The Wisdom of the Measurement Crowd: Building the Internet Yellow Pages a Knowledge Graph for the Internet

Romain Fontugne
IIJ Research Laboratory
Tokyo, Japan
iyp@ihr.live

Malte Tashiro
IIJ Research Laboratory / SOKENDAI
Tokyo, Japan

Raffaele Sommese
University of Twente
Enschede, The Netherlands

Mattijs Jonker
University of Twente
Enschede, The Netherlands

Zachary S. Bischof
Georgia Tech
Atlanta, GA, USA

Emile Aben
RIPE NCC
Amsterdam, The Netherlands

Abstract

The Internet measurement community has significantly advanced our understanding of the Internet by documenting its various components. Subsequent research often builds on these efforts, using previously published datasets. This process is fundamental for researchers, but a laborious task due to the diverse data formats, terminologies, and areas of expertise involved. Additionally, the time-consuming task of merging datasets is undertaken only if the expected benefits are worthwhile, posing a barrier to simple exploration and innovation. In this paper we present the Internet Yellow Pages (IYP), a knowledge graph for Internet resources. By leveraging the flexibility of graph databases and ontology-based data integration, we compile datasets (currently 46) from diverse and independent sources into a single harmonized database where the meaning of each entity and relationship is unequivocal. Using simple examples, we illustrate how IYP allows us to seamlessly navigate data coming from numerous underlying sources. As a result, IYP significantly reduces time to insight, which we demonstrate by reproducing two past studies and extending them by incorporating additional datasets available in IYP. Finally, we discuss how IYP can foster the sharing of datasets as it provides a universal platform for querying and describing data. This is a seminal effort to bootstrap what we envision as a community-driven project where dataset curation and ontology definitions evolve with the Internet measurement community.

CCS Concepts

• **Networks** → **Network measurement**; • **Information systems** → *Information integration*; *Graph-based database models.*

Keywords

Internet measurement, knowledge graph, Internet topology, DNS, routing

IMC '24, November 4–6, 2024, Madrid, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3688444

ACM Reference Format:
Romain Fontugne, Malte Tashiro, Raffaele Sommese, Mattijs Jonker, Zachary S. Bischof, and Emile Aben. 2024. The Wisdom of the Measurement Crowd: Building the Internet Yellow Pages a Knowledge Graph for the Internet. In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC '24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3646547.3688444

1 Introduction

The Internet measurement community generates a considerable amount of public datasets, which provide not only valuable insights into the current Internet but also a solid foundation for studying its evolution. The process of continuously building on these new insights is fundamental to research, yet it is a laborious task that involves convoluted processing and integration of diverse and often large datasets. Consequently, we find ourselves ill-equipped to promptly answer specific questions that fall within the realm of knowledge amassed by the measurement community. The community's efforts to encourage researchers to share their source code are a testimony to this challenge. To an extent, challenges in integrating Internet data are an impediment to innovation, as researchers tend to merge datasets only when they anticipate significant benefits from doing so.

Unfortunately, our community lacks a systematic approach for compiling and sharing its collective knowledge, an issue that has been tackled in other research fields. Two prime examples of this are the Gene Ontology, a large database representing the current understanding of gene functions [6, 32], and the DrugBank [20, 42], a collaborative database of drug properties, interactions, side effects, and indications. Both help facilitate research and new discoveries in biology and healthcare by combining the knowledge of numerous organizations.

Inspired by successful initiatives from other research fields, we introduce the Internet Yellow Pages (IYP), a knowledge graph for the Internet. The primary goal of IYP is to integrate various Internet datasets, providing researchers with quick and homogeneous access to the collective knowledge of the community. To achieve this, IYP relies on two main components: an ontology and a graph database. The ontology describes networking and Internet measurement terminologies, facilitating the integration of diverse datasets by unifying all datapoints and their relationships. The graph database leverages this data unification and provides IYP with the capabilities to efficiently store, query and analyze data from various datasets. Consequently, IYP constitutes a single harmonized database where the meaning of each entity and relationship is unequivocal. There are different ways to access IYP. We maintain a public instance of IYP and release database snapshots on a weekly basis.

To highlight the benefits of IYP, we reproduce two previous studies that focus on RPKI deployment [39] and DNS robustness [3]. The reproduction of these studies results in a set of IYP queries demonstrating how IYP can help researchers in navigating a multitude of diverse datasets and significantly reduce time to insight. In addition, by leveraging the availability of other datasets within IYP, we are able to quickly explore questions left as future work in previous studies, extending their findings.

Using IYP streamlines the way we explore Internet data and disseminate insights. We have found that IYP queries are remarkably efficient at expressing interesting discoveries and sharing insights. These queries typically consist of just a few lines of code. This makes them easily shareable while still precisely conveying patterns within the knowledge graph that may span multiple underlying datasets. Additionally, the results of queries can be refreshed every time the public instance of IYP is updated. We demonstrate this by providing two Jupyter notebooks containing the IYP queries from the two reproduced studies and code to query the public instance, resulting in weekly reports for these two studies.

In this paper, we outline our efforts to bootstrap IYP, integrating 46 datasets from 23 organizations. We envision the future IYP as a community-driven project where dataset curation and ontology definitions evolve in tandem with the Internet measurement community.

2 IYP design

The ultimate goal of IYP is to provide a universal platform to compile and unify the combined knowledge of the Internet measurement community, a frequent use case for knowledge graphs [25]. In a nutshell, knowledge graphs are databases that use a graph data structure model with built-in semantics, providing an efficient way to merge and integrate datasets. Each entity in the database is represented by a node and relationships between entities are modeled as links. The semantics of each node or link are described by an ontology.

An ontology focuses on the meaning of the elements that compose the database regardless of the database structure [34], as opposed to relational databases' schema describing the structure of a specific database. A relational database schema is typically tailored for a specific application and ensures that common queries provide very efficient data access. However, the integration of new data in a relational database requires updating its schema, a tedious task that may incur significant changes to the database and may break backward compatibility. For knowledge graphs, one could circumvent this problem by mimicking a graph model schema, but this has proven to produce schemas that are both hard and inefficient to query due to the numerous joins required to traverse the graph and the difficulty for relational databases to scale with the number of joins [21, 24, 36]. These challenges have spurred the development of graph databases and ontologies to streamline data modeling and integration while maintaining data interpretability. Similarly to popular large knowledge graphs [20, 37], IYP leverages these two technologies to maintain a large, coherent, and extensible database.

The rest of this section describes our process for constructing the IYP knowledge graph. First, we outline our guidelines for importing a dataset into IYP (Section 2.1). Based on these datasets, we propose an ontology for Internet resources and measurement terminologies (Section 2.2) and then explain how we use this ontology to build the IYP knowledge graph (Section 2.3).

2.1 Datasets

Our knowledge graph is based on open datasets documenting Internet resources. Most of these datasets are created by research groups and Internet companies. Table 1 lists multiple examples of such datasets. Since these datasets constitute the main substance of our knowledge graph, the selection of datasets is paramount. Integrating a large number of datasets to achieve a comprehensive knowledge graph is tempting, but can be counterproductive if it includes too much erroneous or stale data.

When adding new datasets, we try to be as inclusive as possible using the following five rules of thumb.

Format. For building IYP, fortunately, there is an abundance of well-structured datasets made available by the measurement community. Unlike many knowledge graphs, we do not require Natural Language Processing (NLP) techniques to extract data [2, 40]. All the datasets we import are well-structured, and most are available in a CSV or JSON format, making the extraction of values straightforward. Using NLP — which is a delicate and error-prone task — could be an interesting extension for IYP, but we first focus on making the best use of datasets already made available by the community.

Recognition. Any dataset is to some extent noisy and erroneous, which impacts the usability of the knowledge graph. When importing datasets, we do not attempt to sanitize them as doing so could introduce another source of noise or, worse, corrupt the original datasets. Instead, we aim to import reliable datasets that are generally recognized by the measurement community; we import only datasets that have either gone through peer review (e.g., ASdb [43] and AS hegemony values [12]), are commonly used by multiple independent research groups if there is no peer-reviewed equivalent (e.g., APNIC population estimate [8, 10, 13]), are facts from authoritative sources (e.g., AS names from RIPE NCC and RPKI data), or are observations made from renowned institutions (e.g., top domain names from Cloudflare and Cisco).

Freshness. Importing stale datasets could cause the collected knowledge to be out of sync, making interpretation of the data challenging. To avoid this problem, we integrate only datasets that are frequently updated. There is no hard limit on the datasets update frequency, as it depends on the nature of the data and the availability of similar datasets updated at a higher frequency. For example, currently the integrated dataset with the lowest update frequency is Stanford's AS classification, ASdb [43]. This dataset is updated every six months. However, since we do not expect AS business types to change frequently, we elected to include this dataset in IYP.

Originality. Redundant datasets are not a fundamental problem for IYP, but they could become a practical issue. Redundant data unnecessarily complicates the database and code maintenance. Hence,

Table 1: Example of datasets imported in IYP. The complete list currently covers 46 datasets (see Table 8).

Organization	Dataset	Description	Frequency	License
BGPKIT	pfx2asn	Originating AS per prefix seen in all RIS and RouteViews collectors.	Daily	BGPKIT AUA
CAIDA	ASRank	Ranking of ASes based on customer-cone.	Monthly	CAIDA AUA
Cloudflare Radar	top/ases (API)	ASes that queried a domain name the most, derived from Cloudflare 1.1.1.1 data.	-	CC BY-NC 4.0
IHR	AS Hegemony	Inter-dependence of ASes based on BGP data.	Daily	CC BY-NC 4.0
OpenINTEL	tranco1m	DNS resolution for Tranco Top 1M domain names.	Daily	CC BY-NC 4.0
PCH	Routing snapshots	BGP data collected from PCH.	Daily	CC BY-NC-SA 3.0
PeeringDB	ix (API)	Information related to IXPs and their members.	-	PeeringDB AUA
Stanford	ASdb	Classification of ASes by business type.	6-month	-

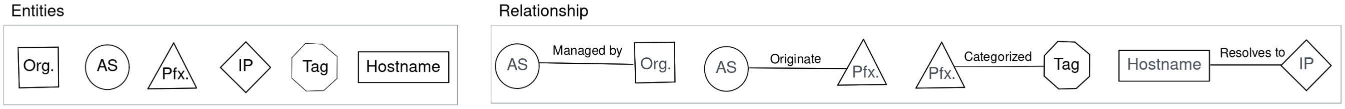
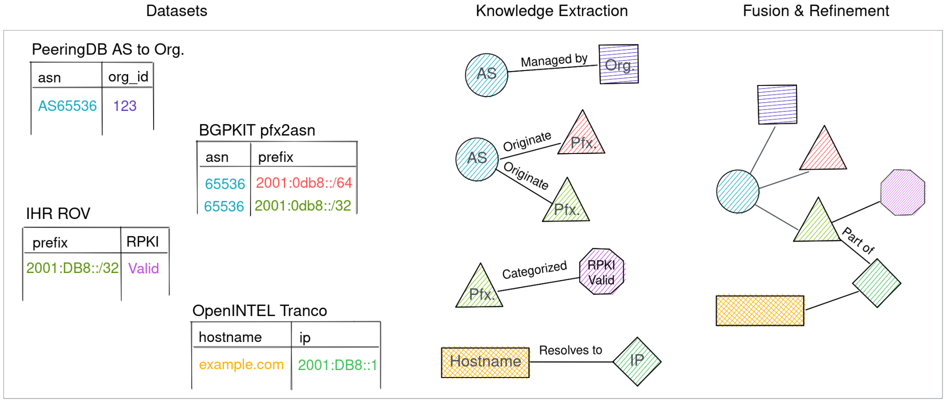


Figure 1: Example ontology for networking data. Entities describe the types of the nodes in the knowledge graph and relationships describe how two entities relate to each other.



PeeringDB AS to Org.

asn	org_id
AS65536	123

BGPKIT pfx2asn

asn	prefix
65536	2001:0db8::/64
65536	2001:0db8::/32

IHR ROV

prefix	RPKI
2001:DB8::/32	Valid

OpenINTEL Tranco

hostname	ip
example.com	2001:DB8::1

Figure 2: Overview of the knowledge graph construction steps. Starting from multiple heterogeneous datasets (Datasets column). Data is extracted from datasets and formalized using the ontology of Figure 1 (Knowledge Extraction column). Then datasets are combined and additional refinements are performed (Fusion & Refinement column).

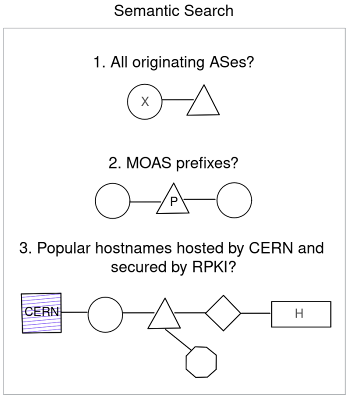


Figure 3: Example searches looking only for patterns in the knowledge graph (1, 2) and a pattern including a specific node (3).

we avoid importing datasets that are inferred from the same data source. For example, CAIDA, IHR, and BGPKIT prefix-to-AS mappings [7, 9, 18] are all derived from RouteViews and RIPE RIS BGP data. Since the processing for these datasets is straightforward we expect little differences between them. Currently, we import only BGPKIT's prefix-to-AS mapping [7] given that it is the only one that uses all RIS and RouteViews collectors and is updated daily.

Shareable. IYP is a public knowledge graph, so we include only datasets with a license that permits us to reshare the data. As discussed in Section 3.1, snapshots of IYP are made public so that one can download and run the whole database locally to import and analyze confidential data. Since each dataset is released with a different license, we provide links to the licenses of all datasets[1] to comply with their terms of use and so that users can more easily find which dataset is appropriate for their purposes.

The list of datasets supported in IYP continues to grow. Currently, IYP includes 46 datasets from 23 organizations (see Table 8). The full list of datasets is available on IYP's documentation page.[2]

2.2 Ontology

The benefits of knowledge graphs come from the integration of formal semantics, also known as ontology, and the data. The ontology describes all entities and their relationships found in the datasets.

[1] https://github.com/InternetHealthReport/internet-yellow-pages/blob/main/ACKNOWLEDGMENTS.md.

[2] https://github.com/InternetHealthReport/internet-yellow-pages/blob/main/documentation/data-sources.md.

It is composed of terms and definitions that derive directly from the domain-specific vocabulary and shared knowledge. This is the glue between the data providers, IYP, and the users, making sure that all stakeholders can easily and precisely interpret the data.

For IYP, we aim to build an ontology that models common networking knowledge. For example:

- *An AS is managed by an organization*;
- *An AS originates a prefix in BGP*;
- *A hostname resolves to an IP address.*

The proposed ontology describes entities, relationships, and properties. Entities delineate the different fields found in datasets and consequently the type of nodes in the knowledge graph (see Entities in Figure 1). These are often network resources, such as *Hostname, IP address, AS, Prefix* but can include other types of information (e.g., *Organization*). IYP currently has 24 distinct entities (see Table 6 in the Appendix). This number is slowly growing as we add more datasets with new fields to IYP. The up-to-date list of entities is available online.[3]

Relationships describe how pairs of entities are related to each other (e.g., *resolves to, originates*, and *managed by*). These are usually implicit in the imported dataset. For instance, a prefix-to-AS mappings dataset is usually expressed as a list of *<prefix, ASN>* pairs meaning that a prefix is originated by a certain AS. In the knowledge graph, these relationships are explicit and clearly described by the ontology. IYP currently has 24 distinct types of relationships shown in Table 7, the up-to-date list of relationships is documented online.[4]

Properties are attributes that characterize a specific entity or relationship. For example, each AS has an *asn* property that uniquely identifies an AS. Generally speaking, entities (nodes) have a small number of properties that are used for identification. On the other hand, relationships may have quite a large number of properties. Fields in datasets that are not described by the ontology entities and relationships are modeled as properties. For example, PeeringDB provides detailed information about IXP members. The membership is modeled by a *member of* relationship between AS and IXP nodes, but circumstantial details (e.g., peering policy or typical traffic levels) are modeled as properties of the relationships.

Furthermore, we systematically add relationship properties in order to document the origin of the data. While importing a dataset, each time we create a new link we annotate the link with the following properties:

Organization: The name of the organization that provides and maintains the dataset. This property can be used to select or discard data from a particular organization.

Dataset name: A unique name for the original dataset. This property is extremely important as it enables tracking the exact source of the data, conveying the limitations and appropriate use of the data. It can also be used to select or discard a specific dataset.

Information URL: If available, this provides a link to a human-readable description of the original dataset.

Dataset URL: The URL from which we retrieved the dataset.

Modification time: If available, the time when the dataset was last modified.

Fetch time: The time at which we imported the dataset into IYP.

These properties are particularly important for tracking the origin of datapoints in the knowledge graph and, selecting or filtering certain datasets for analysis.

2.3 Graph construction

Using the above ontology, the knowledge graph construction is straightforward. We identify entities and relationships in selected datasets and model each dataset as a sub-graph (Knowledge Extraction in Figure 2). The whole graph falls into place by aligning the same entities found in different datasets, and is finalized by adding simple relationships from common networking knowledge (Fusion and Refinement in Figure 2).

For each dataset, we implement a custom script that maps the dataset schema to the entities and relationships of the IYP ontology. Although this is a trivial step for many datasets, we should cautiously verify that the meaning of the schema matches the IYP ontology and avoid ambiguities. For fields that are not yet covered by the ontology, we either extend the ontology, or we store them as relationship properties (not described in IYP ontology). Hence users that are acquainted with the original dataset can still access the whole dataset.

We avoid creating duplicate nodes (i.e., nodes representing the same entities) by enforcing canonical forms of certain identifiers (IP address, IP prefix, ASN, country code). For example, in Figure 2 the IHR dataset contains the 2001:DB8::/32 prefix and the BGPKIT dataset contains the 2001:0db8::/32 prefix. These are both representing the same IPv6 prefix written differently, which would create two nodes for the same prefix. To avoid duplicates we translate these values to their canonical form, hence both entries will be modeled as a prefix node 2001:db8::/32. This ensures that a node in the knowledge graph uniquely identifies the same network resource that may appear in various forms in multiple datasets.

In contrast, for relationships we leverage the possibility to have the semantically same link multiple times but obtained from different datasets. For example, we may have two datasets that indicate that a prefix is originated from a certain origin AS. In IYP this is modeled as two distinct links that connect the same nodes. We can differentiate these two links by looking at their *Dataset name* property (see Section 2.2). Consequently, an imported dataset is entirely mapped to a set of links in the knowledge graph. This allows us to easily select or discard a certain dataset. It is also very handy to compare two similar datasets.

Apart from the translation of identifiers to their canonical form, we purposely make no changes to the datasets and import them as-is. We are aware that certain datasets are erroneous and accept that no dataset is perfect. Pretending to clean up all datasets would be foolish, as we could accidentally introduce new errors, give the false impression that the database is perfect, and confuse users on the methodology and limitations for each dataset. Instead, we encourage anyone to report errors found in IYP directly to the data providers, so that the original dataset can be fixed, which

[3] https://github.com/InternetHealthReport/internet-yellow-pages/blob/main/documentation/node_types.md.

[4] https://github.com/InternetHealthReport/internet-yellow-pages/blob/main/documentation/relationship_types.md.

benefits the community at large. In addition, users that understand the limitations of the original dataset can rely on the fidelity of our knowledge graph to the original data. Users can also decide to discard a specific undesirable dataset when querying the knowledge database.

The final touch to the IYP knowledge graph is the addition of common knowledge that is usually implicit in datasets (Refinement in Figure 2). After all datasets have been imported, we add a property to all IP address and prefix nodes to explicitly describe their address family (IPv4 or IPv6). We also link each IP address node to the prefix node corresponding to the longest prefix match, and each prefix to its covering prefix, so that one can easily navigate between IP addresses and prefixes. Similarly we also link URL nodes to corresponding hostname nodes. Finally, we make sure that every country node has a two- and three-letter country code as well as a common name. All these additions are safe to implement and simplify queries.

3 IYP in practice

This section provides implementation details, describes how we share IYP, and ends with an illustrative example and simple queries.

3.1 Implementation

IYP is based on Neo4j, a native graph database that has an outstanding adoption in multiple domains [31, 33], thanks to its powerful querying language, Cypher [14], and active community. We publish[5] all our scripts for downloading, parsing, and merging datasets. The crawler for each dataset is implemented independently, hence one can easily build a customized knowledge graph.

As shown in the examples below, IYP entities and relationships are written following the Neo4j naming convention.[6] Both are preceded by a colon (:) to indicate that they refer to a node or relationship type. Entities are written in camel-case, beginning with an upper-case character (e.g., :DomainName). Relationships are written in upper-case, using an underscore to separate words (e.g., :RESOLVES_TO).

IYP is made publicly available in two different ways. We deploy a public, read-only Neo4j instance that can be queried by anyone.[7] It has two main interfaces, a graphical web UI and an API. We also provide a Neo4j docker image and weekly snapshots of the database so that users can run their own instance of IYP and locally conduct intensive analysis or experiments with IYP. A local instance is especially suitable for integrating and analyzing confidential data with IYP. We are currently building a complete IYP knowledge graph four times a month (on the 1st, 8th, 15th, and 22nd of the month) in order to provide frequently updated snapshots. Each snapshot is about 4GB compressed and takes up to 40GB when loaded in Neo4j. Apart from disk space the requirements for running IYP are very low — our public instance runs in a virtual machine with 4 vCPU and 8GB of RAM.

3.2 Sneak peek

Figure 4 is an example taken from IYP demonstrating the seamless and intuitive fusion of datasets that IYP enables. The visualization is produced by Neo4j's web GUI which also allows interaction with the graph. One should be able to read and understand the meaning of the graph only with common networking knowledge.

This example (Figure 4) is obtained by following a few relationships starting from the yellow top-left node which represents the DNS zone cut [16] for the *nytimes.com* DomainName. The RANK relationship going to the *Tranco Top 1M* node shows that this domain name is ranked in the Tranco ranking. The exact rank is encoded in a property of the RANK relationship not shown in Figure 4.

The top branch starting from the PART_OF relationship connected to the *nytimes.com* HostName node represents details about the zone apex [16] (resolvable FQDN at the origin of the zone). The TARGET relationship to that node indicates that a ping Atlas measurement is targeting this hostname and one of the Atlas probes participating in this measurement is shown next to it. The probe node properties (not shown in Figure 4) contain all information provided by Atlas (e.g., probe ID, tags, type of probe). The RESOLVES_TO relationship from the *nytimes.com* hostname (pink node) points to the resolved IP address that is part of a prefix that is originated by AS54113 — named Fastly — and categorized as a content delivery network (by BGP.Tools as indicated by a property of the CATEGORIZED link but not shown in Figure 4). The prefix is also categorized as being anycast (by BGP.Tools) and RPKI Valid (by IHR). The RPKI ROA for this prefix is shown by the ROUTE_ORIGIN_AUTHORIZATION relationship between the prefix and the AS.

Going back to the *nytimes.com* DomainName node (yellow node) and following the middle branch starting from the MANAGED_BY relationship, we find one of the authoritative DNS nameservers to which this domain name is delegated. This nameserver resolves to an IPv6 address that is part of an IRR Valid prefix, announced by AS16509, named Amazon-02.

Finally the last branch below the *nytimes.com* DomainName node (yellow node) and starting from the relationship QUERIED_FROM indicates that AS7018 (AT&T a Tier1 network as shown by the NAME and CATEGORIZED relationships) is frequently querying this domain name (data provided by Cloudflare Radar). We also observe that AS7018 is peering with the RIS BGP collector rrc00.

The main purpose here is to illustrate the benefits of the knowledge graph to unify and ease the interpretability of the data provided by the 13 datasets underlying this example. The availability of numerous datasets in IYP is remarkably helpful in uncovering orthogonal datasets that may not be initially considered but prove to be insightful. For example when studying a specific set of IP addresses and discovering the existence of Atlas measurements targeting these IP addresses, or finding that the corresponding prefix is RPKI invalid or anycasted. The example is not comprehensive, there are more nameservers, IP addresses, prefixes, and ASes that relate to the *nytimes.com* domain name. There are also other datasets that could be demoed and uncovered here — for example IXPs from PeeringDB, or related resources in delegated files — but we have limited the number of relationships to keep the example simple.

[5] https://github.com/InternetHealthReport/internet-yellow-pages.

[6] https://neo4j.com/docs/cypher-manual/current/syntax/naming.

[7] https://iyp.iijlab.net.

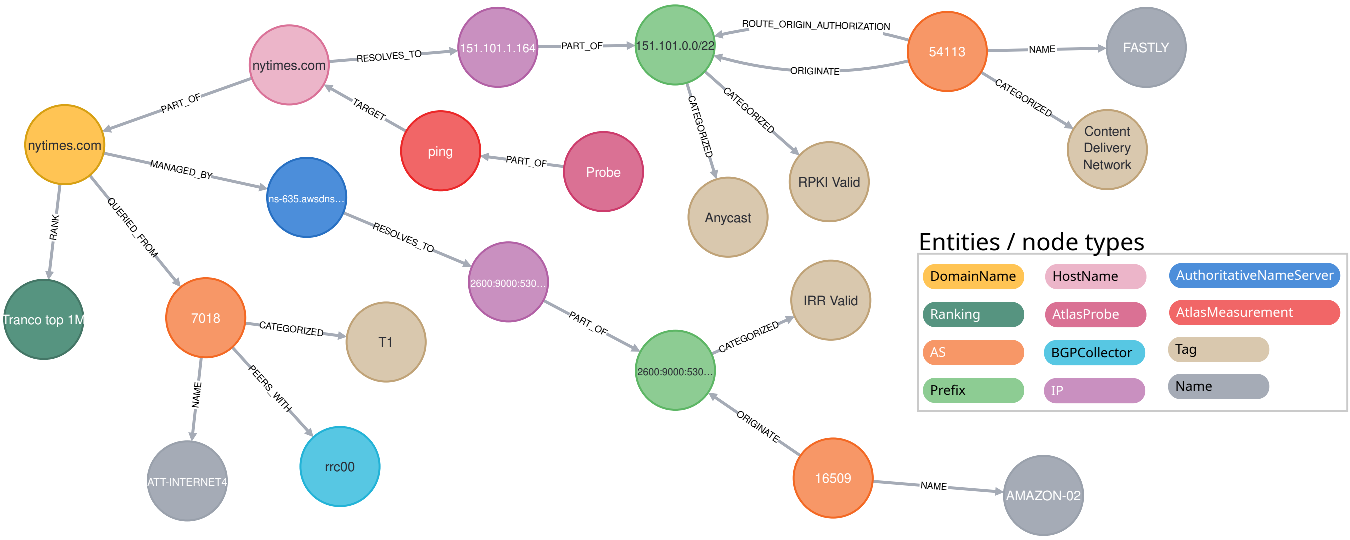


Figure 4: Example of data related to the *nytimes.com* domain name (DNS zone cut, left yellow node) and corresponding hostname (resolvable FQDN, pink node labelled *nytimes.com*). The displayed data is obtained from 13 different datasets illustrating data unification in knowledge graphs. Notice that the meaning of certain relationships (e.g., PART_OF) is based on the connected entities.

3.3 Semantic search

The above example illustrates an exploration of the knowledge graph starting from a certain node in the graph. In order to automate data analysis one can systematically query the graph. Knowledge graphs allow for semantic search, which expresses the user intention rather than the traditional literal match of query keywords. For example, a user searching for AS7018 in the dataset can query for an AS node with the ASN property set to 7018. This is radically different from looking for all instances of '7018' in the database.

A user can search for meaningful patterns in the graph. Going back to the simple example of Figure 2, a user can find all ASes originating prefixes by looking for the pattern where an AS and a prefix node are linked by a *originate* relationship (see 1 in Figure 3). Another example, is to search for multiple origin AS (MOAS) prefixes which consists of looking for prefixes that are connected to two different ASes in the graph (see 2 in Figure 3). Notice that these patterns contain no lexical elements. There is no keyword to look for; they are expressing the user intention exclusively with terms from the ontology.

A tutorial on Cypher, Neo4j's querying language, is beyond the scope of this paper, but to showcase real IYP queries we provide Cypher queries for the two above examples (1 and 2 from Figure 3) in Listing 1 and 2.

```
// Select ASes originating prefixes
MATCH (x:AS)-[:ORIGINATE]-(:Prefix)
// Return the AS's ASN
RETURN DISTINCT x.asn
```

Listing 1: Cypher query to find all originating ASes in IYP (see 1 in Figure 3). Lines starting with // are comments.

```
// Find Prefixes with two originating ASes
MATCH (x:AS)-[:ORIGINATE]-(p:Prefix)-[:ORIGINATE]-(y:AS)
// Make sure that the ASNs of the two ASes are different
WHERE x.asn <> y.asn
// Return the prefix attribute of the Prefix node
RETURN DISTINCT p.prefix
```

Listing 2: Cypher query to find all Multiple Origin AS (MOAS) prefixes in IYP (see 2 in Figure 3). Lines starting with // are comments.

The MATCH clause is followed by a search pattern. The pattern is composed of nodes and relationships indicated respectively by parenthesis and squared brackets surrounded by dashes. For example, this is an AS node, (:AS), and this is a originate relationship, -[:ORIGINATE]-. Consequently the pattern of Listing 1, (x:AS)-[:ORIGINATE]-(:Prefix), stands for *an AS originates a prefix*. The x in the AS node is a variable that we can refer to later in the query. We use it to fetch the ASN property of the AS node in the return clause. Listing 2 shows that a WHERE clause can be used to specify constraints.

The third search in Figure 3 is a more complex example where the pattern starts from a specific node and is branching. We refer interested readers to Listing 3 for the corresponding Cypher query.

The above queries demonstrate how IYP conceals the superfluous details of underlying datasets and focuses only on data semantics. We demonstrate this further in the next section by reproducing past studies with a set of IYP queries.

```
// Find RPKI valid prefixes managed by CERN
MATCH (org:Organization)-[:MANAGED_BY]-(:AS)-[:ORIGINATE]-(pfx:Prefix)-[:CATEGORIZED]-(:Tag {label:'RPKI Valid'})
WHERE org.name = 'CERN'
// Find find popular hostnames in these prefixes (refered as pfx)
MATCH (pfx)-[:PART_OF]-(:IP)-[:RESOLVES_TO {reference_name:'openintel.tranco1m'}]-(h:HostName)
// Return the hostname's name
RETURN distinct h.name
```

Listing 3: Cypher query to find the popular hostnames corresponding to prefixes originated by ASes managed by CERN and that are RPKI valid.

Table 2: Comparison between original RiPKI results and reproduced results with IYP.

	RPKI Invalid	RPKI covered	Top 100k	Bottom 100k	CDN
RiPKI (2015)	0.09%	6%	4%	5.5%	0.9%
IYP (2024)	0.12%	52.2%	55.2%	61.5%	68.4%

4 Time to insight

To illustrate the ability of IYP to quickly provide insights about the Internet, we revisit two studies: *RiPKI* [39], a study of the deployment of RPKI in 2015; and *DNS Robustness* [3], a study of best practices and consolidation of the DNS infrastructure.

These two studies are typical examples of insights that users can obtain using IYP. We selected these two because they are related, both providing insights into popular domain names, and are over six years old. Therefore, we take this opportunity to refresh these results and expand upon them with additional datasets. However, the main objective of this section is to illustrate how IYP helps in reducing time to insight. We only reproduce the main results of these two papers, it is not our intention to revise the detailed interpretation and recommendations provided by both papers.

We reproduce both studies using the same procedure. First, we identify the key results reported by the study and the meaning of these results. Then, we craft an IYP query for each of the key results. To avoid overcomplicating queries with aggregation functions, some queries only extract the required data, which we then aggregate with a few lines of Python code. The queries are executed on a recent snapshot of IYP (2024/05/01). For each study we provide a Jupyter notebook[8] containing the queries and code to reproduce these results. Executing the notebooks refreshes the results using the latest data available in the IYP public instance. By publicly sharing these notebooks, we hope that researchers can reuse provided IYP queries and build on top of them.

One caveat is that both papers defined popular domain names using the Alexa Top 1 million list, which was retired in 2022. Instead we use the Tranco Top 1 million list [22], a popular alternative to Alexa that is designed to improve agreement across different lists and stability over time. The DNS Robustness study is based on zone files that are not available in IYP. Instead we use another DNS dataset provided by OpenINTEL that is available in IYP. The other datasets employed in these studies are similar to the ones integrated into IYP.

4.1 Reproduction: RiPKI

4.1.1 Summary of original study. The RiPKI study [39] quantifies the number of popular domain names that are protected by RPKI. RPKI is a public-key infrastructure designed to help secure Internet inter-domain routing and limit the impact of BGP hijacks and operational mishaps. This study relies on four datasets collected in 2014 and 2015:

- Alexa Top 1 million (not available in IYP, substituted by Tranco Top 1M).
- Domain name address resolution using Google DNS (similar to the OpenINTEL DNS data available in IYP).
- RIPE RIS data to map IP to prefixes and originating ASes (available in IYP).
- RPKI data (available in IYP).

The RiPKI approach is to retrieve the IP addresses for the Alexa Top 1 million domains by querying Google's public resolver. Then it finds the IP prefix and originating ASes for these IP addresses in RIPE RIS's BGP data. And finally, it reports the percentage of prefix/origin AS pairs that are registered in RPKI.

4.1.2 Original results: The tragic story of RPKI. The RiPKI paper makes several key observations (summarized in Table 2). As expected, only a very low percentage (0.09%) of prefix/origin AS pairs are invalid (i.e., the origin AS in BGP is different from the ones in RPKI). For the whole Alexa Top 1M list, they found that only 6% of the prefixes are covered by RPKI (valid and invalid). And surprisingly, prefixes for the bottom 100k domains are better covered by RPKI than the prefixes for the Top 100k domains, respectively 5.5% and 4.0%. The paper justifies that this is due to the very low adoption of RPKI (0.9%) by CDNs in 2015.

4.1.3 Updated results: The happier story of RPKI. Using IYP we update these results and report the status of RPKI for popular domains in 2024. As mentioned above here we use the Tranco list instead of Alexa and the DNS resolution data is provided by OpenINTEL.

The queries for this study search for patterns similar to the top branch of the graph shown in Figure 4. We look for domain names ranked in Tranco, retrieve their IP addresses, the corresponding routed prefixes, and check their RPKI status. For instance, we provide the query to find the number of RPKI invalid prefixes for domain names in Tranco in Listing 4.

Using IYP we found that the percentage of invalid announcements has stayed very low (0.12%), and that 75% of them are due to a wrong maximum prefix length in ROAs. However, we obtain significantly different results for the overall RPKI coverage. We find that more than half (52.2%) of the prefixes for Tranco Top 1M are covered by RPKI, almost 9 times more than in 2015. The percentage

[8] https://github.com/InternetHealthReport/iyp-notebooks.

```
// Resolve IP addresses from the Tranco Top 1 million list and count the number of RPKI invalid prefixes
MATCH (:Ranking {name:'Tranco top 1M'})-[:RANK]-(d:DomainName)-[:PART_OF]-(h:HostName)-[:RESOLVES_TO]-(:IP)-[:PART_OF]-
        (pfx:Prefix)-[:CATEGORIZED]-(t:Tag)
WHERE d.name = h.name AND t.label STARTS WITH 'RPKI Invalid'
RETURN COUNT(DISTINCT pfx)
```

Listing 4: Find RPKI invalid prefixes for domain names in Tranco list. The RPKI status is provided as a tag that is either 'RPKI Invalid' or 'RPKI Invalid,more specific' hence the STARTS WITH condition in the WHERE clause.

for CDNs, the Top 100k, and bottom 100k websites have increased proportionally. The surprising finding of the original study is however still valid, we observe a lower RPKI coverage for the Top 100k domains (55.2%) than for the bottom 100k domains (61.5%).

Using the 'Content Delivery Network' tag provided by BGP.Tools (see Figure 4) we found that the adoption of RPKI by CDNs has also drastically increased since 2015, from 0.9% as reported by the RiPKI study to 68.4%.

The significant increase in RPKI adoption observed for popular domains between 2015 and 2024 is explained by the global RPKI uptake in the recent years. In 2015 the RPKI deployment was almost non-existent but it is now covering over 50% of the IPv4 global routing table [23]. This increase explains the difference between the results reported in the original *RiPKI* paper [39] and the ones observed with recent data in IYP.

4.1.4 Discussion. Each of the reproduced values are obtained with one or two queries that are usually three lines long (see Listing 4). This is a significant improvement over the time-consuming steps that the authors of the RiPKI paper had to go through, which include collecting DNS data, as well as downloading and parsing BGP and RPKI data.

As IYP inherently connects various datasets, we can further leverage this capability to explore more data, thereby enriching our insights into RPKI deployment. For example, modifying the query that reports the RPKI coverage for CDN prefixes can get us the RPKI deployments for each tag provided by BGP.Tools or any other AS classification integrated in IYP. Doing so we found utterly disparate RPKI deployments based on BGP.Tools tags. Networks classified as Academic and Government have the lowest RPKI adoption (respectively 16% and 21%) while DDoS Mitigation networks are among the highest (76%). This could be explained, on the one hand, by the prevalence of legacy space in academic and government networks and the legal barriers impeding the addition of these prefixes into RPKI [5], and on the other hand by the RPKI security benefits that are useful for DDoS mitigation.

Again, our goal here is not to study RPKI deployment — as this line of work may benefit from more in-depth analysis and interpretation — but use a brief analysis to demonstrate the effectiveness of IYP to combine multiple datasets and to provide insights quickly.

4.2 Reproduction: DNS robustness

4.2.1 Summary of original study. The second study focuses on the robustness of the DNS ecosystem for popular domain names [3]. It is mainly based on two longitudinal datasets spanning from 2009 to 2018:

- Alexa Top 1 million (not available in IYP, substituted by Tranco).
- DNS zone files for the .com, .net, and .org Top Level Domain (TLD) names (not available in IYP, substituted by OpenINTEL [26, 35]).

This study surveys the implementation of DNS best practices for popular .com, .net, and .org domain names. Although we reproduced all results of this study using IYP (see Table 3 and 4 and the accompanying Jupyter notebook), we focus here only on key differences with original results.

4.2.2 Nameserver replicas (Ref. [3] §4.1).

Original results: In order to avoid a single point of failure in DNS resolution, RFC 1034 and 2182 requires each DNS zone to maintain two nameservers in two different locations. This study finds that a large fraction of the studied domain names meet this requirement (around 39%) or exceed it by deploying more than two nameservers (around 20%), but still about a third (28%) do not meet these requirements (13% are discarded due to limitations of using only three zone files).

Updated results: To reproduce these results we have replicated the same limitations as the original study's dataset using DNS data from OpenINTEL; focusing only on second level domain names (SLDs) in the .com, .net, and .org zones (49% of the Tranco list) and discarding domains that have no glue record in these zones. We found that in May 2024 the percentage of domain names that exceed the requirements has tripled since 2018 while both the percentage for the domain names that meet and do no meet the requirements has decreased significantly (see Table 3).

Despite the large differences in the absolute values, our results corroborate with the consistent increasing trend of SLDs exceeding requirements observed in the original study from 2009 to 2018. This is also in line with a more recent study showing a marked increase of replicas and adoption of anycast from 2017 to 2021 [30].

4.2.3 Shared infrastructure (Ref. [3] §5).

Original results: The DNS robustness study ends with the analysis of shared DNS infrastructure. It investigates the number of popular domain names that share exactly the same set of nameservers and found that half the domain names (median value) share a set of namesservers with at least 163 other domain names. The largest group contains 9k domain names that share the exact same set of nameservers (see Table 4). Grouping nameservers by /24 prefixes, they report larger groups of domain names, respectively, 3k and 71k for the median and maximum values.

Updated results: Results obtained in 2024 with IYP are similar (see Table 4). We observe much less concentration of domains when grouped by nameservers but more when grouped by /24 prefixes. The largest group contains 114k domain names whose nameservers

Table 3: DNS best practice: Comparison between original DNS robustness results and reproduced results with IYP for .com, .net, and .org domain names.

	Coverage .com, .net, .org	Discarded SLDs	Meet NS Requirements.	Exceed NS Requirements	Not meet NS Requirements	In-zone Glue
DNS Robustness (2009-2018)	56 %	12-15 %	≈ 39 %	≈20 %	28%	69-73%
IYP (2024/05/01)	49 %	10 %	18 %	67 %	4%	76%

Table 4: DNS shared infrastructure: Comparison between original DNS robustness results and reproduced results using IYP for .com, .net, and .org domain names (Med. = Median, Max. = Maximum).

	Grouped by NS		Grouped by /24	
	Med.	Max.	Med.	Max.
DNS Robustness (2018)	163	9k	3k	71k
IYP (2024/05/01)	9	6k	3.9k	114k

Table 5: DNS shared infrastructure: Extended results using IYP for all SLDs and BGP prefixes.

IYP (2024/05/01)	Med.	Max.
.com/.net/.org grouped by BGP prefix	4.1k	114k
All Tranco grouped by BGP prefix	6k	187k
All Tranco grouped by NS	15	25k

are all in the same /24 prefixes which confirms the increasing trend observed by the original study.

4.2.4 Discussions. Both the RiPKI and this DNS study are good showcases for IYP, because the approach for both studies essentially consists in merging different datasets in order to reveal new insights. A common approach in Internet measurements and a task at which IYP excels.

Noteworthy for us is the future work mentioned in the original DNS paper [3]. The prospect of studying more than 3 TLDs, identifying anycast prefixes, and using BGP data to refine the study is left for future work. These are analysis that can be time consuming, for example parsing BGP data, or need very specific datasets (e.g., anycast classification) but readily available in IYP. In fact the IYP queries are even simpler when we are not replicating the limitations of the original study (i.e., 3 TLDs and /24 grouping) as we don't need conditions to limit the query to certain domain names or compute the /24 equivalent for each nameserver (interested readers can see the queries in appendix, Listing 5 and Listing 6). The results of using BGP prefixes instead of /24 prefixes for nameservers of .com, .net, and .org are almost identical (Table 5), the median value is at 4.1k instead of 3.9k and the maximum is unchanged, meaning that the assumption about /24 prefixes made in the original paper is sound. By removing the limit on the 3 TLDs (covering only half of the list) and considering all domain names in Tranco, we find that the group of domain names sharing the same nameservers has increased; the median value almost doubled (9 to 15) which is expected as the number of studied domain names doubled. However, the largest group has significantly increased (6k to 25k) suggesting that domain names from other TLDs are more consolidated. Similar observations are made when grouping the BGP prefixes of nameservers for all Tranco domain names (Table 5).

5 New insights

We now present new findings we discovered while working on the reproduction studies. These additional results highlight the versatility of IYP for Internet data analysis. With all datasets pre-parsed and uniformly structured, IYP allows us to seamlessly explore the wide variety of available datasets from different perspectives.

Each of the papers presented in Section 4 studies a different aspect of the robustness of popular domains, one focusing on the best practices for the nameservers and the other focusing on the deployment of RPKI of popular services. Reproducing these studies, we found interesting ways to combine them and potential extensions.

5.1 Combining RiPKI and DNS Robustness

5.1.1 RPKI coverage for nameservers. The DNS robustness study focuses only on DNS best practices, but one may also question the deployment of RPKI for nameservers.

An IYP query for doing so is similar to the one we used for the reproduction of the RiPKI study (see Section 4.1). However, instead of fetching the IP address of popular hostnames, we fetch the IP addresses of the corresponding nameservers and then the corresponding BGP prefixes and RPKI status (similar to the central branch starting from MANAGED_BY in Figure 4).

We found that 48% of all the prefixes hosting nameservers for Tranco domain names are covered by RPKI, which is a bit lower than the overall RPKI coverage observed for prefixes hosting popular hostnames (52.2% in Table 2) and significantly lower than the 68.4% of CDN prefixes that host the services. However, the concentration of nameservers in a few DNS provider using RPKI means that 84% of the Tranco domain names are managed by nameservers covered by RPKI.

Overall, we find that RPKI coverage is driven by a few ASes and prefixes in which nameservers and hostnames are concentrated. Also, the RPKI deployment in the DNS infrastructure is lagging behind that of content providers.

5.1.2 Web hosting consolidation and RPKI. Given the consolidation of the DNS infrastructure reported in the DNS robustness study, one may question the consolidation of hosting providers and its impact on the RPKI results reported earlier (Table 2). RiPKI reports only percentages of prefixes covered by RPKI, it is not investigating the distribution of domains across prefixes (which made sense at that time given the low value originally reported). To address this, we simply count hostnames instead of prefixes in our queries (see

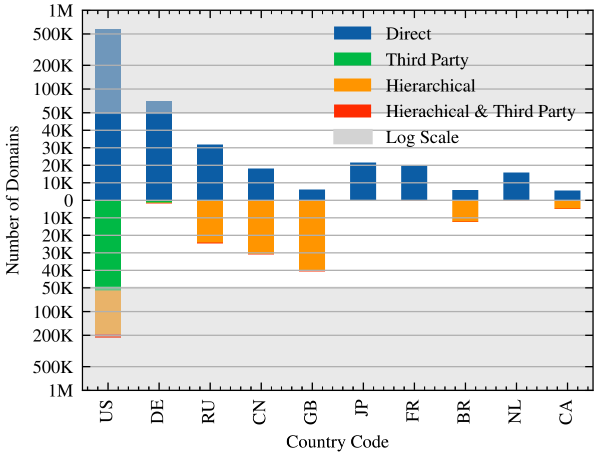


Figure 5: Country-based SPoF in the DNS chain of Tranco and Cisco's Top 1M domain names.

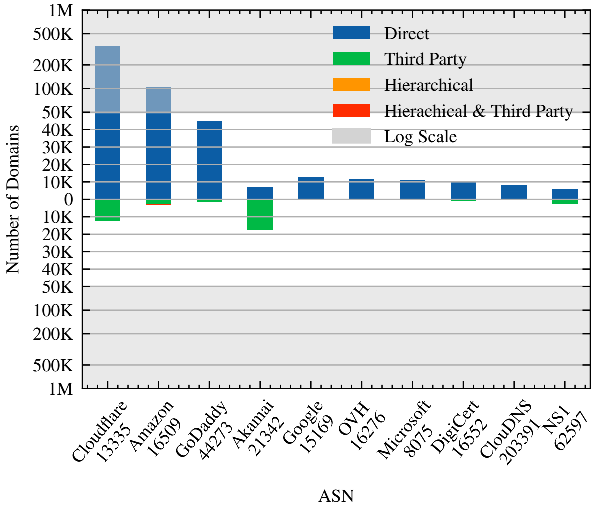


Figure 6: AS-based SPoF in the DNS chain of Tranco and Cisco's Top 1M domain names.

the RETURN statement in Listing 4). We found that, despite 52.2% (Table 2) of all prefixes for Tranco being covered by RPKI, many domains map to the same RPKI-covered prefixes, resulting in 78.8% of Tranco domains actually covered by RPKI. Similarly for CDN prefixes, while we report in Table 2 a coverage of 68.4%, the RPKI-covered prefixes represent 96% of the domains hosted by these CDN providers. These results highlight that prefixes in which domains are concentrated tend to have better RPKI coverage, especially for CDN prefixes.

5.2 SPoF in DNS chain

Another extension that we implemented is for the DNS robustness study. In Section 4.2 we followed the same methodology as the original paper, but used IYP. However, this methodology only focuses on direct dependencies of domain names to their nameservers, ignoring third-party and hierarchical dependencies that happen when performing a complete DNS resolution. For example, a resolver has to query the DNS root and .com nameservers to resolve example.com. With the same intention of understanding robustness in DNS infrastructure, we are able to extend the analysis beyond direct dependencies using several imported datasets (Tranco, Cisco Umbrella, pfx2asn from BGPKIT, DNS dependency graph [35], delegated files from NRO). This approach helped us identify cascading single points of failure (SPoF) that spread over the DNS chain dependencies while being aware of DNS working mechanisms (i.e., usage of glue records). We define hierarchical dependencies as dependencies due to the inner nature of the hierarchical nature of DNS; root servers (excluded here), TLD, SLD, third level, etc. Third-party dependencies are due to *outsourcing*; e.g., foo.com has ns.bar.com as nameserver and bar.com has ns.goo.com as nameserver, and so on.

We investigate SPoF in the DNS chain of Tranco and Cisco's Top 1 Million List at two different levels of granularity: country and AS. A country-based SPoF suggests sovereignty consequences that are associated to DNS (computed using the country code of ASes found in RIRs delegated files), while the AS-based SPoF focuses on consolidation of DNS providers.

Figure 5 shows that direct dependencies dominate the DNS ecosystem, but there is also a significant extent of third-party dependency towards the US. This indicates that DNS hosting services mostly rely on ASes registered in the US for operating their DNS infrastructure. Furthermore, we observe a large hierarchical dependency on Russia, China, and the UK. This is due to numerous domains managed by nameservers that may be outside the country, but where the country-code top-level domains are under the control of an organization within the country.

A more detailed picture emerges in Figure 6 in which we see that different ASes offer different types of services (direct and indirect) to the DNS ecosystem. For example, on one hand, Akamai is mainly seen as a third-party dependency, meaning that it hosts mainly DNS services for DNS hosting companies rather than end-customer nameservers. On the other hand, GoDaddy is focused primarily on DNS services for end customers (direct dependencies).

Overall, this inceptive analysis shows that DNS is a complex ecosystem with an intricate set of interdependencies not always immediately clear from an end-user perspective. IYP enables us to explore and delve deep into these interdependencies thanks to the graph data structure and availability of the different datasets needed for this exploration.

6 Embracing IYP

After developing and experimenting with IYP, it has become highly integrated into our research routine. It is now our preferred way to quickly access and investigate the multiple datasets we use in our projects. It is also our go-to tool to satisfy curiosity during brainstorming sessions or manual data analysis.

In this section we share our experience using IYP for research, describe potential implications for data sharing and reproducibility, and finally discuss our vision for IYP and its role in the Internet measurement community.

6.1 Lessons learned

Local instance. In our daily routine we frequently use both the IYP public instance and personal local instances, as we found that they each serve different purposes. The IYP public instance is read-only and is useful for sporadic queries to address questions raised during discussions or exploratory work. However, in-depth analysis requires numerous queries or queries scanning a large amount of data, which are better served via a local instance of IYP. With the help of docker and the IYP weekly snapshots, it is easy to start a local instance of IYP, requiring no particular hardware and able to run on any modern laptop. A local instance offers the benefits of adding new nodes and relationships to the knowledge graph, whether they are from additional (and possibly confidential) datasets or serving as supplementary knowledge that would facilitate subsequent analysis. For example, for the SPoF analysis conducted in Section 5.2 we added temporal SPoF relationships in the knowledge graph to maintain that information in the database itself. Or, more commonly, one can simply tag the set of studied resources so that subsequent queries are simplified.

Query precision. Queries can express different levels of details. For example, the query in Listing 1 is generic. It searches for any prefix originated by an AS regardless of where the data comes from. As mentioned in Section 2.2, each relationship in IYP has a property that exposes the name of the dataset behind that relationship. This property is helpful to check a posteriori the provenance of the data, or to formulate a precise query targeting only certain datasets. Precise queries give better control on the underlying datasets, which may be preferred for projects involving dataset limitations. In general, generic queries are preferred in exploratory data analysis for discovering unexpected relationships. The outcome for these two type of queries may also vary over time, precise queries depend solely on the selected datasets while generic ones may depend on multiple datasets. Hence precise queries may have a more expected output but could have a shorter lifespan.

Augmented datasets. After integrating more and more datasets into IYP, we realized that adding a new dataset brings mutual benefits to both IYP and the data provider. Not only will a new dataset expand IYP's accumulated knowledge, it will also significantly augment the new dataset by connecting it to all the datasets already available in IYP. For example, the OpenINTEL *tranco1m* dataset (see Table 1) is essentially a list of hostnames and IP addresses, but the integration to IYP extends it to the corresponding BGP prefixes, origin AS, IRR and RPKI status, delegated prefixes, nameservers, and any other entities related to these, enabling insightful studies like the two reproduced in Section 4.

Datasets comparison. Thanks to the data unification of the knowledge graph, IYP has also proven to be an effective tool for comparing datasets. Though we generally avoid importing redundant datasets, there are instances where this can be useful in IYP. For example, BGPKIT's pfx2asn and IHR's ROV datasets both map prefixes to their origin AS. By querying the differences between these two datasets, we discovered an error that affected results for certain IPv6 prefixes in the BGPKIT dataset. Following our own recommendations (Section 2.3), we addressed this by contacting the data provider, leading for the error to be fixed at the origin and corrected in subsequent IYP snapshots. This error would have been difficult to discover and would have required significant work to compare datasets directly, but is rather trivial using IYP.

6.2 Sharing queries

We believe that IYP can also help facilitate the sharing of research methods and reproducibility. We encourage researchers using IYP to include their queries and the date of the snapshots used to produce their results. This can either be included in an appendix to a research paper or an online document such as the two Jupyter notebooks reproducing the two studies of Section 4. Additionally, the use of version control is preferred to accommodate any changes to IYP in the future.

Sharing queries is straightforward, yet it empowers anyone to retrieve the exact same data, which is crucial for researchers replicating or building upon past studies. Furthermore, we have found that sharing queries may be even more useful than sharing datasets. Since the same query can be executed on newer snapshots of IYP, it simplifies the process of reproducing and updating previous works. For instance, the two Jupyter notebooks we share with this paper query the IYP public instance. Re-executing the notebooks will produce all the results presented in Section 4, but will use the most recent data by default. This makes the studies reproducible on-demand.

6.3 Community engagement

This paper represents the results of our efforts to build a practical knowledge graph for Internet resources. The curated list of datasets currently available in IYP and the ontology are based on our experience with networking data, discussions with different stakeholders, and numerous experiments with IYP. Although we are planning to continue improving IYP and operate the public instance, we envision IYP as a community-driven project where dataset curation and ontology definitions evolve with the Internet measurement community. Inspired from the Wikidata community [1], maintaining the knowledge graph underlying many Wikimedia projects, we believe that the long-term success of IYP resides in the engagement with the Internet measurement community for discussing new datasets and terms for the ontology. Consequently, we invite researchers, data providers, and interested parties to participate in discussions in the IYP Github pages.[9]

7 Limitations

IYP has several limitations. This section discuss those that are inherent to its design or common to large data collections.

Learning curve. Despite our efforts to make IYP as accessible as possible (e.g., providing database snapshots, docker images, example queries), there are still some requirements before being able to get started using IYP. The first is Cypher, which is the most intuitive graph querying language we have experimented with, but not broadly used in the measurement community. Fortunately the Neo4j community has produced rich educational material for Cypher, which helps to significantly lower the barrier to entry. The second is getting accustomed with the datasets available in IYP

[9] https://github.com/InternetHealthReport/internet-yellow-pages.

and their modeling. We provide documentation for all imported datasets as well as example queries for different tasks. Users should be aware that documentation is inevitably growing with the number of integrated datasets. For future work, we are also considering the use of Large Language Models to bridge the gap between users and crafting queries for IYP.

Longitudinal analysis. IYP is not particularly well-suited for longitudinal analysis. This is not a limitation of knowledge graphs per se. However, this significantly complicates queries for the graph. Consequently, we opted to construct knowledge graphs that represent snapshots in time. This decision is a trade off between features and usability. We conducted a longitudinal study (not presented here) by running multiple IYP instances representing different snapshots in time. While feasible, we found the process cumbersome as it involves fetching data from multiple instances and merging results ourselves. A variant of IYP including temporal dynamics could be an interesting follow up project.

Data quality. The data quality of IYP is closely related to the quality of imported datasets. While IYP can assist in verifying consistency, completeness, and accuracy of datasets (see Section 6), users should be aware of the original datasets' limitations to accurately interpret results and maximize the utility of IYP. This challenge is inherent to data analysis as a whole, and we believe that documenting and sharing experiences is the most effective way to tackle it.

Maintainability. IYP relies on data collected by various organizations, each facing different challenges in consistently providing data over time. We aim to import datasets that are sustainable (Section 2.1). Some projects, such as RouteViews and RIPE RIS, have demonstrated incredible longevity. However, it is essential to acknowledge that profound changes can occur to imported datasets and the IYP ontology. The unification of data in IYP helps to handle some of these changes but still efforts will be necessary to maintain IYP consistency over time.

8 Related Work

The benefits and challenges in sharing Internet measurement data has been discussed for at least two decades. Allman et al. [3] proposed the Scalable Internet Measurement Repository, a database designed to facilitate the sharing of data, including dataset metadata and user information. This inspired development of the Internet Measurement Data Catalog (DatCat) [29] from CAIDA. These initiative were followed by the DHS project, IMPACT [17] (formerly PREDICT), focusing on cybersecurity datasets and ensuring that data is shared in a controlled manner.These are prime examples of data sharing platforms that foster collaboration and innovation. However, they do not support directly querying the various datasets, as possible in IYP.

More similar to IYP's approach, the Internet Geographic Database (iGDB) [4] is a database combining physical and logical (e.g., AS and IP-level) Internet topology data, enabling visualization and geographic analysis. In addition, MISP is a collaborative platform with various taxonomies for sharing threat intelligence [38]. Although these tools are tailored for specific use-cases and datasets, we plan to explore ways to incorporate their datasets into IYP, or integrate IYP into these tools.

The community has also proposed ways to reduce time to insight for data analysis. For example, BGPstream [27] facilitates access to BGP data by unifying access to the RIS and RouteViews archives. These types of tools are complementary to IYP. One may find insights with IYP and perform a longitudinal analysis with these tools, or use IYP to enrich the information provided by them.

The literature on knowledge graphs is vast, we refer interested reader to surveys [19, 28, 41] that provide an overview of the different techniques of constructing and using knowledge graphs. To the best of our knowledge, IYP is the first knowledge graph focused on Internet data.

9 Ethics

Open datasets are the core of IYP. We ensure for all datasets that their license or AUA permits integration into IYP. We compile all datasets with their respective license, required citations, and links to the original data in one central acknowledgment page, even if not required by all datasets.[10] Therefore, to the best of our knowledge, this work does not raise any ethical issues.

10 Conclusion

In this paper, we presented IYP, a knowledge graph for Internet resources. By unifying various Internet measurement datasets into a single harmonized database, IYP enables us to quickly gain insights about the Internet. We demonstrated this by reproducing and extending two studies on RPKI and DNS. Additionally, we discussed how IYP can enhance results sharing and reproducibility of past studies.

IYP is poised to evolve with the Internet measurement community, incorporating new terminology into its ontology and integrating new datasets. Moreover, IYP paves the way for exploring the numerous knowledge graph applications to Internet data, including knowledge reasoning [11], recommender systems [15], and various applications based on knowledge graph embeddings [41].

Acknowledgments

The Internet Yellow Pages would not have been possible without the extraordinary efforts made by the Internet measurement community to collect and maintain public datasets. We want to extend our gratitude to all organizations that are providing data to IYP, and the people that helped us integrating these datasets. We would like to especially thank, Mingwei Zhang for his support with BGPKIT and Cloudflare data; George Michaelson for his help with the APNIC datasets; Ben Cartwright-Cox for assisting us to access to BGP.Tools data; Zhiyi Chen for his help to integrate the AS to Organization mapping dataset; Roopesh Saravanan, Justin Loye, Friedemann Lipphardt for the integration of various datasets; Ryan Poon for automating some of the IYP processes; and the Google Summer of Code program for supporting some of these contributors. Finally, special thanks to kc claffy, whose unwavering encouragement and support were an invaluable source of strength for this research journey.

[10] https://github.com/InternetHealthReport/internet-yellow-pages/blob/main/ACKNOWLEDGMENTS.md.

References

[1] 2024. Wikidata:Community portal. https://www.wikidata.org/wiki/Wikidata:Community_portal

[2] Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. 2020. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* 8 (2020), 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

[3] Mark Allman. 2018. Comments on DNS Robustness. In *Internet Measurement Conference (IMC'18)*. ACM, 84–90. https://doi.org/10.1145/3278532.3278541

[4] Scott Anderson, Loqman Salamatian, Zachary S. Bischof, Alberto Dainotti, and Paul Barford. 2022. iGDB: Connecting the Physical and Logical Layers of the Internet. In *Internet Measurement Conference (IMC'22)*. ACM, 433–448. https://doi.org/10.1145/3517745.3561443

[5] ARIN. 2024. Legacy Resources at ARIN. https://www.arin.net/resources/guide/legacy/

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 1 (2000), 25–29. https://doi.org/10.1038/75556

[7] BGPKIT. [n. d.]. pfx2as. https://data.bgpkit.com/pfx2as

[8] Zachary S. Bischof, Kennedy Pitcher, Esteban Carisimo, Amanda Meng, Rafael Bezerra Nunes, Ramakrishna Padmanabhan, Margaret E. Roberts, Alex C. Snoeren, and Alberto Dainotti. 2023. Destination Unreachable: Characterizing Internet Outages and Shutdowns. In *SIGCOMM 2023 Conference (SIGCOMM'23)*. ACM, 608–621. https://doi.org/10.1145/3603269.3604883

[9] CAIDA. [n. d.]. RouteViews Prefix to AS mappings. https://catalog.caida.org/dataset/routeviews_prefix2as

[10] Esteban Carisimo, Alexander Gamero-Garrido, Alex C. Snoeren, and Alberto Dainotti. 2021. Identifying ASes of State-Owned Internet Operators. In *Internet Measurement Conference (IMC'21)*. ACM, 687–702. https://doi.org/10.1145/3487552.3487822

[11] Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications* 141 (2020), 112948. https://doi.org/10.1016/j.eswa.2019.112948

[12] Romain Fontugne, Anant Shah, and Emile Aben. 2018. The (thin) Bridges of AS Connectivity: Measuring Dependency using AS Hegemony. In *Passive and Active Network Measurement (PAM'18)*. Springer, 216–227. https://doi.org/10.1007/978-3-319-76481-8_16

[13] Romain Fontugne, Anant Shah, and Kenjiro Cho. 2020. Persistent Last-mile Congestion: Not so Uncommon. In *Internet Measurement Conference (IMC'20)*. ACM, 420–427. https://doi.org/10.1145/3419394.3423648

[14] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *International Conference on Management of Data (SIGMOD'18)*. ACM, 1433–1445. https://doi.org/10.1145/3183713.3190657

[15] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A Survey on Knowledge Graph-Based Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering* 34, 8 (2020), 3549–3568. https://doi.org/10.1109/TKDE.2020.3028705

[16] Paul E. Hoffman, Andrew Sullivan, and Kazunori Fujiwara. 2015. DNS Terminology. RFC 7719. https://doi.org/10.17487/RFC7719

[17] IMPACT. 2024. Information Marketplace for Policy and Analysis of Cyber-risk & Trust. https://www.impactcybertrust.org

[18] Internet Health Report. [n. d.]. ip2asn: Mapping IP address to originating ASN. https://github.com/InternetHealthReport/ip2asn

[19] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2021. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems* 33, 2 (2021), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

[20] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. 2023. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research* 52, D1 (11 2023), D1265–D1275. https://doi.org/10.1093/nar/gkad976

[21] Petri Kotiranta, Marko Junkkari, and Jyrki Nummenmaa. 2022. Performance of Graph and Relational Databases in Complex Queries. *Applied Sciences* 12, 13 (2022), 6490. https://doi.org/10.3390/app12136490

[22] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Network and Distributed Systems Security Symposium (NDSS'19)*. https://doi.org/10.14722/ndss.2019.23386

[23] Doug Madory and Job Snijders. 2024. RPKI ROV Deployment Reaches Major Milestone. https://www.kentik.com/blog/rpki-rov-deployment-reaches-major-milestone/

[24] Neo4j. [n. d.]. Transition from relational to graph database. https://neo4j.com/docs/getting-started/appendix/graphdb-concepts/graphdb-vs-rdbms/

[25] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. *Queue* 17, 2 (2019), 48–75. https://doi.org/10.1145/3329781.3332266

[26] OpenINTEL. 2024. Active DNS Measurement Project. https://openintel.nl/

[27] Chiara Orsini, Alistair King, Danilo Giordano, Vasileios Giotsas, and Alberto Dainotti. 2016. BGPStream: A Software Framework for Live and Historical BGP Data Analysis. In *Internet Measurement Conference (IMC'16)*. ACM, 429–444. https://doi.org/10.1145/2987443.2987482

[28] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218

[29] Colleen Shannon, David Moore, Ken Keys, Marina Fomenkov, Bradley Huffaker, and kc claffy. 2005. The Internet Measurement Data Catalog. *ACM SIGCOMM Computer Communication Review* 35, 5 (2005), 97–100. https://doi.org/10.1145/1096536.1096552

[30] Raffaele Sommese, Gautam Akiwate, Mattijs Jonker, Giovane C. M. Moura, Marco Davids, Roland van Rijswijk-Deij, Geoffrey M. Voelker, Stefan Savage, kc claffy, and Anna Sperotto. 2021. Characterization of Anycast Adoption in the DNS Authoritative Infrastructure. In *Network Traffic Measurement and Analysis Conference (TMA'21)*. IFIP.

[31] Neil Swainston, Riza Batista-Navarro, Pablo Carbonell, Paul D. Dobson, Mark Dunstan, Adrian J. Jervis, Maria Vinaixa, Alan R. Williams, Sophia Ananiadou, Jean-Loup Faulon, Pedro Mendes, Douglas B. Kell, Nigel S. Scrutton, and Rainer Breitling. 2017. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS ONE* 12, 7 (2017), 1–14. https://doi.org/10.1371/journal.pone.0179130

[32] The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* 47, D1 (2019), D330–D338. https://doi.org/10.1093/nar/gky1055

[33] Santiago Timón-Reina, Mariano Rincón, and Rafael Martínez-Tomás. 2021. An overview of graph databases and their applications in the biomedical domain. *Database* 2021 (2021), baab026. https://doi.org/10.1093/database/baab026

[34] Michael Uschold. 2015. Ontology and database schema: What's the difference? *Applied Ontology* 10, 3-4 (2015), 243–258. https://doi.org/10.3233/AO-150158

[35] UTwente. 2024. DNS Dependency Graph Dataset Repository. https://dnsgraph.dacs.utwente.nl

[36] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. 2010. A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective. In *Annual ACM Southeast Conference (ACMSE'10)*. 1–6. https://doi.org/10.1145/1900008.1900067

[37] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489

[38] Cynthia Wagner, Alexandre Dulaunoy, Gérard Wagener, and Andras Iklody. 2016. MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform. In *Workshop on Information Sharing and Collaborative Security (WISCS '16)*. ACM, 49–56. https://doi.org/10.1145/2994539.2994542

[39] Matthias Wählisch, Robert Schmidt, Thomas C. Schmidt, Olaf Maennel, Steve Uhlig, and Gareth Tyson. 2015. RiPKI: The Tragic Story of RPKI Deployment in the Web Ecosystem. In *ACM Workshop on Hot Topics in Networks (HotNets-XIV)*. 1–7. https://doi.org/10.1145/2834050.2834102

[40] Chengbin Wang, Xiaogang Ma, Jianguo Chen, and Jingwen Chen. 2018. Information Extraction and Knowledge Graph Construction from Geoscience Literature. *Computers & Geosciences* 112 (2018), 112–120. https://doi.org/10.1016/j.cageo.2017.12.007

[41] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

[42] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36, suppl_1 (2008), D901–D906. https://doi.org/10.1093/nar/gkm958

[43] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A System for Classifying Owners of Autonomous Systems. In *Internet Measurement Conference (IMC'21)*. ACM, 703–719. https://doi.org/10.1145/3487552.3487853

Table 6: List of entities (node types) in IYP ontology with their description. Important properties are mentioned in the descriptions but these are not comprehensive.

Entity	Description
AS	Autonomous System, uniquely identified with the *asn* property.
AtlasMeasurement	RIPE Atlas Measurement, uniquely identified with the *id* property.
AtlasProbe	RIPE Atlas probe, uniquely identified with the *id* property.
AuthoritativeNameServer	Authoritative DNS nameserver for a set of domain names, uniquely identified with the *name* property.
BGPCollector	A RIPE RIS or RouteViews BGP collector, uniquely identified with the *name* property.
CaidaIXID	Unique identifier for IXPs from CAIDA's IXP dataset.
Country	Represent an economy, uniquely identified by either its two or three character code (properties *country_code* and *alpha*3).
DomainName	Any DNS domain name that is not a FQDN (see HostName), uniquely identified by the *name* property.
Estimate	Represent a report that approximate a quantity, for example the World Bank population estimate.
Facility	Co-location facility for IXPs and ASes, uniquely identified by the *name* property.
HostName	A fully qualified domain name uniquely identified by the *name* property.
IP	An IPv4 or IPv6 address uniquely identified by the *ip* property. The *af* property (address family) provides the IP version of the prefix.
IXP	An Internet Exchange Point, loosely identified by the *name* property or using related IDs (see the EXTERNAL_ID relationship).
Name	Represent a name that could be associated to a network resource (e.g., an AS), uniquely identified by the *name* property.
OpaqueID	Represent the opaque-id value found in RIR's delegated files. Resources related to the same opaque-id are registered to the same resource holder. Uniquely identified by the *id* property.
Organization	Represent an organization and is loosely identified by the *name* property or using related IDs (see the EXTERNAL_ID relationship).
PeeringdbFacID	Unique identifier for a Facility as assigned by PeeringDB.
PeeringdbIXID	Unique identifier for an IXP as assigned by PeeringDB.
PeeringdbNetID	Unique identifier for an AS as assigned by PeeringDB.
PeeringdbOrgID	Unique identifier for an Organization as assigned by PeeringDB.
Prefix	An IPv4 or IPv6 prefix uniquely identified by the *prefix* property. The *af* property (address family) provides the IP version of the prefix.
Ranking	Represent a specific ranking of Internet resources (e.g., CAIDA's ASRank or Tranco ranking). The rank value for each resource is given by the RANK relationship.
Tag	The output of a classification. A tag can be the result of a manual or automated classification. Uniquely identified by the *label* property.
URL	The full URL for an Internet resource, uniquely identified by the *url* property.

```
 // List /24 prefixes of nameservers for .com/.net/.org domain names in Tranco
MATCH  (r:Ranking {name:'Tranco top 1M'})-[:RANK]-(d:DomainName)-[:MANAGED_BY]-(a:AuthoritativeNameServer)
-[:RESOLVES_TO]-(i:IP {af:4})
WHERE d.name ENDS WITH '.com' OR d.name ENDS WITH '.net' OR d.name ENDS WITH '.org'
RETURN d, COLLECT(DISTINCT REDUCE(pfx = "", n IN SPLIT(i.ip, '.')[0..3] | pfx + n + ".")) AS pfx
```

Listing 5: Query used to reproduce results from DNS Robustness [3] on shared infrastructure and using /24 grouping (Table 4). This query reproduces the original paper's setup by selecting only .com, .net, and .org domain names and computing /24 prefixes corresponding to each nameserver.

```
// List prefixes of nameservers for all domain names in Tranco
MATCH  (r:Ranking {name:'Tranco top 1M'})-[:RANK]-(d:DomainName)-[:MANAGED_BY]-(a:AuthoritativeNameServer)
-[:RESOLVES_TO]-(i:IP {af:4})-[:PART_OF]-(pfx:Prefix)
RETURN d, COLLECT(DISTINCT pfx)
```

Listing 6: Query extending results from DNS Robustness [3] on shared infrastructure by looking at all domain names in Tranco and the BGP prefix of the corresponding nameservers (Table 5).

Table 7: List of relationships in IYP ontology with their description. Some properties are mentioned in the descriptions but these are not comprehensive.

Relationship	Description
ALIAS_OF	Equivalent to the CNAME record in DNS. It relates two HostNames.
ASSIGNED	Represent the allocation by a RIR of a network resource (AS, Prefix) to a resource holder (see OpaqueID). Or represent the assigned IP address of an AtlasProbe.
AVAILABLE	Relate ASes and Prefixes to RIRs (in the form of an OpaqueID) meaning that the resource is not allocated and available at the related RIR.
CATEGORIZED	Relate a network resource (AS, Prefix, URL) to a Tag, meaning that the resource has been classified accordingly to the Tag. The *reference_name* property provide the name of the original dataset/classifier.
COUNTRY	Relate any node to its corresponding country. This relation may have different meaning depending on the original dataset (e.g., geo-location or registration).
DEPENDS_ON	Relate an AS or Prefix to an AS, meaning the reachability of the AS/Prefix depends on a certain AS.
EXTERNAL_ID	Relate a node to an identifier commonly used by an organization. For example, PeeringDB assigns unique identifiers to IXPs (see PeeringdbIXID).
LOCATED_IN	Location of a resource at a specific geographical or topological location. For example, co-location Facility for an IXP or AS for an AtlasProbe.
MANAGED_BY	Entity in charge of a network resource. For example an AS is managed by an Organization, a DomainName is managed by an AuthoritativeNameServer.
MEMBER_OF	Represent the membership to an organization. For example, an AS is member of an IXP.
NAME	Relate an entity to its usual or registered name. For example, the name of an AS.
ORIGINATE	Relate a Prefix to an AS, meaning that the prefix is seen as being originated from that AS in BGP.
PARENT	Relate two DomainNames and represent a zone cut between the parent zone and the more specific zone.
PART_OF	Represent that one entity is a part of another. For example, an IP address is a part of an IP Prefix, a HostName is a part of a DomainName.
PEERS_WITH	Represent the connection between two ASes as seen in BGP. It also include peerings between ASes and BGPCollectors.
POPULATION	Indicate that an AS hosts a certain fraction of the population of a country. It also represent the estimated population of a country.
QUERIED_FROM	Relate a DomainName to an AS or Country, meaning that the AS or Country appears in the Top 100 AS or Country to query the most the DomainName (as reported by Cloudflare radar).
RANK	Relate a resource to a Ranking, meaning that the resource appears in the Ranking. The *rank* property gives the exact rank position.
RESERVED	Indicate that an AS or Prefix is reserved for a certain purpose by RIRs or IANA.
RESOLVES_TO	Relate a HostName to an IP address, meaning that a DNS resolution resolved the corresponding IP.
ROUTE_ORIGIN _AUTHORIZATION	Relate an AS and a Prefix, meaning that the AS is authorized to originate the Prefix by RPKI.
SIBLING_OF	Relate ASes or Organization together, meaning that they represent the same entity.
TARGET	Relate an AtlasMeasurement to an IP, HostName, or AS, meaning that an Atlas measurement is setup to probe that resource.
WEBSITE	Relate a URL to an Organization, Facility, IXP, AS, representing a common website for the resource.

Table 8: List of data providers and datasets currently integrated into IYP. Some organizations provide multiple datasets.

Organization	Dataset Name / Description	URL
Alice-LG	IXP route server looking glass snapshots	https://github.com/alice-lg/alice-lg
	AMS-IX	https://lg.ams-ix.net
	BCIX	https://lg.bcix.de
	DE-CIX	https://lg.de-cix.net
	IX.br	https://lg.ix.br
	LINX	https://alice-rs.linx.net
	Megaport	https://lg.megaport.com
	Netnod	https://lg.netnod.se
APNIC	AS population estimate	https://stats.labs.apnic.net/aspop
BGPKIT	as2rel, peer-stats, pfx2as	https://data.bgpkit.com
BGP.Tools	AS names, AS tags	https://bgp.tools/kb/api
	Anycast prefix tags	https://github.com/bgptools/anycast-prefixes
CAIDA	AS Rank	https://doi.org/10.21986/CAIDA.DATA.AS-RANK
	IXPs Dataset	https://www.caida.org/catalog/datasets/ixps
Cisco	Umbrella Popularity List	https://s3-us-west-1.amazonaws.com/umbrella-static/index.html
Citizen Lab	URL testing lists	https://github.com/citizenlab/test-lists
Cloudflare	Cloudflare Radar API endpoints radar/dns/top/ases, radar/dns/top/locations, radar/ranking/top, radar/datasets	https://radar.cloudflare.com
Emile Aben	AS names	https://github.com/emileaben/asnames
IHR	Country Dependency, AS Hegemony, ROV	https://ihr.iijlab.net
Internet Intelligence Lab	AS to Organization Mapping	https://github.com/InetIntel/Dataset-AS-to-Organization-Mapping
NRO	Extended allocation and assignment reports	https://www.nro.net/about/rirs/statistics
OpenINTEL	tranco1m, umbrella1m, ns	https://data.openintel.nl/data
	DNS Dependency Graph	https://dnsgraph.dacs.utwente.nl
Packet Clearing House	Daily routing snapshots	https://www.pch.net/resources/Routing_Data
PeeringDB	API endpoints: fac, ix, ixlan, netfac, org	https://www.peeringdb.com
RIPE NCC	AS names, RPKI	https://ftp.ripe.net/ripe
	RIPE Atlas measurement information	https://atlas.ripe.net
SimulaMet	rDNS data	https://rir-data.org
Stanford	ASdb dataset	https://asdb.stanford.edu
Tranco	Tranco list	https://tranco-list.eu
Virginia Tech	RoVista	https://rovista.netsecurelab.org
World Bank	Indicators API: Country Population Indicator	https://www.worldbank.org