

# Internet Yellow Pages:

Striving for better Internet Infrastructure data

**Romain Fontugne**  
IIJ Research Lab  
**Emile Aben**  
RIPE NCC

« The journey, not the destination matters » :  
The Geopolitics of Internet Routes

December 16<sup>th</sup>, 2022





# Knowledge seekers

Researchers, network/IXP operators, peering coordinators constantly look for knowledge about the Internet.

Going to RIPEstat, bgp.he.net, PeeringDB, CAIDA, whois, etc...

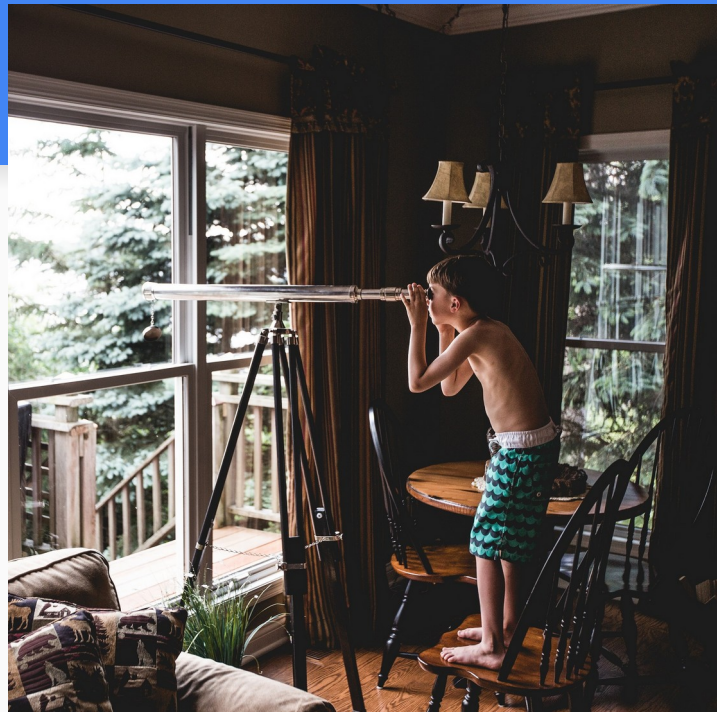
These provide complementary views, sometimes with different semantics.



# Striving for better data

Our Goal:

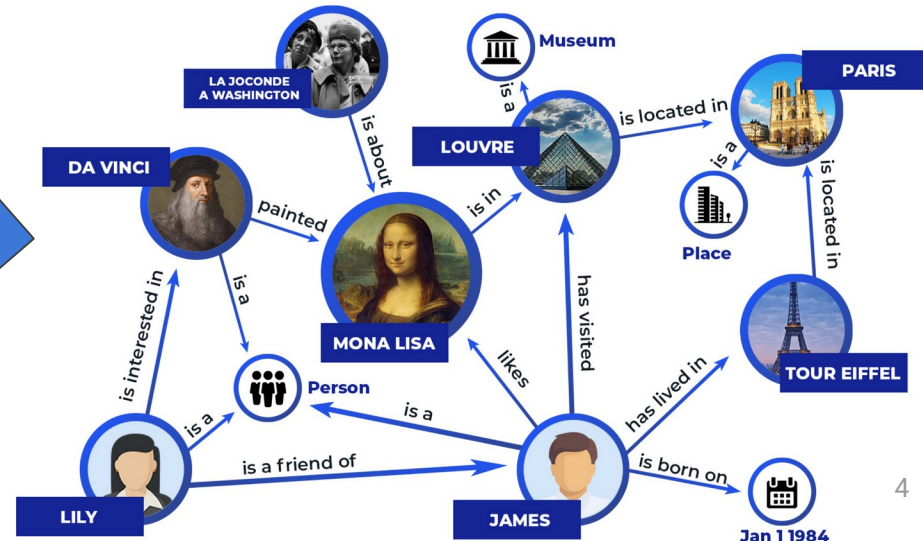
- Provide a place for **sharing knowledge** and connect data from different sources
- **Open** to anyone, easily accessible, easy to contribute to
- **Structured**: not a repo with tons of data dumps
- **Extensible**: no fixed database schema



# Knowledge graph

Structure data in a graph

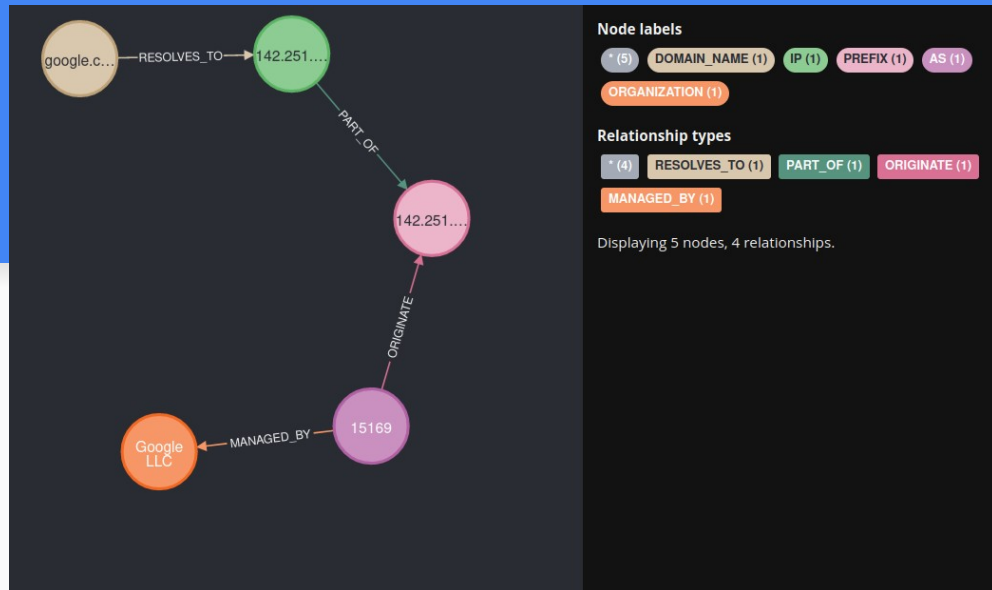
Inspired from wikidata.org,  
Google/Facebook knowledge graph, world factbook



# “Things, not strings”

Ontology - Modeling a domain:

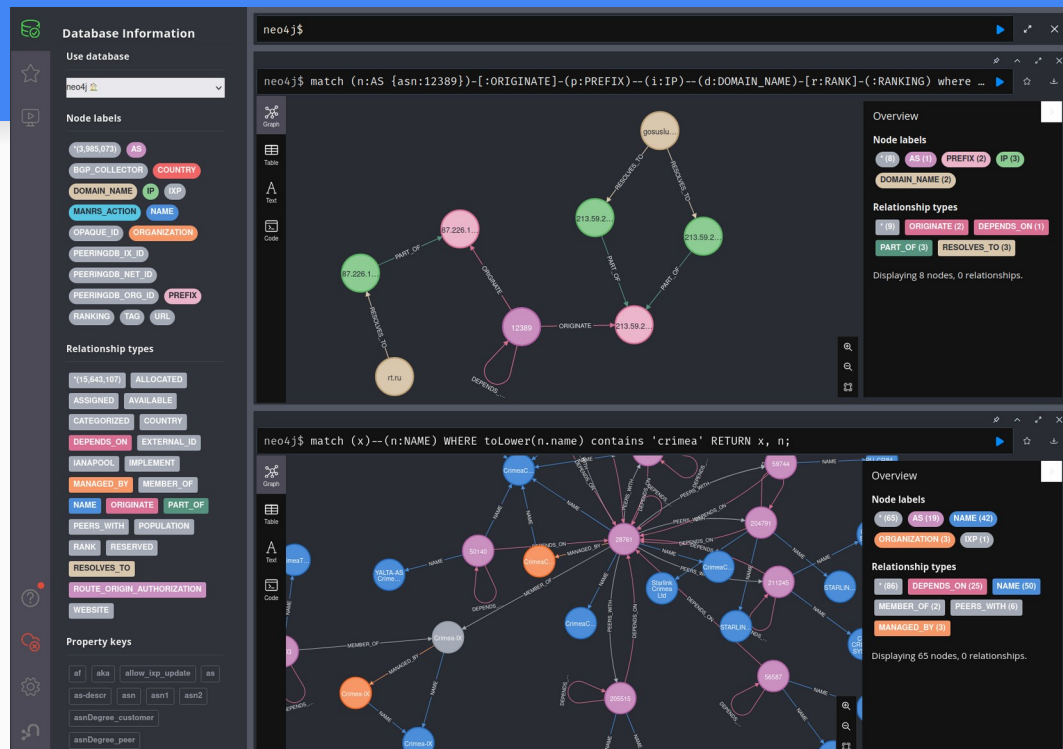
- Expressed in a machine readable format
- Enumeration of entities in this domain and how they relate to each other
- Shared & evolves with a community






# IYP: Current status

- Base on Neo4j
- 12 data sources:  
APNIC, BGPKIT, Bgp.tools, CAIDA, Cloudflare, IHR, MANRS, OpenINTEL, PeeringDB, RIPE NCC, NRO, Tranco
- 17 node types, 21 link types
- About 4M nodes, 15M links



# AS5511's country? (delegated-view)

neo4j\$ `//country in delegated match (net:AS {asn:5511})-[r]-(cc:COUNTRY) return net, cc, r limit 1`

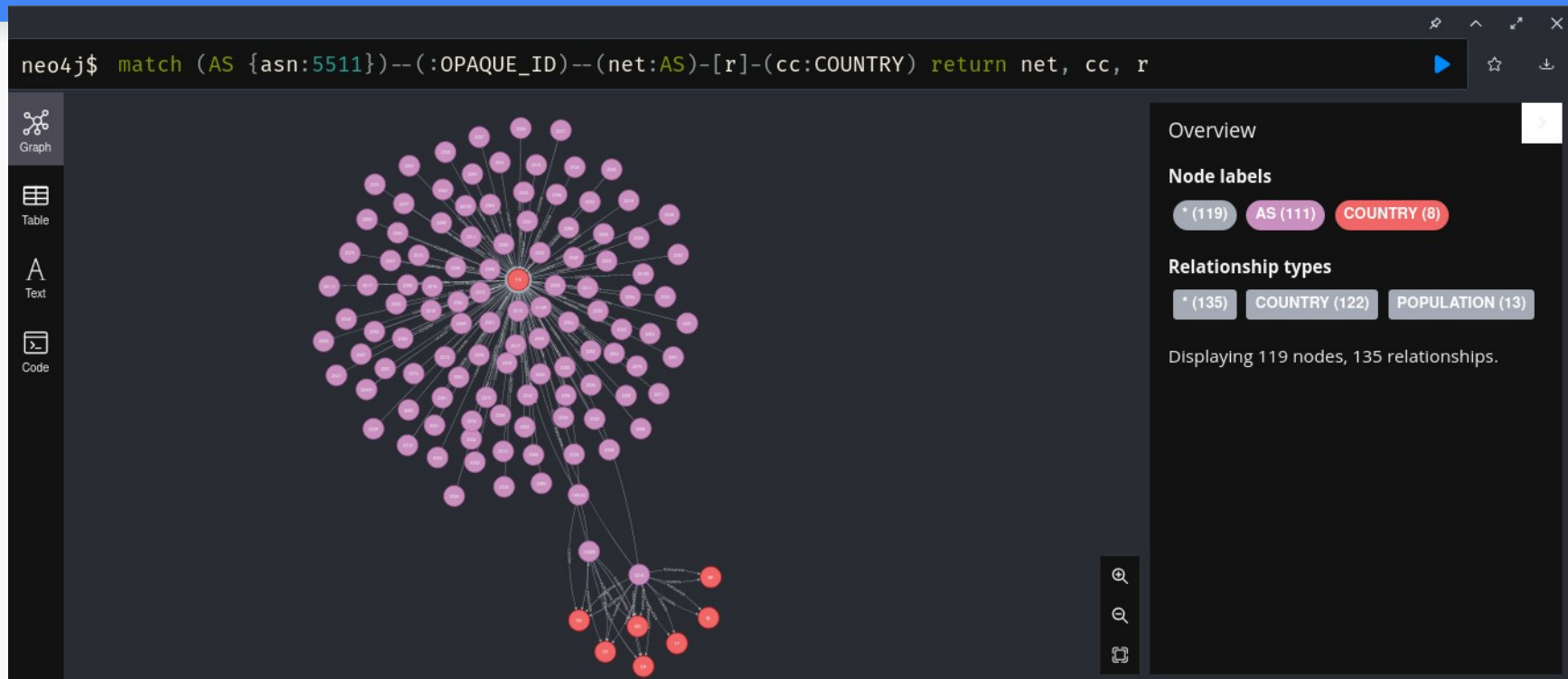


The graph view displays a single relationship between two nodes. On the left is a purple circular node labeled '5511'. On the right is a red circular node labeled 'FR'. A thick, light-brown arrow points from the '5511' node to the 'FR' node, with the label 'COUNTRY' centered on the arrow.

Relationship properties

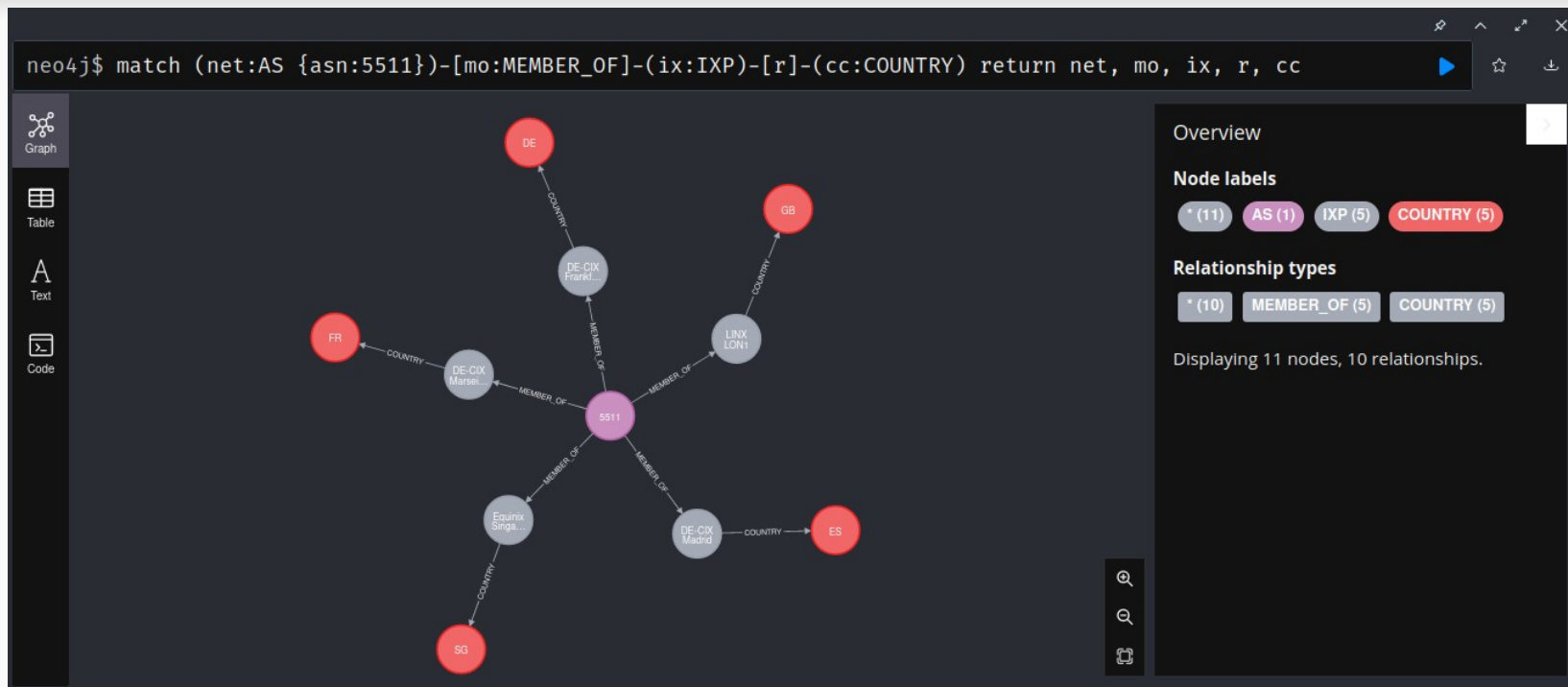
COUNTRY	
<id>	2180
reference_org	NRO
reference_time	"2022-12-09T00:00:00Z"
reference_url	<a href="https://ftp.ripe.net/pub/stats/ripenncc/nro-stats/latest/nro-delegated-stats">https://ftp.ripe.net/pub/stats/ripenncc/nro-stats/latest/nro-delegated-stats</a>
registry	ripenncc

# AS5511's country? (delegated-view2)

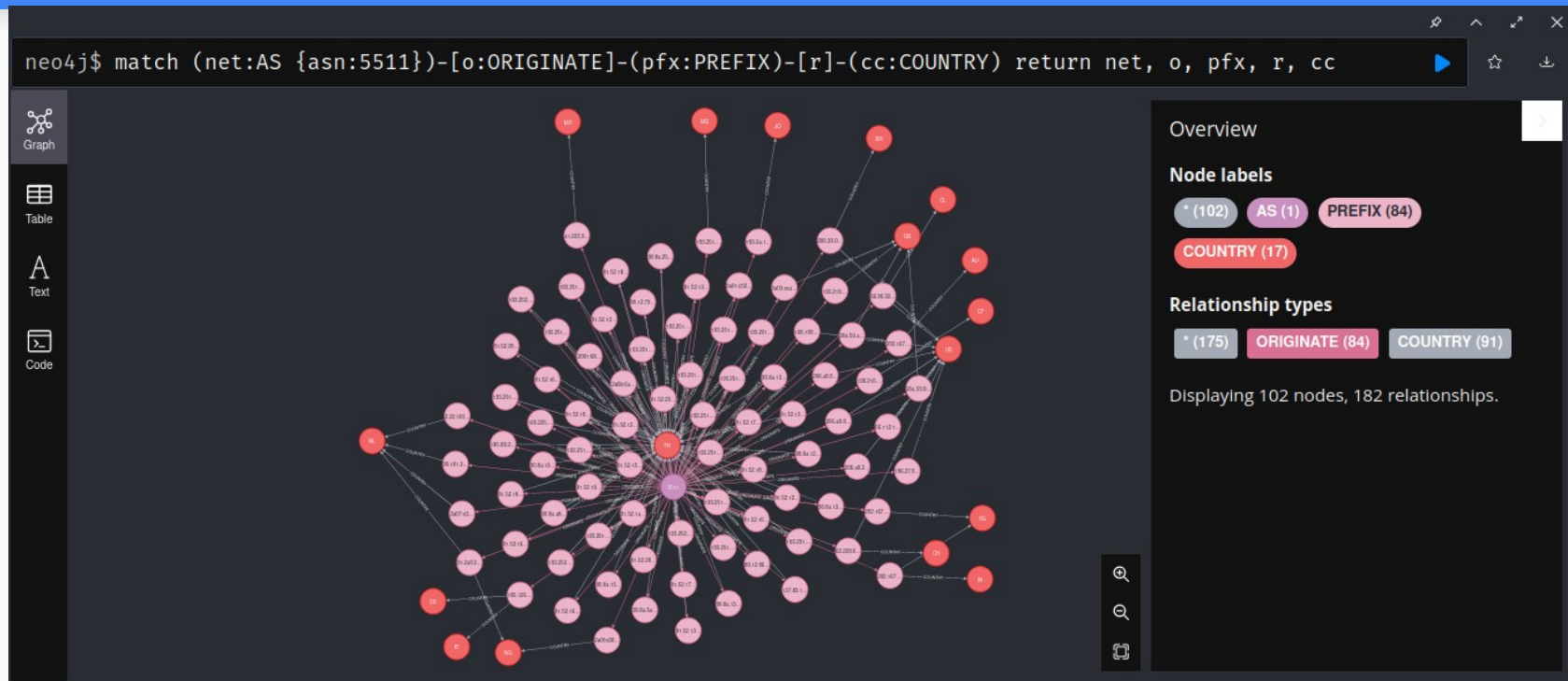




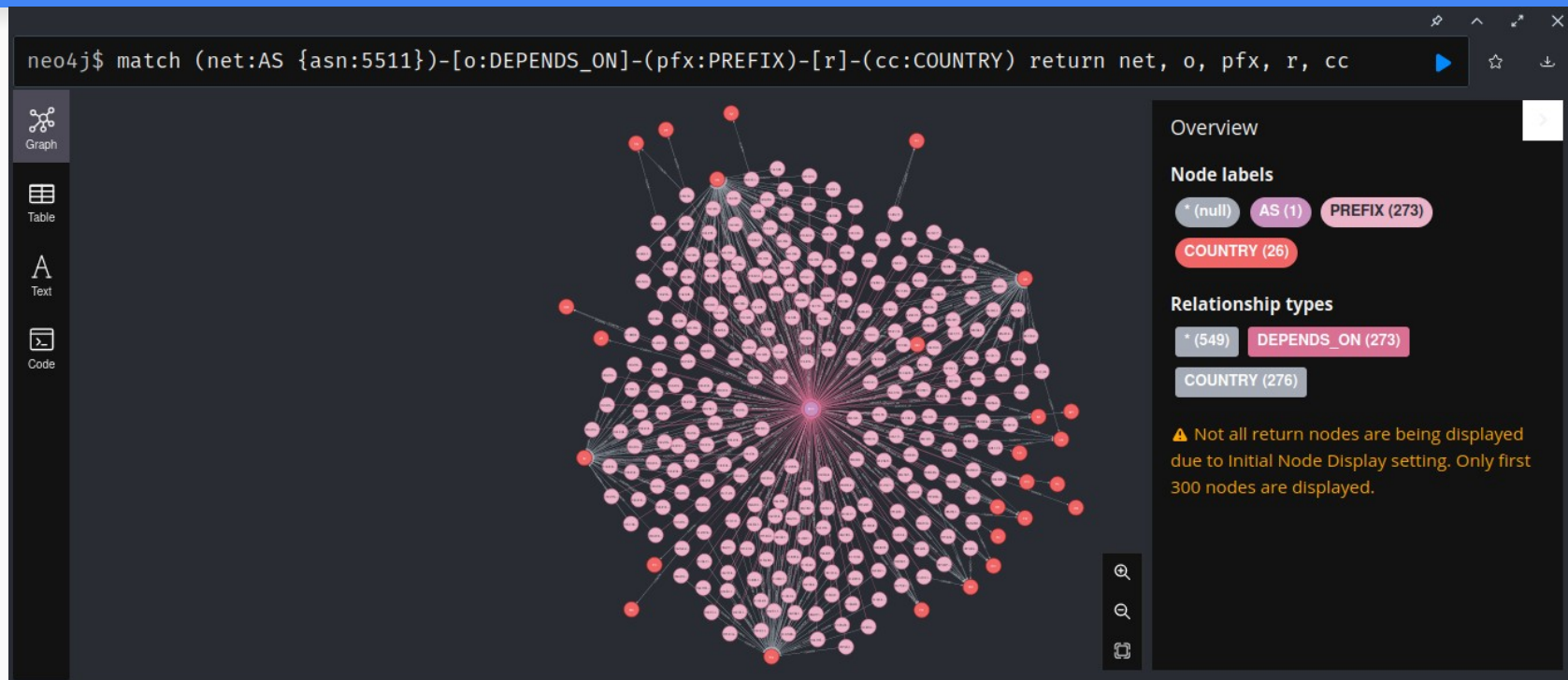
# AS5511's country? (PeeringDB-view)



# AS5511's country? (BGP/Maxmind-view)



# AS5511's country? (IHR/Maxmind-view)





# AS12389 domain names?

```
neo4j$ match (n:AS {asn:12389})-[:ORIGINATE]-(p:PREFIX)--(i:IP)--(d:DOMAIN_NAME)-[r:RANK]-(:RANKING) where r.rank<1000 return n,p,i,d
```

The graph displays the following nodes and relationships:

- Nodes:**
  - rt.ru** (Orange circle, DOMAIN\_NAME)
  - gosuslu...** (Orange circle, DOMAIN\_NAME)
  - 87.226.1...** (Green circle, IP)
  - 213.59.2...** (Green circle, IP)
  - 213.59.2...** (Green circle, IP)
  - 87.226.1...** (Pink circle, PREFIX)
  - 213.59.2...** (Pink circle, PREFIX)
  - 12389** (Purple circle, AS)
- Relationships:**
  - rt.ru** --RESOLVES\_TO--> **87.226.1...**
  - gosuslu...** --RESOLVES\_TO--> **213.59.2...**
  - gosuslu...** --RESOLVES\_TO--> **213.59.2...**
  - 87.226.1...** --PART\_OF--> **87.226.1...**
  - 213.59.2...** --PART\_OF--> **213.59.2...**
  - 213.59.2...** --PART\_OF--> **213.59.2...**
  - 87.226.1...** --ORIGINATE--> **12389**
  - 213.59.2...** --ORIGINATE--> **12389**

**Overview**

**Node labels**

- \* (8)
- AS (1)
- PREFIX (2)
- IP (3)
- DOMAIN\_NAME (2)

**Relationship types**

- \* (9)
- ORIGINATE (2)
- DEPENDS\_ON (1)
- PART\_OF (3)
- RESOLVES\_TO (3)

Displaying 8 nodes, 0 relationships.

# AS12389's siblings domain names?

neo4j\$ match (oid:OPAQUE\_ID)--(net:AS)--(pfx:PREFIX)--(ip:IP)--(dname:DOMAIN\_NAME)-[r]-(:RANKING), (net)-[reference\_org:'RIPE NC...]

**Overview**

**Node labels**

- \* (15)
- AS (3)
- PREFIX (3)
- IP (3)
- DOMAIN\_NAME (3)
- NAME (3)

**Relationship types**

- \* (15)
- DEPENDS\_ON (3)
- ORIGINATE (3)
- NAME (3)
- PART\_OF (3)
- RESOLVES\_TO (3)

Displaying 15 nodes, 0 relationships.

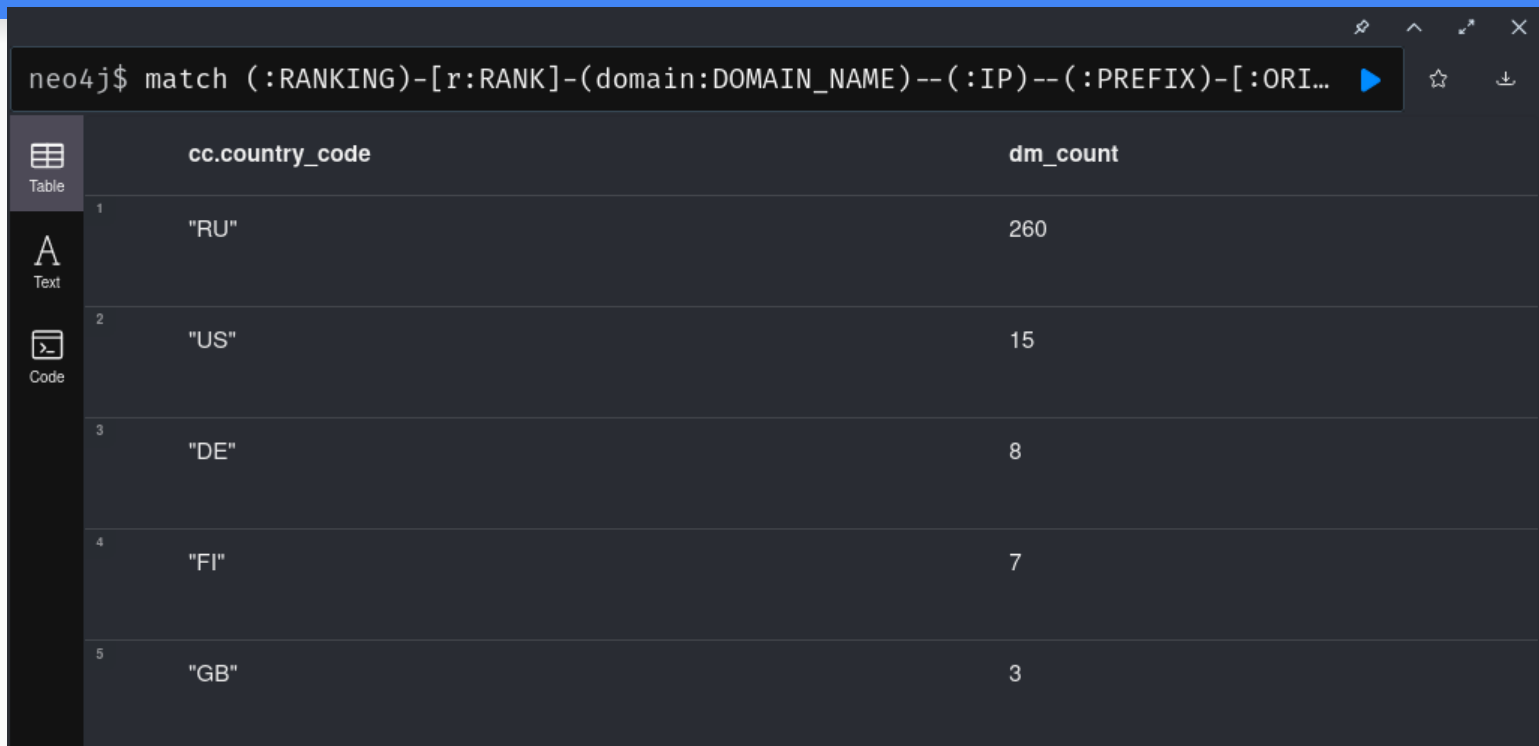
# Countries hosting top .fr domain names?

```
neo4j$ // .fr top 10k match (:RANKING)-[r:RANK]-(domain:DOMAIN_NAME)--(:IP)--(...
```

	cc.country_code	dm_count
1	"FR"	24
2	"US"	18
3	"GB"	1
4	"NL"	1
5	"ZZ"	1



# Countries hosting for top .ru domain names?



The image shows a Neo4j Cypher query interface. The query entered is `match (:RANKING)-[r:RANK]-(domain:DOMAIN_NAME)--(:IP)--(:PREFIX)-[:ORI...`. The results are displayed in a table with two columns: `cc.country_code` and `dm_count`. The table lists the top 5 countries by domain count for .ru domains: RU (260), US (15), DE (8), FI (7), and GB (3).

```
neo4j$ match (:RANKING)-[r:RANK]-(domain:DOMAIN_NAME)--(:IP)--(:PREFIX)-[:ORI...
```

	cc.country_code	dm_count
1	"RU"	260
2	"US"	15
3	"DE"	8
4	"FI"	7
5	"GB"	3

# What's next?

- Planning weekly database publication
- More datasets coming soon
- Discuss more with the community (dataset/applications/ontology)
- Graph analysis
  - AS/prefix/IP classification (node embedding)
  - Singularities in the graph
  - Data inconsistencies / semantically incorrect

# Get Involved!



Tell us:

- What are your favorite datasets?
- How would you use IYP?

Consider:

- Adding your own data
- Browsing/querying IYP
- Starting your own IYP

<https://github.com/InternetHealthReport/internet-yellow-pages>

romain@iij.ad.jp

@romain\_fontugne



# Benefits of using IYP for research

- IYP-query-based research study means:
  - Reproducibility (open data)
  - Easy data sharing
  - Constantly updated results
  - Future proof: new dataset, same query

# A lot of seekers

- Dedicated websites:
  - RIPEstat, bgp.he.net, bgp.tools, Cloudflare's radar...
- Numerous users:
  - RIPEstat: 1.5+ million unique clients/IPs (daily)
  - IHR/ASRank: Thousands unique visitors per months



Internet Health Report

@ihr\_alerts

Cogent disconnecting from one of Russia's largest ISP, Transtelecom

[ihr.iijlab.net/ihr/en-us/netw...](https://ihr.iijlab.net/ihr/en-us/netw...)



Find what's up in your world



Home

Reports

Documentation

API

Contact

AS20485 - TRANSTELECOM Joint Stock Company TransTeleCom, RU



## Tweet Analytics



Internet Health Report @ihr\_alerts · Mar 6



Cogent disconnecting from one of Russia's largest ISP, Transtelecom

[ihr.iijlab.net/ihr/en-us/netw...](https://ihr.iijlab.net/ihr/en-us/netw...)



45



40



2

12:14 A

View

32 Ret

Impressions ⓘ

121K

Engagements ⓘ

2,977

Detail expands ⓘ

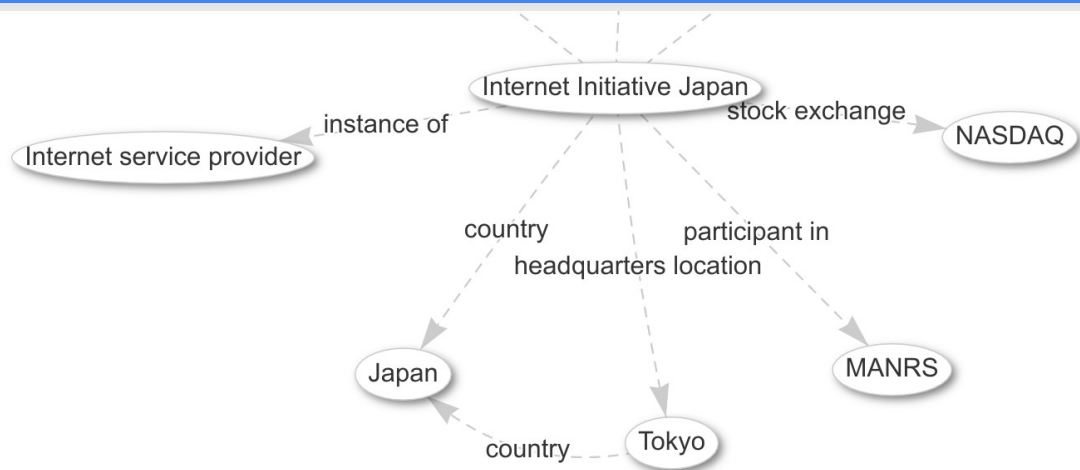
1,362

New followers ⓘ

Profile visits ⓘ

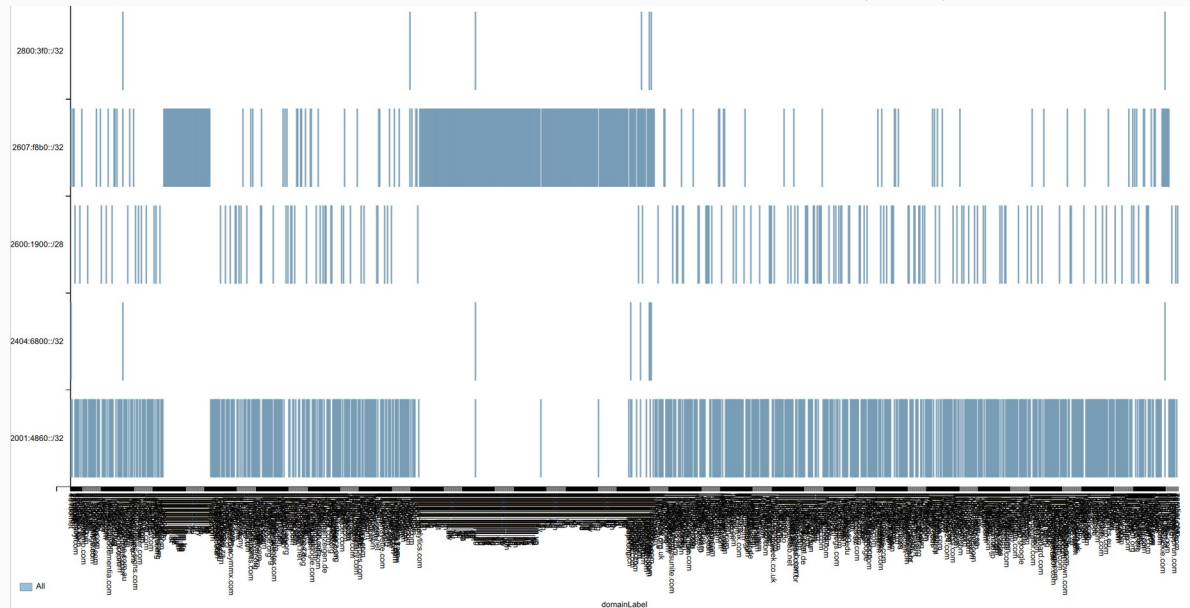
# Wikidata 101

- Item (node, QID):  
IIJ, ISP, Tokyo, Japan, MANRS
- Property (link, PID):  
Instance of, country, participant in
- Statement (one hop):  
IIJ - instance of - ISP



# Query example 2: Google/route server

Popular domains for Google prefixes seen at DECIX route server (IPv6): <https://tinyurl.com/yfszs7qq>



<https://tinyurl.com/ygqhpnjm>

[illegible]



# Query example 4: Prefixes across IXPs (Netflix)

Netflix IPv6 prefixes per IXP: <https://tinyurl.com/yzn5tkwl>



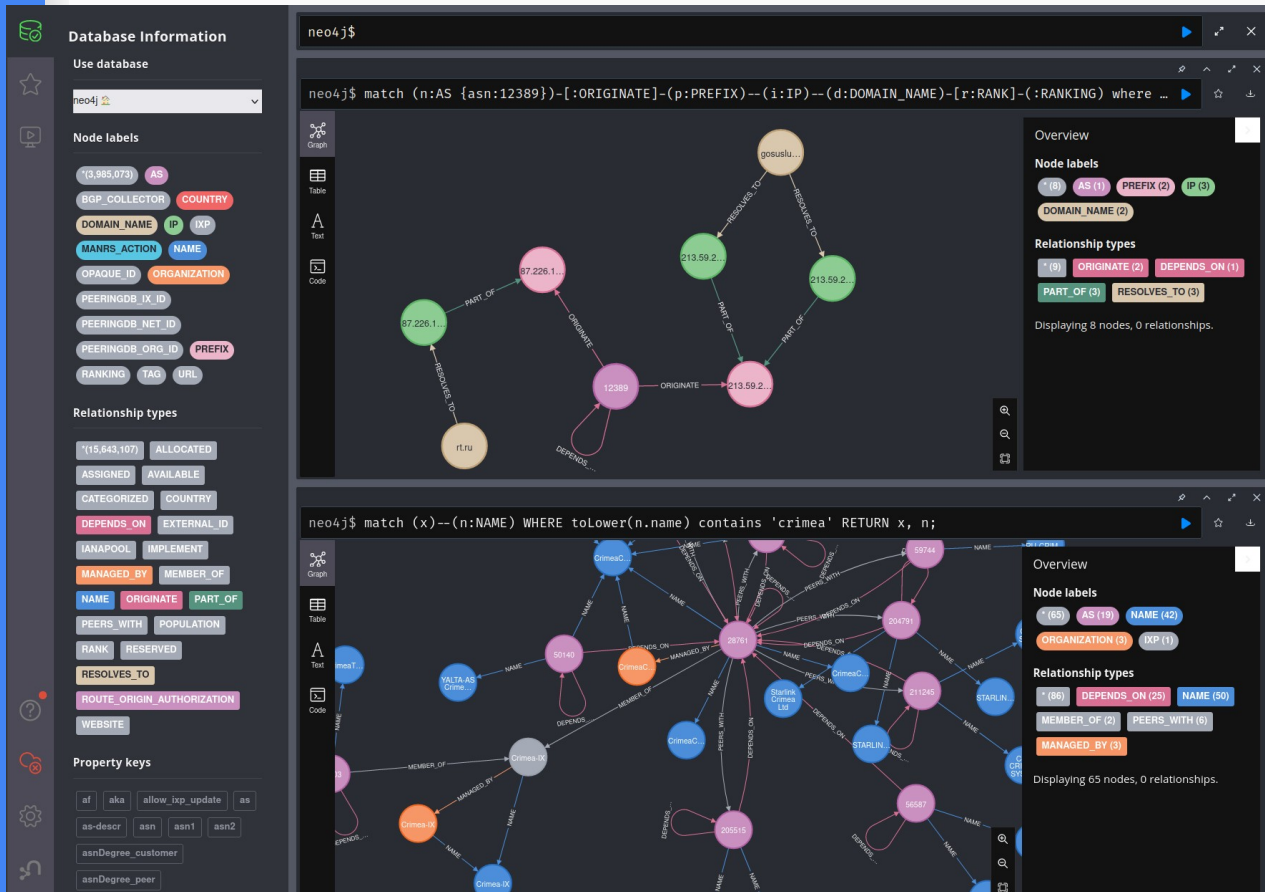
# Neo4j

# Graph database

## Highly scalable, schema free

## Community & tools:

- Machine learning / AI
- Cypher query language
- Multiple web interfaces



# Why not relational DB?

- Graph DB

- + bottom-up development  
what data is available? What can we learn from it?
- + easier to start with
- + easier to JOIN datasets
- + easy to add new datasets
- more efforts to get a first usable graph
- hard to integrate time series?

- Relational DB

- + top-down approach  
Designed for a specific application
- + good when you know what you are looking for
- + fast for a few datasets
- harder to JOIN multiple datasets
- harder to import new datasets make it evolves
-