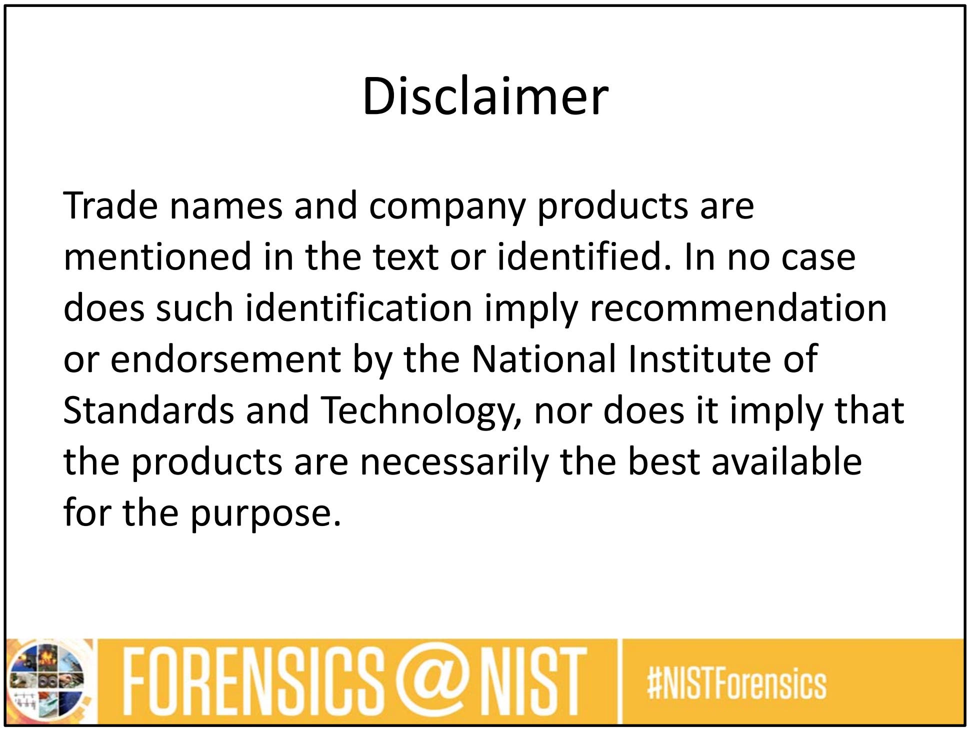
NIST National Institute of Standards and Technology U.S. Department of Commerce

The NSRL and Video Games

Why I Get to Buy Video Games at the Office

FORENSICS @ NIST #NISTForensics

Hi everyone, my name is Austin Snelick, and I work with the National Software Reference Library, or NSRL, here at NIST.

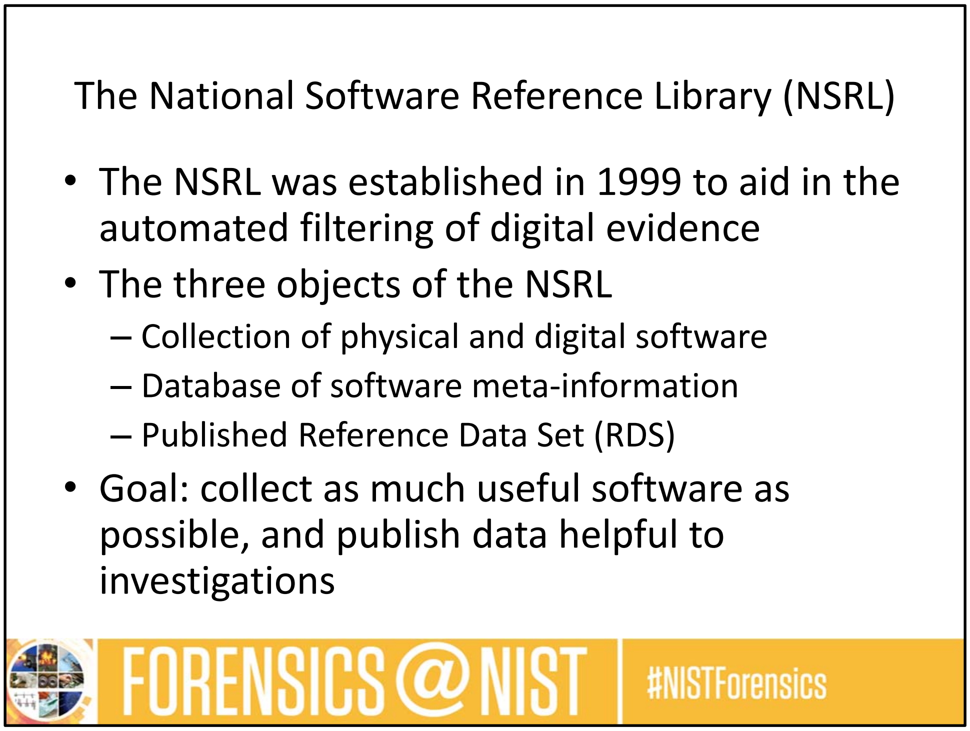


Disclaimer

Trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

FORENSICS @ NIST #NISTForensics

Just a quick disclaimer; I'll be mentioning names and products during this talk, but I or the NSRL or NIST are not endorsing any of them.



The National Software Reference Library (NSRL)

- The NSRL was established in 1999 to aid in the automated filtering of digital evidence
- The three objects of the NSRL
  - Collection of physical and digital software
  - Database of software meta-information
  - Published Reference Data Set (RDS)
- Goal: collect as much useful software as possible, and publish data helpful to investigations

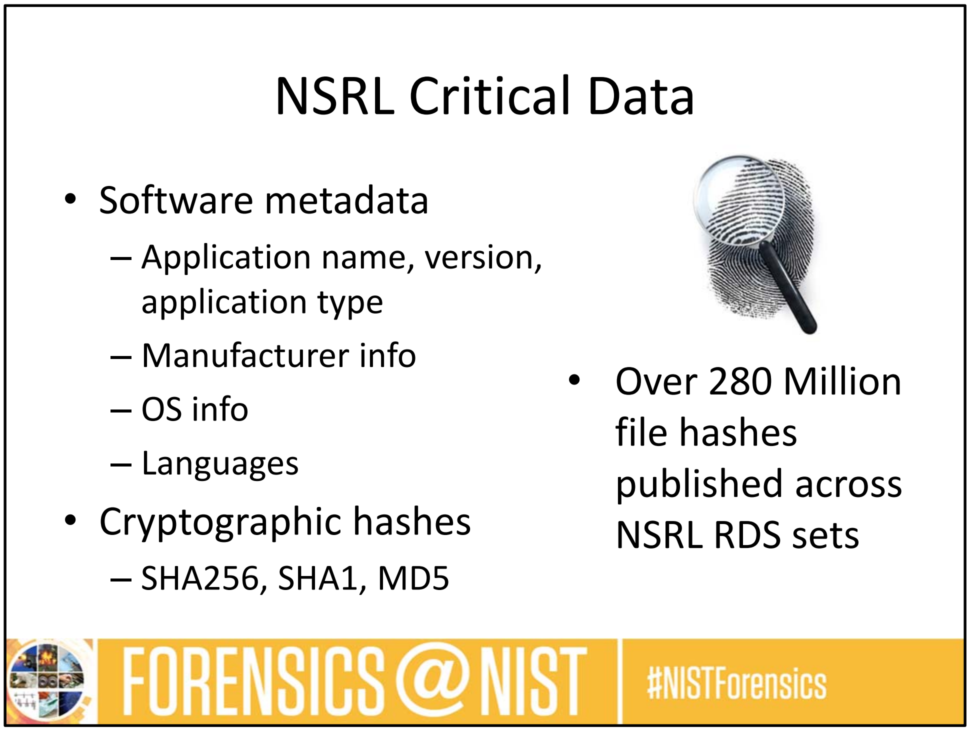
FORENSICS@NIST #NISTForensics

A little background about the NSRL is needed so you can understand why it is actually important to the law enforcement community, and others, that we collect games, and to hopefully convince you that we're all not just sitting around playing games all day up in the NSRL.

The NSRL was established in 1999, to aid in the automated filtering of digital evidence.

Conceptually, the NSRL is 3 objects.
First, it is a collection of physical and digital software.
Second, the NSRL is a database of meta-information about the software, such as application name, publisher, date published, and many others.
And third, the NSRL is a subset of the database, the Reference Data Set (which we refer to as the RDS), which is the data that we publish.

In essence, we try to get as much software as we can get our hands on, physically or digitally, and extract every file, get the useful meta-information from the software, and publish the data that would be important in an investigative setting.



NSRL Critical Data

- Software metadata
  - Application name, version, application type
  - Manufacturer info
  - OS info
  - Languages
- Cryptographic hashes
  - SHA256, SHA1, MD5
- Over 280 Million file hashes published across NSRL RDS sets

FORENSICS @ NIST | #NISTForensics

For each piece of software we get, we need to collect information like: the application name, versioning information, the application type, who manufactured the software, what operating systems does the software run on, what languages does it support, and some other data.

In the traditional NSRL way of doing business, this metadata collection is done by our lab crew, who will read this information off a physical piece of software, like a CD, or from the website where the software was downloaded if it was a digital download, and they then enter that information manually into our database.

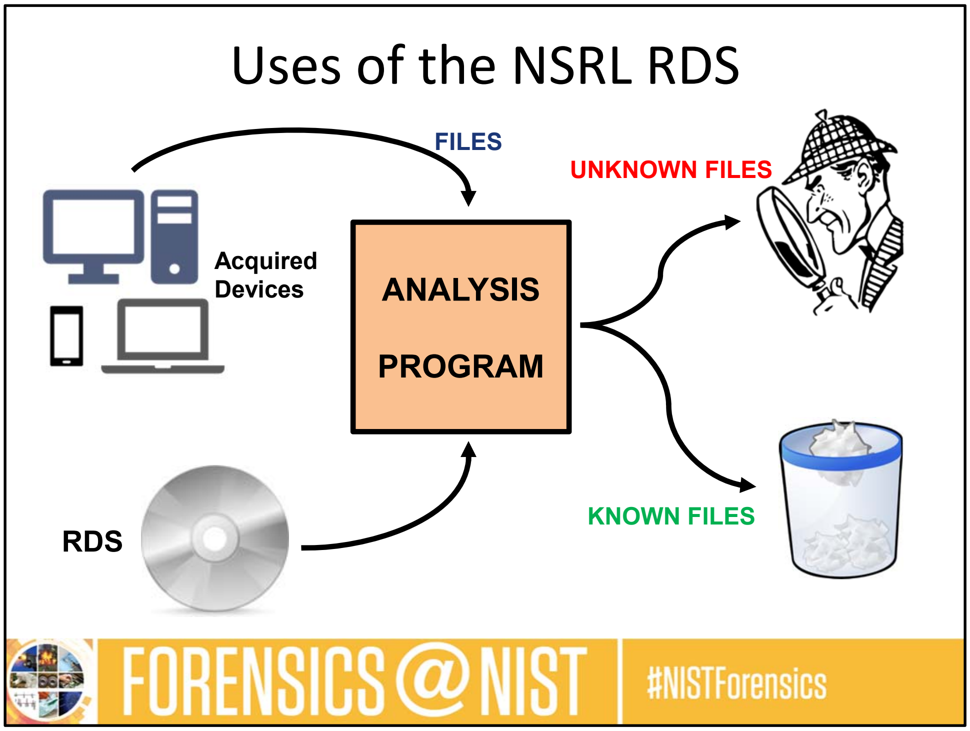
Once we have collected a piece of software and its metadata, we then create a media image of the software, and store that image in our digital software repository.

From that image, we then unpack the software to find as many files as we possibly can, and for every file that we uncover, we generate cryptographic hashes.

In simplest terms, a cryptographic hash of a file is like a fingerprint, that can uniquely identify the file.

So if we make a SHA256 hash of a file, we know that only the exact same file, down to the byte level, would have the exact same SHA256 hash value.

Currently, the NSRL has in its collection over 280 million hashed files published across our RDS sets.



Uses of the NSRL RDS

Acquired Devices

FILES

ANALYSIS PROGRAM

RDS

UNKNOWN FILES

KNOWN FILES

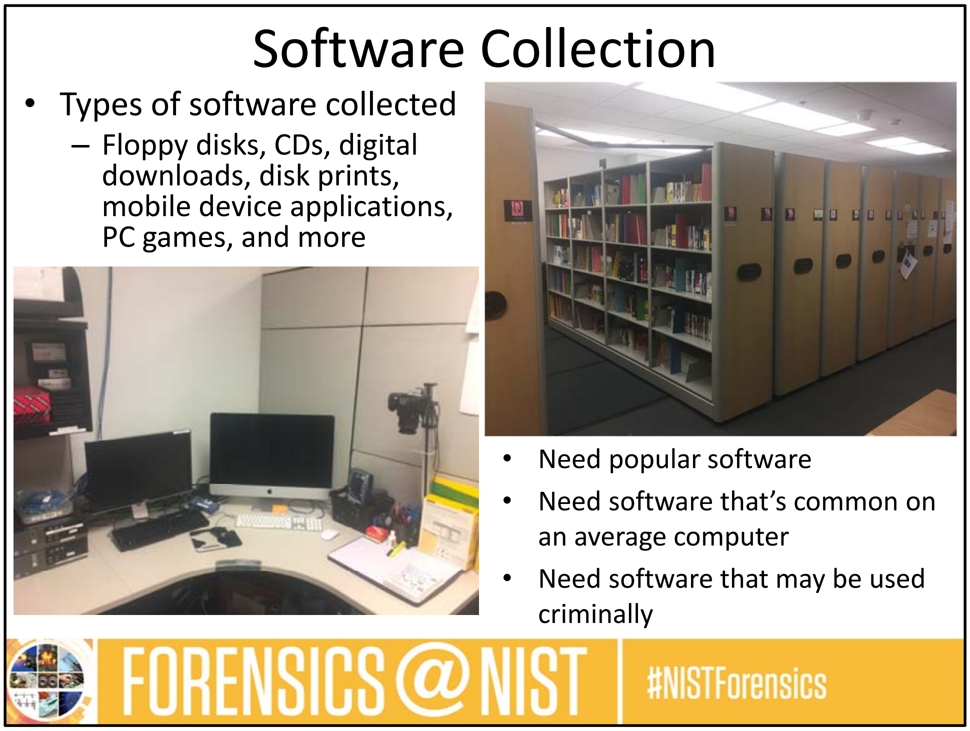
FORENSICS @ NIST #NISTForensics

So, how is this set of file hashes actually useful in the real world?

In an investigation, the examiner will have obtained a device, which may be a laptop, phone, or hard drive.
Using an analysis program, the files stored on the device are hashed and compared against the file hashes that we have collected and published in our Reference Data Set.
The result of the comparison between the acquired device, and the RDS file hashes is a set of matched files, and unmatched files.
Matched files are what we would call known files, as they are things that we have collected information about and publish in the RDS.
Unmatched files are what we would call unknown files, as we don’t know anything about these files.
In an investigation, the examiner will typically want to focus on the unknown files, as these files are likely to contain incriminating or exonerating evidence important to an investigation, if in fact this evidence does exist on the acquired device.
In some cases, examiners are looking for specific files on an acquired device, but I’ll talk more about this later.

Being able to focus on the unknown files in an investigation can really cut down on the time it takes to analyze a device.
That means it is important for us to not only collect as much software as we can get

our hands on, but also make sure we are collecting relevant, popular software that we would be likely to find on anyone’s computer.



Software Collection

- Types of software collected
  - Floppy disks, CDs, digital downloads, disk prints, mobile device applications, PC games, and more

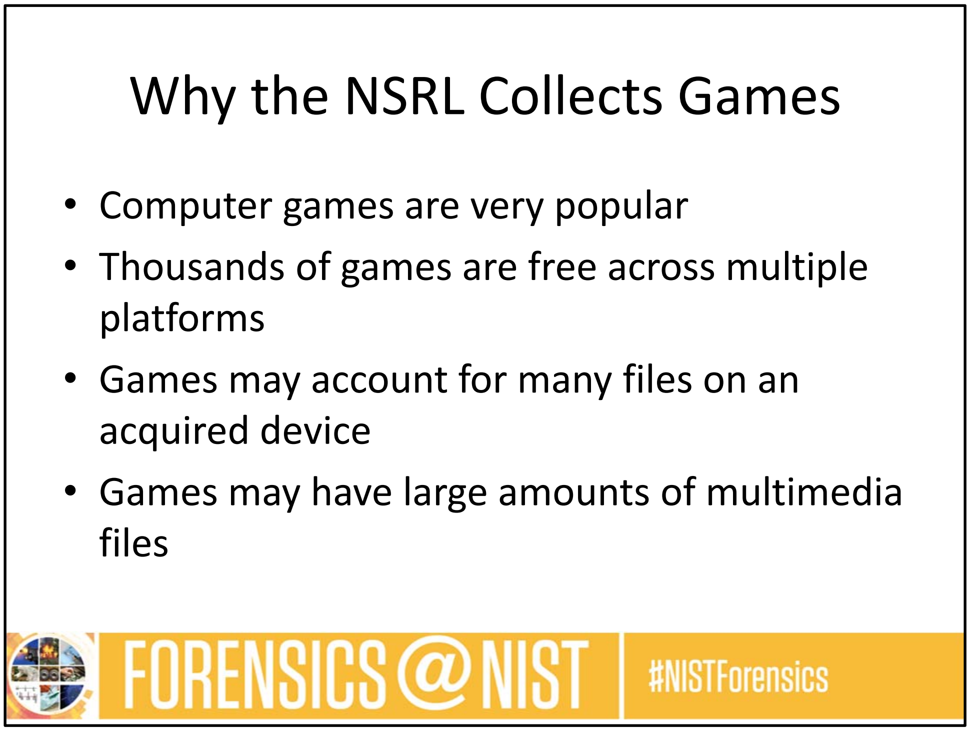
- Need popular software
- Need software that's common on an average computer
- Need software that may be used criminally

FORENSICS @ NIST #NISTForensics

When the NSRL first started back in 1999, software was only available physically, so we collected floppy disks and CDs.
As software evolved it became available to us as digital downloads.
In recent years, we expanded our collection to include disk print data, which helps us find files from software after it has been installed.
So, this would help us find what files get created when installing a piece of software.
As smart phones have become a larger part of everyday lives, we started collecting mobile apps for both Apple iOS and Android phones.
And lately we have been putting more emphasis on collecting PC games.
Games have been in the RDS for a while now, but more recently we have created ingest processes to collect games from some of the larger gaming platforms, on a regular basis.

When we decide what kind of software to collect, we have to keep in mind what will be most useful to our customers.
With the goal of the RDS to be a source for discovering and separating known and unknown files, we want to maximize the number of known files.
We want to have a set of files that are popular, and software that is common for the average person to have on their computer.
For example, operating systems and operating system files are software that any functioning computer is going to have.
We also receive requests to collect more specific software, which may be used in

cases where an investigator wants to look at the known files, because it’s software that may be used in a criminal manner.
For instance, we’ve collected keyloggers, and we’ve also collected flight simulation software, which we collected after the Malaysian airlines flight disappearance a few years back.
On our own, we also have to think about how criminals might use software in a criminal way.
For example, 3D printing software, which alone is harmless, but could be used to print something like a gun, which would not be tracible via a serial number, or have any documentation.
So we do have to keep in mind how criminals might be using software maliciously when we decide on what things to collect.

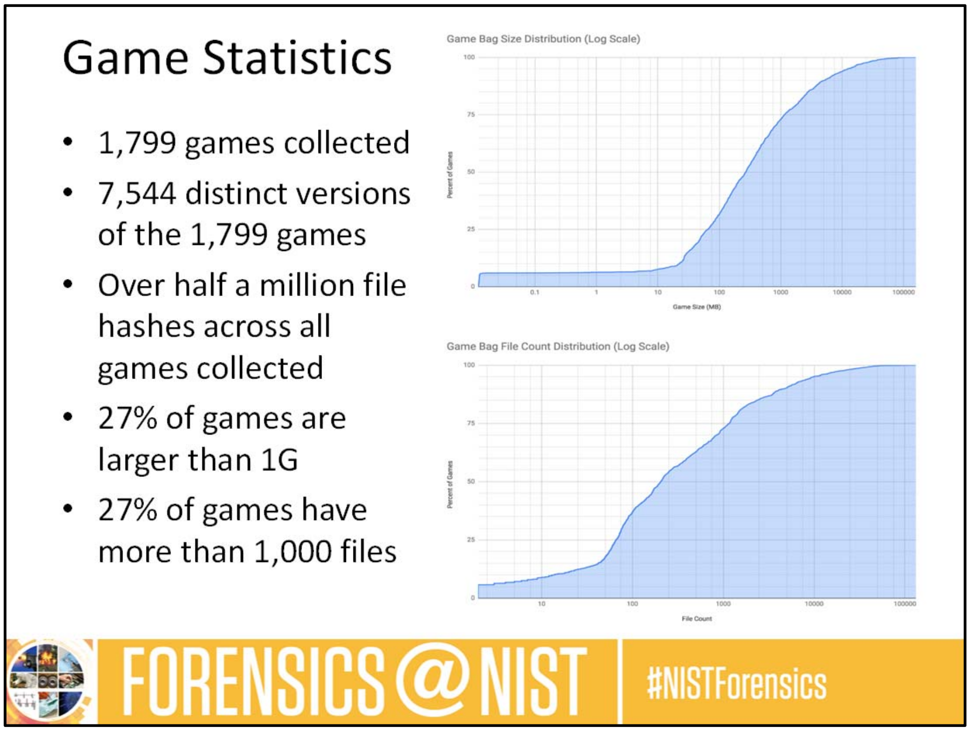


Why the NSRL Collects Games

- Computer games are very popular
- Thousands of games are free across multiple platforms
- Games may account for many files on an acquired device
- Games may have large amounts of multimedia files

FORENSICS @ NIST #NISTForensics

So, why do investigators care about PC games being a part of the NSRL's RDS?
Well, the answer goes back to how we make decisions on what software to collect.
Computer games are very popular, as thousands of games are available for free amongst multiple platforms.
For gamers, the majority of their hard drive is going to be used to store games.
The size of games, and the number of files a single game can have is also an important factor.
We've seen games that are over 150Gs in size, and some games that have over 100,000 files in them, which are just huge games.
We also care about the types of files that are in games.
Some games may have a large number of multimedia files in them, and for investigations that focus on searching a computer for 'illegal' images (and I'll let you take a guess on what some illegal images might be), it can be a huge waste of time for an examiner to have to look through a lot of images that come from games.
So if we can flag a lot of these multimedia images as known files when comparing to our RDS, that can really cut down on examination time.

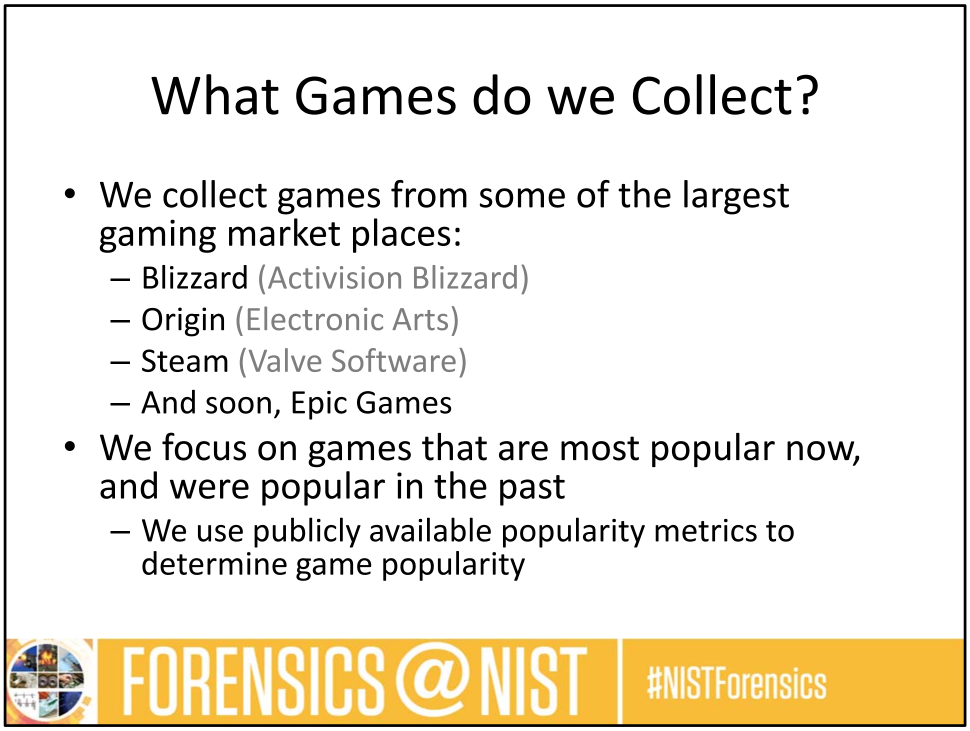


Game Statistics

- 1,799 games collected
- 7,544 distinct versions of the 1,799 games
- Over half a million file hashes across all games collected
- 27% of games are larger than 1G
- 27% of games have more than 1,000 files

Some statistics on the games we've collected so far:
We have collected about 18 hundred games, of which we have multiple versions of many of the games.
In total, we have over 75 hundred distinct versions of the 18 hundred games.
In total we have found over half a million files in the 18 hundred games, which equates to half a million more hashes in our RDS.
And these numbers come from just the games that we currently have published, but we are planning on adding over a million more game hashes in our next RDS publication.

The two charts I have on the right are distributions for game size and file counts from the games we have collected.
It should be noted that these charts are on a log scale.
Of all the games we have, about 27% are greater than 1G in size.
We've also found that a large majority of games have greater than 100 files in them, and about 27% of games have 1,000 or more files in them, which is a good chunk.
If you could imagine that even if half of those 1,000 files was multimedia images, that's a lot of time that would go towards having to look at each image, if they don't get matched with the RDS.



What Games do we Collect?

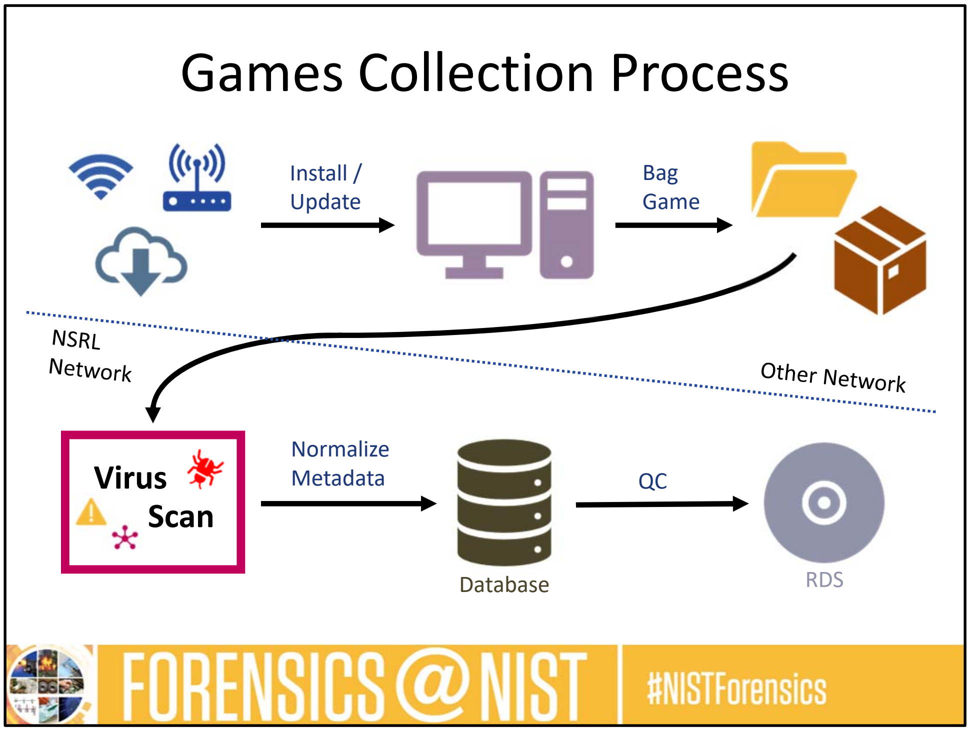
- We collect games from some of the largest gaming market places:
  - Blizzard (Activision Blizzard)
  - Origin (Electronic Arts)
  - Steam (Valve Software)
  - And soon, Epic Games
- We focus on games that are most popular now, and were popular in the past
  - We use publicly available popularity metrics to determine game popularity

FORENSICS@NIST #NISTForensics

It isn't only important for us to collect a large volume of games, but we need to prioritize popular games, as popular games will likely have higher hit rates in investigations.
We collect games from some of the largest gaming marketplaces, Origin, Steam, Blizzard, and soon Epic Games, which will get us Fortnite, which is among the most popular games at the moment if you haven't heard of it.

We would like to note that some of the game platform companies have donated subscription access, and have actively been working with us to make our collection process easier.
So we do have to thank them for working with us and making our lives a little easier.

To determine which games are popular and important for us to download, we use publicly available popularity metrics to get the number of people who were playing a game at a single point in time.
This is useful not only for what is popular now, but for games that were popular in the past that people may still have on their hard drives, even if they don't currently play the game.
For example, a game that may have been popular 5 years ago might not be played a lot today, but it may still be on a large number of computers.
So we don't only want the popular games now, but we also want past popular games that people may still have on their computers.



Games Collection Process

Install / Update

Bag Game

NSRL Network

Other Network

Virus Scan

Normalize Metadata

Database

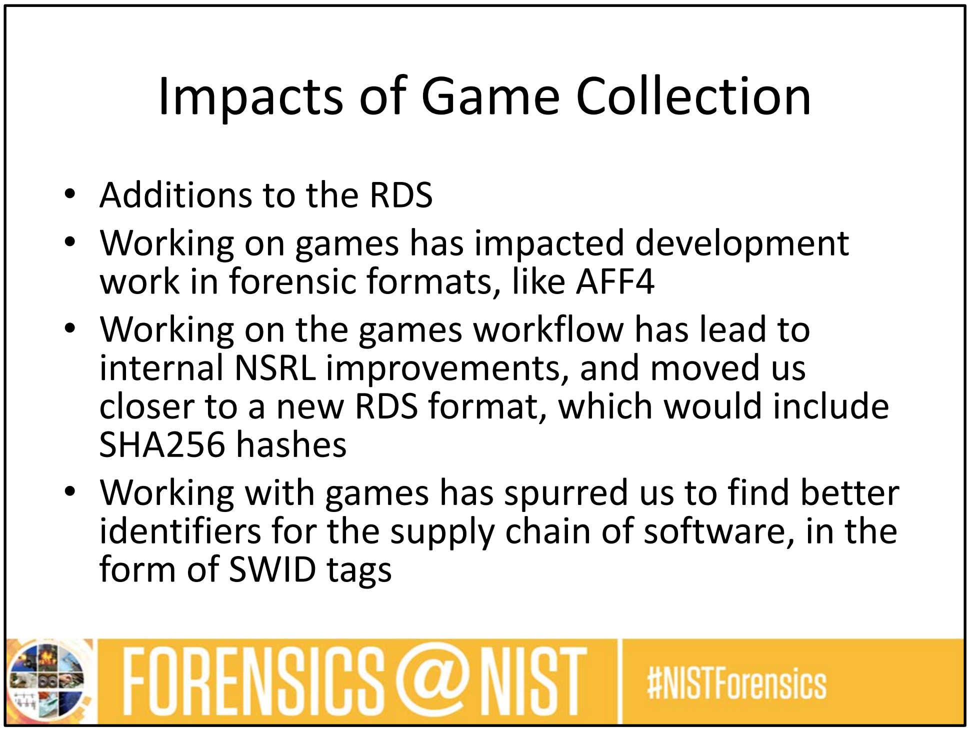
QC

RDS

FORENSICS @ NIST #NISTForensics

Collecting these games is not as simple as just downloading and hashing.
Since we collect games from multiple platforms, we are presented with a few challenges.
First, each platform stores their data differently; so in some cases we can scrape the market place for the meta information we need, other times we have to get that data from the downloads.
Also, each platform uses different popularity metrics.
So we need a tailored ingest process for each market place we collect from.
To make things easier when ingesting the data we collect, we store everything in a Library of Congress bag format, that we ingest into our software repository.

This diagram is our collection process for games.
We first install or update a game to a computer in our lab, using a network external to the NIST network.
We then collect the metadata we need from the game, and the put the game into the bag format.
Next, we copy that game bag to our internal NSRL network, and run a virus scan on it.
We then normalize the metadata, so that it is in a format that we can insert into our database.
And finally, when we are ready to publish an RDS, we do some quality control checks on the data, so that it is consistent with our RDS specifications.

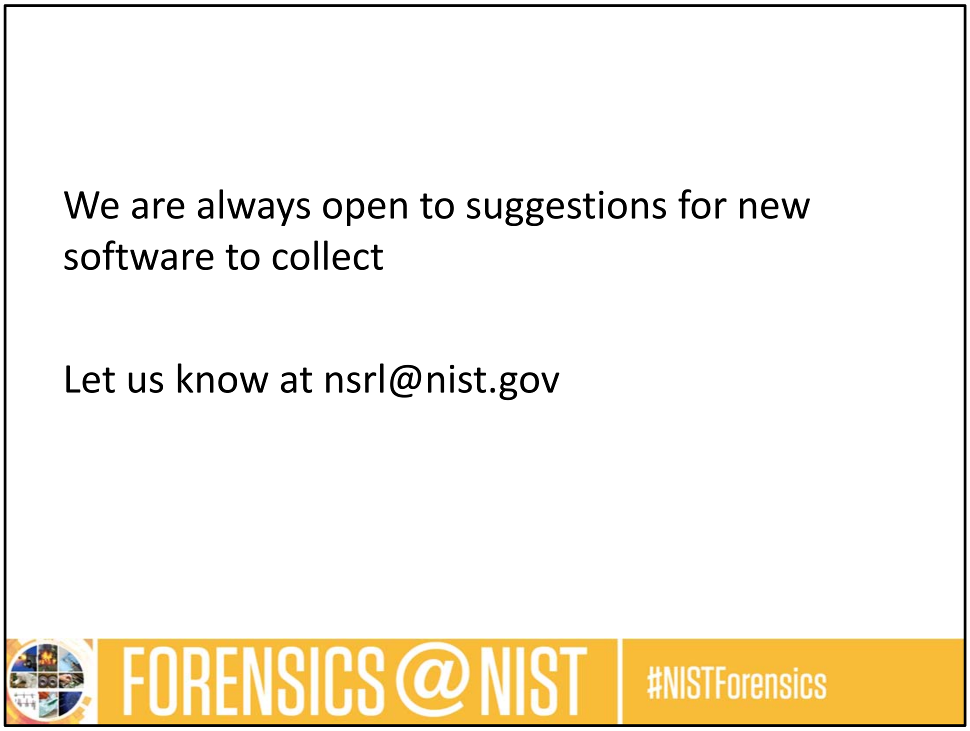


Impacts of Game Collection

- Additions to the RDS
- Working on games has impacted development work in forensic formats, like AFF4
- Working on the games workflow has lead to internal NSRL improvements, and moved us closer to a new RDS format, which would include SHA256 hashes
- Working with games has spurred us to find better identifiers for the supply chain of software, in the form of SWID tags

FORENSICS @ NIST #NISTForensics

Including games does more than just give us more things to add to our data set and RDS.
Working on the game infrastructure has impacted development work in forensic formats, like AFF4.
Working on our games workflow has also lead to improvements to our NSRL infrastructure.
We've updated the way we process the software we collect, and have made improvements to our database, which have gotten us closer to producing a new RDS format that would include SHA256 hashes.
Working with games has also spurred us to find better identifiers for the supply chain of software, in the form of SWID tags.



We are always open to suggestions for new software to collect

Let us know at nsrl@nist.gov

FORENSICS @ NIST | #NISTForensics

So, games are just the newest addition to our collection, but we are always looking forward to expanding as software evolves,
and we are more than happy to take suggestions on what software to collect, not just games but anything.

So if you have suggestions for us on what software to collect, want to donate some software that you have, or just want to learn more about the NSRL project, you can let us know at nsrl@nist.gov and we'd be more than happy to hear from you.

Thank you