Statically Typed String Sanitation Inside a Python

Nathan Fulton Cyrus Omar Jonathan Aldrich
Carnegie Mellon University
Pittsburgh, PA
{nathanfu, comar, aldrich}@cs.cmu.edu

ABSTRACT

Web applications must ultimately command systems like web browsers and database engines using strings. Strings derived from improperly sanitized user input can thus be a vector for command injection attacks. In this paper, we introduce *regular string types*, which classify strings known statically to be in a specified regular language. These types come equipped with common operations like concatenation, substitution and coercion, so they can be used to implement, in essentially a conventional manner, the portions of a web application or web application framework that must directly construct command strings. Simple type annotations at key interfaces can be used to statically verify that sanitization has been performed correctly without introducing redundant run-time checks. We specify this type system as a minimal typed lambda calculus, λ_{RS}.

To be practical, adopting a specialized type system like this should not require the adoption of a new programming language. Instead, we favor extensible type systems: new type system fragments like this should be implemented as libraries atop a mechanism that guarantees that they can be safely composed. We support this by 1) specifying a translation from λ_{RS} to a language containing only strings and regular expressions, then, taking Python as such a language, 2) implement the type system together with the translation as a library using `atlang`, an extensible static type system for Python (being developed by the authors).

1. INTRODUCTION

Command injection vulnerabilities are among the most common and severe security vulnerabilities in modern web applications [11]. They arise because web applications, at their boundaries, control external systems using commands represented as strings. For example, web browsers are controlled using HTML and Javascript sent from a server as a string, and database engines execute SQL queries also sent as strings. When these commands include data derived from user input, care must be taken to ensure that the user cannot subverts the intended command by carefully crafting the data they send. For example, a SQL query constructed using string concatenation exposes a SQL injection vulnerability:

```
'SELECT * FROM users WHERE name="' + name + '"'
```

If a malicious user enters the name `'"; DROP TABLE users --'`, the entire database could be erased.

To avoid this problem, the program must *sanitize* user input. For example, in this case, the developer (or, more often, a framework) might define a function `sanitize` that escapes double quotes and existing backslashes with a backslash, which SQL treats safely. Guaranteeing that user input has already been sanitized like this before it is used to construct a command is challenging. Note that this function is not idempotent, so it should only be called once.

We observe that many such sanitization techniques can be understood using *regular languages* [6]. For example, `name` must be a string in the language described by the regular expression `([^"\]|(\")|(\\))*` – a sequence of characters other than quotation marks and backslashes; these can only appear escaped. This concrete syntax for regular expression patterns can be understood to desugar, in a standard way, to the syntax for regular expressions shown in Figure 1, where $r \cdot r$ is sequencing and $r + r$ is disjunction. We will work with this "core" for simplicity.

In this paper, we present a static type system that tracks the regular language a string belongs to. For example, the output of `sanitize` will be a string in the regular language described by the regular expression above (we describe such a regular language, following convention, as $\mathcal{L}\{r\}$). By leveraging closure and decidability properties of regular languages, the language of a string is tracked through uses of a number of operations, including *replacement* of substrings matching a given regular expression. This makes it simple to implement sanitation procedures like the one just described in an essentially conventional manner. The result is a system where the fact that a string has been correctly sanitized is manifest in its type. Missing calls to sanitization functions are detected statically, and, importantly, so are *incorrectly implemented sanitization functions* (i.e. these functions need not be trusted). These guarantees require run-time checks only when going from less precise to more precise types (e.g. at the edges of the system, where user input has not yet been validated).

We will begin in Sec. 2 by specifying this type system minimally, as a conservative extension of the simply typed lambda calculus called λ_{RS}. This allows us to specify the guarantees that the type system provides precisely. We also formally specify a translation from this calculus to a typed calculus with only standard strings and regular expressions,

$$r \;::=\; \epsilon \mid . \mid a \mid r \cdot r \mid r + r \mid r* \qquad a \in \Sigma$$

Figure 1: Regular expressions over the alphabet Σ.

$$\begin{array}{lll}
\sigma & ::= & \sigma \to \sigma \qquad \text{source types} \\
& \mid & \mathsf{stringin}[r] \\
\\
e & ::= & x \qquad \text{source terms} \\
& \mid & \lambda x.e \\
& \mid & e(e) \\
& \mid & \mathsf{rstr}[s] \qquad s \in \Sigma^* \\
& \mid & \mathsf{rconcat}(e; e) \\
& \mid & \mathsf{rreplace}[r](e; e) \\
& \mid & \mathsf{rcoerce}[r](e) \\
& \mid & \mathsf{rcheck}[r](e; x.e; e)
\end{array}$$

Figure 2: Syntax of λ_{RS}.

intending it as a guide to language implementors interested in building this feature into their own languages. This also demonstrates that no additional space overhead is required.

Waiting for a language designer to build this feature in is unsatisfying in practice. Moreover, we also face a "chicken-and-egg problem": justifying its inclusion into a commonly used language benefits from empirical demonstrations that it is useful, but this is difficult to do if developers have no way to use it in practice. As such, we take the position that a better path forward for the community is to work within a programming language where such type system fragments can be introduced modularly and orthogonally, as libraries.

In Sec. 4, we show how to implement the type system fragment we have specified using `atlang`, an extensible static type system implemented as a library inside Python. `atlang` leverages local type inference to control the semantics of literal forms, so reguar string types can be introduced using string literals without any run-time overhead. Coercions that are known to be safe due to a sublanguage relationship are performed implicitly, also without run-time overhead. This results in a *usably secure* system: working with regular strings differs little from working with standard strings.

We conclude after discussing related work in Sec. 5.

2. REGULAR STRING TYPES, MINIMALLY

In this section, we define a minimal typed lambda calculus with regular string types called λ_{RS}. The syntax of λ_{RS} is specified in Figure 2, its static semantics in Figure 3 and its evaluation semantics in Figure 4. This will serve as the source language for our translation to a calculus with only standard strings and regular expressions, defined in Sec. **??**.

There are two type constructors in λ_{RS}, $\to$ and $\mathsf{stringin}$. Arrow types classify functions, which are introduced via lambda abstraction, $\lambda x.e$, and can be applied, written $e(e)$, in the usual way [5]. Regular string types are of the form $\mathsf{stringin}[r]$, where r is a regular expression. Values of such regular string types take the form $\mathsf{rstr}[s]$, where s is a string (i.e. $s \in \Sigma^\star$, defined in the usual way). The rule S-T-StringIn-I ensures that $s \in \mathcal{L}\{r\}$. The remaining operations on terms of type $\mathsf{stringin}[r]$ preserve this property.

outline of §2 goes here

2.1 The Language of Constrained Strings: λ_{RS}

overview

2.1.1 Concatenation and String Decomposition

2.1.2 Coercion

2.1.3 Replacement

PHP

premier

2.1.4 Definition of Substitution and Replacement for Regular Languages

2.1.5 Safety

2.1.6 The Security Theorem

2.2 A Target Language with a Regular Expression Library: λ_P

2.2.1 Safety

2.2.2 Correctness

2.3 Translation from λ_{RS} to λ_P

2.3.1 Correctness of Translation

2.3.2 Preservation of Security

- S-T-Concat: the type of concatenated regular strings is obtained by concatenating the regular expressions.
- S-T-Replace: fix this descriptionthe term $\mathsf{rreplace}$ evaluates to the result of substituting every string matching r in s_1 with s_2. This operation corresponds to the $\mathsf{str\_replace}$ function of PHP. The type of these expressions is defined in terms of an extra-linguistic $\mathsf{lreplace}$ function, which is defined later in this section. The expression $\mathsf{lreplace}(r; r_1; r_2)$ is obtained by starting with r_1, and replacing any subexpression matching r with r_2.
- S-T-SafeCoerce: fix this description allows only safe coercions between string types by expoiting the decidability of language inclusion.
- S-T-Check: fix this descriptionThe check rules allow casts between strings that cannot be guaranteed at compile time, but might pass a runtime check.
- strcase

Definition 1 (Definition of lhead(r))**.** The function $\mathsf{lhead}(r)$ is defined in terms of the structure of r:

- $\mathsf{lhead}(ar') = a$ where $a \in \Sigma$
- $\mathsf{lhead}(r_1 \cup r_2) = \mathsf{lhead}(r_1) \cup \mathsf{lhead}(r_2)$
- $\mathsf{lhead}(r*) = \mathsf{lhead}(r)$
- $\mathsf{lhead}(.) = a_1 \cup a_2 \cup ... \cup a_n$ for all $a_i \in \Sigma$ where $|\Sigma| = n$
- $\mathsf{lhead}(\epsilon) = \epsilon$.

Definition 2 (Definition of $\mathsf{ltail}(r)$)**.** The function $\mathsf{ltail}(r)$ is defined in terms of $\mathsf{lhead}(r)$. Note that $\mathsf{lhead}(r) = a_1 \cup a_2 \cup ... \cup a_i$. We define $\mathsf{ltail}(r) = \delta_{a_1}(r) \cup \delta_{a_2}(r) \cup ... \cup \delta_{a_i}(r)$ where $\delta_s(r)$ is the Brzozowski derivative of r with respect to s [1].

The premier operation for manipulating strings in λ_{RS} is string substitution, which is a familiar operation to any programmer who has used regular expressions. The replacement operation replaces all instances of a pattern in one string with another string; for instance, $\mathsf{subst}(a|b; a; c) = c$. In order to compute the type resulting from substitution, we also need to compute the result of replacing one language with another inside a given language. Finally, just for convenience, we provide a coerce operation. The introduction of coercion requires handling of runtime errors.

The underlying language λ_P has only one type for strings. We prove that whenever a term is translated from λ_{RS} to λ_P, correctness is preserved. The only exception is in the case of unsafe casts in λ_{RS}, which are unnecessary but are included to demonstrate that the regex library of λ_P may be used to insert dynamic checks whenever even when developers are not careful about using statically checked operations.

A brief outline of this section follows:

- In §2.1-2.3, we define λ_{RS}, λ_P and the translation from one to the other.
- In §2.4 we state some properties about regular expressions which are needed in our correctness proofs.
- In §2.5 we prove type safety for λ_P as well as both type safety and correctness for λ_{RS}.
- In §2.6 we prove that translation preserves the correctness result about λ_{RS}.

2.4 Definition of λ_{RS}

The λ_{RS} system extends the simply-typed lambda calculus with regular expression types

$\boxed{\Psi \vdash e : \sigma}$ $\quad \Psi ::= \emptyset \mid \Psi, x : \sigma$

$$\text{S-T-Var}\ \frac{x : \sigma \in \Psi}{\Psi \vdash x : \sigma} \qquad \text{S-T-Abs}\ \frac{\Psi, x : \sigma_1 \vdash e : \sigma_2}{\Psi \vdash \lambda x.e : \sigma_1 \to \sigma_2}$$

$$\text{S-T-App}\ \frac{\Psi \vdash e_1 : \sigma_2 \to \sigma \qquad \Psi \vdash e_2 : \sigma_2}{\Psi \vdash e_1(e_2) : \sigma} \qquad \text{S-T-Stringin-I}\ \frac{s \in \mathcal{L}\{r\}}{\Psi \vdash \mathsf{rstr}[s] : \mathsf{stringin}[r]}$$

$$\text{S-T-Concat}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r_1] \qquad \Psi \vdash e_2 : \mathsf{stringin}[r_2]}{\Psi \vdash \mathsf{rconcat}(e_1; e_2) : \mathsf{stringin}[r_1 \cdot r_2]}$$

$$\text{S-T-Case}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r] \qquad \Psi \vdash e_2 : \sigma \qquad \Psi, x : \mathsf{stringin}[\mathsf{lhead}(r)], y : \mathsf{stringin}[\mathsf{ltail}(r)] \vdash e_3 : \sigma}{\Psi \vdash \mathsf{strcase}(e_1; e_2; x.y.e_3) : \sigma}$$

$$\text{S-T-Replace}\ \frac{\Psi \vdash e_1 : \mathsf{stringin}[r_1] \qquad \Psi \vdash e_2 : \mathsf{stringin}[r_2] \qquad \mathsf{lreplace}(r; r_1; r_2) = r'}{\Psi \vdash \mathsf{rreplace}[r](e_1; e_2) : \mathsf{stringin}[r']}$$

$$\text{S-T-SafeCoerce}\ \frac{\Psi \vdash e : \mathsf{stringin}[r'] \qquad \mathcal{L}\{r'\} \subseteq \mathcal{L}\{r\}}{\Psi \vdash \mathsf{rcoerce}[r](e) : \mathsf{stringin}[r]}$$

$$\text{S-T-Check}\ \frac{\Psi \vdash e_0 : \mathsf{stringin}[r_0] \qquad \Psi, x : \mathsf{stringin}[r] \vdash e_1 : \sigma \qquad \Psi \vdash e_2 : \sigma}{\Psi \vdash \mathsf{rcheck}[r](e_0; x.e_1; e_2) : \sigma}$$

Figure 3: Typing rules for λ_{RS}. The typing context Ψ is standard.

$\boxed{e \Downarrow e}$

$$\text{S-E-Abs}\ \frac{}{\lambda x.e \Downarrow \lambda x.e} \qquad \text{S-E-App}\ \frac{e_1 \Downarrow \lambda x.e_3 \qquad e_2 \Downarrow e_2' \qquad [e_2'/x]e_3 \Downarrow v}{e_1(e_2) \Downarrow v}$$

$$\text{S-E-RStr}\ \frac{}{\mathsf{rstr}[s] \Downarrow \mathsf{rstr}[s]} \qquad \text{S-E-Concat}\ \frac{e_1 \Downarrow \mathsf{rstr}[s_1] \qquad e_2 \Downarrow \mathsf{rstr}[s_2]}{\mathsf{rconcat}(e_1; e_2) \Downarrow \mathsf{rstr}[s_1 s_2]}$$

$$\text{S-E-Case-}\epsilon\ \frac{e_1 \Downarrow \mathsf{rstr}[\epsilon] \qquad e_2 \Downarrow v_2}{\mathsf{strcase}(e_1; e_2; e_3) \Downarrow v_2}$$

$$\text{S-E-Case-Concat}\ \frac{e_1 \Downarrow \mathsf{rstr}[ps] \qquad e_3 \Downarrow x.y.e_4 \qquad [p/x][s/y]e_4 \Downarrow v}{\mathsf{strcase}(e_1; e_2; e_3) \Downarrow v}$$

$$\text{S-E-Replace}\ \frac{e_1 \Downarrow \mathsf{rstr}[e_1] \qquad e_2 \Downarrow \mathsf{rstr}[s_2] \qquad \mathsf{subst}(r; s_1; s_2) = s}{\mathsf{rreplace}[r](e_1; e_2) \Downarrow \mathsf{rstr}[s]}$$

$$\text{S-E-SafeCoerce}\ \frac{e \Downarrow \mathsf{rstr}[s]}{\mathsf{rcoerce}[r](e) \Downarrow \mathsf{rstr}[s]}$$

$$\text{S-E-Check-Ok}\ \frac{e \Downarrow \mathsf{rstr}[s] \qquad s \in \mathcal{L}\{r\} \qquad [\mathsf{rstr}[s]/x]e_1 \Downarrow v}{\mathsf{rcheck}[r](e; x.e_1; e_2) \Downarrow v}$$

$$\text{S-E-Check-NotOk}\ \frac{e \Downarrow \mathsf{rstr}[s] \qquad s \notin \mathcal{L}\{r\}}{\mathsf{rcheck}[r](e; x.e_1; e_2) \Downarrow e_2}$$

Figure 4: Big step semantics for λ_{RS}

.

$\boxed{\Theta \vdash \iota : \tau}$ $\quad \Theta ::= \emptyset \mid \Theta, x : \tau$

$$\text{P-T-Var}\ \frac{x : \tau \in \Theta}{\Theta \vdash x : \tau} \qquad \text{P-T-Abs}\ \frac{\Theta, x : \tau_1 \vdash \iota_2 : \tau_2}{\Theta \vdash \lambda x.\iota_2 : \tau_1 \to \tau_2}$$

$$\text{P-T-App}\ \frac{\Theta \vdash \iota_1 : \tau_2 \to \tau \qquad \Theta \vdash \iota_2 : \tau_2}{\Theta \vdash \iota_1(\iota_2) : \iota} \qquad \text{P-T-String}\ \frac{}{\Theta \vdash \mathsf{str}[s] : \mathsf{string}}$$

$$\text{P-T-Regex}\ \frac{}{\Theta \vdash \mathsf{rx}[r] : \mathsf{regex}} \qquad \text{P-T-Concat}\ \frac{\Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta \vdash \iota_2 : \mathsf{string}}{\Theta \vdash \mathsf{concat}(\iota_1; \iota_2) : \mathsf{string}}$$

$$\text{P-T-Case}\ \frac{\Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta \vdash \iota_2 : \tau \qquad \Theta, x : \mathsf{string}, y : \mathsf{string} \vdash \iota_3 : \tau}{\Theta \vdash \mathsf{pstrcase}(\iota_1; \iota_2; \iota_3) : \tau}$$

$$\text{P-T-Replace}\ \frac{\Theta \vdash \iota_1 : \mathsf{regex} \qquad \Theta \vdash \iota_2 : \mathsf{string} \qquad \Theta \vdash \iota_3 : \mathsf{string}}{\Theta \vdash \mathsf{preplace}(\iota_1; \iota_2; \iota_3) : \mathsf{string}}$$

$$\text{P-T-Check}\ \frac{\Theta \vdash \iota_r : \mathsf{regex} \qquad \Theta \vdash \iota_1 : \mathsf{string} \qquad \Theta, x : \mathsf{string} \vdash \iota_2 : \sigma \qquad \Theta \vdash \iota_3 : \sigma}{\Theta \vdash \mathsf{check}(\iota_r; \iota_1; x.\iota_2; \iota_3) : \sigma}$$

Figure 5: Typing rules for λ_P. The typing context Θ is standard.

2.5 Definition of λ_P

The system λ_P is a straight-forward extension of a simply typed lambda calculus with a string type and a regular expression tpye. We include two operations which are available in the regular expression library of any modern programming language. The check operation ensures that an expression recognizes a string, and the replace operation is string replacement.

2.6 Definition of Translation

The translation from λ_{RS} to λ_P is defined in figure 5. The coercion cases are most interesting. If the safety of coercion in manifest in the types of the expressions, then no runtime check is inserted. If the safety of coercion is not manifest in the types, then a check is inserted.

In practice, the type of a replacement rarely matches a specification. Therefore, it is convienant in an implementation to always insert the appropriate coercion, and then only raise type errors when an automatically inserted coercion actually requires the insertion of a runtime check. Alternatively, this policy may be codified in the type system itself using subtyping [3].

2.7 Properties of Regular Languages

Our type safety proof for language S replies on a relationship between string substitution and language substitution given in lemma 7. We also rely upon several other properties of regular languages. Throughout this section, we fix an alphabet Σ over which strings s and regular expressions r are defined. throughout the paper, $\mathcal{L}\{r\}$ refers to the language recognized by the regular expression r. This distinction between the regular expression and its language – typically elided in the literature – makes our definition and proofs about systems S and P more readable.

$\boxed{\iota \Downarrow \iota}$

$$\text{P-E-Abs}\ \frac{}{\lambda x.e \Downarrow \lambda x.e} \qquad \text{P-E-App}\ \frac{\iota_1 \Downarrow \lambda x.\iota_3 \qquad \iota_2 \Downarrow \iota_2' \qquad [\iota_3/x]\iota_2' \Downarrow v}{\iota_1(\iota_2) \Downarrow v}$$

$$\text{P-E-Str}\ \frac{}{\mathsf{str}[s] \Downarrow \mathsf{str}[s]} \qquad \text{P-E-Rx}\ \frac{}{\mathsf{rx}[r] \Downarrow \mathsf{rx}[r]} \qquad \text{P-E-Concat}\ \frac{\iota_1 \Downarrow \mathsf{str}[s_1] \qquad \iota_2 \Downarrow \mathsf{str}[s_2]}{\mathsf{concat}(\iota_1;\iota_2) \Downarrow \mathsf{str}[s_1 s_2]}$$

$$\text{P-E-Case-}\epsilon\ \frac{\iota_1 \Downarrow \mathsf{str}[\,] \qquad \iota 2 \Downarrow \iota_2}{\mathsf{pstrcase}(\iota_1;\iota_2;\iota_3) \Downarrow v_2}$$

$$\text{P-E-Case-Concat}\ \frac{\iota_1 \Downarrow \mathsf{str}[ps] \qquad x.y.\iota_3 \Downarrow \iota_4 \qquad [p/x][s/y]\iota_4 \Downarrow v}{\mathsf{pstrcase}(\iota_1;\iota_2;\iota_3) \Downarrow v}$$

$$\text{P-E-Replace}\ \frac{\iota_1 \Downarrow \mathsf{rx}[r] \qquad \iota_2 \Downarrow \mathsf{str}[s_2] \qquad \iota_3 \Downarrow \mathsf{str}[s_3] \qquad \mathsf{subst}(r;s_2;s_3) = s}{\mathsf{preplace}(\iota_1;\iota_2;\iota_3) \Downarrow \mathsf{str}[s]}$$

$$\text{P-E-Check-OK}\ \frac{\iota \Downarrow \mathsf{str}[s] \qquad s \in \mathcal{L}\{r\} \qquad [\mathsf{str}[s]/x]\iota_1 \Downarrow \iota_3}{\mathsf{check}(\mathsf{rx}[r];\iota;x.\iota_1;\iota_2) \Downarrow \iota_3}$$

$$\text{P-E-Check-NotOK}\ \frac{\iota \Downarrow \mathsf{str}[s] \qquad s \notin \mathcal{L}\{r\}}{\mathsf{check}(\mathsf{rx}[r];\iota;x.\iota_1;\iota_2) \Downarrow \iota_2}$$

Figure 6: Big step semantics for of λ_P

.

$$\begin{array}{lll} \theta & ::= & \theta \to \theta \qquad \text{target types} \\ & | & \mathsf{string} \\ & | & \mathsf{regex} \\ \\ \mathrm{P} & ::= & x \qquad \text{target terms} \\ & | & \lambda x.\iota \\ & | & \iota\iota \\ & | & \mathsf{str}[s] \\ & | & \mathsf{rx}[r] \\ & | & \mathsf{concat}(\iota;\iota) \\ & | & \mathsf{preplace}(\iota;\iota;\iota) \\ & | & \mathsf{check}(\mathsf{rx}[r];\iota;\iota;\iota) \end{array}$$

Figure 7: Syntax for the target language, λ_P, containing strings and statically constructed regular expressions.

Lemma 3 (Properties of Regular Languages and Expressions.)**.** *The following are properties of regular expressions which are necessary for our proofs: If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then $s_1 s_2 \in \mathcal{L}\{r_1 r_2\}$. For all strings s and regular expressions r, either $s \in \mathcal{L}\{r\}$ or $s \notin \mathcal{L}\{r\}$. Regular languages are closed under difference, right quotient, reversal, and string homomorphism.*

If any of these properties are unfamiliar, the reader may refer to a standard text on the subject [6].

Definition 4 (`subst`)**.** The relation $\mathsf{subst}(r;s_1;s_2) = s$ produces a string s in which all substrings of s_1 matching r are replaced with s_2.

Definition 5 (`lreplace`)**.** The relation $\mathsf{lreplace}(r;r_1;r_2) =$

$\boxed{\Psi \vdash [\![e]\!] = \iota}$

$$\text{Tr-Var}\ \frac{}{\Psi \vdash [\![x]\!] = x} \qquad \text{Tr-Abs}\ \frac{\Psi \vdash [\![e]\!] = \iota}{\Psi \vdash [\![\lambda x.e]\!] = \lambda x.\iota}$$

$$\text{Tr-App}\ \frac{\Psi \vdash [\![e_1]\!] = \iota_1 \qquad \Psi \vdash [\![e_2]\!] = \iota_2}{\Psi \vdash [\![e_1(e_2)]\!] = \iota_1(\iota_2)}$$

$$\text{Tr-Case}\ \frac{\Psi \vdash [\![e_1]\!] = \iota_1 \qquad \Psi \vdash [\![e_2]\!] = \iota_2 \qquad \Psi \vdash [\![e_3]\!] = \iota_3}{\Psi \vdash [\![\mathsf{strcase}(e_1;e_2;e_3)]\!] = \mathsf{pstrcase}(\iota_1;\iota_2;\iota_3)}$$

$$\text{Tr-string}\ \frac{}{\Psi \vdash [\![\mathsf{rstr}[s]]\!] = \mathsf{str}[s]}$$

$$\text{Tr-Concat}\ \frac{\Psi \vdash [\![e_1]\!] = \iota_1 \qquad \Psi \vdash [\![e_2]\!] = \iota_2}{\Psi \vdash [\![\mathsf{rconcat}(e_1;e_2)]\!] = \mathsf{concat}(\iota_1;\iota_2)}$$

$$\text{Tr-Subst}\ \frac{\Psi \vdash [\![e_1]\!] = \iota_1 \qquad \Psi \vdash [\![e_2]\!] = \iota_2}{\Psi \vdash [\![\mathsf{rreplace}[r](e_1;e_2)]\!] = \mathsf{replace}(\mathsf{rx}[r];\iota_1;\iota_2)}$$

$$\text{Tr-SafeCoerce}\ \frac{\Psi \vdash [\![e]\!] = \iota}{\Psi \vdash [\![\mathsf{rcoerce}[r'](e)]\!] = \iota}$$

$$\text{Tr-Check}\ \frac{\Psi \vdash [\![e]\!] = \iota_1 \qquad \Psi \vdash [\![e_1]\!] = \iota_1?? \qquad \Psi \vdash [\![e_2]\!] = \iota_2}{\Psi \vdash [\![\mathsf{rcheck}[r](e;x.e_1;e_2)]\!] = \mathsf{check}(\mathsf{rx}[r];\iota;x.\iota_1;\iota_2)}$$

Figure 8: Translation from source terms (e) to target terms (ι). The translation is type-directed.

r' relates r, r_1, and r_2 to a language r' containing all strings of r_1 except that any substring $s_{pre}ss_{post} \in \mathcal{L}\{r_1\}$ where $s \in \mathcal{L}\{r\}$ is replaced by the set of strings $s_{pre}s_2s_{post}$ for all $s_2 \in \mathcal{L}\{r_2\}$ (the prefix and postfix positions may be empty).

Proposition 6 (Closure.)**.** *If $\mathcal{L}\{r\}, \mathcal{L}\{r_1\}$ and $\mathcal{L}\{r_2\}$ are regular expressions, then $\mathcal{L}\{\mathsf{lreplace}(r;r_1;r_2)\}$ is also a regular language.*

Proof. The theorem follows from closure under difference, right quotient, reversal and string homomorphism. □

Proposition 7 (Substitution Correspondence.)**.** *If $s_1 \in \mathcal{L}\{r_1\}$ and $s_2 \in \mathcal{L}\{r_2\}$ then $\mathsf{subst}(r;s_1;s_2) \in \mathcal{L}\{\mathsf{lreplace}(r;s_1;s_2)\}$.*

Proof. The theorem follows from the definitions of subst and lreplace; note that language substitutions over-approximate string substitutions. □

2.8 Safety of the Source and Target Languages

In this section, we establish type safety for the source (λ_{RS}) and target (λ_P) languages. In addition to type safety, we also prove a stronger correctness property for λ_{RS}.

Our first two theorems establish that our rules preserve the well-formedness of regular expressions. The standard

lemmas required for safety of the simply typed lambda claculus are also required; proofs for these do not differ substntially from [5].

Lemma 8. *If* $\Psi \vdash e : \mathsf{stringin}[r]$ *then* r *is a well-formed regular expression.*

Proof. The only non-trivial case is S-T-Replace, which follows from lemma 6. □

Lemma 9. *If* $\Theta \vdash \iota : \mathsf{regex}$ *then* $\iota \Downarrow \mathsf{rx}[r]$ *such that* r *is a well-formed regular expression.*

We now prove safety for the string fragment of the source and target languages. The easiest property is type safety of λ_P, which follows almost directly from type safety for the simply typed lambda calculus.

Theorem 10. *Let* ι *be a term in the target language. If* $\Theta \vdash \iota : \tau$ *then* $\iota \Downarrow \iota'$ *and* $\Theta \vdash \iota' : \tau$.

Safety for the string fragment of λ_{RS} is more involved because it involves validating that the type system's definition is justified by our theorems about regular languages. We state the theorem for the string fragment only to demontrate this point; full type safety for λ_{RS} follows easily from safety for the string fagment. Note that type safety does *not* guarantee that strings of a type are in the correct language. We isolate this property so that we may reason about its preservation under translation to λ_P, which does not have regular expression types.

Theorem 11 (Type Safety for the String Fragment of λ_{RS}.)**.** *If* $\Psi \vdash e : \sigma$ *then* $e \Downarrow e'$ *and* $\Psi \vdash e' : \mathsf{stringin}[r]$.

Proof. By induction on the typing relation. The S-T-Concat case requires Lemma 3 and the S-T-Replace case appeals to Lemma 7. □

Theorem 12 (Canonical Forms for String Fragment of λ_{RS}.)**.** *If* $\Psi \vdash e : \mathsf{stringin}[r]$ *then* $e \Downarrow \mathsf{rstr}[s]$

In addition to safety, λ_{RS} requires a correctness result ensuring that well-typed terms of regular string type are in the language associated with their type.

Theorem 13 (Correctness of Input Sanitation for λ_{RS}.)**.** *If* $\Psi \vdash e : \mathsf{stringin}[r]$ *and* $e \Downarrow \mathsf{rstr}[s]$ *then* $s \in \mathcal{L}\{r\}$.

Proof. Follows directly from type safety, canonical forms for λ_{RS}, and inversion of the typing relation for λ_{RS}. □

2.9 Translation Correctness

The main theorem of this paper is Theorem 15, which establishes that Theorem 13 is preserved under translation into the target language λ_P.

Establishing this result requires an additional theorem establishing a relationship between canonical forms for the string fragments of λ_{RS} and λ_P.

Theorem 14 (Translation Correctness.)**.** *If* $\Psi \vdash e : \mathsf{stringin}[r]$ *then there exists an* ι *such that* $[\![e]\!] = \iota$, $\iota \Downarrow \mathsf{str}[s]$, *and* $e \Downarrow \mathsf{rstr}[s]$.

Proof. The proof proceeds by induction on the typing relation for e and an appropriate choice of ι; in each case, the choice is obvious and the rest of the proof proceeds by our type safety theorems as well as an appeal to the induction hypothesis. □

Finally, our main theorem establishes that input sanitation correctness of λ_{RS} is preserved under the translation into λ_P.

Theorem 15 (Correctness of Input Sanitation for Translated Terms.)**.** *If* $[\![e]\!] = \iota$ *and* $\Psi \vdash e : \mathsf{stringin}[r]$ *then* $\iota \Downarrow \mathsf{str}[s]$ *for* $s \in \mathcal{L}\{r\}$.

Proof. By theorem 14, $\iota \Downarrow \mathsf{str}[s]$ implies that $e \Downarrow \mathsf{rstr}[s]$. By theorem 13, the above property together with the assumption that e is well-typed implies that $s \in \mathcal{L}\{r\}$. □

3. IMPLEMENTATION IN ATLANG

3.1 An Example

4. IMPLEMENTATION IN ATLANG

In the previous section, we specified a type system and a translation semantics to a language containing only strings and regular expressions. In this section, we take Python to be such a target language. Python does not have a static type system, however, so to implement these semantics, we will leverage `atlang`, an extensible type system for Python (being developed by the authors). By using `atlang`, which builds on Python's quotations and reflection facilities, we can implement these semantics as a library, rather than as a new dialect of Python.

```
1  from atlib import fn, string_in
2
3  @fn
4  def sanitize(s : string_in[r'.*']):
5    return (s.replace(r'"', '&quot;')
6             .replace(r'<', '&lt;')
7             .replace(r'>', '&gt;'))
8
9  @fn
10 def results_query(s : string_in[r'[^"]*']):
11   return 'SELECT * FROM users WHERE name="' + s + '"'
12
13 @fn
14 def results_div(s : string_in[r'[^<>]*']):
15   return '<div>Results for ' + s + '</div>'
16
17 @fn
18 def main(db):
19   input = sanitize(user_input())
20   results = db.execute(results_query(input))
21   return results_div(input) + format(results)
```

Figure 9: Regular string types in atlang, a library that enables static type checking for Python.

Figure **??** demonstrates the use of string types in atlang. The sanitize function takes an arbitrary string and returns a string without double quotes or left and right brackets. In this example, we use HTML escape sequences.

The main function receives user input and passes this input to a sanitize function, which replaces all double quotes and brackets with HTML escape sequences.

The result of applying sanitize to input is appended to two functions which construct a safe query and safe output. The arguments to the result and output construction functions constitute *specifications*. In the case of `results_query`, this specification ensures that user input is always interpreted as a string literal by the SQL server. In the case of `results_div`, this specification ensures that user input does not contain

any HTML tags, which is a conservative but effective policy for preventing XSS attacks.

Note that input does not actually meet these specifications without additional machinery. The type of input is quite large and does not actually equal the specified domains of the query or output construction methods. This mismatch is common – in fact, nearly universal. Therefore, our implementation includes a simple subtyping relation between regular expression types. This subtyping relation is justified theoretically by the fact that language inclusion is decidable; see [3] for a formal definition of the subtyping relation. Additionally, our extension remains composable because subtyping is defined on a type-by-type basis; see [4] for a discussion of subtyping in Atlang (referred to there as Ace).

4.1 Implementation of the Regular Expression Type

4.2 Example Usage

Figure 9 demonstrates the use of two type constructors, `fn` and `string_in`, both of which we have included in `atlib`, the standard library for `atlang`. The `fn` type constructor can be used to annotate functions that should be statically checked by `atlang`. The function `sanitize` on lines 3-7 specifies one argument, `s`, of type `string_in[r'.*']`.fix sizes

The sanitize function takes an arbitrary string and returns a string without double quotes or left and right brackets. In this example, we use HTML escape sequences.

The main function receives user input and passes this input to a sanitize function, which replaces all double quotes and brackets with HTML escape sequences.

The result of applying sanitize to input is appended to two functions which construct a safe query and safe output. The arguments to the result and output construction functions constitute *specifications*. In the case of results_query, this specification ensures that user input is always interpreted as a string literal by the SQL server. In the case of results_div, this specification ensures that user input does not contain any HTML tags, which is a conservative but effective policy for preventing XSS attacks.

Note that input does not actually meet these specifications without additional machinery. The type of input is quite large and does not actually equal the specified domains of the query or output construction methods. This mismatch is common – in fact, nearly universal. Therefore, our implementation includes a simple subtyping relation between regular expression types. This subtyping relation is justified theoretically by the fact that language inclusion is decidable; see [3] for a formal definition of the subtyping relation. Additionally, our extension remains composable because subtyping is defined on a type-by-type basis; see [4] for a discussion of subtyping in Atlang (referred to there as Ace).

4.3 Implementation of the Regular Expression Type

We implemented a variation on the type system presented in this paper. The only significant difference is that we only support replacements where s_2 is the empty string. Therefore, our implementation respects the system presented in section 2 only modulo the definition of lreplace.

Atlang translates programs using type definitions, which may extend both the static and dynamic semantics of the language. New types are defined as Python classes; figure 4.1 contains the source code of our implementation.

```
1  class string_in(atlang.Type):
2    def __init__(self, rx):
3      rx = rx_normalize(rx)
4      atlang.Type.__init__(idx=rx)
5
6    def ana_Str(self, ctx, node):
7      if not in_lang(node.s, self.idx):
8        raise atlang.TypeError("...", node)
9
10   def trans_Str(self, ctx, node):
11     return astx.copy(node)
12
13   def syn_BinOp_Add(self, ctx, node):
14     left_t = ctx.syn(node.left)
15     right_t = ctx.syn(node.right)
16     if isinstance(left_t, string_in):
17       left_rx = left_t.idx
18       if isinstance(right_t, string_in):
19         right_rx = right_t.idx
20         return string_in[lconcat(left_rx, right_rx)]
21     raise atlang.TypeError("...", node)
22
23   def trans_BinOp_Add(self, ctx, node):
24     return astx.copy(node)
25
26   def syn_Method_replace(self, ctx, node):
27     [rx, exp] = node.args
28     if not isinstance(rx, ast.Str):
29       raise atlang.TypeError("...", node)
30     rx = rx.s
31     exp_t = ctx.syn(exp)
32     if not isinstance(exp_t, string_in):
33       raise atlang.TypeError("...", node)
34     exp_rx = exp_t.idx
35     return string_in[lreplace(self.idx, rx, exp_rx)]
36
37   def trans_Method_replace(self, ctx, node):
38     return astx.quote(
39       """__import__(re); re.sub(%0, %1, %2)""",
40       astx.Str(s=node.args[0]),
41       astx.copy(node.func.value),
42       astx.copy(node.args[1]))
43
44   def syn_Method_check(self, ctx, node):
45     [rx] = node.args
46     if not isinstance(rx, ast.Str):
47       raise atlang.TypeError("...", node)
48     return string_in[rx.s]
49
50   def trans_Method_check(self, ctx, node):
51     return astx.quote(
52       """__import__(string_in_helper);
53       string_in_helper.coerce(%0, %1)""",
54       astx.Str(s=other_t.idx),
55       astx.copy(node))
56
57   def check_Coerce(self, ctx, node, other_t):
58     # coercions can only be defined between
59     # types with the same type constructor,
60     if rx_sublang(other_t.idx, self.idx):
61       return other_t
62     else: raise atlang.TypeError("...", node)
```

Figure 10: Implementation of the `string_in` type constructor in atlang.

The string_in type has an indexing regular expression `idx`. Our translation is defined by the trans_ methods while the syn_ methods define our type checker. Atlang defers type checking and translation to these methods whenever an expression of type string_in is encountered.

5. RELATED WORK AND ALTERNATIVE APPROACHES

The input sanitation problem is well-understood. There exist a large number of techniques and technologies, proposed by both practitioners and researchers, for preventing injection-style attacks. In this section, we explain how our approach to the input sanitation problem differs from each of these approaches. More important than these differences, however, is our more general assertion that language extensibility is a promising approach toward consideration of security goals in programming language design.

Unlike *frameworks and libraries* provided by languages such as Haskell and Ruby, our type system provides a *static* guarantee that input is always properly sanitized before use. Doing so requires reasoning about the operations on regular languages corresponding to standard operations on strings; we are unaware of any production system which contains this form of reasoning. Therefore, even where frameworks and libraries provide a viable interface or wrapper around input sanitation, our approach is complementary because it ensures the correctness of the framework or library itself. Furthermore, our approach is more general than database abstraction layers because our mechanism is applicable to all forms of command injection (e.g. shell injection or remote file inclusion).

A number of research languages provide static guarantees that a program is free of input sanitation vulnerabilities [2]. Unlike this work, our solution to the input sanitation problem has a very low barrier to adoption; for instance, our implementation conservatively extends Python – a popular language among web developers. We also believe our general approach is better-positioned for security, where continuously evolving threats might require frequent addition of new analyses; in these cases, the composability and generality of our approach is a substantial advantage.

The Wyvern programming language provides a general framework for composing language extensions [10][9]. Our work identifies one particular extension, and is therefore complementary to Wyvern and related work on extensible programming languages. We are also unaware of any extensible programming languages which emphasize applications to security concerns.

Incorporating regular expressions into the type system is not novel. The XDuce system [8, 7] checks XML documents against schema using regular expressions. Similarly, XHaskell [12] focuses on XML documents. We differ from this and related work in at least three ways:

- Our system is defined within an extensible type system.
- We demonstrate that regular expression types are applicable to the web security domain, whereas previous work on regular expression types focused on XML schema.
- Although our static replacement operation is definable in some languages with regular expression types, we are the first to expose this operation and connect the semantics of regular language replacement with the semantics of string substitution via a type safety and compilation correctness argument.

In conclusion, our contribution is a type system, implemented within an extensible type system, for checking the correctness of input sanitation algorithms.

6. CONCLUSION

Composable analyses which complement existing approaches constitute a promising approach toward the integration of security concerns into programming languages. In this paper, we presented a system with both of these properties and defined a security-preserving transformation. Unlike other approaches, our solution complements existing, familiar solutions while providing a strong guarantee that traditional library and framework-based approaches are implemented and utilized correctly.

7. REFERENCES

[1] J. A. Brzozowski. Derivatives of regular expressions. *J. ACM*, 11(4):481–494, Oct. 1964.

[2] A. Chlipala. Static checking of dynamically-varying security policies in database-backed applications. In *OSDI'10: Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation*, Oct. 2010.

[3] N. Fulton. Security through extensible type systems. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, pages 107–108, New York, NY, USA, 2012. ACM.

[4] N. Fulton. A typed lambda calculus for input sanitation. Undergraduate thesis in mathematics, Carthage College, 2013.

[5] R. Harper. *Practical Foundations for Programming Languages*. Cambridge University Press, 2012.

[6] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.

[7] H. Hosoya and B. C. Pierce. XDuce: A statically typed XML processing language. *ACM Transactions on Internet Technology*, 3(2):117–148, May 2003.

[8] H. Hosoya, J. Vouillon, and B. C. Pierce. Regular Expression Types for XML. In *ICFP '00*, 2000.

[9] L. Nistor, D. Kurilova, S. Balzer, B. Chung, A. Potanin, and J. Aldrich. Wyvern: A simple, typed, and pure object-oriented language. In *Proceedings of the 5th Workshop on MechAnisms for SPEcialization, Generalization and inHerItance*, MASPEGHI '13, pages 9–16, New York, NY, USA, 2013. ACM.

[10] C. Omar, D. Kurilova, L. Nistor, B. Chung, A. Potanin, and J. Aldrich. Safely composable type-specific languages. In R. Jones, editor, *ECOOP 2014 âĂŞ Object-Oriented Programming*, volume 8586 of *Lecture Notes in Computer Science*, pages 105–130. Springer Berlin Heidelberg, 2014.

[11] OWASP. Open web application security project top 10.

[12] M. Sulzmann and K. Lu. Xhaskell – adding regular expression types to haskell. In O. Chitil, Z. HorvÃąth, and V. ZsÃşk, editors, *Implementation and Application of Functional Languages*, volume 5083 of *Lecture Notes in Computer Science*, pages 75–92. Springer Berlin Heidelberg, 2008.