# FormatFuzzer: Effective Fuzzing of Binary File Formats

Rafael Dutra
rafael.dutra@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Rahul Gopinath
rahul.gopinath@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Andreas Zeller
zeller@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

## ABSTRACT

Effective fuzzing of programs that process structured binary inputs, such as multimedia files, is a challenging task, since those programs expect a very specific input format. Existing fuzzers, however, are mostly *format-agnostic,* which makes them versatile, but also ineffective when a specific format is required.

We present FormatFuzzer, a generator for *format-specific fuzzers.* FormatFuzzer takes as input a *binary template* (a format specification used by the 010 Editor) and compiles it into C++ code that acts as *parser, mutator,* and highly efficient *generator* of inputs conforming to the rules of the language.

The resulting format-specific fuzzer can be used in multiple ways: (1) It can be used as a standalone producer or mutator in black-box settings, where no guidance from the program is available. (2) By providing mutable *decision seeds,* it can be easily *integrated* with arbitrary *format-agnostic fuzzers* such as AFL to make them format-aware. (3) It can be integrated with *parser-enabled* fuzzers to extend their capabilities using one the several hundred binary templates available. In our evaluation on complex formats such as MP4 or ZIP, FormatFuzzer showed to be a highly effective producer of valid inputs that also detected previously unknown memory errors in `ffmpeg` and `timidity`.

## CCS CONCEPTS

• **Software and its engineering → Software testing and debugging**; *Translator writing systems and compiler generators*; *Parsers*; *Syntax*.

## KEYWORDS

structure-aware fuzzing, file format specifications, binary files, grammars, parser generators, generator-based fuzzing

**ACM Reference Format:**
Rafael Dutra, Rahul Gopinath, and Andreas Zeller. 2021. FormatFuzzer: Effective Fuzzing of Binary File Formats. In *Proceedings of The 28th ACM Conference on Computer and Communications Security (CCS 2021).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Feedback-directed fuzzing tools such as AFL [47] and libFuzzer [34] have shown great success in finding bugs and vulnerabilities in widely used software. Such tools are generally *format-agnostic*, meaning that they assume no knowledge about the input format when producing or mutating inputs. This makes them easy to set up and deploy. However, it also limits their usage to settings in which (1) Sample input files to be mutated are available; (2) Instrumentation and feedback from the program under test is available; (3) The cost of producing myriads of invalid inputs is bearable.

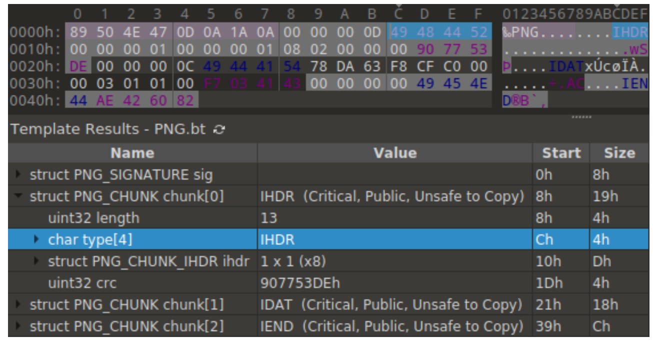


**Figure 1: 010 Editor displaying a PNG file.**

The alternative to a format-agnostic fuzzer is to create a *format-specific* fuzzer, making use of format knowledge to produce valid inputs. However, creating such a fuzzer comes with considerable effort, in particular when one wants to reuse the guidance and analysis capabilities of feedback-directed fuzzers.

In this paper, we present a novel approach that combines the flexibility of format-agnostic fuzzers with the efficiency of format-specific fuzzers. Our FormatFuzzer[1] framework leverages existing *binary templates*—specifications for input formats, from JPEG to PCAP—to (1) produce or mutate valid file inputs, even in black--box settings where no guidance from the program is available; (2) integrate with arbitrary *format-agnostic fuzzers* such as AFL to make them format aware; and (3) to integrate with *parser-enhanced* fuzzers such as AFLSmart to extend their capabilities to the several hundred available binary formats available. In all settings, using FormatFuzzer *increases coverage in the program under test,* covering code not reached by a format-agnostic fuzzer, and thus increasing the chance of detecting bugs and vulnerabilities.

How does FormatFuzzer work? The inputs for FormatFuzzer are *existing input specifications.* The 010 Editor [37] is a commercial editor which allows users to view and edit binary files with detailed structural information, as shown in Figure 1. To parse its inputs, the editor relies on human written specification files, known as *binary templates* [36] to specify the different sections of a file. Those

[1]Actual tool name omitted here for anonymization.


*CCS 2021, 14 - 19 November, 2021, Seoul, South Korea*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

```
typedef struct {
    byte btRed;
    byte btGreen;
    byte btBlue;
} PNG_PALETTE_PIXEL;

struct PNG_CHUNK_PLTE (int32 chunkLen) {
    PNG_PALETTE_PIXEL plteChunkData[chunkLen/3];
};
```

**Listing 1: PNG binary template (extract).**

binary templates are written in a C-like language and have been community-developed for over ten years, resulting in a repositoty of more than 160 different specifications for popular binary file formats [35].

As an example of a binary template, consider Listing 1, showing an excerpt of the original template for PNG files. We see that (like a C program), it defines *types* such as `PNG_PALETTE_PIXEL` consisting of three byte values. The following `struct` fragment defines a data chunk of such pixels named `PNG_CHUNK_PLTE`, whose overall (parameterized) length is `chunkLen`. The full PNG format contains dozens of such chunk definitions, all formalized within the PNG binary template; these also include executable code that computes context-sensitive information such as checksums (whose inference is a major roadblock for format-agnostic fuzzers).

FormatFuzzer takes as input one of these binary templates—say, a PNG template. It then *compiles* it into:

(1) A high-efficiency *generator* that produces outputs in the format specified (i.e., a PNG fuzzer); and
(2) A *parser* that reads in existing inputs in the format specified (i.e., a PNG parser).

The *parser* can immediately be used in existing *parser-enhanced* fuzzers such as AFLSmart [31], effectively giving us a format-specific fuzzer that evolves, say, given PNG seed inputs; this can be repeated with hundreds of different templates, producing hundreds of format-specific, coverage-guided fuzzers. The *generator,* on the other hand, can be used as a *standalone* generator of inputs that all conform to the format specifications. This is useful in black-box settings, where feedback-driven fuzzers cannot be applied. Furthermore, the combination of *parser* and *generator* allows FormatFuzzer to *mutate* existing inputs in a way that ensures their validity. Finally, FormatFuzzer allows to *integrate format-agnostic fuzzers and make them format-aware,* either by having them apply smart (format-aware) mutations on the inputs, or by having them mutate *decision seeds*—strings of bytes that represent decisions taken during parsing and generating. Both integrations bring together the best of format-agnostic guidance and format-aware mutations, and increase the reach of format-agnostic fuzzers.

If this sounds too good to be true, that is because it is. The hundreds of existing binary templates work well for *parsing;* but to make effective *generators,* one has to further refine them. These refinements, however, are easy to write and the feature-full language of binary templates is rich enough to express virtually all needed features of binary file formats, including length fields, checksums, complex structures and constraints between variables. Extending

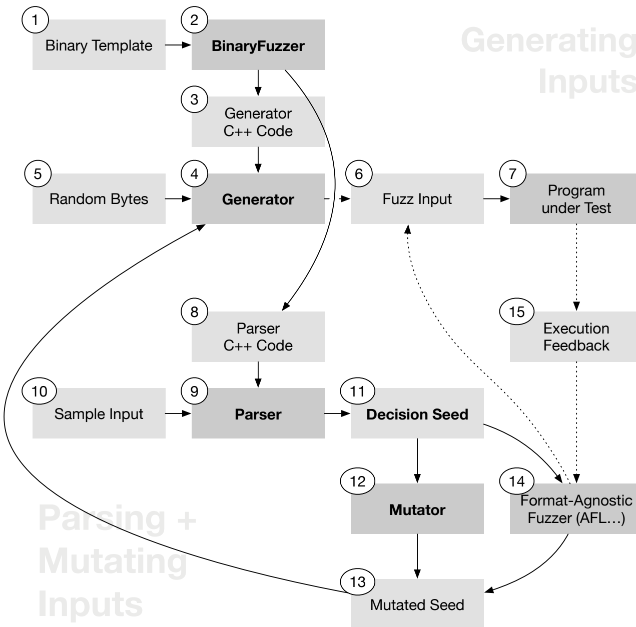


**Figure 2: FormatFuzzer overview. From a *binary template* (1) that specifies an input format *F*, FormatFuzzer (2) generates C++ code (3) that compiles into a highly efficient *generator* (4) for *F*. In a *black-box testing* setting, the generator uses a source of random bytes (5) to generate fuzz input (6) in the format *F* for the program under test (7). From the same binary template, FormatFuzzer also generates C++ code (8) for a *parser* (9) for *F*. In a *mutation* setting, this parser transforms a sample input (10) into a *decision seed* (11), which encodes the decisions made while parsing the sample into a string of bytes. FormatFuzzer can *mutate* this seed (12) (for example, replacing one chunk); feeding the mutated seed (13) into the generator yields a mutated fuzz input still conforming to *F*. Finally, for an *evolutionary* setting, the mutator can also be a format-agnostic fuzzer (14) such as AFL, which then mutates and evolves seed strings based on execution feedback (15). Such a fuzzer can also apply out-of-format mutations to the fuzz input directly.**

a binary template and compiling it is a one-time effort for each format; and once one has, say, a full-fledged PNG specification, one gets a PNG fuzzer, a PNG parser, a PNG mutator, and a variant of your favorite format-agnostic, feedback-directed fuzzer that produces thousands of valid PNG files per second—all for free.

In the remainder of the paper, we detail the techniques behind FormatFuzzer and how they improve the state of the art in fuzzing. Specifially, we make the following contributions:

(1) **A novel description format for generating binary inputs.** There are hundreds of well-curated *binary templates* available that describe as many binary file formats. In Section 2, we describe their structure; Section 3 shows how to extend them with features for *generating* valid inputs.

(2) **Decision seeds to synchronize generators and parsers.** Our FormatFuzzer takes binary templates to produce C++ code (Figure 2) that compiles into specialized and highly efficient generators and parsers, including several context-sensitive features (Section 4). Generator and parser synchronize over *decision seeds*—a sequence of bytes reflecting decisions taken during generating or parsing (Section 5). After parsing an input, the resulting decision seed allows the generator to re-take the same decisions (but now for generating), which reproduces the input exactly.
(3) **Making format-agnostic fuzzers input-format aware.** Mutating bits and bytes in the decision seed allows to explore the entire domain of inputs. If a format-agnostic fuzzer is thus set to mutate and evolve decision seeds, FormatFuzzer can then translate any such mutation into a valid binary input. This way, our approach can be integrated with any format-agnostic fuzzer, leveraging its mutation and evolution strategies (Section 6). The format-agnostic fuzzer can also apply the format-specific *smart mutations* provided by FormatFuzzer, again achieving format awareness.
(4) **A detailed evaluation of all FormatFuzzer elements.** We evaluated FormatFuzzer on several different programs and file formats (Section 7) and found that:
   (a) FormatFuzzer very quickly produces valid inputs, as standalone generator as well as mutator.
   (b) FormatFuzzer integrates well with format-agnostic fuzzers such as AFL, reaching lines otherwise not covered. In our initial fuzzing experiments, FormatFuzzer detected a number of previously unknown (and potentially exploitable) segmentation faults and aborts in `ffmpeg` and `timidity`.
   (c) FormatFuzzer improves the efficiency of parser-enabled fuzzers even with unchanged stock binary templates.

FormatFuzzer is available as open source and enjoys a quickly growing user community. For details, see the FormatFuzzer project page at https://uds-se.github.io/FormatFuzzer/.

## 2 THE BINARY TEMPLATE LANGUAGE

The *binary template language* [38] was designed to fully specify how any binary format could be parsed. It is a C-like language with all the usual control-flow constructs (`if`, `else`, `while`, `switch`, etc.), as well as the ability to call functions and perform arbitrary computations. The language also allows early termination with a `return` statement in the case of irrecoverable parsing errors, as can be seen in Listing 4.

The main peculiarity is in regards to variable definitions. When a variable is defined, such as `uint32 length` in Line 2 of Listing 2, the appropriate number of bytes (in this case, 4) is read from the file to initialize the value of the variable. The current file position is also advanced by the same number of bytes. The exception is when a variable is defined with the keyword `local`, such as the variable `start` in Line 3. Such local variables do not read from the file or advance the file position, but can be freely used in computations and control-flow decisions by the binary template.

The template language allows the definitions of structs to represent the hierarchical structure of files. Listing 2 defines a `struct PNG_CHUNK` which can contain some internal fields, such as `length`, `type` and `crc`, as well as other structs. For example, if the `type` variable is `"IHDR"`, the chunk will contain a `struct PNG_CHUNK_IHDR`, defined in Listing 3. Besides native types (such as `uint32` and `string`) and `struct` types, variables can also have `enum` types, where the list of possible values is explicitly provided, such as the type `PNG_COLOR_SPACE_TYPE` defined in Listing 3.

Another important feature provided by the binary template language is *lookahead functions.* These functions allow the parser to *peek* at a few later bytes in the file without advancing the file position, and can be used for control-flow decisions . For instance, peeking at the first bytes of a `struct` ahead of time can help determine which `struct` type should be parsed. Examples of lookahead functions are `ReadByte()` and `ReadBytes()`. Examples of both are provided in Listings 3 and 6.

The language provides additional functions allowing greater control during parsing, making the language expressive enough to support complex constructs from hundreds of binary formats. Some examples are `FEof()`, which checks for the end of file, `FTell()`, which returns the current file position, `FSeek()`, which allows jumping to a different file position for non-sequential parsing, and `FileSize()`, which returns the file size. Advanced lookahead functions such as `FindFirst()` and `FindAll()` allow searching the file for specific tokens.

We would like to note that our choice for binary templates as the base for FormatFuzzer is purely pragmatic: Binary templates do exist for hundreds of formats; they are reasonably well documented; and as we will show in this paper, they are not too hard to adapt for effective fuzzing. Whether all the subtleties of input formats can be specified in a more elegant or useful way is yet to be proven.

## 3 BINARY TEMPLATES FOR GENERATION

In order to use binary templates to *generate* new files from scratch, some decisions need to be made during the generation process. Most notably, a value will have to be chosen for every variable that is declared, and also for every lookahead function that is called. If such values were to be chosen uniformly at random from the entire range of possible values (for example, $2^{32}$ possible values for a `uint32`), most variables would end up having nonsensical values, rendering the file invalid. Therefore, we extend the language of binary templates in a number of aspects:

**Choices of valid values.** Our first extension allows the specification of an *array of valid values to choose from.* When a variable is declared, this array of choices can is assigned as an *initialization list.* Line 12 of Listing 3 tells the generator to pick a value for `bits` uniformly from five possible choices.

**Enumerations.** To handle enumerations, the generator chooses a value uniformly from the set of the enumerated values (e.g., `PNG_COLOR_SPACE_TYPE` in Listing 3), which is usually the appropriate behavior. But it is also possible to specify a different set of choices through an initialization list.

**Lookaheads.** For lookahead functions, we extend the language to allow a new argument specifying the possible values to choose from. One example is the array `colors` passed as an argument to the `ReadByte()` call in Line 10 of Listing 3. Additional extensions will be detailed in Section 3.

```
typedef struct {
    uint32 length <min=1, max=16>;
    local int64 start = FTell();
    char   type[4];
    if (type == "IHDR")
        PNG_CHUNK_IHDR  ihdr;
    else if (type == "PLTE")
        PNG_CHUNK_PLTE  plte(length);
    /* ... */
    else if (length > 0 && type != "IEND")
        ubyte   data[length];
    local int64 end = FTell();
    local uint32 correct = end - start - 4;
    if (length != correct) { /* Fix it */
        FSeek(start - 4);
        local int evil = SetEvilBit(false);
        uint32 length = { correct };
        SetEvilBit(evil);
        FSeek(end);
    }
    local uint32 crc_calc = Checksum(CHECKSUM_CRC32, start,
         end-start);
    uint32 crc = { crc_calc };
    if (crc != crc_calc)
        Warning("Bad CRC %08x; expected: %08x", crc, crc_calc
             );
} PNG_CHUNK;
```

**Listing 2: PNG binary template (extract). Changes made to enable generation are highlighted in blue.**

```
typedef enum <byte> {
    GrayScale=0, TrueColor=2, Indexed=3,
    AlphaGrayScale=4, AlphaTrueColor=6
} PNG_COLOR_SPACE_TYPE;

typedef struct {
    uint32  width <min=1, max=24>;
    uint32  height <min=1, max=24>;
    local byte colors[] = { GrayScale, TrueColor, Indexed,
      /* ... */ };
    switch (ReadByte(FTell() + 1, colors)) { /* color_type
      */
    case GrayScale:
        ubyte   bits = { 1, 2, 4, 8, 16 };
        break;
    case TrueColor:
        ubyte   bits = { 8, 16 };
        break;
    case Indexed:
        ubyte   bits = { 1, 2, 4, 8 };
        break;
    /* ... */
    }
    PNG_COLOR_SPACE_TYPE color_type;
    PNG_COMPR_METHOD   compr_method;
    PNG_FILTER_METHOD  filter_method;
    PNG_INTERLACE_METHOD interlace_method;
} PNG_CHUNK_IHDR;
```

**Listing 3: Contents of the IHDR chunk, where `bits` depends on `color_type`.**

**Allowing invalid choices.** Finally, we also allow the generator to deviate from the specified behavior and pick any value from the entire range of $2^8$ possible values for `ubyte` with a small probability of about 1%, by performing an action that we call an "evil" decision. Such evil decisions can be turned on and off during generation by calling a function `SetEvilBit()`. When enabled, they can give the fuzzer greater diversity of generated files by allowing the fuzzer to explore a small number of invalid choices during generation.

Let us now put these extensions into action and show what changes had to be made to the binary template in order to support the generation of valid inputs. Our running example, PNG, is representative of the changes that are required for other file formats as well. All our changes are highlighted in blue in the code listings.

## 3.1 Magic Values

The simplest common feature of binary formats is the use of *magic values*, where certain bytes in the file need to have a specific value. For example, a PNG file needs to start with a signature containing the bytes `"\x89504E470D0A1A0A"`. In the original PNG template, there was already a check which compares the signature to the expected value and terminates the template if the signature is invalid, as shown in Listing 4. In this case, FormatFuzzer is able to automatically mine such comparisons (with operators `!=` or `==`) when analyzing the source code of the binary template and remember the magic values that should be used. For example, the value `0x8950` will be remembered as a good value to use for index `0` in the array `btPngSignature`.

```
typedef struct {
    uint16 btPngSignature[4];
} PNG_SIGNATURE;

local int evil = SetEvilBit(false);
PNG_SIGNATURE sig;
SetEvilBit(evil);
if (sig.btPngSignature[0] != 0x8950 ||
        sig.btPngSignature[1] != 0x4E47 ||
        sig.btPngSignature[2] != 0x0D0A ||
        sig.btPngSignature[3] != 0x1A0A) {
    Warning("File is not a PNG image.");
    return -1;
}
```

**Listing 4: PNG signature: the magic bytes are mined automatically.**

The only change made to the binary template for the signature generation was adding two calls to `SetEvilBit()`, which disable evil decisions temporarily for the signature generation, re-enabling them afterwards. This is done because any file with an incorrect PNG signature is not a valid PNG file and would be immediately rejected by the target program.

## 3.2 Size Fields

Another common feature in binary formats is the presence of size fields, whose value correspond to the size of a certain structure in the file, or some particular position or index inside the file. Such size fields are context-sensitive features, and binary formats using such features are not expressible using context-free grammars. It is particularly crucial that size fields are generated correctly, because a wrong value for a size field would completely change the interpretation of the remaining bytes of the file.

In the PNG format, the chunk `length` is the first field in every chunk, as seen in Line 2 of Listing 2. There, we have extended the definition of the `length` variable with additional metadata `min=1` and `max=16` to restrict the choices of values for this variable to a small range. This is important since this variable will be used as the length of an array, as seen in Line 11 of Listing 2. If no such bounds are specified, FormatFuzzer still samples values for integer variables from a skewed distribution that prioritizes small positive integers, but still allows arbitrary large or negative integers with a smaller probability.

```
1 local uint32 seq_num = 0;
2 struct PNG_CHUNK_FDAT {
3     uint32 sequence_number = { seq_num++ };
4     ubyte frame_data[length-4];
5 };
```

**Listing 5: Sequence numbers require a global counter.**

One challenge in generating PNG chunks which is also common in other file formats is that the `length` field appears before the data it refers to. And it is not always possible to predict beforehand what the length of certain data should be before generating that data. In addition, a bad choice for `length` could lead to the length of some array becoming too large or negative, such as when an integer underflow happens in Line 4 of Listing 5. The strategy we use to tackle those issues is to use the initial value sampled for `length` only as a hint. So if, for example, this integer underflow happens when defining the array `frame_data`, we ignore the value of `length` and choose some small positive integer to use as the length of the array. This behavior can be enabled in the binary template with a call to the function `ChangeArrayLength()`.

Now, since the chosen value of `length` is used only as a hint, the generated data may end up being larger or smaller than expected. This needs to be corrected. Otherwise the chunk will have a different length than what is specified in the `length` variable, leading to all following chunks being interpreted incorrectly. Therefore, after generating the data inside a chunk, we check if the length is correct. In case of a discrepancy, we go back and overwrite the value of `length` with the correct value, as shown in Lines 13 to 20 of Listing 2. There, the `FSeek()` function is used to jump to a different position in the file. Note that we temporarily disable evil decisions when fixing the value of `length`, because any wrong value would completely change the file parse tree.

## 3.3 Complex Constraints

Next we discuss how to satisfy some more complex constraints between variables in the template.

**Relationship between variables.** One common pattern happens when the value of one variable `x` restricts the possible values for another variable `y`. If `x` is generated first, then we can simply use the value of `x` in control-flow decisions or initialization lists that will define how `y` is generated. For example, a PNG chunk of type `"IEND"` must have length 0. This can be ensured by using the value of `type` in a control-flow decision in Line 10 of Listing 2 to ensure no `data` will be generated if `type` is `"IEND"`.

In case `x` will only be generated after `y`, we can still look first at the value of `x` using a lookahead function. Listing 3 shows the implementation of the IHDR chunk, where the possible values for the `bits` variable are restricted by the value of the next variable `color_type`. By using the lookahead function `ReadByte()` in Line 10, we can first look at the value at position `FTell()+1`, one byte after the current position. This will be the value of `color_type`, and can be used in a `switch` to determine the possible values for `bits`.

In generation mode, `ReadByte()` will use the array `colors` provided as an argument to choose which value should be generated at position `FTell()+1` in the file. Once a value is chosen by the lookahead function, it is fixed and cannot be overwritten. Therefore, when `color_type` is declared later in Line 22, it is guaranteed to have the same value that was previously chosen by `ReadByte()`.

**Checksums.** Checksums are a common context-sensitive feature of binary formats. Luckily, the binary template language already provides a helper function `Checksum()` which implements common checksum algorithms. Therefore, checksums can be handled simply by computing the correct checksum value and specifying it as the desired value for the checksum variable, as shown in Line 22 of Listing 2.

**Global state.** Some features, such as sequence numbers which are incremented every time, require some global state to be generated correctly. This can be handled by defining a `local` variable of global scope, as in Listing 5.

## 3.4 Chunk Ordering

A big challenge in generating valid files is choosing a valid sequence of chunk types. The set of possible types for the next chunk is usually context-sensitive, depending on which chunks have already been generated. For instance, the chunks IHDR, IDAT and IEND are mandatory. The first chunk must be IHDR and the last chunk must be IEND. Some optional chunks, such as tIME, can appear before or after IDAT. Other optional chunks, such as bKGD, can only appear before IDAT.

If we are only interested in parsing (not generating) PNG files, the following implementation suffices to parse a sequence of chunks.

```
while (!FEof())
    PNG_CHUNK chunk;
```

However, to generate a valid sequence of PNG chunks, we need a way to specify the set of possible choices and allow those choices to change for each new generated chunk.

To solve this problem, we have extended the lookahead function `ReadBytes()` with 3 new arguments: an array of preferred

```
1  ChangeArrayLength();
2  local char chunk_type[4];
3  local string preferred[] = { "IHDR" };
4  local string possible[] = { "IHDR" };
5  while (ReadBytes(chunk_type, FTell() + 4, 4, preferred,
       possible, 0.25)) {
6      PNG_CHUNK chunk;
7      switch (chunk_type) {
8      case "IHDR":
9          switch (chunk.ihdr.color_type) {
10         case GrayScale:
11             local string preferred[] = { "IDAT" };
12             local string possible[] = { "tIME", "tEXt", "pHYs
       ", /*...*/ "bKGD", "IDAT" };
13             break;
14         /* ... */
15         }
16         break;
17     case "IDAT":
18         local string preferred[] = { "IEND" };
19         possible -= ("IDAT", "pHYs", /*...*/ "bKGD");
20         possible += "IEND";
21         break;
22     /* ... */
23     case "IEND":
24         local string preferred[0];
25         local string possible[0];
26         break;
27     }
28 }
```

**Listing 6: Use of `ReadBytes()` to define chunk ordering.**

values, an array of possible values and a probability to pick from the preferred values. We have also added new operators `-=` and `+=` which can add or remove values from such arrays. Listing 6 shows a binary template that can generate a sequence of PNG chunks satisfying the ordering constraints. Here, function `ReadBytes()` in Line 5 needs to choose the 4-byte value which will be written to position `FTell()+4` in the file, which is the location of the `type` array on the next chunk. The chosen value is also returned in the `chunk_type` variable, which is passed by reference to the function.

The `preferred` array is used to specify what is the next mandatory chunk, while the `possible` array includes all chunks which could be inserted in the next position. With a probability of 0.25, `ReadBytes()` will try to pick a value from `preferred`, to ensure that it makes progress towards finishing the generation and does not spend too much time creating lots of optional chunks. With the complementary probability of $1 - 0.25$, `ReadBytes()` will pick a value from the `possible` array (or even a completely random value if an evil decision is made). If the `preferred` array is empty, `ReadBytes()` can choose a value from `possible`, but it is also allowed to not choose any value, not write any value to the file and return `false`, leaving the `while` loop. This signals that the file is now complete and no more chunks will be added.

With no evil decisions, a valid chunk order is guaranteed. The first chunk must be IHDR as the only option specified in Lines 3 and 4. If an IHDR is generated with `color_type` equal to `GrayScale`, the next preferred chunk will be IDAT (Line 11), with several possible chunks available (Line 12), including tIME and bKGD. After IDAT is generated, IEND will be the next preferred chunk (Line 18) and the `possible` array is updated with operators `-=` and `+=` in Lines 19 and 20 to remove and add new values. For instance, bKGD is removed because a bKGD chunk is not allowed after IDAT. Finally, when an IEND is generated, both arrays become empty (Lines 24 and 25) to signal that `ReadBytes()` must return `false`.

### 3.5 Compressed and Encoded Data

With all the changes shown so far, our modified binary template is able to generate almost all chunks of a PNG file correctly. Only the IDAT chunk is still a challenge, because it contains a zlib-compressed datastream of the image pixel data. In order to handle compressed (or otherwise encoded) data, the compression and decompression functions should be specified such that the generator can generate the uncompressed data first and then apply the compression function to it to obtain the datastream that will be written to the file. We have manually implemented this strategy for the PNG IDAT chunk by calling the appropriate functions from the zlib library, allowing FormatFuzzer to generate fully valid PNGs. PNG was the only format so far to require this special handling for compression, because interestingly we found that in other formats simply generating random bytes for the compressed datastreams was sufficient to obtain streams that could be successfully decompressed with high probability.

## 4 IMPLEMENTATION

In the domain of language-specific fuzzers, FormatFuzzer is a *generator compiler,* bringing together the versatility of language specifications and the efficiency of compiled generator code. FormatFuzzer *compiles* a binary template to C++ code for generating and parsing inputs in the format specified by the template, leveraging existing libraries for parsing [11] and processing [10] binary templates. The generated C++ code for a given format can be compiled into a standalone executable or a shared library that can be loaded by other fuzzers, such as AFL++ [15]. Each `struct` definition in the binary template is implemented as a C++ class. We also create C++ classes to handle native types, such as `uint32` and arrays. Every time a variable is declared in the template, this triggers a call to the `generate()` method of the corresponding class.

Our implementation allows setting a maximum size in bytes for the generated file. This is important for fuzzing because smaller files are both faster to produce and to process with the target program. In our experiments, we have set the maximum file size to 64 kB, since we found that even with only a few kilobytes the files can already exhibit the whole variety of chunk types and structures inside them. During generation, any attempt to generate too much data, such as defining an array with a huge length, will throw an exception, aborting the generation. Generation is also aborted if the generator attempts to consume more bytes than available in the decision seed or performs some invalid operation, such as accessing a non-existing field from a `struct` (this can happen for some formats, since control-flow decisions can determine which fields are generated inside a `struct`).

We have also integrated into FormatFuzzer quality assurance tests to verify the correct behavior of our generators and parsers. We use round-trip tests to make sure that whenever a file is successfully generated, then it can also be correctly parsed, and vice versa. We also ensure that our fuzzers can correctly parse a corpus of real files from each format, which were manually collected from GitHub.

# 5 DECISION SEEDS

The methodology followed by FormatFuzzer is the following: To generate a file, we first provide an array of random bytes, called the *decision seed*. This array is the source of randomness for generator decisions. Such decision seeds can also be mined by FormatFuzzer while *parsing*.

## 5.1 Using Decision Seeds for Generation

Every time the generator needs to make a random decision, such as which value should be used for a variable or lookahead function, it will read the required number of bytes from the decision seed.

For example, the `ReadByte()` call in Line 10 of Listing 3 would first read the next byte $b_0$ from the seed to decide whether to take an evil decision. If $b_0 \bmod 128 = 127$, which happens with probability $1/128$, it will perform an evil decision. In that case, the generator reads the next byte $b_1$ from the seed and will use $b_1$ as the value chosen by `ReadByte()`, which covers the whole range of possible values for the type `byte`. In the more common scenario when no evil decision is made, the generator still reads the next byte $b_1$ from the seed, but this time computes $b_1 \bmod 5$ and uses the result as an index to choose from one of the 5 values in the `colors` array, which are the valid values for a `PNG_COLOR_SPACE_TYPE`.

Given a decision seed, the generation process is completely deterministic. The decision seed encodes the essential features of a given file in a representation which is tailored for fuzzing by exhaustively exploring the file structure. The correspondence between the decision seeds and the generated files can be thought of as a partial function $f : \mathit{Seeds} \rightharpoonup \mathit{Files}$. For some values of the seed, the generation can fail. For example, say we have taken an evil decision such as an invalid `color_type` in the PNG IHDR chunk. This makes it impossible to later generate the data for another chunk. But if the generation succeeds, it associates the seed $s$ to a unique well-defined file $f(s)$. The resulting file $f(s)$ is not guaranteed to conform to the format specification (for instance, due to evil decisions), but it should be valid with high probability. This mapping $f$ is not injective, but ideally every valid file should be obtainable by the generator, i.e. $\mathit{Valid} \subseteq f(\mathit{Seeds})$.

## 5.2 Creating Decision Seeds During Parsing

When FormatFuzzer runs in parsing mode, we obtain not only the parse tree for the input file, but also a *decision seed* that can be used to generate the same input file, as shown in Figure 2. One important axiom of our framework is that a file can be successfully parsed by FormatFuzzer if and only if the file can be generated from some appropriate decision seed. And the generator and parser must always agree on the same parse tree for the file. That is why we cannot allow evil decisions when fixing the `length` field in Line 17 of Listing 2. If the generator were to overwrite `length` with an evil value, the generator and the parser would disagree about the starting position of the next chunk.

# 6 FUZZING STRATEGIES

The FormatFuzzer generators and parsers can be used for numerous fuzzing strategies. We discuss a few in this section.

## 6.1 Generating Random Files

The simplest approach to fuzzing with FormatFuzzer is to use completely random seeds (say, from `/dev/urandom`) to generate files in a black-box manner. A good fraction of the generated files will be valid according to the format specification and the files will come from a *semantically diverse distribution* in terms of covered features. For example, as discussed in Listing 3, all possible values for `color_type` will be chosen with equal probability, as will the possible values of `bits` for a given `color_type`.

## 6.2 Mutating Inputs

Another fuzzing strategy enabled by FormatFuzzer is the use of *smart mutations,* where chunks can be abstracted, replaced, deleted or inserted into a file, as seen in Figure 3. Here, we use 'chunk' generically to refer to any `struct` or variable defined in the binary template. We define novel smart mutation operations which work over decision seeds, thus allowing contextual information to be taken into account when producing the mutated file.

*6.2.1 Smart Abstraction.* The first operator we discuss is *smart abstraction*, which allows abstracting one specific chunk $c_1$ into a new random version $c_2$. This operation works by first parsing the original file and obtaining its decision seed. Then, we generate a new file by repeating all the decisions that were made before the target chunk $c_1$ and then using random bytes from `/dev/urandom` to produce the target chunk. Once we detect that the target chunk has been generated, we go back to using the same decision bytes that were used after chunk $c_1$ in the original file. We have found that smart abstractions were the most useful in reaching new coverage, and they are only possible because of our FormatFuzzer framework, which enables synchronized generation and parsing.

*6.2.2 Smart Replacement.* The next operator *smart replacement*, which replaces a chunk $c_1$ from $\mathit{file}_1$ with a different chunk $c_2$ of the same type from $\mathit{file}_2$. Again, we first parse both files to obtain the decision seeds $\mathit{seed}_1$ and $\mathit{seed}_2$. Then, a new mutated seed is created by replacing the decision bytes which created chunk $c_1$ with the decision bytes that create chunk $c_2$, leaving the remaining parts of the seed unchanged, as depicted in the *smart replace* operation of Figure 3. This mutated seed is then used by the generator to produce the mutated file.

To see why performing these operations on the decision seeds can help to generate valid files, consider the case where $c_1$ is some internal field of a PNG chunk. If $c_1$ is replaced by a new instance $c_2$ of the same type, but a different value and different size, then the generator will still compute the correct checksum and size for the modified chunk, while a simple copying of the contents of $c_2$ into $c_1$ would leave the checksum and size invalid. This is especially important for successful smart mutations of formats such as MP4,

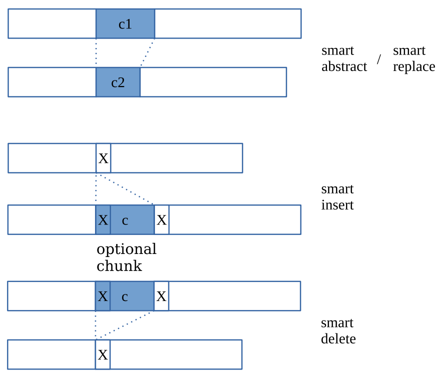


**Figure 3: Smart mutations in FormatFuzzer. Here, X indicates the use of lookahead functions.**

which consist of structs called *boxes* which recursively contain other *boxes*.

*6.2.3 Smart Deletions.* Our *smart delete* operation consists of deleting a chunk from a file, by removing the decision bytes which were responsible for the generation of this chunk. Such a deletion operation only makes sense if we identify that the target chunk *c* is optional. We define a chunk *c* to be *optional* when, right before this chunk is generated, the template calls a lookahead function. This signals that, depending on the result of this lookahead call, we might decide not to generate *c* at all. For example, the variable `chunk` declared in Line 6 of Listing 6 is considered optional, because its declaration comes after a call to `ReadBytes()`. We allow smart deletion to be performed over a chunk *c* only when a lookahead function has been called right before and right after the generation of *c*, as shown in Figure 3. This way, with the new mutated seed, the call to the lookahead function will consume the bytes that were originally consumed at the lookahead call that came after chunk *c*.

*6.2.4 Smart Insertions.* Conversely, we define a *smart insert* as the inverse of the *smart delete* operation, trying to insert a new chunk *c* into a file. Here, the original file must have a call to a lookahead function at the insertion position, so that the result of this call could be used to decide whether chunk *c* should be created. The inserted chunk *c* must also be optional. When performing smart insertions and replacements, FormatFuzzer also checks whether the correct number of decision bytes has been consumed in the generation of the new chunk. In case the mutations do not work well because the new chunk does not fit in the desired position, this can be identified and reported by FormatFuzzer. For example, attempting to copy a bKGD chunk (which specifies a background color) will not work if one file uses color type `TrueColor`, which requires 3 values, while the other uses `GrayScale`, which requires only one.

*6.2.5 Cross-File Operations.* FormatFuzzer implements procedures to parse a list of files and remember all the information about the chunks so they can be used later for smart mutations. There is also a procedure to apply one smart mutation to a chosen file. Here, we randomly choose which kind of mutation will be applied and which chunks will be involved.

## 6.3 Integrating with Format-Agnostic Fuzzers

Besides generation and mutation, FormatFuzzer can also integrate with existing *format-agnostic* fuzzers such as AFL [47] and *parser-enabled* fuzzers such as AFLSmart [31] in different ways. For our experiments, we have integrated FormatFuzzer with AFL++ [15] version 2.60c and with AFLSmart as follows.

**AFL+BFGen** uses AFL to mutate and evolve the *decision seeds*, which will be later fed into the FormatFuzzer generator to produce inputs for the target program. AFL can use the coverage feedback from the program to learn how to effectively mutate the decision seeds. The advantage of working on such seeds is that they are simpler than binary files, since each byte corresponds to a unique decision, and those decisions are made sequentially in the order in which they appear in the seed. Since the FormatFuzzer generator already takes care of the correct input structure, such as computing the correct checksums, inserting the appropriate magic values and setting the correct size fields, AFL can focus exclusively on the high-level decisions that represent the file.

**AFL+BFMut** lets AFL mutate the files that will be given to the target program as usual, but adds new mutation operations which are the smart mutations provided by FormatFuzzer. Every input saved in the AFL queue is parsed by FormatFuzzer so it can be used as a source for smart mutations.

**AFLSmartBF** is a modification of AFLSmart, a *parser-enabled* fuzzer which parses files according to a format specification and uses the learned chunk boundaries to perform smart mutations, such as copying one chunk into a different position. The only change we made in AFLSmartBF is that instead of using Peach [13] format specifications for parsing, we make it use binary templates via the FormatFuzzer parser.

# 7 EVALUATION

Our evaluation focuses on the following research questions:

**RQ1** How much effort is it to set up a binary template file for input generation?
**RQ2** How *efficient* is FormatFuzzer in producing inputs?
**RQ3** How *accurate* is FormatFuzzer in producing inputs?
**RQ4** How efficient is FormatFuzzer as a standalone black-box fuzzer?
**RQ5** How efficient is the integration of FormatFuzzer with a format-agnostic fuzzer?
**RQ6** How does FormatFuzzer compare against other format-aware fuzzers?
**RQ7** Can FormatFuzzer be used with original binary templates?
**RQ8** Does FormatFuzzer find real bugs?

To answer these questions, we ran a series of experiments. All our experiments were conducted on an Intel Xeon CPU E5-4650L machine with 64 cores and 756 GB RAM, running Debian 10. All fuzzers were run for eight hours on a single processor.

## 7.1 RQ1: Effort for Extending Template Files

We start with RQ1: *How much effort is it to set up a binary template file for input generation?* To answer this question, Table 1 lists the number of lines of code required for a sample of format

**Table 1: Number of lines of code required for each format.**

| Format | Original | Modified Template | Generated C++ |
|---|---|---|---|
| GIF | 203 | 234 | 2,144 |
| PCAP | 219 | 270 | 1,388 |
| MIDI | 261 | 282 | 1,771 |
| PNG | 388 | 492 | 2,664 (+107) |
| ZIP | 668 | 772 | 2,608 |
| MP4 | 806 | 1,420 | 4,556 |
| JPG | 1,631 | 1,624 | 5,575 |

specifications. We list the sizes of the original binary template and modified version, which supports generation. For most formats the number of changes is small compared to the amount of code we can already leverage from existing parsing-only binary templates. From our experience, a couple of days is sufficient for updating most formats for generation. However, a few complex formats composed of several different chunk types (e.g. MP4) took more than one week to modify. Note, though, that this effort can hardly be avoided: A random fuzzer will only very occasionally generate a valid MP4 file, let alone systematically cover all chunk types; and a hand-coded MP4 generator will require the same specification effort (and neither result in a parser nor a mutator).

Table 1 also lists the number of lines of format-specific C++ code which were automatically generated by FormatFuzzer. We have found that the use of a domain-specific language of binary templates makes our specifications much more succinct and readable than if they were developed from scratch in a general-purpose language such as C++. Only for the PNG format did we have to edit the C++ code directly to support data compression; in future versions of FormatFuzzer, native support for compression will make such manual changes unnecessary.

**Takeaway 1.** *Extending binary templates is less work than writing a generator from scratch, and provides a parser and mutator on top.*

## 7.2 RQ2: Generator Speed

For RQ2 (*How* efficient *is FormatFuzzer in producing inputs?*), we measure the speed of our generators and parsers and classify the validity of the generated files by feeding them to a target program. The question of how to define the validity of files is also non-trivial. Different programs that process a given format tend to not agree about which inputs they accept, and they often do not follow the official specification. Programs such as media players tend to be forgiving and continue processing the input even after identifying some data corruption. In line with our main goal of reaching deeper code in the application, we classify a file as invalid only when it triggers a critical error which prevents further processing of the file. For example, for images we used the command `identify -verbose` from ImageMagick and checked the output to see if the program could successfully print detailed information about the image, even if some errors were reported.

Table 2 summarizes our results. For each format-aware fuzzer, we have generated 10,000 files, then also parsed those files and tested them with a target program to check for validity (validation command). Here, we see that FormatFuzzer can both generate and parse inputs at a speed of thousands per second for all the tested file formats. This is comparable or sometimes even faster than the speed with which format-agnostic fuzzers can execute inputs. Therefore, our generators and parsers can be easily integrated into existing fuzzers without becoming a performance bottleneck.

**Takeaway 2.** *FormatFuzzer is very fast, generating and parsing thousands of inputs per second.*

## 7.3 RQ3: Input Validity

Table 2 also shows which fraction of the generations produced an output and which fraction of those outputs was classified as valid, addressing RQ3: *How* accurate *is FormatFuzzer in producing inputs?* The results show that our fuzzers succeed almost all the time in producing inputs and those inputs have a high chance of being accepted by the taget program. The validity is even higher if evil decisions are disabled; however evil decisions can be useful to trigger some unexpected behavior. The ZIP format achieved the lowest validity of 27%. Here, we should take into account that for validation we used the command `unzip -t`, which actually extracts all the archive files in memory and validates their checksums. If we were to use the less stringent validation command `unzip -l`, which simply lists the names and metadata of all the files contained in the zip archive (but may still allow some of the files to be corrupted), the validity is actually 59%.

**Takeaway 3.** *The large majority of files generated by FormatFuzzer is valid.*

To put this validity into perspective, let us compare against inputs produced by the format-agnostic AFL fuzzer, listed in the "AFL" column. We see that even with sample files, AFL produces far fewer valid input files.

**Takeaway 4.** *FormatFuzzer produces far more valid inputs than a format-agnostic fuzzer.*

## 7.4 RQ4: Black-Box Fuzzing

Let us now evaluate the *effectiveness* of FormatFuzzer, notably in achieving *coverage* in our test subjects. Generally speaking, the higher the coverage of a test generator, the higher its chance to detect bugs—in particular, since if code is not covered, bugs will not be detected.

We start with RQ4: *How efficient is FormatFuzzer as a standalone black-box fuzzer?* We evaluate two black-box settings:

**BFGen** is using FormatFuzzer as a standalone input generator, requiring no knowledge or feedback from the program under test. (Note that this is a setting in which only a specification-based fuzzer like FormatFuzzer can succeed.)

**BFMut** is using FormatFuzzer to parse existing input files and apply smart mutations, again without guidance. (To the best of our knowledge, this is a capability unique to FormatFuzzer.)

Table 3 lists the programs with the exact command used for fuzzing. In all fuzzing experiments we have used a timeout of eight hours. For all techniques which require an initial corpus of inputs, we have used the same corpus of files, which were small files from each format manually downloaded from github. Each corpus had an average of 12 files (minimum 5, maximum 34).

**Table 2: Average performance of the fuzzers in terms of speed and validity.**

| | Speed (files per $s$) | | | With Evil | | Without Evil | | AFL | |
|---|---|---|---|---|---|---|---|---|---|
| Format | Generated | Parsed | File Size | Success | Valid | Success | Valid | Valid | Validation Command |
| PNG | 6,061 | 2,470 | 648 | 97% | 88% | 100% | 95% | 0% | `identify -verbose - \| grep Elapsed` |
| JPG | 4,157 | 5,283 | 1,328 | 99% | 80% | 100% | 96% | 63% | `identify -verbose - \| grep Elapsed` |
| GIF | 5,553 | 4,045 | 5,260 | 100% | 40% | 100% | 42% | 68% | `identify -verbose - \| grep Elapsed` |
| MIDI | 3,446 | 20,382 | 570 | 96% | 75% | 100% | 76% | 51% | `! timidity - -Ol \| grep ^-:` |
| MP4 | 3,081 | 2,526 | 2,070 | 99% | 82% | 100% | 89% | 36% | `ffmpeg -y -i - -c:v mpeg4 -c:a copy out.mp4` |
| ZIP | 14,533 | 13,507 | 2,029 | 96% | 27% | 98% | 30% | 1% | `yes \| unzip -P "" -t` ⟨input file⟩ |
| PCAP | 4,038 | 5,283 | 2,451 | 96% | 92% | 100% | 100% | 2% | `tcpdump -nr -` |

**Table 3: Programs under test**

| Format | Program | Invocation |
|---|---|---|
| PNG | `libpng 1.6.37-3` | `readpng` |
| JPG | `libjpeg-turbo 1_2.0.6-4` | `djpeg` ⟨input file⟩ |
| GIF | `gif2png 2.5.14` | `gif2png` ⟨input file⟩ |
| MIDI | `TiMidity++ 2.14.0-8` | `timidity - -o - -Ow` |
| MP4 | `FFmpeg 4.4` | `ffmpeg -y -i - out.mp4` |
| ZIP | `UnZip 6.0-26` | `unzip -P "" -t` ⟨input file⟩ |
| PCAP | `tcpdump 4.99.0-2`, `libpcap 1.10.0-2` | `tcpdump -nr -` |

**Table 4: Coverage (%) of FormatFuzzer in black-box settings**

| Setting | PNG | JPG | GIF | MIDI | MP4 | ZIP | PCAP |
|---|---|---|---|---|---|---|---|
| BFGen | 22.3 | 24.2 | 68.5 | 12.4 | 5.4 | 33.7 | 11.5 |
| BFMut | 22.5 | 24.2 | 70.7 | 10.3 | 7.0 | 34.8 | 7.9 |
| BFGen \ BFMut | 0.2 | 0.0 | 0.3 | 2.1 | 0.5 | 0.3 | 4.8 |
| BFMut \ BFGen | 0.4 | 0.1 | 2.5 | 0.1 | 2.1 | 1.4 | 1.2 |

Table 4 lists the resulting line coverage (average over all runs). One interesting aspect is that in both settings, all inputs are generated syntactically valid by construction; and indeed, inspection shows that error-handling code is hardly covered. (Generally speaking, reaching error-handling code is the easy part of fuzzing; if one wants to cover error-handling code, too, one could easily introduce subtle lexical mutations into the files generated by FormatFuzzer).

**Takeaway 5.** *FormatFuzzer achieves decent coverage even in black-box settings, without any guidance by the program under test.*

**Takeaway 6.** *By itself, FormatFuzzer focuses on code that handles* valid *inputs.*

The lower half of Table 4 highlights the *coverage difference* between the two settings; here, $A \setminus B$ denotes lines covered in $A$, but not in $B$. The advantage of BFGen over BFMut is prominent in MIDI and PCAP, where the binary template covers exotic contents not found in our sample files. The additional coverage of BFMut over BfGen is likely due to some features in the corpus of input files which are not described in the binary templates.

**Takeaway 7.** *Some features in the spec are unlikely to be found in files in the wild.*

## 7.5 RQ5: Integrating Format-Agnostic Fuzzers

In practice, we will frequently encounter settings where feedback from the program under test is available, and where FormatFuzzer therefore can integrate with existing format-agnostic fuzzers. Let us thus address RQ5: *How efficient is the integration of FormatFuzzer with a format-agnostic fuzzer?* We have integrated FormatFuzzer with the popular AFL++ [15] fuzzer in two different ways, as presented in Section 6.3:

**AFL+BFGen** is using AFL to mutate the *decision seed* which is then fed into FormatFuzzer and used to generate inputs for the target program. Since only the decision seed is mutated, all generated inputs will be conforming to the given format.

**AFL+BFMut** is having AFL use FormatFuzzer to perform smart mutations on the input file. FormatFuzzer parses the files in the AFL queue and remembers information about the chunks present in those files, which it can later use to perform smart mutations. Note that regular AFL mutations are also used, which can yield invalid inputs and helps covering error handling code.

**Table 5: Coverage (%) of FormatFuzzer integrated with AFL**

| Setting | PNG | JPG | GIF | MIDI | MP4 | ZIP | PCAP |
|---|---|---|---|---|---|---|---|
| AFL | 17.5 | 29.2 | 73.1 | 13.1 | 11.3 | 36.3 | 25.5 |
| AFL+BFGen | 23.8 | 27.9 | 71.9 | 12.8 | 9.7 | 36.6 | 22.4 |
| AFL+BFMut | 26.0 | 32.8 | 73.3 | 12.4 | 10.7 | 37.4 | 23.6 |
| AFL+BFGen \ AFL | 7.5 | 2.3 | 0.3 | 0.3 | 0.7 | 1.2 | 1.2 |
| AFL+BFMut \ AFL | 8.5 | 3.6 | 0.3 | 0.2 | 0.7 | 1.4 | 1.0 |
| AFL \ AFL+BFGen | 1.2 | 3.6 | 1.5 | 0.6 | 2.3 | 0.9 | 4.3 |
| AFL \ AFL+BFmut | 0.0 | 0.0 | 0.1 | 0.9 | 1.3 | 0.3 | 2.9 |
| AFL+BFGen \ BFGen | 1.4 | 3.6 | 3.9 | 0.8 | 4.2 | 3.1 | 10.9 |
| AFL+BFMut \ BFMut | 3.5 | 8.6 | 2.6 | 2.1 | 3.8 | 2.6 | 15.8 |

Table 5 lists the resulting coverage. For all subjects, the set differences show that both strategies increase coverage over plain AFL, showing that FormatFuzzer is effective in making format-agnostic fuzzers format aware. For four out of seven subjects, the integration also achieves a higher coverage in absolute terms.

**Takeaway 8.** *Format-aware fuzzing with FormatFuzzer reaches lines that format-agnostic fuzzing does not.*

However, AFL by itself also covers lines that the integration of AFL and FormatFuzzer did not reach. One reason for this is again AFL producing several *invalid* inputs, covering error-handling code, which AFL+BFGen avoids by construction.

**Takeaway 9.** *Format-aware and format-agnostic fuzzing complement each other.*

Comparing the AFL integration against FormatFuzzer standalone, we also see that integrating AFL's coverage guidance into format-aware fuzzing also improves coverage over a pure black-box setting. Hence, such integration is a preferred setting if sample inputs and coverage feedback are available.

**Takeaway 10.** *Integrating FormatFuzzer with coverage-guided fuzzers improves coverage over black-box settings.*

## 7.6 RQ6: Alternate Format-Aware Strategies

Our next evaluation concerns RQ6: *How does FormatFuzzer compare against other format-aware fuzzers?* The competitor here is AFLSmart [31], which uses format information from *Peach* specifications to determine chunk boundaries in input files, thus also resulting in smarter mutations. AFLSmart uses input specifications to *parse* inputs only, but not to *generate* valid inputs.

We first evaluate AFLSmart on its own (AFLSmart) and then evaluate the variant of AFLSmart in which we exchange the *Peach* specifications against binary templates (AFLSmartBF). This allows us to factor out the influence of individual input specifications.

**Table 6: Coverage (%) of FormatFuzzer and AFLSmart**

| Setting | PNG | JPG | GIF | MIDI | MP4 | ZIP | PCAP |
|---|---|---|---|---|---|---|---|
| AFLSmart | 18.1 | 29.2 | –/–[2] | 12.4 | 11.4 | 36.1 | 24.5 |
| AFLSmartBF | 17.6 | 29.2 | 73.4 | 14.2 | 12.2 | 36.7 | 25.1 |
| AFLSmart \ AFL+BFMut | 0.3 | 0.0 | –/– | 0.2 | 1.2 | 0.2 | 2.8 |
| AFL+BFMut \ AFLSmart | 8.2 | 3.6 | –/– | 0.2 | 0.7 | 1.4 | 1.7 |
| AFLSmart \ AFLSmartBF | 0.5 | 0.0 | –/– | 0.7 | 0.4 | 0.0 | 1.4 |
| AFLSmartBF \ AFLSmart | 0.0 | 0.0 | –/– | 2.5 | 1.4 | 0.7 | 2.0 |

The coverage results are listed in Table 6. Comparing AFLSmart against AFL+BFMut (Table 5), we see that the AFL+FormatFuzzer integration outperforms AFLSmart in five out of seven subjects. As we again see in the differences, either approach covers lines that the other does not.

**Takeaway 11.** *Using input formats for parsing, mutating, and generating, as FormatFuzzer does, yields additional coverage over using them for parsing only, as AFLSmart does.*

Are these differences due to the input format specs used—that is, *Peach* for AFLSmart, and binary templates for FormatFuzzer? The results for AFLSmartBF show no significant differences, except for MIDI, where the binary template covers more details already during parsing. Hence, the differences between AFLSmart and FormatFuzzer are mostly due to the respective strategies, not the format specs used.

**Takeaway 12.** *The combination of different fuzzing strategies is likely to yeald more coverage.*

## 7.7 RQ7: Using Unchanged Templates

The AFLSmartBF setting from Section 7.6 uses only the *parsing* capabilities of FormatFuzzer. Hence, one would be able to use it with *unchanged* templates (without adding generative capabilities) as well. This is what we evaluate to address RQ7: *Can FormatFuzzer be used with original binary templates?*

With the AFLSmartBF$_0$ setting, we repeat the experiment from AFLSmartBF, but using *unchanged* binary templates.

[2]AFLSmart crashed with a segmentation fault.

**Table 7: Coverage (%) with stock binary templates**

| Setting | PNG | JPG | GIF | MIDI | MP4 | ZIP | PCAP |
|---|---|---|---|---|---|---|---|
| AFLSmartBF$_0$ | 18.1 | 29.4 | 73.2 | 13.8 | 11.2 | 37.3 | 25.2 |
| AFLSmartBF \ AFLSmartBF$_0$ | 0.0 | 0.0 | 0.1 | 2.5 | 1.3 | 0.3 | 1.6 |
| AFL+BFMut \ AFLSmartBF$_0$ | 8.2 | 3.5 | 0.1 | 0.2 | 0.6 | 1.2 | 1.3 |

Table 7 shows that the AFLSmartBF$_0$ setting yields the same results as AFLSmartBF, indicating a cost-effective way to unlock format-aware fuzzing with hundreds of binary templates.

**Takeaway 13.** *Integrating FormatFuzzer into AFLSmart allows to use unchanged stock binary templates for very good performance.*

The overall best performing setting, however, is still AFL+BFMut, which uses a *generative* template. The last line of Table 7 shows its additional coverage over AFLSmartBF$_0$.

## 7.8 RQ8: Bugs Found

We'd like to close our evaluation with RQ8: *Does FormatFuzzer find real bugs?* The answer is: Yes! In our initial fuzzing experiments with FormatFuzzer, we have found and reported:

- 13 distinct segmentation faults (by different stack traces) and 8 distinct aborts in `ffmpeg` (MP4), which turned out to be at least 6 distinct bugs which have already been fixed to date by the FFmpeg developers. These occur when `ffmpeg` is compiled with pthreads (which is the default option) and the fuzzer runs with a memory limit of 200 MB (using AFL's `-m` option). We will file appropriate CVEs for those bugs.
- Two new segmentation faults, in addition to a couple of abort errors in `timidity` (MIDI). We will update the paper when those bugs are fixed.

Note that `ffmpeg` "is part of the workflow of hundreds of other software projects, and its libraries are a core part of software media players such as VLC, and has been included in core processing for YouTube and iTunes." [1]

**Takeaway 14.** *FormatFuzzer finds bugs in relevant software.*

## 7.9 Discussion

Leveraging input format specifications pays off. In every single setting, FormatFuzzer explored additional coverage—and thus additional chances to find bugs and vulnerabilities. FormatFuzzer can be used as a standalone generator, notably in black-box settings; the integration with feedback-driven fuzzers such as AFL combines the best of format-aware and feedback-directed fuzzing.

The evaluation also shows the *versatility* of FormatFuzzer. Its individual components (parsing/mutating/generating) can either be used standalone (say, *generating* in a black-box setting), or partially integrated with existing format-aware tools (say, *parsing* using existing binary templates with AFLSmart), or fully integrated with format-agnostic fuzzers (where the popular AFL tool can be replaced by any even more performant format-agnostic fuzzer).

This versatility and modularity makes FormatFuzzer a *platform* for quickly creating and integrating novel fuzzing strategies and adapting them towards the setting at hand—in contrast to *monolithic* fuzzers, where altering or assessing individual design choices is much harder. As our evaluation shows, both format-aware and format-agnostic strategies have their specific strengths, further motivating the need for an integrative platform.

One aspect not addressed in our evaluation is that by construction already, FormatFuzzer gives the tester *control* over what should be tested. By commenting out particular parts of a binary template, FormatFuzzer allows to focus on specific features—say, those that were recently changed or otherwise are critical. This adds to the versatility of fuzzing with FormatFuzzer.

# 8 RELATED WORK

## 8.1 Fuzzing with Input Specifications

Using language specifications for generating inputs is an old idea, especially for context-free grammars. The different grammars such as regular, context-free, context-sensitive, and unconstrained grammars were invented by Chomsky for linguistic applications [7] in 1950s specifically for parsing. Using grammars for input generation was first suggested by Burkhardt [5], Hanford [21] and Purdom [32] in the late 1960s and early 1970s. Logic languages [19] such as Prolog [44] which are known from 1970s allow turning the entire program inside out, using the program itself to generate the kind of inputs that it accepts.

## 8.2 Context-Free Grammars

Grammar-based testing took off in the new millennium when researchers rediscovered the utility of fuzzing, and how grammars can improve the effectiveness of fuzzing. One of the first to recognize their use in fuzzing was Godefroid [17] who augmented whitebox fuzzing with grammars. Other noteworthy grammar fuzzers include Gramfuzz [20], Grammarinator [23], Dharma [27], Domato [16], and CSS Fuzz [33], as well as PolyGlot [6], which augments context-free grammars with semantic annotations. LangFuzz [24] uses a language specification to collect code fragments which can be applied as smart mutations. All these work on context-free languages, which are not sufficient to specify binary formats.

## 8.3 Other Specification Languages

Besides context-free grammars, fuzzers have used alternate input specifications, such as regular languages (i.e. finite state automata) [9, 43], or constraint languages [12]. MoWF [30] leverages a specification expressed as a constraint over the input space. It uses selective symbolic execution to identify uncovered branches, and can repair length fields, checksums and other validation fields. The fuzzer by Pan et al. [29] leverages higher order attribute grammars to describe length, checksums, and other validation fields, and uses them to generate file format inputs such as PNGs. Parsifal [25] targets both parsing and generation of binary formats. It is, however, limited to fixed size formats. Nail [4] defines a parser generator specification, and targets binary formats (typically protocols) that contain offset fields and checksums, and unlike Parsifal, can handle more complex constructs. Underwood [40] extends context-free grammars using attribute grammars, and provides a mapping from binary format specifications to attribute grammars. Such grammars can be used for parsing as well as generation.

In principle, FormatFuzzer could make use of any of these format specifications, and still leverage its unique capabilities such as generating synchronized parsers and generators, or integrating format-agnostic fuzzers. By using binary templates, however, FormatFuzzer can build on hundreds of format specifications that have been created and refined by an enthusiastic community for more than two decades now.

## 8.4 Fuzzing Strategies

CSmith [46] shows that embedding the entire language specification into the generator can lead to effective fuzzing. However, with that, one loses the generality of being able to target other kinds of inputs. Other notable research on grammar-based fuzzers include LangFuzz [24], Nautilus [3], Blendfuzz [45], Skyfire [41], and Superion [42], which extends AFL. *Parameter sequences* in Zest [28] encode generator choices like our *decision seeds* and can be used for generator-based testing in the style of QuickCheck [8]. However, obtaining these by parsing existing files, as well as mutating them (Section 6.2) is a novel contribution of FormatFuzzer.

## 8.5 Fuzzing Binary Formats

The WEIZZ [14] fuzzer targets chunk-based binary formats. During fuzzing, it tries to learn how the chunk-based format is specified. Peach [13] is another fuzzer that targets binary formats. AFLSmart [31] is a fuzzer that provides chunk aware mutations to the input, and hence targets binary formats. It does this by maintaining a virtual structure of the input being fuzzed in memory. AFLSmart is the closest related work to FormatFuzzer. However, since its format specifications (*Peach pits*) had to be developed from scratch for use with AFLSmart, they are less complete and comprehensive than our binary templates. Since FormatFuzzer can use its binary templates not only for parsing, but also for generation, this enables additional fuzzing strategies detailed in this paper.

# 9 CONCLUSION

If one seriously wants to fuzz binary inputs, the best way is to create a dedicated fuzzer. Creating such fuzzers takes significant effort, though. With FormatFuzzer, however, one can leverage *hundreds of existing binary templates* to make existing fuzzers format-aware. FormatFuzzer can transform binary templates into *parsers,* which can be integrated into parser-enabled fuzzers. Extending existing binary templates for generation unlocks fuzzing in black-box settings that are infeasible for format-agnostic fuzzers. By having format-agnostic fuzzers work on decision seeds rather than inputs, FormatFuzzer can make any fuzzer format-aware.

While FormatFuzzer is a highly efficient choice for domain-specific fuzzing, its modularity also makes it a useful platform for creating future fuzzers. Besides extending FormatFuzzer with further formats, our future work will focus on the following topics:

**Black-box strategies.** Recent advances in grammar-based fuzzing, such as systematically achieving grammar coverage [22] or learning and leveraging probability distributions [39] could easily be adopted for binary template fuzzing, too.

**Search-based fuzzing.** The *decision seed* format opens way for easily integrating alternate fuzzing strategies. Of particular interest is *genetic optimization* as part of *search-based testing,* since FormatFuzzer already implements mutation and crossover operations.

**Mining binary formats.** Recent techniques for mining context-free grammars from parsers [18, 26] could be adapted to binary formats, simplifying template construction.

**Mining binary constraints.** Extending a (mined) syntactic template, we can track processing of individual input elements using dynamic tainting and dynamic analysis, and extract context-sensitive constraints from input processors.

**Engaging the community.** We will be creating tutorials and other material to engage the community in writing binary templates. The Wikipedia list of file formats [2] lists more than 1,000 formats—so there is still lots to do!

FormatFuzzer and all supporting material are available as open source. For more information on FormatFuzzer, see

https://uds-se.github.io/FormatFuzzer/

## REFERENCES


[1] [n.d.]. Wikipedia: ffmpeg. https://en.wikipedia.org/wiki/FFmpeg. Accessed May 2021.

[2] [n.d.]. Wikipedia: List of File Formats. https://en.wikipedia.org/wiki/List_of_file_formats. Accessed May 2021.

[3] Cornelius Aschermann, Tommaso Frassetto, Thorsten Holz, Patrick Jauernig, Ahmad-Reza Sadeghi, and Daniel Teuchert. 2019. NAUTILUS: Fishing for Deep Bugs with Grammars. In *Proceedings of NDSS 2019*. https://www.ndss-symposium.org/ndss-paper/nautilus-fishing-for-deep-bugs-with-grammars/

[4] Julian Bangert and Nickolai Zeldovich. 2014. Nail: A practical tool for parsing and generating data formats. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*. 615–628.

[5] W. H. Burkhardt. 1967. Generating test programs from syntax. *Computing* 2, 1 (March 1967), 53–73. https://doi.org/10.1007/BF02235512

[6] Yongheng Chen, Rui Zhong, Hong Hu, Hangfan Zhang, Yupeng Yang, Dinghao Wu, and Wenke Lee. 2021. One Engine to Fuzz 'em All: Generic Language Processor Testing with Semantic Validation (to appear). In *Proceedings of the 42nd IEEE Symposium on Security and Privacy (Oakland)*. San Francisco, CA.

[7] Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2 (1956), 113–124. https://chomsky.info/wp-content/uploads/195609-.pdf

[8] Koen Claessen and John Hughes. 2011. QuickCheck: a lightweight tool for random testing of Haskell programs. *ACM SIGPLAN Notices* 46, 4 (2011), 53–64.

[9] Baojiang Cui, Shurui Liang, Shilei Chen, Bing Zhao, and Xiaobing Liang. 2014. A novel fuzzing method for Zigbee based on finite state machine. *International Journal of Distributed Sensor Networks* 10, 1 (2014), 762891.

[10] James "d0c_s4vage" Johnson. 2020. GitHub - d0c-s4vage/pfp: pfp - Python Format Parser - a python-based 010 Editor template interpreter. https://github.com/d0c-s4vage/pfp. Accessed February 16, 2020.

[11] James "d0c_s4vage" Johnson. 2020. GitHub - d0c-s4vage/py010parser: A modified pycparser to parse 010 templates. https://github.com/d0c-s4vage/py010parser. Accessed February 16, 2020.

[12] Kyle Dewey, Jared Roesch, and Ben Hardekopf. 2014. Language fuzzing using constraint logic programming. ACM, 725–730.

[13] Michael Eddington. 2011. Peach fuzzing platform. *Peach Fuzzer* 34 (2011).

[14] Andrea Fioraldi, Daniele Cono D'Elia, and Emilio Coppa. 2020. WEIZZ: Automatic grey-box fuzzing for structured binary formats. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 1–13.

[15] Andrea Fioraldi, Dominik Maier, Heiko Eißfeldt, and Marc Heuse. 2020. AFL++: Combining Incremental Steps of Fuzzing Research. In *14th USENIX Workshop on Offensive Technologies (WOOT 20)*. USENIX Association.

[16] Ivan Fratric. 2019. *Domato A DOM fuzzer*. https://github.com/googleprojectzero/domato

[17] Patrice Godefroid, Adam Kiezun, and Michael Y. Levin. 2008. Grammar-based Whitebox Fuzzing. ACM, New York, NY, USA, 206–215.

[18] Rahul Gopinath, Björn Mathis, and Andreas Zeller. 2020. Mining input grammars from dynamic control flow. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 172–183.

[19] Claude Cordell Green. 1970. *The application of theorem proving to question-answering systems*. Number 96. Management Information Services.

[20] Tao Guo, Puhan Zhang, Xin Wang, and Qiang Wei. 2013. Gramfuzz: Fuzzing testing of web browsers based on grammar analysis and structural mutation. In *2013 Second International Conference on Informatics & Applications (ICIA)*. IEEE, 212–215.

[21] Kenneth V. Hanford. 1970. Automatic Generation of Test Cases. *IBM Syst. J.* 9, 4 (Dec. 1970), 242–257. https://doi.org/10.1147/sj.94.0242

[22] Nikolas Havrikov and Andreas Zeller. 2019. Systematically Covering Input Structure. In *Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering* (San Diego, California) *(ASE '19)*. IEEE Press, 189–199. https://doi.org/10.1109/ASE.2019.00027

[23] Renáta Hodován, Ákos Kiss, and Tibor Gyimóthy. 2018. Grammarinator: a grammar-based open source fuzzer. In *Proceedings of the 9th ACM SIGSOFT International Workshop on Automating TEST Case Design, Selection, and Evaluation*. ACM, 45–48.

[24] Christian Holler, Kim Herzig, and Andreas Zeller. 2012. Fuzzing with Code Fragments. In *Proceedings of the 21st USENIX Conference on Security Symposium* (Bellevue, WA) *(Security'12)*. USENIX Association, Berkeley, CA, USA, 38–38.

[25] Olivier Levillain. 2014. Parsifal: A pragmatic solution to the binary parsing problems. In *2014 IEEE Security and Privacy Workshops*. IEEE, 191–197.

[26] Björn Mathis, Rahul Gopinath, Michaël Mera, Alexander Kampmann, Matthias Höschele, and Andreas Zeller. 2019. Parser-directed fuzzing. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 548–560.

[27] Mozilla. 2019. *Dharma: A generation-based, context-free grammar fuzzer*. https://blog.mozilla.org/security/2015/06/29/dharma/

[28] Rohan Padhye, Caroline Lemieux, Koushik Sen, Mike Papadakis, and Yves Le Traon. 2019. Semantic fuzzing with zest. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 329–340.

[29] Fan Pan, Ying Hou, Zheng Hong, Lifa Wu, and Haiguang Lai. 2013. Efficient Model-based Fuzz Testing Using Higher-order Attribute Grammars. *JSW* 8, 3 (2013), 645–651.

[30] Van-Thuan Pham, Marcel Böhme, and Abhik Roychoudhury. 2016. Model-based whitebox fuzzing for program binaries. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*. 543–553.

[31] Van-Thuan Pham, Marcel Böhme, Andrew Edward Santosa, Alexandru Razvan Caciulescu, and Abhik Roychoudhury. 2019. Smart greybox fuzzing. *IEEE Transactions on Software Engineering* (2019).

[32] Paul Purdom. 1972. A sentence generator for testing parsers. *BIT Numerical Mathematics* 12, 3 (1972), 366–375. https://doi.org/10.1007/BF01932308

[33] Jesse Ruderman. 2007. *Introducing jsfunfuzz*. http://www.squarefree.com/2007/08/02/introducing-jsfunfuzz/

[34] Kosta Serebryany. 2016. Continuous fuzzing with libfuzzer and addresssanitizer. In *2016 IEEE Cybersecurity Development (SecDev)*. IEEE, 157–157.

[35] SweetScape Software. [n.d.]. 010 Editor - Binary Template Repository - Download Binary Templates. https://www.sweetscape.com/010editor/repository/templates/. Accessed February 16, 2020.

[36] SweetScape Software. [n.d.]. 010 Editor - Binary Templates - Parsing Binary Files. https://www.sweetscape.com/010editor/templates.html. Accessed February 16, 2020.

[37] SweetScape Software. [n.d.]. 010 Editor - Pro Text/Hex Editor | Edit 160+ Formats | Fast & Powerful. https://www.sweetscape.com/010editor/. Accessed February 16, 2020.

[38] SweetScape Software. [n.d.]. 010 Editor Manual - Writing Templates. https://www.sweetscape.com/010editor/manual/IntroTemplates.htm. Accessed February 16, 2020.

[39] Ezekiel Soremekun, Esteban Pavese, Nikolas Havrikov, Lars Grunske, and Andreas Zeller. 2020. Inputs from Hell: Learning Input Distributions for Grammar-Based Test Generation. *IEEE Transactions on Software Engineering* (2020), 1–1. https://doi.org/10.1109/TSE.2020.3013716

[40] William Underwood. 2012. Grammar-Based Specification and Parsing of Binary File Formats. *International Journal of Digital Curation* 7 (03 2012), 95–106. https://doi.org/10.2218/ijdc.v7i1.217

[41] Junjie Wang, Bihuan Chen, Lei Wei, and Yang Liu. 2017. Skyfire: Data-driven seed generation for fuzzing. IEEE, 579–594.

[42] Junjie Wang, Bihuan Chen, Lei Wei, and Yang Liu. 2019. Superion: grammar-aware greybox fuzzing. In *Proceedings of the 41st International Conference on Software Engineering*. IEEE Press, 724–735.

[43] Ming-Hung Wang, Han-Chi Wang, You-Ru Chen, and Chin-Laung Lei. 2017. Automatic Test Pattern Generator for Fuzzing Based on Finite State Machine. *Security and Communication Networks* 2017 (2017).

[44] David HD Warren, Luis M Pereira, and Fernando Pereira. 1977. Prolog-the language and its implementation compared with Lisp. *ACM SIGPLAN Notices* 12, 8 (1977), 109–115.

[45] Jingbo Yan, Yuqing Zhang, and Dingning Yang. 2013. Structurized grammar-based fuzz testing for programs with highly structured inputs. *Security and Communication Networks* 6, 11 (2013), 1319–1330.

[46] Xuejun Yang, Yang Chen, Eric Eide, and John Regehr. 2011. Finding and understanding bugs in C compilers. In *ACM SIGPLAN Notices*, Vol. 46. ACM, 283–294.

[47] Michał Zalewski. [n.d.]. American Fuzzy Lop. http://lcamtuf.coredump.cx/afl. Accessed October 1, 2016.