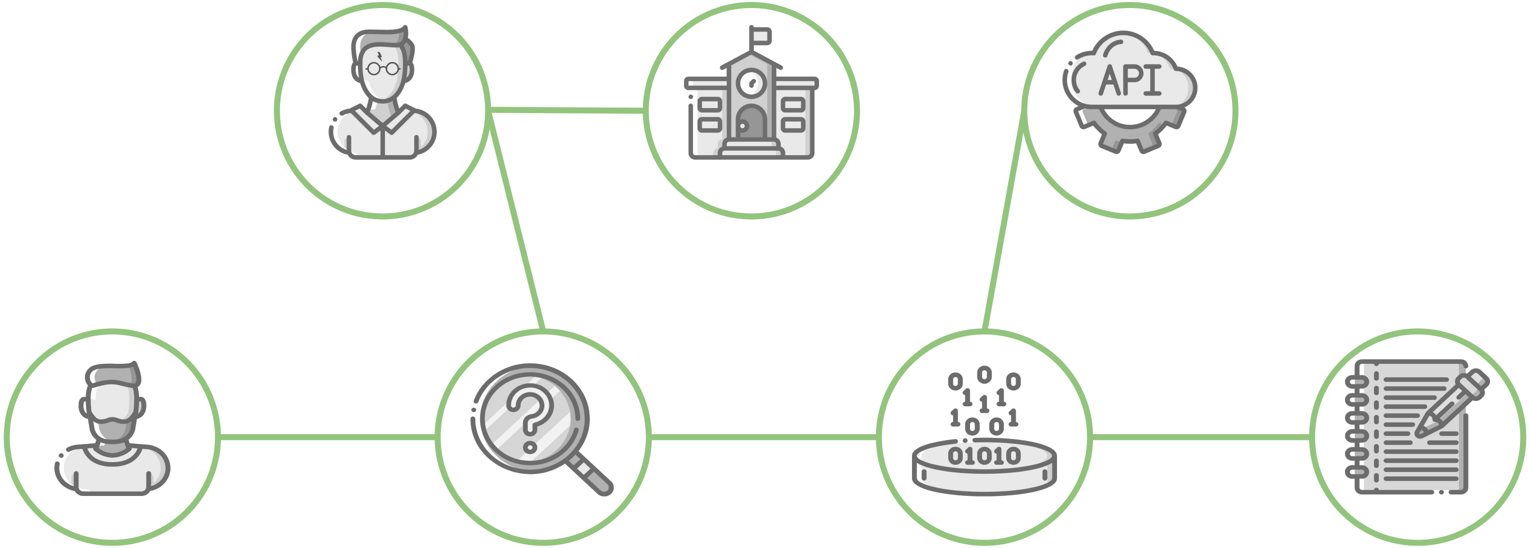


The Marauder’s Map

“The Marauder's Map was a magical document that revealed all of Hogwarts School of Witchcraft and Wizardry. [It] showed **every inch of the grounds**, as well as **all the secret passages** [...]. It was also capable of **accurately identifying each person**, and was **not fooled by [...] invisibility cloaks**; **even the Hogwarts ghosts were not exempt.**”

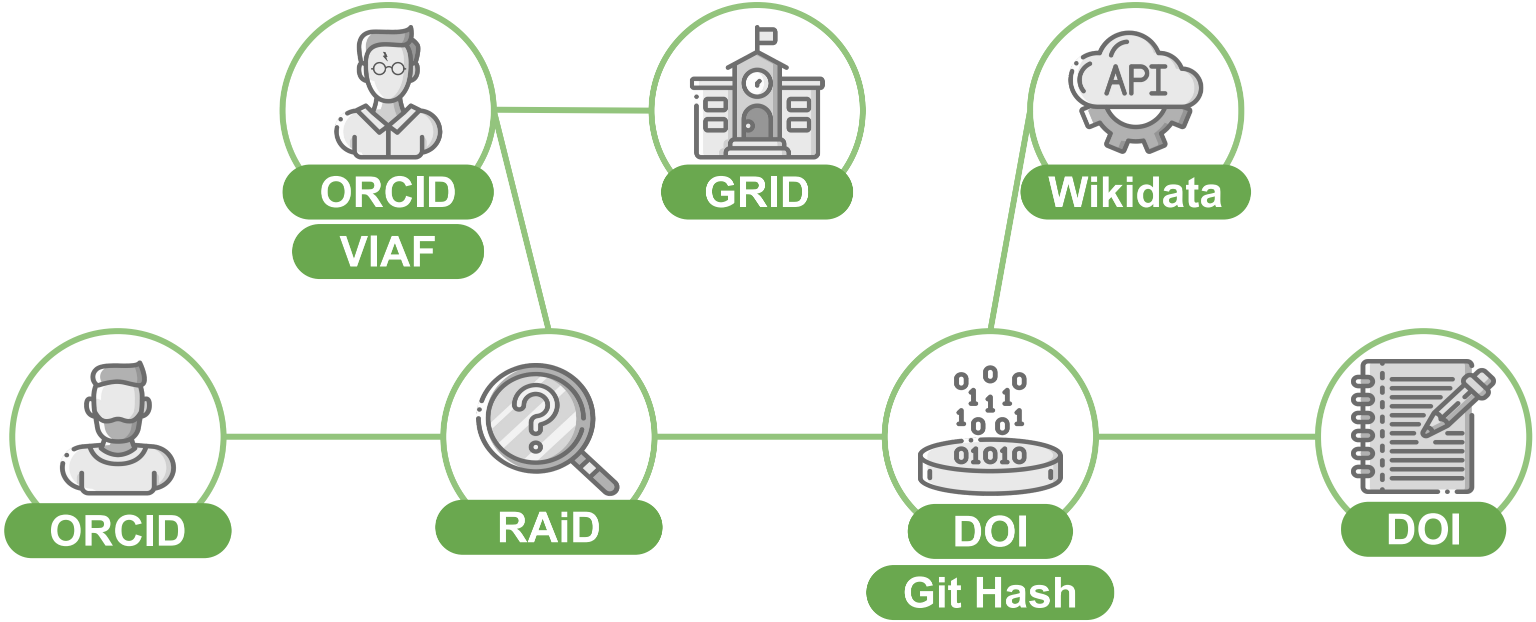
https://harrypotter.fandom.com/wiki/Marauder%27s_Map

FREYA's PID Graph



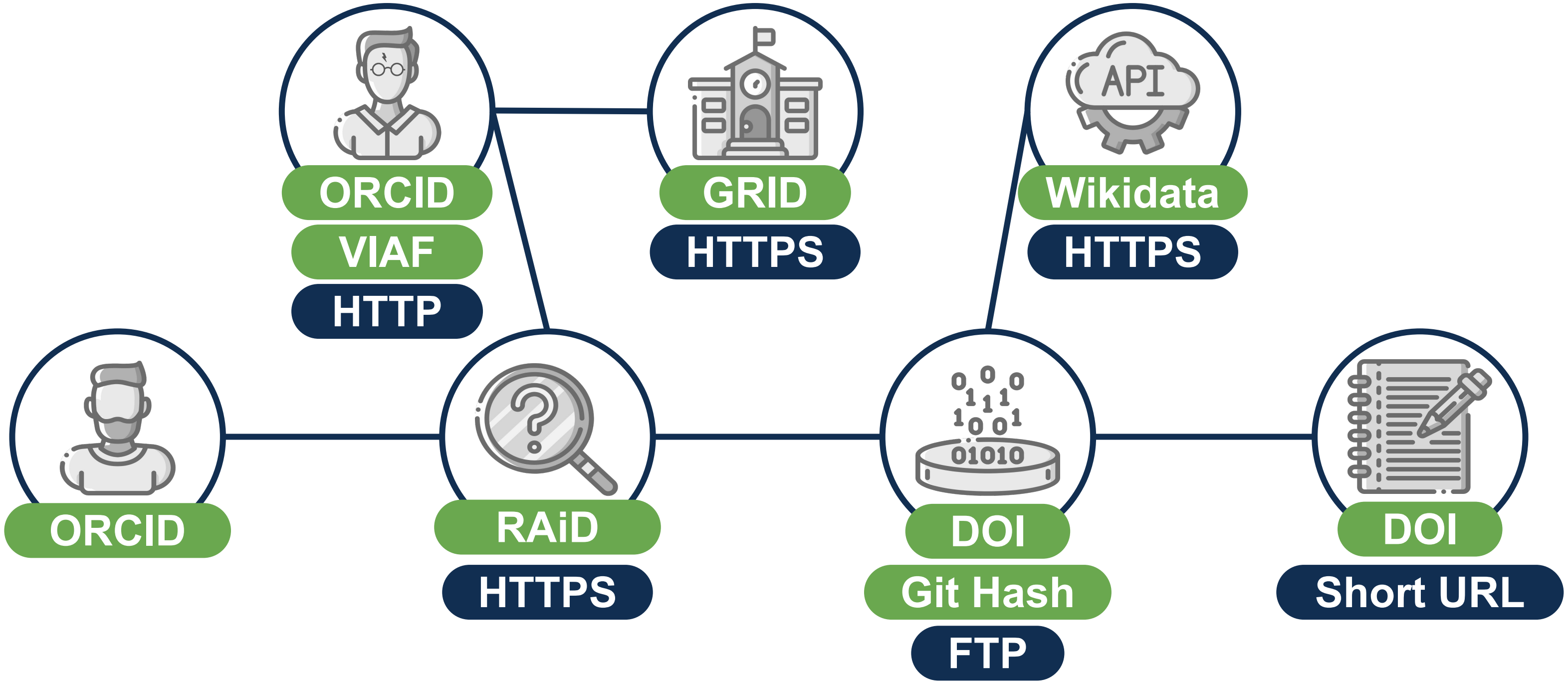
https://www.flaticon.com/authors/smashicons
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

FREYA’s PID Graph



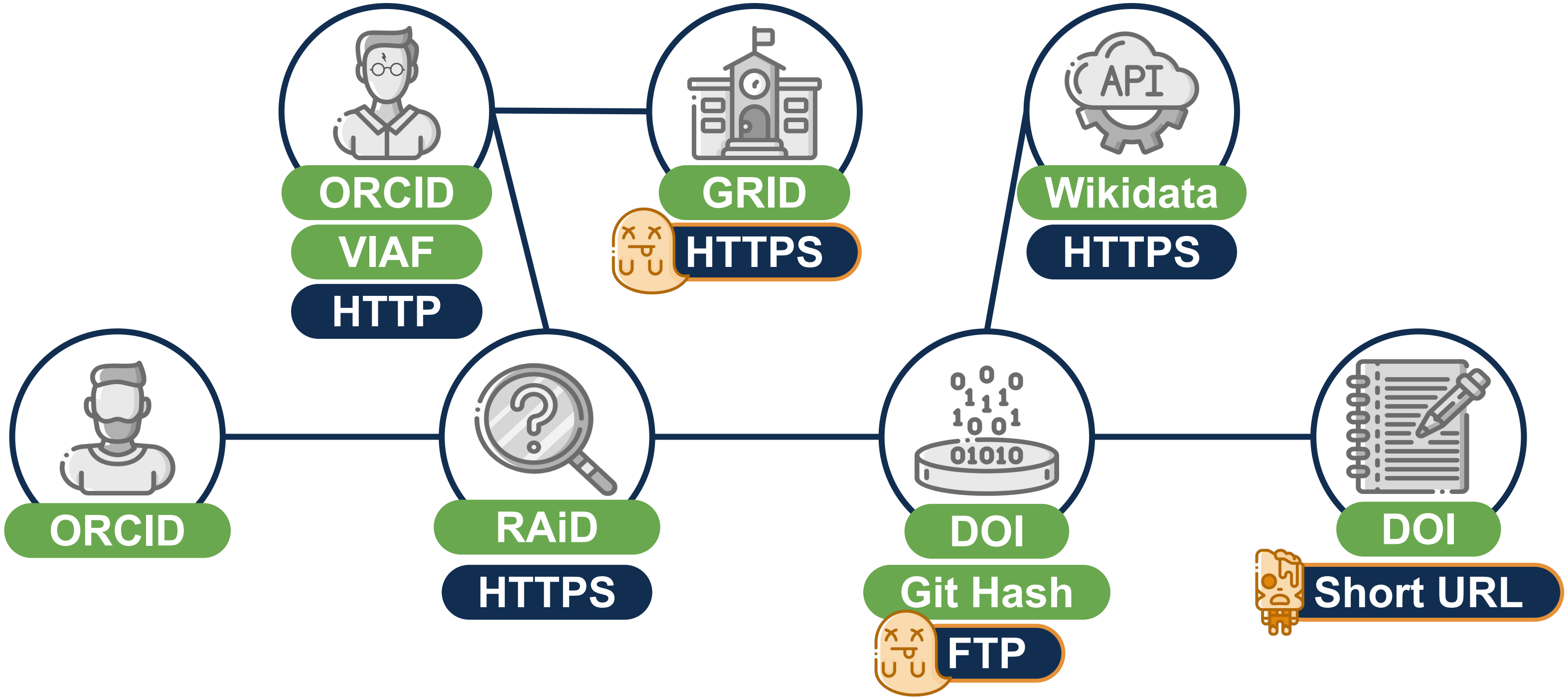
https://www.flaticon.com/authors/smashicons
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

Cobaltmetrics’ PID+URI Graph



https://cobaltmetrics.com/docs/page/uri-transmutation
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

Cobaltmetrics’ PID+URI Graph



https://cobaltmetrics.com/docs/page/uri-transmutation
https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

Enabling Data Discovery

- **Information retrieval task**, including but not limited to:
  - Searching for documents in a collection
  - Searching for information in a document
- How do we measure **performance**?
  - Online metrics: click-through rate, zero-result rate, etc.
  - Offline metrics: precision, recall, etc.

Enabling Data Discovery

- Development of a **common metadata model**?
    - Maybe, but let machines reconcile models
- **Expert curation** to enhance metadata?
    - Yes, reasoning systems are still limited
- Linking of research objects through **identifiers**?
    - Yes, PIDs foster discoverability for bots and humans

mischief
managed