Measuring the Quality of Number Resource Registration Data

Methodology and First Results

Robert Kisteleki
Science Group Manager, RIPE NCC
robert@ripe.net

The need to measure data quality

Why have we started such measurements?

- Because we want to be aware of the quality of our data:
  - In order to demonstrate that our data quality is good
  - Or if it isn't, we want to know what's wrong and fix it
- Because it's our duty to be good shepherds of data!
  - Therefore our strategy also focuses on data quality

Historical background

RIPE NCC's registration data has lots of history:

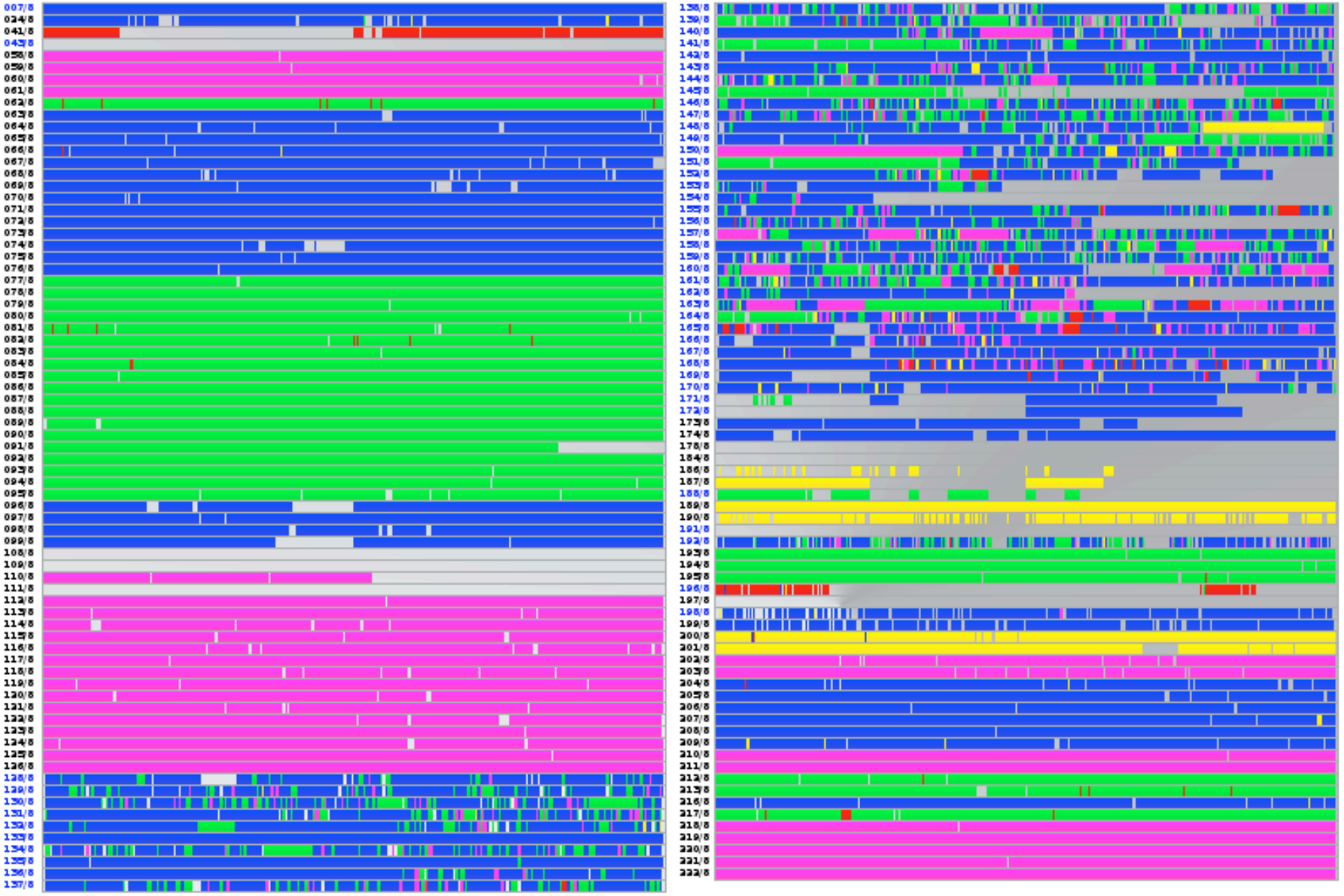
- Legacy data
- Policy changes over time
- One-time events like Early Registration Transfer (ERX, in 2003-2004) and the birth of AfriNIC
- Database entries managed (or not) by LIRs/users

Historical background

Back in the day, IPv4 objects were voluntarily put into RIPE DB, and some of these entries are "stuck" there:

- We cannot delete them (easily)
- Sometimes we cannot even update them
- "Registry of last resort", anyone?

Spot the ERX space!



RIPE NCC ARIN APNIC LACNIC AFRINIC

Other incentives

We need this already and want to prepare for things to come:

- 2007-01 (“Contractual Requirements for PI …”)
- 2009-01 (“Global Policy for the allocation of IPv4 blocks to RIRs”)
- Intra- and inter RIR transfers (if/when they happen)
- Resource Certification

So what is RDQ?

Registration Data Quality is:

- A RIPE NCC internal effort to measure, monitor and enhance internal data quality
- Main goals are:
  - Be aware of exactly what resources we are responsible for, and evaluate our confidence level in this.
  - To be able to answer who is the current legitimate holder of a specific resource and what is our confidence level in this.

Focus

At this stage we're focusing on IPv4 related data

- Mostly because of the many aspects of IPv4 run-out
- This is also considered to be the "dirtiest" pool of all
  - Compared to ASN and IPv6 data
- This has the most historical artefacts, too

At a later stage we intend to include IPv6 and AS numbers too.

Approach

There are two phases in the effort, to answer the two “naïve questions”:

- Question 1: Is the RIPE NCC responsible for this (IPv4) resource?
- Question 2: If the answer to question 1 is yes, who is the legitimate holder?

Question 1:
"Is the RIPE NCC responsible…?"

We've built a framework to support the effort:

- It can evaluate all known and available databases (internal and external)
- It can check if all of these have consistent information
- It can do this evaluation periodically:
  - As of "now" as time goes on
  - Historically – how did the quality change over time?

We're measuring regularly, and have started to fix the inconsistencies.

Question 1:
"Is the RIPE NCC responsible…?"

Databases we're including in this phase:

- Internal registration databases
- RIPE NCC reverse DNS services
- "Stats" files from all five RIRs
- Historical transfer information (2003-2005)
- IANA assignments/allocations pages

Not included in this stage of the evaluation:

- RIPE DB
- Other RIR's "whois" databases

Question 1:
"Is the RIPE NCC responsible…?"

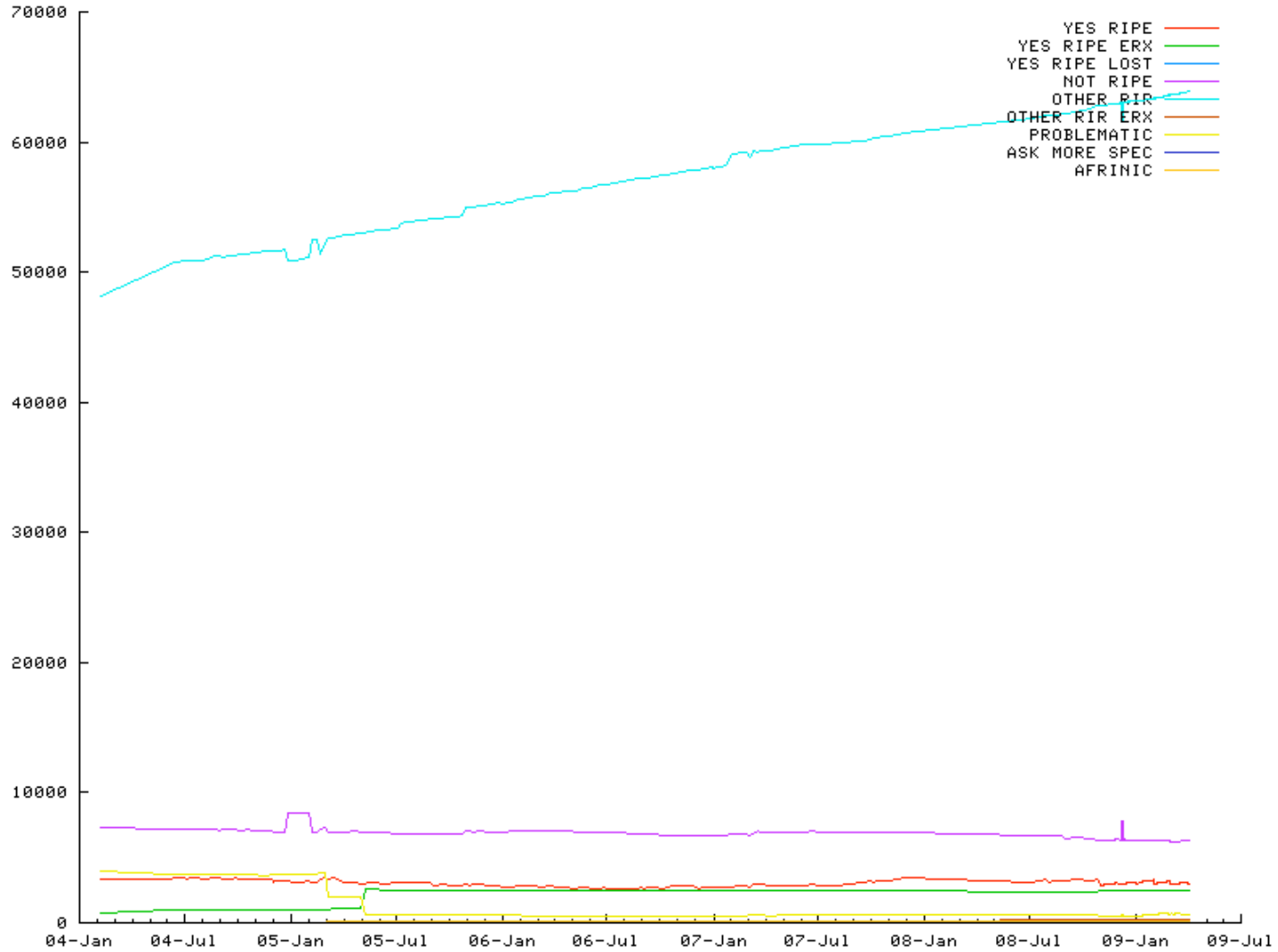
Our approach is simple:

- Each data source is asked whether it thinks a resource (IPv4 prefix) is NCC space or not.
  - The answer is basically: Yes (Y), No (N), Don't know (U)
- Responses from all data sources are combined to form a "gene sequence" for that prefix:
  - Ie. YYUUUUYYYY is good, YUNUUUYNYN not so much
- We group all prefixes having the same gene sequence
  - Prefixes with the same gene sequence have similar properties
  - Therefore we can fix groups of issues, not individual ones

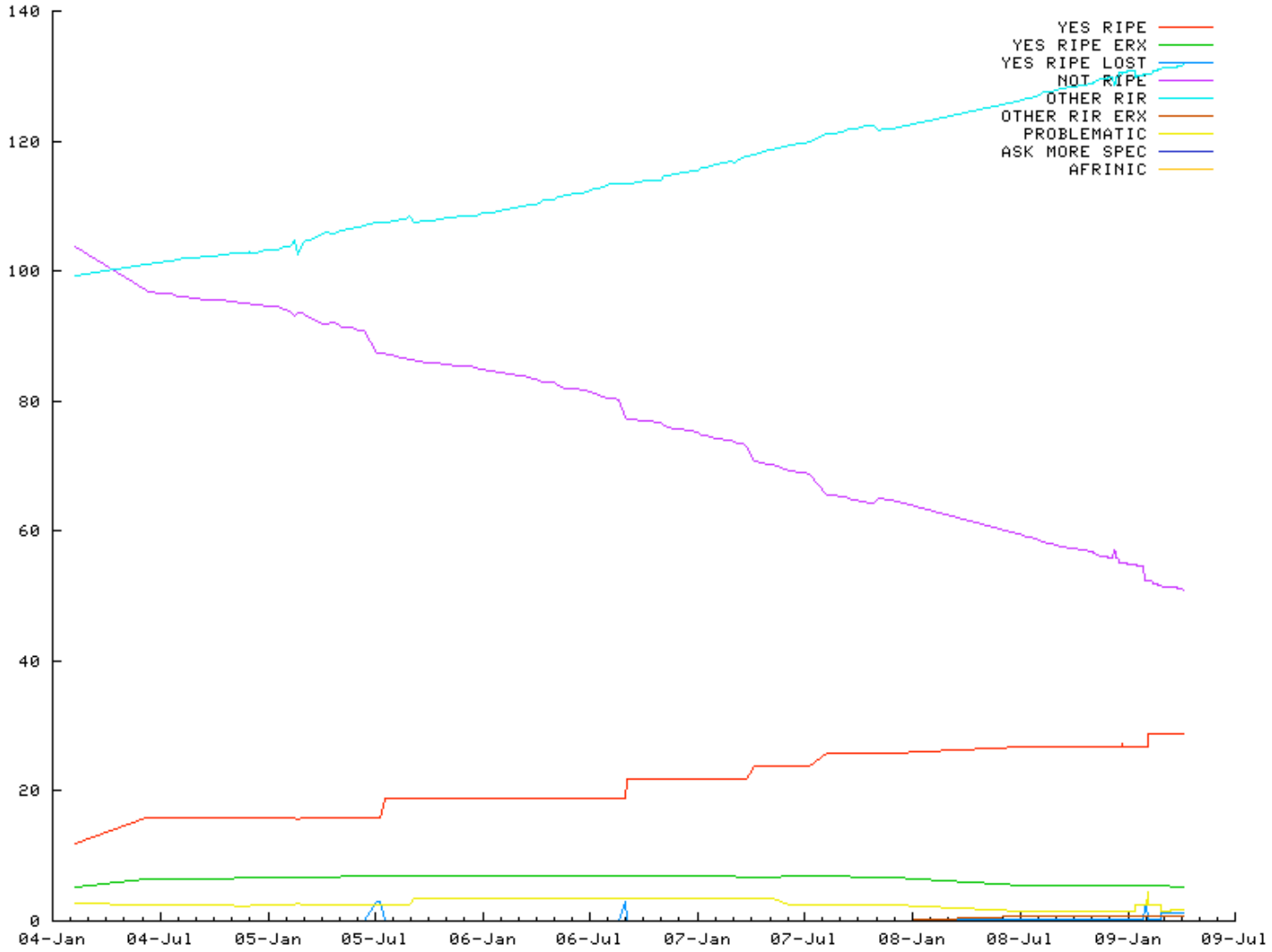
Current approximate numbers

Category	Prefixes	In terms of /8s
RIPE NCC	**~5.500**	**~34**
• *Allocated to RIPE NCC by IANA*	*~3.000*	*~29*
• *Transferred to RIPE NCC from other RIRs*	*~2.500*	*~5*
Claimed by other RIRs	**~64.000**	**~132**
Transferred to other RIRs from RIPE NCC	**~300**	**~1**
Unallocated by IANA / unclaimed by RIRs	**~6.300**	**~48**
Inconsistent	**~830**	**~5**
• *Legacy & ERX space*	*~800*	*~5*
• *Allocated to RIRs*	*~30*	*< 0.1*
Other (like 10/8, 127/8)	**3**	**3**
Total	**~77.000**	**223**

Changes over time…
(Number of prefixes)

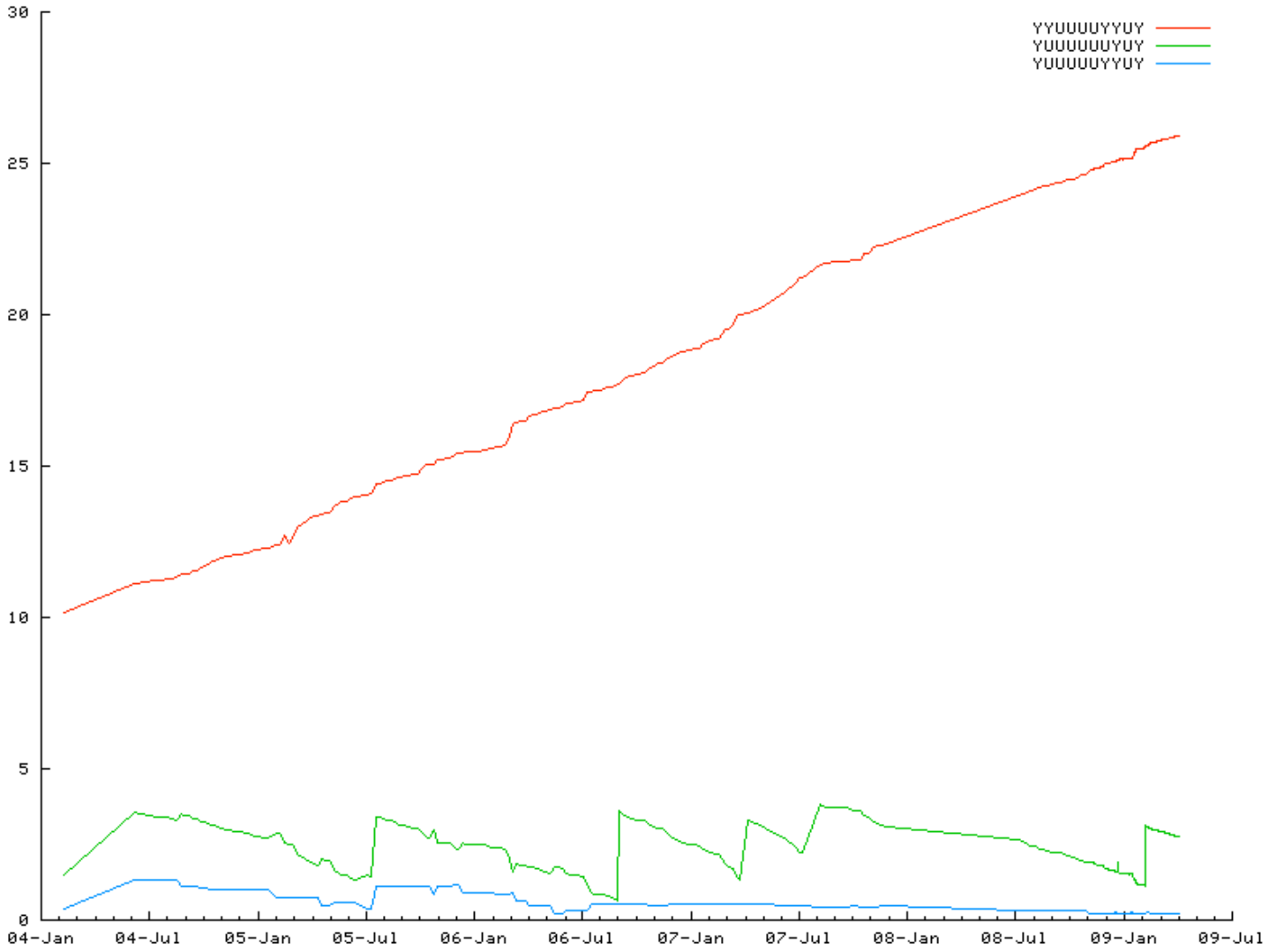


Changes over time…
(Size of address space)

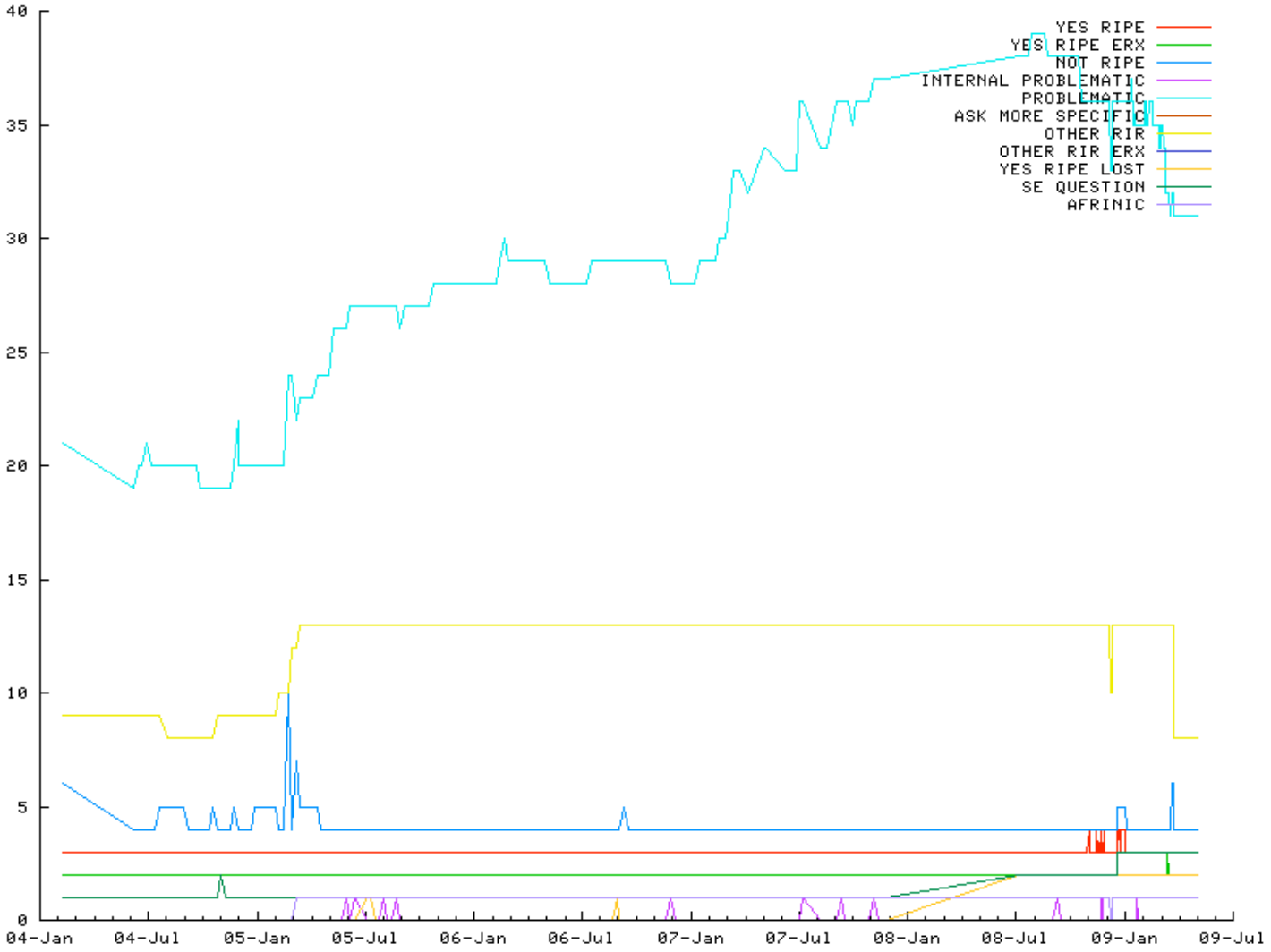


Changes over time…

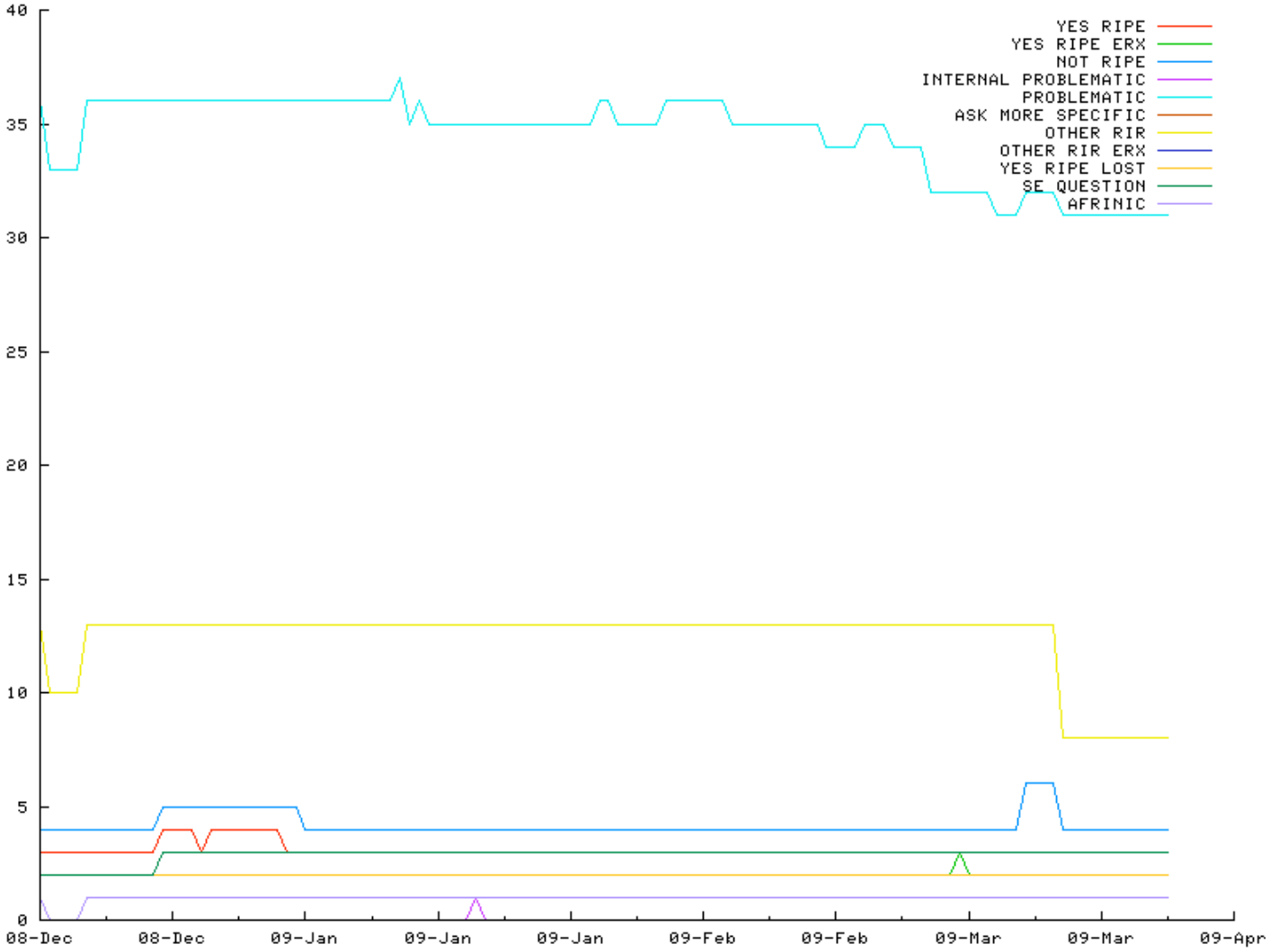
(Size of “allocated to RIPE NCC” address space)



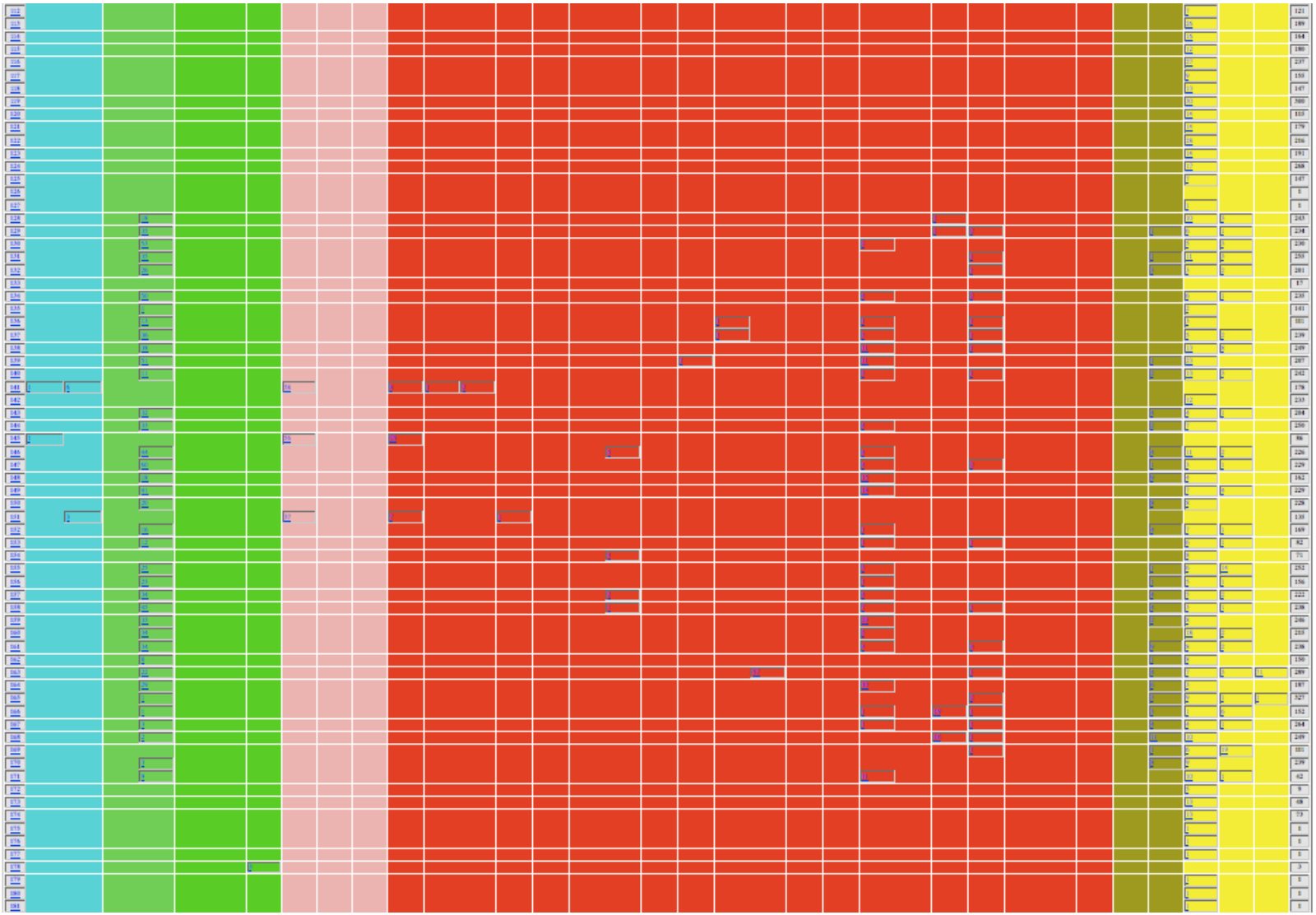
Changes over time…
(Number of gene sequences per category)



Changes over time…
(Number of gene sequences per category)



The Matrix



There's more to do…

Question 2 ("Who is the legitimate holder? What is our confidence level in the answer?"):

- We are in the process of building the framework for this phase.
- It will try to match actual data (like contact information) found in different databases:
  - This will likely include the RIPE DB
  - Probably will incorporate routing information too
  - Difficult because of differences in database structures, historical artifacts and because results are not quantifiable.

Further inspiration from RDQ

The whole exercise led us to think about “higher level” consistency checks, eg. Inter-RIR consistency checks:

- ERX space
- Reverse DNS services
- “whois” databases

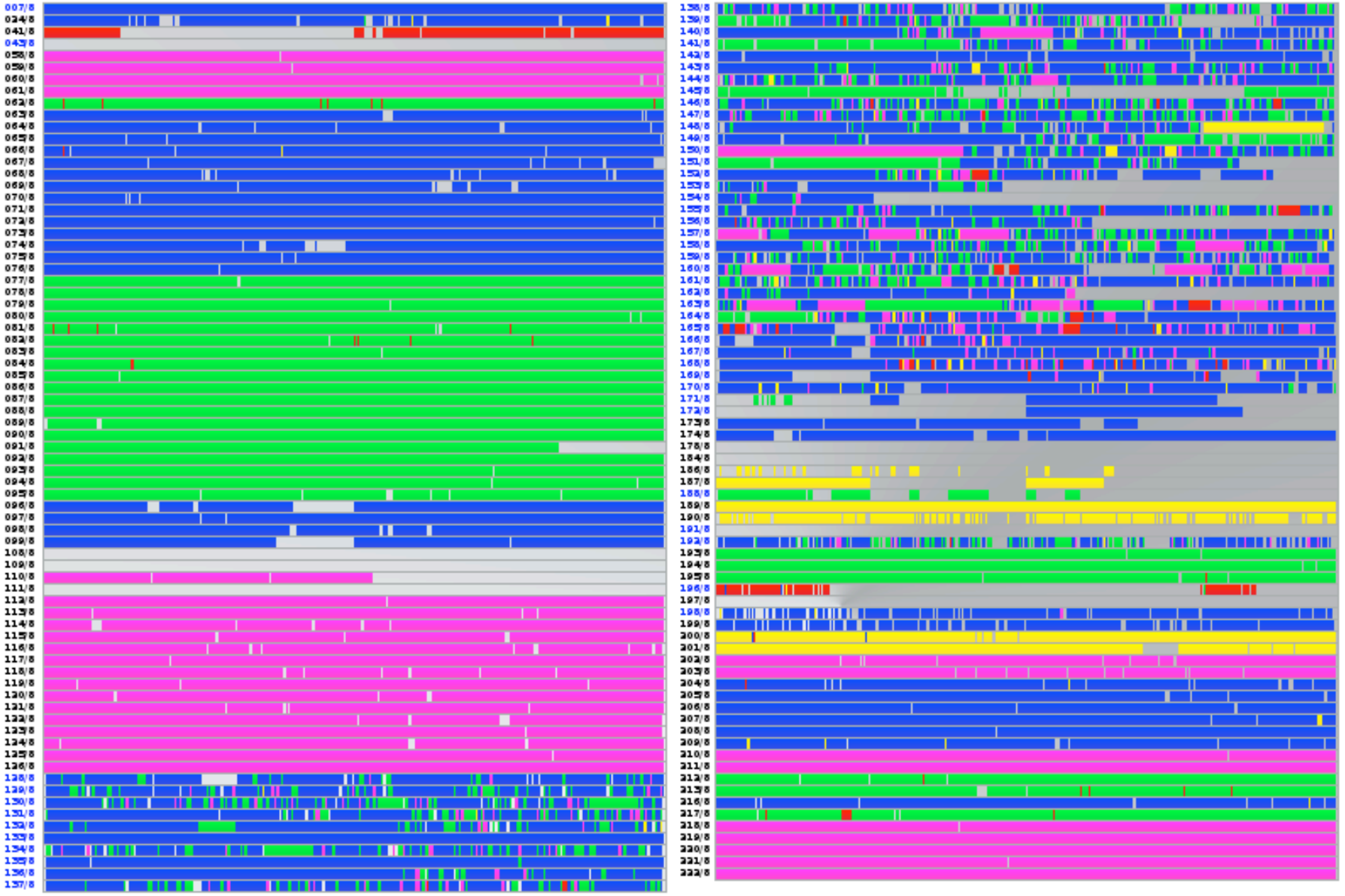
We cannot do all of these alone, as most of these need large scale access to other RIR’s and IANA’s databases!

Finally

Summary:

- Quality of our registration data is important for us
- We are working on continually measuring and enhancing our data
- Most issues are in the legacy and ERX space
- We want to be prepared to answer “tough” questions too – eg. ERX related ones
  - These are the least well maintained
  - Therefore they are more prone to misuse, as “clean” IPv4 runs out.

Spot the ERX space!



RIPE NCC ARIN APNIC LACNIC AFRINIC

Questions?